DISSERTATION

NEW DEVELOPMENTS ON LINEAR REGRESSION WITH RANDOM DESIGN AND HIGH-DIMENSIONAL MEDIATION ANALYSIS

Submitted by

Zifeng Zhang

Department of Statistics

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2025

Doctoral Committee:

Advisor: Haonan Wang
Co-Advisor: Wen Zhou

Piotr Kokoszka
F. Jay Breidt
Jie Luo

ABSTRACT

NEW DEVELOPMENTS ON LINEAR REGRESSION WITH RANDOM DESIGN AND HIGH-DIMENSIONAL MEDIATION ANALYSIS

Linear regression is arguably the most widely used statistical method. In this thesis, we study the robustness of the least squares estimator when regressors are random and the errors are correlated with unknown correlation structure. We further investigate the small-sample robustness of the least squares estimator and offer a new geometric perspective on the F-test. Motivated by the Baron-Kenny approach, we also apply linear models to high-dimensional mediation analysis with the treatment-by-mediator interaction.

In linear regression with fixed regressors and correlated errors, the conventional wisdom is to modify the variance-covariance estimator to accommodate the known correlation structure of the errors. We depart from the literature by showing that with random regressors, linear regression inference is robust to correlated errors with unknown correlation structure. The existing theoretical analyses for linear regression are no longer valid because even the asymptotic normality of the least-squares coefficients breaks down in this regime. We first prove the asymptotic normality of the t statistics by establishing their Berry–Esseen bounds based on a novel probabilistic analysis of self-normalized statistics. We then study the local power of the corresponding t tests and show that, perhaps surprisingly, error correlation can even enhance power in the regime of weak signals. Overall, our results show that linear regression is applicable more broadly than the conventional theory suggests, and further demonstrate the value of randomization to ensure robustness of inference.

Next, we explore the small sample robustness of the least squares estimator by the F and t tests. The F distribution is one of the most widely applied statistical tools in small sample inference, and it has been recognized that its definition does not necessarily require

normality, but merely a spherical distribution. While existing literature has touched upon the relationship between the F and spherical distributions, these discussions remain either incomplete or not rigorously structured. We provide a geometric perspective that clearly delineates the relationship between F and spherical distributions, and introduce a novel definition of the F distribution. Perhaps surprisingly, based on this new definition, in the linear model, the validity of the ordinary least squares F-test and t-test is preserved under spherical symmetry of the design matrix, even if the error terms have non-zero means, heteroscedasticity, strong correlations, or heavy tails.

Finally, we apply the linear model (Baron-Kenny approach) in the mediation analysis under high-dimensional setting with interaction. Mediation analysis has been commonly applied in various fields, including economics, finance, and genomic and genetic research. A key challenge in this domain is the inference of natural direct and indirect effects in the presence of potential interactions between treatment and high-dimensional mediators. These interactions often give rise to moderator effects, which are further complicated by the intricate dependencies among the mediators. In this paper, we introduce a new inference procedure that addresses this challenge. By incorporating a non-convex penalty into the outcome model, our method effectively identifies important mediators while accounting for their interactions with the treatments, which admits the guaranteed oracle property. Leveraging the oracle property, we can exploit a projection onto the mediator model, guided by the estimated important direction in the mediator space. We establish the asymptotic normality of both natural indirect and direct effects for inference. Additionally, we develop an algorithm that utilizes the overlapping group SCAD penalty to promote heredity structure among the main effects and interactions, which comes with provable guarantees. Our extensive numerical studies, comparing our method with other existing approaches across various scenarios, demonstrate its effectiveness. To illustrate the practical application of our methods, we conduct a study investigating the impact of childhood trauma on cortisol

stress reactivity. Using DNA methylation loci as mediators, we uncover several new loci that remain undetected when interactions are ignored.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisors, Prof. Wen Zhou and Prof. Haonan Wang, for their continuous support and guidance throughout my Ph.D. journey. Their patience, encouragement, and immense knowledge have been instrumental to my academic and personal growth. Without their help, I would never have been able to finish this work. I am also sincerely thankful to my collaborators, Prof. Peng Ding, Prof. Fan Yang, and Prof. Lihua Lei. It has been a great honor to learn from and collaborate with them. Their insightful guidance and generous support have been a constant source of confidence and inspiration in my academic pursuits.

I am especially grateful to my committee members, Prof. Piotr Kokoszka, Prof. F. Jay Breidt, and Prof. Jie Luo, for their continuous interest in my research and for generously offering their time, guidance, and thoughtful feedback during the preparation and review of this dissertation.

I am grateful to the Department of Statistics at Colorado State University for providing financial support during my Ph.D. studies. I have greatly enjoyed my time in Fort Collins and deeply appreciate the faculty and staff for their kind assistance and support over the years.

I would also like to thank my friends and fellow cohorts — Xiangdong, Hongyang, Hanxiao, Wenqin, Mengting, Lyuou, Yao, Bingying, Austin, Justin, Lane, Maddie, and many others — for their help both in my studies and in daily life.

Finally, I am forever indebted to my parents for their unconditional love, support, and encouragement. Their guidance and belief in me have been a constant source of strength throughout this journey.

DEDICATION

I would like to dedicate this dissertation to my family.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Linear regression $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ is arguably the quintessential practice in statistics dating back to Gauss and Legendre in the nineteenth century. The fundamental method of ordinary least squares (OLS)

$$\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \tag{1.0.1}$$

has been widely employed for two centuries across numerous scientific disciplines. See Plackett (1949); Seal (1967) for a historical review and Seber and Lee (2003); Cameron and Trivedi (2005); Hayashi (2011); MacKinnon (2012); Cohen, Cohen, West, and Aiken (2013); Kleinbaum, Kupper, Nizam, and Rosenberg (2013); Montgomery, Peck, and Vining (2021) for a variety of applications. From the well-known Gauss-Markov Theorem (Gauss, 1855; Markov, 1912; David & Neyman, 1938), OLS is the best linear unbiased estimator (BLUE) if the observation errors are uncorrelated and have equal variances. However, correlated noise is ubiquitous, especially when the data are collected in a temporal domain, a spatial domain, a social network, or are from repeated measures, etc.

Although OLS remains unbiased in the presence of correlated noise, the estimation may be inefficient (Taylor, 1977; Ali & Ponnapalli, 1990; Koreisha & Fang, 2001; Kariya & Kurata, 2004). There has been a plethora of methods dealing with different correlation structures among observation errors. Ideally, when the correlation structure in $\boldsymbol{\varepsilon}$ is known, the generalized least squares (GLS) is BLUE (Aitken, 1936) with an extension for singular $\mathbb{C}\mathrm{ov}(\boldsymbol{\varepsilon})$ in Zyskind and Martin (1969). In practice, $\mathbb{C}\mathrm{ov}(\boldsymbol{\varepsilon})$ is usually unknown so the GLS is not immediately applicable. A number of alternative approaches have evolved, in which some estimated $\mathbb{C}\mathrm{ov}(\boldsymbol{\varepsilon})$ replaces the true yet unknown error covariance ,e.g., two-stage Aitken estimator.

Under serial correlation in $\mathbb{C}\mathrm{ov}(\boldsymbol{\varepsilon})$, the most widely used statistic for testing AR(1) autoregressive process residuals is perhaps the DW-statistic originated from Durbin and Watson (1950, 1951), and modified in Durbin (1970). Later the DW test was shown to be asymptotically equivalent to Lagrange multiplier test see Aldrich (1978); Godfrey (1978); Breusch and Pagan (1980), which allows for higher order serial correlation. After the test of autocorrelation, Cochrane-Orcutt (CO) type of estimation (Cochrane & Orcutt, 1949; Prais & Winsten, 1954; Beach & MacKinnon, 1978) can be applied for feasible GLS. See Sections 12.2–12.5 in Wooldridge (2012) for implementation. Besides modeling the serial correlation, one can also apply the heteroskedasticity and autocorrelation consistent (HAC) standard error adjustment for OLS from Newey and West (1987); Andrews (1991).

In the presence of spatial correlation in $\mathbb{C}\mathrm{ov}(\boldsymbol{\varepsilon})$, a simple test for the existence of spatial correlation is the Moran's I (MI) statistic from Moran (1950); Cliff and Ord (1972) or Lagrange multiplier test from Burridge (1980). Different estimation methods are proposed such as maximum likelihood estimation for different spatial regression models (Besag, 1974; Ord, 1975; K. V. Mardia & Marshall, 1984; Warnes & Ripley, 1987; K. Mardia & Watkins, 1989; Case, 1991), generalized method of moments (Conley, 1999; Kelejian & Prucha, 1999), or Bayesian hierarchical models (Banerjee, Gelfand, Finley, & Sang, 2008; Banerjee, Carlin, & Gelfand, 2014). Some recent works discuss the standard error adjustment such as Barrios, Diamond, Imbens, and Kolesár (2012); Müller and Watson (2022a, 2022b). For more details, see Anselin (2022) or textbooks Anselin (1988); Gelfand, Diggle, Guttorp, and Fuentes (2010); Møller (2013); Cressie (2015); Schabenberger and Gotway (2017).

If clusters of units are sampled inducing a block-diagonal correlation structure in $\mathbb{C}\mathrm{ov}(\boldsymbol{\varepsilon})$ (Campbell, 1977; Scott & Holt, 1982; Davis, 2002; Schochet, Pashley, Miratrix, & Kautz, 2022), different estimations are obtained by imposing certain parametric assumptions, such as the maximum likelihood estimator for the model of repeated measurements in Diggle (1988), the feasible GLS for the random effects model in Section 14.2 of Wooldridge (2012), and for fixed effects panel and multilevel models with autocorrelation in Hansen (2007).

Beyond estimation, robust standard error adjustments for the OLS inference such as Kloek (1981); Liang and Zeger (1986); Arellano et al. (1987); Bester, Conley, and Hansen (2011); Cameron, Gelbach, and Miller (2011) are proposed with discussions in Moulton (1986); Barrios et al. (2012); Abadie, Athey, Imbens, and Wooldridge (2017); Su and Ding (2021).

If we impose certain invariance assumption on errors, randomization-based procedures can be applied. Typically under exchangeable errors, Freedman and Lane (1983); M. J. Anderson (2001); Winkler, Ridgway, Webster, Smith, and Nichols (2014) proposed and compared different permutation tests on regression residuals in the spirit of E. J. Pitman (1937); E. J. G. Pitman (1938). We refer to Section A.2 of Lei and Bickel (2021) for a thorough literature review of the permutation test. Particularly, under normally distributed errors with exchangeability, Halperin (1951); McElroy (1967); Arnold (1979) discussed the performance of OLS procedure. Recently, Lei and Bickel (2021) developed a cyclic permutation test combining marginal rank test with exchangeability. Y. S. Wang, Lee, Toulis, and Kolar (2021); Wen, Wang, and Wang (2022) also extended residual permutation to moderate and high dimensions. For invariance assumption beyond exchangeability, Toulis (2019) proposed a general randomization-based framework.

The aforementioned methods for correlated errors contributed significantly to the developments in a wide range of research areas and are still widely used. Owing to their virtue of exploiting the correlation patterns, the estimation and inference are improved substantially. However, in many empirical applications, the correlation pattern is arbitrary, neither serial, spatial, clustered, nor exchangeable, leading to the elusive justification of which approach to use. Since we only observe a single error vector $\boldsymbol{\varepsilon}$, the correct identification of the structure of $\mathbb{C}\mathrm{ov}(\boldsymbol{\varepsilon})$ can be quite intangible. The concern of the misspecification under arbitrary and unknown $\mathbb{C}\mathrm{ov}(\boldsymbol{\varepsilon})$ may tempt us to apply the initial approach, OLS. In some cases, the intricate correlation in $\mathbb{C}\mathrm{ov}(\boldsymbol{\varepsilon})$ is even overlooked, and the best prototype, OLS estimator, is applied directly.

Instead of putting stress on the abuse of OLS, in Chapter 2, we compare the validity of OLS confidence intervals under fixed and random designs with correlated errors; see Figure 2.1a. It turns out that random design is surprisingly a remedy for correlated errors, which guarantees the valid inference without any adjustments. In fact, this phenomenon echoes some previous literature such as page 328 of Greenwald (1983), page 233 of Angrist and Pischke (2009), Lemma 4 of Barrios et al. (2012), and some ideas in the design-based literature (Abadie, Athey, Imbens, & Wooldridge, 2020; Su & Ding, 2021) dating back to Neyman (1923). Even more surprisingly, in the presence of correlated errors and random design, we observe a power gain with small signals while a power loss with large signals. To demystify these phenomena, we carefully characterize the inference of the OLS procedure under random design and unknown correlation in $\boldsymbol{\varepsilon}$ by developing the Berry–Esseen bound and local power approximation.

Technically, researchers have traditionally tackled the problem of correlated errors by fixing $\mathbf{X}$ first and then transforming those correlated errors into the classic i.i.d. errors. However, with unknown correlation structure in $\boldsymbol{\varepsilon}$, even we assume the independence between $\boldsymbol{\varepsilon}$ and $\mathbf{X}$, we have $\operatorname{cov}(\widehat{\boldsymbol{\beta}}_{\text{OLS}}|\mathbf{X}) = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbb{C}\text{ov}(\boldsymbol{\varepsilon})\mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}$, which reveals the interaction between $\mathbf{X}$ and $\mathbb{C}\text{ov}(\boldsymbol{\varepsilon})$; see Kruskal (1968); Bloomfield and Watson (1975); Scott and Holt (1982); Greenwald (1983); Krämer and Donninger (1987); Abadie et al. (2017) for theoretical explorations and Griffiths and Beesley (1984); Moulton (1990); Liang and Zeger (1993) for simulation evidence. With random $\mathbf{X}$, the interaction between $\mathbf{X}$ and $\mathbb{C}\text{ov}(\boldsymbol{\varepsilon} \mid \mathbf{X})$ becomes even more intricate, complicating the analysis of the validity of OLS inference. With arbitrarily correlated errors, the asymptotic normality of $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\text{OLS}} - \boldsymbol{\beta})$ may fail, and the standard error may not converge to a constant.

In Chapter 2, we tackle this problem by conditioning on the correlated error $\boldsymbol{\varepsilon}$ and utilize the randomness in $\mathbf{X}$, which reverses the conventional order. Such approach provides a convenient quantification of the performance of OLS estimator. Then the Berry–Esseen bound for the OLS t-statistic is obtained by the self-normalization of $\boldsymbol{\varepsilon}$ inspired by the spirit in Kiefer, Vogelsang, and Bunzel (2000); Lobato (2001). Despite the inconsistency of the OLS standard error, the self-normalization cancels the effects of the correlation in $\mathbb{C}\mathrm{ov}(\boldsymbol{\varepsilon})$. Then the randomness in $\mathbf{X}$ dominates the self-normalized $\boldsymbol{\varepsilon}$ and provides asymptotic normality. This technical idea echoes the potential outcome framework in causal inference dated back to Neyman (1923), which distinguishes our approach from traditional approaches.

Although Chapter 2 broadens the applicability of OLS inference in the presence of correlated errors, its focus is limited to the large sample regime. When the sample size is small with random design, the validity of the F and t tests within the OLS framework is rarely explored. In Chapter 3, we find that, perhaps surprisingly, if the design matrix enjoys spherical symmetry, such as the standard normal distribution, even the errors have non-zero mean, heteroskedasticity, strong correlations, or heavy tails, the F and t tests in the OLS procedure remain valid. To demystify this phenomenon, we develop a new perspective on the F distribution based on a geometric interpretation. This novel perspective not only clarifies the relationship between F and spherical distribution, but also reveals symmetric roles of $\mathbf{X}$ and $\boldsymbol{\varepsilon}$ in the regression model, which guarantees the robustness of the F and t tests against arbitrary $\boldsymbol{\varepsilon}$ when the design has spherical symmetry. Practically, our findings suggest that after modeling the response variable to the best of our knowledge, randomization of the variable of interest still yields valid inference, even when the errors contain unmodeled structures. This insight echoes the famous quote "*block what you can and randomize what you cannot.*" in page 103 of Box, Hunter, Hunter, et al. (1978).

As an application of the linear model, in Chapter 4, we further investigate high-dimensional mediation analysis with treatment-by-mediator interaction, using the structural equation modeling (SEM) framework (Baron & Kenny, 1986). Although there is a growing body of work on high-dimensional mediators (Clark-Boucher et al., 2023), applied across fields such as economics, finance, and genomics, there remains a need for inferential procedures which can effectively characterize the causal effects under varying treatment levels, *i.e.* , in the presence of the treatment-by-mediator interaction. The *difference method* in Guo et al. (2023) was not feasible with interaction, and the modified debiased Lasso (Zhou et al., 2020) method imposed strong assumptions on the correlation structure among mediators. To perform inference on the *natural indirect effects* and *natural direct effects*, we first adopt a nonconvex penalized least squares approach for the outcome model, establishing the oracle property, which ensures selection of the active mediators with probability tending to one. We then apply the *product method* to the mediator model, using the selected mediators from the first step, and establish the asymptotic normality of the resulting estimators for both natural indirect and direct effects. To promote the *heredity* principle—where interaction terms are included only when their corresponding main effects are present—we propose an overlapping group SCAD method that retains the oracle property. Finally, we apply our method to study how DNA methylation mediates the causal effect of childhood trauma on adult stress reactivity.

1.1 Outline

In this dissertation, we study the robustness of least squares inference under correlated errors in both large sample and small sample regimes, as well as high-dimensional mediation analysis with treatment-by-mediator interactions. In Chapter 2, we focus on the robustness of the least squares estimator in the large sample regime. We develop the Berry–Esseen bound for the OLS t statistic, illustrating its validity under unknown correlations in $\boldsymbol{\varepsilon}$ with random design $\mathbf{X}$. We further develop a local power approximation to unveil the phase transition phenomenon as the signal strength varies under correlated errors. In Chapter 3, we turn to the small sample regime and examine the robustness of the least squares tests. By scrupulously examining the connection between F distribution and spherical distribution, we develop a novel geometric perspective on the F-test in the regression model. This perspective demonstrates the validity of both the F- and t-tests under arbitrary errors, provided the design matrix exhibits spherical symmetry. In Chapter 4, we investigate high-dimensional mediation analysis with treatment-by-mediator interactions. We propose a two-step procedure for estimating causal effects. In the first step, we establish the oracle property of a nonconvex penalized estimator; in the second step, we show the asymptotic normality of the causal effect estimators. We also propose an overlapping group SCAD method to promote heredity. We list a short summary and discussion of future work in Chapter 5. A brief summary and discussion of future directions are presented in Chapter 5. Technical proofs, as well as additional results from simulation studies and real data analysis, are provided in Appendix A.

Chapter 2

With random regressors, least squares inference is robust to correlated errors with unknown correlation structure

2.1 Linear regression: fixed design, random design, and error distribution

2.1.1 Literature review and our perspective

Linear regression [1] is widely used in many disciplines and has attracted continued interest in statistics research; see Lei and Bickel (2021 Appendix A) for a recent review. The classic linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ assumes that the $n \times d$ covariate matrix $\mathbf{X}$ is fixed and the n-dimensional error vector $\boldsymbol{\varepsilon}$ has independent and identically distributed normal components with mean 0 and variance σ^2. Under this model, (a) the ordinary least squares (OLS) estimator $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$ is normal with mean $\boldsymbol{\beta}$ and covariance $\mathrm{cov}(\widehat{\beta}) = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$, (b) $\widehat{\sigma}^2 = \mathbf{y}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}})\mathbf{y}/(n-d)$ is unbiased for σ^2, with $\widehat{\sigma}^2/\sigma^2 \sim \chi^2_{n-d}/(n-d)$, where $\mathbf{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$ is the projection matrix onto the column space of $\mathbf{X}$, and (c) $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}^2$ are independent. The results (a)–(c) justify the statistical inference based on the pivotal quantity $T_j = L_j^{-1}(\widehat{\beta}_j - \beta_j) \sim t_{n-d}$, where β_j and $\widehat{\boldsymbol{\beta}}_j$ are the jth coordinate of $\boldsymbol{\beta}$ and $\widehat{\boldsymbol{\beta}}$, respectively, and $L_j = \widehat{\sigma}\{\mathbf{e}_j^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{e}_j\}^{1/2}$ is the standard error of $\widehat{\beta}_j$ with $\mathbf{e}_j$ being the jth basis vector in the d-dimensional space.

There is a large literature on relaxing the assumption of independent and identically distributed normal errors. First, we can relax the normality assumption but can still show $T_j \to \mathcal{N}(0,1)$ in distribution by the law of large numbers and central limit theorem. The

[1] This paper was published in *Biometrika*, 112 (1), asae054.

change from the t quantiles to normal quantiles is small when n is large compared with d. Second, we can relax the homoskedasticity assumption on the errors. With heteroskedastic errors, Eicker (1967) and White (1980) proposed a heteroskedasticity-robust covariance estimator. Third, we can further allow for dependence among the errors. With clustered errors, Liang and Zeger (1986) proposed to use the cluster-robust covariance estimator. With time series errors, Newey and West (1987) proposed to use the autocorrelation-robust covariance estimator. With spatial or network correlated errors, we can construct the corresponding robust covariance estimators.

As reviewed above, the literature focuses on modifying the standard error in constructing the t statistic, under various known correlation structures of the errors. Departing from the literature, we study the robustness of the original OLS inference procedure with respect to correlated errors with unknown correlation structure. We do not modify the original definition of the t statistic but show that $T_j \to \mathcal{N}(0,1)$ in distribution still holds under the assumption of random regressors even if the errors have an unknown correlation structure. With correlated errors, the asymptotic normality of $\widehat{\beta}_j$ breaks down in general. However, the central limit theorem for the t-statistic T_j can still hold in that regime. Intuitively, T_j has a ratio form and the correlation effect of the error $\boldsymbol{\varepsilon}$ cancels out because it appears in both the numerator and the denominator. To make this argument rigorous, we will first show that the key stochastic component in T_j is approximately $\mathbf{X}^{\mathrm{T}}\boldsymbol{\varepsilon}/\|\boldsymbol{\varepsilon}\|$ and then show the self-normalized error $\boldsymbol{\varepsilon}/\|\boldsymbol{\varepsilon}\|$ is nearly uniformly distributed over the unit sphere as long as the correlation is not extremely strong. Due to these two facts, the randomness of T_j is approximately driven by the sample mean of the rows of $\mathbf{X}$, which follows the central limit theorem with random regressors. See Section 2.3 for more details.

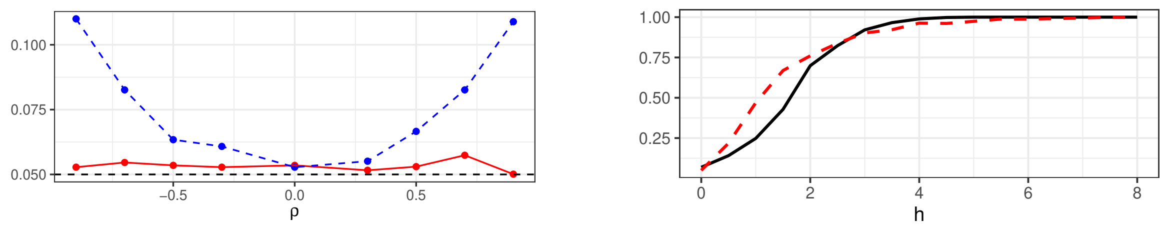
In short, our theory demonstrates that OLS inference is valid even with correlated errors, as long as the regressors are random. With fixed regressors, the correlated errors will invalidate OLS inference in general. Therefore, with fixed regressors or conditional on random regressors, the p-values from OLS can be non-uniform under the null hypotheses due to correlated errors. However, averaged over the randomness of the regressors, the p-values become uniform under the null hypotheses even if the errors are correlated in unknown ways. Overall, our theory shows that OLS inference is applicable more broadly than the classic theory suggests.

Importantly, the regime of random regressors arises naturally from randomized experiments, in which the experimenter has control over the distribution of the treatment. Therefore, our theory further demonstrates the value of randomization to ensure robustness of inference. Our setting with random regressors is reminiscent of the framework of randomization-based inference. In that literature, the focus was the robustness of inference with misspecified models (W. Lin, 2013). In contrast, we focus on the robustness of inference with correlated errors.

2.1.2 A simulated example to motivate the theory

To motivate the development of the theory, we start with the following simple yet non-trivial example. We generate data from the linear model $y_i = x_i\beta_1 + \boldsymbol{\varepsilon}_i$ for $i = 1, \dots, n$ with $n = 100$, where the x_i's are independent and identically distributed Rademacher random variables, each with a probability of 1/2 being either $+1$ or -1, and the $\boldsymbol{\varepsilon}_i$'s are multivariate normal with $\text{cov}(\boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_j) = V_{ij} = \rho^{|i-j|}$. We will vary ρ from -0.9 to 0.9 in the simulation to investigate the impact of the strength of correlation on inference. This simple model is not completely unrealistic. For instance, if x_i is the unit-level randomized treatment status with $+1$ for the treatment and -1 for the control, then the average treatment effect equals $E(y_i \mid x_i = 1) - E(y_i \mid x_i = -1) = 2\beta_1$.

Based on $10,000$ replications, we calculate the non-coverage probabilities of the confidence interval $\widehat{\beta}_1 \pm 1.96L_1$ of the true parameter $\beta_1 = 1$. While the x_i's are regenerated for each replication under the random design, they are kept unchanged across replications under the fixed design. Figure 2.1(a) shows that under the random design, the confidence interval remains valid, but under the fixed design, it is not valid due to the correlated errors.



(a) **(b)**

Figure 2.1: (a) Non-coverage probabilities of random (solid) and fixed (dashed) design. (b) Empirical power of independent and identically distributed (solid) and correlated (dashed) errors.

Moreover, we study the local power of the one-sided test $\mathcal{I}(\widehat{\beta}_1/L_1 > 1.64)$. We consider the regime of $\beta_1 \asymp (n-1)^{-1/2}h$ with h varying from 0 to 8, and set $\mathbb{C}\mathrm{ov}(\boldsymbol{\varepsilon}) = \mathbf{V}_{1,\rho} = (1-\rho)\mathbf{I}_n + \rho\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}} \in \mathbb{R}^{n\times n}$ with the diagonal entries all being 1 and off-diagonal entries all being ρ. In Figure. 2.1(b), we set $\rho = 0.9$. Perhaps surprisingly, Figure 2.1(b) shows that the t test has larger empirical power with correlated errors compared with independent errors when the signal is small.

Figure 2.1 reveals some new phenomena which only appear with random regressors. To demystify Figure 2.1(a), we will demonstrate the validity of T_j under classic OLS by establishing its Berry–Esseen bound in Section 2.3. To demystify Figure 2.1(b), we will study the local power function of the t test in Section 2.4. We first introduce the regularity conditions for our theory below. For a random variable A, let $\|A\|_{\psi_2} = \inf[t > 0 : E\{\exp(A^2/t^2)\} \leq 2]$.

Assumption 2.1.1. *Define the average variance as* $\sigma^2 = n^{-1}\sum_{i=1}^n \mathrm{Var}(\boldsymbol{\varepsilon}_i)$ *with possibly nonconstant values of* $\mathrm{Var}(\boldsymbol{\varepsilon}_i)$. *Define* $\mathbf{V} = \sigma^{-2}\mathbb{C}\mathrm{ov}(\boldsymbol{\varepsilon})$ *such that* $\mathrm{tr}(\mathbf{V}) = n$, *which equals the correlation matrix of the errors when* $\mathrm{Var}(\boldsymbol{\varepsilon}_i) = \sigma^2$ *for all* $i = 1, \ldots, n$. *The* (n, d) *satisfies* $d/n^{1/2} \to 0$ *as* $n \to \infty$.

(i) $\boldsymbol{\varepsilon} = \sigma\mathbf{V}^{1/2}\mathbf{w}$, *where* $\mathbf{V} \in \mathbb{R}^{n\times n}$ *is positive definite, and* $\mathbf{w} = (w_1, \ldots, w_n)^{\mathrm{T}} \in \mathbb{R}^n$ *has independent sub-Gaussian entries* w_i *with zero mean, unit variance, and* $\max_{1\leq i\leq n}\|w_i\|_{\psi_2} \leq K_w$ *for some constant* $K_w > 0$.

(ii) $\mathbf{X} = \mathbf{Z}\boldsymbol{\Sigma}^{1/2}$, *where* $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ *is positive definite, and* $\mathbf{Z} \in \mathbb{R}^{n\times d}$ *has independent sub-Gaussian entries* z_{ij} *with zero mean, unit variance, and* $\max_{1\le i\le n, 1\le j\le d} \|z_{ij}\|_{\psi_2} \le K_z$ *for some constant* $K_z > 0$.

(iii) $\mathbf{Z}$ *and* $\mathbf{w}$ *are independent, and* $\mathbb{P}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z} \text{ is singular}) = 0$.

(iv) $V_{\mathrm{lo}} \le \mathrm{Var}(\boldsymbol{\varepsilon}_i) \le V_{\mathrm{up}}$ *for some constants* $V_{\mathrm{lo}}, V_{\mathrm{up}} > 0$, *for all* $i = 1, \ldots, n$.

(v) $\mathbb{P}(\widehat{\sigma}^2 = 0) = 0$.

Assumption 2.1.1(i) excludes heavy-tailed errors. Assumption 2.1.1(ii) emphasizes the condition on random regressors, and specifies the rows of $\mathbf{X}$, denoted by $\mathbf{x}_i$ for $i = 1, \ldots, n$, as independent with zero mean and covariance $\boldsymbol{\Sigma}$. Assumption 2.1.1(iii) imposes the standard assumption of independence between the regressors and errors, and rules out degeneracy in the regressors. Assumption 2.1.1(iv) allows for heteroskedasticity but bounds the relative heteroskedasticity across units. Assumption 2.1.1(v) rules out the possibility of degenerate residuals, which is useful for simplifying the proofs.

As the correlation structure $\mathbf{V}$ only actives in the error term $\boldsymbol{\varepsilon}$, the effects of $\mathbf{V}$ on the projection estimator are essentially induced by $\boldsymbol{\varepsilon}$. Before stepping into the main topic, we first present two probabilistic characterizations of $\boldsymbol{\varepsilon}$ in Section 2.2 which provides some insights for the following discussions.

We use the following notation throughout the paper. For sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \lesssim b_n$ and $a_n \gtrsim b_n$ if there exists a positive integer N such that for all $n > N$, we have $a_n \le C_1 b_n$ and $a_n \ge C_2 b_n$ for some absolute constants C_1 and C_2, respectively. Let $\Phi(\cdot)$ denote the cumulative distribution function of $\mathcal{N}(0,1)$, and let z_α denote the α upper quantile of $\mathcal{N}(0,1)$. Let $\lambda_{\min}(\mathbf{V})$, $\lambda_{\max}(\mathbf{V})$ and $\lambda_i(\mathbf{V})$ denote the smallest, the largest, and the ith largest eigenvalues of the matrix $\mathbf{V}$, respectively.

2.2 Probabilistic characterizations of correlated errors

Lemmas 2.2.1 and 2.2.2 below characterize $\boldsymbol{\varepsilon}/\|\boldsymbol{\varepsilon}\|_2$ and $\delta = \boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}/(n\sigma^2) = \mathbf{w}^{\mathrm{T}}\mathbf{V}\mathbf{w}/n$, respectively.

Lemma 2.2.1. *Assume Assumption 2.1.1 (i), (iv), and $\mathbb{P}(\boldsymbol{\varepsilon} = \mathbf{0}_n) = 0$. Define $\mathbf{v} = \boldsymbol{\varepsilon}/\|\boldsymbol{\varepsilon}\|_2$ with entries v_i, $i = 1, \ldots, n$. Then for any $t > 0$, we have*

$$\mathbb{P}(n^{1/2}|v_i| \geq t) \leq 4\exp\{-c\lambda_{\min}(\mathbf{V})t^2\},$$

where c is an absolute constant as a function of K_w, V_{lo}, and V_{up}.

Lemma 2.2.1 extends Theorem 3.4.6 in Vershynin (2018) for the uniform distribution over the sphere with radius $n^{1/2}$. We adopt a similar proving technique but establish a stronger result for a general vector $\boldsymbol{\varepsilon}$ with correlation structure $\mathbf{V}$. If we further assume $\lambda_{\min}(\mathbf{V}) \geq c_{\min}$ for some absolute constant $c_{\min} > 0$, then Lemma 2.2.1 ensures that the ratio between $|v_i|$ and $n^{-1/2}$ has a sub-Gaussian tail, with $\|v_i\|_{\psi_2} \lesssim (nc_{\min})^{-1/2}$, which further ensures that the v_i's behave like $n^{-1/2}$. This is a key property to prove Theorems 2.3.1 and 2.4.1.

Lemma 2.2.2. *Under Assumption 2.1.1 (i), suppose that $\lambda_{\max}(\mathbf{V}) \asymp n^{\iota}$ where $\iota \in [0, 1]$. (a) If $\iota \in [0, 1)$, then $\delta \to 1$ in probability. (b) If $\iota = 1$, suppose that $\lambda_i(\mathbf{V})/n \to \alpha_i \in (0, 1]$, $i = 1, \ldots, K$, $\sum_{i=1}^{K} \alpha_i \in (0, 1]$, and $\lambda_i(\mathbf{V})/n \to 0$ for $i = K + 1, \ldots, n$ with fixed K. Under Assumption 2.1.1 (iv), we have $\delta \to \sum_{i=1}^{K} \alpha_i Z_i^2 + (1 - \sum_{i=1}^{K} \alpha_i)$ in distribution, where $Z_1, \ldots, Z_K$ are independent and identically distributed $\mathcal{N}(0, 1)$.*

Lemma 2.2.2 plays a crucial role in the local power analysis. It reveals that the increase in $\lambda_{\max}(\mathbf{V})$ leads to the increase in the variation in $\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}/(n\sigma^2)$. As long as $\lambda_{\max}(\mathbf{V}) = o(n)$, $\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}/(n\sigma^2)$ still converges to 1. If $\lambda_{\max}(\mathbf{V}) \asymp n$, such as the low-rank structured $\mathbf{V}$, $\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}/(n\sigma^2)$ weakly converges to a right skewed distribution. The large variation in $\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}/(n\sigma^2)$ suggests the non-robustness from correlation, which is the source of the power phase transition in Figure 2.1(b).

2.3 Validity of OLS inference with correlated errors

The key theoretical result to ensure the robustness of the classic OLS inference is the asymptotic normality of the t statistic. Theorem 2.3.1 below gives the Berry–Esseen bound on T_j.

Theorem 2.3.1 (Berry–Esseen bound on T_j)**.** *Under Assumption 2.1.1, we have*

$$\sup_{t\in\mathbb{R}} \left|\mathbb{P}(T_j \le t) - \Phi(t)\right| \lesssim \lambda_{\min}^{-3/2}(\mathbf{V}) \cdot \max(d, \log n) \cdot n^{-1/2}. \tag{2.3.1}$$

If $\lambda_{\min}(\mathbf{V}) \ge c_{\min} > 0$ for an absolute constant $c_{\min}$, the bound in (2.3.1) converges to 0 as long as $d/n^{1/2} \to 0$, which is required by Assumption 2.1.1 and matches the condition invoked by Bickel and Freedman (1982) to prove the asymptotic normality of the least squares coefficient with fixed regressors. Even if the errors are strongly correlated with $\lambda_{\min}(\mathbf{V}) \to 0$, the bound in (2.3.1) is still useful for establishing the central limit theorem of T_j as long as the bound converges to 0.

Although Theorem 2.3.1 looks similar to the classic Berry–Esseen bound with fixed regressors and independent errors, the mathematical details differ fundamentally. In particular, if we standardize the OLS coefficient by its true standard error, $T_j' = \{\sigma^2 \mathbf{e}_j^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{e}_j\}^{-1/2}(\widehat{\beta}_j - \beta_j)$ does not satisfy the central limit theorem if the errors have a general correlation structure. Only when we standardize the OLS coefficient by its estimated standard error, $T_j = \{\widehat{\sigma}^2 \mathbf{e}_j^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{e}_j\}^{-1/2}(\widehat{\beta}_j - \beta_j)$ satisfies the central limit theorem. We revisit the simulation example in Section 2.1.2 to illustrate this phenomenon. Panels A and B of Figure 2.2 show the empirical densities of T_1' and T_1, respectively. In the simulation, we set $\beta_1 = 1$ and $\mathbb{C}\mathrm{ov}(\boldsymbol{\varepsilon}) = \mathbf{V} = \mathbf{V}_{1,\rho}$, with $\rho = 0$ for independent errors and $\rho = 0.9$ for equally correlated errors. In Panel A, the empirical density of T_1' does not match that of $\mathcal{N}(0,1)$ when the errors are correlated, whereas in Panel B, the empirical density of T_1 matches that of $\mathcal{N}(0,1)$ regardless of the correlation of the errors.

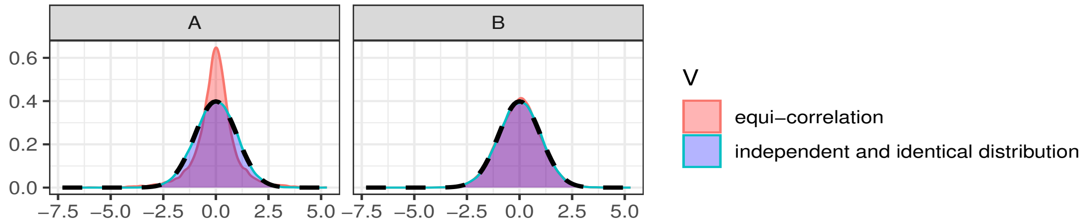


Figure 2.2: Empirical densities of $T_1' = \{\sigma^2 \mathbf{e}_1^{\mathrm{T}} (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} \mathbf{e}_1\}^{-1/2} (\widehat{\beta}_1 - \beta_1)$ in Panel A and empirical densities of $T_1 = \{\widehat{\sigma}^2 \mathbf{e}_1^{\mathrm{T}} (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} \mathbf{e}_1\}^{-1/2} (\widehat{\beta}_1 - \beta_1)$ in Panel B, under different correlation structures of the errors. The black dashed curves are the $\mathcal{N}(0,1)$ density.

To understand the central limit theorem ensured by Theorem 2.3.1, we provide some heuristics below. Let $\widehat{\mathbf{\Sigma}} = n^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X} = n^{-1}\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}$ denote the empirical second moment of covariates. Then we can rewrite T_j as

$$T_j = \{\widehat{\sigma}^2 \mathbf{e}_j^{\mathrm{T}} (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} \mathbf{e}_j\}^{-1/2} (\widehat{\beta}_j - \beta_j) = (\mathbf{e}_j^{\mathrm{T}} \widehat{\mathbf{\Sigma}}^{-1} \mathbf{e}_j)^{-1/2} \mathbf{e}_j^{\mathrm{T}} \widehat{\mathbf{\Sigma}}^{-1} \cdot \widehat{\sigma}^{-1} \cdot (n^{-1/2}\mathbf{X}^{\mathrm{T}}\boldsymbol{\varepsilon}),$$

where the first term $(\mathbf{e}_j^{\mathrm{T}} \widehat{\mathbf{\Sigma}}^{-1} \mathbf{e}_j)^{-1/2} \mathbf{e}_j^{\mathrm{T}} \widehat{\mathbf{\Sigma}}^{-1}$ is unrelated to $\mathbf{V}$, while $\widehat{\sigma}^{-1}$ and $n^{-1/2}\mathbf{X}^{\mathrm{T}}\boldsymbol{\varepsilon}$ are related to $\mathbf{V}$. The classic theory of OLS proves (a) the consistency of $\widehat{\sigma}$ and (b) the asymptotic normality of $n^{-1/2}\mathbf{X}^{\mathrm{T}}\boldsymbol{\varepsilon}$. Then Slutsky's theorem ensures the validity of the inference based on the asymptotic normality of T_j. However, both (a) and (b) break down if the errors have an unknown correlation structure $\mathbf{V}$. Nevertheless, the asymptotic normality of T_j still holds even though (a) and (b) do not hold. The theoretical justification of the asymptotic normality of T_j is completely different from the classic theory. We will provide the heuristics for the asymptotic normality of $\widehat{\sigma}^{-1} \cdot (n^{-1/2}\mathbf{X}^{\mathrm{T}}\boldsymbol{\varepsilon})$. sAssume d is small compared with n. Approximately, we have

$$\widehat{\sigma}^{-1} \cdot (n^{-1/2}\mathbf{X}^{\mathrm{T}}\boldsymbol{\varepsilon}) \approx \{\boldsymbol{\varepsilon}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}})\boldsymbol{\varepsilon}\}^{-1/2} \cdot \mathbf{X}^{\mathrm{T}}\boldsymbol{\varepsilon} \approx (\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon})^{-1/2} \cdot \mathbf{X}^{\mathrm{T}}\boldsymbol{\varepsilon} = \mathbf{X}^{\mathrm{T}} \frac{\boldsymbol{\varepsilon}}{\|\boldsymbol{\varepsilon}\|}.$$

Lemma 2.2.1 ensures that the entries of the self-normalized vector $\boldsymbol{\varepsilon}/\|\boldsymbol{\varepsilon}\|$ are around $n^{-1/2}$. This key probabilistic result then ensures

$$\widehat{\sigma}^{-1} \cdot (n^{-1/2}\mathbf{X}^{\mathrm{T}}\boldsymbol{\varepsilon}) \approx n^{-1/2}\sum_{i=1}^{n}\mathbf{x}_i, \tag{2.3.2}$$

which is asymptotically normal by the standard central limit theorem with random $\mathbf{x}_i$'s.

Remark 2.3.1. *Consider the extreme case with a single cluster so that all the $\boldsymbol{\varepsilon}_i$'s are correlated. The central limit theorem for $\widehat{\boldsymbol{\beta}}$ breaks down and the cluster-robust covariance estimator degenerates to* 0 *(Liang & Zeger, 1986), making the corresponding inference useless. By contrast, the standard OLS inference based on T_j can still be valid as long as Assumption 2.1.1 holds.*

Remark 2.3.2. *The Berry–Esseen bound in Theorem 2.3.1 relies crucially on the assumption of random regressors. Chetverikov, Hahn, Liao, and Santos (2023) reported a similar robustness property of the classic OLS inference. Our result is also related to the randomization-based inference with randomized treatment (Barrios et al., 2012; W. Lin, 2013; Abadie, Athey, Imbens, & Wooldridge, 2023). However, our theory is fundamentally different. Their theories deal with the regime of asymptotically normal estimators and consistent variance estimators, whereas our theory can deal with the regime in which the asymptotic normality of the OLS coefficient breaks down and our proof relies on the concentration properties of the self-normalized vector $\boldsymbol{\varepsilon}/\|\boldsymbol{\varepsilon}\|$ as shown in Lemma 2.2.1.*

2.4 Power analysis under a local alternative hypothesis

Based on the asymptotic normality in Theorem 2.3.1, the one-sided test for the null hypothesis of $H_0 : \beta_j = 0$ is $\mathcal{I}(L_j^{-1}\widehat{\beta}_j > z_\alpha)$. We will further study the local power of this test under the alternative hypothesis of

$$H_1 : \beta_j = h\left(\frac{\sigma^2\mathbf{e}_j^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{e}_j}{n-d}\right)^{1/2} \text{ with } h > 0. \tag{2.4.1}$$

The choice of the alternative hypothesis H_1 in (2.4.1) is motivated by the form of L_j to simplify the form of the asymptotic power function, which will be clear in (2.4.3) below.

Theorem 2.4.1 (power)**.** *Under Assumption 2.1.1 and H_1 in* (2.4.1)*, we have*

$$\left|\mathbb{P}(L_j^{-1}\widehat{\beta}_j > z_\alpha) - \pi(h, \mathbf{V})\right| \lesssim \lambda_{\min}^{-3/2}(\mathbf{V}) \cdot \max(d, \log n) \cdot (\log n)^{1/2} \cdot n^{-1/2}, \tag{2.4.2}$$

where the asymptotic power function equals $\pi(h, \mathbf{V}) = E\{\Phi(h\delta^{-1/2} - z_\alpha)\}$, with the expectation taken over $\delta = \boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}/(n\sigma^2) = \mathbf{w}^{\mathrm{T}}\mathbf{V}\mathbf{w}/n$.

In the classic regime with independent errors, $\delta \to 1$ in probability and the asymptotic power function reduces to $\Phi(h - z_\alpha)$. In the regime with strongly correlated errors, δ converges to a random variable as shown in Lemma 2.2.2. Therefore, the asymptotic power function has the form of $\pi(h, \mathbf{V})$ in Theorem 2.4.1.

We provide some heuristics for the asymptotic power function. The statistic $L_j^{-1}\widehat{\beta}_j$ decomposes as $L_j^{-1}\widehat{\beta}_j = L_j^{-1}(\widehat{\beta}_j - \beta_j) + L_j^{-1}\beta_j$, where (a) the first term, $T_j = L_j^{-1}(\widehat{\beta}_j - \beta_j)$, is approximately $\mathcal{N}(0, 1)$ by Theorem 2.3.1; (b) the second term is approximately

$$L_j^{-1}\beta_j = h\left(\frac{\sigma^2\mathbf{e}_j^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{e}_j}{n-d}\right)^{1/2} \bigg/ \left\{\frac{\widehat{\sigma}^2\mathbf{e}_j^{\mathrm{T}}(n^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{e}_j}{n}\right\}^{1/2} \approx h\delta^{-1/2},$$

because $n^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X} \approx \boldsymbol{\Sigma}$ and $\widehat{\sigma}^2 = \boldsymbol{\varepsilon}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}})\boldsymbol{\varepsilon}/(n-d) \approx \boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}/(n-d)$; and (c) the first and second terms are asymptotically independent. We can use (a)–(c) to derive the asymptotic power function

$$\mathrm{pr}(L_j{}^{-1}\widehat{\beta}_j > z_\alpha) = \mathrm{pr}(T_j > z_\alpha - L_j{}^{-1}\beta_j) \approx \Phi(h\delta^{-1/2} - z_\alpha), \tag{2.4.3}$$

with random δ. Therefore, the final asymptotic power function needs to take expectation over δ, as stated in Theorem 2.4.1.

Although δ has mean 1, it can have large variability around 1 with strongly correlated errors as shown in Lemma 2.2.2. Compared with the power function for independent errors,

$\Phi(h - z_\alpha)$, the integrand for correlated errors, $\Phi(h\delta^{-1/2} - z_\alpha)$, has a larger value if $\delta < 1$ and a smaller value if $\delta > 1$. Averaged over δ, whether correlated errors benefit or harm power depends on the variability of δ relative to h. Overall, with small h, correlated errors benefit power, whereas with large h, they harm power. To gain insights into this phenomenon, we simplify the power function under normal errors with the exchangeable correlation structure $\mathbf{V}_{1,\rho}$ below.

Corollary 2.4.1 (power function under exchangeable correlation structure)**.** *Under Assumption 2.1.1 and H_1 in (2.4.1), if $\mathbf{w} \sim \mathcal{N}(0, \mathbf{I}_n)$ and $\mathbf{V} = \mathbf{V}_{1,\rho}$ with $\rho \in [0, 1 - c_{\min}]$ for an absolute constant $c_{\min} \in (0, 1)$, then*

$$\left|\mathbb{P}(L_j^{-1}\widehat{\beta}_j > z_\alpha) - \pi(h, \rho)\right| \lesssim (\log n)^{1/2} \cdot \max(d, \log n) \cdot n^{-1/2},$$

where $\pi(h, \rho) = E\{\Phi(h(\rho\chi_1^2 + 1 - \rho)^{-1/2} - z_\alpha)\}$, with expectation taken over χ_1^2.

The asymptotic power function $\pi(h, \rho)$ in Corollary 2.4.1 is a special case of the general $\pi(h, \mathbf{V})$ in Theorem 2.4.1, with δ replaced by its asymptotic distribution $\rho\chi_1^2 + 1 - \rho$. Corollary 2.4.1 offers insights into the dependence of power on the correlation structure. Figure 2.3 highlights the region of (h, ρ) with $\pi(h, \rho) - \pi(h, 0) > 0$ such that the t-test based on OLS is more powerful with correlated errors than with independent errors. It shows that with small h, correlated errors improve the power, whereas with large h, correlated errors harm the power.

We can also use Figure 2.4 to better explain the phase transition under general correlation structure $\mathbf{V}$, which demonstrates the competition between the integral variable δ and the integrand $\Phi(\delta^{-1/2}h - z_\alpha)$. Compared with the power from independent errors $\Phi(h - z_\alpha)$, the integrand produces power loss if $\delta > 1$ such as δ_1 in Figure 2.4, as $\Phi(\delta_1^{-1/2}h - z_\alpha) - \Phi(h - z_\alpha) < 0$. The integrand tends to gain power if $\delta < 1$ such as δ_2 in Figure 2.4, as $\Phi(\delta_2^{-1/2}h - z_\alpha) - \Phi(h - z_\alpha) > 0$.

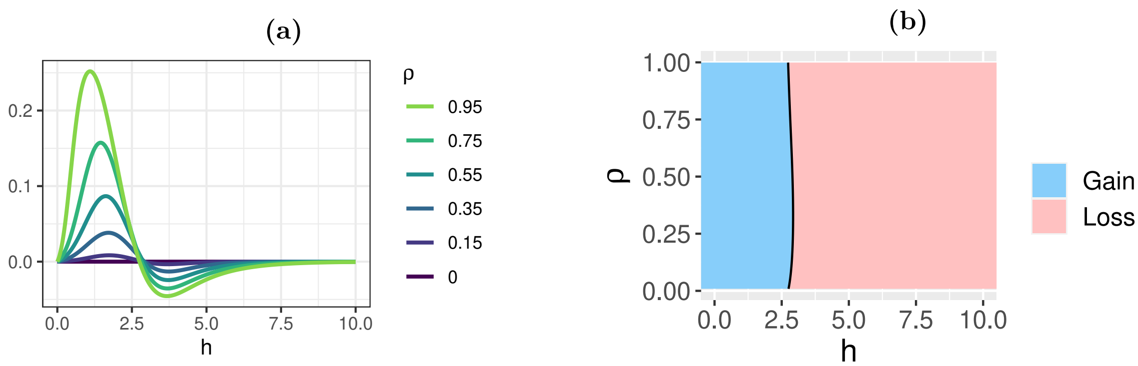
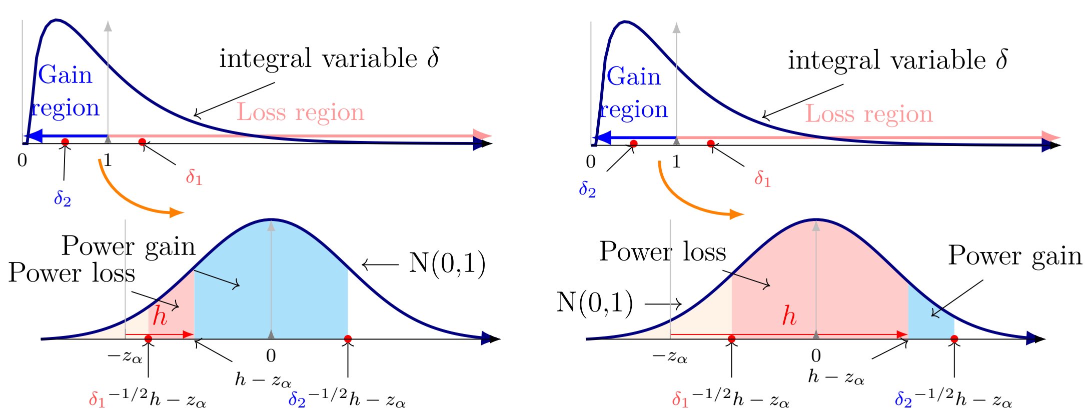


Figure 2.3: (a) $\pi(h, \rho) - \pi(h, 0)$ as a function of h, given different values of ρ. (b) Region with power gain such that $\pi(h, \rho) - \pi(h, 0) > 0$ and region with power loss such that $\pi(h, \rho) - \pi(h, 0) < 0$.

In Figure 2.4(a), when h is small, under the $\mathcal{N}(0, 1)$ density, the area indicating power loss due to $\delta > 1$ is confined to $(-z_\alpha, h - z_\alpha)$. In contrast, the area for power gain from $\delta < 1$ surrounds zero where $\Phi()$ increases fastest. Such power gain from the fast increment in $\Phi()$ is large enough which leads to the ultimate power gain after the integral of δ. On the other hand, in Figure 2.4(b), with large h, the power loss from $\delta < 1$ aligns with the normal density's mode, whereas the power gain from $\delta < 1$ is pushed to the right tail $(h - z_\alpha, \infty)$, where $\Phi()$ increases slowly. Such power gain is too small to compensate the predominated power loss under the normal density, which leads to the overall power loss after the integral of δ.



(a) Power Gain with Small Signals **(b)** Power Loss with Large Signals

Figure 2.4: The intuition of the relative power transition.

Conceptually, since the test statistic is robust while the standard error is sensitive to the correlated errors, the competition between the variation in the standard error stemming from the dependent noise, and the normal distribution from self-normalization and random $\mathbf{X}$, yield this surprising phenomenon.

2.5 Fixed, random, or mixed regressors

With random regressors, we have demonstrated the robustness of OLS inference with respect to correlated errors. With fixed regressors, the theory breaks down. We can construct a counterexample. For instance, if $\mathbf{y} = \beta \mathbf{1}_n + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{V}_{1,\rho})$, then $\widehat{\beta} - \beta \sim \mathcal{N}(0, \rho + n^{-1}(1-\rho))$ is bounded in probability, and $L_1 \approx \{n^{-2}\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}\}^{1/2}$ converges to 0 in probability. Therefore, wrongly assuming asymptotic normality of $L_1^{-1}(\widehat{\beta} - \beta)$ does not give valid inference.

When the regressors contain both fixed and random components, the OLS inference for the coefficients of the fixed components is not valid whereas that of the random components is still valid asymptotically. Consider model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $n = 100$, $\boldsymbol{\beta} = [1, 1, 1]^{\mathrm{T}}$, $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3] \in \mathbb{R}^{100\times 3}$ where $\mathbf{X}_1 = \mathbf{1}_n$ is the intercept. Both $\mathbf{X}_2$ and $\mathbf{X}_3$ have independent and identically distributed Rademacher entries. Yet $\mathbf{X}_2$ is fixed across replications and $\mathbf{X}_3$ is regenerated for each replication. The errors satisfy $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{V})$, where $V_{ij} = \rho^{|i-j|}$ with ρ varying from -0.9 to 0.9. Figure 2.5 demonstrates the non-coverage probabilities of the confidence interval $\widehat{\beta}_j \pm 1.96 L_j$ and the densities of T_j. The OLS inference for the coefficient of the random regressor $\mathbf{X}_3$ is valid because T_3 is close to $\mathcal{N}(0, 1)$. By contrast, the OLS inference for other coefficients is not valid.

2.6 Appendix

In this Section, we provide additional conclusions and proofs for the main paper. Section 2.6.1 proves the two lemmas in Section 2.2 of the main paper and documents technical lemmas used throughout this supplementary file. Section 2.6.3 presents the proof of the

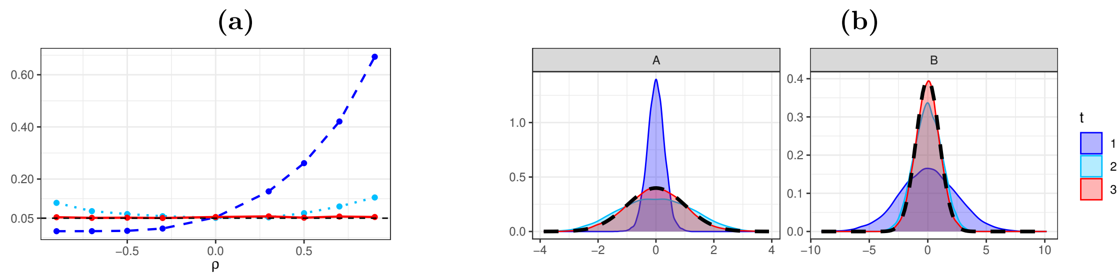


Figure 2.5: (a) Non-coverage probabilities of β_1 (dashed), β_2 (dotted), and β_3 (solid). (b) Empirical densities of T_j for $j = 1, 2, 3$ when $\rho = -0.9$ (Panel A) and 0.7 (Panel B). The black dashed curves are the $\mathcal{N}(0, 1)$ density.

Berry–Esseen bound in Theorem 2.3.1. Sections 2.6.4–2.6.6 provide the proofs for all the results about power, with Section 2.6.4 for Theorem 2.4.1, Section 2.6.5 for Corollary 2.4.1, and Section 2.6.6 for properties of $\pi(h, \rho) - \pi(h, 0)$, respectively. Finally, Section 2.6.7 provides tail behaviors of $\|\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta}\|_2^2$ and L_j^2 to highlight the role of self-normalization for the robustness of inference.

Notation. For sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \asymp b_n$ if there are constants m, M and N such that $0 < m < |a_n/b_n| < M < \infty$ for all $n > N$. We write $a_n \lesssim b_n$ if $a_n \leq Cb_n$ for some C independent of n. We also write $|a_n| \lesssim 1$ if $a_n = O(1)$. We write $a_n \ll b_n$ if $a_n/b_n \to 0$. We write $a_n = o(1)$ if $\lim_{n\to\infty} a_n = 0$. For a vector $\mathbf{a}$, let $\|\mathbf{a}\|_1$ and $\|\mathbf{a}\|_2$ denote the $\mathcal{L}_1$ and $\mathcal{L}_2$ norms, respectively. Let $\mathbf{1}_n \in \mathbb{R}^n$ denote the vector whose entries are all 1's. Let $\mathrm{diag}\{a_1, \dots, a_n\}$ denote the diagonal matrix in $\mathbb{R}^{n\times n}$ with diagonal entries $a_1, \dots, a_n$. Given an $n \times d$ matrix $\mathbf{M}$, let $s_{\min}(\mathbf{M})$ and $s_{\max}(\mathbf{M})$ denote the minimum and maximum nonzero singular values of $\mathbf{M}$, respectively. Let $\|\mathbf{M}\|$ and $\|\mathbf{M}\|_{\mathrm{F}}$ denote the operator norm and the Frobenius norm of $\mathbf{M}$, respectively. Let $\mathbf{P}_{\mathbf{M}}$ denote the projection matrix of the column space of $\mathbf{M}$, that is, $\mathbf{P}_{\mathbf{M}} = \mathbf{M}(\mathbf{M}^{\mathrm{T}}\mathbf{M})^{-1}\mathbf{M}^{\mathrm{T}}$ if $\mathbf{M}^{\mathrm{T}}\mathbf{M}$ is nonsingular. For a square matrix $\mathbf{M}$, let $\det(\mathbf{M})$ denote the determinant of $\mathbf{M}$ and let $\mathrm{diag}(\mathbf{M}) \in \mathbb{R}^{d\times d}$ denote a diagonal matrix with diagonal entries from $\mathbf{M}$. Let $\mathrm{diag}\{\mathbf{M}_1, \dots, \mathbf{M}_K\}$ denote the block diagonal matrix with diagonal blocks $\mathbf{M}_1, \dots, \mathbf{M}_K$. For any symmetric matrix $\mathbf{M} \in \mathbb{R}^{d\times d}$, let $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$ denote the minimum and maximum eigenvalues of $\mathbf{M}$, respectively; let $\lambda_i(\mathbf{M})$ denote the ith largest eigenvalue of $\mathbf{M}$ for $i = 1, \dots, d$. Let $\mathbf{I}_d$

denote the $d \times d$ identity matrix. Let $\operatorname{tr}(\mathbf{M}) = \sum_{i=1}^{d} M_{ii}$ denote the trace of $\mathbf{M}$. Let $\mathbf{e}_i \in \mathbb{R}^d$ denote the vector with a 1 in the i-th position and 0's elsewhere. Let $\stackrel{d}{=}$ denote equal in distribution. For random vectors $\{\mathbf{X}_n\}_{n=1}^{\infty}$, $\mathbf{X}$, and $\mathbf{Y}$, let $\mathbf{X}_n \xrightarrow{L} \mathbf{X}$ and $\mathbf{X}_n \xrightarrow{P} \mathbf{X}$ denote the convergence in law and convergence in probability, respectively. Let $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}$ denote that $\mathbf{X}$ is independent of $\mathbf{Y}$. Let $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Let $\sim$ denote following certain distribution, for example, $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. For a random variable X, define $\|X\|_{\psi_p} = \inf\{t > 0 : \mathbb{E}[\exp(|X|^p/t^p)] \leq 2\}$, where $p = 1$ represents the sub-Exponential norm and $p = 2$ represents the sub-Gaussian norm. We use c, C, and $\widetilde{c}$ to denote positive and generic absolute constants. With slight abuse of notation, $\mathcal{E}_1$, $\mathcal{E}_2$, and $\mathcal{E}_\eta$, referring to different events, have different meanings across different sections.

2.6.1 Proofs of Lemmas 2.2.1 and 2.2.2

2.6.1.1 Proof of Lemma 2.2.1

The proof of Lemma 2.2.1 uses the idea similar to that of Theorem 3.4.6 in Vershynin (2018). Let $v_i = \mathbf{e}_i^{\mathrm{T}}\mathbf{v}$, where $\mathbf{e}_i \in \mathbb{R}^n$ is the ith canonical basis of $\mathbb{R}^n$. Hence

$$
\begin{aligned}
\mathbb{P}\left\{\sqrt{n}\,|\mathbf{e}_i^{\mathrm{T}}\mathbf{v}| \geq t\right\} &= \mathbb{P}\left\{\frac{\left|\mathbf{e}_i^{\mathrm{T}}\mathbf{V}^{1/2}\mathbf{w}\right|}{\sqrt{\mathbf{w}^{\mathrm{T}}\mathbf{V}\mathbf{w}}} \geq \frac{t}{\sqrt{n}}\right\} \\
&\leq \mathbb{P}\left\{\frac{\left|\mathbf{e}_i^{\mathrm{T}}\mathbf{V}^{1/2}\mathbf{w}\right|}{\sqrt{\lambda_{\min}(\mathbf{V})\mathbf{w}^{\mathrm{T}}\mathbf{w}}} \geq \frac{t}{\sqrt{n}}\right\} \\
&\leq \mathbb{P}\left\{\sqrt{\mathbf{w}^{\mathrm{T}}\mathbf{w}} < \frac{\sqrt{n}}{2}\right\} + \mathbb{P}\left\{\sqrt{\mathbf{w}^{\mathrm{T}}\mathbf{w}} \geq \frac{\sqrt{n}}{2}, \frac{\left|\mathbf{e}_i^{\mathrm{T}}\mathbf{V}^{1/2}\mathbf{w}\right|}{\sqrt{\mathbf{w}^{\mathrm{T}}\mathbf{w}}} \geq \frac{t\sqrt{\lambda_{\min}(\mathbf{V})}}{\sqrt{n}}\right\} \\
&\leq \mathbb{P}\left\{\mathbf{w}^{\mathrm{T}}\mathbf{w} < \frac{n}{4}\right\} + \mathbb{P}\left\{\left|\mathbf{e}_i^{\mathrm{T}}\mathbf{V}^{1/2}\mathbf{w}\right| \geq \frac{\sqrt{\lambda_{\min}(\mathbf{V})}t}{2}\right\}.
\end{aligned}
$$

By the Bernstein's inequality in Lemma 2.6.4 below, we have

$$
\mathbb{P}\left\{\mathbf{w}^{\mathrm{T}}\mathbf{w} < \frac{n}{4}\right\} = \mathbb{P}\left\{\mathbf{w}^{\mathrm{T}}\mathbf{w} - n < -\frac{3}{4}n\right\} \leq 2\exp\left(-cn\right) \leq \exp[-c\lambda_{\min}(\mathbf{V})n], \tag{2.6.1}
$$

where the last step in (2.6.1) follows from $\lambda_{\min}(\mathbf{V}) \leq 1$. By Lemma 2.6.3 and Assumption 2.1.1 (i), (iv), we have $\|\mathbf{e}_i^{\mathrm{T}}\mathbf{V}^{1/2}\mathbf{w}\|_{\psi_2}^2 \leq CK_w^2\mathbf{e}_i^{\mathrm{T}}\mathbf{V}\mathbf{e}_i \leq CK_w^2 V_{\mathrm{up}}$ for an absolute constant C, which implies that

$$\mathbb{P}\left\{\left|\mathbf{e}_i^{\mathrm{T}}\mathbf{V}^{1/2}\mathbf{w}\right| \geq \frac{\sqrt{\lambda_{\min}(\mathbf{V})}t}{2}\right\} \leq 2\exp\left[-c\lambda_{\min}(\mathbf{V})t^2\right], \tag{2.6.2}$$

where c in (2.6.2) absorbs the constants V_{up} and K_w. Comparing $\exp[-c\lambda_{\min}(\mathbf{V})n]$ in (2.6.1) and $\exp[-c\lambda_{\min}(\mathbf{V})t^2]$ in (2.6.2), we consider two cases:

1. If $t^2 \leq n$, then $\exp[-c\lambda_{\min}(\mathbf{V})n] \leq \exp[-c\lambda_{\min}(\mathbf{V})t^2]$ and therefore,

$$\mathbb{P}\left\{\sqrt{n}\,|\mathbf{e}_i^{\mathrm{T}}\mathbf{v}| \geq t\right\} \leq 4\exp[-c\lambda_{\min}(\mathbf{V})t^2].$$

2. If $t^2 > n$, then $t/\sqrt{n} > 1$. So we have $\mathbb{P}\{\sqrt{n}|\mathbf{e}_i^{\mathrm{T}}\mathbf{v}| \geq t\} \leq \mathbb{P}\{|\mathbf{e}_i^{\mathrm{T}}\mathbf{v}| > 1\}$. However, we always have

$$|\mathbf{e}_i^{\mathrm{T}}\mathbf{V}^{1/2}\mathbf{w}| \leq \sqrt{\mathbf{e}_i^{\mathrm{T}}\mathbf{e}_i}\sqrt{\mathbf{w}^{\mathrm{T}}\mathbf{V}\mathbf{w}} = \sqrt{\mathbf{w}^{\mathrm{T}}\mathbf{V}\mathbf{w}},$$

   which implies that $|\mathbf{e}_i^{\mathrm{T}}\mathbf{v}| \leq 1$. Hence we have $\mathbb{P}\{\sqrt{n}|\mathbf{e}_i^{\mathrm{T}}\mathbf{v}| \geq t\} = 0$, which is still bounded by $4\exp[-c\lambda_{\min}(\mathbf{V})t^2]$.

Combining case 1 and case 2, we obtain the desired result.

2.6.1.2 Proof of Lemma 2.2.2

The proof of Lemma 2.2.2 (a) is derived from the standard Hanson–Wright inequality (Theorem 6.2.1 in Vershynin (2018)). For Lemma 2.2.2 (b), the proof follows from the asymptotic theory results reviewed in Section A.3.1.

Part (a). For any $r \geq 0$,

$$
\begin{aligned}
& \mathbb{P}\left\{\left|\frac{\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}}{n\sigma^2} - 1\right| \geq \max(r, r^2)\right\} \\
= \;& \mathbb{P}\left\{\left|\mathbf{w}^{\mathrm{T}}\mathbf{V}\mathbf{w} - \mathbb{E}\left(\mathbf{w}^{\mathrm{T}}\mathbf{V}\mathbf{w}\right)\right| \geq \max(r, r^2)n\right\} && (2.6.3) \\
\leq \;& 2\exp\left\{-c\min\left[\frac{\max^2\left(r^2, r\right)n^2}{\operatorname{tr}(\mathbf{V}^2)}, \frac{\max\left(r^2, r\right)n}{\lambda_{\max}\left(\mathbf{V}\right)}\right]\right\} && (2.6.4) \\
\leq \;& 2\exp\left\{-c\min\left[\max^2\left(r^2, r\right), \max\left(r^2, r\right)\right]\min\left[\frac{n^2}{\operatorname{tr}(\mathbf{V}^2)}, \frac{n}{\lambda_{\max}(\mathbf{V})}\right]\right\} \\
\leq \;& 2\exp\left[-\frac{cnr^2}{\lambda_{\max}(\mathbf{V})}\right], && (2.6.5)
\end{aligned}
$$

where (2.6.3) follows from $\boldsymbol{\varepsilon} = \sigma\mathbf{V}^{1/2}\mathbf{w}$ in Assumption 2.1.1 (i) and $\operatorname{tr}(\mathbf{V}) = n$, (2.6.4) follows from the Hanson-Wright inequality in Lemma 2.6.5 below, with c only depending on K_w, and (2.6.5) follows from the fact that for any $r \geq 0$, $\min[\max^2(r^2, r), \max(r^2, r)] = r^2$ and $\operatorname{tr}(\mathbf{V}^2) \leq n\lambda_{\max}(\mathbf{V})$, so that $n^2/\operatorname{tr}(\mathbf{V}^2) \geq n^2/[n\lambda_{\max}(\mathbf{V})] = n/\lambda_{\max}(\mathbf{V})$.

Part (b). Let $\mathbf{V} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^{\mathrm{T}}$, where $\mathbf{Q} = [\mathbf{q}_1 \;\ldots\; \mathbf{q}_K \;\mathbf{q}_{K+1} \;\ldots\; \mathbf{q}_n]$ with $\mathbf{q}_i \in \mathbb{R}^n$ denoting the eigenvector corresponding to $\lambda_i(\mathbf{V})$ for $i = 1, \ldots, n$, and $\boldsymbol{\Lambda} = \operatorname{diag}\{\lambda_1(\mathbf{V}), \ldots, \lambda_K(\mathbf{V}), \lambda_{K+1}(\mathbf{V}), \ldots, \lambda_n(\mathbf{V})\}$. Then $\mathbf{V} = \mathbf{V}_K + \mathbf{V}_{-K}$ where

$$
\mathbf{V}_K = \sum_{i=1}^{K} \lambda_i(\mathbf{V})\mathbf{q}_i\mathbf{q}_i^{\mathrm{T}}, \;\; \mathbf{V}_{-K} = \sum_{j=K+1}^{n} \lambda_j(\mathbf{V})\mathbf{q}_j\mathbf{q}_j^{\mathrm{T}}. \tag{2.6.6}
$$

Thus, $\mathbf{w}^{\mathrm{T}}\mathbf{V}\mathbf{w} = \mathbf{w}^{\mathrm{T}}\mathbf{V}_K\mathbf{w} + \mathbf{w}^{\mathrm{T}}\mathbf{V}_{-K}\mathbf{w}$.

First, we consider $n^{-1}(\mathbf{w}^{\mathrm{T}}\mathbf{V}_{-K}\mathbf{w})$. For any $r \geq 0$,

$$
\begin{aligned}
& \mathbb{P}\{|n^{-1}(\mathbf{w}^{\mathrm{T}}\mathbf{V}_{-K}\mathbf{w}) - (1 - \alpha_f)| \geq r\} \\
\leq \;& \mathbb{P}\{|n^{-1}(\mathbf{w}^{\mathrm{T}}\mathbf{V}_{-K}\mathbf{w}) - n^{-1}\mathbb{E}(\mathbf{w}^{\mathrm{T}}\mathbf{V}_{-K}\mathbf{w})| + |n^{-1}\mathbb{E}(\mathbf{w}^{\mathrm{T}}\mathbf{V}_{-K}\mathbf{w}) - (1 - \alpha_f)| \geq r\} \\
= \;& \mathbb{P}\{|\mathbf{w}^{\mathrm{T}}\mathbf{V}_{-K}\mathbf{w} - \mathbb{E}(\mathbf{w}^{\mathrm{T}}\mathbf{V}_{-K}\mathbf{w})| \geq n[r - o(1)]\} && (2.6.7) \\
\leq \;& 2\exp\left(\frac{-cn}{\lambda_{K+1}(\mathbf{V})}\min\left\{[r - o(1)]^2, [r - o(1)]\right\}\right), && (2.6.8)
\end{aligned}
$$

where (2.6.7) follows from $\mathrm{tr}(\mathbf{V}) = n$ and

$$\begin{aligned}\left|\frac{\mathbb{E}(\mathbf{w}^{\mathrm{T}}\mathbf{V}_{-K}\mathbf{w})}{n} - (1-\alpha_f)\right| &= \left|\frac{\sum_{i=K+1}^{n}\lambda_i(\mathbf{V})}{n} - (1-\alpha_f)\right| \\ &= \left|\left[1 - \frac{\sum_{i=1}^{K}\lambda_i(\mathbf{V})}{n}\right] - (1-\alpha_f)\right| \\ &= o(1),\end{aligned}$$

and (2.6.8) follows from similar proofs to (2.6.4) and (2.6.5), using the Hanson-Wright inequality in Lemma 2.6.5 below.

Next we consider $n^{-1}(\mathbf{w}^{\mathrm{T}}\mathbf{V}_K\mathbf{w})$. Denote $[\mathbf{q}_1 \ \dots \ \mathbf{q}_K]^{\mathrm{T}} = [\boldsymbol{\varphi}_1 \ \dots \ \boldsymbol{\varphi}_n] \in \mathbb{R}^{K\times n}$, where $\boldsymbol{\varphi}_i = [q_{i1} \ \dots \ q_{iK}]^{\mathrm{T}} \in \mathbb{R}^K$ for $i = 1,\dots,n$. By the Cauchy–Schwarz inequality, we have

$$|q_{ij}| = |\mathbf{e}_i^{\mathrm{T}}\mathbf{q}_j| = |\mathbf{e}_i^{\mathrm{T}}\mathbf{V}^{1/2}\mathbf{V}^{-1/2}\mathbf{q}_j| \le (\mathbf{q}_j^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{q}_j)^{1/2}(\mathbf{e}_i^{\mathrm{T}}\mathbf{V}\mathbf{e}_i)^{1/2} = [\lambda_j^{-1}(\mathbf{V})V_{ii}]^{1/2},$$

for $j = 1,\dots,K$, which implies that

$$\boldsymbol{\varphi}_i^{\mathrm{T}}\boldsymbol{\varphi}_i = q_{i1}^2 + \cdots + q_{iK}^2 \le V_{ii}[\lambda_1^{-1}(\mathbf{V}) + \ldots + \lambda_K^{-1}(\mathbf{V})] \le \lambda_K^{-1}(\mathbf{V})KV_{ii}. \tag{2.6.9}$$

Under Assumption 2.1.1 (iv), $V_{ii} \asymp O(1),\ i = 1,\dots,n.$ Define

$$\mathbf{S}_n := \begin{bmatrix}\boldsymbol{\varphi}_1 & \dots & \boldsymbol{\varphi}_n\end{bmatrix}\mathbf{w} = \sum_{i=1}^{n} w_i\boldsymbol{\varphi}_i \tag{2.6.10}$$

so that

$$\frac{\mathbf{w}^{\mathrm{T}}\mathbf{V}_K\mathbf{w}}{n} = \mathbf{S}_n^{\mathrm{T}}\mathrm{diag}\left\{\frac{\lambda_1(\mathbf{V})}{n},\dots,\frac{\lambda_K(\mathbf{V})}{n}\right\}\mathbf{S}_n. \tag{2.6.11}$$

We will use the Cramér–Wold device to show the asymptotic normality of $\mathbf{S}_n$, that is, we will show that for any real vector $\mathbf{b} \in \mathbb{R}^K$, $\mathbf{b}^{\mathrm{T}}\mathbf{S}_n = \sum_{i=1}^{n} w_i\mathbf{b}^{\mathrm{T}}\boldsymbol{\varphi}_i \xrightarrow{L} \mathcal{N}(0, \mathbf{b}^{\mathrm{T}}\mathbf{I}_K\mathbf{b})$. Since $\boldsymbol{\varphi}_i,\ i = 1,\dots,n$ changes with $\mathbf{V}$ as n changes, $\{w_i\mathbf{b}^{\mathrm{T}}\boldsymbol{\varphi}_i\}_{i=1}^{n}$ forms a triangular array as n increases. So we will use the Lyapounov Central Limit Theorem (Lemma 2.6.1) for this triangular array. Here are some basic facts:

1. For each n, $w_i\mathbf{b}^{\mathrm{T}}\boldsymbol{\varphi}_i,\ i=1,\ldots,n$, are independent;

2. $\mathbb{E}(w_i\mathbf{b}^{\mathrm{T}}\boldsymbol{\varphi}_i)=0$ and $\mathbb{E}(w_i\mathbf{b}^{\mathrm{T}}\boldsymbol{\varphi}_i)^2=(\mathbf{b}^{\mathrm{T}}\boldsymbol{\varphi}_i)^2\leq(\mathbf{b}^{\mathrm{T}}\mathbf{b})(\boldsymbol{\varphi}_i^{\mathrm{T}}\boldsymbol{\varphi}_i)<\infty$;

3. we have

$$
\begin{aligned}
&\left[\sum_{i=1}^{n}\mathbb{E}(w_i\mathbf{b}^{\mathrm{T}}\boldsymbol{\varphi}_i)^2\right]^{-3/2}\sum_{i=1}^{n}\mathbb{E}\left|w_i\mathbf{b}^{\mathrm{T}}\boldsymbol{\varphi}_i\right|^3\\
=\;&(\mathbf{b}^{\mathrm{T}}\mathbf{b})^{-3/2}\sum_{i=1}^{n}\mathbb{E}|w_i\mathbf{b}^{\mathrm{T}}\boldsymbol{\varphi}_i|^3 \qquad (2.6.12)\\
\leq\;&(\mathbf{b}^{\mathrm{T}}\mathbf{b})^{-3/2}(\max_i\mathbb{E}|w_i|^3)\sum_{i=1}^{n}|\mathbf{b}^{\mathrm{T}}\boldsymbol{\varphi}_i|^3\\
\leq\;&(\mathbf{b}^{\mathrm{T}}\mathbf{b})^{-3/2}C\sum_{i=1}^{n}(\mathbf{b}^{\mathrm{T}}\mathbf{b})^{3/2}(\boldsymbol{\varphi}_i^{\mathrm{T}}\boldsymbol{\varphi}_i)^{3/2} \qquad (2.6.13)\\
\leq\;&CK^{3/2}V_{\mathrm{up}}^{3/2}n\lambda_K^{-3/2}(\mathbf{V})\longrightarrow 0, \qquad (2.6.14)
\end{aligned}
$$

where (2.6.12) follows from

$$
\sum_{i=1}^{n}\mathbb{E}\left(w_i\mathbf{b}^{\mathrm{T}}\boldsymbol{\varphi}_i\right)^2=\mathbf{b}^{\mathrm{T}}\left(\sum_{i=1}^{n}\boldsymbol{\varphi}_i\boldsymbol{\varphi}_i^{\mathrm{T}}\right)\mathbf{b}=\mathbf{b}^{\mathrm{T}}\begin{bmatrix}\mathbf{q}_1 & \ldots & \mathbf{q}_K\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\mathbf{q}_1 & \ldots & \mathbf{q}_K\end{bmatrix}\mathbf{b}=\mathbf{b}^{\mathrm{T}}\mathbf{b},
$$

(2.6.13) follows from the fact that under Assumption 2.1.1 (i), $\max_i\mathbb{E}|w_i|^3$ is bounded by C related to K_w and the Cauchy–Schwarz inequality, (2.6.14) follows from (2.6.9) and $\sum_{i=1}^{n}V_{ii}^{3/2}\leq nV_{\mathrm{up}}^{3/2}$. Since $\lambda_K(\mathbf{V})\asymp n$, (2.6.14) converges to zero.

By the Lyapounov Central Limit Theorem in Lemma 2.6.1 below, we have

$$
(\mathbf{b}^{\mathrm{T}}\mathbf{b})^{-1/2}\sum_{i=1}^{n}w_i\mathbf{b}^{\mathrm{T}}\boldsymbol{\varphi}_i\xrightarrow{L}\mathcal{N}(0,1)\Rightarrow\mathbf{b}^{\mathrm{T}}\mathbf{S}_n=\sum_{i=1}^{n}w_i\mathbf{b}^{\mathrm{T}}\boldsymbol{\varphi}_i\xrightarrow{L}\mathcal{N}(0,\mathbf{b}^{\mathrm{T}}\mathbf{I}_K\mathbf{b}).
$$

By the Cramér–Wold Theorem in Lemma 2.6.2 below, we have $\mathbf{S}_n\xrightarrow{L}[Z_1\ \ldots\ Z_K]^{\mathrm{T}}$, where $Z_1,\ldots,Z_K\overset{iid}{\sim}\mathcal{N}(0,1)$. By $n^{-1/2}\lambda_i^{1/2}(\mathbf{V})\to\alpha_f^{(i)1/2}$ for $i=1,\ldots,K$, Slutsky's theorem, and the definition of $\mathbf{S}_n$ in (2.6.10), we have

$$
\operatorname{diag}\left\{\sqrt{\frac{\lambda_1(\mathbf{V})}{n}},\ldots,\sqrt{\frac{\lambda_K(\mathbf{V})}{n}}\right\}\mathbf{S}_n\xrightarrow{L}\operatorname{diag}\left\{\sqrt{\alpha_f^{(1)}},\ldots,\sqrt{\alpha_f^{(K)}}\right\}[Z_1\ \ldots\ Z_K]^{\mathrm{T}}.
$$

By the continuous mapping theorem and (2.6.11) , $n^{-1}\mathbf{w}^{\mathrm{T}}\mathbf{V}_K\mathbf{w} \xrightarrow{L} \sum_{i=1}^{K} \alpha_f^{(i)} Z_i^2$. Finally using Slutsky's theorem again, we obtain

$$\frac{\mathbf{w}^{\mathrm{T}}\mathbf{V}\mathbf{w}}{n} = \frac{\mathbf{w}^{\mathrm{T}}\mathbf{V}_K\mathbf{w}}{n} + \frac{\mathbf{w}^{\mathrm{T}}\mathbf{V}_{-K}\mathbf{w}}{n} \xrightarrow{L} \sum_{i=1}^{K} \alpha_f^{(i)} Z_i^2 + 1 - \alpha_f.$$

2.6.2 Other Technical Lemmas

2.6.2.1 Lemmas for the Asymptotic Theory

Lemma 2.6.1. *(Lyapounov Central Limit Theorem, Lehmann and Romano (2005, Corollary 11.2.1)). Suppose for each* n, $\xi_{n,1}, \ldots, \xi_{n,r_n}$ *are independent with* $\mathbb{E}(\xi_{n,i}) = 0$ *and* $\sigma_{n,i}^2 = \mathbb{E}(\xi_{n,i}^2) < \infty$. *Let* $s_n^2 = \sum_{i=1}^{r_n} \sigma_{n,i}^2$. *Assume that for some* $\eta > 0$, *it holds that*

$$\lim_{n\to\infty} \sum_{i=1}^{r_n} \frac{1}{s_n^{2+\eta}} \mathbb{E}(|\xi_{n,i}^{2+\eta}|) = 0.$$

Then $\sum_{i=1}^{r_n} X_{n,i}/s_n \xrightarrow{L} \mathcal{N}(0,1)$.

Lemma 2.6.2. *(Cramér–Wold Theorem, Billingsley (1995, Theorem 29.4)) For random vectors* $\{\boldsymbol{\xi}_n\}_{n=1}^{\infty} \subset \mathbb{R}^K$ *and* $\boldsymbol{\xi} \in \mathbb{R}^K$, *a necessary and sufficient condition for* $\boldsymbol{\xi}_n \xrightarrow{L} \boldsymbol{\xi}$ *is that* $\mathbf{b}^{\mathrm{T}}\boldsymbol{\xi}_n \xrightarrow{L} \mathbf{b}^{\mathrm{T}}\boldsymbol{\xi}$ *for all* $\mathbf{b} \in \mathbb{R}^K$.

2.6.2.2 Lemmas about Concentration Inequalities

Lemma 2.6.3. *(Sums of independent sub-Gaussian variables, Vershynin (2018, Proposition 2.6.1)) Let* $\xi_1, \ldots, \xi_n$ *be independent, mean zero, sub-Gaussian random variables. Then* $\sum_{i=1}^{n} \xi_i$ *is also a sub-Gaussian random variable, and*

$$\left\| \sum_{i=1}^{n} \xi_i \right\|_{\psi_2}^2 \leq C \sum_{i=1}^{n} \|\xi_i\|_{\psi_2}^2,$$

where C *is an absolute constant.*

Lemma 2.6.4. *(Bernstein's inequality, Vershynin (2018, Theorem 2.8.1)) Let $\xi_1, \ldots, \xi_n$ be independent, mean zero, sub-exponential random variables. Then, for every $t \geq 0$, we have,*

$$\mathbb{P}\left(\sum_{i=1}^{n} \xi_i \geq t\right) \leq \exp\left\{-c \min\left(\frac{t^2}{\sum_{i=1}^{n} \|\xi_i\|_{\psi_1}^2}, \frac{t}{\max_i \|\xi_i\|_{\psi_1}}\right)\right\}, \tag{2.6.15}$$

where c is an absolute constant. The bound for $\mathbb{P}(\sum_{i=1}^{n} \xi_i \leq -t)$ is the same as (2.6.15).

Lemma 2.6.5. *(Hanson–Wright inequality, Vershynin (2018, Theorem 6.2.1)) Let $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)^{\mathrm{T}} \in \mathbb{R}^n$ be a random vector with independent, mean zero, sub-Gaussian coordinates. Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be a fixed matrix. Then, for every $t \geq 0$, we have*

$$\mathbb{P}\left\{\left|\boldsymbol{\xi}^{\mathrm{T}} \mathbf{A} \boldsymbol{\xi} - \mathbb{E}\left(\boldsymbol{\xi}^{\mathrm{T}} \mathbf{A} \boldsymbol{\xi}\right)\right| \geq t\right\} \leq 2 \exp\left[-c \min\left(\frac{t^2}{\max_i^4 \|\xi\|_{\psi_2} \|\mathbf{A}\|_{\mathrm{F}}^2}, \frac{t}{\max_i^2 \|\xi\|_{\psi_2} \|\mathbf{A}\|}\right)\right].$$

Lemma 2.6.6. *(Vershynin (2012, Theorem 5.39)) Let $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n]^{\mathrm{T}}$ be an $n \times d$ matrix whose rows $\mathbf{z}_i^{\mathrm{T}}$ are independent sub-Gaussian vectors such that $\mathbb{E}(\mathbf{z}_i \mathbf{z}_i^{\mathrm{T}}) = \mathbf{I}_d$. Then for every $\alpha \geq 0$, with probability at least $1 - 2\exp(-\widetilde{c}_1 \alpha^2)$, we have*

$$\sqrt{n} - \widetilde{C}\sqrt{d} - \alpha \leq s_{\min}(\mathbf{Z}) \leq s_{\max}(\mathbf{Z}) \leq \sqrt{n} + \widetilde{C}\sqrt{d} + \alpha,$$

where $\widetilde{C}$, $\widetilde{c}_1 > 0$ depend on $\max_{1 \leq i \leq n} \|\mathbf{z}_i\|_{\psi_2}$ and $\|\mathbf{z}_i\|_{\psi_2} = \sup_{\mathbf{t} \in \mathbb{R}^d} \|\mathbf{t}^{\mathrm{T}} \mathbf{z}_i\|_{\psi_2}$. Additionally, if $\mathbf{Z}$ has independent entries z_{ij} for $i = 1, \ldots, n$ and $j = 1, \ldots, d$, then $\widetilde{C}$, $\widetilde{c}_1 > 0$ depend on $K_z = \max_{1 \leq i \leq n, 1 \leq j \leq d} \|z_{ij}\|_{\psi_2}$.

Lemma 2.6.7. *Assume $\mathbf{Z} \in \mathbb{R}^{n \times d}$ satisfies Assumption 2.1.1 (ii) and (iii). Let $\widetilde{C}$ and $\widetilde{c}_1$ be absolute constants depending only on K_z.*

(i) For $0 < \alpha < n^{1/2} - \widetilde{C} d^{1/2}$, we have

$$\mathbb{P}\left\{\lambda_{\max}\left[(\mathbf{Z}^{\mathrm{T}} \mathbf{Z})^{-1}\right] > \left(\sqrt{n} - \widetilde{C}\sqrt{d} - \alpha\right)^{-2}\right\} \leq 2 \exp\left(-\widetilde{c}_1 \alpha^2\right). \tag{2.6.16}$$

(ii) For $\alpha > 0$, we have

$$\mathbb{P}\left\{\lambda_{\min}\left[(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\right] < \left(\sqrt{n} + \widetilde{C}\sqrt{d} + \alpha\right)^{-2}\right\} \leq 2\exp\left(-\widetilde{c}_1\alpha^2\right). \tag{2.6.17}$$

Proof of Lemma 2.6.7. Since $\mathbb{P}\{\mathbf{Z}^{\mathrm{T}}\mathbf{Z} \text{ is singular}\} = 0$ in Assumption 2.1.1 (iii), we have $\lambda_{\min}[(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}] > 0$.

For (2.6.16), since $\lambda_{\max}[(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}] = 1/\lambda_{\min}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})$, we have

$$\begin{aligned}
& \mathbb{P}\left\{\lambda_{\max}\left[(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\right] > \left(\sqrt{n} - \widetilde{C}\sqrt{d} - \alpha\right)^{-2}\right\} \\
\leq\ & \mathbb{P}\left\{\lambda_{\min}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right) < \left(\sqrt{n} - \widetilde{C}\sqrt{d} - \alpha\right)^{2}\right\} \\
\leq\ & \mathbb{P}\left\{\sqrt{\lambda_{\min}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)} < \left|\sqrt{n} - \widetilde{C}\sqrt{d} - \alpha\right|\right\} \\
=\ & \mathbb{P}\left\{s_{\min}\left(\mathbf{Z}\right) < \sqrt{n} - \widetilde{C}\sqrt{d} - \alpha\right\} && (2.6.18) \\
\leq\ & 2\exp\left(-\widetilde{c}_1\alpha^2\right), && (2.6.19)
\end{aligned}$$

where (2.6.19) follows from $\alpha < \sqrt{n} - \widetilde{C}\sqrt{d}$ so $|\sqrt{n} - \widetilde{C}\sqrt{d} - \alpha| = \sqrt{n} - \widetilde{C}\sqrt{d} - \alpha$, and (2.6.19) follows from Lemma 2.6.6.

Since $\lambda_{\min}[(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}] = 1/\lambda_{\max}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})$, we can prove (2.6.17) in a similar way. □

Lemma 2.6.8. *Let $\widetilde{\mathbf{v}} \in \mathbb{R}^n$ be any random vector distributed on the unit sphere S^{n-1}. Assume $\mathbf{Z} \in \mathbb{R}^{n\times d}$ satisfies Assumption 2.1.1 (ii), and $\mathbf{Z}$ is independent of $\widetilde{\mathbf{v}}$. Let $\widetilde{c}_2$ be an absolute constant depending only on K_z. Then for any $\kappa > 0$, we have*

$$\mathbb{P}\left\{\widetilde{\mathbf{v}}^{\mathrm{T}}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\widetilde{\mathbf{v}} > d + \kappa\right\} \leq \exp\left[-\widetilde{c}_2\min\left(\kappa^2/d, \kappa\right)\right], \tag{2.6.20}$$

$$\mathbb{P}\left\{\widetilde{\mathbf{v}}^{\mathrm{T}}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\widetilde{\mathbf{v}} < d + \kappa\right\} \leq \exp\left[-\widetilde{c}_2\min\left(\kappa^2/d, \kappa\right)\right]. \tag{2.6.21}$$

Proof of Lemma 2.6.8. For (2.6.20), by the independence of $\widetilde{\mathbf{v}}$ and $\mathbf{Z}$, we have

$$\mathbb{P}\left\{\widetilde{\mathbf{v}}^{\mathrm{T}}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\widetilde{\mathbf{v}} > d + \kappa\right\}$$

$$= \mathbb{E}_{\widetilde{\mathbf{v}}}\mathbb{E}_{\mathbf{Z}}\mathcal{I}\left\{\widetilde{\mathbf{v}}^{\mathrm{T}}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\widetilde{\mathbf{v}} > d+\kappa\right\}$$

$$= \mathbb{E}_{\widetilde{\mathbf{v}}}\mathbb{P}_{\mathbf{Z}}\left\{\widetilde{\mathbf{v}}^{\mathrm{T}}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\widetilde{\mathbf{v}} > d+\kappa\right\},$$

where the inner expectation $\mathbb{E}_{\mathbf{Z}}$ is taken with respect to $\mathbf{Z}$ while treating $\widetilde{\mathbf{v}}$ as a constant.

If for any given $\widetilde{\mathbf{v}} \in S^{n-1}$, $\mathbb{P}\{\widetilde{\mathbf{v}}^{\mathrm{T}}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\widetilde{\mathbf{v}} > d+\kappa\} \leq \exp[-\widetilde{c}_2\min(\kappa^2/d,\kappa)]$, then we can prove (2.6.20). We have $\widetilde{\mathbf{v}}^{\mathrm{T}}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\widetilde{\mathbf{v}} = \sum_{j=1}^{d}(\widetilde{\mathbf{v}}^{\mathrm{T}}\boldsymbol{z}_j)^2$ where $\boldsymbol{z}_j \in \mathbb{R}^n$, $j = 1,\ldots,d$, are the jth column of $\mathbf{Z}$. For any given $\widetilde{\mathbf{v}} \in S^{n-1}$, $\widetilde{\mathbf{v}}^{\mathrm{T}}\boldsymbol{z}_j$, $j = 1,\ldots,d$, are independent sub-Gaussian variables with zero mean, unit variance, and bounded sub-Gaussian norm $\|\widetilde{\mathbf{v}}^{\mathrm{T}}\boldsymbol{z}_j\|_{\psi_2}^2 \leq CK_z^2$. Therefore, $[(\widetilde{\mathbf{v}}^{\mathrm{T}}\boldsymbol{z}_j)^2-1]$, $j = 1,\ldots,d$ are independent sub-exponential variables with zero mean and bounded sub-exponential norm $\|[(\widetilde{\mathbf{v}}^{\mathrm{T}}\boldsymbol{z}_j)^2-1]\|_{\psi_1} \leq C\|(\widetilde{\mathbf{v}}^{\mathrm{T}}\boldsymbol{z}_j)^2\|_{\psi_1} = C\|\widetilde{\mathbf{v}}^{\mathrm{T}}\boldsymbol{z}_j\|_{\psi_2}^2 \leq CK_z^2$. Thus given any $\widetilde{\mathbf{v}} \in S^{n-1}$, apply Lemma 2.6.4 and we have $\mathbb{P}\{\widetilde{\mathbf{v}}^{\mathrm{T}}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\widetilde{\mathbf{v}} > d+\kappa\} \leq \exp[-\widetilde{c}_2\min(\kappa^2/d,\kappa)]$, which completes the proof of (2.6.21).

By a similar argument, we can prove (2.6.21). □

Lemma 2.6.9. *(Z. Lin and Bai (2010, 2.1.b.)) Let $\Phi(\cdot)$ and $\phi(\cdot)$ denote the cumulative distribution function and probability density function of $\mathcal{N}(0,1)$. For all $x > 0$, we have*

$$\left(\frac{1}{x}-\frac{1}{x^3}\right)\phi(x) < \frac{x}{1+x^2}\phi(x) < 1-\Phi(x) < \frac{1}{x}\phi(x).$$

2.6.2.3 Technical Lemmas for Theorem 2.4.1

Lemma 2.6.10. *For $\mathbf{w} \in \mathbb{R}^n$ under Assumption 2.1.1 (i), we have $\mathbb{P}(\mathbf{w}^{\mathrm{T}}\mathbf{w} \leq n/2) \leq \exp(-\widetilde{c}_4 n)$, where $\widetilde{c}_4$ is a constant depending only on K_w.*

Proof of Lemma 2.6.10. It follows from Lemma 2.6.4 as $\mathbb{P}(\mathbf{w}^{\mathrm{T}}\mathbf{w} \leq n/2) = \mathbb{P}\{\sum_{i=1}^{n}(w_i^2-1) \leq -n/2\} \leq \exp(-\widetilde{c}_4 n)$. □

Lemma 2.6.11. *Define $\gamma = \sqrt{\widetilde{c}_5\log n}$ with $\widetilde{c}_5 \geq 1/\{2\min(\widetilde{c}_1,\widetilde{c}_2)\}$, where $\widetilde{c}_1$ is from Lemma 2.6.7 and $\widetilde{c}_2$ is from Lemma 2.6.8. If $\mathbf{w} \in \mathbb{R}^n$ and $\mathbf{V} \in \mathbb{R}^{n\times n}$ are from Assumption 2.1.1 (i) and $n > \exp(1/\widetilde{c}_5)$, then*

$$\mathbb{P}\{\mathbf{w}^{\mathrm{T}}\mathbf{V}\mathbf{w} - n \geq \gamma^2 n\} \leq 2/\sqrt{n}.$$

Proof of Lemma 2.6.11. We have

$$\mathbb{P}\left\{\mathbf{w}^{\mathrm{T}}\mathbf{V}\mathbf{w} - n \geq \gamma^2 n\right\} \leq 2\exp\left\{-\frac{1}{2\widetilde{c}_5}\frac{\min\left(\gamma^4, \gamma^2\right)n}{\lambda_{\max}(\mathbf{V})}\right\} \leq 2\exp\left[-\frac{1}{2\widetilde{c}_5}\frac{n\gamma^2}{\lambda_{\max}(\mathbf{V})}\right] \leq \frac{2}{\sqrt{n}},$$

where the first inequality follows from the same arguments as in (2.6.3)-(2.6.5), with c in (2.6.4) replaced by $(2\widetilde{c}_5)^{-1}$ for a simpler form of γ, the second inequality uses $n > \exp(1/\widetilde{c}_5)$ so that $\gamma \geq 1$, and the last inequality uses $n/\lambda_{\max}(\mathbf{V}) \geq 1$ and the definition of γ. $\square$

Lemma 2.6.12. *If v_i and c are from Lemma 2.2.1, then under Assumption 2.1.1 (i), (iv), and $\mathbb{P}\{\boldsymbol{\varepsilon} = \mathbf{0}_n\} = 0$, we have*

$$\mathbb{P}\left(\exists\ i \in \{1, \ldots, n\} \text{ s.t. } |v_i| > \sqrt{\frac{3}{2c\lambda_{\min}(\mathbf{V})}}\sqrt{\frac{\log n}{n}}\right) \leq \frac{4}{\sqrt{n}}.$$

Proof of Lemma 2.6.12. We have

$$\begin{aligned}
& \mathbb{P}\left(\exists\ i \in \{1, \ldots, n\} \text{ s.t. } |v_i| > \sqrt{\frac{3}{2c\lambda_{\min}(\mathbf{V})}}\sqrt{\frac{\log n}{n}}\right) \\
\leq\ & \sum_{i=1}^{n} \mathbb{P}\left\{\sqrt{n}\,|\mathbf{e}_i^{\mathrm{T}}\mathbf{v}| > \sqrt{\frac{3}{2c\lambda_{\min}(\mathbf{V})}}\sqrt{\log n}\right\} \\
\leq\ & \sum_{i=1}^{n} 4\exp\left(-\frac{3}{2}\log n\right) \\
=\ & \frac{4n}{n^{3/2}} = \frac{4}{\sqrt{n}},
\end{aligned}$$

where the first inequality follows from the union bound and the second inequality follows from Lemma 2.2.1. $\square$

2.6.2.4 Technical Lemmas for Corollary 2.4.1

Lemma 2.6.13. *Suppose the block diagonal matrix $\boldsymbol{\Lambda}'$ has the form $\boldsymbol{\Lambda}' = \mathrm{diag}\{(1-\rho_1)\mathbf{I}_{n_1}, \ldots, (1-\rho_K)\mathbf{I}_{n_K}\} \in \mathbb{R}^{n\times n}$ for some constants $\rho_k \in [0, 1)$, $k = 1, \ldots, K$ and a fixed integer K. The block sizes n_k's satisfy $\sum_{k=1}^{K} n_k = n$ and $|n_k/n - r_k| \leq 1/\sqrt{n}$ for some constants $r_k \in (0, 1]$, $k = 1, \ldots, K$, such that $\sum_{k=1}^{K} r_k = 1$. Suppose $\mathbf{w} = [\mathbf{w}_1^{\mathrm{T}}, \ldots, \mathbf{w}_K^{\mathrm{T}}]^{\mathrm{T}} \sim \mathcal{N}(\mathbf{0}_n, \mathbf{I}_n)$, where*

$\mathbf{w}_k \in \mathbb{R}^{n_k}$, $k = 1, \ldots, K$ *are sub-vectors of* $\mathbf{w}$. *If* $n \geq (4/\min_{1\leq k\leq K} r_k)^2$, *then we have*

$$\mathbb{P}\left\{\left|\frac{\mathbf{w}^{\mathrm{T}}\mathbf{\Lambda}'\mathbf{w}}{n} - \sum_{k=1}^{K} r_k(1-\rho_k)\right| > \frac{1}{2}\sum_{k=1}^{K} r_k(1-\rho_k)\right\} \leq 2K\exp\left[-\widetilde{c}_6\left(\min_{1\leq k\leq K} r_k\right)n\right]. \quad (2.6.22)$$

Proof of Lemma 2.6.13. One the left-hand side of (2.6.22), we have

$$\left|\frac{\mathbf{w}^{\mathrm{T}}\mathbf{\Lambda}'\mathbf{w}}{n} - \sum_{k=1}^{K} r_k(1-\rho_k)\right| > \frac{1}{2}\sum_{k=1}^{K} r_k(1-\rho_k),$$

which implies that

$$\sum_{k=1}^{K}(1-\rho_k)|\mathbf{w}_k^{\mathrm{T}}\mathbf{w}_k - nr_k| \geq \left|\mathbf{w}^{\mathrm{T}}\mathbf{\Lambda}'\mathbf{w} - \sum_{k=1}^{K} nr_k(1-\rho_k)\right| > \sum_{k=1}^{K}\frac{1}{2}nr_k(1-\rho_k), \quad (2.6.23)$$

by $\mathbf{w}^{\mathrm{T}}\mathbf{\Lambda}'\mathbf{w} = \sum_{k=1}^{K}(1-\rho_k)\mathbf{w}_k^{\mathrm{T}}\mathbf{w}_k$. Then (2.6.23) further implies that

$$\exists\ k \in \{1, \ldots, K\},\ \text{s.t. } (1-\rho_k)|\mathbf{w}_k^{\mathrm{T}}\mathbf{w}_k - nr_k| > \frac{1}{2}nr_k(1-\rho_k). \quad (2.6.24)$$

Based on (2.6.24), we apply the union bound to obtain

$$\mathbb{P}\left\{\left|\frac{\mathbf{w}^{\mathrm{T}}\mathbf{\Lambda}'\mathbf{w}}{n} - \sum_{k=1}^{n} r_k(1-\rho_k)\right| > \frac{1}{2}\sum_{k=1}^{K} r_k(1-\rho_k)\right\} \leq \sum_{k=1}^{K}\mathbb{P}\left\{|\mathbf{w}_k^{\mathrm{T}}\mathbf{w}_k - nr_k| > \frac{n}{2}r_k\right\}. \quad (2.6.25)$$

For each $k \in \{1, \ldots, K\}$, we have

$$\begin{aligned}
& \mathbb{P}\left\{|\mathbf{w}_k^{\mathrm{T}}\mathbf{w}_k - nr_k| > \frac{n}{2}r_k\right\} \\
\leq\ & \mathbb{P}\left\{|\mathbf{w}_k^{\mathrm{T}}\mathbf{w}_k - n_k| + |n_k - nr_k| > \frac{n}{2}r_k\right\} \\
\leq\ & \mathbb{P}\left\{|\mathbf{w}_k^{\mathrm{T}}\mathbf{w}_k - n_k| > \frac{n}{4}r_k\right\} && (2.6.26) \\
\leq\ & 2\exp\left[-\widetilde{c}_6\min\left(\frac{n^2r_k^2}{n_k\|w_1^2-1\|_{\psi_1}^2}, \frac{nr_k}{\|w_1^2-1\|_{\psi_1}}\right)\right] && (2.6.27) \\
\leq\ & 2\exp\left[-\widetilde{c}_6\left(\min_{1\leq k\leq K} r_k\right)n\right], && (2.6.28)
\end{aligned}$$

where (2.6.26) follows from $|n_k/n - r_k| \le n^{-1/2}$ and $n(r_k/2 - 1/\sqrt{n}) \ge nr_k/4$ resulted from $1/\sqrt{n} \le r_k/4$, (2.6.27) follows from the Bernstein's inequality in Lemma 2.6.4, and (2.6.28) follows from $r_k/(n_k/n) \ge r_k/(r_k + 1/\sqrt{n}) \ge r_k/(r_k + r_k/4) = 4/5$, with $\|w_1^2 - 1\|_{\psi_1}$ and 4/5 absorbed by $\widetilde{c}_6$. Combining (2.6.28) with (2.6.25) completes the proof. □

Lemma 2.6.14. *Suppose that $\mathbf{\Lambda}'$, ρ_k, r_k, and $\mathbf{w}$ are from Lemma 2.6.13 and $|n_k/n - r_k| \le 1/\sqrt{n} \le r_k/4$ for $k = 1, \ldots, K$. Then we have*

$$\mathbb{E}\left[\frac{\left|\sum_{k=1}^K \frac{n_k}{n}\rho_k w_k^2 - \sum_{k=1}^K r_k \rho_k w_k^2\right|}{\sqrt{\sum_{k=1}^K \rho_k r_k w_k^2 + \sum_{k=1}^K r_k(1-\rho_k)}}\right] \le \sqrt{\frac{K}{n\left(\min_{1\le k\le K} r_k\right)}}, \tag{2.6.29}$$

$$\mathbb{E}\left[\frac{\left|\frac{\mathbf{w}^{\mathrm{T}}\mathbf{\Lambda}'\mathbf{w}}{n} - \sum_{k=1}^K r_k(1-\rho_k)\right|}{\sqrt{\sum_{k=1}^K \rho_k r_k w_k^2 + \sum_{k=1}^K r_k(1-\rho_k)}}\right] \le \frac{20}{\sqrt{3}\widetilde{c}_6}\frac{\sqrt{\sum_{k=1}^K r_k(1-\rho_k)}}{\left(\min_{1\le k\le K} r_k\right)\sqrt{n}}. \tag{2.6.30}$$

Proof of Lemma 2.6.14. For (2.6.29), if $\rho_1 = \cdots = \rho_K = 0$, the left-hand side is zero.

If there exists $\rho_k > 0$, the denominator on the left-hand side is lower bounded by $\{(\min_{1\le k\le K} r_k)\sum_{k=1}^K \rho_k w_k^2\}^{1/2}$ and the numerator on the left-hand side satisfies $|\sum_{k=1}^K (n_k/n)\rho_k w_k^2 - \sum_{k=1}^K r_k\rho_k w_k^2| \le \sum_{k=1}^K |(n_k/n) - r_k|\rho_k w_k^2 \le n^{-1/2}\sum_{k=1}^K \rho_k w_k^2$. Combining the bounds of the numerator and denominator on the left-hand side, we have

$$\text{LHS} \le \frac{\mathbb{E}\sqrt{\sum_{k=1}^K \rho_k w_k^2}}{\sqrt{n\left(\min_{1\le k\le K} r_k\right)}} \le \frac{\sqrt{\sum_{k=1}^K \rho_k \mathbb{E} w_k^2}}{\sqrt{n\left(\min_{1\le k\le K} r_k\right)}} \le \sqrt{\frac{K}{n\left(\min_{1\le k\le K} r_k\right)}},$$

where the second inequality follows from Jensen's inequality.

For (2.6.30), on the left-hand side, the denominator is lower bounded by $[\sum_{k=1}^K r_k(1-\rho_k)]^{1/2}$ and the numerator satisfies

$$\left|\frac{\mathbf{w}^{\mathrm{T}}\mathbf{\Lambda}'\mathbf{w}}{n} - \sum_{k=1}^K r_k(1-\rho_k)\right| = \left|\sum_{k=1}^K (1-\rho_k)\frac{\mathbf{w}_k^{\mathrm{T}}\mathbf{w}_k}{n} - \sum_{k=1}^K r_k(1-\rho_k)\right| \le \sum_{k=1}^K (1-\rho_k) r_k \left|\frac{\mathbf{w}_k^{\mathrm{T}}\mathbf{w}_k}{nr_k} - 1\right|.$$

Combining the bounds of the numerator and denominator on the left-hand side, we have

$$\text{LHS} \le \sqrt{\sum_{k=1}^K r_k(1-\rho_k)}\max_{1\le k\le K}\mathbb{E}\left|\frac{\mathbf{w}_k^{\mathrm{T}}\mathbf{w}_k}{nr_k} - 1\right|. \tag{2.6.31}$$

It remains to bound $\mathbb{E}|(nr)^{-1}\mathbf{w}_k^{\mathrm{T}}\mathbf{w}_k - 1|$ for $k = 1, \ldots, K$. Since $|(nr)^{-1}\mathbf{w}_k^{\mathrm{T}}\mathbf{w}_k - 1| \geq 0$, we have

$$
\begin{aligned}
& \mathbb{E}\left|(nr_k)^{-1}\mathbf{w}_k^{\mathrm{T}}\mathbf{w}_k - 1\right| \\
= & \int_0^\infty \mathbb{P}\left\{\left|(nr_k)^{-1}\mathbf{w}_k^{\mathrm{T}}\mathbf{w}_k - 1\right| > t\right\} \mathrm{d}t \\
\leq & \int_0^\infty \mathbb{P}\left\{|\mathbf{w}_k^{\mathrm{T}}\mathbf{w}_k - n_k| + |n_k - nr_k| > tnr_k\right\} \mathrm{d}t \\
\leq & \int_0^\infty \mathbb{P}\left\{|\mathbf{w}_k^{\mathrm{T}}\mathbf{w}_k - n_k| > 2^{-1}tnr_k\right\} \mathrm{d}t + \int_0^\infty \mathbb{P}\left\{|n_k - nr_k| > 2^{-1}tnr_k\right\} \mathrm{d}t. \quad (2.6.32)
\end{aligned}
$$

By Lemma 2.6.4, the first term in (2.6.32) is bounded by

$$
\begin{aligned}
& \int_0^\infty \mathbb{P}\{|\sum_{n_k \text{ terms}} (w_i^2 - 1)| > 2^{-1}tnr_k\}\mathrm{d}t \\
\leq & \; 2\int_0^\infty \exp[-\widetilde{c}_6 n \min(t^2 r_k^2, tr_k)]\mathrm{d}t \quad (2.6.33) \\
\leq & \; 2\int_0^{1/r_k} \exp(-\widetilde{c}_6 n t^2 r_k^2)\mathrm{d}t + 2\int_{1/r_k}^\infty \exp(-\widetilde{c}_6 n t r_k)\mathrm{d}t \\
\leq & \; \frac{\sqrt{\pi}}{r_k\sqrt{\widetilde{c}_6}\sqrt{n}} + \frac{2}{r_k\widetilde{c}_6 n} \leq \frac{4}{r_k\widetilde{c}_6\sqrt{n}}, \quad (2.6.34)
\end{aligned}
$$

where in the last inequality, we set $\widetilde{c}_6 < 1$ for a simpler upper bound, which does not affect the validity of the tail bound in (2.6.33).

The second term in (2.6.32) is bounded by

$$
\int_0^\infty \mathbb{P}\left\{\left|\frac{n_k}{n} - r_k\right| > \frac{tr_k}{2}\right\} \mathrm{d}t \leq \int_0^{\frac{2}{\sqrt{n}r_k}} \mathbb{P}\left\{\left|\frac{n_k}{n} - r_k\right| > \frac{tr_k}{2}\right\} \mathrm{d}t \leq \frac{2}{\sqrt{n}r_k} \leq \frac{2}{\sqrt{n}r_k\widetilde{c}_6}, \quad (2.6.35)
$$

where the first inequality follows from $|n_k/n - r_k| \leq 1/\sqrt{n}$.

Combining (2.6.32), (2.6.34), and (2.6.35), we have

$$
\max_{1\leq k\leq K} \mathbb{E}\left|\frac{\mathbf{w}_k^{\mathrm{T}}\mathbf{w}_k}{nr_k} - 1\right| \leq \frac{6}{\widetilde{c}_6\sqrt{n}\,(\min_{1\leq k\leq K} r_k)}. \quad (2.6.36)
$$

Plugging (2.6.36) into (2.6.31) completes the proof. □

Lemma 2.6.15. *If $n \geq e^2$ and $w_1, \ldots, w_K$ follow i.i.d. $\mathcal{N}(0,1)$, given $0 \leq \rho_K \leq \cdots \leq \rho_1 \leq 1 - c_{\min}$, $c_{\min} \in (0,1)$ and r_k, n_k from Lemma 2.6.13, we have*

$$\mathbb{P}\left\{\sum_{k=1}^{K} r_k \rho_k w_k^2 - \sum_{k=1}^{K} r_k \rho_k > \frac{\log n}{2\widetilde{c}_6}\right\} \leq \frac{1}{\sqrt{n}}, \tag{2.6.37}$$

$$\mathbb{P}\left\{\sum_{k=1}^{K} \frac{n_k}{n} \rho_k w_k^2 - \sum_{k=1}^{K} \frac{n_k}{n} \rho_k > \frac{\log n}{2\widetilde{c}_6}\right\} \leq \frac{1}{\sqrt{n}}. \tag{2.6.38}$$

Proof of Lemma 2.6.15. For (2.6.37), if $\rho_1 = 0$, the left-hand side is zero. If $\rho_1 > 0$, then

$$\begin{aligned}
& \mathbb{P}\left(\sum_{k=1}^{K} r_k \rho_k w_k^2 - \sum_{k=1}^{K} r_k \rho_k > \frac{\log n}{2\widetilde{c}_6}\right) \\
\leq \;& \exp\left[-\widetilde{c}_6 \min\left\{\frac{\left(\frac{\log n}{2\widetilde{c}_6}\right)^2}{\sum_{k=1}^{K} \rho_k^2 r_k^2}, \frac{\frac{\log n}{2\widetilde{c}_6}}{\max_{1\leq k\leq K}(\rho_k r_k)}\right\}\right] && (2.6.39) \\
\leq \;& \exp\left[-\widetilde{c}_6 \min\left\{\left(\frac{\log n}{2\widetilde{c}_6}\right)^2, \left(\frac{\log n}{2\widetilde{c}_6}\right)\right\}\right] && (2.6.40) \\
= \;& \exp\left(-\widetilde{c}_6 \frac{\log n}{2\widetilde{c}_6}\right) = \frac{1}{n^{1/2}}, && (2.6.41)
\end{aligned}$$

where (2.6.39) follows from Lemma 2.6.4, (2.6.40) follows from $\sum_{k=1}^{K} \rho_k^2 r_k^2 \leq \sum_{k=1}^{K} \rho_k r_k \leq \sum_{k=1}^{K} r_k = 1$ and $\max_{1\leq k\leq K} \rho_k r_k \leq 1$. In (2.6.41), we set $\widetilde{c}_6 < 1$ which does not affect the validity of the tail bound in (2.6.39). Since we assumed $n \geq e^2$, $\widetilde{c}_6 < 1$ implies $n \geq e^{2\widetilde{c}_6}$. Hence we have $\{\log n/(2\widetilde{c}_6)\}^2 \geq \log n/(2\widetilde{c}_6)$.

With similar arguments, we can prove (2.6.38). □

2.6.3 Proof of Theorem 2.3.1

Before giving the formal proof, we first provide some intuition for Theorem 2.3.1. Recall that $\mathbf{v} = \boldsymbol{\varepsilon}/\|\boldsymbol{\varepsilon}\|_2 = (\mathbf{w}^{\mathrm{T}}\mathbf{V}\mathbf{w})^{-1/2}\mathbf{V}^{1/2}\mathbf{w}$ defined in Lemma 2.2.1 of the main paper. We also define

$$\mathbf{u}_j = \frac{\boldsymbol{\Sigma}^{-1/2}\mathbf{e}_j}{\left(\mathbf{e}_j^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{e}_j\right)^{1/2}} \in \mathbb{R}^d, \tag{2.6.42}$$

for $j = 1, \ldots, d$ such that $\|\mathbf{u}_j\|_2 = 1$. Then the t statistic T_j can be rewritten as

$$\begin{aligned}
T_j &= \frac{\widehat{\beta}_j - \beta_j}{L_j} \\
&= \frac{\mathbf{e}_j^{\mathrm{T}} \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \mathbf{X}^{\mathrm{T}}\boldsymbol{\varepsilon}}{\left\{\frac{\boldsymbol{\varepsilon}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}})\boldsymbol{\varepsilon}}{n-d}\right\}^{1/2} \left\{\mathbf{e}_j^{\mathrm{T}} \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \mathbf{e}_j\right\}^{1/2}} \\
&= \frac{(n-d)^{1/2}\mathbf{e}_j^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2} \left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1} \mathbf{Z}^{\mathrm{T}}\mathbf{V}^{1/2}\mathbf{w}}{\left\{\mathbf{e}_j^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2} \left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1} \boldsymbol{\Sigma}^{-1/2}\mathbf{e}_j\right\}^{1/2} \left\{\mathbf{w}^{\mathrm{T}}\mathbf{V}^{1/2} \left(\mathbf{I}_n - \mathbf{P}_{\mathbf{Z}}\right) \mathbf{V}^{1/2}\mathbf{w}\right\}^{1/2}} \qquad (2.6.43) \\
&= \mathbf{u}_j^{\mathrm{T}} \left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1} \mathbf{Z}^{\mathrm{T}}\mathbf{v} \frac{(n-d)^{1/2}}{\left\{\mathbf{u}_j^{\mathrm{T}} \left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1} \mathbf{u}_j\right\}^{1/2}} \frac{1}{\left\{\mathbf{v}^{\mathrm{T}} \left(\mathbf{I}_n - \mathbf{P}_{\mathbf{Z}}\right) \mathbf{v}\right\}^{1/2}}, \qquad (2.6.44)
\end{aligned}$$

where (2.6.43) follows from Assumption 2.1.1 (i) and (ii). If n is large compared with d, we have $(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1} \approx n^{-1}\mathbf{I}_d$, $\{\mathbf{u}_j^{\mathrm{T}}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{u}_j\}^{-1/2} \approx n^{-1/2}$, and $\mathbf{v}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P}_{\mathbf{Z}})\mathbf{v} \approx 1$.

Then from (2.6.44), we have

$$T_j \approx \mathbf{u}_j^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{v} = \sum_{i=1}^{n} \left(\mathbf{u}_j^{\mathrm{T}}\mathbf{z}_i\right) \mathbf{v}_i, \qquad (2.6.45)$$

where $\mathbf{z}_i \in \mathbb{R}^d$ is the ith row of $\mathbf{Z}$. By Lemma 2.2.1, the self-normalized quantities $v_i = \boldsymbol{\varepsilon}_i/\|\boldsymbol{\varepsilon}\|_2$ for $i = 1, \ldots, n$ are approximately $n^{-1/2}$, even if $\boldsymbol{\varepsilon}$ has an unknown correlation structure. Hence, T_j in (2.6.45) is approximately a sum of independent elements $u_j^{\mathrm{T}} z_i$ for $i = 1, \ldots, n$, with weights of $n^{-1/2}$. Thus, $\mathbf{u}_j^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{v}$ in (2.6.45) can be treated as an asymptotically normal approximation for T_j. Unlike the traditional approach, the $n^{-1/2}$ weight is provided by the self-normalized quantity v_i, which is robust to correlated errors.

The following lemma depicts different pieces of the Berry–Esseen bound in Theorem 2.3.1.

Lemma 2.6.16. *Under Assumption 2.1.1, for any $\eta > 0$ and $t \in \mathbb{R}$, we have*

$$\sup_{t\in\mathbb{R}} |\mathbb{P}(T_j \leq t) - \Phi(t)| \leq \mathbb{P}(\mathcal{E}_\eta) + \Delta + (2\pi)^{-1/2}\eta, \qquad (2.6.46)$$

where

$$\begin{aligned} \mathcal{E}_\eta &= \left\{ \left| T_j - \mathbf{u}_j^{\mathrm{T}} \mathbf{Z}^{\mathrm{T}} \mathbf{v} \right| \geq \eta \right\}, & (2.6.47) \\ \Delta &= \sup_{t \in \mathbb{R}} \left| \mathbb{P}\left(\mathbf{u}_j^{\mathrm{T}} \mathbf{Z}^{\mathrm{T}} \mathbf{v} \leq t \right) - \Phi(t) \right|, & (2.6.48) \end{aligned}$$

and $\mathbf{v} = \|\boldsymbol{\varepsilon}\|_2^{-1} \boldsymbol{\varepsilon} = (\mathbf{w}^{\mathrm{T}} \mathbf{V} \mathbf{w})^{-1/2} \mathbf{V}^{1/2} \mathbf{w}$ *is defined in Lemma 2.2.1.*

The upper bound (2.6.46) has three components. The $\mathbb{P}(\mathcal{E}_\eta)$ term is the approximation error of using $\mathbf{u}_j^{\mathrm{T}} \mathbf{Z}^{\mathrm{T}} \mathbf{v}$ to approximate T_j. The Δ term is the Berry–Esseen bound of using $\mathbf{u}_j^{\mathrm{T}} \mathbf{Z}^{\mathrm{T}} \mathbf{v}$ to approximate $\mathcal{N}(0,1)$. The $(2\pi)^{-1/2}\eta$ term is the accumulated error from the previous approximations.

Proof of Lemma 2.6.16.

Upper bound of $\mathbb{P}(T_j \leq t) - \Phi(t)$. For any $t \in \mathbb{R}$, we have

$$\begin{aligned} \mathbb{P}\left(\frac{\widehat{\beta}_j - \beta_j}{L_j} \leq t \right) &\leq \mathbb{P}(\mathcal{E}_\eta) + \mathbb{P}\left(\frac{\widehat{\beta}_j - \beta_j}{L_j} \leq t \text{ and } \mathcal{E}_\eta^c \right) \\ &\leq \mathbb{P}(\mathcal{E}_\eta) + \mathbb{P}\left(\mathbf{u}_j^{\mathrm{T}} \mathbf{Z}^{\mathrm{T}} \mathbf{v} \leq t + \eta \right), & (2.6.49) \end{aligned}$$

where the last step in (2.6.49) follows from the fact that on the event $\mathcal{E}_\eta^c$, we have

$$\mathbf{u}_j^{\mathrm{T}} \mathbf{Z}^{\mathrm{T}} \mathbf{v} \leq \left| \mathbf{u}_j^{\mathrm{T}} \mathbf{Z}^{\mathrm{T}} \mathbf{v} - \frac{\widehat{\beta}_j - \beta_j}{L_j} \right| + \frac{\widehat{\beta}_j - \beta_j}{L_j} \leq \eta + t.$$

Then by the definition of Δ in (2.6.48), we have

$$\begin{aligned} \mathbb{P}\left(\mathbf{u}_j^{\mathrm{T}} \mathbf{Z}^{\mathrm{T}} \mathbf{v} \leq t + \eta \right) &\leq \left| \mathbb{P}\left(\mathbf{u}_j^{\mathrm{T}} \mathbf{Z}^{\mathrm{T}} \mathbf{v} \leq t + \eta \right) - \Phi(t+\eta) \right| + \Phi(t+\eta) \\ &\leq \Delta + \Phi(t+\eta). & (2.6.50) \end{aligned}$$

Combining (2.6.49) and (2.6.50), we have

$$\begin{aligned}
& \mathbb{P}\left(\frac{\widehat{\beta}_j - \beta_j}{L_j} \le t\right) - \Phi(t) \\
\le\ & \mathbb{P}(\mathcal{E}_\eta) + \mathbb{P}\left(\mathbf{u}_j^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{v} \le t + \eta\right) - \Phi(t) && (2.6.51) \\
\le\ & \mathbb{P}(\mathcal{E}_\eta) + \Delta + \Phi(t+\eta) - \Phi(t) && (2.6.52) \\
\le\ & \mathbb{P}(\mathcal{E}_\eta) + \Delta + (2\pi)^{-1/2}\eta, && (2.6.53)
\end{aligned}$$

where (2.6.51) follows from (2.6.49), (2.6.52) follows from (2.6.50), and (2.6.53) follows from the mean value theorem with $\mathrm{d}\Phi(x)/\mathrm{d}x \le (2\pi)^{-1/2}$ for all $x \in \mathbb{R}$.

Lower bound of $\mathbb{P}(T_j \le t) - \Phi(t)$. For any $t \in \mathbb{R}$, similar to (2.6.49), we have

$$\begin{aligned}
\mathbb{P}\left(\mathbf{u}_j^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{v} \le t - \eta\right) &\le \mathbb{P}(\mathcal{E}_\eta) + \mathbb{P}\left(\mathbf{u}_j^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{v} \le t - \eta \text{ and } \mathcal{E}_\eta^c\right) \\
&\le \mathbb{P}(\mathcal{E}_\eta) + \mathbb{P}\left(\frac{\widehat{\beta}_j - \beta_j}{L_j} \le t\right), && (2.6.54)
\end{aligned}$$

where the last step in (2.6.54) follows from the fact that on the even $\mathcal{E}_\eta^c$, we have

$$\frac{\widehat{\beta}_j - \beta_j}{L_j} \le \left|\frac{\widehat{\beta}_j - \beta_j}{L_j} - \mathbf{u}_j^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{v}\right| + \mathbf{u}_j^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{v} \le \eta + t - \eta = t.$$

Similar to (2.6.50), we have

$$\begin{aligned}
\mathbb{P}\left(\mathbf{u}_j^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{v} \le t - \eta\right) &= \Phi(t-\eta) + \mathbb{P}\left(\mathbf{u}_j^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{v} \le t - \eta\right) - \Phi(t-\eta) \\
&\ge \Phi(t-\eta) - \left|\mathbb{P}\left(\mathbf{u}_j^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{v} \le t - \eta\right) - \Phi(t-\eta)\right| \\
&\ge \Phi(t-\eta) - \Delta. && (2.6.55)
\end{aligned}$$

Similar to (2.6.51)–(2.6.53), we have

$$\mathbb{P}\left(\frac{\widehat{\beta}_j - \beta_j}{L_j} \le t\right) - \Phi(t)$$

$$\geq \quad -\mathbb{P}(\mathcal{E}_\eta) + \mathbb{P}\left(\mathbf{u}_j^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{v} \leq t - \eta\right) - \Phi(t) \tag{2.6.56}$$

$$\geq \quad -\mathbb{P}(\mathcal{E}_\eta) - \Delta + \Phi(t-\eta) - \Phi(t) \tag{2.6.57}$$

$$\geq \quad -\mathbb{P}(\mathcal{E}_\eta) - \Delta - (2\pi)^{-1/2}\eta, \tag{2.6.58}$$

where (2.6.56) follows from (2.6.54), (2.6.57) follows from (2.6.55), and (2.6.58) follows from the mean value theorem.

Combining (2.6.53) and (2.6.58), we have (2.6.46). □

2.6.3.1 Proof of Theorem 2.3.1

We prove Theorem 2.3.1 by bounding Δ, $\mathbb{P}(\mathcal{E}_\eta)$, and $(2\pi)^{-1/2}\eta$ in (2.6.46), separately.

Step 1. Bound Δ in (2.6.48).

In (2.6.45), for any given v, the $(\mathbf{u}_j^{\mathrm{T}}\mathbf{z}_i)v_i$ terms are independent with

$$E\{(\mathbf{u}_j^{\mathrm{T}}\mathbf{z}_i)v_i\} = 0,\ \mathrm{Var}\{(\mathbf{u}_j^{\mathrm{T}}\mathbf{z}_i)v_i\} = v_i^2,\ E(|\mathbf{u}_j^{\mathrm{T}}\mathbf{z}_i v_i|^3) = |v_i|^3 E(|\mathbf{u}_j^{\mathrm{T}}\mathbf{z}_i|^3) \leq C_1|v_i|^3,$$

where the upper bound for the third moment follows from Lemma 2.6.3:

$$\|\mathbf{u}_j^{\mathrm{T}}\mathbf{z}_i\|_{\psi_2}^2 \leq C\sum_{k=1}^d \|u_{kj}z_{ik}\|_{\psi_2}^2 = C\sum_{k=1}^d u_{kj}^2\|z_{ik}\|_{\psi_2}^2 \leq CK_z^2\sum_{k=1}^d u_{kj}^2 = CK_z^2 = C_1, \tag{2.6.59}$$

with K_z absorbed by C_1 in the last step.

So applying Berry-Esseen bound conditional on $\mathbf{v}$ yields that,

$$\begin{aligned}
\Delta &= \sup_{t\in\mathbb{R}}\left|\mathbb{E}_{\mathbf{v}}\left[\mathbb{E}_{\mathbf{Z}}\mathcal{I}\left\{\mathbf{u}_j^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{v} \leq t\right\}\right] - \Phi(t)\right| && (2.6.60)\\
&= \sup_{t\in\mathbb{R}}\left|\mathbb{E}_{\mathbf{v}}\left[\mathbb{E}_{\mathbf{Z}}\mathcal{I}\left\{\mathbf{u}_j^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{v} \leq t\right\} - \Phi(t)\right]\right| \\
&\leq \mathbb{E}_{\mathbf{v}}\sup_{t\in\mathbb{R}}\left|\mathbb{E}_{\mathbf{Z}}\mathcal{I}\left\{\mathbf{u}_j^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{v} \leq t\right\} - \Phi(t)\right| && (2.6.61)\\
&\leq C_1\mathbb{E}_{\mathbf{v}}\left(\sum_{i=1}^n |v_i|^3\right), && (2.6.62)
\end{aligned}$$

where (2.6.60) follows from $\mathbf{v} \perp\!\!\!\perp \mathbf{Z}$ and the inner expectation $\mathbb{E}_{\mathbf{Z}}$ is taken with respect to $\mathbf{Z}$ while treating $\mathbf{v}$ as a constant, (2.6.61) follows from from Jensen's inequality, and (2.6.62) follows from the Berry-Esseen bound for non-identically distributed summands and $\sum_{i=1}^{n} v_i^2 = 1$. It remains to bound $\mathbb{E}_{\mathbf{v}}(\sum_{i=1}^{n} |v_i|^3)$ in (2.6.62).

From Lemma 2.2.1, we have $\|v_i\|_{\psi_2} \lesssim [n\lambda_{\min}(\mathbf{V})]^{-1/2}$ so that $\mathbb{E}(|v_i|^3) \lesssim [n\lambda_{\min}(\mathbf{V})]^{-3/2}$ and $\mathbb{E}(\sum_{i=1}^{n} |v_i|^3) \lesssim [\lambda_{\min}(\mathbf{V})]^{-3/2} n^{-1/2}$. Therefore,

$$\Delta \leq C_1 \left[\lambda_{\min}(\mathbf{V})\right]^{-3/2} n^{-1/2}. \tag{2.6.63}$$

Step 2. Bound $\mathbb{P}(\mathcal{E}_\eta)$ and $(2\pi)^{-1/2}\eta$ together.

The η in $\mathcal{E}_\eta$ is the approximation error of $\mathbf{u}_j^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{v}$ for $(\widehat{\beta}_j - \beta_j)/L_j$. This approximation error also appears in $(2\pi)^{-1/2}\eta$ originated from $\Phi(t+\eta) - \Phi(t)$ in (2.6.52) and $\Phi(t) - \Phi(t-\eta)$ in (2.6.57). So we will find η such that both $\mathbb{P}(\mathcal{E}_\eta)$ and $(2\pi)^{-1/2}\eta$ approach zero with the desired rate $n^{-1/2} \max(d, \log n)$.

Finding such η is started by decomposing $\mathbb{P}(\mathcal{E}_\eta)$ as

$$\mathbb{P}(\mathcal{E}_\eta) = p\{\mathcal{E}_\eta \cap (\mathcal{E}_1 \cup \mathcal{E}_2)\} + \mathbb{P}\left(\mathcal{E}_\eta \cap \mathcal{E}_1^c \cap \mathcal{E}_2^c\right) \leq \mathbb{P}\left(\mathcal{E}_1\right) + \mathbb{P}\left(\mathcal{E}_2\right) + \mathbb{P}\left(\mathcal{E}_\eta \cap \mathcal{E}_1^c \cap \mathcal{E}_2^c\right), \tag{2.6.64}$$

where

$$\begin{aligned}
\mathcal{E}_1 &= \left\{\lambda_{\max}\left[(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\right] > (n^{1/2} - \widetilde{C}d^{1/2} - \alpha)^{-2}\right\} \\
&\cup \left\{\lambda_{\min}\left[(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\right] < (n^{1/2} + \widetilde{C}d^{1/2} + \alpha)^{-2}\right\}, \\
\mathcal{E}_2 &= \left\{\mathbf{v}^{\mathrm{T}}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\mathbf{v} > d + \kappa\right\}.
\end{aligned}$$

The following proofs to find η include (a) finding α in $\mathcal{E}_1$ and κ in $\mathcal{E}_2$ such that $\mathbb{P}(\mathcal{E}_1)$ and $\mathbb{P}(\mathcal{E}_2)$ approach 0 with rate $n^{-1/2}$, and (b) finding $\eta \asymp n^{-1/2} \max(d, \log n)$ with α and κ from (a) such that $\mathbb{P}(\mathcal{E}_\eta \cap \mathcal{E}_1^c \cap \mathcal{E}_2^c) = 0$:

1. Lemmas 2.6.7 and 2.6.8 imply that $\mathbb{P}\left(\mathcal{E}_1\right) \leq 4\exp\left(-\widetilde{c}_3\alpha^2\right)$ and $\mathbb{P}\left(\mathcal{E}_2\right) \leq \exp\{-\widetilde{c}_3 \min(\kappa^2/d, \kappa)\}$, where $\widetilde{c}_3 = \min\left(\widetilde{c}_1, \widetilde{c}_2\right)$ in Lemmas 2.6.7 and 2.6.8.

   To obtain the $n^{-1/2}$ rate, for $\mathbb{P}(\mathcal{E}_1)$, by choosing $\exp\left(-\widetilde{c}_3\alpha^2\right) = n^{-1/2}$, we find $\alpha = \{\log n/(2\widetilde{c}_3)\}^{1/2} = (c_3 \log n)^{1/2}$ where we define $c_3 = 1/(2\widetilde{c}_3)$.

   Similarly, for $\mathbb{P}(\mathcal{E}_2)$, let $\exp\{-\widetilde{c}_3 \min\left(\kappa^2/d, \kappa\right)\} = n^{-1/2}$. If $\kappa < d$, we choose $\exp\left(-\widetilde{c}_3\kappa^2/d\right) = n^{-1/2}$ so that $\kappa = (c_3 d \log n)^{1/2} = d^{1/2}\alpha$. If $\kappa \geq d$, we choose $\exp\left(-\widetilde{c}_3\kappa\right) = n^{-1/2}$ so that $\kappa = c_3 \log n = \alpha^2$. That is, $\kappa = \alpha \max(d^{1/2}, \alpha)$.

2. With the α and κ from (a), we will find η such that $\mathbb{P}(\mathcal{E}_\eta \cap \mathcal{E}_1^c \cap \mathcal{E}_2^c) = 0$ and $\eta \asymp n^{-1/2}\max(d, \log n)$.

   On the event $\mathcal{E}_\eta$, we have

$$\left|\frac{\widehat{\beta}_j - \beta_j}{L_j} - \mathbf{u}_j^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{v}\right| = \left|\mathbf{u}_j^{\mathrm{T}}\left[\frac{(n-d)^{1/2}}{\left\{\mathbf{u}_j^{\mathrm{T}}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{u}_j\right\}^{1/2}}\frac{\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}}{\left\{\mathbf{v}^{\mathrm{T}}\left(\mathbf{I}_n - \mathbf{P}_{\mathbf{Z}}\right)\mathbf{v}\right\}^{1/2}} - \mathbf{I}_d\right] Z^{\mathrm{T}}v\right|.$$

   If we define

$$\tilde{\mathbf{I}}_d = \frac{(n-d)^{1/2}}{\left\{\mathbf{u}_j^{\mathrm{T}}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{u}_j\right\}^{1/2}}\frac{\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}}{\left\{\mathbf{v}^{\mathrm{T}}\left(\mathbf{I}_n - \mathbf{P}_{\mathbf{Z}}\right)\mathbf{v}\right\}^{1/2}}, \tag{2.6.65}$$

   then we have

$$\begin{aligned}
&\left|\frac{\widehat{\beta}_j - \beta_j}{L_j} - \mathbf{u}_j^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{v}\right| \\
= \;& \left|\mathbf{u}_j^{\mathrm{T}}\left(\tilde{\mathbf{I}}_d - \mathbf{I}_d\right)\mathbf{Z}^{\mathrm{T}}\mathbf{v}\right| \\
= \;& \left|\frac{\mathbf{u}_j^{\mathrm{T}}}{\|\mathbf{u}_j\|_2}\left(\tilde{\mathbf{I}}_d - \mathbf{I}_d\right)\frac{\mathbf{Z}^{\mathrm{T}}\mathbf{v}}{\|\mathbf{Z}^{\mathrm{T}}\mathbf{v}\|_2}\right| \|\mathbf{u}_j\|_2 \|\mathbf{Z}^{\mathrm{T}}\mathbf{v}\|_2 \\
\leq \;& \|\tilde{\mathbf{I}}_d - \mathbf{I}_d\| \|\mathbf{Z}^{\mathrm{T}}\mathbf{v}\|_2 \\
= \;& \max\left\{\left|\lambda_{\max}\left(\tilde{\mathbf{I}}_d - \mathbf{I}_d\right)\right|, \left|\lambda_{\min}\left(\tilde{\mathbf{I}}_d - \mathbf{I}_d\right)\right|\right\}\left(\mathbf{v}^{\mathrm{T}}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\mathbf{v}\right)^{1/2} \\
= \;& \max\left[\left|\lambda_{\max}\left(\tilde{\mathbf{I}}_d\right) - 1\right|, \left|\lambda_{\min}\left(\tilde{\mathbf{I}}_d\right) - 1\right|\right]\left(\mathbf{v}^{\mathrm{T}}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\mathbf{v}\right)^{1/2}.
\end{aligned} \tag{2.6.66}$$

On the even $\mathcal{E}_\eta \cap \mathcal{E}_1^c \cap \mathcal{E}_2^c$, by some algebra, we have

$$
\begin{aligned}
& \lambda_{\max}(\tilde{\mathbf{I}}_d) - 1 \\
\leq \; & \frac{n^{1/2}\left\{\frac{nd}{\left(n^{1/2}-\widetilde{C}d^{1/2}-\alpha\right)^2} - d + \frac{n\max(d,\alpha^2)}{\left(n^{1/2}-\widetilde{C}d^{1/2}-\alpha\right)^2}\right\}^{1/2} + 3\alpha n^{1/2} + 3\widetilde{C}(dn)^{1/2}}{\left(n^{1/2}-\widetilde{C}d^{1/2}-\alpha\right)^2\left\{1-\frac{d+\max(d,\alpha^2)}{\left(n^{1/2}-\widetilde{C}d^{1/2}-\alpha\right)^2}\right\}^{1/2}}, \\
& \lambda_{\min}(\tilde{\mathbf{I}}_n) - 1 \geq -\frac{(nd)^{1/2} + 3\alpha n^{1/2} + 3\widetilde{C}(nd)^{1/2} + (\alpha+\widetilde{C}d^{1/2})^2}{\left(n^{1/2}+\widetilde{C}d^{1/2}+\alpha\right)^2}, \\
& \left(\mathbf{v}^{\mathrm{T}}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\mathbf{v}\right)^{1/2} \leq \sqrt{2}\max\left(d^{1/2},\alpha\right).
\end{aligned}
$$

By $d = o(n^{1/2})$ in Assumption 2.1.1, in $\lambda_{\max}(\tilde{\mathbf{I}}_d) - 1$, we have $n/(n^{1/2}-\widetilde{C}d^{1/2}-\alpha)^2 \to 1$ and $\{d+\max(d,\alpha^2)\}/(n^{1/2}-\widetilde{C}d^{1/2}-\alpha)^2 \to 0$. Also in $\lambda_{\min}(\tilde{\mathbf{I}}_d) - 1$, we have $n^{1/2}/(n^{1/2}+\widetilde{C}d^{1/2}+\alpha) \to 1$. Hence there exists a positive integer N such that for $n > N$, on the event $\mathcal{E}_\eta \cap \mathcal{E}_1^c \cap \mathcal{E}_2^c$, we have

$$
\begin{aligned}
\left|\frac{\widehat{\beta}_j - \beta_j}{L_j} - \mathbf{u}_j\mathbf{Z}^{\mathrm{T}}\mathbf{v}\right| \;&\leq\; \max\left\{\left|\lambda_{\max}(\tilde{\mathbf{I}}_d) - 1\right|, \left|\lambda_{\min}(\tilde{\mathbf{I}}_d) - 1\right|\right\}\left(\mathbf{v}^{\mathrm{T}}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\mathbf{v}\right)^{1/2} \\
&\leq\; C_2\frac{\max\left(d,\alpha^2\right)}{n^{1/2}}, \qquad (2.6.67)
\end{aligned}
$$

where C_2 is an absolute constant depending on $\widetilde{C}$. If η is slightly greater than the upper bound in (2.6.67), $\mathbb{P}(\mathcal{E}_\eta \cap \mathcal{E}_1^c \cap \mathcal{E}_2^c) = 0$.

So we finally choose

$$
\eta = (C_2+1)\frac{\max\left(d,\alpha^2\right)}{n^{1/2}} = (C_2+1)\frac{\max\left(d, c_3\log n\right)}{n^{1/2}}. \qquad (2.6.68)
$$

The η in (2.6.68) has the desired rate $n^{-1/2}\max(d,\log n)$. Plugging $\mathbb{P}(\mathcal{E}_\eta \cap \mathcal{E}_1^c \cap \mathcal{E}_2^c) = 0$ into (2.6.64), we also have

$$
\mathbb{P}\left(\mathcal{E}_\eta\right) \leq \mathbb{P}(\mathcal{E}_1) + \mathbb{P}(\mathcal{E}_1) \leq \frac{4}{n^{1/2}} + \frac{1}{n^{1/2}} = \frac{5}{n^{1/2}}. \qquad (2.6.69)
$$

Step 3. Combine the results. Combining (2.6.63), (2.6.68), and (2.6.69), we have

$$\mathbb{P}\left(\mathcal{E}_{\eta}\right)+\Delta+(2\pi)^{-1/2}\eta \lesssim \frac{\max\left(d,\log n\right)}{n^{1/2}\lambda_{\min}^{3/2}\left(\mathbf{V}\right)},$$

which completes the proof by (2.6.46). □

2.6.4 Proof of Theorem 2.4.1

If $h=0$, then $\widehat{\beta}_j/L_j$ is T_j in Theorem 2.3.1. So in this section, we will focus on $h>0$. We analyze the power function $\mathbb{P}(\widehat{\beta}_j/L_j>z_\alpha)=E\{\mathcal{I}(\widehat{\beta}_j/L_j>z_\alpha)\}$ by first conditioning on w and then averaging over the randomness of w.

Specifically, rewrite $\widehat{\beta}_j/L_j$ as $(\widehat{\beta}_j-\beta_j)/L_j+\beta_j/L_j$. Since $\beta_j=h\sigma(\mathbf{e}_j^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{e}_j)^{1/2}(n-d)^{-1/2}$, we have

$$\frac{\beta_j}{L_j}=\frac{h}{\left\{\mathbf{v}^{\mathrm{T}}\left(\mathbf{I}_n-\mathbf{P}_{\mathbf{Z}}\right)\mathbf{v}\right\}^{1/2}\left\{n\mathbf{u}_j^{\mathrm{T}}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{u}_j\right\}^{1/2}\left(\mathbf{w}^{\mathrm{T}}\mathbf{V}\mathbf{w}/n\right)^{1/2}}, \tag{2.6.70}$$

which can be approximated by $h(\mathbf{w}^{\mathrm{T}}\mathbf{V}\mathbf{w}/n)^{-1/2}$.

From (2.6.44), we have

$$\frac{\widehat{\beta}_j-\beta_j}{L_j}=\mathbf{u}_j^{\mathrm{T}}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{v}\frac{\sqrt{n-d}}{\sqrt{\mathbf{u}_j^{\mathrm{T}}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{u}_j}}\frac{1}{\sqrt{\mathbf{v}^{\mathrm{T}}\left(\mathbf{I}_n-\mathbf{P}_{\mathbf{Z}}\right)\mathbf{v}}}, \tag{2.6.71}$$

which can be approximated by $\mathbf{u}_j^{\mathrm{T}}\mathbf{Z}\mathbf{v}$, where $\mathbf{v}$. From Theorem 2.3.1, $(\widehat{\beta}_j-\beta_j)/L_j$ converges weakly to $\mathcal{N}(0,1)$ with the randomness from $\mathbf{Z}$ and the entries in $\mathbf{v}$ are approximately $1/\sqrt{n}$.

Since $\mathbf{Z}\perp\!\!\!\perp\mathbf{w}$ in Assumption 2.1.1 (iii), $h(\mathbf{w}^{\mathrm{T}}\mathbf{V}\mathbf{w}/n)^{-1/2}$ and $\mathbf{u}_j^{\mathrm{T}}\mathbf{Z}\mathbf{v}$ are approximately independent, and so are β_j/L_j and $(\widehat{\beta}_j-\beta_j)/L_j$. Such observation motivates us to characterize the asymptotic normality of $(\widehat{\beta}_j-\beta_j)/L_j$ given $\mathbf{w}$ first. Then we include the randomness of $\mathbf{w}$ in $h(\mathbf{w}^{\mathrm{T}}\mathbf{V}\mathbf{w}/n)^{-1/2}$.

However, not every $\mathbf{w}$ can lead to the asymptotic normality of $\mathbf{u}_j^{\mathrm{T}}\mathbf{Z}\mathbf{v}$ in $(\widehat{\beta}_j-\beta_j)/L_j$. We need to define a proper set of $\mathbf{w}$ under which $\mathbf{u}_j^{\mathrm{T}}\mathbf{Z}\mathbf{v}$ still converges to $\mathcal{N}(0,1)$. Hence we define

$$\mathcal{E}_{\text{prop}} = \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3, \tag{2.6.72}$$

with

$$\begin{aligned}
\mathcal{E}_1 &= \{\mathbf{w}^{\mathrm{T}}\mathbf{w} > n/2\}, & (2.6.73)\\
\mathcal{E}_2 &= \left\{\mathbf{w}^{\mathrm{T}}\mathbf{V}\mathbf{w} - n < \gamma^2 n\right\}, & (2.6.74)\\
\mathcal{E}_3 &= \left\{\forall\, i \in \{1, \ldots, n\},\ |v_i| \leq \sqrt{3/[2c\lambda_{\min}(\mathbf{V})]}\sqrt{\log n/n}\right\}, & (2.6.75)
\end{aligned}$$

and $\gamma = \sqrt{\widetilde{c}_5 \log n}$ in $\mathcal{E}_2$ is from Lemma 2.6.11, v_i and c in $\mathcal{E}_3$ are from Lemma 2.2.1. Event $\mathcal{E}_1$ requires $\mathbf{w}^{\mathrm{T}}\mathbf{w}$ not to be too small to avoid the degenerate case. Event $\mathcal{E}_2$ requires $\mathbf{w}^{\mathrm{T}}\mathbf{V}\mathbf{w}$ not to deviate from $\mathbb{E}(\mathbf{w}^{\mathrm{T}}\mathbf{V}\mathbf{w}) = n$ too far. Event $\mathcal{E}_3$ requires $|v_i|$ not to surpass $n^{-1/2}$ too much which may fail the asymptotic normality of $\mathbf{u}_j^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{v}$.

Given $\mathbf{w} \in \mathcal{E}_{\text{prop}}$, we characterize the approximations in (2.6.70) and (2.6.71) by the following events:

$$\begin{aligned}
\mathcal{E}_{\eta_1}^{\mathbf{w}} &= \left\{\left|\frac{\beta_j}{L_j} - \frac{h}{\sqrt{\delta}}\right| \geq \eta_1 \frac{h}{\sqrt{\delta}}\right\}, & (2.6.76)\\
\mathcal{E}_{\eta_2}^{\mathbf{w}} &= \left\{\left|\frac{\widehat{\beta}_j - \beta_j}{L_j} - \mathbf{u}_j^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{v}\right| \geq \eta_2\right\}, & (2.6.77)
\end{aligned}$$

where the superscript $\mathbf{w}$ refers to the fixed value of $\mathbf{w}$, $\delta = \mathbf{w}^{\mathrm{T}}\mathbf{V}\mathbf{w}/n$ is defined in Theorem 2.4.1, and the approximation errors are reflected by η_1 and η_2, .

The following lemma is an intermediate result for proving Theorem 2.4.1.

Lemma 2.6.17. *Under Assumption 2.1.1, we have*

$$\begin{aligned}
&\left|\mathbb{P}\left\{\frac{\widehat{\beta}_j}{L_j} > z_\alpha\right\} - \pi(h, \mathbf{V})\right|\\
\leq\ & 2\mathbb{E}_{\mathbf{w}}\left[\mathcal{I}\left\{\mathbf{w} \in \mathcal{E}_{\text{prop}}^c\right\}\right]\\
+\ & \mathbb{E}_{\mathbf{w}}\left[\mathcal{I}\{\mathbf{w} \in \mathcal{E}_{\text{prop}}\}\left(\mathbb{P}_{\mathbf{Z}}\left\{\mathcal{E}_{\eta_1}^{\mathbf{w}}\right\} + \mathbb{P}_{\mathbf{Z}}\left\{\mathcal{E}_{\eta_2}^{\mathbf{w}}\right\} + \Delta_{\mathbf{w}} + \Gamma_{\mathbf{w},\eta_1,\eta_2}(h)\right)\right], \qquad (2.6.78)
\end{aligned}$$

where

$$\Delta_{\mathbf{w}} \;=\; \sup_{t\in\mathbb{R}}\left|\mathbb{P}_{\mathbf{Z}}\left\{\mathbf{u}_j^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{v}\le t\right\}-\Phi(t)\right|, \tag{2.6.79}$$

$$\Gamma_{\mathbf{w},\eta_1,\eta_2}(h) \;=\; \overline{\Phi}\left(z_\alpha-\frac{h}{\sqrt{\delta}}-\frac{\eta_1 h}{\sqrt{\delta}}-\eta_2\right)-\overline{\Phi}\left(z_\alpha-\frac{h}{\sqrt{\delta}}+\frac{\eta_1 h}{\sqrt{\delta}}+\eta_2\right), \tag{2.6.80}$$

where $\delta=\mathbf{w}^{\mathrm{T}}\mathbf{V}\mathbf{w}/n$ *is defined in Theorem 2.4.1, the* $\mathbb{P}_{\mathbf{Z}}$ *in* $\mathbb{P}_{\mathbf{Z}}\{\mathcal{E}^{\mathbf{w}}_{\eta_1}\}$ *and* $\mathbb{P}_{\mathbf{Z}}\{\mathcal{E}^{\mathbf{w}}_{\eta_2}\}$ *is the probability measure for* $\mathbf{Z}$ *with given* $\mathbf{w}$*, and* $\overline{\Phi}(t)=1-\Phi(t)$*.*

Proof of Lemma 2.6.17. By $\mathbf{w}\perp\!\!\!\perp\mathbf{Z}$ in Assumption 2.1.1 (iii), we have

$$\begin{aligned}
&\left|\mathbb{P}\left\{\widehat{\beta}_j/L_j>z_\alpha\right\}-\pi(h,\mathbf{V})\right|\\
=\;&\left|\mathbb{E}_{\mathbf{w}}\left[\mathbb{E}_{\mathbf{Z}}\left(\mathcal{I}\left\{\widehat{\beta}_j/L_j>z_\alpha\right\}\right)\right]-\mathbb{E}_{\mathbf{w}}\left[\Phi\left(\frac{h}{\sqrt{\delta}}-z_\alpha\right)\right]\right|\\
\le\;&\mathbb{E}_{\mathbf{w}}\left|\mathbb{E}_{\mathbf{Z}}\left(\mathcal{I}\left\{\widehat{\beta}_j/L_j>z_\alpha\right\}\right)-\Phi\left(\frac{h}{\sqrt{\delta}}-z_\alpha\right)\right|\\
=\;&\mathbb{E}_{\mathbf{w}}\left|\mathbb{P}_{\mathbf{Z}}\left\{\widehat{\beta}_j/L_j>z_\alpha\right\}-\overline{\Phi}\left(z_\alpha-\frac{h}{\sqrt{\delta}}\right)\right|\\
\le\;&2\mathbb{E}_{\mathbf{w}}\left(\mathcal{I}\left\{\mathbf{w}\in\mathcal{E}^c_{\mathrm{prop}}\right\}\right)\\
+\;&\mathbb{E}_{\mathbf{w}}\left[\mathcal{I}\left\{\mathbf{w}\in\mathcal{E}_{\mathrm{prop}}\right\}\left|\mathbb{P}_{\mathbf{Z}}\left\{\frac{\widehat{\beta}_j}{L_j}>z_\alpha\right\}-\overline{\Phi}\left(z_\alpha-\frac{h}{\sqrt{\delta}}\right)\right|\right].
\end{aligned} \tag{2.6.81}$$

Next we will bound the second term in (2.6.81).

Given $\mathbf{w}\in\mathcal{E}_{\mathrm{prop}}$, we have

$$\begin{aligned}
&\mathbb{P}_{\mathbf{Z}}\left\{\mathbf{u}_j^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{v}-\eta_2+\frac{h}{\sqrt{\delta}}-\eta_1\frac{h}{\sqrt{\delta}}>z_\alpha\right\}\\
\le\;&\mathbb{P}_{\mathbf{Z}}\left\{\mathbf{u}_j^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{v}-\eta_2+\frac{h}{\sqrt{\delta}}-\eta_1\frac{h}{\sqrt{\delta}}>z_\alpha\text{ and }(\mathcal{E}^{\mathbf{w}}_{\eta_1})^c\cap(\mathcal{E}^{\mathbf{w}}_{\eta_2})^c\right\}+\mathbb{P}_{\mathbf{Z}}\left\{\mathcal{E}^{\mathbf{w}}_{\eta_1}\right\}+\mathbb{P}_{\mathbf{Z}}\left\{\mathcal{E}^{\mathbf{w}}_{\eta_2}\right\}\\
\le\;&\mathbb{P}_{\mathbf{Z}}\left\{\frac{\widehat{\beta}_j}{L_j}>z_\alpha\right\}+\mathbb{P}_{\mathbf{Z}}\left\{\mathcal{E}^{\mathbf{w}}_{\eta_1}\right\}+\mathbb{P}_{\mathbf{Z}}\left\{\mathcal{E}^{\mathbf{w}}_{\eta_2}\right\},
\end{aligned} \tag{2.6.82}$$

where the last step follows from the fact that on the event $(\mathcal{E}^{\mathbf{w}}_{\eta_1})^c \cap (\mathcal{E}^{\mathbf{w}}_{\eta_2})^c$,

$$\begin{aligned} \frac{h}{\sqrt{\delta}} - \frac{\beta_j}{L_j} &\leq \left|\frac{\beta_j}{L_j} - \frac{h}{\sqrt{\delta}}\right| < \eta_1 \frac{h}{\sqrt{\delta}}, \\ \mathbf{u}_j^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{v} - \frac{\widehat{\beta}_j - \beta_j}{L_j} &\leq \left|\frac{\widehat{\beta}_j - \beta_j}{L_j} - \mathbf{u}_j^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{v}\right| < \eta_2. \end{aligned}$$

Then by the definition of $\Delta_{\mathbf{w}}$ in (2.6.79), we have

$$\begin{aligned} & \mathbb{P}_{\mathbf{Z}}\left\{\mathbf{u}_j^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{v} - \eta_2 + \frac{h}{\sqrt{\delta}} - \eta_1\frac{h}{\sqrt{\delta}} > z_\alpha\right\} \\ = \;& \overline{\Phi}\left(z_\alpha - \frac{h}{\sqrt{\delta}} + \frac{\eta_1 h}{\sqrt{\delta}} + \eta_2\right) \\ + \;& \mathbb{P}_{\mathbf{Z}}\left\{\mathbf{u}_j^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{v} > z_\alpha - \frac{h}{\sqrt{\delta}} + \frac{\eta_1 h}{\sqrt{\delta}} + \eta_2\right\} - \overline{\Phi}\left(z_\alpha - \frac{h}{\sqrt{\delta}} + \frac{\eta_1 h}{\sqrt{\delta}} + \eta_2\right) \\ \geq \;& \overline{\Phi}\left(z_\alpha - \frac{h}{\sqrt{\delta}} + \frac{\eta_1 h}{\sqrt{\delta}} + \eta_2\right) \\ - \;& \left|\mathbb{P}_{\mathbf{Z}}\left\{\mathbf{u}_j^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{v} > z_\alpha - \frac{h}{\sqrt{\delta}} + \frac{\eta_1 h}{\sqrt{\delta}} + \eta_2\right\} - \overline{\Phi}\left(z_\alpha - \frac{h}{\sqrt{\delta}} + \frac{\eta_1 h}{\sqrt{\delta}} + \eta_2\right)\right| \\ \geq \;& \overline{\Phi}\left(z_\alpha - \frac{h}{\sqrt{\delta}} + \frac{\eta_1 h}{\sqrt{\delta}} + \eta_2\right) - \Delta_{\mathbf{w}}. \end{aligned} \tag{2.6.83}$$

Combine (2.6.82) and (2.6.83) to obtain

$$\begin{aligned} & \mathbb{P}_{\mathbf{Z}}\left\{\widehat{\beta}_j/L_j > z_\alpha\right\} - \overline{\Phi}\left(z_\alpha - h/\sqrt{\delta}\right) \\ \geq \;& \overline{\Phi}\left(z_\alpha - \frac{h}{\sqrt{\delta}} + \frac{\eta_1 h}{\sqrt{\delta}} + \eta_2\right) - \overline{\Phi}\left(z_\alpha - \frac{h}{\sqrt{\delta}}\right) - \Delta_{\mathbf{w}} - \mathbb{P}_{\mathbf{Z}}\left\{\mathcal{E}^{\mathbf{w}}_{\eta_1}\right\} - \mathbb{P}_{\mathbf{Z}}\left\{\mathcal{E}^{\mathbf{w}}_{\eta_2}\right\} \\ \geq \;& -\Gamma_{\mathbf{w},\eta_1,\eta_2}(h) - \Delta_{\mathbf{w}} - \mathbb{P}\left\{\mathcal{E}^{\mathbf{w}}_{\eta_1}\right\} - \mathbb{P}\left\{\mathcal{E}^{\mathbf{w}}_{\eta_2}\right\} \end{aligned} \tag{2.6.84}$$

Similar to (2.6.82), we have

$$\begin{aligned} \mathbb{P}_{\mathbf{Z}}\left\{\frac{\widehat{\beta}_j}{L_j} > z_\alpha\right\} &\leq \mathbb{P}_{\mathbf{Z}}\left\{\frac{\widehat{\beta}_j - \beta_j}{L_j} + \frac{\beta_j}{L_j} > z_\alpha \text{ and } (\mathcal{E}^{\mathbf{w}}_{\eta_1})^c \cap (\mathcal{E}^{\mathbf{w}}_{\eta_2})^c\right\} + \mathbb{P}_{\mathbf{Z}}\left\{\mathcal{E}^{\mathbf{w}}_{\eta_1}\right\} + \mathbb{P}_{\mathbf{Z}}\left\{\mathcal{E}^{\mathbf{w}}_{\eta_2}\right\} \\ &\leq \mathbb{P}\left\{\mathbf{u}_j^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{v} + \eta_2 + \frac{h}{\sqrt{\delta}} + \eta_1\frac{h}{\sqrt{\delta}} > z_\alpha\right\} + \mathbb{P}_{\mathbf{Z}}\left\{\mathcal{E}^{\mathbf{w}}_{\eta_1}\right\} + \mathbb{P}_{\mathbf{Z}}\left\{\mathcal{E}^{\mathbf{w}}_{\eta_2}\right\}, \end{aligned} \tag{2.6.85}$$

where the last step follows from the fact that on the event $(\mathcal{E}_{\eta_1}^{\mathbf{w}})^c \cap (\mathcal{E}_{\eta_2}^{\mathbf{w}})^c$,

$$\begin{aligned} \frac{\beta_j}{L_j} - \frac{h}{\sqrt{\delta}} &\leq \left|\frac{\beta_j}{L_j} - \frac{h}{\sqrt{\delta}}\right| < \eta_1 \frac{h}{\sqrt{\delta}}, \\ \frac{\widehat{\beta}_j - \beta_j}{L_j} - \mathbf{u}_j^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{v} &\leq \left|\frac{\widehat{\beta}_j - \beta_j}{L_j} - \mathbf{u}_j^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{v}\right| < \eta_2. \end{aligned}$$

Similar to (2.6.83) , we have

$$\begin{aligned} & \mathbb{P}_{\mathbf{Z}}\left\{\mathbf{u}_j^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{v} + \eta_2 + \frac{h}{\sqrt{\delta}} + \eta_1\frac{h}{\sqrt{\delta}} > z_\alpha\right\} \\ = & \overline{\Phi}\left(z_\alpha - \frac{h}{\sqrt{\delta}} - \frac{\eta_1 h}{\sqrt{\delta}} - \eta_2\right) \\ + & \mathbb{P}_{\mathbf{Z}}\left\{\mathbf{u}_j^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{v} > z_\alpha - \frac{h}{\sqrt{\delta}} - \frac{\eta_1 h}{\sqrt{\delta}} - \eta_2\right\} - \overline{\Phi}\left(z_\alpha - \frac{h}{\sqrt{\delta}} - \frac{\eta_1 h}{\sqrt{\delta}} - \eta_2\right) \\ \leq & \overline{\Phi}\left(z_\alpha - \frac{h}{\sqrt{\delta}} - \frac{\eta_1 h}{\sqrt{\delta}} - \eta_2\right) \\ + & \left|\mathbb{P}_{\mathbf{Z}}\left\{\mathbf{u}_j^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{v} > z_\alpha - \frac{h}{\sqrt{\delta}} - \frac{\eta_1 h}{\sqrt{\delta}} - \eta_2\right\} - \overline{\Phi}\left(z_\alpha - \frac{h}{\sqrt{\delta}} - \frac{\eta_1 h}{\sqrt{\delta}} - \eta_2\right)\right| \\ \leq & \overline{\Phi}\left(z_\alpha - \frac{h}{\sqrt{\delta}} - \frac{\eta_1 h}{\sqrt{\delta}} - \eta_2\right) + \Delta_{\mathbf{w}}. \end{aligned} \tag{2.6.86}$$

Similar to (2.6.84), combine (2.6.85) and (2.6.86) to obtain

$$\begin{aligned} & \mathbb{P}_{\mathbf{Z}}\left\{\widehat{\beta}_j/L_j > z_\alpha\right\} - \overline{\Phi}\left(z_\alpha - h/\sqrt{\delta}\right) \\ \leq & \overline{\Phi}\left(z_\alpha - \frac{h}{\sqrt{\delta}} - \frac{\eta_1 h}{\sqrt{\delta}} - \eta_2\right) - \overline{\Phi}\left(z_\alpha - \frac{h}{\sqrt{\delta}}\right) + \Delta_{\mathbf{w}} + \mathbb{P}_{\mathbf{Z}}\left\{\mathcal{E}_{\eta_1}^{\mathbf{w}}\right\} + \mathbb{P}_{\mathbf{Z}}\left\{\mathcal{E}_{\eta_2}^{\mathbf{w}}\right\} \\ \leq & \Gamma_{\mathbf{w},\eta_1,\eta_2}(h) + \Delta_{\mathbf{w}} + \mathbb{P}\left\{\mathcal{E}_{\eta_1}^{\mathbf{w}}\right\} + \mathbb{P}\left\{\mathcal{E}_{\eta_2}^{\mathbf{w}}\right\}. \end{aligned} \tag{2.6.87}$$

Therefore, given $\mathbf{w} \in \mathcal{E}_{\text{prop}}$, combining (2.6.84) and (2.6.87), we have

$$\left|\mathbb{P}_{\mathbf{Z}}\left\{\frac{\widehat{\beta}_j}{L_j} > z_\alpha\right\} - \overline{\Phi}\left(z_\alpha - \frac{h}{\sqrt{\delta}}\right)\right| \leq \Gamma_{\mathbf{w},\eta_1,\eta_2}(h) + \Delta_{\mathbf{w}} + \mathbb{P}_{\mathbf{Z}}\left\{\mathcal{E}_{\eta_1}^{\mathbf{w}}\right\} + \mathbb{P}_{\mathbf{Z}}\left\{\mathcal{E}_{\eta_2}^{\mathbf{w}}\right\}. \tag{2.6.88}$$

Plugging (2.6.88) into the second term of (2.6.81) completes the proof. □

The following lemma bounds $\mathbb{P}_{\mathbf{Z}}\{\mathcal{E}_{\eta_1}^{\mathbf{w}}\}$ and $\mathbb{P}_{\mathbf{Z}}\{\mathcal{E}_{\eta_2}^{\mathbf{w}}\}$ given $\mathbf{w} \in \mathcal{E}_{\text{prop}}$ in (2.6.78).

Lemma 2.6.18. *Let*

$$\eta_1 = (C_5+1)\frac{\max(d^{1/2},\gamma)}{n^{1/2}} \text{ and } \eta_2 = (C_2+1)\frac{\max(d,\gamma^2)}{n^{1/2}}, \tag{2.6.89}$$

where C_5 is a sufficiently large constant depending on $\widetilde{C}$ in Lemma 2.6.7, C_2 is from (2.6.68), *and $\gamma = (\widetilde{c}_5 \log n)^{1/2}$ is from in Lemma 2.6.10. Under Assumption 2.1.1, given $\mathbf{w} \in \mathcal{E}_{\text{prop}}$ defined in* (2.6.72), *there exists a positive integer N such that for any $n > N$, we have*

$$\mathbb{P}_{\mathbf{Z}}\left\{\mathcal{E}_{\eta_1}^{\mathbf{w}}\right\} \le \frac{5}{\sqrt{n}}, \quad \mathbb{P}_{\mathbf{Z}}\left\{\mathcal{E}_{\eta_2}^{\mathbf{w}}\right\} \le \frac{5}{\sqrt{n}}.$$

Proof of Lemma 2.6.18. We first show the upper bound of $\mathbb{P}_{\mathbf{Z}}\{\mathcal{E}_{\eta_2}^{\mathbf{w}}\} \lesssim n^{-1/2}$ following (2.6.65)–(2.6.69) from the proof of Theorem 2.3.1. Specifically, we have

$$\mathbb{P}_{\mathbf{Z}}\{\mathcal{E}_{\eta_2}^{\mathbf{w}}\} \le \mathbb{P}_{\mathbf{Z},0} + \mathbb{P}_{\mathbf{Z},\eta_2},$$

where

$$\mathbb{P}_{\mathbf{Z},0} = \mathbb{P}_{\mathbf{Z}}\left\{(\mathcal{E}_4^{\mathbf{w}})^c \cup (\mathcal{E}_5^{\mathbf{w}})^c \cup (\mathcal{E}_6^{\mathbf{w}})^c\right\}, \quad \mathbb{P}_{\mathbf{Z},\eta_2} = \mathbb{P}_{\mathbf{Z}}\left\{\mathcal{E}_{\eta_2}^{\mathbf{w}} \cap \mathcal{E}_4^{\mathbf{w}} \cap \mathcal{E}_5^{\mathbf{w}} \cap \mathcal{E}_6^{\mathbf{w}}\right\}, \tag{2.6.90}$$

with

$$\begin{aligned}
\mathcal{E}_4^{\mathbf{w}} &= \left\{\lambda_{\min}\left[(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\right] \ge \left(\sqrt{n} + \widetilde{C}\sqrt{d} + \gamma\right)^{-2}\right\}, \\
\mathcal{E}_5^{\mathbf{w}} &= \left\{\lambda_{\max}\left[(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\right] \le \left(\sqrt{n} - \widetilde{C}\sqrt{d} - \gamma\right)^{-2}\right\}, \\
\mathcal{E}_6^{\mathbf{w}} &= \left\{\mathbf{v}^{\mathrm{T}}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\mathbf{v} \le d + \kappa\right\}.
\end{aligned} \tag{2.6.91}$$

With fixed $\mathbf{w}$, from Lemmas 2.6.7 and 2.6.8, we still have

$$\mathbb{P}_{\mathbf{Z},0} \leq 4\exp\left(-\widetilde{c}_1\gamma^2\right) + \exp\left[-\widetilde{c}_2 \min\left(\frac{\kappa^2}{d}, \kappa\right)\right]$$

Plugging in $\gamma = \sqrt{\widetilde{c}_5 \log n}$ defined in Lemma 2.6.12, we have $\exp(-\widetilde{c}_1\gamma^2) \leq n^{-1/2}$. To make the term $\exp\{-\widetilde{c}_2 \min(\kappa^2/d, \kappa)\} \leq n^{-1/2}$ we let $\kappa = \sqrt{d}\gamma$ if $\kappa < d$ and $\kappa = \gamma^2$ if $\kappa \geq d$. Combining γ and κ, we have $\mathbb{P}_{\mathbf{Z},0} \leq 5/\sqrt{n}$.

For the probability $\mathbb{P}_{\mathbf{Z},\eta_2}$, note that on the intersection of the four events of $\mathbb{P}_{\mathbf{Z},\eta_2}$ in (2.6.90), $|L_j^{-1}(\widehat{\beta}_j - \beta_j) - \mathbf{u}_j^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{v}| \geq \eta_2$ implies that

$$\max\left[\left|\lambda_{\max}\left(\widetilde{\mathbf{I}}_d\right) - 1\right|, \left|\lambda_{\min}\left(\widetilde{\mathbf{I}}_d\right) - 1\right|\right] \sqrt{\mathbf{v}^{\mathrm{T}}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\mathbf{v}} \geq \eta_2,$$

where $\widetilde{\mathbf{I}}_d$ is defined in (2.6.65). Follow (2.6.66)–(2.6.69) in the proof of Theorem 2.3.1, we obtain that if $n > \max[16, 4(\widetilde{C}+1)^2]\max(d,\gamma^2)$ and $\eta_2 = 8(3\widetilde{C}+6)\max(d,\gamma^2)/\sqrt{n}$, $\mathbb{P}_{\mathbf{Z},\eta_2} = 0$. Therefore $\mathbb{P}_{\mathbf{Z}}\{\mathcal{E}_{\eta_2}^{\mathbf{w}}\} \leq 5/\sqrt{n}$.

We then bound $\mathbb{P}_{\mathbf{Z}}\left\{\mathcal{E}_{\eta_1}^{\mathbf{w}}\right\}$. From (2.6.70), we have

$$\mathcal{E}_{\eta_1}^{\mathbf{w}} = \left\{\frac{h}{\sqrt{\delta}}\left|\frac{1}{\sqrt{1-\mathbf{v}^{\mathrm{T}}\mathbf{P}_{\mathbf{Z}}\mathbf{v}}\sqrt{n\mathbf{u}_j^{\mathrm{T}}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{u}_j}} - 1\right| \geq \eta_1\frac{h}{\sqrt{\delta}}\right\} \subset \mathcal{E}_7^{\mathbf{w}},$$

where

$$\mathcal{E}_7^{\mathbf{w}} = \left\{\left|(1-\mathbf{v}^{\mathrm{T}}\mathbf{P}_{\mathbf{Z}}\mathbf{v})^{-1/2}\left[n\mathbf{u}_j^{\mathrm{T}}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{u}_j\right]^{-1/2} - 1\right| \geq \eta_1\right\}.$$

Hence we have

$$\mathbb{P}_{\mathbf{Z}}\{\mathcal{E}_{\eta_1}^{\mathbf{w}}\} \leq \mathbb{P}\{\mathcal{E}_7^{\mathbf{w}}\} \leq \mathbb{P}_{\mathbf{Z},0} + \mathbb{P}_{\mathbf{Z},+\eta_1} + \mathbb{P}_{\mathbf{Z},-\eta_1},$$

where $\mathbb{P}_{\mathbf{Z},0}$ is defined in (2.6.90), $\mathbb{P}_{\mathbf{Z},+\eta_1}$ is the probability of

$$\left\{(1-\mathbf{v}^{\mathrm{T}}\mathbf{P}_{\mathbf{Z}}\mathbf{v})^{-1/2}\left[n\mathbf{u}_j^{\mathrm{T}}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{u}_j\right]^{-1/2} - 1 \geq \eta_1\right\} \cap \mathcal{E}_4^{\mathbf{w}} \cap \mathcal{E}_5^{\mathbf{w}} \cap \mathcal{E}_6^{\mathbf{w}}$$

while $\mathbb{P}_{\mathbf{Z},-\eta_1}$ is the probability of

$$\left\{(1-\mathbf{v}^{\mathrm{T}}\mathbf{P}_{\mathbf{Z}}\mathbf{v})^{-1/2}\left[n\mathbf{u}_j^{\mathrm{T}}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{u}_j\right]^{-1/2}-1\leq-\eta_1\right\}\cap\mathcal{E}_4^{\mathbf{w}}\cap\mathcal{E}_5^{\mathbf{w}}\cap\mathcal{E}_6^{\mathbf{w}}$$

with $\mathcal{E}_4^{\mathbf{w}}$, $\mathcal{E}_5^{\mathbf{w}}$, and $\mathcal{E}_6^{\mathbf{w}}$ defined in (2.6.91). Therefore,

$$\mathbb{P}_{\mathbf{Z},+\eta_1}\leq\mathbb{P}_{\mathbf{Z}}\left\{\frac{1}{\sqrt{1-\frac{d+\kappa}{\left(\sqrt{n}-\widetilde{C}\sqrt{d}-\gamma\right)^2}}}\frac{1}{\sqrt{\frac{n}{\left(\sqrt{n}+\widetilde{C}\sqrt{d}+\gamma\right)^2}}}-1\geq\eta_1\right\}$$

If n is large enough compared with $\sqrt{d}$ and γ, we have $(d+\kappa)/(\sqrt{n}-\widetilde{C}\sqrt{d}-\gamma)^2\approx 0$ and $n/(\sqrt{n}+\widetilde{C}\sqrt{d}+\gamma)^2\approx 1$. Hence by some algebra, if $n>\max[16,4(\widetilde{C}+1)^2]\max(d,\gamma^2)$, we have

$$\frac{1}{\sqrt{1-\frac{d+\kappa}{\left(\sqrt{n}-\widetilde{C}\sqrt{d}-\gamma\right)^2}}}\frac{1}{\sqrt{\frac{n}{\left(\sqrt{n}+\widetilde{C}\sqrt{d}+\gamma\right)^2}}}-1\leq\frac{\sqrt{2}(5+\widetilde{C})}{\sqrt{n}}\max\left(\sqrt{d},\gamma\right).$$

If we set $\eta_1>\sqrt{2}(5+\widetilde{C})\max(\sqrt{d},\gamma)/\sqrt{n}$, $\mathbb{P}_{\mathbf{Z},+\eta_1}=0$.

For $\mathbb{P}_{\mathbf{Z},-\eta_1}$, we have

$$\begin{aligned}
&\mathbb{P}_{\mathbf{Z},-\eta_1}\\
\leq\;&\mathbb{P}_{\mathbf{Z}}\left\{\lambda_{\max}\left[(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\right]\leq\left(\sqrt{n}-\widetilde{C}\sqrt{d}-\gamma\right)^{-2},\left[n\mathbf{u}_j^{\mathrm{T}}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{u}_j\right]^{-1/2}-1\leq-\eta_1\right\}\\
\leq\;&\mathbb{P}\left\{\frac{1}{\sqrt{\frac{n}{\left(\sqrt{n}-\widetilde{C}\sqrt{d}-\gamma\right)^2}}}-1\leq-\eta_1\right\}\\
=\;&\mathbb{P}\left\{\frac{\widetilde{C}\sqrt{d}+\gamma}{\sqrt{n}}\geq\eta_1\right\}\\
\leq\;&\mathbb{P}\left\{\frac{(\widetilde{C}+1)}{\sqrt{n}}\max\left(\sqrt{d},\gamma\right)\geq\eta_1\right\}.
\end{aligned}$$

If we set $\eta_1>\sqrt{2}(5+\widetilde{C})\max(\sqrt{d},\gamma)/\sqrt{n}$, then $\mathbb{P}_{\mathbf{Z},-\eta_1}=0$.

So we choose $\eta_1 = \sqrt{2}(6+\widetilde{C})\max(\sqrt{d},\gamma)/\sqrt{n}$ which yields $\mathbb{P}_{\mathbf{Z}}\{\mathcal{E}^{\mathbf{w}}_{\eta_1}\} \leq \mathbb{P}_{\mathbf{Z},0} \leq 5/\sqrt{n}$. Together with the bounds for $\mathbb{P}_{\mathbf{Z}}\left\{\mathcal{E}^{\mathbf{w}}_{\eta_2}\right\}$, we have proved Lemma 2.6.18. □

The following lemma bounds $\Delta_{\mathbf{w}}$ given $\mathbf{w} \in \mathcal{E}_3$ in (2.6.78).

Lemma 2.6.19. *Under Assumption 2.1.1, recalling* $\mathbf{w} \in \mathcal{E}_3$ *defined in* (2.6.75) *and* $\Delta_{\mathbf{w}}$ *defined in* (2.6.79)*, we have*

$$\Delta_{\mathbf{w}} = \sup_{t\in\mathbb{R}} \left|\mathbb{P}\left\{\mathbf{u}_j^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{v} \leq t\right\} - \Phi(t)\right| \lesssim \frac{(\log n)^{3/2}}{\lambda_{\min}^{3/2}(\mathbf{V})\sqrt{n}}.$$

Proof of Lemma 2.6.19. Recall that $\mathbf{u}_j^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{v} = \sum_{i=1}^n \mathbf{u}_j^{\mathrm{T}}\mathbf{z}_i v_i$, where $\mathbf{z}_i \in \mathbb{R}^d$ is the ith row of $\mathbf{Z}$. As $\mathbf{w}$ is given, $\mathbf{v}$ defined in Lemma 2.2.1 is given, which implies that $\mathbf{u}_j^{\mathrm{T}}\mathbf{z}_i v_i, i = 1, \ldots, n$ are independent with zero mean, $\mathbb{E}(\mathbf{u}_j^{\mathrm{T}}\mathbf{z}_i v_i)^2 = v_i^2$, and

$$\mathbb{E}\left|\mathbf{u}_j^{\mathrm{T}}\mathbf{z}_i v_i\right|^3 = |v_i|^3\,\mathbb{E}\left|\mathbf{u}_j^{\mathrm{T}}\mathbf{z}_i\right|^3 \leq \mathbb{E}\left|\mathbf{u}_j^{\mathrm{T}}\mathbf{z}_i\right|^3 \left[\frac{3}{2c\lambda_{\min}(\mathbf{V})}\right]^{3/2} \frac{(\log n)^{3/2}}{n^{3/2}},$$

where the last step follows from the fact that $\mathbf{w} \in \mathcal{E}_3$ defined in (2.6.75). By (2.6.59), $\mathbb{E}|\mathbf{u}_j^{\mathrm{T}}\mathbf{z}_i|^3 \leq C_1$ where C_1 is related to K_z. From all these facts, the result follows by the Berry-Esseen bound for independent variables. □

The following lemma bounds $\Gamma_{\mathbf{w},\eta_1,\eta_2}(h)$ given $\mathbf{w} \in \mathcal{E}_1 \cap \mathcal{E}_2$ in (2.6.78).

Lemma 2.6.20. *Under Assumption 2.1.1, given* $\mathbf{w} \in \mathcal{E}_1 \cap \mathcal{E}_2$ *defined in* (2.6.73) *and* (2.6.74)*, suppose that* η_1 *and* η_2 *are from Lemma 2.6.18 and* $\gamma = \sqrt{\tilde{c}_5 \log n}$ *is defined in Lemma 2.6.10. Let* $n > \exp(1/\tilde{c}_5)$ *and* $\sqrt{n} \geq 8(3\widetilde{C}+6)\max(d,\gamma^2)$*. Then*

$$\Gamma_{\mathbf{w},\eta_1,\eta_2}(h) \leq \sqrt{\frac{2}{\pi}}\left[\frac{4}{\sqrt{\lambda_{\min}(\mathbf{V})}}\gamma\left(1+z_\alpha+\sqrt{2\log n}\right)\eta_1 + \eta_2\right].$$

Proof of Lemma 2.6.20. Define

$$\tau(z_\alpha,\gamma) = 2\sqrt{1+\gamma^2}\left(1+z_\alpha+\sqrt{2\log n}\right). \tag{2.6.92}$$

We will discuss $h \geq \tau(z_\alpha, \gamma)$ and $h < \tau(z_\alpha, \gamma)$ separately.

If $h \geq \tau(z_\alpha, \gamma)$, we have

$$\Gamma_{\mathbf{w},\eta_1,\eta_2}(h) \leq 1 - \overline{\Phi}\left(z_\alpha - \frac{h}{\sqrt{\delta}} + \eta_1 \frac{h}{\sqrt{\delta}} + \eta_2\right) = \Phi\left(z_\alpha - \frac{h}{\sqrt{\delta}} + \eta_1 \frac{h}{\sqrt{\delta}} + \eta_2\right). \quad (2.6.93)$$

Given $\sqrt{n} \geq 8(3\widetilde{C} + 6)\max(d, \gamma^2)$, we have $\eta_2 \leq 1$ and $\sqrt{n} \geq 2\sqrt{2}(6 + \widetilde{C})\max(\sqrt{d}, \gamma)$, that is, $\eta_1 \leq 1/2$. Hence

$$\begin{aligned} & \Phi\left(z_\alpha - \frac{h}{\sqrt{\delta}} + \eta_1 \frac{h}{\sqrt{\delta}} + \eta_2\right) \\ \leq\ & \Phi\left(z_\alpha - \frac{h}{\sqrt{\delta}} + \frac{1}{2}\frac{h}{\sqrt{\delta}} + 1\right) \\ \leq\ & \Phi\left(z_\alpha - \frac{1}{2}\frac{h}{\sqrt{1+\gamma^2}} + 1\right), \end{aligned} \quad (2.6.94)$$

where the last inequality follows from $\mathbf{w} \in \mathcal{E}_2$ defined in (2.6.74). Since $h \geq \tau(z_\alpha, \gamma)$ and $n \geq 2$, we have $h/(2\sqrt{1+\gamma^2}) - 1 - z_\alpha > 1$. From the Gaussian tail in Lemma 2.6.9, we have

$$\begin{aligned} & \Phi\left(z_\alpha - \frac{1}{2}\frac{h}{\sqrt{1+\gamma^2}} + 1\right) \\ \leq\ & \frac{1}{\sqrt{2\pi}}\left(\frac{h}{2\sqrt{1+\gamma^2}} - 1 - z_\alpha\right)^{-1} \exp\left[-\frac{1}{2}\left(\frac{h}{2\sqrt{1+\gamma^2}} - 1 - z_\alpha\right)^2\right] \\ \leq\ & \frac{1}{n}, \end{aligned} \quad (2.6.95)$$

where (2.6.95) follows from $1/\sqrt{2\pi} < 1$, $1/[h/(2\sqrt{1+\gamma^2}) - 1 - z_\alpha] < 1$, and $h \geq \tau(z_\alpha, \gamma)$.

Combining (2.6.93)–(2.6.95), we have that for $h \geq \tau(z_\alpha, \gamma)$,

$$\Gamma_{\mathbf{w},\eta_1,\eta_2}(h) \leq 1/n. \quad (2.6.96)$$

If $h < \tau(z_\alpha, \gamma)$, we have

$$\Gamma_{\mathbf{w},\eta_1,\eta_2}(h) \leq \sqrt{2/\pi}\left(h\eta_1/\sqrt{\delta} + \eta_2\right) \quad (2.6.97)$$

$$\leq \sqrt{2/\pi}\left(\sqrt{2/\lambda_{\min}(\mathbf{V})}\tau(z_\alpha,\gamma)\eta_1+\eta_2\right) \tag{2.6.98}$$

$$\leq \sqrt{\frac{2}{\pi}}\left[\frac{4}{\sqrt{\lambda_{\min}(\mathbf{V})}}\gamma\left(1+z_\alpha+\sqrt{2\log n}\right)\eta_1+\eta_2\right]. \tag{2.6.99}$$

where (2.6.97) follows from the mean value theorem, (2.6.98) follows from $\delta \geq \lambda_{\min}(\mathbf{V})\mathbf{w}^{\mathrm{T}}\mathbf{w}/n$ and $\mathbf{w}^{\mathrm{T}}\mathbf{w}/n > 1/2$ in $\mathcal{E}_1$ defined in (2.6.73), (2.6.99) follows from (2.6.92) and the fact that $n > \exp(1/\widetilde{c}_5)$ implies $\gamma > 1$, so $1+\gamma^2 \leq 2\gamma^2$.

Comparing (2.6.99) and (2.6.96), we choose the larger one (2.6.99) to complete the proof.

□

2.6.4.1 Proof of Theorem 2.4.1

Proof of Theorem 2.4.1. We prove Theorem 2.4.1 by bounding the terms in (2.6.78). From Lemmas 2.6.10–2.6.12, we have

$$\mathbb{E}_{\mathbf{w}}\left[\mathcal{I}\left\{\mathbf{w}\in\mathcal{E}_{\text{prop}}^c\right\}\right] = \mathbb{P}\left\{\mathbf{w}\in\mathcal{E}_{\text{prop}}^c\right\} \leq \mathbb{P}\left(\mathcal{E}_1^c\right)+\mathbb{P}\left(\mathcal{E}_2^c\right)+\mathbb{P}\left(\mathcal{E}_3^c\right) \leq \exp\left(-\widetilde{c}_4 n\right)+2/\sqrt{n}+4/\sqrt{n}. \tag{2.6.100}$$

From Lemma 2.6.18, we have

$$\mathbb{E}_{\mathbf{w}}\left[\mathcal{I}\left\{\mathbf{w}\in\mathcal{E}_{\text{prop}}\right\}\left(\mathbb{P}_{\mathbf{Z}}\left\{\mathcal{E}_{\eta_1}^{\mathbf{w}}\right\}+\mathbb{P}_{\mathbf{Z}}\left\{\mathcal{E}_{\eta_2}^{\mathbf{w}}\right\}\right)\right] \leq 10/\sqrt{n}. \tag{2.6.101}$$

From Lemma 2.6.19, we have

$$\mathbb{E}_{\mathbf{w}}\left[\mathcal{I}\left\{\mathbf{w}\in\mathcal{E}_{\text{prop}}\right\}\Delta_{\mathbf{w}}\right] \lesssim \frac{(\log n)^{3/2}}{\lambda_{\min}^{3/2}(\mathbf{V})\sqrt{n}}. \tag{2.6.102}$$

From Lemma 2.6.20, we have

$$\begin{aligned}\mathbb{E}_{\mathbf{w}}\left[\mathcal{I}\left\{\mathbf{w}\in\mathcal{E}_{\text{prop}}\right\}\Gamma_{\mathbf{w},\eta_1,\eta_2}(h)\right] &\leq \sqrt{\frac{2}{\pi}}\left[\frac{4}{\sqrt{\lambda_{\min}(\mathbf{V})}}\gamma\left(1+z_\alpha+\sqrt{2\log n}\right)\eta_1+\eta_2\right]\\ &\lesssim \frac{\log n\max(d,\log n)}{\sqrt{\lambda_{\min}(\mathbf{V})n}},\end{aligned} \tag{2.6.103}$$

where $\eta_1 = \sqrt{2}(6+\widetilde{C})\max(\sqrt{d},\gamma)/\sqrt{n}$ and $\eta_2 = 8(3\widetilde{C}+6)\max(d,\gamma^2)/\sqrt{n}$ are defined in (2.6.89), and $\gamma = \sqrt{\widetilde{c}_5 \log n}$ is defined in Lemma 2.6.10.

If $\sqrt{n} > \max\{8(3\widetilde{C}+6), \exp[1/(2\widetilde{c}_5)]\}\max(d, \widetilde{c}_5 \log n)$, it satisfies all the conditions of n in Lemmas 2.6.10-2.6.12 and 2.6.18-2.6.20. Collecting all the bounds in (2.6.100)–(2.6.103) completes the proof. □

2.6.5 Proofs of Corollary 2.4.1

We will prove a more general corollary below with Corollary 2.4.1 being a special case when $K = 1$:

Corollary 2.6.1. (Block-diagonal correlation structure) *Under Assumption 2.1.1, assume* $\mathbf{w} \sim \mathcal{N}(\mathbf{0}_n, \mathbf{I}_n)$. *Let* $\mathbf{V} = \mathrm{diag}\{\mathbf{V}_1 \ \ldots \ \mathbf{V}_K\} \in \mathbb{R}^{n\times n}$, *where* $\mathbf{V}_k = \rho_k \mathbf{1}_{n_k}\mathbf{1}_{n_k}^{\mathrm{T}} + (1-\rho_k)\mathbf{I}_{n_k}$ *for* $k = 1, \ldots, K$, *and* K *is a constant integer. The sizes of diagonal blocks in* $\mathbf{V}$ *satisfy* $\sum_{k=1}^K n_k = n$, *and for any* $k = 1, \ldots, K$, $|n_k/n - r_k| \le 1/\sqrt{n}$, *where* $r_k \in (0, 1]$ *for* $k = 1, \ldots, K$, *and* $\sum_{k=1}^K r_k = 1$. *Given* $n \ge (4/\min_{1\le k\le K} r_k)^2$ *and any absolute constant* $c_{\min} \in (0, 1)$, *for any* $h \ge 0$ *and* $\rho_k \in [0, 1-c_{\min}]$ *for* $k = 1, \ldots, K$, *we have*

$$\left|\mathbb{P}\left\{\frac{\widehat{\beta}_j}{L_j} > z_\alpha\right\} - \pi(h, \rho_1, \ldots, \rho_K, r_1, \ldots, r_K)\right| \lesssim \frac{\sqrt{\log n}\max(d, \log n)}{\{\min_{1\le k\le K} r_k\}\sqrt{n}}, \tag{2.6.104}$$

where

$$\pi(h, \rho_1, \ldots, \rho_K, r_1, \ldots, r_K) = \mathbb{E}\Phi\left(\frac{h}{\sqrt{\sum_{k=1}^K r_k\left[\rho_k Z_k^2 + 1 - \rho_k\right]}} - z_\alpha\right),$$

and $Z_1, \ldots, Z_K \overset{iid}{\sim} \mathcal{N}(0, 1)$.

In this corollary and Corollary 2.4.1 in the main paper, the normality assumption is not essential. we adopt it only for theoretical convenience. From (2.4.2) in Theorem 2.4.1, we only need to show that

$$|\pi(h, \mathbf{V}) - \pi(h, \rho_1, \ldots, \rho_K, r_1, \ldots, r_K)| \lesssim \frac{(\log n)^{1/2}\max(d, \log n)}{n^{1/2}}. \tag{2.6.105}$$

In Corollary 2.6.1, let $\mathbf{V} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^{\mathrm{T}}$ denote the eigen-decomposition of $\mathbf{V}$. The diagonal matrix $\boldsymbol{\Lambda}$ consists of eigenvalues of $\mathbf{V}$, which are $n_k\rho_k + 1 - \rho_k$ with multiplicity 1, and $1-\rho_k$ with multiplicity $n_k - 1$ for $k = 1, \ldots, K$.

Under the condition $\mathbf{w} \sim \mathcal{N}(\mathbf{0}_n, \mathbf{I}_n)$ and using the fact that $\mathbf{Q}^{\mathrm{T}}\mathbf{w}$ follows the same distribution as $\mathbf{w}$, we have

$$\pi(h, \mathbf{V}) = \mathbb{E}\left\{\Phi\left(\frac{h}{(\mathbf{w}^{\mathrm{T}}\mathbf{V}\mathbf{w}/n)^{1/2}} - z_\alpha\right)\right\} = \mathbb{E}\left\{\Phi\left(\frac{h}{(\mathbf{w}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{w}/n)^{1/2}} - z_\alpha\right)\right\}.$$

Note that $\mathbf{w}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{w} = \sum_{k=1}^{K} n_k\rho_k w_k^2 + \mathbf{w}^{\mathrm{T}}\boldsymbol{\Lambda}'\mathbf{w}$, where $\boldsymbol{\Lambda}' = \mathrm{diag}\{(1-\rho_1)\mathbf{I}_{n_1}, \ldots, (1-\rho_K)\mathbf{I}_{n_K}\} \in \mathbb{R}^{n\times n}$. Since

$$\frac{\mathbf{w}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{w}}{n} = \sum_{k=1}^{K}\frac{n_k}{n}\rho_k w_k^2 + \frac{\mathbf{w}^{\mathrm{T}}\boldsymbol{\Lambda}'\mathbf{w}}{n} \to \sum_{k=1}^{K} r_k\left(\rho_k w_k^2 + 1 - \rho_k\right)$$

in distribution by the Portmanteau Theorem, we have

$$\pi(h, \mathbf{V}) \to E\left\{\Phi\left(h\left\{\sum_{k=1}^{K} r_k\left(\rho_k w_k^2 + 1 - \rho_k\right)\right\}^{-1/2} - z_\alpha\right)\right\}. \tag{2.6.106}$$

The following proof characterizes the convergence rate of (2.6.106).

Proof of Corollary 2.6.1. First from the definition of $\pi(h, \mathbf{V})$ and $\pi(h, \rho_1, \ldots, \rho_K, r_1, \ldots, r_K)$, we get

$$\begin{aligned}
& |\pi(h, \mathbf{V}) - \pi(h, \rho_1, \ldots, \rho_K, r_1, \ldots, r_K)| \\
\leq\ & \mathbb{E}\left|\Phi\left(\frac{h}{\sqrt{\mathbf{w}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{w}/n}} - z_\alpha\right) - \Phi\left(\frac{h}{\sqrt{\sum_{k=1}^{K} r_k\left[\rho_k w_k^2 + 1 - \rho_k\right]}} - z_\alpha\right)\right| \\
=\ & \mathbb{E}\mathcal{I}_{\{\mathcal{E}_1\}}\left|\Phi\left(\frac{h}{\sqrt{\mathbf{w}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{w}/n}} - z_\alpha\right) - \Phi\left(\frac{h}{\sqrt{\sum_{k=1}^{K} r_k\left[\rho_k w_k^2 + 1 - \rho_k\right]}} - z_\alpha\right)\right| + I_1 \\
\leq\ & 2\mathbb{P}\left\{\mathcal{E}_1\right\} + I_1,
\end{aligned} \tag{2.6.107}$$

where

$$\mathcal{E}_1 = \left\{ \left| \frac{\mathbf{w}^{\mathrm{T}} \mathbf{\Lambda}' \mathbf{w}}{n} - \sum_{k=1}^{K} r_k (1-\rho_k) \right| > \frac{1}{2} \sum_{k=1}^{K} r_k (1-\rho_k) \right\},$$

and

$$I_1 = \mathbb{E}\mathcal{I}_{\{\mathcal{E}_1^c\}} \left| \Phi\left(\frac{h}{\sqrt{\mathbf{w}^{\mathrm{T}} \mathbf{\Lambda} \mathbf{w}/n}} - z_\alpha \right) - \Phi\left(\frac{h}{\sqrt{\sum_{k=1}^{K} r_k \left[\rho_k w_k^2 + 1 - \rho_k\right]}} - z_\alpha \right) \right|.$$

Next we will bound I_1 given different ranges of h.

Case 1. If $h > \left(z_\alpha + \sqrt{\log n}\right) \sqrt{5/2 + \log n/(2\widetilde{c}_6)}$, where $\widetilde{c}_6$ is from Lemma 2.6.13, we have

$$\begin{aligned} I_1 &= \mathbb{E}\mathcal{I}_{\{\mathcal{E}_1^c\}} \left| \Phi\left(\frac{h}{\sqrt{\mathbf{w}^{\mathrm{T}} \mathbf{\Lambda} \mathbf{w}/n}} - z_\alpha \right) - 1 + 1 - \Phi\left(\frac{h}{\sqrt{\sum_{k=1}^{K} r_k \left[\rho_k w_k^2 + 1 - \rho_k\right]}} - z_\alpha \right) \right| \\ &= \mathbb{E}\mathcal{I}_{\{\mathcal{E}_1^c\}} \left| \Phi\left(z_\alpha - \frac{h}{\sqrt{\mathbf{w}^{\mathrm{T}} \mathbf{\Lambda} \mathbf{w}/n}} \right) - \Phi\left(z_\alpha - \frac{h}{\sqrt{\sum_{k=1}^{K} r_k \left[\rho_k w_k^2 + 1 - \rho_k\right]}} \right) \right| \\ &\leq I_2 + I_3, \end{aligned}$$

where

$$I_2 = \mathbb{E}\mathcal{I}_{\{\mathcal{E}_1^c\}} \Phi\left(z_\alpha - \frac{h}{\sqrt{\mathbf{w}^{\mathrm{T}} \mathbf{\Lambda} \mathbf{w}/n}} \right), \ I_3 = \mathbb{E}\mathcal{I}_{\{\mathcal{E}_1^c\}} \Phi\left(z_\alpha - \frac{h}{\sqrt{\sum_{k=1}^{K} r_k \left[\rho_k w_k^2 + 1 - \rho_k\right]}} \right).$$

To bound I_2, note that on the event $\mathcal{E}_1^c$, we have

$$\frac{\mathbf{w}^{\mathrm{T}} \mathbf{\Lambda}' \mathbf{w}}{n} \leq \frac{3}{2} \sum_{k=1}^{K} r_k (1-\rho_k) \leq \frac{3}{2} \sum_{k=1}^{K} r_k = \frac{3}{2} \ \Rightarrow \ \frac{\mathbf{w}^{\mathrm{T}} \mathbf{\Lambda} \mathbf{w}}{n} \leq \sum_{k=1}^{K} \frac{n_k}{n} \rho_k w_k^2 + \frac{3}{2}.$$

Hence

$$I_2 \leq \mathbb{E}\Phi\left(z_\alpha - \frac{h}{\sqrt{\sum_{k=1}^{K} \frac{n_k}{n} \rho_k w_k^2 + \frac{3}{2}}} \right)$$

$$
\begin{aligned}
&= \mathbb{E}\mathcal{I}\left\{\sum_{k=1}^K \frac{n_k}{n}\rho_k w_k^2 - \sum_{k=1}^K \frac{n_k}{n}\rho_k > \frac{\log n}{2\tilde{c}_6}\right\}\Phi\left(z_\alpha - \frac{h}{\sqrt{\sum_{k=1}^K \frac{n_k}{n}\rho_k w_k^2 + \frac{3}{2}}}\right) \\
&\quad + \mathbb{E}\mathcal{I}\left\{\sum_{k=1}^K \frac{n_k}{n}\rho_k w_k^2 - \sum_{k=1}^K \frac{n_k}{n}\rho_k \le \frac{\log n}{2\tilde{c}_6}\right\}\Phi\left(z_\alpha - \frac{h}{\sqrt{\sum_{k=1}^K \frac{n_k}{n}\rho_k w_k^2 + \frac{3}{2}}}\right)
\end{aligned} \tag{2.6.108}
$$

By Lemma 2.6.15, the first term in (2.6.108) is bounded by

$$
\mathbb{P}\left\{\sum_{k=1}^K \frac{n_k}{n}\rho_k w_k^2 - \sum_{k=1}^K \frac{n_k}{n}\rho_k > \frac{\log n}{2\tilde{c}_6}\right\} \le \frac{1}{\sqrt{n}}.
$$

The second term in (2.6.108) satisfies

$$
\begin{aligned}
&\mathbb{E}\mathcal{I}\left\{\sum_{k=1}^K \frac{n_k}{n}\rho_k w_k^2 - \sum_{k=1}^K \frac{n_k}{n}\rho_k \le \frac{\log n}{2\tilde{c}_6}\right\}\Phi\left(z_\alpha - \frac{h}{\sqrt{\sum_{k=1}^K \frac{n_k}{n}\rho_k w_k^2 + \frac{3}{2}}}\right) \\
\le\ & \mathbb{E}\Phi\left(z_\alpha - \frac{h}{\sqrt{\sum_{k=1}^K \frac{n_k}{n}\rho_k + \frac{\log n}{2\tilde{c}_6} + \frac{3}{2}}}\right) \\
\le\ & \mathbb{E}\Phi\left(z_\alpha - \frac{h}{\sqrt{\frac{\log n}{2\tilde{c}_6} + \frac{5}{2}}}\right) && (2.6.109) \\
\le\ & \frac{1}{\sqrt{2\pi n}}, && (2.6.110)
\end{aligned}
$$

where (2.6.109) follows from $\sum_{k=1}^K n_k\rho_k/n \le 1$, (2.6.110) follows from the case 1 $h > \left(z_\alpha + \sqrt{\log n}\right)\sqrt{5/2 + \log n/(2\tilde{c}_6)}$ and $\Phi(-\sqrt{\log n}) = 1 - \Phi(\sqrt{\log n}) \le (2\pi n)^{-1/2}$ from Lemma 2.6.9. Combining the bounds of the first and the second term, $I_2 \le n^{-1/2} + (2\pi n)^{-1/2}$.

By similar arguments with Lemma 2.6.15, one has $I_3 \le n^{-1/2} + (2\pi n)^{-1/2}$ which implies that if $h > \left(z_\alpha + \sqrt{\log n}\right)\sqrt{5/2 + \log n/(2\tilde{c}_6)}$, $I_1 \lesssim 1/\sqrt{n}$ in (2.6.107).

Case 2. If $0 \le h \le \left(z_\alpha + \sqrt{\log n}\right)\sqrt{5/2 + \log n/(2\tilde{c}_6)}$, we can bound I_1 in (2.6.107) with Lagrange mean value theorem:

$$
I_1 \le \frac{h}{\sqrt{2\pi}}\mathbb{E}\mathcal{I}_{\{\mathcal{E}_1^c\}}\left|\frac{1}{\sqrt{\frac{\mathbf{w}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{w}}{n}}} - \frac{1}{\sqrt{\sum_{k=1}^K r_k\left[\rho_k w_k^2 + 1 - \rho_k\right]}}\right|
$$

$$= \frac{h}{\sqrt{2\pi}} \mathbb{E}\mathcal{I}_{\{\mathcal{E}_1^c\}} \frac{\left|\sqrt{\frac{\mathbf{w}^{\mathrm{T}}\mathbf{\Lambda}\mathbf{w}}{n}} - \sqrt{\sum_{k=1}^K r_k \left[\rho_k w_k^2 + 1 - \rho_k\right]}\right|}{\sqrt{\frac{\mathbf{w}^{\mathrm{T}}\mathbf{\Lambda}\mathbf{w}}{n}}\sqrt{\sum_{k=1}^K r_k \left[\rho_k w_k^2 + 1 - \rho_k\right]}}.$$

Note that on the event $\mathcal{E}_1^c$, the denominator $\mathbf{w}^{\mathrm{T}}\mathbf{\Lambda}\mathbf{w}/n \geq \mathbf{w}^{\mathrm{T}}\mathbf{\Lambda}'\mathbf{w}/n \geq \frac{1}{2}\sum_{k=1}^K r_k(1-\rho_k)$, and $\sum_{k=1}^K r_k\left(\rho_k w_k^2 + 1 - \rho_k\right) \geq \sum_{k=1}^K r_k(1-\rho_k)$ so that

$$\begin{aligned} I_1 &\leq \frac{h}{\sqrt{2\pi}} \frac{\sqrt{2}}{\sum_{k=1}^K r_k(1-\rho_k)} \mathbb{E}\left|\sqrt{\frac{\mathbf{w}^{\mathrm{T}}\mathbf{\Lambda}\mathbf{w}}{n}} - \sqrt{\sum_{k=1}^K r_k \left[\rho_k w_k^2 + 1 - \rho_k\right]}\right| \\ &= \frac{h}{\sqrt{2\pi}} \frac{\sqrt{2}}{\sum_{k=1}^K r_k(1-\rho_k)} \mathbb{E}\frac{\left|\frac{\mathbf{w}^{\mathrm{T}}\mathbf{\Lambda}\mathbf{w}}{n} - \sum_{k=1}^K r_k\rho_k w_k^2 - \sum_{k=1}^K r_k(1-\rho_k)\right|}{\sqrt{\frac{\mathbf{w}^{\mathrm{T}}\mathbf{\Lambda}\mathbf{w}}{n}} + \sqrt{\sum_{k=1}^K r_k\left[\rho_k w_k^2 + 1 - \rho_k\right]}}. \end{aligned}$$

Since $\mathbf{w}^{\mathrm{T}}\mathbf{\Lambda}\mathbf{w}/n = \sum_{k=1}^K (n_k/n)\rho_k w_k^2 + \mathbf{w}^{\mathrm{T}}\mathbf{\Lambda}'\mathbf{w}/n$, we split the expression above into two parts,

$$\begin{aligned} I_1 &\leq \frac{h}{\sqrt{2\pi}} \frac{\sqrt{2}}{\sum_{k=1}^K r_k(1-\rho_k)} \mathbb{E}\left[\frac{\left|\sum_{k=1}^K \frac{n_k}{n}\rho_k w_k^2 - \sum_{k=1}^K r_k\rho_k w_k^2\right|}{\sqrt{\frac{\mathbf{w}^{\mathrm{T}}\mathbf{\Lambda}\mathbf{w}}{n}} + \sqrt{\sum_{k=1}^K \rho_k r_k w_k^2 + \sum_{k=1}^K r_k(1-\rho_k)}}\right] \\ &\quad + \frac{h}{\sqrt{2\pi}} \frac{\sqrt{2}}{\sum_{k=1}^K r_k(1-\rho_k)} \mathbb{E}\left[\frac{\left|\frac{\mathbf{w}^{\mathrm{T}}\mathbf{\Lambda}'\mathbf{w}}{n} - \sum_{k=1}^K r_k(1-\rho_k)\right|}{\sqrt{\frac{\mathbf{w}^{\mathrm{T}}\mathbf{\Lambda}\mathbf{w}}{n}} + \sqrt{\sum_{k=1}^K \rho_k r_k w_k^2 + \sum_{k=1}^K r_k(1-\rho_k)}}\right] \\ &\leq \frac{h}{\sqrt{2\pi}} \frac{\sqrt{2}}{\sum_{k=1}^K r_k(1-\rho_k)} \left[\sqrt{\frac{K}{n\left(\min_{1\leq k\leq K} r_k\right)}} + \frac{20}{\sqrt{3}\widetilde{c}_6} \frac{\sqrt{\sum_{k=1}^K r_k(1-\rho_k)}}{\left(\min_{1\leq k\leq K} r_k\right)\sqrt{n}}\right] \quad (2.6.111) \\ &\leq \frac{hK}{\sqrt{\pi}c_{\min}\left(\min_{1\leq k\leq K} r_k\right)\sqrt{n}} + \frac{20h}{\sqrt{3\pi}\widetilde{c}_6\sqrt{c_{\min}}\left(\min_{1\leq k\leq K} r_k\right)\sqrt{n}} \quad (2.6.112) \\ &\lesssim \frac{\log n}{\sqrt{n}\left(\min_{1\leq k\leq K} r_k\right)}, \quad (2.6.113) \end{aligned}$$

where (2.6.111) follows from Lemma 2.6.14, (2.6.112) follows from $\sum_{k=1}^K r_k(1-\rho_k) \geq c_{\min}\sum_{k=1}^K r_k = c_{\min}$, and (2.6.113) uses the fact that $0 \leq h \leq \{z_\alpha + (\log n)^{1/2}\}\{5/2 + \log n/(2\widetilde{c}_6)\}^{1/2}$.

Collecting the results from **Case 1** and **Case 2**, we conclude that for any $h \geq 0$, $I_1 \lesssim \log n/\{\sqrt{n}\left(\min_{1\leq k\leq K} r_k\right)\}$. Additionally, by Lemma 2.6.13, $\mathbb{P}\left\{\mathcal{E}_1\right\} \leq 2K\exp\left[-\widetilde{c}_6\left(\min_{1\leq k\leq K} r_k\right)n\right]$. Plugging the bounds of I_1 and $\mathbb{P}\{\mathcal{E}_1\}$ into (2.6.107), we have

$$|\pi(h,\mathbf{V}) - \pi(h,\rho_1,\ldots,\rho_K,r_1,\ldots,r_K)| \lesssim \exp\left[-\widetilde{c}_6\left(\min_{1\leq k\leq K} r_k\right)n\right] + \frac{\log n}{\sqrt{n}\left(\min_{1\leq k\leq K} r_k\right)}$$

$$\lesssim \frac{\log n}{n^{1/2}},$$

which concludes the proof of (2.6.105) and the final bound in (2.6.104). $\square$

2.6.6 Some Properties of $\pi(h,\rho) - \pi(h,0)$ in Corollary 2.4.1

Recall the power approximation $\pi(h,\rho) = \mathbb{E}\{\Phi(h(\rho\chi_1^2 + 1 - \rho)^{-1/2} - z_\alpha)\}$ as defined in Corollary 2.4.1, where $\rho \in [0, 1 - c_{\min}]$ with a pre-given $c_{\min} \in (0,1)$ is the correlation in $\mathbf{V}_{1,\rho} = \rho \mathbf{1}_n \mathbf{1}_n^{\mathrm{T}} + (1-\rho)\mathbf{I}_n$, and $h \in [0,\infty)$ is the signal in β_j. Define the power difference between $\rho \geq 0$ and $\rho = 0$ as

$$\Delta\pi(h,\rho) = \pi(h,\rho) - \pi(h,0). \tag{2.6.114}$$

Lemma 2.6.21. *For $\Delta\pi(h,\rho)$ defined in* (2.6.114)*, we have*

1. *Power gain with small signal: given $\rho \in (0, 1-c_{\min}]$, if $h \leq \frac{1}{2}\sqrt{1-\rho}(z_\alpha + \sqrt{z_\alpha^2 + 12})$, then $\Delta\pi(h,\rho) > 0$.*

2. *Power loss with large signal: given $\rho \in (0, 1-c_{\min}]$, if $h > \max(2z_\alpha, \{z_\alpha \int_1^\infty f(t)[1 - (\rho t + 1 - \rho)^{-1/2}]\mathrm{d}t\}^{-1}\mathbb{P}\{\chi_1^2 < 1\})$, then $\Delta\pi(h,\rho) < 0$, where $f(t)$ is the density of χ_1^2.*

3. *Diminishing power difference: given $\rho \in [0, 1-c_{\min}]$, we have $\lim_{h\to\infty} \Delta\pi(h,\rho) = 0$.*

Proof of Lemma 2.6.21.

1. For part 1 of Lemam 2.6.21, denote the χ_1^2 random variable by T, (2.6.114) reduces to $\Delta\pi(h,\rho) = \mathbb{E}_T[\Phi(h/\sqrt{\rho T + 1 - \rho} - z_\alpha) - \Phi(h - z_\alpha)]$. The integrand has second order derivative

$$\begin{aligned}
&\frac{\partial^2\left[\Phi\left(h/\sqrt{\rho T + 1 - \rho} - z_\alpha\right) - \Phi\left(h - z_\alpha\right)\right]}{\partial T^2} \\
=\;& \frac{-h\rho^2\phi\left(h/\sqrt{\rho T + 1 - \rho} - z_\alpha\right)}{4(\rho T + 1 - \rho)^{5/2}}\left[\left(h/\sqrt{\rho T + 1 - \rho}\right)^2 - z_\alpha\left(h/\sqrt{\rho T + 1 - \rho}\right) - 3\right],
\end{aligned}$$

where $\phi(\cdot)$ is the density of $\mathcal{N}(0,1)$. If $T > \rho^{-1}\{[2h/(z_\alpha + \sqrt{z_\alpha^2 + 12})]^2 - (1-\rho)\}$, the second order derivative is positive. Hence if $[2h/(z_\alpha + \sqrt{z_\alpha^2 + 12})]^2 - (1-\rho) \le 0$, the second order derivative is positive on the support $T > 0$. By Jensen's inequality, $\Phi\left(h/\sqrt{\rho T + 1 - \rho} - z_\alpha\right) - \Phi(h - z_\alpha)$ is strictly convex, which implies that $\Delta\pi(h,\rho) > 0$.

2. For part 2 of Lemma 2.6.21, rewrite $\Delta\pi(h,\rho)$ as

$$\begin{aligned}\Delta\pi(h,\rho) \quad &= \quad \int_0^1 f(t)\left[\Phi\left(\frac{h}{\sqrt{\rho t + 1 - \rho}} - z_\alpha\right) - \Phi(h - z_\alpha)\right]\mathrm{d}t \\ &\quad - \int_1^\infty f(t)\left[\Phi(h - z_\alpha) - \Phi\left(\frac{h}{\sqrt{\rho t + 1 - \rho}} - z_\alpha\right)\right]\mathrm{d}t. \quad (2.6.115)\end{aligned}$$

For the first term in (2.6.115), if $h > z_\alpha$, by Lemma 2.6.9, it is bounded by

$$\int_0^1 f(t)\left[1 - \Phi(h - z_\alpha)\right]dt \le \frac{\mathbb{P}\{\chi_1^2 < 1\}\,\phi(h - z_\alpha)}{h - z_\alpha}. \qquad (2.6.116)$$

For the second term in (2.6.115), we have

$$\begin{aligned} &\int_1^\infty f(t)\left[\Phi(h - z_\alpha) - \Phi\left(\frac{h}{\sqrt{\rho t + 1 - \rho}} - z_\alpha\right)\right]\mathrm{d}t \\ \ge\ &\int_1^\infty f(t)\phi(h - z_\alpha)\left(h - z_\alpha - \frac{h}{\sqrt{\rho t + 1 - \rho}} + z_\alpha\right)\mathrm{d}t \qquad (2.6.117) \\ \ge\ & h\phi(h - z_\alpha)\int_1^\infty f(t)\left[1 - (\rho t + 1 - \rho)^{-1/2}\right]\mathrm{d}t, \qquad (2.6.118)\end{aligned}$$

where (2.6.117) follows from the mean value theorem and the fact that if $h > 2z_\alpha$ and $t > 1$, we have $|h/\sqrt{\rho t + 1 - \rho} - z_\alpha| \le h - z_\alpha$.

Plugging (2.6.116) and (2.6.118) into (2.6.115), if $h > 2z_\alpha$, we have

$$\Delta\pi(h,\rho) \le \phi(h - z_\alpha)\left\{\frac{\mathbb{P}(\chi_1^2 < 1)}{z_\alpha} - h\int_1^\infty f(t)\left[1 - (\rho t + 1 - \rho)^{-1/2}\right]\mathrm{d}t\right\},$$

which is negative when $h > \{z_\alpha \int_1^\infty f(t)[1 - (\rho t + 1 - \rho)^{-1/2}]\mathrm{d}t\}^{-1}\mathbb{P}\{\chi_1^2 < 1\}$.

3. For part 3 in Lemma 2.6.21, by the monotone convergence theorem, for any $\rho \in [0, 1-c_{\min}]$, we have $\lim_{h\to\infty} \pi(h,\rho) = \mathbb{E}\{\lim_{h\to\infty} \Phi(h(\rho\chi_1^2 + 1 - \rho)^{-1/2} - z_\alpha)\} = 1$ which completes the proof.

$\square$

2.6.7 Tail Behaviors of the Estimation Error and Standard Error

In this section, we characterize the tail behaviors of the estimation error $\|\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta}\|_2^2$ and the squared standard error L_j^2, which are both sensitive to the correlation structure $\mathbf{V}$. We also demonstrate the important role of self-normalization for the valid inference.

Theorem 2.6.1. *Under (i)-(iii) in Condition 2.1.1, for absolute constants $C, c_1, c_2 > 0$ depending only on K_w and K_z, whenever $\sqrt{n} > C\sqrt{d}$,*

(i) with probability at least $1-\eta$, for $c_1 \exp[-c_2(\sqrt{n} - C\sqrt{d})^2] \le \eta \le 1/3$,

$$\|\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta}\|_2^2 \lesssim \log\left(\frac{1}{\eta}\right) \tau_1(n,d,\sigma^2,\boldsymbol{\Sigma})(1+\tau_3(n,\mathbf{V},\eta)); \tag{2.6.119}$$

(ii) with probability at least $1-\eta$, for $c_1 \exp\{-c_2 \min[n/\lambda_{\max}(\mathbf{V}), d]\} \le \eta \le 3/4$,

$$\|\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta}\|_2^2 \gtrsim \tau_2(n,d,\sigma^2,\boldsymbol{\Sigma})\left[1-\tau_3(n,\mathbf{V},\eta)\right]; \tag{2.6.120}$$

where $\tau_1(n,d,\sigma^2,\boldsymbol{\Sigma}) = \lambda_{\min}^{-1}(\boldsymbol{\Sigma})(\sqrt{n} - C\sqrt{d})^{-2}\sigma^2 d$, $\tau_2(n,d,\sigma^2,\boldsymbol{\Sigma}) = \lambda_{\max}^{-1}(\boldsymbol{\Sigma})(\sqrt{n} + C\sqrt{d})^{-2}\sigma^2 d$, and

$$\tau_3(n,\mathbf{V},\eta) \asymp \max\left[\sqrt{\log\left(\frac{1}{\eta}\right)\frac{\lambda_{\max}(\mathbf{V})}{n}}, \log\left(\frac{1}{\eta}\right)\frac{\lambda_{\max}(\mathbf{V})}{n}\right].$$

Theorem 2.6.1 provides concentration inequalities which quantify how $\|\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta}\|_2^2$ deviates from its convergence rate. Under linear model, $\tau_1(n,d,\sigma^2,\boldsymbol{\Sigma})$ in (2.6.119) and $\tau_2(n,d,\sigma^2,\boldsymbol{\Sigma})$ in (2.6.120) can be treated as approximations of the rate of the mean squared error (MSE).

Term $1+\tau_3(n,\mathbf{V},\eta)$ in (2.6.119) inflates the MSE while $1-\tau_3(n,\mathbf{V},\eta)$ in (2.6.120) shrinks the MSE.

The correlation structure $\mathbf{V}$ inflates or shrinks the MSE via $n^{-1}\lambda_{\max}(\mathbf{V})$ in $\tau_3(n,\mathbf{V},\eta)$. If $\boldsymbol{\varepsilon}_i$ are i.i.d. or governed by an AR(1) process, $\lambda_{\max}(\mathbf{V}) \asymp O(1)$, while if $\boldsymbol{\varepsilon}_i$ are generated by a factor model or equicorrelated $\mathbf{V}$, then $\lambda_{\max}(\mathbf{V}) \asymp n$ and the tails of $\|\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}-\boldsymbol{\beta}\|_2^2$ are enlarged by $\lambda_{\max}(\mathbf{V})$ drastically. Such sensitivity towards correlation is a result that $\|\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}-\boldsymbol{\beta}\|_2^2$ contains $\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}/(n\sigma^2)$. In essence, $\|\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}-\boldsymbol{\beta}\|_2^2$ is a quadratic form $\boldsymbol{\varepsilon}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-2}\mathbf{X}^{\mathrm{T}}\boldsymbol{\varepsilon}$ where the matrix $\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-2}\mathbf{X}^{\mathrm{T}}$ concentrates around $[1/(n\sigma^2)](\sigma^2 d/n)\boldsymbol{\Sigma}^{-1}$. The term $1/(n\sigma^2)$ multiplied by $\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}$ is sensitive toward $\mathbf{V}$ as demonstrated by Lemma 2.2.2, which produces $\tau_3(n,\mathbf{V},\eta)$.

To quantify the effects of $\mathbf{V}$ on the tail behavior of L_j, we have the following theorem:

Theorem 2.6.2. *Under Conditions (i)-(iii) in Condition 2.1.1, for absolute constants $C, c_1, c_2 > 0$ depending only on K_w and K_z,*

(i) with probability at least $1-\eta$, for $c_1\exp[-c_2(\sqrt{n}-C\sqrt{d})^2] \le \eta \le 3/4$ and $\sqrt{n} > C\sqrt{d}$

$$L_j^2 \lesssim \tau_4(n,d,\sigma^2,\boldsymbol{\Sigma})\,(1+\tau_3(n,\mathbf{V},\eta))\,; \tag{2.6.121}$$

(ii) with probability at least $1-\eta$, for $c_1\exp\{-c_2\min[n/\lambda_{\max}(\mathbf{V}),(\sqrt{n}-C\sqrt{d})^2]\} \le \eta \le 3/4$ and $\sqrt{n} \ge Cd$,

$$L_j^2 \gtrsim \tau_4(n,d,\sigma^2,\boldsymbol{\Sigma})\,(1-\tau_3(n,\mathbf{V},\eta))\,; \tag{2.6.122}$$

where $\tau_4(n,d,\sigma^2,\boldsymbol{\Sigma}) = (\sigma^2\mathbf{e}_j^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{e}_j)/(n-d)$ and $\tau_3(n,\mathbf{V},\eta)$ is defined in Theorem 2.6.1.

Theorem 2.6.2 provides concentration inequalities of L_j^2. If rows of $\mathbf{X}$ are i.i.d. from $\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma})$ and they are independent of $\boldsymbol{\varepsilon}$ following $\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I}_n)$, then $\mathbb{E}(L_j^2) = (n-d-1)^{-1}\sigma^2\mathbf{e}_j^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{e}_j$. Hence similar to Theorem 2.6.1, we could treat $\tau_4(n,d,\sigma^2,\boldsymbol{\Sigma})$ as a center approximation while $1+\tau_3(n,\mathbf{V},\eta)$ and $1-\tau_3(n,\mathbf{V},\eta)$ correspond to the inflation or shrinkage of $\tau_4(n,d,\sigma^2,\boldsymbol{\Sigma})$. The correlation structure $\mathbf{V}$ contributes to the deviation through

$\lambda_{\max}(\mathbf{V})$ in $\tau_3(n, \mathbf{V}, \eta)$, with the same pattern as discussed in Theorem 2.6.1. As we pointed out after Theorem 2.6.1, the variation in L_j originates from the quadratic form $\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}/(n\sigma^2)$, which echoes Lemma 2.2.2.

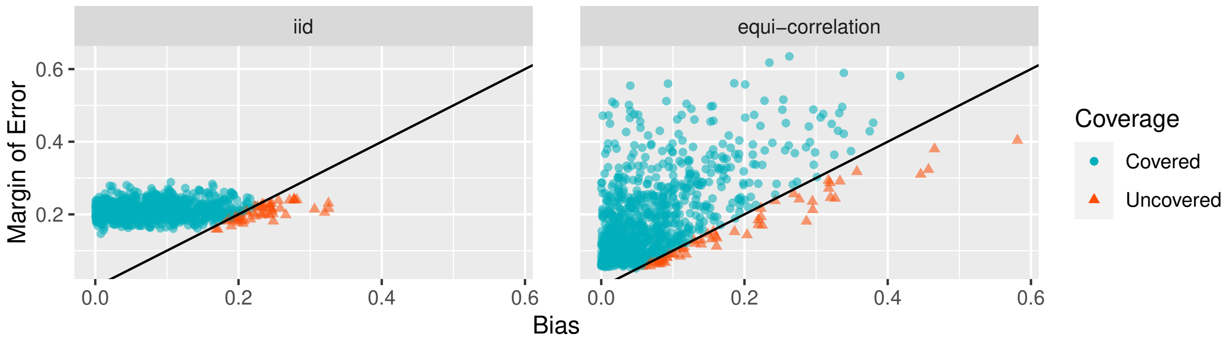


Figure 2.6: The solid line is $y = x$. With confidence interval $\widehat{\beta}_1 \pm z_{0.975}L_1$, the blue round point represents the coverage of the true value of β_1, and the red triangle point represents the failure of the coverage of the true value.

Although both of the OLS estimation error $\|\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta}\|_2$ and standard error L_j exhibit larger variation from $\mathbf{V}$, the validity of the OLS t-statistic comes from the ratio of $|\widehat{\beta}_j - \beta_j|$ and L_j, which leads to the self-normalization and cancels the effects from $\mathbf{V}$. In the setting of Figure 2.2, we compare the bias $|\widehat{\beta}_1 - \beta_1|$ and the margin of error $z_{0.975}L_1$ for OLS estimator in Figure 2.6. If $|\widehat{\beta}_1 - \beta_1| < z_{0.975}L_1$, then we have valid confidence interval. Under the correlation structure $\mathbf{V}_{1,\rho}$ in $\boldsymbol{\varepsilon}$, the left panel shows $\rho = 0$ for the independent errors, and the right panel shows $\rho = 0.9$ for the equally correlated errors. The bias and margin of error of $\widehat{\beta}_1$ under 1000 replications are presented. Both independent and equally correlated errors achieve approximately 95% coverage probability. However, the margin of error is more concentrated under independence and more dispersed under strong correlation. The variability of the margin of error comes from L_j. As shown in Theorem 2.6.2, the variation in L_j caused by $\mathbf{V}$ invalidates the traditional approach to obtaining asymptotic normality, such as White (1980); Andrews (1991); Newey and West (1994); Barrios et al. (2012); Kuchibhotla, Rinaldo, and Wasserman (2020), which sought for a consistent estimator $\widehat{\sigma}^2$

for σ^2. However, the bias in Figure 2.6 under correlated errors is adaptive to the variation in the margin of error, which indicates that the ratio of bias and standard error, $T_j = (\widehat{\beta}_j - \beta_j)/L_j$, results in a cancellation of the effects from $\mathbf{V}$, and the asymptotic distribution of T_j becomes pivotal. Such inconsistent studentized statistics are commonly used as self-normalized estimator in econometrics (Kiefer et al., 2000; Lobato, 2001).

2.6.7.1 Proof of Theorem 2.6.1

Recall that

$$\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{\varepsilon} = \sigma\boldsymbol{\Sigma}^{-1/2}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{V}^{1/2}\mathbf{w} \tag{2.6.123}$$

which yields that

$$\|\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta}\|_2^2 = \sigma^2\mathbf{w}^{\mathrm{T}}\mathbf{V}^{1/2}\mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\boldsymbol{\Sigma}^{-1}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{V}^{1/2}\mathbf{w}. \tag{2.6.124}$$

We will show part (i) of Theorem 2.6.1 first. Observe that

$$\begin{aligned} & \|\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta}\|_2^2 \\ \leq\ & \lambda_{\max}\left(\boldsymbol{\Sigma}^{-1}\right)\lambda_{\max}^2\left[\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\right]\sigma^2\mathbf{w}^{\mathrm{T}}\mathbf{V}^{1/2}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\mathbf{V}^{1/2}\mathbf{w} \\ =\ & \sigma^2\lambda_{\min}^{-1}\left(\boldsymbol{\Sigma}\right)\lambda_{\max}^2\left[\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\right]\mathbf{v}^{\mathrm{T}}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\mathbf{v}\cdot\mathbf{w}^{\mathrm{T}}\mathbf{V}\mathbf{w}, \end{aligned}$$

where $\mathbf{v}$ is defined in Lemma 2.2.1.

Then for any $t > 0$,

$$\begin{aligned} & \mathbb{P}\left\{\|\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta}\|_2^2 \geq t\right\} \\ \leq\ & \mathbb{P}\left\{\frac{\sigma^2}{\lambda_{\min}\left(\boldsymbol{\Sigma}\right)}\lambda_{\max}^2\left[\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\right]\mathbf{v}^{\mathrm{T}}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\mathbf{v}\cdot\mathbf{w}^{\mathrm{T}}\mathbf{V}\mathbf{w} \geq t\right\} \\ \leq\ & \mathbb{P}_1 + \mathbb{P}_2 + \mathbb{P}_3, \end{aligned} \tag{2.6.125}$$

where for any $\alpha \in \left(0, \sqrt{n} - \widetilde{C}\sqrt{d}\right)$ and any $\phi > 0$, we define

$$\begin{aligned}
\mathcal{E}_1 &= \left\{\lambda_{\max}\left[(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\right] > \left(\sqrt{n} - \widetilde{C}\sqrt{d} - \alpha\right)^{-2}\right\}, \\
\mathcal{E}_2 &= \left\{\mathbf{v}^{\mathrm{T}}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\mathbf{v} > d + \max\left(\phi^2, \phi\right) d\right\},
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{P}_1 &= \mathbb{P}\{\mathcal{E}_1\} \\
&\leq \mathbb{P}\left\{\lambda_{\max}\left[(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\right] > 1/\left(\sqrt{n} - \widetilde{C}\sqrt{d} - \alpha\right)^2\right\} \\
&\overset{(2.6.16)}{\leq} 2\exp\left(-\widetilde{c}_1\alpha^2\right),
\end{aligned} \tag{2.6.126}$$

$$\mathbb{P}_2 = \mathbb{P}\{\mathcal{E}_2\} \leq \mathbb{P}\left\{\mathbf{v}^{\mathrm{T}}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\mathbf{v} > d + \max\left(\phi^2, \phi\right) d\right\} \overset{(2.6.20)}{\leq} \exp\left(-\widetilde{c}_2\phi^2 d\right), \tag{2.6.127}$$

$$\begin{aligned}
\mathbb{P}_3 &= \mathbb{P}\left(\left\{\frac{\sigma^2}{\lambda_{\min}(\mathbf{\Sigma})}\lambda_{\max}^2\left[(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\right]\mathbf{v}^{\mathrm{T}}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\mathbf{v}\cdot\mathbf{w}^{\mathrm{T}}\mathbf{V}\mathbf{w} \geq t\right\} \cap \mathcal{E}_1^c \cap \mathcal{E}_2^c\right) \\
&\leq \mathbb{P}\left\{\frac{\sigma^2}{\lambda_{\min}(\mathbf{\Sigma})}\frac{\left[1 + \max\left(\phi^2, \phi\right)\right] d}{\left(\sqrt{n} - \widetilde{C}\sqrt{d} - \alpha\right)^4}\cdot\mathbf{w}^{\mathrm{T}}\mathbf{V}\mathbf{w} \geq t\right\}.
\end{aligned}$$

By choosing

$$t = \frac{\sigma^2}{\lambda_{\min}(\mathbf{\Sigma})}\frac{\left[1 + \max\left(\phi^2, \phi\right)\right] d}{\left(\sqrt{n} - \widetilde{C}\sqrt{d} - \alpha\right)^4}\cdot\left[1 + \max\left(r^2, r\right)\right] n,$$

for any $r > 0$, we have

$$\mathbb{P}_3 \leq \mathbb{P}\left\{\mathbf{w}^{\mathrm{T}}\mathbf{V}\mathbf{w} - n \geq \max\left(r^2, r\right) n\right\} \overset{(i)}{\leq} 2\exp\left[-\widetilde{c}_3 r^2 \frac{n}{\lambda_{\max}(\mathbf{V})}\right], \tag{2.6.128}$$

where we applied Hanson-Wright inequality and $\mathbb{E}\left(\mathbf{w}^{\mathrm{T}}\mathbf{V}\mathbf{w}\right) = \mathrm{tr}(\mathbf{V}) = n$ in the second inequality.

Combining (2.6.125), (2.6.126), (2.6.127), we find that for any $\alpha \in (0, \sqrt{n} - \widetilde{C}\sqrt{d})$ and any $\phi, r > 0$,

$$\begin{aligned} &\mathbb{P}\left\{\|\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta}\|_2^2 \geq \frac{\sigma^2}{\lambda_{\min}(\boldsymbol{\Sigma})} \frac{nd\left[1 + \max\left(\phi^2, \phi\right)\right]}{\left(\sqrt{n} - \widetilde{C}\sqrt{d} - \alpha\right)^4} \cdot \left[1 + \max\left(r^2, r\right)\right]\right\} \\ \leq\ & 2\exp\left(-\widetilde{c}_1\alpha^2\right) + \exp\left(-\widetilde{c}_2\phi^2 d\right) + 2\exp\left[-\widetilde{c}_3 r^2 \frac{n}{\lambda_{\max}(\mathbf{V})}\right]. \end{aligned} \tag{2.6.129}$$

Let $2\exp\left(-\widetilde{c}_1\alpha^2\right) = \exp\left(-\widetilde{c}_2\phi^2 d\right) = 2\exp\left[-\widetilde{c}_3 r^2 \frac{n}{\lambda_{\max}(\mathbf{V})}\right] = \eta/3$ and we obtain that $\alpha = \sqrt{\widetilde{c}_1^{-1}\log(6/\eta)}$, $\phi = \sqrt{\widetilde{c}_2^{-1} d^{-1} \log\left(\frac{3}{\eta}\right)}$, $r = \sqrt{\widetilde{c}_3^{-1} n^{-1} \lambda_{\max}(\mathbf{V}) \log(6/\eta)}$. To simplify (2.6.129), we enlarge the lower bound of $\|\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta}\|_2^2$ to obtain a sufficient condition. Note that $\eta > 6\exp[-\frac{\widetilde{c}_1}{4}(\sqrt{n} - 2\widetilde{C}\sqrt{d})^2]$ implies $\alpha = \sqrt{\widetilde{c}_1^{-1}\log(6/\eta)} < \sqrt{n}/2 - \widetilde{C}\sqrt{d}$, which satisfies that range of α to be less than $\sqrt{n} - \widetilde{C}\sqrt{d}$ and also implies $n(\sqrt{n} - \widetilde{C}\sqrt{d} - \alpha)^{-4} \leq 16(\sqrt{n} - \widetilde{C}\sqrt{d})^{-2}$. To make $\alpha > 0$, we also need $\sqrt{n} > 2\widetilde{C}\sqrt{d}$. Without loss of generality, assume $\widetilde{c}_2 < 1$ so that $1 + \max(\phi^2, \phi) \leq 1 + \max(d\phi^2, \sqrt{d}\phi) = 1 + d\phi^2 \leq 2d\phi^2 = \widetilde{c}_2^{-1} 2\log(3/\eta) \leq \widetilde{c}_2^{-1} 4\log(1/\eta)$, where the last step is obtained by assuming $\eta \leq 1/3$. Note that $\eta \leq 1/3$ also implies $\sqrt{\widetilde{c}_3^{-1} n^{-1}\lambda_{\max}(\mathbf{V})\log(1/\eta)} \leq r \leq \sqrt{4\widetilde{c}_3^{-1} n^{-1}\lambda_{\max}(\mathbf{V})\log(1/\eta)}$. Plugging all these inflated terms into (2.6.129) and letting $c_1 = 6, c_2 = \widetilde{c}_1/4$, we have for $c_1\exp[-c_2(\sqrt{n} - 2\widetilde{C}\sqrt{d})^2] < \eta \leq 1/3$

$$\mathbb{P}\left\{\|\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta}\|_2^2 \geq \frac{64}{\widetilde{c}_2}\log\left(\frac{1}{\eta}\right)\tau_1(n, d, \sigma^2, \boldsymbol{\Sigma})\left[1 + \tau_3(n, \mathbf{V}, \eta)\right]\right\} \leq \eta$$

where $\tau_3(n, \mathbf{V}, \eta) = \max(r, r^2) \asymp \max[\sqrt{n^{-1}\lambda_{\max}(\mathbf{V})\log(1/\eta)}, n^{-1}\lambda_{\max}(\mathbf{V})\log(1/\eta)]$, and $\tau_1(n, d, \sigma^2, \boldsymbol{\Sigma}) = \lambda_{\min}^{-1}(\boldsymbol{\Sigma})(\sqrt{n} - C\sqrt{d})^{-2}\sigma^2 d$ with $C = 2\widetilde{C}$.

Now we move to Part (ii). Using similar argument with symmetric modifications, one can prove Part(ii) of Theorem 2.6.1. Observe that

$$\begin{aligned} &\|\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta}\|_2^2 \\ \geq\ & \lambda_{\min}\left(\boldsymbol{\Sigma}^{-1}\right)\lambda_{\min}^2\left[(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\right]\sigma^2\mathbf{w}^{\mathrm{T}}\mathbf{V}^{1/2}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\mathbf{V}^{1/2}\mathbf{w} \end{aligned}$$

$$= \quad \frac{\sigma^2}{\lambda_{\max}(\mathbf{\Sigma})}\lambda_{\min}^2\left[(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\right]\mathbf{v}^{\mathrm{T}}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\mathbf{v}\cdot\mathbf{w}^{\mathrm{T}}\mathbf{V}\mathbf{w}$$

With a slight abuse of notation, we define

$$\begin{aligned}
\mathcal{E}_1 &= \left\{\lambda_{\min}\left[(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\right] < \left(\sqrt{n}+\widetilde{C}\sqrt{d}+\alpha\right)^{-2}\right\}, \\
\mathcal{E}_2 &= \left\{\mathbf{v}^{\mathrm{T}}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\mathbf{v} < d - \max\left(\phi^2,\phi\right)d\right\},
\end{aligned}$$

and follow similar arguments in (2.6.126) and (2.6.127). We choose

$$t = \frac{\sigma^2 n}{\lambda_{\max}(\mathbf{\Sigma})}\frac{\left[1-\max\left(\phi^2,\phi\right)\right]d}{\left(\sqrt{n}+\widetilde{C}\sqrt{d}+\alpha\right)^4}\cdot\left[1-\max\left(r^2,r\right)\right].$$

Notice that α can be any positive number but we need $\phi, r \in (0,1)$ to maintain positiveness of t.

Similar to (2.6.125) and (2.6.129), we find that for any $\alpha > 0$ and any $\phi, r \in (0,1)$:

$$\begin{aligned}
&\mathbb{P}\left\{\|\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}-\boldsymbol{\beta}\|_2^2 \le \frac{\sigma^2}{\lambda_{\max}(\mathbf{\Sigma})}\frac{nd\left[1-\max\left(\phi^2,\phi\right)\right]}{\left(\sqrt{n}+\widetilde{C}\sqrt{d}+\alpha\right)^4}\cdot\left[1-\max\left(r^2,r\right)\right]\right\} \\
\le\ & 2\exp\left(-\widetilde{c}_1\alpha^2\right)+\exp\left(-\widetilde{c}_2\phi^2 d\right)+2\exp\left\{-\widetilde{c}_3 r^2\frac{n}{\lambda_{\max}(\mathbf{V})}\right\}. \qquad (2.6.130)
\end{aligned}$$

We still assign $2\exp(-\widetilde{c}_1\alpha^2) = \exp(-\widetilde{c}_2\phi^2 d) = 2\exp\left[-\widetilde{c}_3 r^2\frac{n}{\lambda_{\max}(\mathbf{V})}\right] = \eta/3$ so that $\alpha = \sqrt{\widetilde{c}_1^{-1}\log(6/\eta)}$, $\phi = \sqrt{\widetilde{c}_2^{-1}d^{-1}\log\left(\frac{3}{\eta}\right)}$, $r = \sqrt{\widetilde{c}_3^{-1}n^{-1}\lambda_{\max}(\mathbf{V})\log(6/\eta)}$. To simplify (2.6.130), we reduce the upper bound of $\|\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}-\boldsymbol{\beta}\|_2^2$ to obtain a sufficient condition. By letting $\eta > 6\exp[-\widetilde{c}_3 n/\lambda_{\max}(\mathbf{V})]$, we have $r < 1$ as desired. By letting $\eta > 6\exp[-0.25\min(\widetilde{c}_1,\widetilde{c}_2)d]$ and $\sqrt{n} > (9\widetilde{C}+5)\sqrt{d}$, we have $\phi < 1$, $1-\max(\phi^2,\phi) \ge 0.5$, and $n(\sqrt{n}+\widetilde{C}\sqrt{d}+\alpha)^{-4} \ge 0.9^4(\sqrt{n}+\widetilde{C}\sqrt{d})^{-2}$. Bounding $\eta \le 3/4$, then we have $\sqrt{\widetilde{c}_3^{-1}n^{-1}\lambda_{\max}(\mathbf{V})\log(1/\eta)} \le r \le \sqrt{\widetilde{c}_3^{-1}n^{-1}\lambda_{\max}(\mathbf{V})C'\log(1/\eta)}$ where C' is a constant such that $\log(6/\eta) \le C'\log(1/\eta)$. Collecting them together, if $c_1\exp\{-c_2\min[n/\lambda_{\max}(\mathbf{V}), d]\} \le \eta \le 3/4$ and $\sqrt{n} > C\sqrt{d}$ where

$c_1 = 6, c_2 = \min\{0.25\widetilde{c}_1, 0.25\widetilde{c}_2, \widetilde{c}_3\}$ and $C = (9\widetilde{C} + 5)$, we have the desired result

$$\mathbb{P}\left\{\|\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta}\|_2^2 \leq 0.9^4 0.5\tau_2(n, d, \sigma^2, \boldsymbol{\Sigma})\left[1 - \tau_3(n, \mathbf{V}, \eta)\right]\right\} \leq \eta$$

where $\tau_3(n, \mathbf{V}, \eta) \asymp \max[\sqrt{n^{-1}\log(1/\eta)\lambda_{\max}(\mathbf{V})}, n^{-1}\log(1/\eta)\lambda_{\max}(\mathbf{V})]$. □

2.6.7.2 Proof of Theorem 2.6.2

Recall $\mathbf{v} = \mathbf{V}^{1/2}\mathbf{w}/\|\mathbf{V}^{1/2}\mathbf{w}\|_2$ in Lemma 2.2.1. Observe that

$$\begin{aligned}
L_j^2 &= S^2\mathbf{e}_j^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{e}_j \\
&= \frac{\boldsymbol{\varepsilon}^{\mathrm{T}}\left(\mathbf{I} - \mathbf{P}_{\mathbf{X}}\right)\boldsymbol{\varepsilon}}{n-d}\mathbf{e}_j^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{e}_j \\
&= \sigma^2\mathbf{e}_j^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\boldsymbol{\Sigma}^{-1/2}\mathbf{e}_j\frac{\mathbf{w}^{\mathrm{T}}\mathbf{V}^{1/2}\left(\mathbf{I} - \mathbf{P}_{\mathbf{Z}}\right)\mathbf{V}^{1/2}\mathbf{w}}{n-d} \\
&= \sigma^2\frac{\mathbf{e}_j^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\boldsymbol{\Sigma}^{-1/2}\mathbf{e}_j}{n-d}\cdot\mathbf{v}^{\mathrm{T}}\left(\mathbf{I} - \mathbf{P}_{\mathbf{Z}}\right)\mathbf{v}\cdot\mathbf{w}^{\mathrm{T}}\mathbf{V}\mathbf{w}. \qquad (2.6.131)
\end{aligned}$$

We prove part (ii) of Theorem 2.6.2 first and part (i) follows similar arguments.

From (2.6.131), we have

$$\sigma^2\frac{\mathbf{e}_j^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{e}_j}{n-d}\lambda_{\min}\left[(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\right]\left\{1 - \lambda_{\max}\left[(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\right]\mathbf{v}^{\mathrm{T}}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\mathbf{v}\right\}\mathbf{w}^{\mathrm{T}}\mathbf{V}\mathbf{w} \leq L_j^2.$$

Hence for any $t > 0$,

$$\mathbb{P}\left\{L_j^2 \leq t\right\} \leq \mathbb{P}\left\{\mathcal{E}_0\right\} \leq \mathbb{P}_1 + \mathbb{P}_2 + \mathbb{P}_3,$$

where with a slight abuse of notation, we define

$$\mathcal{E}_0 = \left\{\sigma^2\frac{\mathbf{e}_j^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{e}_j}{n-d}\lambda_{\min}\left[(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\right]\left\{1 - \lambda_{\max}\left[(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\right]\mathbf{v}^{\mathrm{T}}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\mathbf{v}\right\}\mathbf{w}^{\mathrm{T}}\mathbf{V}\mathbf{w} \leq t\right\}$$

and for $\alpha \in \left(0, \sqrt{n} - C\sqrt{d}\right)$, we define

$$\mathcal{E}_1 = \left\{\lambda_{\min}\left[(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\right] < \left(\sqrt{n} + \widetilde{C}\sqrt{d} + \alpha\right)^{-2} \text{ or } \lambda_{\max}\left[(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\right] > \left(\sqrt{n} - \widetilde{C}\sqrt{d} - \alpha\right)^{-2}\right\},$$

$$\mathcal{E}_2 = \left\{\mathbf{v}^{\mathrm{T}}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\mathbf{v} > d + \max\left(\phi^2, \phi\right) n\right\},$$

and

$$\mathbb{P}_1 = \mathbb{P}\left\{\mathcal{E}_1\right\} \overset{(2.6.16),(2.6.17)}{\leq} 4\exp\left(-\widetilde{c}_1\alpha^2\right), \tag{2.6.132}$$

$$\begin{aligned}
\mathbb{P}_2 = \mathbb{P}\left\{\mathcal{E}_2\right\} &\overset{(2.6.20)}{\leq} \exp\left\{-\widetilde{c}_2 \min\left[\max^2\left(\phi^2, \phi\right) n^2/d, \max\left(\phi^2, \phi\right) n\right]\right\} \\
&\leq \exp\left\{-\widetilde{c}_2 \min\left[\max^2\left(\phi^2, \phi\right), \max\left(\phi^2, \phi\right)\right] \min\left(n^2/d, n\right)\right\} \\
&\overset{(i)}{=} \exp\left(-\widetilde{c}_2\phi^2 n\right),
\end{aligned} \tag{2.6.133}$$

$$\begin{aligned}
\mathbb{P}_3 &= \mathbb{P}\left\{\mathcal{E}_0 \cap \mathcal{E}_1^c \cap \mathcal{E}_2^c\right\} \\
&\overset{(ii)}{\leq} \mathbb{P}\left\{\sigma^2 \frac{\mathbf{e}_j^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{e}_j}{n-d} \cdot \left[1 - \frac{d + \max\left(\phi^2, \phi\right) n}{\left(\sqrt{n} - \widetilde{C}\sqrt{d} - \alpha\right)^2}\right] \frac{\mathbf{w}^{\mathrm{T}}\mathbf{V}\mathbf{w}}{\left(\sqrt{n} + \widetilde{C}\sqrt{d} + \alpha\right)^2} \leq t\right\}.
\end{aligned}$$

Step (i) in (2.6.133) follows from $\min\left[\max^2\left(\phi^2, \phi\right), \max\left(\phi^2, \phi\right)\right] = \phi^2$ and $n/d \geq 1$ and we need $0 < \phi < n^{-1}[(\sqrt{n} - \widetilde{C}\sqrt{d} - \alpha)^2 - d]$ in the step (ii) to maintain positiveness.

By setting

$$t = \sigma^2 \frac{\mathbf{e}_j^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{e}_j}{n-d} \cdot \left[1 - \frac{d + \max\left(\phi^2, \phi\right) n}{\left(\sqrt{n} - \widetilde{C}\sqrt{d} - \alpha\right)^2}\right] \frac{\left[1 - \max\left(r^2, r\right)\right] n}{\left(\sqrt{n} + \widetilde{C}\sqrt{d} + \alpha\right)^2},$$

with $0 < r < 1$, we have

$$\mathbb{P}_3 \leq \mathbb{P}\left\{\mathbf{w}^{\mathrm{T}}\mathbf{V}\mathbf{w} - n \leq -n\max(r, r^2)\right\} \leq 2\exp\left[-\widetilde{c}_3 r^2 \frac{n}{\lambda_{\max}(\mathbf{V})}\right], \tag{2.6.134}$$

where the second inequity follows from Hanson-Wright inequality.

Combining (2.6.132), (2.6.133), and (2.6.134), we find that for any $\alpha \in (0, \sqrt{n} - \widetilde{C}\sqrt{d})$, $r \in (0,1)$, $\phi \in (0, n^{-1}[(\sqrt{n} - \widetilde{C}\sqrt{d} - \alpha)^2 - d])$,

$$\begin{aligned} &\mathbb{P}\left\{L_j^2 \leq \sigma^2 \frac{\mathbf{e}_j^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{e}_j}{n-d} \cdot \left[1 - \frac{d + \max(\phi^2, \phi)\, n}{\left(\sqrt{n} - \widetilde{C}\sqrt{d} - \alpha\right)^2}\right] \frac{[1 - \max(r^2, r)]\, n}{\left(\sqrt{n} + \widetilde{C}\sqrt{d} + \alpha\right)^2}\right\} \\ \leq\ & 4\exp\left(-\widetilde{c}_1\alpha^2\right) + \exp\left(-\widetilde{c}_2\phi^2 n\right) + 2\exp\left[-\widetilde{c}_3 r^2 \frac{n}{\lambda_{\max}(\mathbf{V})}\right]. \end{aligned} \tag{2.6.135}$$

By setting $4\exp(-\widetilde{c}_1\alpha^2) = \exp(-\widetilde{c}_2\phi^2 n) = 2\exp[-\widetilde{c}_3 r^2 n\lambda_{\max}^{-1}(\mathbf{V})] = \eta/3$, we have $\alpha = \sqrt{\widetilde{c}_1^{-1}\log(12/\eta)}$, $\phi = \sqrt{\widetilde{c}_2^{-1} n^{-1}\log(3/\eta)}$, $r = \sqrt{\widetilde{c}_3^{-1} n^{-1}\lambda_{\max}(\mathbf{V})\log(6/\eta)}$. We also have $\eta > 6\exp[-\widetilde{c}_3 n\lambda_{\max}^{-1}(\mathbf{V})]$, which implies $r < 1$. If we additionally assume $\sqrt{n} \geq \max(\widetilde{C}/0.9, \sqrt{\widetilde{c}_2})d$ and $\eta \geq 12\exp\{-\frac{9}{169}\min(\widetilde{c}_1, \widetilde{c}_2)[\sqrt{n} - \frac{13}{3}\widetilde{C}\sqrt{d}]^2\}$, we have $\alpha < \sqrt{n} - \widetilde{C}\sqrt{d}$, $\phi \leq n^{-1}[0.1(\sqrt{n} - \widetilde{C}\sqrt{d} - \alpha)^2 - d]$ and $0.4\tau_4(n, d, \sigma^2, \mathbf{\Sigma}) \leq \tau_4(n, d, \sigma^2, \mathbf{\Sigma})\{1 - [d + \max(\phi, \phi^2)n](\sqrt{n} - \widetilde{C}\sqrt{d} - \alpha)^{-2}\}[1 - \max(r, r^2)]n(\sqrt{n} + \widetilde{C}\sqrt{d} + \alpha)^{-2}$. If $\eta \leq 3/4$, we have $r \asymp \sqrt{n^{-1}\lambda_{\max}(\mathbf{V})\log(1/\eta)}$. Define $c_1 = 12, c_2 = \min[9\widetilde{c}_1/169, 9\widetilde{c}_2/169, \widetilde{c}_3]$ and $C = \max[13\widetilde{C}/3, \widetilde{C}/0.9, \sqrt{\widetilde{c}_2}]$. If $\sqrt{n} \geq Cd$ and $c_1\exp\{-c_2\min[n\lambda_{\max}^{-1}(\mathbf{V}), (\sqrt{n} - C\sqrt{d})^2]\} \leq \eta \leq 3/4$, we have

$$\mathbb{P}\left\{L_j^2 \leq 0.4\tau_4(n, d, \sigma^2, \mathbf{\Sigma})\left(1 - \tau_3(n, \mathbf{V}, \eta)\right)\right\} \leq \eta.$$

and $\tau_3(n, \mathbf{V}, \eta) \asymp \max[\sqrt{n^{-1}\lambda_{\max}(\mathbf{V})\log(1/\eta)}, n^{-1}\lambda_{\max}(\mathbf{V})\log(1/\eta)]$.

Now we move to the proof of Part (i) of Theorme 2.6.2.

Define event

$$\mathcal{E}_1 = \left\{\lambda_{\max}\left[(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\right] > \left(\sqrt{n} - \widetilde{C}\sqrt{d} - \alpha\right)^{-2}\right\}$$

where $0 < \alpha < \sqrt{n} - \widetilde{C}\sqrt{d}$ with probability

$$\mathbb{P}\left\{\mathcal{E}_1\right\} \overset{(2.6.16)}{\leq} 2\exp\left(-\widetilde{c}_1\alpha^2\right). \tag{2.6.136}$$

For any $t > 0$, we have

$$\begin{aligned}
\mathbb{P}\left\{L_j^2 \geq t\right\} &\leq \mathbb{P}\left\{\mathcal{E}_1\right\} + \mathbb{P}\left\{\sigma^2 \frac{\mathbf{e}_j^{\mathrm{T}}\mathbf{\Sigma}^{-1/2}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{\Sigma}^{-1/2}\mathbf{e}_j}{(n-d)} \cdot \mathbf{w}^{\mathrm{T}}\mathbf{V}\mathbf{w} \geq t \cap \mathcal{E}_1^c\right\} \\
&\leq \mathbb{P}\left\{\mathcal{E}_1\right\} + \mathbb{P}\left\{\sigma^2 \frac{\mathbf{e}_j^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{e}_j}{(n-d)} \cdot \frac{\mathbf{w}^{\mathrm{T}}\mathbf{V}\mathbf{w}}{\left(\sqrt{n} - \widetilde{C}\sqrt{d} - \alpha\right)^2} \geq t\right\}.
\end{aligned}$$

Set

$$t = \frac{\sigma^2\mathbf{e}_j^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{e}_j}{(n-d)} \cdot \frac{n\left[1 + \max\left(r^2, r\right)\right]}{\left(\sqrt{n} - \widetilde{C}\sqrt{d} - \alpha\right)^2}.$$

Then for any $r > 0$, we have

$$\begin{aligned}
&\mathbb{P}\left\{\sigma^2 \frac{\mathbf{e}_j^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{e}_j}{(n-d)} \frac{\mathbf{w}^{\mathrm{T}}\mathbf{V}\mathbf{w}}{\left(\sqrt{n} - \widetilde{C}\sqrt{d} - \alpha\right)^2} \geq t\right\} \\
\leq\ & \mathbb{P}\left\{\mathbf{w}^{\mathrm{T}}\mathbf{V}\mathbf{w} - n \geq n\max\left(r^2, r\right)\right\} \\
\leq\ & 2\exp\left[-\widetilde{c}_3 r^2 \frac{n}{\lambda_{\max}(\mathbf{V})}\right],
\end{aligned} \tag{2.6.137}$$

where the last inequality follows from Hanson-Wright inequality.

Combining (2.6.136) and (2.6.137), we find that for any $\alpha \in (0, \sqrt{n} - \widetilde{C}\sqrt{d})$ and $r > 0$,

$$\begin{aligned}
&\mathbb{P}\left\{L_j^2 \geq \frac{\sigma^2\mathbf{e}_j^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{e}_j}{(n-d)} \frac{n\left[1 + \max\left(r^2, r\right)\right]}{\left(\sqrt{n} - \widetilde{C}\sqrt{d} - \alpha\right)^2}\right\} \\
\leq\ & 2\exp\left(-\widetilde{c}_1\alpha^2\right) + 2\exp\left[-\widetilde{c}_3 r^2 \frac{n}{\lambda_{\max}(\mathbf{V})}\right].
\end{aligned} \tag{2.6.138}$$

Let $2\exp(-\widetilde{c}_1\alpha^2) = 2\exp[-\widetilde{c}_3 r^2 n\lambda_{\min}^{-1}(\mathbf{V})] = \eta/2$ and we obtain that $\alpha = \sqrt{\widetilde{c}_1^{-1}\log(4/\eta)}$ and $r = \sqrt{n^{-1}\widetilde{c}_3^{-1}\lambda_{\max}(\mathbf{V})\log(4/\eta)}$. If $\eta \geq 4\exp[-\frac{\widetilde{c}_1}{4}(\sqrt{n}-2\widetilde{C}\sqrt{d})^2]$, we have $\alpha < \sqrt{n}-\widetilde{C}\sqrt{d}$ and $n(\sqrt{n}-\widetilde{C}\sqrt{d}-\alpha)^{-2} \leq 4$. We also need $\sqrt{n} > \widetilde{C}\sqrt{d}$ to maintain the positiveness of α and we let $\eta \leq 3/4$ so that $r \asymp \sqrt{n^{-1}\lambda_{\max}(\mathbf{V})\log(1/\eta)}$. Hence if $c_1\exp[-c_2(\sqrt{n}-C\sqrt{d})^2] \leq \eta \leq 3/4$ and $\sqrt{n} > C\sqrt{d}$, where $c_1 = 4, c_2 = \widetilde{c}/4$ and $C = 2\widetilde{C}$, we have

$$\mathbb{P}\left\{L_j^2 \geq 4\tau_4(n,d,\sigma^2,\mathbf{\Sigma})\left(1+\tau_3(n,\mathbf{V},\eta)\right)\right\} \leq \eta,$$

with $\tau_3(n,\mathbf{V},\eta) \asymp \max[\sqrt{n^{-1}\lambda_{\max}(\mathbf{V})\log(1/\eta)}, n^{-1}\lambda_{\max}(\mathbf{V})\log(1/\eta)]$. $\square$

Chapter 3

F Distribution in Linear Regression: A New Lens Through One Century Journey

3.1 Background

3.1.1 Review on F distribution

George Snedecor once said that Student's t distribution is the "*distribution that revolutionized the statistics of small samples*" (p.55 in Snedecor and Cochran (1980)). This also leads to the standard belief in the derivation of F distribution using the ratio of independent χ^2 random variables. As ubiquitously appeared in textbooks on probability theory and statistics (Johnson, Kotz, & Balakrishnan, 1995; Casella & Berger, 2024), the F distribution $F(v_1, v_2)$ is defined by

$$F = \frac{\chi^2(v_1)/v_1}{\chi^2(v_2)/v_2},$$

where $\chi^2(v_1)$ and $\chi^2(v_2)$ are independent χ^2 distributions with degrees of freedom v_1 and v_2, respectively. Derived from the ratio of random variables, the density of $F(v_1, v_2)$ is

$$f_F(x) = \frac{v_1^{v_1/2} v_2^{v_2/2}}{B(\frac{v_1}{2}, \frac{v_2}{2})} \frac{x^{v_1/2-1}}{(v_1 x + v_2)^{(v_1+v_2)/2}}, \quad x \in [0, \infty),$$

where $B(\cdot, \cdot)$ is the beta function.

The χ^2 ratio definition of the F distribution arose from comparing two population variances, σ_1^2 and σ_2^2, using the sample variances, s_1^2 and s_2^2. In the epoch-making paper presented at the International Congress of Mathematicians in Toronto (Fisher, 1924), Sir Ronald A. Fisher first explained the unsatisfactory performance of $s_1^2 - s_2^2$ for small samples. Then he used the notion of

$$e^{2z} = \frac{s_1^2}{s_2^2} = \frac{\sigma_1^2\chi^2(n_1-1)/(n_1-1)}{\sigma_2^2\chi^2(n_2-1)/(n_2-1)}$$

to define the z distribution, which is independent of σ_1^2 and σ_2^2 under $H_0 : \sigma_1^2 = \sigma_2^2$.

Later in 1932, P.C. Mahalanobis first tabulated the values of $e^{2z} = s_1^2/s_2^2$ in Mahalanobis (1932), which is the well-known F statistic in current practice. The notation F was chosen to represent the variance ratio s_1^2/s_2^2 by Snedecor (1934) in honor of Fisher's contribution. For more anecdotes about F distribution, refer to Hald (2008); Gorroochurn (2016).

This line of work established the foundation of model comparison for small samples. From then on, the F-test has become one of the most-used *exact tests* with diverse applications.

In Fisher's first book, *Statistical Methods for Research Workers* (Fisher, 1928), Chaps. VII and VIII introduced the analysis of variance (ANOVA) for the inference of intraclass correlation and regression problems, which reflected the ubiquity of the F-test. The mathematical foundation of the ANOVA was elaborated in Irwin (1931, 1934); Cochran (1934). The ANOVA became the basic tool for experimental design after the publication of Fisher's book *The Design of Experiments* (Fisher, 1935), such as factorial designs, random effects models, and Latin square designs. See Scheffé (1959); Wu and Hamada (2011) for detailed usage.

The ANOVA is also widely used in linear regression since the experimental data can be modeled by the regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. The F distribution in Proposition 3.2.2 plays a fundamental role in drawing simultaneous inference on a number of parametric functions, such as the F-test of $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{c}$ and the confidence region of $\mathbf{C}\boldsymbol{\beta}$ (p.172 of R. Anderson and Bancroft (1952), Scheffé (1959), Chap.4 of Rao, Rao, Statistiker, Rao, and Rao (1973), and Chap.4 of Seber and Lee (2003)). The F-test and ANOVA can also be applied to the multiple comparison for $\boldsymbol{\beta}$ (Scheffé, 1953; Olshen, 1973), the confidence band for $\mathbf{X}\boldsymbol{\beta}$ (Lieberman, Miller Jr, & Hamilton, 1967; Bohrer, 1973; Lane & DuMouchel, 1994), the simultaneous prediction band for future values of Y (Carlstein, 1986; Limam & Thomas, 1988), and stepwise model selection (Efroymson (1960); Wilkinson (1970); N. Draper, Guttman, and Kanemasu (1971); Pope and Webster (1972), Chap.15 of N. R. Draper and Smith (1998)).

The ANOVA F-test is based on three key assumptions: independence, homogeneity of variance, and normality. The consequences of dependent individuals on Type I and Type II errors were discussed in Glass, Peckham, and Sanders (1972); Wickens and Keppel (2004). The effects of heteroscedasticity were discussed in Box (1954a, 1954b); Brown and Forsythe (1974) with corresponding procedures such as Welch (1951); Brunner, Dette, and Munk (1997). The results of the violation of normality were discussed in David and Johnson (1951); Box (1953); Box and Watson (1962); Atiqullah (1964).

The F-test can be applied to many other fields, such as the functional F-test (Chap .9 in Ramsay and Silverman (1997); Shen and Faraway (2004)), the bias analysis of preliminary tests (Bancroft, 1944), the tests of goodness of fit for exponentially distributed lifetimes (Epstein, 1960a, 1960b), the test of uniformity based on order statistics (p.335 in D'Agostino (2017)), the trend analysis in time series (Chaps. 3 and 4 in T. W. Anderson (2011)), etc.

3.1.2 Symmetric Distribution

As pointed out on p.163 of Stigler (2016), "*Fisher recognized that his tests—the various F-tests—required only spherical symmetry under the null hypothesis to work. Normality and independence imply spherical symmetry, but those are not necessary conditions.*" This observation highlights the connection between the F distribution and the spherically symmetric distribution. To better illustrate this connection, this section discusses some important properties of the spherically symmetric distribution. Readers who wish to proceed more quickly may skip to Section 3.1.3.

Definition 3.1.1 (Definition 2.1 of Fang (2018)). *A vector $\mathbf{w} \in \mathbb{R}^n$ is said to have a spherically symmetric distribution (or simply spherical distribution) if and only if for every orthogonal matrix $\mathbf{\Phi} \in \mathbb{R}^{n\times n}$, we have $\mathbf{w} \overset{d}{=} \mathbf{\Phi}\mathbf{w}$, i.e. , $\mathbf{w}$ and $\mathbf{\Phi}\mathbf{w}$ have the same distribution.*

The multivariate normal distribution with zero mean and identity covariance, *i.e.* , $\mathcal{N}(\mathbf{0}_n, \mathbf{I}_n)$, and the uniform distribution on the unit sphere $\text{Unif}(S^{n-1})$, where $S^{n-1} = \{\mathbf{w} \in \mathbb{R}^n : \|\mathbf{w}\|_2 = 1\}$, are both examples of the spherical distributions.

Define the characteristic function of the random vector $\mathbf{w} \in \mathbb{R}^n$ to be $\psi_{\mathbf{w}}(\mathbf{t}) = \mathbb{E}[\exp(\mathbf{t}^{\mathrm{T}}\mathbf{w})]$ for $\mathbf{t} \in \mathbb{R}^n$. The $\psi_{\mathbf{w}}(\mathbf{t})$ of the spherical distribution has the following property:

Proposition 3.1.1 (Theorem 2.1 in Fang (2018))**.** *The following statements are equivalent:*

(i) $\mathbf{w} \in \mathbb{R}^n$ *is the spherical distribution defined in Definition 3.1.1;*

(ii) $\psi_{\mathbf{w}}(\mathbf{t}) = \phi(\mathbf{t}^{\mathrm{T}}\mathbf{t})$ *for some function* $\phi()$ *of a scalar variable.*

Then we have the following corollary:

Corollary 3.1.1. *If* $\mathbf{w} \in \mathbb{R}^n$ *follows the spherical distribution defined in Definition 3.1.1 and* $\mathbf{\Gamma} \in \mathbb{R}^{n \times p}$ *with* $1 \leq p \leq n$ *is a fixed matrix such that* $\mathbf{\Gamma}^{\mathrm{T}}\mathbf{\Gamma} = \mathbf{I}_p$*, then* $\mathbf{\Gamma}^{\mathrm{T}}\mathbf{w} \in \mathbb{R}^p$ *also follows the spherical distribution.*

Proof. For $\mathbf{t} \in \mathbb{R}^p$, we have

$$\psi_{\mathbf{\Gamma}^{\mathrm{T}}\mathbf{w}}(\mathbf{t}) = \mathbb{E}[\exp(\mathbf{t}^{\mathrm{T}}\mathbf{\Gamma}^{\mathrm{T}}\mathbf{w})] = \psi_{\mathbf{w}}(\mathbf{\Gamma}\mathbf{t}).$$

By Proposition 3.1.1, we have

$$\psi_{\mathbf{w}}(\mathbf{\Gamma}\mathbf{t}) = \phi(\mathbf{t}^{\mathrm{T}}\mathbf{\Gamma}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{t}) = \phi(\mathbf{t}^{\mathrm{T}}\mathbf{t}).$$

So we have

$$\psi_{\mathbf{\Gamma}^{\mathrm{T}}\mathbf{w}}(\mathbf{t}) = \phi(\mathbf{t}^{\mathrm{T}}\mathbf{t}).$$

By Proposition 3.1.1, $\mathbf{\Gamma}^{\mathrm{T}}\mathbf{w} \in \mathbb{R}^p$ follows the spherical distribution. $\square$

The normal distribution $\mathcal{N}(\mathbf{0}_n, \mathbf{I}_n)$ and the uniform distribution on the unit sphere $\mathrm{Unif}(S^{n-1})$ have the following relationship:

Proposition 3.1.2 (Theorem 1.5.6 of Muirhead (2009))**.** *If* $\mathbf{w} \sim \mathcal{N}(\mathbf{0}_n, \mathbf{I}_n)$*, then we have* $\|\mathbf{w}\|_2 \perp\!\!\!\perp \mathbf{w}/\|\mathbf{w}\|_2$*, i.e. ,* $\|\mathbf{w}\|_2$ *and* $\mathbf{w}/\|\mathbf{w}\|_2$ *are independent, where* $\|\mathbf{w}\|_2^2 \sim \chi^2(n)$ *and* $\mathbf{w}/\|\mathbf{w}\|_2 \sim \mathrm{Unif}(S^{n-1})$*.*

In fact, every spherical distribution has the similar property:

Proposition 3.1.3 (Corollary after Theorem 2.2 in Fang (2018))**.** *If the random vector* $\mathbf{w} \in \mathbb{R}^n$ *is the spherical distribution defined in Definition 3.1.1, then we have* $\mathbf{w} \overset{d}{=} r\boldsymbol{\theta}$*, where* $r \perp\!\!\!\perp \boldsymbol{\theta}$*,* $\boldsymbol{\theta} \sim \mathrm{Unif}(S^{n-1})$*, and* $r \in [0, \infty)$ *is a random variable.*

3.1.3 Organization of the Paper

The rest of the paper is organized as follows. Section 3.2 presents two standard results about F distribution to be reexamined. Specifically, Section 3.2.1 revisits the classic definition of the F distribution, which is the ratio of the quadratic forms under normal distribution. Section 3.2.2 revisits the classic F-test and t-test in the normal linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Our contributions are detailed in the following sections:

- In Section 3.3, the definition of F distribution in Section 3.2.1 will be renewed beyond normality. We clearly characterize the relationship between the F distribution and the spherical distribution. The juxtaposition of the classic definition in Section 3.2.1 and the new definition in Section 3.3 provides a geometric view of the F distribution.

- In Section 3.4, the finite sample F and t tests presented in Section 3.2.2 will be renewed to the scenarios where $\boldsymbol{\varepsilon}$ can exhibit nonzero mean, strong correlation, heteroscedasticity, and heavy tails, if $\mathbf{X}$ has spherical symmetry. The juxtaposition of the classic F and t tests in Section 3.2.2 and the new extensions in Section 3.4 reveals the symmetric roles of $\mathbf{X}$ and $\boldsymbol{\varepsilon}$ in the linear model.

Section 3.5 presents some simulations grounded in real-world contexts. Section 3.6 concludes the paper with a discussion.

3.2 Two Standard Results to Be Renewed

Before we introduce our new results in Section 3.3 and Section 3.4, we review two standard results of the F distribution.

3.2.1 The First Standard Result to Be Renewed

Conventionally, the F distribution can be generated by the ratio of the quadratic form of the normal distribution:

Proposition 3.2.1 (Classic F distribution)**.** *Suppose $\boldsymbol{\xi} \sim \mathcal{N}(0, \mathbf{I}_n)$. For any fixed matrix $\mathbf{Z} = [\mathbf{Z}_1\ \mathbf{Z}_2]$ with $\mathbf{Z}_1 \in \mathbb{R}^{n\times q}$ and $\mathbf{Z}_2 \in \mathbb{R}^{n\times(p-q)}$ such that $\operatorname{rank}(\mathbf{Z}) = p$ and $1 \le q \le p < n$, we have*

$$F = \frac{[\boldsymbol{\xi}^{\mathrm{T}}(\mathbf{P}_{\mathbf{Z}} - \mathbf{P}_{\mathbf{Z}_2})\boldsymbol{\xi}]\,/q}{[\boldsymbol{\xi}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P}_{\mathbf{Z}})\boldsymbol{\xi}]\,/(n-p)} \sim F(q, n-p), \tag{3.2.1}$$

where $\mathbf{P}_{\mathbf{Z}} = \mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}$ is the projection matrix of $\mathcal{C}(\mathbf{Z})$, the column space of $\mathbf{Z}$, and $\mathbf{P}_{\mathbf{Z}_2}$ is similarly defined. If $q = p$, then $\mathbf{P}_{\mathbf{Z}_2} = \mathbf{O}_{n\times n}$.

Proof. It is a standard result (Rao et al., 1973) but we still give a short proof for completeness. We can write $\mathbf{P}_{\mathbf{Z}} - \mathbf{P}_{\mathbf{Z}_2} = \boldsymbol{\Gamma}_1\boldsymbol{\Gamma}_1^{\mathrm{T}}$ and $\mathbf{I}_n - \mathbf{P}_{\mathbf{Z}} = \boldsymbol{\Gamma}_2\boldsymbol{\Gamma}_2^{\mathrm{T}}$, where $\boldsymbol{\Gamma}_1 \in \mathbb{R}^{n\times q}$ and $\boldsymbol{\Gamma}_2 \in \mathbb{R}^{n\times(n-p)}$ are orthonormal bases of $\mathcal{C}(\mathbf{Z}) \cap \mathcal{C}(\mathbf{Z}_2)^{\perp}$ and $\mathcal{C}(\mathbf{Z})^{\perp}$, respectively, such that $\boldsymbol{\Gamma}_1^{\mathrm{T}}\boldsymbol{\Gamma}_2 = \mathbf{O}$. Then we have $[\boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2]^{\mathrm{T}}\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}_{q+n-p}, \mathbf{I}_{q+n-p})$. By the definition of χ^2 distribution and F distribution, we have the result. □

To produce the χ^2 distributions in F, it seems that Proposition 3.2.1 relies on the normality of $\boldsymbol{\xi}$. However, as noticed in Chap.V of Fisher (1935), Kelker (1970); Stigler (2004), and p.163 in Stigler (2016), spherical distribution is enough:

Corollary 3.2.1. *Suppose* $\boldsymbol{\theta} \sim \mathrm{Unif}(S^{n-1})$ *and* $\mathbf{Z}$ *is under the same condition in Proposition 3.2.1. Then we have*

$$F = \frac{\left[\boldsymbol{\theta}^{\mathrm{T}}(\mathbf{P}_{\mathbf{Z}} - \mathbf{P}_{\mathbf{Z}_2})\boldsymbol{\theta}\right]/q}{\left[\boldsymbol{\theta}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P}_{\mathbf{Z}})\boldsymbol{\theta}\right]/(n-p)} \sim F(q, n-p). \tag{3.2.2}$$

Proof. Replace $\boldsymbol{\xi}$ in (3.2.1) by $\boldsymbol{\theta} = \boldsymbol{\xi}/\|\boldsymbol{\xi}\|_2$, we still have the same distribution. From Proposition 3.1.2, we have $\boldsymbol{\theta} \sim \mathrm{Unif}(S^{n-1})$. □

In (3.2.2), $\boldsymbol{\theta} \sim \mathrm{Unif}(S^{n-1})$ and $\mathbf{Z} \in \mathbb{R}^{n\times p}$ is an arbitrary fixed matrix with full column rank. We are curious about the result after switching the roles of $\boldsymbol{\theta}$ and $\mathbf{Z}$; that is, if we fix $\boldsymbol{\theta}$ at any arbitrary point on S^{n-1} while set $\mathbf{Z}$ to be random with spherical symmetry, do we still have exact $F(q, n-p)$? Theorem 3.3.1 in Section 3.3 answers this question.

3.2.2 The Second Standard Result to Be Renewed

The next standard result is from the F-test and t-test in the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{3.2.3}$$

where $\mathbf{y} \in \mathbb{R}^n$ is the response vector, $\mathbf{X} \in \mathbb{R}^{n\times p}$ is the design matrix, $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is the error vector, and $\boldsymbol{\beta} \in \mathbb{R}^p$ is the parameter vector to be estimated. If $\mathbf{X}$ has full column rank, we define the ordinary least squares (OLS) estimator as

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}. \tag{3.2.4}$$

Proposition 3.2.2 (Classic F-test and t-test in the normal linear model)**.** *Under model* (3.2.3)*, suppose that* $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2\mathbf{I}_n)$ *and* $\mathbf{X} \in \mathbb{R}^{n\times p}$ *is fixed with* $\mathrm{rank}(\mathbf{X}) = p < n$*. Let*

$\mathbf{C} \in \mathbb{R}^{q\times p}$ *be a fixed matrix such that* $\operatorname{rank}(\mathbf{C}) = q \leq p$*. Then we have*

$$\begin{aligned} F &= \frac{\{(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^{\mathrm{T}}\mathbf{C}^{\mathrm{T}}\left[\mathbf{C}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{C}^{\mathrm{T}}\right]^{-1}\mathbf{C}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\}/q}{\{\mathbf{y}^{\mathrm{T}}\left(\mathbf{I}_n-\mathbf{P}_{\mathbf{X}}\right)\mathbf{y}\}/(n-p)} \\ &\sim F(q, n-p), \end{aligned} \tag{3.2.5}$$

and for $j = 1, \ldots, p$*, we have*

$$T_j = \frac{\widehat{\beta}_j - \beta_j}{\text{S.E.}(\widehat{\beta}_j)} \sim t(n-p), \tag{3.2.6}$$

where

$$\text{S.E.}(\widehat{\beta}_j) = \sqrt{(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}_{[j,j]}}\sqrt{\frac{\mathbf{y}^{\mathrm{T}}(\mathbf{I}_n-\mathbf{P}_{\mathbf{X}})\mathbf{y}}{n-p}},$$

$(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}_{[j,j]}$ *is the* j*th diagonal entry of* $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$*,* $\widehat{\beta}_j$ *and* β_j *are the* j*th component of* $\widehat{\boldsymbol{\beta}}$ *and* $\boldsymbol{\beta}$*, respectively.*

Proof. It is a standard result (Rao et al., 1973) but we still give a short proof for completeness. Since $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{\varepsilon}$, with a slight abuse of notation, we have

$$(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^{\mathrm{T}}\mathbf{C}^{\mathrm{T}}\left[\mathbf{C}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{C}^{\mathrm{T}}\right]^{-1}\mathbf{C}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) = \boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\Gamma}_1\boldsymbol{\Gamma}_1^{\mathrm{T}}\boldsymbol{\varepsilon},$$

where $\boldsymbol{\Gamma}_1 = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{C}^{\mathrm{T}}[\mathbf{C}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{C}^{\mathrm{T}}]^{-1/2} \in \mathbb{R}^{n\times q}$ is an orthonormal basis of a subspace of $\mathcal{C}(\mathbf{X})$. Since $(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}})\mathbf{y} = (\mathbf{I}_n - \mathbf{P}_{\mathbf{X}})\boldsymbol{\varepsilon}$, we have

$$\mathbf{y}^{\mathrm{T}}(\mathbf{I}_n-\mathbf{P}_{\mathbf{X}})\mathbf{y} = \boldsymbol{\varepsilon}^{\mathrm{T}}(\mathbf{I}_n-\mathbf{P}_{\mathbf{X}})\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\Gamma}_2\boldsymbol{\Gamma}_2^{\mathrm{T}}\boldsymbol{\varepsilon},$$

where $\boldsymbol{\Gamma}_2 \in \mathbb{R}^{n\times(n-p)}$ is an orthonormal basis of $\mathcal{C}(\mathbf{X})^{\perp}$.

Then we have $[\boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2]^{\mathrm{T}}(\boldsymbol{\varepsilon}/\sigma) \sim \mathcal{N}(\mathbf{0}_{q+n-p}, \mathbf{I}_{q+n-p})$ and

$$F = \frac{(\frac{\boldsymbol{\varepsilon}}{\sigma})^{\mathrm{T}}\boldsymbol{\Gamma}_1\boldsymbol{\Gamma}_1^{\mathrm{T}}(\frac{\boldsymbol{\varepsilon}}{\sigma})/q}{(\frac{\boldsymbol{\varepsilon}}{\sigma})^{\mathrm{T}}\boldsymbol{\Gamma}_2\boldsymbol{\Gamma}_2^{\mathrm{T}}(\frac{\boldsymbol{\varepsilon}}{\sigma})/(n-p)} \sim F(q, n-p). \tag{3.2.7}$$

When $\mathbf{C} \in \mathbb{R}^{1\times p}$ has 1 in the jth entry and zeros elsewhere, by $t^2(n-p) = F(1, n-p)$ and $-T_j \overset{d}{=} T_j$, we have (3.2.6). $\square$

Similar to Corollary 3.2.1, Corollary 3.2.2 replaces the normal distribution with $\mathrm{Unif}(S^{n-1})$ in Proposition 3.2.2:

Corollary 3.2.2. *Suppose the error $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2\mathbf{I}_n)$ from the linear model (3.2.3) in Proposition 3.2.2 is replaced by $\boldsymbol{\theta} \sim \mathrm{Unif}(S^{n-1})$, while other conditions in Proposition 3.2.2 remain unchanged. Then (3.2.5) and (3.2.6) still hold.*

Proof. In (3.2.7), we replace $\boldsymbol{\varepsilon}/\sigma \sim \mathcal{N}(\mathbf{0}_n, \mathbf{I}_n)$ with $(\boldsymbol{\varepsilon}/\sigma)/\|\boldsymbol{\varepsilon}/\sigma\|_2 \sim \mathrm{Unif}(S^{n-1})$. Then we have the same conclusion. $\square$

In Proposition 3.2.2 and Corollary 3.2.2, the error vector follows the spherical distribution while the design matrix is an arbitrary fixed matrix with full column rank. Our question is, if the design matrix is random with spherical symmetry while the error vector has nonzero mean, heteroscedasticity, strong correlation, heavy tails, or is even fixed, do the F-test in (3.2.5) and t-test in (3.2.6) remain valid? In fact, the benefit of randomized design has been noticed since Fisher's field experiments at the Rothamsted Experimental Station. As pointed out on p.164 of Stigler (2016): "*(Fisher recognized that) the design randomization itself could induce a discretized spherical symmetry, granting approximated validity to procedures that seemed to require much stronger conditions. ... This subtle point was not widely grasped.*" Our Theorem 3.4.1 in Section 3.4 affirms the validity of the F and t tests under random design with spherical symmetry.

3.3 A New Geometric Proof of Finite Sample F Result

Our first main result below gives a new definition of the F distribution, which is parallel to Proposition 3.2.1 and Corollary 3.2.1. Yet we fix $\boldsymbol{\theta}$ to be an arbitrary unit vector $\boldsymbol{\vartheta}$ while set $\mathbf{Z}$ to be random.

Theorem 3.3.1. *Suppose that* $\mathbf{Z} = [\mathbf{Z}_1\ \mathbf{Z}_2]$ *is a random matrix with* $\mathbf{Z}_1 \in \mathbb{R}^{n\times q}$ *and* $\mathbf{Z}_2 \in \mathbb{R}^{n\times(p-q)}$ *such that* $1 \le q \le p < n$, $\mathbb{P}\{\mathrm{rank}(\mathbf{Z}) < p\} = 0$, *and* $\mathbf{Z} \overset{d}{=} \mathbf{\Phi Z}$ *for any orthogonal matrix* $\mathbf{\Phi} \in \mathbb{R}^{n\times n}$. *Then for any fixed* $\boldsymbol{\vartheta}$ *with* $\|\boldsymbol{\vartheta}\|_2 = 1$, *we have*

$$F = \frac{\boldsymbol{\vartheta}^{\mathrm{T}}(\mathbf{P}_{\mathbf{Z}} - \mathbf{P}_{\mathbf{Z}_2})\boldsymbol{\vartheta}/q}{\boldsymbol{\vartheta}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P}_{\mathbf{Z}})\boldsymbol{\vartheta}/(n-p)} \sim F(q, n-p). \tag{3.3.1}$$

Proof. First, we assume that a random vector $\boldsymbol{\theta} \in \mathbb{R}^n$ follows $\mathrm{Unif}(S^{n-1})$, which is independent of $\mathbf{Z}$, *i.e.* , $\boldsymbol{\theta} \sim \mathrm{Unif}(S^{n-1}) \perp\!\!\!\perp \mathbf{Z}$ and we define

$$F' = \frac{\boldsymbol{\theta}^{\mathrm{T}}(\mathbf{P}_{\mathbf{Z}} - \mathbf{P}_{\mathbf{Z}_2})\boldsymbol{\theta}/q}{\boldsymbol{\theta}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P}_{\mathbf{Z}})\boldsymbol{\theta}/(n-p)}. \tag{3.3.2}$$

By Corollary 3.2.1, we have $F'|\mathbf{Z} \sim F(q, n-p)$ for any given $\mathbf{Z}$. Hence we have $F' \perp\!\!\!\perp \mathbf{Z}$ and the marginal distribution of F' also follows $F(q, n-p)$.

Next, we show that that $F' \perp\!\!\!\perp \boldsymbol{\theta}$ in (3.3.2) by proving $(\boldsymbol{\theta}|F' = f') \sim \mathrm{Unif}(S^{n-1})$ for any $f' \ge 0$. For any orthogonal matrix $\mathbf{\Phi} \in \mathbb{R}^{n\times n}$, from (3.3.2), we have

$$\begin{aligned} F' &= \frac{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{\Phi\Phi}^{\mathrm{T}}(\mathbf{P}_{\mathbf{Z}} - \mathbf{P}_{\mathbf{Z}_2})\mathbf{\Phi\Phi}^{\mathrm{T}}\boldsymbol{\theta}/q}{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{\Phi\Phi}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P}_{\mathbf{Z}})\mathbf{\Phi\Phi}^{\mathrm{T}}\boldsymbol{\theta}/(n-p)} \\ &= \frac{\overline{\boldsymbol{\theta}}^{\mathrm{T}}(\mathbf{P}_{\overline{\mathbf{Z}}} - \mathbf{P}_{\overline{\mathbf{Z}}_2})\overline{\boldsymbol{\theta}}/q}{\overline{\boldsymbol{\theta}}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P}_{\overline{\mathbf{Z}}})\overline{\boldsymbol{\theta}}/(n-p)}, \end{aligned} \tag{3.3.3}$$

where in the second equality, we define $\overline{\boldsymbol{\theta}} = \mathbf{\Phi}^{\mathrm{T}}\boldsymbol{\theta}$ and $\overline{\mathbf{Z}} = \mathbf{\Phi}^{\mathrm{T}}\mathbf{Z}$. By the spherical symmetry of $\mathrm{Unif}(S^{n-1})$ and $\mathbf{Z}$, we have $\overline{\boldsymbol{\theta}} \sim \mathrm{Unif}(S^{n-1})$, which is independent of $\overline{\mathbf{Z}} = [\mathbf{\Phi}^{\mathrm{T}}\mathbf{Z}_1\ \mathbf{\Phi}^{\mathrm{T}}\mathbf{Z}_2] = [\overline{\mathbf{Z}}_1\ \overline{\mathbf{Z}}_2] \overset{d}{=} \mathbf{Z}$. Comparing (3.3.2) and (3.3.3), we have $(\boldsymbol{\theta}|F' = f') \overset{d}{=} (\overline{\boldsymbol{\theta}}|F' = f')$, that is, $(\boldsymbol{\theta}|F' = f') \overset{d}{=} (\mathbf{\Phi}^{\mathrm{T}}\boldsymbol{\theta}|F' = f')$ for any orthogonal matrix $\mathbf{\Phi}$. Additionally, given $F' = f'$ in (3.3.2), we still have $\|\boldsymbol{\theta}\|_2 = 1$. By Proposition 3.1.3, we have $(\boldsymbol{\theta}|F' = f') \sim \mathrm{Unif}(S^{n-1})$ for any $f' \ge 0$. Hence $F' \perp\!\!\!\perp \boldsymbol{\theta}$.

Therefore, given any fixed realization $\boldsymbol{\vartheta}$ of $\boldsymbol{\theta}$ from $\mathrm{Unif}(S^{n-1})$ in (3.3.2), $(F'|\boldsymbol{\theta} = \boldsymbol{\vartheta})$ has the form (3.3.1) and follows the marginal distribution $F(q, n-p)$. □

Remark 3.3.1. *The proof of Theorem 3.3.1 only uses the definitions of F distribution and* $\mathrm{Unif}(S^{n-1})$ *without manipulating the density function. Similar approach can be found in the proof of the distribution of eigenvectors for square matrices with i.i.d.* $\mathcal{N}(0,1)$ *entries (Rudelson & Vershynin, 2015). Interestingly, in* (3.3.2), F', $\boldsymbol{\theta}$, *and* $\mathbf{Z}$ *are pairwise independent but not mutually independent.*

The conclusion of Theorem 3.3.1 remains valid if $\boldsymbol{\vartheta}$ *is any nonzero vector in* $\mathbb{R}^n$. *This is because, by dividing both the numerator and the denominator in* (3.3.1) *by* $\|\boldsymbol{\vartheta}\|_2^2$, *we standardize* $\boldsymbol{\vartheta}$ *in* (3.3.1) *such that* $\boldsymbol{\vartheta} \in S^{n-1}$.

According to the conclusion of Theorem 3.3.1, if $\boldsymbol{\vartheta}$ *follows any distribution on the unit sphere that is independent of* $\mathbf{Z}$, *we still have* $F \perp\!\!\!\perp \boldsymbol{\vartheta}$ *and* $F \sim F(q, n-p)$.

Combining (3.3.1) in Theorem 3.3.1 and (3.2.2) in Corollary 3.2.1, we realize that the F distribution can be obtained by the spherical symmetry either from the space of $\mathbf{Z}$ or from the vector $\boldsymbol{\theta}$. From a geometric point of view, the F distribution can be interpreted as the squared length ratio between $(\mathbf{P_Z} - \mathbf{P_{Z_2}})\boldsymbol{\theta}$ and $(\mathbf{I}_n - \mathbf{P_Z})\boldsymbol{\theta}$, *i.e.* , the vector $\boldsymbol{\theta}$ on S^{n-1} projected onto two orthogonal spaces: $\mathcal{C}(\mathbf{Z}) \cap \mathcal{C}(\mathbf{Z}_2)^{\perp}$ and $\mathcal{C}(\mathbf{Z})^{\perp}$. Whether the spaces are fixed and the distribution of the vector $\boldsymbol{\theta}$ has rotational invariance, or the vector is fixed at $\boldsymbol{\vartheta}$ and the distribution of the space basis $\mathbf{Z}$ has rotational invariance, the distribution of the squared length ratio, after adjusted by the degrees of freedom, remains the F distribution.

In Figure 3.1, we present a simplified visualization for the case where $n = 3$ and $q = p = 2$ such that $\mathbf{P_{Z_2}} = \mathbf{O}$. The F distribution is determined by $\|\mathbf{P_Z}\boldsymbol{\theta}\|_2 / \|(\mathbf{I}_n - \mathbf{P_Z})\boldsymbol{\theta}\|_2$. By fixing the plane $\mathcal{C}(\mathbf{Z})$ and allowing $\boldsymbol{\theta}$ to vary on S^2 with rotational invariance, or by fixing $\boldsymbol{\theta}$ on S^2 and allowing the plane $\mathcal{C}(\mathbf{Z})$ to vary with rotational invariance, we obtain the same $F(2, 1)$ distribution.

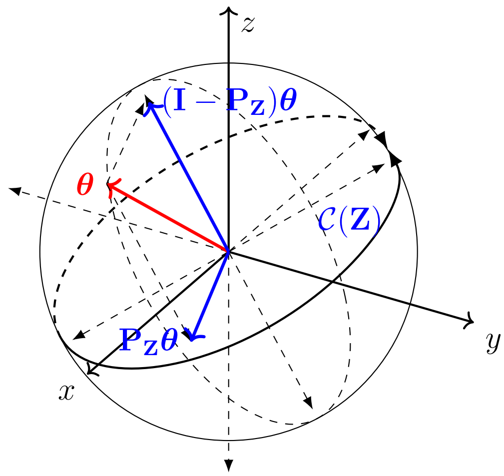


Figure 3.1: A geometric view of F distribution

3.4 Application in Linear Regression

In this section, we will provide a theorem parallel to Proposition 3.2.2 and Corollary 3.2.2. Unlike the fixed $\mathbf{X}$ in the linear model (3.2.3), we assume

$$\mathbf{X} = \mathbf{Z}\boldsymbol{\Sigma}^{1/2}, \tag{3.4.1}$$

where $\mathbf{Z} \in \mathbb{R}^{n\times p}$ is random, and $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ is fixed and positive definite, accounting for an extra adjustment for the covariance within each row of $\mathbf{X}$.

The following theorem shows that if the random $\mathbf{Z}$ is independent of $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ and has spherical symmetry, the OLS F-test and t-test still hold, even when $\boldsymbol{\varepsilon}$ has nonzero mean, heteroscedasticity, strong correlation, or heavy tails. The randomness in $\mathbf{X}$ is a remedy for the model violation in $\boldsymbol{\varepsilon}$, which echoes the statement in Stigler (2004): "*Fisher knew that with a rich enough randomization set, the act of randomization induced a spherically symmetric distribution conditional upon the data.*"

Theorem 3.4.1 (The new F-test and t-test parallel to Proposition 3.2.2)**.** *Under the model* (3.2.3) *and* (3.4.1)*, suppose* $\mathbb{P}\{\text{rank}(\mathbf{Z}) < p\} = 0$*,* $\mathbb{P}\{\boldsymbol{\varepsilon} = \mathbf{0}_n\} = 0$*, and* $\boldsymbol{\varepsilon} \perp\!\!\!\perp \mathbf{Z}$*. For any orthogonal matrix* $\boldsymbol{\Phi} \in \mathbb{R}^{n\times n}$*, we have* $\boldsymbol{\Phi}\mathbf{Z} \overset{d}{=} \mathbf{Z}$*. Let* $\mathbf{C} \in \mathbb{R}^{q\times p}$ *be a fixed matrix with*

$\mathrm{rank}(\mathbf{C}) = q$ *and* $1 \leq q \leq p < n$. *Then the* F *distribution in* (3.2.5) *and the* t *distribution in* (3.2.6) *from Proposition 3.2.2 still hold.*

Before the proof of Theorem 3.4.1, we collect some results about the Frisch–Waugh–Lovell Theorem (Frisch & Waugh, 1933; Lovell, 1963) summarized in Chap.7 of Ding (2024), which is used frequently in the proof of Theorem 3.4.1.

Proposition 3.4.1 (The Frisch–Waugh–Lovell Theorem). *Suppose that we have* $\mathbf{y} \in \mathbb{R}^n$ *and* $\mathbf{W} = [\mathbf{W}_1, \mathbf{W}_2] \in \mathbb{R}^{n\times p}$ *with* $\mathbf{W}_1 \in \mathbb{R}^{n\times q}$, $\mathbf{W}_2 \in \mathbb{R}^{n\times(p-q)}$, $\mathrm{rank}(\mathbf{W}) = p$, *and* $1 \leq q \leq p \leq n$. *Define* $\widehat{\boldsymbol{\gamma}} = (\mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{W}^{\mathrm{T}}\mathbf{y} \in \mathbb{R}^p$ *as the projection coefficient of* $\mathbf{y}$ *onto* $\mathcal{C}(\mathbf{W})$, *although there is not necessarily a linear model between* $\mathbf{W}$ *and* $\mathbf{y}$. *Also,* $\widehat{\boldsymbol{\gamma}}$ *has sub-vectors* $\widehat{\boldsymbol{\gamma}} = [\widehat{\boldsymbol{\gamma}}_1^{\mathrm{T}}, \widehat{\boldsymbol{\gamma}}_2^{\mathrm{T}}]^{\mathrm{T}}$, *where* $\widehat{\boldsymbol{\gamma}}_1 \in \mathbb{R}^q$ *and* $\widehat{\boldsymbol{\gamma}}_2 \in \mathbb{R}^{p-q}$ *correspond to* $\mathbf{W}_1$ *and* $\mathbf{W}_2$, *respectively. Define* $\widetilde{\mathbf{y}} = (\mathbf{I}_n - \mathbf{P}_{\mathbf{W}_2})\mathbf{y}$ *and* $\widetilde{\mathbf{W}}_1 = (\mathbf{I}_n - \mathbf{P}_{\mathbf{W}_2})\mathbf{W}_1$. *Then we have*

(i) $\mathrm{rank}(\widetilde{\mathbf{W}}_1) = q$ *and*

$$\widehat{\boldsymbol{\gamma}}_1 = (\widetilde{\mathbf{W}}_1^{\mathrm{T}}\widetilde{\mathbf{W}}_1)^{-1}\widetilde{\mathbf{W}}_1^{\mathrm{T}}\mathbf{y} = (\widetilde{\mathbf{W}}_1^{\mathrm{T}}\widetilde{\mathbf{W}}_1)^{-1}\widetilde{\mathbf{W}}_1^{\mathrm{T}}\widetilde{\mathbf{y}};$$

(ii) For $(\mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1}$, *the top left* $q \times q$ *sub-matrix is* $(\widetilde{\mathbf{W}}_1^{\mathrm{T}}\widetilde{\mathbf{W}}_1)^{-1}$ *and the bottom right* $(p-q)\times(p-q)$ *sub-matrix is* $[\mathbf{W}_2^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P}_{\mathbf{W}_1})\mathbf{W}_2]^{-1}$;

(iii) $\mathbf{P}_{\mathbf{W}} = \mathbf{P}_{\mathbf{W}_2} + \mathbf{P}_{\widetilde{\mathbf{W}}_1}$, $\mathbf{P}_{\mathbf{W}_2}\mathbf{P}_{\widetilde{\mathbf{W}}_1} = \mathbf{O}$;

(iv) $(\mathbf{I}_n - \mathbf{P}_{\mathbf{W}})\mathbf{y} = (\mathbf{I}_n - \mathbf{P}_{\widetilde{\mathbf{W}}_1})\widetilde{\mathbf{y}}$.

Now we are ready for the proof:

Proof of Theorem 3.4.1.

Step 1: algebraic simplification: Before the main proof, we first simplify F defined in (3.2.5) into a more concise form.

Let $\mathbf{D}_1 = \mathbf{C} \in \mathbb{R}^{q\times p}$. Since $\text{rank}(\mathbf{D}_1) = \text{rank}(\mathbf{C}) = q \le p$, there exists $\mathbf{D}_2 \in \mathbb{R}^{(p-q)\times p}$ such that

$$\mathbf{D}_3 = \begin{bmatrix} \mathbf{D}_1 \\ \mathbf{D}_2 \end{bmatrix} \in \mathbb{R}^{p\times p} \tag{3.4.2}$$

is invertible. Then we define a transformed design matrix as

$$\mathbf{W} = \mathbf{X}\mathbf{D}_3^{-1} = [\mathbf{W}_1, \mathbf{W}_2] \in \mathbb{R}^{n\times p}, \tag{3.4.3}$$

where $\mathbf{W}_1 \in \mathbb{R}^{n\times q}$ and $\mathbf{W}_2 \in \mathbb{R}^{n\times(p-q)}$ are the first q columns and the rest $p-q$ columns of $\mathbf{W}$, respectively. Plugging $\mathbf{W}$ in the linear model (3.2.3), we have

$$\mathbf{y} = \mathbf{W}\mathbf{D}_3\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{W}_1\mathbf{D}_1\boldsymbol{\beta} + \mathbf{W}_2\mathbf{D}_2\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{3.4.4}$$

Define $\widetilde{\mathbf{y}} = (\mathbf{I}_n - \mathbf{P}_{\mathbf{W}_2})\mathbf{y}$, $\widetilde{\mathbf{W}}_1 = (\mathbf{I}_n - \mathbf{P}_{\mathbf{W}_2})\mathbf{W}_1$, and $\widetilde{\boldsymbol{\varepsilon}} = (\mathbf{I}_n - \mathbf{P}_{\mathbf{W}_2})\boldsymbol{\varepsilon}$ so that

$$\widetilde{\mathbf{y}} = \widetilde{\mathbf{W}}_1\mathbf{D}_1\boldsymbol{\beta} + \widetilde{\boldsymbol{\varepsilon}}. \tag{3.4.5}$$

Now we will use $\mathbf{W}$ and $\widetilde{\mathbf{W}}_1$ to replace $\mathbf{X}$ in (3.2.5). Recalling $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$ from (3.2.4), we have

$$\begin{aligned} \mathbf{D}_3\widehat{\boldsymbol{\beta}} &= [(\mathbf{D}_3^{-1})^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{D}_3^{-1}]^{-1}(\mathbf{D}_3^{-1})^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} \\ &= (\mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{W}^{\mathrm{T}}\mathbf{y}, \end{aligned}$$

which is regressing $\mathbf{y}$ against the new design matrix $\mathbf{W}$ in (3.4.4). By Proposition 3.4.1 (i), the first q entries in $\mathbf{D}_3\widehat{\boldsymbol{\beta}}$, *i.e.* , $\mathbf{D}_1\widehat{\boldsymbol{\beta}}$ satisfies $\mathbf{D}_1\widehat{\boldsymbol{\beta}} = (\widetilde{\mathbf{W}}_1^{\mathrm{T}}\widetilde{\mathbf{W}}_1)^{-1}\widetilde{\mathbf{W}}_1^{\mathrm{T}}\widetilde{\mathbf{y}}$. By (3.4.5) and $\widetilde{\mathbf{W}}_1^{\mathrm{T}}\widetilde{\boldsymbol{\varepsilon}} = \widetilde{\mathbf{W}}_1^{\mathrm{T}}\boldsymbol{\varepsilon}$, we have

$$\mathbf{C}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \mathbf{D}_1(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = (\widetilde{\mathbf{W}}_1^{\mathrm{T}}\widetilde{\mathbf{W}}_1)^{-1}\widetilde{\mathbf{W}}_1^{\mathrm{T}}\boldsymbol{\varepsilon}. \tag{3.4.6}$$

By Proposition 3.4.1 (ii), the top left $q \times q$ sub-matrix of $(\mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1}$ is $(\widetilde{\mathbf{W}}_1^{\mathrm{T}}\widetilde{\mathbf{W}}_1)^{-1}$. Also by the definition of $\mathbf{W}$ in (3.4.3), we have

$$(\mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1} = [(\mathbf{D}_3^{-1})^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{D}_3^{-1}]^{-1} = \mathbf{D}_3(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{D}_3^{\mathrm{T}}.$$

By the definition of $\mathbf{D}_3$ in (3.4.2), the top left $q \times q$ sub-matrix of $(\mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1}$ is

$$\mathbf{D}_1(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{D}_1^{\mathrm{T}} = \mathbf{C}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{C}^{\mathrm{T}}.$$

Therefore, we have

$$\mathbf{C}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{C}^{\mathrm{T}} = (\widetilde{\mathbf{W}}_1^{\mathrm{T}}\widetilde{\mathbf{W}}_1)^{-1}. \tag{3.4.7}$$

Combining (3.4.6) and (3.4.7), for the numerator of the F in (3.2.5), we have

$$\begin{aligned} & (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}}\mathbf{C}^{\mathrm{T}}\left[\mathbf{C}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{C}^{\mathrm{T}}\right]^{-1}\mathbf{C}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ = \; & \boldsymbol{\varepsilon}^{\mathrm{T}}\mathbf{P}_{\widetilde{\mathbf{W}}_1}\boldsymbol{\varepsilon} \\ = \; & \boldsymbol{\varepsilon}^{\mathrm{T}}(\mathbf{P}_{\mathbf{W}} - \mathbf{P}_{\mathbf{W}_2})\boldsymbol{\varepsilon}, \end{aligned}$$

where the last step is from Proposition 3.4.1 (iii). For the denominator of (3.2.5), we have $(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}})\mathbf{y} = (\mathbf{I}_n - \mathbf{P}_{\mathbf{X}})\boldsymbol{\varepsilon}$. Also by $\mathcal{C}(\mathbf{X}) = \mathcal{C}(\mathbf{W})$, we have $\mathbf{P}_{\mathbf{X}} = \mathbf{P}_{\mathbf{W}}$. Thus F in (3.2.5) becomes

$$\begin{aligned} F &= \frac{\boldsymbol{\varepsilon}^{\mathrm{T}}(\mathbf{P}_{\mathbf{W}} - \mathbf{P}_{\mathbf{W}_2})\boldsymbol{\varepsilon}/q}{\boldsymbol{\varepsilon}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P}_{\mathbf{W}})\boldsymbol{\varepsilon}/(n-p)} \\ &= \frac{\left(\frac{\boldsymbol{\varepsilon}}{\|\boldsymbol{\varepsilon}\|_2}\right)^{\mathrm{T}}(\mathbf{P}_{\mathbf{W}} - \mathbf{P}_{\mathbf{W}_2})\left(\frac{\boldsymbol{\varepsilon}}{\|\boldsymbol{\varepsilon}\|_2}\right)/q}{\left(\frac{\boldsymbol{\varepsilon}}{\|\boldsymbol{\varepsilon}\|_2}\right)^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P}_{\mathbf{W}})\left(\frac{\boldsymbol{\varepsilon}}{\|\boldsymbol{\varepsilon}\|_2}\right)/(n-p)}. \end{aligned} \tag{3.4.8}$$

Step 2: probabilistic arguments: Now we prove the F distribution in (3.2.5) with the help of Theorem 3.3.1. Given any fixed realization of $\boldsymbol{\varepsilon}$, by $\boldsymbol{\varepsilon} \perp\!\!\!\perp \mathbf{Z}$, $F|\boldsymbol{\varepsilon}$ has the form (3.4.8). For arbitrary orthogonal matrix $\boldsymbol{\Phi} \in \mathbb{R}^{n\times n}$, since $\boldsymbol{\Phi}\mathbf{Z} \overset{d}{=} \mathbf{Z}$, we have $\boldsymbol{\Phi}\mathbf{Z}\boldsymbol{\Sigma}^{1/2}\mathbf{D}_3^{-1} \overset{d}{=} \mathbf{Z}\boldsymbol{\Sigma}^{1/2}\mathbf{D}_3^{-1}$, *i.e.* , $\mathbf{\Phi W} \stackrel{d}{=} \mathbf{W}$. By Theorem 3.3.1, $F|\boldsymbol{\varepsilon}$ in (3.4.8) follows $F(q, n-p)$. Hence in (3.4.8), we have $\boldsymbol{\varepsilon} \perp\!\!\!\perp F$, and the marginal distribution of F follows $F(q, n-p)$.

For T_j in (3.2.6) with the special case of $\mathbf{C} = \mathbf{e}_j^{\mathrm{T}} \in \mathbb{R}^{1\times p}$, where $\mathbf{e}_j$ is the jth standard basis in $\mathbb{R}^p$, following the same algebraic procedures in (3.4.2)–(3.4.8), we have

$$T_j = \frac{(\widetilde{\mathbf{W}}_1^{\mathrm{T}}\widetilde{\mathbf{W}}_1)^{-1}\widetilde{\mathbf{W}}_1^{\mathrm{T}}\boldsymbol{\varepsilon}}{\sqrt{(\widetilde{\mathbf{W}}_1^{\mathrm{T}}\widetilde{\mathbf{W}}_1)^{-1}}\sqrt{\frac{\boldsymbol{\varepsilon}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P}_{\mathbf{W}})\boldsymbol{\varepsilon}}{n-p}}},$$

which implies that $T_j^2 \sim F(1, n-p)$. Also since $-\mathbf{X} \stackrel{d}{=} \mathbf{X}$ and

$$T_j = \frac{\mathbf{e}_j^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{\varepsilon}}{\sqrt{\mathbf{e}_j^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{e}_j}\sqrt{\frac{\boldsymbol{\varepsilon}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}})\boldsymbol{\varepsilon}}{n-p}}},$$

we have $-T_j \stackrel{d}{=} T_j$. Using the fact that $t^2(n-p) \stackrel{d}{=} F(1, n-p)$, we have the conclusion.

□

3.4.1 Extension: Mixed Regressors

In practice, $\mathbf{X}$ may include both fixed and random regressors. In this section, Corollary 3.4.1 shows that if the random regressors have spherical symmetry, the F-test and t-test for the coefficients of the random regressors are still valid.

Specifically, in the linear model (3.2.3), let $\mathbf{X} = [\mathbf{X}_0\ \mathbf{X}_1] \in \mathbb{R}^{n\times p}$ with $\mathrm{rank}(\mathbf{X}) = p$, where $\mathbf{X}_0 \in \mathbb{R}^{n\times s}$ is fixed and $\mathbf{X}_1 \in \mathbb{R}^{n\times d}$ is random with $p = s + d$. Let $\boldsymbol{\beta} = [\boldsymbol{\beta}_0^{\mathrm{T}}\ \boldsymbol{\beta}_1^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^p$ where $\boldsymbol{\beta}_0 \in \mathbb{R}^s$ and $\boldsymbol{\beta}_1 \in \mathbb{R}^d$ correspond to $\mathbf{X}_0$ and $\mathbf{X}_1$, respectively. So the model in (3.2.3) becomes

$$\mathbf{y} = \mathbf{X}_0\boldsymbol{\beta}_0 + \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}. \tag{3.4.9}$$

The OLS estimator in (3.2.4) can be written as $\widehat{\boldsymbol{\beta}} = [\widehat{\boldsymbol{\beta}}_0^{\mathrm{T}}\ \widehat{\boldsymbol{\beta}}_1^{\mathrm{T}}]^{\mathrm{T}}$.

Our interest is the inference on $\widehat{\boldsymbol{\beta}}_1$. So we define

$$\mathbf{C} = [\mathbf{O}_{q\times s}, \mathbf{C}_1] \in \mathbb{R}^{q\times(s+d)}, \tag{3.4.10}$$

where $\mathbf{C}_1 \in \mathbb{R}^{q\times d}$ is any fixed matrix with $\mathrm{rank}(\mathbf{C}_1) = q$. Then we have $\mathbf{C}\widehat{\boldsymbol{\beta}} = \mathbf{C}_1\widehat{\boldsymbol{\beta}}_1$. Suppose $1 \le q \le d \le d + s = p < n$. If $s = 0$, $\mathbf{X}$ is simplified to $\mathbf{X}_1$ and $\mathbf{C} = \mathbf{C}_1$. The random $\mathbf{X}_1$ follows the model (3.4.1), *i.e.* , $\mathbf{X}_1 = \mathbf{Z}\boldsymbol{\Sigma}^{1/2}$, where $\mathbf{Z} \in \mathbb{R}^{n\times d}$ is random and $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ is positive definite, providing an extra adjustment for the row-wise covariance.

Corollary 3.4.1 (The F-test and t-test with mixed regressors)**.** *Suppose* $\mathbb{P}\{\mathrm{rank}(\mathbf{X}) < s+d\} = 0$ *and* $\boldsymbol{\varepsilon} \perp\!\!\!\perp \mathbf{Z}$*. If there exists an orthonormal basis of* $\mathcal{C}(\mathbf{X}_0)^\perp$ *which forms the matrix* $\boldsymbol{\Gamma} \in \mathbb{R}^{n\times(n-s)}$*, such that for any orthogonal matrix* $\boldsymbol{\Phi} \in \mathbb{R}^{(n-s)\times(n-s)}$*, we have* $\boldsymbol{\Phi}\boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{Z} \overset{d}{=} \boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{Z}$ *and* $\mathbb{P}\{\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\varepsilon} = \mathbf{0}_{n-s}\} = 0$*, then the F distribution in* (3.2.5) *still holds. The t distribution for* T_j *in* (3.2.6) *also holds if* $j = s + k$ *with* $k = 1, \dots, d$*.*

Proof of Corollary 3.4.1. We will simplify F in (3.2.5) to get rid of $\mathbf{X}_0$. Then we can apply Theorem 3.4.1 to obtain the results. By the definition of $\boldsymbol{\Gamma} \in \mathbb{R}^{n\times(n-s)}$, we have

$$\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_0} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^{\mathrm{T}}. \tag{3.4.11}$$

Let $\mathbf{y}^\star = \boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{y} \in \mathbb{R}^{n-s}$, $\mathbf{X}_1^\star = \boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{X}_1 \in \mathbb{R}^{(n-s)\times d}$, $\mathbf{Z}^\star = \boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{Z} \in \mathbb{R}^{(n-s)\times d}$, and $\boldsymbol{\varepsilon}^\star = \boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\varepsilon} \in \mathbb{R}^{n-s}$. Multiplying $\boldsymbol{\Gamma}^{\mathrm{T}}$ on both sides of (3.4.9), we have

$$\mathbf{y}^\star = \mathbf{X}_1^\star\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}^\star. \tag{3.4.12}$$

Define the OLS estimator in (3.4.12) as

$$\widehat{\boldsymbol{\beta}}_1^\star = (\mathbf{X}_1^{\star\mathrm{T}}\mathbf{X}_1^\star)^{-1}\mathbf{X}_1^{\star\mathrm{T}}\mathbf{y}^\star. \tag{3.4.13}$$

We will apply Theorem 3.4.1 to (3.4.12) after completing the algebra in the following (3.4.14)–(3.4.17).

From Proposition 3.4.1 (i) and (3.4.11), we have

$$\widehat{\boldsymbol{\beta}}_1 = [\mathbf{X}_1^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_0})\mathbf{X}_1]^{-1}\mathbf{X}_1^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_0})\mathbf{y} = \widehat{\boldsymbol{\beta}}_1^\star, \tag{3.4.14}$$

which implies that

$$\mathbf{C}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \mathbf{C}_1(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) = \mathbf{C}_1(\widehat{\boldsymbol{\beta}}_1^\star - \boldsymbol{\beta}_1). \tag{3.4.15}$$

From Proposition 3.4.1 (ii) and the definition of $\mathbf{C}$ in (3.4.10), we have

$$\begin{aligned} \mathbf{C}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{C}^{\mathrm{T}} &= \mathbf{C}_1[\mathbf{X}_1^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_0})\mathbf{X}_1]^{-1}\mathbf{C}_1^{\mathrm{T}} \\ &= \mathbf{C}_1(\mathbf{X}_1^{\star\mathrm{T}}\mathbf{X}_1^\star)^{-1}\mathbf{C}_1^{\mathrm{T}}, \end{aligned}$$

where the second equality follows from (3.4.11).

Since $\mathbf{P}_{(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_0})\mathbf{X}_1}$ and $\mathbf{P}_{\mathbf{X}_1^\star}$ are the projection matrices of $\mathcal{C}((\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_0})\mathbf{X}_1)$ and $\mathcal{C}(\mathbf{X}_1^\star)$, respectively, by (3.4.11), we have

$$\mathbf{P}_{(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_0})\mathbf{X}_1} = \boldsymbol{\Gamma}\mathbf{P}_{\mathbf{X}_1^\star}\boldsymbol{\Gamma}^{\mathrm{T}}. \tag{3.4.16}$$

Then we have

$$\begin{aligned} (\mathbf{I}_n - \mathbf{P}_{\mathbf{X}})\mathbf{y} &= [\mathbf{I}_n - \mathbf{P}_{(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_0})\mathbf{X}_1}](\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_0})\mathbf{y} \\ &= (\mathbf{I}_n - \boldsymbol{\Gamma}\mathbf{P}_{\mathbf{X}_1^\star}\boldsymbol{\Gamma}^{\mathrm{T}})\boldsymbol{\Gamma}\mathbf{y}^\star \\ &= \boldsymbol{\Gamma}(\mathbf{I}_{n-s} - \mathbf{P}_{\mathbf{X}_1^\star})\mathbf{y}^\star, \end{aligned} \tag{3.4.17}$$

where the first equality follows from Proposition 3.4.1 (iv) and the second equality follows from (3.4.16) and (3.4.11).

Combining (3.4.14), (3.4.15), and (3.4.17), the F in (3.2.5) is equal to

$$\frac{\{(\widehat{\boldsymbol{\beta}}_1^\star - \boldsymbol{\beta}_1)^{\mathrm{T}}\mathbf{C}_1^{\mathrm{T}}\left[\mathbf{C}_1(\mathbf{X}_1^{\star\mathrm{T}}\mathbf{X}_1^\star)^{-1}\mathbf{C}_1^{\mathrm{T}}\right]^{-1}\mathbf{C}_1(\widehat{\boldsymbol{\beta}}_1^\star - \boldsymbol{\beta})\}/q}{\{\mathbf{y}^{\star\mathrm{T}}\left(\mathbf{I}_{n-s} - \mathbf{P}_{\mathbf{X}_1^\star}\right)\mathbf{y}^\star\}/(n-p)},$$

and the T_j in (3.2.6) with $j = s + k$ is equal to

$$\frac{\widehat{\beta}^{\star}_{1k} - \beta_{1k}}{\sqrt{(\mathbf{X}_1^{\star\mathrm{T}}\mathbf{X}_1^{\star})^{-1}_{[k,k]}}\sqrt{\mathbf{y}^{\star\mathrm{T}}(\mathbf{I}_{n-s} - \mathbf{P}_{\mathbf{X}_1^{\star}})\mathbf{y}^{\star}/(n-p)}},$$

where $\widehat{\beta}^{\star}_{1k}$ and β_{1k} are the kth entry of $\widehat{\boldsymbol{\beta}}^{\star}_1$ and $\boldsymbol{\beta}_1$, respectively, and $(\mathbf{X}_1^{\star\mathrm{T}}\mathbf{X}_1^{\star})^{-1}_{[k,k]}$ is the kth diagonal entry of $(\mathbf{X}_1^{\star\mathrm{T}}\mathbf{X}_1^{\star})^{-1}$.

Since for any orthogonal matrix $\boldsymbol{\Phi} \in \mathbb{R}^{(n-s)\times(n-s)}$, we have $\boldsymbol{\Phi}\mathbf{Z}^{\star} = \boldsymbol{\Phi}\boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{Z} \overset{d}{=} \boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{Z} = \mathbf{Z}^{\star}$, we can apply Theorem 3.4.1 to (3.4.12) to obtain the desired results. □

In Corollary 3.4.1, the matrix $\boldsymbol{\Gamma}$ is not hard to find. For any orthonormal basis of $\mathcal{C}(\mathbf{X}_0)^{\perp}$, if the columns of $\mathbf{Z}$ follow the independent spherical distribution in $\mathbb{R}^n$, by Corollary 3.1.1, the columns of $\boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{Z}$ also follow the independent spherical distribution in $\mathbb{R}^{n-s}$. Thus we have $\boldsymbol{\Phi}\boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{Z} \overset{d}{=} \boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{Z}$ for any orthogonal matrix $\boldsymbol{\Phi} \in \mathbb{R}^{(n-s)\times(n-s)}$. Then the result in Corollary 3.4.1 holds as long as $\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\varepsilon} \neq \mathbf{0}_{n-s}$ almost surely.

In the experimental study with limited sample size, as Fisher (1924) pointed out, "*some of the most important statistics do not tend to a normal distribution, and in many other cases, with small samples of the size usually available, the distribution is far from normal.*" Our results provide exact tests which are valid for small samples.

3.5 Numerical Evidence

Assume that a drug affects the biomarker y_i through the linear model $y_i = D_i\beta + \epsilon_i$, where $D_i \sim \mathcal{N}(\mu, \sigma^2)$ is the dose level with pre-determined μ and σ, $\beta \in \mathbb{R}$ is the drug effect to be investigated, and $\epsilon_i \in \mathbb{R}$ contains other effects which might have nonzero mean, strong correlation, heteroscedasticity, and heavy tails. If we can randomly generate D_i via $\mu + Z_i\sigma$ with $Z_i \sim \mathcal{N}(0, 1)$, even with a small sample size, Corollary 3.4.1 guarantees the exact t distribution for the standardized OLS estimator $\widehat{\beta}$.

Specifically, if the participants take the assigned drug with dose level $D_i = \mu + Z_i\sigma$ with $Z_i \sim \mathcal{N}(0, 1)$, the model becomes $y_i = \mu\beta + Z_i\sigma\beta + \epsilon_i$. Then the model for the n participants is

$$\mathbf{y} = \mathbf{X}_0\beta_0 + \mathbf{X}_1\beta_1 + \boldsymbol{\varepsilon},$$

where $\beta_0 = \beta_1 = \beta$, $\mathbf{X}_0 = \mu\mathbf{1}_n$, $\mathbf{X}_1 = \mathbf{Z}\sigma$, and $\mathbf{Z} = [Z_1, \ldots, Z_n]^{\mathrm{T}} \sim \mathcal{N}(\mathbf{0}_n, \mathbf{I}_n)$. Corollary 3.4.1 guarantees that the t-statistic for the random regressor, *i.e.* , $(\widehat{\beta}_1 - \beta_1)/\text{S.E.}(\widehat{\beta}_1)$, follows $t(n-2)$.

In this simulation, set $\mu = 10$, $\sigma = 2$, $\beta = 1$. To emphasize the validity with small samples, set $n = 4$. Over 100,000 replications, configure $\boldsymbol{\varepsilon}$ under four settings: S1: $\boldsymbol{\varepsilon}$ is fixed over all replications from $\mathcal{N}(\mathbf{0}_n, \mathbf{I}_n)$ with seed 12345 (fixed error); S2: $\boldsymbol{\varepsilon}_i$ follows i.i.d. $t(1)$ (heavy tails); S3: $\boldsymbol{\varepsilon} = \mathbf{V}^{1/2}\mathbf{w}$ where $\mathbf{V} \in \mathbb{R}^n$ has diagonal entries $1, 2, \ldots n$ and off-diagonal entries 0.9, and $\mathbf{w}$ has i.i.d. symmetric Bernoulli entries (correlation and heteroscedasticity); S4: $\boldsymbol{\varepsilon} = \boldsymbol{\mu}_\epsilon + \mathbf{V}^{1/2}\mathbf{w}$ where $\boldsymbol{\mu}_\epsilon$ is fixed from $\mathcal{N}(5\mathbf{1}_n, 100\mathbf{I}_n)$ with seed 12345, $\mathbf{V} \in \mathbb{R}^n$ is the same as in S3, and $\mathbf{w}$ has i.i.d. $t(1)$ entries. Figure 3.2 shows the QQ-plots of $(\widehat{\beta}_1 - \beta_1)/\text{S.E.}(\widehat{\beta}_1)$ against $t(n-2)$, which verify the validity of the exact t-test with small samples.

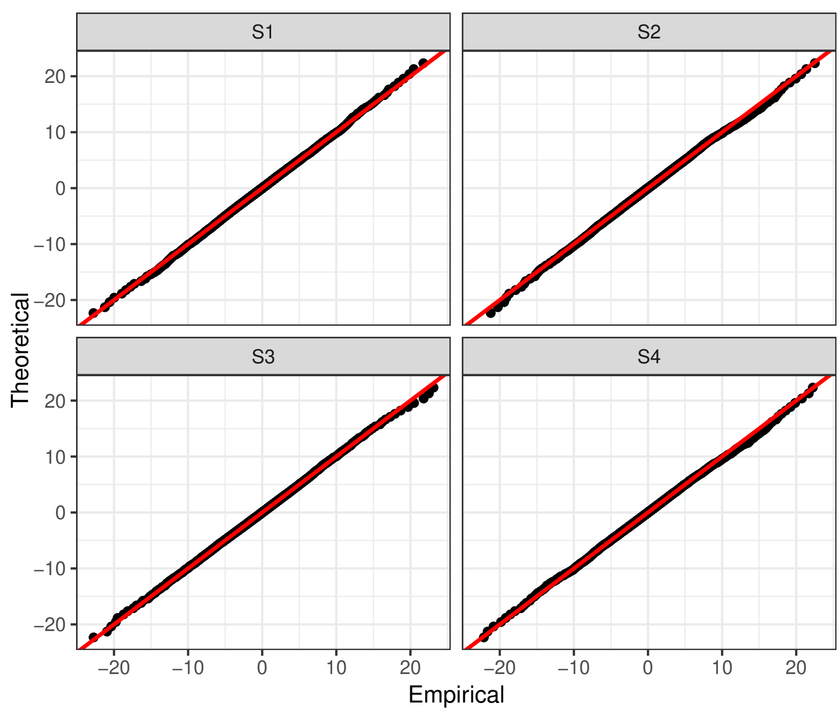


Figure 3.2: QQ-plots of $(\widehat{\beta}_1 - \beta_1)/\text{S.E.}(\widehat{\beta}_1)$ against $t(2)$

3.6 Discussion

In this paper, we generalize the definition of the F distribution using properties of the spherical distribution and provide a geometric interpretation. We extend this to the linear model by developing exact F-tests and t-tests, conditional on the error vector, assuming that the design matrix has spherical symmetry. This approach to utilizing the randomness given data is similar to Ding and Dasgupta (2018); A. Zhao and Ding (2021).

Beyond spherical symmetry in Definition 3.1.1, there are other symmetric distributions. For example, if $\mathbf{\Phi}$ is a permutation matrix, the distribution is *exchangeable*, which is commonly used in permutation tests (Freedman & Lane, 1983; Lei & Bickel, 2021). If $\mathbf{\Phi}$ is a diagonal matrix with entries of 1 or -1, the distribution has *orthant symmetry*, allowing $\mathbf{\Phi}$ to be assigned to a distribution for theoretical investigations (Efron, 1969; Eaton & Efron, 1970). Future studies may explore more general theoretical utilities of the symmetric distribution.

Chapter 4

High-dimensional moderated mediation analysis with heredity

4.1 Introduction

Over the past decade, we have witnessed an explosion in the interest in mediation analysis, which is widely used in economics (Guo et al., 2023), biology (Huang & Pan, 2016; Zhou et al., 2020), and many other disciplines (see MacKinnon (2012)). The Baron-Kenny approach (Baron & Kenny, 1986) provides a convenient framework to understand the causal effects of treatment T on the outcome Y, potentially mediated through a set of mediators $\mathbf{G} = [G_1, \ldots, G_p]^{\mathrm{T}}$. Specifically, the Baron-Kenny approach characterizes the concept of mediation by the following linear outcome and mediator models:

$$\mathbb{E}(Y \mid T, \mathbf{G}) = \theta_0^\star + T\theta_T^\star + \mathbf{G}^{\mathrm{T}}\boldsymbol{\theta}_G^\star, \quad \mathbb{E}(\mathbf{G} \mid T) = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 T.$$

Plugging the mediator model into the outcome model yields $\mathbb{E}(Y \mid T) = \alpha_0 + \alpha_T T$ with $\alpha_T = \theta_T^\star + \boldsymbol{\beta}_1^{\mathrm{T}}\boldsymbol{\theta}_G^\star$. Equipped with the counterfactual framework (Robins & Greenland, 1992; Pearl, 2001; Imai, Keele, & Yamamoto, 2010), the parameters α_T, $\theta_T^\star$, $\boldsymbol{\beta}_1^{\mathrm{T}}\boldsymbol{\theta}_G^\star$ have causal interpretations, known as *total effect*, *natural direct effect* (NDE), and *natural indirect effect* (NIE), respectively. The NDE can be conceptualized as the treatment effect on the outcome at a fixed level of the mediator variables, and the NIE can be understood as the effect on the outcome of changes of the treatment which operate through mediator levels; see Section 4.2.1 for a more detailed explanation. With low-dimensional mediators, the coefficients can be estimated by the conventional least squares. If we denote the estimator of NDE as $\widehat{\theta}_T^\star$, then NIE can be estimated equivalently, under continuous outcome, by $\widehat{\boldsymbol{\beta}}_1^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_G^\star$ and $\widehat{\alpha} - \widehat{\theta}_T^\star$. The former is known as *product method* and the latter is known as *difference method*

(T. VanderWeele, 2015; T. J. VanderWeele, 2016). Then the inference can be conducted by the usual delta method (Sobel, 1982).

An intriguing extension of the original Baron-Kenny approach arises when the outcome model incorporates the treatment-by-mediator interaction, namely, the outcome models becomes

$$\mathbb{E}(Y \mid T, \mathbf{G}) = \theta_0^\star + T\theta_T^\star + \mathbf{G}^{\mathrm{T}}\boldsymbol{\theta}_G^\star + T\mathbf{G}^{\mathrm{T}}\boldsymbol{\theta}_D^\star.$$

For instance, Powers and Swinton (1984); Holland (1988); Hong, Deutsch, and Hill (2015) investigated the case where the treatment was whether the students are encouraged to study for a test with provided study materials, the mediator was the number of study hours, and the outcome was the test performance. The students who received encouragement and materials not only studied longer but also more effectively compared to those without such support. This finding reveals that treatment influences both the mediator and its relationship with the outcome, which was also documented in other studies (T. J. VanderWeele et al., 2012; MacKinnon, Valente, & Gonzalez, 2020). Moreover, the inclusion of the interaction term enhances the power of detecting the mediation effects, particularly when $\boldsymbol{\theta}_G^\star$ and $\boldsymbol{\theta}_D^\star$ exhibit opposite effects; see Section 4.5 for simulated examples.

Given the importance of the treatment-by-mediator interaction, much of the existing literature (Preacher, Rucker, & Hayes, 2007; T. J. VanderWeele & Vansteelandt, 2009, 2010; Valeri & VanderWeele, 2013; T. J. VanderWeele, 2016) explored it in the low-dimensional scenarios. However, with the rise of high-throughput biological data, such as DNA methylation analysis (Song et al., 2021; Guo et al., 2022; Liu et al., 2022), the number of potential mediators p becomes much larger than the sample size n, presenting the challenges of high-dimensional mediation analysis.

4.1.0.0.1 Challenges of high-dimensional mediation analysis with interaction Although high-dimensional mediation analysis is developing rapidly as shown below, only

Huang and Pan (2016); C. Wang, Hu, Blaser, and Li (2020) considered the treatment-by-mediator interaction. Yet both of them lack comprehensive theoretical support. Specifically, the principal component mediation analysis (PCMA) in Huang and Pan (2016) transforms the original correlated mediator $\mathbf{G} = [G_1, \ldots, G_p]^{\mathrm{T}}$ to the uncorrelated mediator $\mathbf{G}^* = [G_1^*, \ldots, G_K^*]^{\mathrm{T}}$ by $\mathbf{G}^* = \widehat{\mathbf{U}}\mathbf{G}$, where $\widehat{\mathbf{U}}^{\mathrm{T}} \in \mathbb{R}^{p \times K}$ is obtained by performing PCA on the residual matrix of the mediator models. Under Gaussian assumption, $\mathbf{G}^*$ has independent entries which allows to fit K low-dimensional regressions separately. However, as pointed out in Section 4.5 of this paper, if $p > n$, the PCA cannot recover 100% of variance. Thus it introduces potential bias for the estimands and limits the usefulness under high-dimensional settings. Meanwhile, C. Wang et al. (2020) introduced a Lasso-type approach, but without establishing theoretical guarantees.

Although Zhou et al. (2020); Guo et al. (2023) provide rigorous theoretical guarantees, they do not address the treatment-by-mediator interaction. Specifically, Guo et al. (2023) adopted the nonconvex penalized least squares in the outcome model to obtain an estimator of $\widehat{\theta}_T^\star$ for the NDE with oracle property. Then the NIE is obtained by the difference method. This approach benefits from the low dimensionality of both the total effect, $\widehat{\alpha}_T$, and the direct effect, $\widehat{\theta}_T^\star$. However, in the presence of the interaction, the difference method does not work. Suppose $T \in \{0, 1\}$ for simplicity, then the NDE is $\theta_T^\star + (\boldsymbol{\beta}_0 + T\boldsymbol{\beta}_1)^{\mathrm{T}}\boldsymbol{\theta}_D^\star$ and NIE is $\boldsymbol{\beta}_1^{\mathrm{T}}(\boldsymbol{\theta}_G^\star + T\boldsymbol{\theta}_D^\star)$ (Valeri and VanderWeele (2013) or (4.2.3), (4.2.4)). The term $\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{\theta}_D^\star$ in NDE cannot be identified solely using the outcome model, rendering the difference method in Guo et al. (2022) infeasible.

The product method can be extended to allow treatment-by-mediator interactions, however, needs to handle the error accumulation from high-dimensional mediators. Take the inference on the NIE $\boldsymbol{\beta}_1^{\mathrm{T}}\boldsymbol{\theta}_G^\star$ as an example. Let $\widehat{\boldsymbol{\beta}}_1 = \boldsymbol{\beta}_1 + \boldsymbol{e}_{\beta_1}$ and $\widehat{\boldsymbol{\theta}}_G^\star = \boldsymbol{\theta}_G^\star + \boldsymbol{e}_{\theta_G}$. Then we have the following decomposition

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_1^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_G^\star - \boldsymbol{\beta}_1^{\mathrm{T}}\boldsymbol{\theta}_G^\star) = \sqrt{n}\boldsymbol{\theta}_G^{\star\mathrm{T}}\boldsymbol{e}_{\beta_1} + \sqrt{n}\widehat{\boldsymbol{\beta}}_1^{\mathrm{T}}\boldsymbol{e}_{\theta_G}. \tag{4.1.1}$$

As pointed out by (Zhou et al., 2020), when $p > n$, $\widehat{\boldsymbol{\theta}}_G^\star$ from Lasso is not recommended since the limiting distribution of $\sqrt{n}\boldsymbol{e}_{\theta_G}$ is intractable. If $\widehat{\boldsymbol{\theta}}_G^\star$ is from the debiased Lasso (Javanmard & Montanari, 2014; van de Geer, Bühlmann, Ritov, & Dezeure, 2014; C.-H. Zhang & Zhang, 2014), the debiased error $\boldsymbol{e}_{\theta_G}$ ignores the sparsity in $\boldsymbol{\theta}_G^\star$ and the error accumulation in $\sqrt{n}\widehat{\boldsymbol{\beta}}_1^{\mathrm{T}}\boldsymbol{e}_{\theta_G}$ is of order p and is intractable. So Zhou et al. (2020) proposed a modified debiased Lasso for the product method, which effectively reduces the error accumulation. However, the prerequisite CLIME procedure (Cai, Liu, & Luo, 2011) needs the the sparsity in $\mathbb{E}\{[\mathbf{G}^{\mathrm{T}}, T\mathbf{G}]^{\mathrm{T}}[G^{\mathrm{T}}, T\mathbf{G}^{\mathrm{T}}]\}$. Given the correlated nature between $T\mathbf{G}$ and $\mathbf{G}$, such an assumption is difficult to satisfy in practice.

4.1.0.0.2 Our threefold contributions

1. From a methodological perspective, we propose a two-step procedure that effectively handles high-dimensional mediation analysis with the treatment-by-mediator interaction under categorical treatments. In the first step, inspired by the work of Guo et al. (2023), we implement the nonconvex penalized least squares procedure in the outcome model. By the sparsity assumption in $\boldsymbol{\theta}_G^\star$ and $\boldsymbol{\theta}_D^\star$, we establish the oracle property in this step, which reduces the high-dimensional mediators to their low-dimensional support. In the second step, we employ the product method to conduct inference on the NIE and the NDE in the presence of interaction.

   As suggested in many applications (X. Li, Sudarsanam, & Frey, 2006; C. Yang, Wan, Yang, Xue, & Yu, 2010; S. Lee & Xing, 2012), the interactions should only be present if the response cannot be explained by additive functions of main effects. This is known as *heredity* (Peixoto, 1987; Hamada & Wu, 1992; Chipman, 1996). In our setting, selecting an interaction TG_j while dropping out the corresponding main effects from G_j is challenging to interpret. Thus we introduce an overlapping group SCAD method, which is solved by the local linear approximation (Zou & Li, 2008), *i.e.* , the overlapping group Lasso. This method not only promotes heredity but also maintains the oracle property.

2. From a theoretical perspective, we establish the oracle property for the outcome model with high-dimensional treatment-by-mediator interactions and categorical T. We further develop the asymptotic normality of the estimators for NIE and NDE.

   To address the error accumulation of the product method in (4.1.1), we observe that if the high-dimensional term $\boldsymbol{e}_{\theta_G}$ is sparse like $\boldsymbol{\theta}_G^\star$, we reduce (4.1.1) to a sum of s terms, where s is the sparsity level in $\boldsymbol{\theta}_G^\star$. The oracle property from the outcome model ensures that $\widehat{\boldsymbol{\theta}}_G^\star$ performs as well as the low-dimensional maximum likelihood estimator (MLE) on the support of $\boldsymbol{\theta}_G^\star$, with entries outside support being zeros with a probability tending to 1. Thus the oracle property successfully identifies the active mediators of order s. Meanwhile, since the sparsity level s is allowed to grow with n, the classic delta method for fixed-dimensional settings is not feasible. We carefully control the error accumulation in (4.1.1) by adjusting the sparsity and signal assumptions.

   Compared with Guo et al. (2022) which utilized the low-dimensionality in the difference method, our work involves the error control in the product method with high-dimensional mediators (4.1.1), necessitating more stringent conditions, $s = o(n^{1/4})$ and the bound of the signal in $\boldsymbol{\theta}_G^\star$. With categorical treatment and the treatment-by-mediator interactions, the Gram matrix of the outcome model is more complicated, leading to the development of new standard errors for the estimators. Compared with the modified debiased Lasso procedure in Zhou et al. (2020), our approach does not require weak dependence among mediators and the sub-Gaussian distribution of the errors. Therefore, our method can be applied under more general dependence conditions in the presence of interaction and milder distribution conditions.

3. From an application perspective, we apply our procedure to investigate the mediator role of DNA methylation that bridges childhood trauma experiences and the adult

stress reactivity. Compared with previous results (Houtepen et al., 2016; van Kesteren & Oberski, 2019; Guo et al., 2022), we discover three DNA methylation loci that remain undetected without considering the treatment-by-mediator interaction.

4.1.0.0.3 Related work on high-dimensional mediation analysis High-dimensional mediation analysis has evolved rapidly. We can broadly divide these methods into three groups: (1) Penalization-based methods: these methods apply different penalties to the outcome model. Examples include the nonconvex penalized least squares procedures Guo et al. (2023) and H. Zhang et al. (2016) (HIMA), the debiased Lasso procedures Zhou et al. (2020) and Gao et al. (2019) (HDMA), and other structures of Lasso-type penalties such as C. Wang et al. (2020); Q. Zhang (2021); Y. Zhao and Luo (2022); (2) Projection-based methods: these methods use the linear transformation of $G_1, \ldots, G_p$ to generate new projected mediators or latent factors. Examples include PCA-based methods Huang and Pan (2016) (PCMA) and Y. Zhao, Lindquist, and Caffo (2020) (SPCMA), and likelihood-based methods Chén et al. (2018) (HDMM) and Derkach, Pfeiffer, Chen, and Sampson (2019) (LVMA); (3) Bayesian-based methods: examples are Song et al. (2020) (BSLMM) and Song et al. (2021) (GMM). For a detailed review, see Clark-Boucher et al. (2023).

Organization. The remainder of the paper is organized as follows. Section 4.2 sets up the high-dimensional mediation model with interaction and the corresponding estimands. Section 4.3 develops the two-step procedure for statistical estimation. Section 4.4 establishes the asymptotic normality of our estimation procedure. Section 4.5 presents numerical studies. Section 4.6 discusses the overlapping group SCAD to achieve heredity in the estimator. Section 4.7 presents a real data example.

Notation. Throughout this article, for constants a and b, let $a \wedge b$ and $a \vee b$ denote the minimum and maximum of a and b, respectively. Define the indicator function $\mathcal{I}\{A\}$ to be 1 if the condition A is satisfied, and 0 otherwise. For a set $\mathcal{S}$, let $|\mathcal{S}|$ denote the number of elements in $\mathcal{S}$. For a vector $\mathbf{a} \in \mathbb{R}^p$, let $\|\mathbf{a}\|_2$ and $\|\mathbf{a}\|_\infty$ denote the ℓ_2 norm and ℓ_∞ norm, respectively, *i.e.* , $\|\mathbf{a}\|_2 = \sqrt{\mathbf{a}^{\mathrm{T}}\mathbf{a}}$ and $\|\mathbf{a}\|_\infty = \max_i |a_i|$. For an index set $\mathcal{I}$ of $\mathbf{a}$, let $\mathbf{a}_{\mathcal{I}}$ denote the sub-vector of $\mathbf{a}$ with entries in $\mathcal{I}$. Let $\mathbf{1}_n \in \mathbb{R}^n$ denote the vector whose entries are all 1's. For a matrix $\mathbf{M} \in \mathbb{R}^{n\times p}$ with the row index set $\mathcal{R} \subset \{1,\ldots,n\}$ and the column index set $\mathcal{C} \subset \{1,\ldots,p\}$, let $\mathbf{M}_{\mathcal{R}\times\mathcal{C}}$ denote the sub-matrix of $\mathbf{M}$ with rows in $\mathcal{R}$ and columns in $\mathcal{C}$. If $\mathcal{R} = \{1,\ldots,n\}$, $\mathbf{M}_{\mathcal{R}\times\mathcal{C}}$ is simplified as $\mathbf{M}_{\mathcal{C}}$. If $\mathcal{C} = \{1,\ldots,p\}$, $\mathbf{M}_{\mathcal{R}\times\mathcal{C}}$ is simplified as $\mathbf{M}_{\mathcal{R}\times[p]}$. For any symmetric matrix $\mathbf{M} \in \mathbb{R}^{p\times p}$, let $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$ denote the minimum and maximum eigenvalues of $\mathbf{M}$, respectively. Let $\otimes$ denote the Kronecker product. Let vec denote the operator which maps $\mathbf{A} \in \mathbb{R}^{m\times n}$ to $\mathbb{R}^{mn}$ by stacking its columns. Let $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. For random vectors $\mathbf{X}$ and $\mathbf{Y}$, let $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}$ denote that $\mathbf{X}$ is independent of $\mathbf{Y}$. For random vectors $\{\mathbf{X}_n\}_{n=1}^{\infty}$ and $\mathbf{X}$, let $\mathbf{X}_n \xrightarrow{L} \mathbf{X}$ and $\mathbf{X}_n \xrightarrow{P} \mathbf{X}$ denote the convergence in law and convergence in probability, respectively. We use C, C_1, C_2, $\ldots$ to denote positive and finite absolute constants that unless otherwise indicated can change from line to line.

4.2 Model

4.2.1 Causal Estimands and Identifying Assumptions

Consider a simple random sample of size n from a superpopulation. We use the vector $\mathbf{T} \in \mathbb{R}^q$ to denote a categorical treatment with $q+1$ levels. The support of $\mathbf{T}$ is given by $\mathcal{T} = \{\mathbf{t}_0, \mathbf{t}_1, \ldots, \mathbf{t}_q\}$ where $\mathbf{t}_0 = \mathbf{0}_q$ represents the control level, and $\mathbf{t}_j$ with 1 in the jth entry and 0's elsewhere represents level j for $j = 1, ..., q$. Let $\mathbf{X} \in \mathbb{R}^r$, $\mathbf{G} \in \mathbb{R}^p$, $Y \in \mathbb{R}$ be the vector of r pretreatment covariates, the vector of p mediators, and the outcome, respectively. We allow high dimensional mediators with $p > n$, but assume $\mathbf{T}$ and $\mathbf{X}$ to be low dimensional. We adopt the potential outcomes framework (Robins & Greenland, 1992; Pearl, 2001; Imai et al., 2010) to define the causal effects of interest. We use $\mathbf{G}(\mathbf{t}_j) = [G_1(\mathbf{t}_j), \ldots, G_p(\mathbf{t}_j)]^{\mathrm{T}}$ to denote the potential values that the mediators would take under treatment assignment $\mathbf{t}_j$, $j = 0, 1, ..., q$. We use $Y(\mathbf{t}_j, \mathbf{g})$ for the potential outcome that Y would take if the treatment and the mediators were set to $\mathbf{t}_j$ and $\mathbf{g}$, respectively.

We compare each treatment level j with the control level. Define the following conditional natural indirect effects (NIE): $\delta_{\mathbf{x},j}(\mathbf{t}_j) = \mathbb{E}\left\{Y(\mathbf{t}_j, \mathbf{G}(\mathbf{t}_j)) - Y(\mathbf{t}_j, \mathbf{G}(\mathbf{t}_0))|\mathbf{X} = \mathbf{x}\right\}$ and $\delta_{\mathbf{x},j}(\mathbf{t}_0) = \mathbb{E}\left\{Y(\mathbf{t}_0, \mathbf{G}(\mathbf{t}_j)) - Y(\mathbf{t}_0, \mathbf{G}(\mathbf{t}_0))|\mathbf{X} = \mathbf{x}\right\}$. The $\delta_{\mathbf{x},j}(\mathbf{t}_j)$ and $\delta_{\mathbf{x},j}(\mathbf{t}_0)$ quantify the indirect effects of treatment level j on the outcome that operates through mediators by changing mediator values from $\mathbf{G}(\mathbf{t}_0)$ to $\mathbf{G}(\mathbf{t}_j)$ while holding treatment at levels j and 0, respectively. Define the following conditional natural direct effects (NDE): $\zeta_{\mathbf{x},j}(\mathbf{t}_j) = \mathbb{E}\{Y(\mathbf{t}_j, \mathbf{G}(\mathbf{t}_j)) - Y(\mathbf{t}_0, \mathbf{G}(\mathbf{t}_j))|\mathbf{X} = \mathbf{x}\}$ and $\zeta_{\mathbf{x},j}(\mathbf{t}_0) = \mathbb{E}\{Y(\mathbf{t}_j, \mathbf{G}(\mathbf{t}_0)) - Y(\mathbf{t}_0, \mathbf{G}(\mathbf{t}_0))|\mathbf{X} = \mathbf{x}\}$. The $\zeta_{\mathbf{x},j}(\mathbf{t}_j)$ and $\zeta_{\mathbf{x},j}(\mathbf{t}_0)$ quantify the direct effects of treatment level j on the outcome not via the mediators by holding mediators at $\mathbf{G}(t_j)$ and $\mathbf{G}(t_0)$, respectively. Given the above definitions, $\delta_{\mathbf{x},j}(\mathbf{t}_j) + \zeta_{\mathbf{x},j}(\mathbf{t}_0) = \delta_{\mathbf{x},j}(\mathbf{t}_0) + \zeta_{\mathbf{x},j}(\mathbf{t}_j) = \tau_{\mathbf{x},j}$, where $\tau_{\mathbf{x},j} = \mathbb{E}\left\{Y(\mathbf{t}_j, \mathbf{G}(\mathbf{t}_j)) - Y(\mathbf{t}_0, \mathbf{G}(\mathbf{t}_0))|\mathbf{X} = \mathbf{x}\right\}$, *i.e.* , the conditional average treatment effect of level j compared to the control level. Under the sequential ignorability assumption (Imai et al., 2010) that (i) $\{Y(\mathbf{t}_{j'}, \mathbf{g}), \mathbf{G}(\mathbf{t}_j)\} \perp\!\!\!\perp \mathbf{T} \mid \mathbf{X} = \mathbf{x}$ and (ii) $Y(\mathbf{t}_{j'}, \mathbf{g}) \perp\!\!\!\perp \mathbf{G}(\mathbf{t}_j) \mid \mathbf{T} = \mathbf{t}_j, \mathbf{X} = \mathbf{x}$ hold for all $\mathbf{t}_j, \mathbf{t}_{j'} \in \mathcal{T}, \mathbf{x} \in \mathcal{X}$ and $\mathbf{g} \in \mathcal{G}$, we have the following identification result: $\mathbb{E}\{Y(\mathbf{t}_j, \mathbf{G}(\mathbf{t}_{j'})) \mid \mathbf{X} = \mathbf{x}\} = \int_{\mathcal{G}} \mathbb{E}(Y \mid \mathbf{G} = \mathbf{g}, \mathbf{T} = \mathbf{t}_j, \mathbf{X} = \mathbf{x})\, \mathrm{d}F_{\mathbf{G}}(\mathbf{g} \mid \mathbf{T} = \mathbf{t}_{j'}, \mathbf{X} = \mathbf{x})$.

4.2.2 Linear Structural Equation Model

Focusing on continuous mediators and outcome, we consider the following linear model:

$$\mathbf{Y} = \theta_0^\star \mathbf{1}_n + \mathbf{T}\boldsymbol{\theta}_T^\star + \mathbf{X}\boldsymbol{\theta}_X^\star + \mathbf{G}\boldsymbol{\theta}_G^\star + (\mathbf{T} : \mathbf{G})\mathrm{vec}(\boldsymbol{\Theta}_D^\star) + \boldsymbol{\varepsilon}_Y = \mathbf{W}\boldsymbol{\theta}_W^\star + \boldsymbol{\varepsilon}_Y, \tag{4.2.1}$$

$$\mathbf{G} = \mathbf{H}[\boldsymbol{\beta}_0\ \mathbf{B}\ \mathbf{A}]^{\mathrm{T}} + \boldsymbol{\varepsilon}_G, \tag{4.2.2}$$

where $\mathbf{Y} = [Y_1\ \ldots\ Y_n]^{\mathrm{T}} \in \mathbb{R}^n$, $\mathbf{T} = [\mathbf{T}_1\ \ldots\ \mathbf{T}_n]^{\mathrm{T}} \in \mathbb{R}^{n\times q}$, $\mathbf{X} = [\mathbf{X}_1\ \ldots\ \mathbf{X}_n]^{\mathrm{T}} \in \mathbb{R}^{n\times r}$, $\mathbf{G} = [\mathbf{G}_1\ \ldots\ \mathbf{G}_n]^{\mathrm{T}} \in \mathbb{R}^{n\times p}$, $\mathbf{T} : \mathbf{G} \in \mathbb{R}^{n\times pq}$ whose rows are $\mathbf{T}_i^{\mathrm{T}} \otimes \mathbf{G}_i^{\mathrm{T}}$, $i = 1, \ldots, n$, $\mathbf{W} = [\mathbf{W}_1\ \ldots\ \mathbf{W}_n]^{\mathrm{T}} \in \mathbb{R}^{n\times(1+q+r+p+pq)}$, $\boldsymbol{\varepsilon}_Y = [\epsilon_{Y_1}\ \ldots\ \epsilon_{Y_n}]^{\mathrm{T}} \in \mathbb{R}^n$, $\mathbf{H} = [\mathbf{1}_n\ \mathbf{T}\ \mathbf{X}] \in \mathbb{R}^{n\times(1+q+r)}$, and $\boldsymbol{\varepsilon}_G = [\boldsymbol{\varepsilon}_{G_1}\ \ldots\ \boldsymbol{\varepsilon}_{G_n}]^{\mathrm{T}} \in \mathbb{R}^{n\times p}$. The random errors ϵ_{Y_i} and ϵ_{G_i} are independent under sequential ignorability, $\mathbb{E}(\epsilon_{Y_i}) = 0$, $\mathrm{Var}(\epsilon_{Y_i}) = \sigma_Y^2$, and $\mathbb{E}(\boldsymbol{\varepsilon}_{G_i}) = \mathbf{0}_p$, $\mathrm{Cov}(\boldsymbol{\varepsilon}_{G_i}) = \boldsymbol{\Sigma}_G$.

Different from many existing works on high-dimensional mediation analysis, we allow the interactions between treatment and mediators in their effects on the outcome by including the $1 \times (pq)$ interaction term $\mathbf{T}_i^{\mathrm{T}} \otimes \mathbf{G}_i^{\mathrm{T}}$, where $\otimes$ is the Kronecker product. For example, if $\mathbf{T}_i = [1, \mathbf{0}_{q-1}^{\mathrm{T}}]^{\mathrm{T}}$, then $\mathbf{T}_i^{\mathrm{T}} \otimes \mathbf{G}_i^{\mathrm{T}} = [\mathbf{G}_i^{\mathrm{T}}, \mathbf{0}_p^{\mathrm{T}}, \ldots, \mathbf{0}_p^{\mathrm{T}}]$. Without this term, the model would invoke the assumption that $\delta_{\mathbf{x},j}(\mathbf{t}_j) = \delta_{\mathbf{x},j}(\mathbf{t}_0)$ and $\zeta_{\mathbf{x},j}(\mathbf{t}_j) = \zeta_{\mathbf{x},j}(\mathbf{t}_0)$ for $j = 1, ..., q$, which could be overly strong in practice as discussed in the introduction. In (4.2.1), $\boldsymbol{\theta}_T^\star$, $\boldsymbol{\theta}_G^\star$ and $\boldsymbol{\theta}_X^\star$ are vectors of coefficients of length q, p and r, respectively; $\boldsymbol{\Theta}_D^\star = [\boldsymbol{\theta}_{D_1}^\star, \ldots, \boldsymbol{\theta}_{D_q}^\star]$ is a p by q matrix of coefficients for interactions. The combined coefficient vector is $\boldsymbol{\theta}_W^\star = [\theta_0^\star \ \boldsymbol{\theta}_T^{\star\mathrm{T}} \ \boldsymbol{\theta}_X^{\star\mathrm{T}} \ \boldsymbol{\theta}_G^{\star\mathrm{T}} \ \boldsymbol{\theta}_{D_1}^{\star\mathrm{T}} \ \ldots \ \boldsymbol{\theta}_{D_q}^{\star\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{1+q+r+p+pq}$. We consider p to be high-dimensional, and constrain q and r to be low-dimensional. We assume throughout that $\boldsymbol{\theta}_G^\star$ and $\boldsymbol{\theta}_{D_j}^\star$, $j = 1, \ldots, q$ are sparse vectors, as in the high dimensional regime with $p > n$, the estimation error of the coefficients is intractable without some structural assumptions. In the mediator model (4.2.2), $\mathbf{B} = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q]$ and $\mathbf{A}$ are $p \times q$ and $p \times r$ coefficient matrices for $\mathbf{T}$ and $\mathbf{X}$, respectively. Further, let $\mathbb{E}(\mathbf{X}_i) = \boldsymbol{\mu}_X$ and $\mathrm{Var}(\mathbf{X}_i) = \boldsymbol{\Sigma}_X$.

Various estimands have been proposed in the literature such as the individual mediation effects $\beta_{j,h}(\theta_{G,h}^\star + \theta_{D_j,h}^\star)$ for $h \in \{1, \ldots, p\}$ (Liu et al., 2022; Dai, Stanford, & LeBlanc, 2022), or the ℓ_2 norm of individual mediation effects $\sum_{h=1}^p \beta_{j,h}^2(\theta_{G,h}^\star + \theta_{D_j,h}^\star)^2$ (Huang & Pan, 2016; C. Wang et al., 2020). Those individual effects are meaningful when mediators are parallel, *i.e.* , conditionally independent given $\mathbf{T}$ and $\mathbf{X}$. When we are away from the case of parallel mediators, the contributions of each mediator can not be separated without further assumptions. In our work, we evaluate the contribution of the mediators as a group and focus on the NIEs and NDEs defined in section 4.2.1. Under models (4.2.1) and (4.2.2), for $j = 1, \ldots, q$, we have

$$\delta_{\mathbf{x},j}(\mathbf{t}_j) = \boldsymbol{\beta}_j^{\mathrm{T}} \left(\boldsymbol{\theta}_G^\star + \boldsymbol{\theta}_{D_j}^\star\right), \ \delta_{\mathbf{x},j}(\mathbf{t}_0) = \boldsymbol{\beta}_j^{\mathrm{T}} \boldsymbol{\theta}_G^\star, \tag{4.2.3}$$

$$\zeta_{\mathbf{x},j}(\mathbf{t}_j) = \theta_{T,j}^\star + (\boldsymbol{\beta}_0 + \boldsymbol{\beta}_j + \mathbf{A}\mathbf{x})^{\mathrm{T}} \boldsymbol{\theta}_{D_j}^\star, \ \zeta_{\mathbf{x},j}(\mathbf{t}_0) = \theta_{T,j}^\star + (\boldsymbol{\beta}_0 + \mathbf{A}\mathbf{x})^{\mathrm{T}} \boldsymbol{\theta}_{D_j}^\star. \tag{4.2.4}$$

4.3 Estimation of the NIE and NDE

As introduced in Section 4.1, the difference method is not feasible in the presence of the treatment-by-mediator interaction. We propose a two-step estimation procedure designed to produce tractable errors in (4.1.1) by the product method. We first estimate $\boldsymbol{\theta}_W^\star$ in the outcome model (4.2.1) under ℓ_2 loss with non-convex penalization to identify the support of $\boldsymbol{\theta}_W^\star$. Next we fit the mediator model (4.2.2) using the seemingly unrelated regressions (SUR) (Zellner, 1962), or equivalently, the ordinary least squares (OLS) to obtain estimates of $\boldsymbol{\beta}_0$, $\mathbf{B}$, and $\mathbf{A}$ which are then projected onto the identified low-dimensional directions according to the nonzero locations in $\widehat{\boldsymbol{\theta}}_W^\star$ from the first step, thereby obtaining the estimates of $\boldsymbol{\beta}_j^{\mathrm{T}}\boldsymbol{\theta}_G^\star$, $\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{\theta}_{D_j}^\star$, $\boldsymbol{\beta}_j^{\mathrm{T}}\boldsymbol{\theta}_{D_j}^\star$, $\mathbf{A}^{\mathrm{T}}\boldsymbol{\theta}_{D_j}^\star$ with tractable error accumulation. Specifically:

Step 1: Estimate $\boldsymbol{\theta}_W^\star$ in the outcome model (4.2.1) by applying the partially penalized least squares method:

$$\widehat{\boldsymbol{\theta}}_W^\star = \underset{\boldsymbol{\theta}_W \in \mathbb{R}^{1+q+r+p+pq}}{\arg\min} \frac{1}{2n}\|\mathbf{Y} - \mathbf{W}\boldsymbol{\theta}_W\|_2^2 + \sum_{h=1}^{p} p_{\lambda_n}(|\theta_{G,h}|) + \sum_{j=1}^{q}\sum_{h=1}^{p} p_{\lambda_n}(|\theta_{D_j,h}|), \tag{4.3.1}$$

where $\widehat{\boldsymbol{\theta}}_W^\star = [\widehat{\theta}_0^\star, \widehat{\boldsymbol{\theta}}_T^{\star\mathrm{T}}, \widehat{\boldsymbol{\theta}}_X^{\star\mathrm{T}}, \widehat{\boldsymbol{\theta}}_G^{\star\mathrm{T}}, \widehat{\boldsymbol{\theta}}_{D_1}^{\star\mathrm{T}}, \ldots, \widehat{\boldsymbol{\theta}}_{D_q}^{\star\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{1+q+r+p+pq}$, and $p_{\lambda_n}(\cdot)$ is the SCAD or MCP penalty. Leveraging the oracle property of $p_{\lambda_n}(\cdot)$ as stated in Theorem 4.4.1, *Step 1* reduces the high-dimensional problem of $\boldsymbol{\theta}_W^\star$ to its low-dimensional support.

Step 2: Estimate $[\boldsymbol{\beta}_0\ \mathbf{B}\ \mathbf{A}]$ in the mediator model (4.2.2) by applying the SUR or OLS:

$$[\widehat{\boldsymbol{\beta}}_0\ \widehat{\mathbf{B}}\ \widehat{\mathbf{A}}]^{\mathrm{T}} = (\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{G} \in \mathbb{R}^{(1+q+r)\times p}. \tag{4.3.2}$$

As a result, the estimators for $\boldsymbol{\beta}_j^{\mathrm{T}}\boldsymbol{\theta}_G^\star$, $\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{\theta}_{D_j}^\star$, $\boldsymbol{\beta}_j^{\mathrm{T}}\boldsymbol{\theta}_{D_j}^\star$, and $\mathbf{A}^{\mathrm{T}}\boldsymbol{\theta}_{D_j}^\star$, $j = 1, \ldots, q$, are

$$\begin{aligned} \widehat{\boldsymbol{\beta}_j^{\mathrm{T}}\boldsymbol{\theta}_G^\star} &= \mathbf{e}_{j+1}^{\mathrm{T}}[\widehat{\boldsymbol{\beta}}_0\ \widehat{\mathbf{B}}\ \widehat{\mathbf{A}}]^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_G^\star,\ \widehat{\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{\theta}_{D_j}^\star} = \mathbf{e}_1^{\mathrm{T}}[\widehat{\boldsymbol{\beta}}_0\ \widehat{\mathbf{B}}\ \widehat{\mathbf{A}}]^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_{D_j}^\star, \\ \widehat{\boldsymbol{\beta}_j^{\mathrm{T}}\boldsymbol{\theta}_{D_j}^\star} &= \mathbf{e}_{j+1}^{\mathrm{T}}[\widehat{\boldsymbol{\beta}}_0\ \widehat{\mathbf{B}}\ \widehat{\mathbf{A}}]^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_{D_j}^\star,\ \widehat{\mathbf{A}^{\mathrm{T}}\boldsymbol{\theta}_{D_j}^\star} = \mathbf{I}_X^{\mathrm{T}}[\widehat{\boldsymbol{\beta}}_0\ \widehat{\mathbf{B}}\ \widehat{\mathbf{A}}]^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_{D_j}^\star, \end{aligned} \tag{4.3.3}$$

where $\mathbf{e}_k \in \mathbb{R}^{1+q+r}$ is the kth standard basis in $\mathbb{R}^{1+q+r}$ for $k = 1, \ldots, 1+q+r$, and the columns in $\mathbf{I}_X \in \mathbb{R}^{(1+q+r)\times r}$ are $\mathbf{e}^{\mathrm{T}}_{1+q+1}, \ldots, \mathbf{e}^{\mathrm{T}}_{1+q+r}$. The estimators of $\delta_{\mathbf{x},j}(\mathbf{t}_j)$, $\delta_{\mathbf{x},j}(\mathbf{t}_0)$, $\zeta_{\mathbf{x},j}(\mathbf{t}_j)$, and $\zeta_{\mathbf{x},j}(\mathbf{t}_0)$ are subsequently obtained according to the identification results (4.2.3) and (4.2.4).

Although $\mathbf{G}$ is high dimensional, and the same is true for $\widehat{\boldsymbol{\beta}}_0$, $\widehat{\mathbf{B}}$, and $\widehat{\mathbf{A}}$, (4.3.3) produces tractable estimation errors because of the sparsity imposed on $\boldsymbol{\theta}^{\star}_G$ and $\boldsymbol{\theta}^{\star}_{D_j}$, $j = 1, \ldots, q$. As long as $\widehat{\boldsymbol{\theta}}^{\star}_G$ and $\widehat{\boldsymbol{\theta}}^{\star}_{D_j}$ can correctly identify the supports of $\boldsymbol{\theta}^{\star}_G$ and $\boldsymbol{\theta}^{\star}_{D_j}$ with overwhelming probability, $[\widehat{\boldsymbol{\beta}}_0\ \widehat{\mathbf{B}}\ \widehat{\mathbf{A}}]^{\mathrm{T}}\widehat{\boldsymbol{\theta}}^{\star}_G$ and $[\widehat{\boldsymbol{\beta}}_0\ \widehat{\mathbf{B}}\ \widehat{\mathbf{A}}]^{\mathrm{T}}\widehat{\boldsymbol{\theta}}^{\star}_{D_j}$ will simply be projections onto low-dimensional directions according to the nonzero locations in $\widehat{\boldsymbol{\theta}}^{\star}_G$ and $\widehat{\boldsymbol{\theta}}^{\star}_{D_j}$. Such correct identification of the supports of $\boldsymbol{\theta}^{\star}_G$ and $\boldsymbol{\theta}^{\star}_{D_j}$ is guaranteed by the oracle property in (4.3.1), and the projection in (4.3.3) actually constrains the high-dimensional problem to the low-dimensional supports of $\boldsymbol{\theta}^{\star}_G$ and $\boldsymbol{\theta}^{\star}_{D_j}$.

4.4 Inference of the NIE and NDE

This section establishes the theoretical guarantees of our two-step approach. Section 4.4.1 presents the oracle property of $\widehat{\boldsymbol{\theta}}^{\star}_W$ from the partially penalized least squares (4.3.1) in *Step 1*. Section 4.4.2 presents the asymptotic normality of the projection estimators $\widehat{\delta}_{\mathbf{x},j}(\mathbf{t}_j)$, $\widehat{\delta}_{\mathbf{x},j}(\mathbf{t}_0)$, $\widehat{\zeta}_{\mathbf{x},j}(\mathbf{t}_j)$, and $\widehat{\zeta}_{\mathbf{x},j}(\mathbf{t}_0)$ in *Step 2*.

4.4.1 Oracle Property

To introduce the oracle property of $\widehat{\boldsymbol{\theta}}^{\star}_W = [\widehat{\theta}^{\star}_0, \widehat{\boldsymbol{\theta}}^{\star\mathrm{T}}_T, \widehat{\boldsymbol{\theta}}^{\star\mathrm{T}}_X, \widehat{\boldsymbol{\theta}}^{\star\mathrm{T}}_G, \widehat{\boldsymbol{\theta}}^{\star\mathrm{T}}_{D_1}, \ldots, \widehat{\boldsymbol{\theta}}^{\star\mathrm{T}}_{D_q}]^{\mathrm{T}}$ from (4.3.1), we first present some notation for the identification of the support. Let $\mathcal{A}_G = \{h \in \{1, \ldots, p\} : \theta^{\star}_{G,h} \neq 0\}$ and $\mathcal{A}_{D_j} = \{h \in \{1, \ldots, p\} : \theta^{\star}_{D_j,h} \neq 0\}$ with cardinalities $|\mathcal{A}_G| = s_G$ and $|\mathcal{A}_{D_j}| = s_{D_j}$ for $j = 1, \ldots, q$. Let $s = s_G + s_{D_1} + \cdots + s_{D_q}$. Then we partition the indices of $\boldsymbol{\theta}^{\star}_W$ into $\mathcal{A}$ and $\mathcal{A}^c$. Specifically, $\mathcal{A} = \{0\} \cup \{(T,1), \ldots, (T,q)\} \cup \{(X,1), \ldots, (X,r)\} \cup \{(G,h) : h \in \mathcal{A}_G\} \cup \bigcup_{j=1}^{q}\{(D_j,h) : h \in \mathcal{A}_{D_j}\}$ which consists of the indices of $\theta^{\star}_0$, $\boldsymbol{\theta}^{\star}_T$, $\boldsymbol{\theta}^{\star}_X$, and the indices of the supports of $\boldsymbol{\theta}^{\star}_G$ and $\boldsymbol{\theta}^{\star}_{D_j}$, $j = 1, \ldots, q$. $|\mathcal{A}| = 1 + q + r + s$. Also define $\mathcal{A}^c = \{(G,h) : h \notin \mathcal{A}_G\} \cup \bigcup_{j=1}^{q}\{(D_j,h) : h \notin \mathcal{A}_{D_j}\}$ which consists of the indices of $\boldsymbol{\theta}^{\star}_G$ and $\boldsymbol{\theta}^{\star}_{D_j}$, $j = 1, \ldots, q$

that are not in the supports. $|\mathcal{A}_G^c| = (1+q)p - s$. Then $\widehat{\boldsymbol{\theta}}_W^\star$ is partitioned into sub-vector $\widehat{\boldsymbol{\theta}}_{W,\mathcal{A}}^\star = [\widehat{\theta}_0^\star, \widehat{\boldsymbol{\theta}}_T^{\star\mathrm{T}}, \widehat{\boldsymbol{\theta}}_X^{\star\mathrm{T}}, \widehat{\boldsymbol{\theta}}_{G,\mathcal{A}_G}^{\star\mathrm{T}}, \widehat{\boldsymbol{\theta}}_{D_1,\mathcal{A}_{D_1}}^{\star\mathrm{T}}, \ldots, \widehat{\boldsymbol{\theta}}_{D_q,\mathcal{A}_{D_q}}^{\star\mathrm{T}}]^\mathrm{T} \in \mathbb{R}^{1+q+r+s}$ whose true value is $\boldsymbol{\theta}_{W,\mathcal{A}}^\star = [\theta_0^\star,$ $\boldsymbol{\theta}_T^{\star\mathrm{T}}$, $\boldsymbol{\theta}_X^{\star\mathrm{T}}$, $\boldsymbol{\theta}_{G,\mathcal{A}_G}^{\star\mathrm{T}}$, $\boldsymbol{\theta}_{D_1,\mathcal{A}_{D_1}}^{\star\mathrm{T}}, \ldots, \boldsymbol{\theta}_{D_q,\mathcal{A}_{D_q}}^{\star\mathrm{T}}]^\mathrm{T}$, and sub-vector $\widehat{\boldsymbol{\theta}}_{W,\mathcal{A}^c}^\star \in \mathbb{R}^{(q+1)p-s}$ whose true value is $\mathbf{0}_{(q+1)p-s}$. Correspondingly, $\mathbf{W} \in \mathbb{R}^{1+q+r+p+pq}$ is partitioned into $\mathbf{W}_\mathcal{A} \in \mathbb{R}^{n\times(1+q+r+s)}$ and $\mathbf{W}_{\mathcal{A}^c} \in \mathbb{R}^{n\times[(1+q)p-s]}$, consisting of the columns in $\mathcal{A}$ and $\mathcal{A}^c$, respectively. Since the rows in $\mathbf{W}_\mathcal{A}$ are independent, we have $n^{-1}(\mathbf{W}_\mathcal{A}^\mathrm{T}\mathbf{W}_\mathcal{A}) \xrightarrow{P} \boldsymbol{\Sigma} \in \mathbb{R}^{(1+q+r+s)\times(1+q+r+s)}$, where the explicit form of $\boldsymbol{\Sigma}$ is deferred to (A.1.2) in the supplementary material. Define a measure of the minimal signal as $d_n = (\min_{h\in\mathcal{A}_G}\{|\theta_{G,h}^\star|\} \wedge \min_{j=1,\ldots,q,h\in\mathcal{A}_{D_j}}\{|\theta_{D_j,h}^\star|\})/2$.

We impose standard assumptions on the penalty function, the treatment assignment, the relationship between $\mathbf{W}_\mathcal{A}$ and $\mathbf{W}_{\mathcal{A}^c}$, the minimal signal, and the moment conditions (Fan & Li, 2001; Guo et al., 2023). The description is deferred to Conditions A.1.1–A.1.5 in the supplementary material.

Theorem 4.4.1. *Under Conditions A.1.1–A.1.5, if $s = o(n^{1/2})$, $r \le C \cdot s$ for an absolute constant $C > 0$, and $p'_{\lambda_n}(d_n) = o((ns)^{-1/2})$, then with probability tending to one, $\widehat{\boldsymbol{\theta}}_{W,\mathcal{A}^c}^\star = \mathbf{0}_{(1+q)p-s}$ and $\|\widehat{\boldsymbol{\theta}}_{W,\mathcal{A}}^\star - \boldsymbol{\theta}_{W,\mathcal{A}}^\star\|_2 = O_p(\sqrt{s/n})$. Additionally, if $s = o(n^{1/3})$, then $\sqrt{n}(\widehat{\boldsymbol{\theta}}_{W,\mathcal{A}}^\star - \boldsymbol{\theta}_{W,\mathcal{A}}^\star) = (1/\sqrt{n})\boldsymbol{\Sigma}^{-1}\mathbf{W}_\mathcal{A}^\mathrm{T}\boldsymbol{\varepsilon}_Y + o_p(\mathbf{1})$, where the vector $o_p(\mathbf{1}) \in \mathbb{R}^{1+q+r+s}$ means that $\|o_p(\mathbf{1})\|_2 = o_p(1)$ and $\boldsymbol{\Sigma}$ is defined by* (A.1.2) *in the supplementary material.*

Remark 4.4.1. *Denote $\widehat{\theta}_{\mathcal{A},k}^\star$ and $\theta_{\mathcal{A},k}^\star$ the kth entry in $\widehat{\boldsymbol{\theta}}_{W,\mathcal{A}}^\star$ and $\boldsymbol{\theta}_{W,\mathcal{A}}^\star$ for $k = 1, \ldots, (1 + q + r + s)$, respectively. Under Conditions A.1.1–A.1.5, if $s = o(n^{1/3})$, $r \le C \cdot s$, and $p'_{\lambda_n}(d_n) = o((ns)^{-1/2})$, then $\sqrt{n}(\widehat{\theta}_{\mathcal{A},k}^\star - \theta_{\mathcal{A},k}^\star) \xrightarrow{L} N(0, \sigma_Y^2\tilde{\mathbf{e}}_k^\mathrm{T}\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{e}}_k)$, where $\tilde{\mathbf{e}}_k$ is the kth standard basis in $\mathbb{R}^{1+q+r+s}$.*

Under standard conditions, even with high-dimensional interactions, Theorem 4.4.1 recovers the oracle property of $\widehat{\boldsymbol{\theta}}_W^\star$ as established in Guo et al. (2023). That is, the estimator $\widehat{\boldsymbol{\theta}}_W^\star$ performs as well as the oracle estimator with $\boldsymbol{\theta}_{W,\mathcal{A}^c}^\star$ known to be $\mathbf{0}_{(1+q)p-s}$, in terms of convergence rate and asymptotic normality of $\widehat{\boldsymbol{\theta}}_{W,\mathcal{A}}^\star$. The overwhelming probability of $\widehat{\boldsymbol{\theta}}_{W,\mathcal{A}^c}^\star = \mathbf{0}_{(1+q)p-s}$ and the low-dimensional $\widehat{\boldsymbol{\theta}}_{W,\mathcal{A}}^\star$ preclude the error accumulation described in (4.1.1). This is the basis for the projection procedure in *Step 2* of our approach.

4.4.2 The Asymptotic Normality of Projection

This section establishes the asymptotic normality of the estimated NIEs, *i.e.* , $\widehat{\delta}_{\mathbf{x},j}(\mathbf{t}_j)$ and $\widehat{\delta}_{\mathbf{x},j}(\mathbf{t}_0)$, and the estimated NDEs, *i.e.* , $\widehat{\zeta}_{\mathbf{x},j}(\mathbf{t}_j)$ and $\widehat{\zeta}_{\mathbf{x},j}(\mathbf{t}_0)$, for $j = 1, \ldots, q$, from the projection procedure (4.3.3) in *Step 2*.

Define $\tilde{\mathbf{I}}_{\mathcal{A}_G} \in \mathbb{R}^{(1+q+r+s)\times s_G}$ and $\tilde{\mathbf{I}}_{\mathcal{A}_{D_j}} \in \mathbb{R}^{(1+q+r+s)\times s_{D_j}}$, $j = 1, \ldots, q$, whose columns are $\tilde{\mathbf{e}}_k$'s where k's correspond to the locations of $\{(G, h) : h \in \mathcal{A}_G\}$ and $\{(D_j, h) : \mathcal{A}_{D_j}\}$, $j = 1, \ldots, q$, in $\mathcal{A}$, respectively.

Theorem 4.4.2. *Under Conditions A.1.1–A.1.5, suppose that* $p'_{\lambda_n}(d_n) = o(1/[\sqrt{n}s])$, $s = o(n^{1/4})$, $(\max_{h\in\mathcal{A}_G}\{|\theta^\star_{G,h}|\} \vee \max_{j=1,\ldots,q,h\in\mathcal{A}_{D_j}}\{|\theta^\star_{D_j,h}|\}) \leq C_\theta$ *for an absolute constant* C_θ, *and* r *is fixed. Additionally, in* (4.2.4), *assume that the entries in* $\mathbf{x} \in \mathbb{R}^r$ *are bounded and* $\lambda_{\min}(\mathbf{\Sigma}_X) \geq c_X > 0$. *Then for treatment level* $j = 1, \ldots, q$, *we have*

$$\sqrt{n}\left[\widehat{\delta}_{\mathbf{x},j}(\mathbf{t}_0) - \delta_{\mathbf{x},j}(\mathbf{t}_0)\right] \xrightarrow{L} N\left(0, \mathbf{V}_{\delta_{\mathbf{x},j}(\mathbf{t}_0)}\right), \quad \sqrt{n}\left[\widehat{\delta}_{\mathbf{x},j}(\mathbf{t}_j) - \delta_{\mathbf{x},j}(\mathbf{t}_j)\right] \xrightarrow{L} N\left(0, \mathbf{V}_{\delta_{\mathbf{x},j}(\mathbf{t}_j)}\right),$$
$$\sqrt{n}\left[\widehat{\zeta}_{\mathbf{x},j}(\mathbf{t}_0) - \zeta_{\mathbf{x},j}(\mathbf{t}_0)\right] \xrightarrow{L} N\left(0, \mathbf{V}_{\zeta_{\mathbf{x},j}(\mathbf{t}_0)}\right), \quad \sqrt{n}\left[\widehat{\zeta}_{\mathbf{x},j}(\mathbf{t}_j) - \zeta_{\mathbf{x},j}(\mathbf{t}_j)\right] \xrightarrow{L} N\left(0, \mathbf{V}_{\zeta_{\mathbf{x},j}(\mathbf{t}_j)}\right),$$

where

$$\begin{aligned}
\mathbf{V}_{\delta_{\mathbf{x},j}(\mathbf{t}_0)} &= \sigma_Y^2 \boldsymbol{\beta}^{\mathrm{T}}_{j,\mathcal{A}_G}\left(\mathbf{\Sigma}^{-1}\right)_{\mathcal{A}_G\times\mathcal{A}_G}\boldsymbol{\beta}_{j,\mathcal{A}_G} + \mathbf{e}^{\mathrm{T}}_{j+1}\mathbf{J}^{-1}\mathbf{e}_{j+1}\boldsymbol{\theta}^{\star T}_{G,\mathcal{A}_G}\mathbf{\Sigma}_{G(\mathcal{A}_G\times\mathcal{A}_G)}\boldsymbol{\theta}^\star_{G,\mathcal{A}_G}, \\
\mathbf{V}_{\delta_{\mathbf{x},j}(\mathbf{t}_j)} &= \sigma_Y^2\left(\boldsymbol{\beta}^{\mathrm{T}}_{j,\mathcal{A}_G}\tilde{\mathbf{I}}^{\mathrm{T}}_{\mathcal{A}_G} + \boldsymbol{\beta}^{\mathrm{T}}_{j,\mathcal{A}_{D_j}}\tilde{\mathbf{I}}^{\mathrm{T}}_{\mathcal{A}_{D_j}}\right)\mathbf{\Sigma}^{-1}\left(\tilde{\mathbf{I}}_{\mathcal{A}_G}\boldsymbol{\beta}_{j,\mathcal{A}_G} + \tilde{\mathbf{I}}_{\mathcal{A}_{D_j}}\boldsymbol{\beta}_{j,\mathcal{A}_{D_j}}\right) + \\
&\quad \mathbf{e}^{\mathrm{T}}_{j+1}\mathbf{J}^{-1}\mathbf{e}_{j+1}\begin{bmatrix}\boldsymbol{\theta}^{\star T}_{G,\mathcal{A}_G} & \boldsymbol{\theta}^{\star T}_{D_j,\mathcal{A}_{D_j}}\end{bmatrix}\begin{bmatrix}\mathbf{\Sigma}_{G(\mathcal{A}_G\times\mathcal{A}_G)} & \mathbf{\Sigma}_{G(\mathcal{A}_G\times\mathcal{A}_{D_j})} \\ \mathbf{\Sigma}_{G(\mathcal{A}_{D_j}\times\mathcal{A}_G)} & \mathbf{\Sigma}_{G(\mathcal{A}_{D_j}\times\mathcal{A}_{D_j})}\end{bmatrix}\begin{bmatrix}\boldsymbol{\theta}^\star_{G,\mathcal{A}_G} \\ \boldsymbol{\theta}^\star_{D_j,\mathcal{A}_{D_j}}\end{bmatrix},
\end{aligned}$$

$$\begin{aligned}
\mathbf{V}_{\zeta_{\mathbf{x},j}(\mathbf{t}_0)} &= \sigma_Y^2\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\boldsymbol{\alpha} + \left(\mathbf{e}^{\mathrm{T}}_1 + \mathbf{x}^{\mathrm{T}}\mathbf{I}^{\mathrm{T}}_X\right)\mathbf{J}^{-1}\left(\mathbf{e}_1 + \mathbf{I}_X\mathbf{x}\right)\boldsymbol{\theta}^{\star\mathrm{T}}_{D_j,\mathcal{A}_{D_j}}\mathbf{\Sigma}_{G(\mathcal{A}_{D_j}\times\mathcal{A}_{D_j})}\boldsymbol{\theta}^\star_{D_j,\mathcal{A}_{D_j}}, \\
\mathbf{V}_{\zeta_{\mathbf{x},j}(\mathbf{t}_j)} &= \sigma_Y^2\left(\boldsymbol{\alpha}^{\mathrm{T}} + \boldsymbol{\beta}^{\mathrm{T}}_{j,\mathcal{A}_{D_j}}\tilde{\mathbf{I}}^{\mathrm{T}}_{\mathcal{A}_{D_j}}\right)\mathbf{\Sigma}^{-1}\left(\boldsymbol{\alpha} + \tilde{\mathbf{I}}_{\mathcal{A}_{D_j}}\boldsymbol{\beta}_{j,\mathcal{A}_{D_j}}\right) + \\
&\quad \left(\mathbf{e}^{\mathrm{T}}_1 + \mathbf{x}^{\mathrm{T}}\mathbf{I}^{\mathrm{T}}_X + \mathbf{e}^{\mathrm{T}}_{j+1}\right)\mathbf{J}^{-1}\left(\mathbf{e}_1 + \mathbf{I}_X\mathbf{x} + \mathbf{e}_{j+1}\right)\boldsymbol{\theta}^{\star\mathrm{T}}_{D_j,\mathcal{A}_{D_j}}\mathbf{\Sigma}_{G(\mathcal{A}_{D_j}\times\mathcal{A}_{D_j})}\boldsymbol{\theta}^\star_{D_j,\mathcal{A}_{D_j}},
\end{aligned}$$

with $\boldsymbol{\alpha} = \tilde{\mathbf{e}}_{j+1} + \tilde{\mathbf{I}}_{\mathcal{A}_{D_j}} \boldsymbol{\beta}_{0,\mathcal{A}_{D_j}} + \tilde{\mathbf{I}}_{\mathcal{A}_{D_j}} \mathbf{A}_{\mathcal{A}_{D_j} \times [r]} \mathbf{x} \in \mathbb{R}^{1+q+r+s}$ *and* $\mathbf{J} \in \mathbb{R}^{(1+q+r)\times(1+q+r)}$ *being the upper-left* $(1+q+r) \times (1+q+r)$ *sub-matrix of* $\boldsymbol{\Sigma}$*.*

Compared to the conditions $s = o(n^{1/3})$ and $p'_{\lambda_n}(d_n) = o(1/\sqrt{sn})$ outlined in Theorem 4.4.1 and referenced in Guo et al. (2023), Theorem 4.4.2 imposes more stringent requirements, namely, $s = o(n^{1/4})$, $p'_{\lambda_n}(d_n) = o(1/[\sqrt{n}s])$, and an upper bound C_θ for the signals. This is due to the need to handle the interactions in the product method (4.3.3). For the difference method in Guo et al. (2023), the estimation of the low-dimensional total effect follows the standard ordinary least squares, which does not require additional assumptions regarding sparsity or signal strength. So the conditions for the asymptotic normality in Guo et al. (2023) are directly inherited from those for achieving oracle property as outlined in Theorem 4.4.1. However, in the product method with the error accumulation (4.1.1), to control the errors in $\widehat{\boldsymbol{\beta}}_j \boldsymbol{e}_{\theta_G}$ with $j = 1, \ldots, q$, Cauchy–Schwarz inequality gives $|\widehat{\boldsymbol{\beta}}_j \boldsymbol{e}_{\theta_G}| \leq \|\widehat{\boldsymbol{\beta}}_{j,\mathcal{A}_G}\|_2 \|\boldsymbol{e}_{\theta_G}\|_2$. Although the errors in $\|\boldsymbol{e}_{\theta_G}\|_2$ are tractable with $s = o(n^{1/3})$ and $p'_{\lambda_n}(d_n) = o(1/\sqrt{sn})$, bounding $\|\widehat{\boldsymbol{\beta}}_{j,\mathcal{A}_G}\|_2$ results in an additional factor $\sqrt{s}$. Similarly, to control the errors in $\boldsymbol{\theta}_G^\star \boldsymbol{e}_{\beta_j}$ from (4.1.1), Cauchy–Schwarz inequality also produces $C_\theta \sqrt{s}$ to bound $\|\boldsymbol{\theta}_G^\star\|_2$. These stronger assumptions are the price for handling the interaction.

Compared with the modified debiased procedure in Zhou et al. (2020), which makes no assumptions on the minimal signal of $\boldsymbol{\theta}_W^\star$, we assume that the minimal signal satisfies $d_n \gg \lambda_n \gg \max\{\sqrt{s/n}, \sqrt{\log p/n}\}$ in Condition A.1.4 as the price for identifying the active mediators. This is a standard condition to obtain the oracle property (Fan & Li, 2001). In contrast, Zhou et al. (2020) utilizes the modified debiased procedure to correct the bias of all mediators, including the inactive ones, which fails to maintain the sparsity and the identification of the active mediators. Moreover, as mentioned in Section 4.1, the Assumption 2 and 3 in Zhou et al. (2020) require weak dependence within $[\mathbf{G}_i^{\mathrm{T}}, \mathbf{T}_i^{\mathrm{T}} \otimes \mathbf{G}_i^{\mathrm{T}}]^{\mathrm{T}}$. Such assumption imposes sparsity restrictions on $\boldsymbol{\beta}_0$, $\mathbf{B}$, $\boldsymbol{\Sigma}_G$, and is impractical to meet given the correlation between $\mathbf{G}_i$ and $\mathbf{T}_i^{\mathrm{T}} \otimes \mathbf{G}_i^{\mathrm{T}}$. The sub-Gaussian assumptions on $\boldsymbol{\varepsilon}_{G_i}$ and ϵ_{Y_i} in Zhou et al. (2020) further limit its practical applicability compared with our moment assumptions for G_{ij} in Condition A.1.5.

To apply our method, $\mathbf{V}_{\delta_{\mathbf{x},j}(\mathbf{t}_0)}$, $\mathbf{V}_{\delta_{\mathbf{x},j}(\mathbf{t}_1)}$, $\mathbf{V}_{\zeta_{\mathbf{x},j}(\mathbf{t}_0)}$, and $\mathbf{V}_{\zeta_{\mathbf{x},j}(\mathbf{t}_0)}$ need to be consistently estimated. We first identify $\widehat{\mathcal{A}}_G$ and $\widehat{\mathcal{A}}_{D_j}$ for $j = 1, \dots, q$ from the indices of the nonzero components in $\widehat{\boldsymbol{\theta}}^{\star}_G$ and $\widehat{\boldsymbol{\theta}}^{\star}_{D_j}$, $j = 1, \dots, q$. By the oracle property of $\widehat{\boldsymbol{\theta}}^{\star}_W$ in (4.3.1), with probability tending to one, we have $\widehat{\mathcal{A}}_G = \mathcal{A}_G$ and $\widehat{\mathcal{A}}_{D_j} = \mathcal{A}_{D_j}$ for $j = 1, \dots, q$. Hence σ^2_Y, $\boldsymbol{\theta}^{\star}_{G,\mathcal{A}_G}$, and $\boldsymbol{\theta}^{\star}_{D_j,\mathcal{A}_{D_j}}$ can be consistently estimated by $\widehat{\sigma}^2_Y = n^{-1}(\mathbf{Y} - \mathbf{W}\widehat{\boldsymbol{\theta}}^{\star}_W)^{\mathrm{T}}(\mathbf{Y} - \mathbf{W}\widehat{\boldsymbol{\theta}}^{\star}_W)$, $\widehat{\boldsymbol{\theta}}^{\star}_{G,\widehat{\mathcal{A}}_G}$, and $\widehat{\boldsymbol{\theta}}^{\star}_{D_j,\widehat{\mathcal{A}}_{D_j}}$ for $j = 1, \dots, q$, respectively. Also, $\mathbf{J}$ and $\boldsymbol{\Sigma}$ can be consistently estimated by $\widehat{\mathbf{J}} = n^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{H}$ with $\mathbf{H}$ from (4.2.2) and $\widehat{\boldsymbol{\Sigma}} = n^{-1}\mathbf{W}^{\mathrm{T}}_{\widehat{\mathcal{A}}}\mathbf{W}_{\widehat{\mathcal{A}}}$. Similarly, for index sets $\mathcal{I}$, $\mathcal{R}$, and $\mathcal{C}$ which can take $\mathcal{A}_G$ or $\mathcal{A}_{D_j}$ for $j = 1, \dots, q$, $\boldsymbol{\beta}_{0,\mathcal{I}}$, $\boldsymbol{\beta}_{j,\mathcal{I}}$, $\mathbf{A}_{\mathcal{R}\times[r]}$, and $\boldsymbol{\Sigma}_{G(\mathcal{R}\times\mathcal{C})}$ can be consistently estimated by $\widehat{\boldsymbol{\beta}}_{0,\widehat{\mathcal{I}}}$, $\widehat{\boldsymbol{\beta}}_{j,\widehat{\mathcal{I}}}$, $\widehat{\mathbf{A}}_{\widehat{\mathcal{R}}\times[r]}$, and $\widehat{\boldsymbol{\Sigma}}_{G(\widehat{\mathcal{R}}\times\widehat{\mathcal{C}})}$, where $\widehat{\boldsymbol{\beta}}_0$, $\widehat{\mathbf{B}} = [\widehat{\boldsymbol{\beta}}_1 \ \dots \ \widehat{\boldsymbol{\beta}}_q]$, $\widehat{\mathbf{A}}$ are from (4.3.2), and $\widehat{\boldsymbol{\Sigma}}_G = n^{-1}\{\mathbf{G} - \mathbf{H}[\widehat{\boldsymbol{\beta}}_0 \ \widehat{\mathbf{B}} \ \widehat{\mathbf{A}}]^{\mathrm{T}}\}^{\mathrm{T}}\{\mathbf{G} - \mathbf{H}[\widehat{\boldsymbol{\beta}}_0 \ \widehat{\mathbf{B}} \ \widehat{\mathbf{A}}]^{\mathrm{T}}\}$.

4.5 Simulation Studies

In Section 4.5.1, we consider a simple setup with moderate p and compare the finite sample performance of the proposed approach with competing methods in varied settings. We then consider a simulation study mimicking the real data structure of our application study with large p in Section 4.5.2.

4.5.1 Simulations with Moderate p

We simulate data according to models (4.2.1) and (4.2.2) with $q = 1$, $n = 300$, and $p = 500$ assuming no confounders for simplicity. We consider a completely randomized design with $T_i = 1$ for $i = 1, ..., n/2$ and $T_i = 0$ for the rest of the subjects. In the mediator model (4.2.2), let $\boldsymbol{\beta}_0 = c_M\boldsymbol{\gamma}_0$ and $\boldsymbol{\beta}_1 = c_M\boldsymbol{\gamma}_1$, where $c_M \in \mathbb{R}$ and $\boldsymbol{\gamma}_0, \boldsymbol{\gamma}_1 \in \mathbb{R}^p$ will be specified later. We generate $\boldsymbol{\varepsilon}_{G_i} \overset{iid}{\sim} \mathcal{N}(\mathbf{0}_p, \boldsymbol{\Sigma}_G)$ where $\boldsymbol{\Sigma}_G$ has diagonal entries 1 and off-diagonal entries 0.5. In the outcome model (4.2.1), we set $\theta^{\star}_0 = 0.5$, $\boldsymbol{\theta}^{\star\mathrm{T}}_G = [1, 0.9, 0.8, 0.7, 0.6, 1, 0.9, 0.8, 0.7, 0.6, \mathbf{0}^{\mathrm{T}}_{p-10}]$, $\boldsymbol{\theta}^{\star}_D = c_D\boldsymbol{\theta}^{\star}_G$. Since $q = 1$, $\boldsymbol{\Theta}^{\star}_D = \boldsymbol{\theta}^{\star}_D$. We generate $\epsilon_{Y_i} \overset{iid}{\sim} N(0, \sigma^2_Y)$ with $\sigma_Y = 0.5$.

We consider the following two settings. In Setting 1, we assess the performance of different methods under varied strengths of the interactions between treatment and mediators in their effects on the outcome. In Setting 2, we assess the performance under varied effect sizes.

Setting 1 We fix $c_M = 1$, $\theta_T^\star = 3.15$ and set $c_D = 0, \pm 0.5, \pm 1, \pm 1.5, \pm 2$. The first ten entries in $\boldsymbol{\gamma}_0$ are -0.1, -0.2, -0.3, …, -1, and the first ten entries in $\boldsymbol{\gamma}_1$ are 0.1, 0.2, 0.3, …, 1. The rest of the entries in $\boldsymbol{\gamma}_0$ and $\boldsymbol{\gamma}_1$ are independently generated from $N(0, 0.1^2)$. Then, the NIEs are $\delta(t_0) := \delta_{\mathbf{x},1}(t_0) = 4.2$ and $\delta(t_1) := \delta_{\mathbf{x},1}(t_1) = 4.2(1 + c_D)$, and the NDEs are $\zeta(t_0) := \zeta_{\mathbf{x},1}(t_0) = 3.15 - 4.2c_D$ and $\zeta(t_1) := \zeta_{\mathbf{x},1}(t_1) = 3.15$, where we ignore the $(\mathbf{x}, 1)$ in the subscript as we assume no confounders and $q = 1$ in this simulation.

Setting 2 We fix $c_D = -2$, $\theta_T^\star = 0$ and set $c_M = 0, \pm 0.2, \pm 0.4, \pm 0.6, \pm 0.8, \pm 1$. The first ten entries in $\boldsymbol{\gamma}_0$ are -0.05, -0.1, -0.15, …, -0.5. The rest of the entries in $\boldsymbol{\gamma}_0$ and $\boldsymbol{\gamma}_1$ are generated the same way as in Setting 1. Then, $\delta(t_0) = \zeta(t_0) = 4.2c_M$, and $\delta(t_1) = \zeta(t_1) = -4.2c_M$.

We compare the performance of our method using the SCAD penalty (Fan & Li, 2001), denoted as M1, with the following four competing methods: M0, the oracle method with known $\mathcal{A}_G$ and $\mathcal{A}_{D_j}$ for $j = 1, \ldots, q$; M2, the *difference method* in Guo et al. (2023); M3, the debiased *product method* in Zhou et al. (2020); M4, the PCMA in Huang and Pan (2016). In the implementation, the tuning parameters in methods M1 and M2 are selected by cross-validation. For method M3, we apply the algorithm in R package `freebird` with $\tau_n = \sqrt{\log p/n}/3$ in the CLIME. Note that M2 and M3 do not consider interactions in the outcome model and only produce one estimate for NIE and one estimate for NDE. For method M4, we perform the eigen-decomposition of $\widehat{\boldsymbol{\Sigma}}_G$ estimated using the OLS residuals from the mediator model. The loading matrix $\widehat{\mathbf{U}} \in \mathbb{R}^{p \times K}$ is formed by the top eigenvectors that explain 80% variance in $\widehat{\boldsymbol{\Sigma}}_G$. We then fit K low-dimensional OLS models following Huang and Pan (2016).

We present the coverage probabilities and the powers of methods M0–M4 in Figure 4.1 for Setting 1 and in Figure 4.2 for Setting 2. Since under Setting 1, $\zeta(t_1) = 3.15$ and $\delta(t_0) = 4.2$

which are both constants, for the power performance in Figure 4.1, we only report the local power curves of $\zeta(t_0) = 3.15 - 4.2c_D$ and $\delta(t_1) = 4.2(1 + c_D)$. Under both Settings 1 and 2, our method M1 performs approximately as well as the oracle method M0 in most cases, providing valid confidence intervals while maintaining reasonable power curves. However, under Setting 1 with small values of c_D, *i.e.* , $c_D = \pm 0.5$, we observe that the coverage probabilities of our method for $\zeta(t_0)$ are slightly lower than 0.95. This is due to the required Condition A.1.4, a setup with very weak signals is not ideal for our method. Since M2 and M3 fail to account for the interaction, they both suffer from invalid coverage probabilities when $|c_D| \neq 0$. In Setting 2 where substantial cancellation occurs between $\boldsymbol{\theta}^\star_G$ and $\boldsymbol{\theta}^\star_D$ with $\boldsymbol{\theta}^\star_D = -2\boldsymbol{\theta}^\star_G$, both M2 and M3 demonstrate close to zero powers as expected. Since M4, the PCMA, is designed for the inference of the NIE but not NDE, we present its performance on the NIE only. After the PCA transformation $\mathbf{G}^* = \widehat{\mathbf{U}}^{\mathrm{T}}\mathbf{G}$, M4 uses $\boldsymbol{\beta}_1^{\mathrm{T}}\widehat{\mathbf{U}}\widehat{\mathbf{U}}^{\mathrm{T}}\boldsymbol{\theta}^\star_G$ and $\boldsymbol{\beta}_1^{\mathrm{T}}\widehat{\mathbf{U}}\widehat{\mathbf{U}}^{\mathrm{T}}(\boldsymbol{\theta}^\star_G + \boldsymbol{\theta}^\star_D)$ to estimate $\delta(t_0)$ and $\delta(t_1)$, respectively. Since $\widehat{\boldsymbol{\Sigma}}_G$ is singular when $p > n$, the loading matrix $\widehat{\mathbf{U}} \in \mathbb{R}^{p\times K}$ from PCA corresponds to 80% of variance in $\widehat{\boldsymbol{\Sigma}}_G$ so that $\widehat{\mathbf{U}}\widehat{\mathbf{U}}^{\mathrm{T}} \neq \mathbf{I}_p$. Such bias caused by the high dimensionality invalidates this method. The weak signals in $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ make the approximation of $\widehat{\boldsymbol{\Sigma}}_G$ even worse. The only scenario where such bias does not lead to biased inference on the NIE is when the true effect equals 0. This occurs for $\delta(t_1)$ when $c_D = -1$ in Setting 1 and for $\delta(t_0)$ and $\delta(t_1)$ when $\boldsymbol{\beta}_1 = \mathbf{0}_p$ in Setting 2.

In the supplementary material, we also present simulation results when $p = 1000$ or when $\boldsymbol{\varepsilon}_{G_i}$ is not normally distributed. The results lead to the same conclusions. In addition, the supplementary material shows the estimation bias when $p_{\lambda_n}(\cdot)$ is replaced by the ℓ_1-penalty (Lasso) in our method under Settings 1 and 2.

4.5.2 Simulations to Mimic the Real Data with Large p

In this section, we consider a setup that mimics the real data structure in Section 4.7 to evaluate the performance of our method with small n and large p, specifically, $n = 120$ and

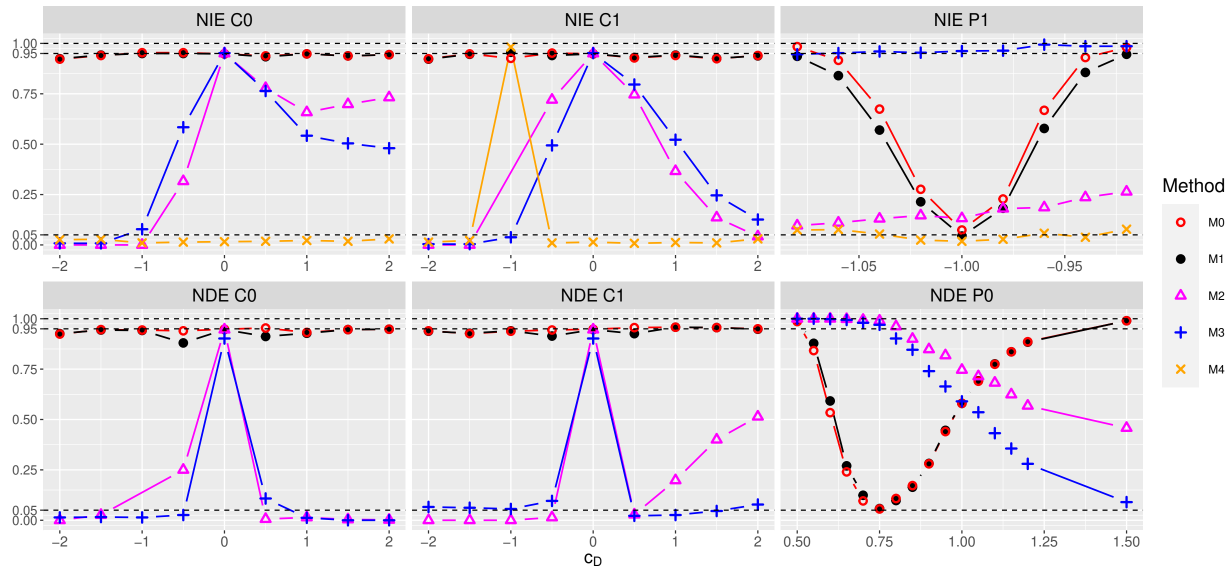


Figure 4.1: The comparison of different methods under Setting 1 when $p = 500$, $\boldsymbol{\varepsilon}_{G_i} \sim \mathcal{N}(\mathbf{0}_p, \boldsymbol{\Sigma}_p)$, and $\epsilon_{Y_i} \sim N(0, \sigma_Y^2)$. Under Setting 1, $\delta(t_0) = 4.2$, $\delta(t_1) = 4.2(1 + c_D)$, $\zeta(t_0) = 3.15 - 4.2c_D$, and $\zeta(t_1) = 3.15$. NIE C0: the coverage probability of $\delta(t_0)$; NIE C1: the coverage probability of $\delta(t_1)$; NIE P1: the empirical power of $\delta(t_1)$; NDE C0: the coverage probability of $\zeta(t_0)$; NDE C1: the coverage probability of $\zeta(t_1)$; NDE P0: the empirical power of $\zeta(t_0)$. M0: the oracle method; M1: our method; M2: the *difference method* in Guo et al. (2023); M3: the debiased *product method* in Zhou et al. (2020); M4: the PCMA in Huang and Pan (2016). Each coverage probability or power is calculated over 500 replications.

$p = 1000$ or 2000. As discussed in Section 4.5.1 , the PCMA from Huang and Pan (2016) is biased when $p > n$. The coverage probabilities of the debiased method also suffer from large p, as presented in Section 6.2 of the supplementary material of Zhou et al. (2020). And therefore, we focus on the comparison of the empirical coverage probabilities between our method (M1) and the *difference method* from Guo et al. (2022) (M2).

We generate two independent covariates $X_{i,1} \sim \text{Ber}(0.5)$ and $X_{i,2} \sim N(0, 1)$. The treatment $\mathbf{T}_i$ are generated using a balanced design, *i.e.* the first $n/3$ are $[1, 0]^{\mathrm{T}}$, the second $n/3$ are $[0, 1]^{\mathrm{T}}$, and the last $n/3$ are $[0, 0]^{\mathrm{T}}$. In the outcome model, since the magnitudes of the effect estimates are large in Table 4.2 and 4.3 in Section 4.7, we set $\theta_0^\star = -2$, $\boldsymbol{\theta}_T^\star = [1.5, -0.5]^{\mathrm{T}}$, $\boldsymbol{\theta}_X^\star = [3, -3]^{\mathrm{T}}$ for ease of demonstration. Let $\boldsymbol{\theta}_{G,\mathcal{A}_G}^\star = [-3, 2, 1, -1, -2, 4, -2, 4, 1, -1]^{\mathrm{T}}$ where $\mathcal{A}_G = \{20, 27, 106, 146, 149, 268, 497, 589, 807, 879\}$. For $\boldsymbol{\theta}_{D_1}^\star$, $\boldsymbol{\theta}_{D_2}^\star \in \mathbb{R}^p$, the nonzero interactions are $\theta_{D_1,106}^\star = 5$, $\theta_{D_1,879}^\star = 4$, and $\theta_{D_2,807}^\star = 5$. Set $\sigma_Y = 1$. In the mediator model, we

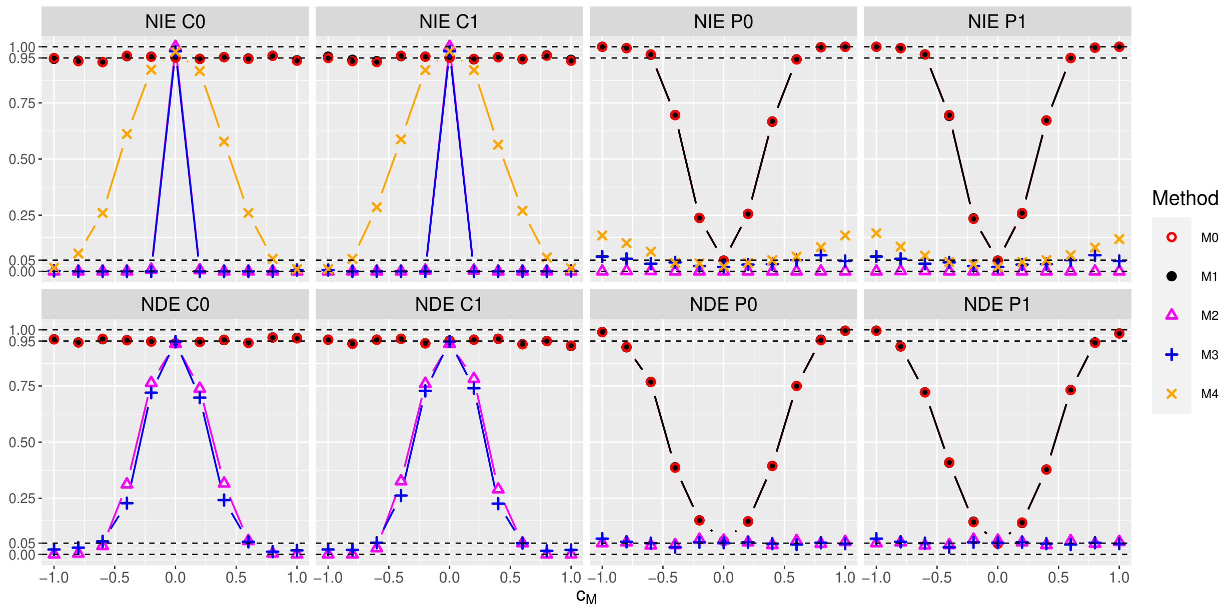


Figure 4.2: The comparison of different methods under Setting 2 when $p = 500$, $\boldsymbol{\varepsilon}_{G_i} = \mathcal{N}(\mathbf{0}_p, \boldsymbol{\Sigma}_G)$, and $\epsilon_{Y_i} \sim N(0, \sigma_Y^2)$. Under Setting 2, $\delta(t_0) = \zeta(t_0) = 4.2c_M$, $\delta(t_1) = \zeta(t_1) = -4.2c_M$. NIE C0: the coverage probability of $\delta(t_0)$; NIE C1: the coverage probability of $\delta(t_1)$; NIE P0: the empirical power of $\delta(t_0)$; NIE P1: the empirical power of $\delta(t_1)$; NDE C0: the coverage probability of $\zeta(t_0)$; NDE C1: the coverage probability of $\zeta(t_1)$; NDE P0: the empirical power of $\zeta(t_0)$; NDE P1: the empirical power of $\zeta(t_1)$. M0: the oracle method; M1: our method; M2: the *difference method* in Guo et al. (2023); M3: the debiased *product method* in Zhou et al. (2020); M4: the PCMA in Huang and Pan (2016). Each coverage probability or power is calculated over 500 replications.

set $\boldsymbol{\beta}_{0,\mathcal{A}_G} = [0.1, 0.2, \ldots, 1]^{\mathrm{T}}$, $\boldsymbol{\beta}_{1,\mathcal{A}_G} = \boldsymbol{\beta}_{2,\mathcal{A}_G} = [-0.05, -0.1, -0.15, \ldots, -0.5]^{\mathrm{T}}$. Other entries in $\boldsymbol{\beta}_0,\ \boldsymbol{\beta}_1,\ \boldsymbol{\beta}_2 \in \mathbb{R}^p$ and $\mathbf{A} \in \mathbb{R}^{p\times r}$ are generated from $N(0, 0.1^2)$. We focus on the NIEs and NDEs when $\mathbf{x} = [0.5, 0]^{\mathrm{T}}$.

The covariance $\boldsymbol{\Sigma}_G$ is generated by fitting the residuals of the first 1000 mediators from the real data in Section 4.7 using the R package `glasso` with a tuning parameter 0.082085, and then turning it into a correlation matrix. In the supplementary material, we also present a case with $\boldsymbol{\Sigma}_G$ generated by an autoregressive structure whose (i, j)-element is $0.5^{|i-j|}$. We consider three different distributions of the noise terms $\boldsymbol{\varepsilon}_{G_i}$ and ϵ_{Y_i}: Distribution 1: $\boldsymbol{\varepsilon}_{G_i} \sim \mathcal{N}(\mathbf{0}_p, \boldsymbol{\Sigma}_G)$ and $\epsilon_{Y_i} \sim N(0, 1)$; Distribution 2: $\boldsymbol{\varepsilon}_{G_i} \sim \boldsymbol{\Sigma}_G^{1/2}\mathbf{w}$, where the entries of $\mathbf{w}$ are i.i.d. $\sqrt{3/4}t(8)$ and $\epsilon_{Y_i} \sim N(0, 1)$; Distribution 3: $\boldsymbol{\varepsilon}_{G_i} \sim \boldsymbol{\Sigma}_G^{1/2}\mathbf{w}$, where the entries of $\mathbf{w}$ are i.i.d. Pois(1) $- 1$ and $\epsilon_{Y_i} \sim \sqrt{3/4}t(8)$. The tuning parameters of both our method and Guo et al. (2022) are selected by high-dimensional BIC (L. Wang, Kim, & Li, 2013).

Table 4.1 presents the empirical coverage probabilities and powers over 500 replications. We mark the power cell in gray if the corresponding coverage probability is below 0.8. Our method (M1) has valid coverage probabilities since we have a correct model in the presence of interactions. Notice that for M2, the powers for $\delta_{\mathbf{x},j}(\mathbf{t}_0)$ and $\delta_{\mathbf{x},j}(\mathbf{t}_j)$ ($\zeta_{\mathbf{x},j}(\mathbf{t}_0)$ and $\zeta_{\mathbf{x},j}(\mathbf{t}_j)$) are identical for both $j = 1, 2$ as a result of ignoring interactions in Guo et al. (2022).

$p = 1000$												
	$\boldsymbol{\varepsilon}_{G_i}$: Normal; ϵ_{Y_i}: Normal				$\boldsymbol{\varepsilon}_{G_i}$: t(8); ϵ_{Y_i}: Normal				$\boldsymbol{\varepsilon}_{G_i}$: Poisson; ϵ_{Y_i}: t(8)			
Effect	Coverage		Power		Coverage		Power		Coverage		Power	
	M1	M2	M1	M2	M1	M2	M1	M2	M1	M2	M1	M2
$\delta_{\mathbf{x},1}(\mathbf{t}_0) = -1.55$	0.958	0.868	0.140	0.314	0.930	0.864	0.164	0.316	0.938	0.876	0.160	0.296
$\delta_{\mathbf{x},1}(\mathbf{t}_1) = -4.30$	0.938	0.744	0.454	0.314	0.940	0.766	0.494	0.316	0.952	0.778	0.498	0.296
$\delta_{\mathbf{x},2}(\mathbf{t}_0) = -1.55$	0.956	0.880	0.188	0.344	0.938	0.876	0.172	0.354	0.924	0.862	0.158	0.314
$\delta_{\mathbf{x},2}(\mathbf{t}_2) = -3.80$	0.946	0.810	0.478	0.344	0.946	0.796	0.450	0.354	0.934	0.828	0.438	0.314
$\zeta_{\mathbf{x},1}(\mathbf{t}_0) = 6.56$	0.958	0.668	1.000	0.852	0.932	0.678	1.000	0.872	0.930	0.680	1.000	0.872
$\zeta_{\mathbf{x},1}(\mathbf{t}_1) = 3.81$	0.934	0.704	0.940	0.852	0.932	0.716	0.952	0.872	0.930	0.728	0.960	0.872
$\zeta_{\mathbf{x},2}(\mathbf{t}_0) = 3.58$	0.930	0.000	0.988	1.000	0.952	0.000	0.990	1.000	0.932	0.000	0.996	1.000
$\zeta_{\mathbf{x},2}(\mathbf{t}_2) = 1.33$	0.934	0.000	0.352	1.000	0.926	0.000	0.418	1.000	0.928	0.000	0.336	1.000
$p = 2000$												
	$\boldsymbol{\varepsilon}_{G_i}$: Normal; ϵ_{Y_i}: Normal				$\boldsymbol{\varepsilon}_{G_i}$: t(8); ϵ_{Y_i}: Normal				$\boldsymbol{\varepsilon}_{G_i}$: Poisson; ϵ_{Y_i}: t(8)			
Effect	Coverage		Power		Coverage		Power		Coverage		Power	
	M1	M2	M1	M2	M1	M2	M1	M2	M1	M2	M1	M2
$\delta_{\mathbf{x},1}(\mathbf{t}_0) = -1.55$	0.928	0.776	0.146	0.310	0.942	0.792	0.148	0.250	0.916	0.792	0.188	0.272
$\delta_{\mathbf{x},1}(\mathbf{t}_1) = -4.30$	0.944	0.488	0.490	0.310	0.938	0.514	0.470	0.250	0.930	0.520	0.494	0.272
$\delta_{\mathbf{x},2}(\mathbf{t}_0) = -1.55$	0.920	0.766	0.158	0.310	0.938	0.788	0.158	0.262	0.924	0.800	0.184	0.280
$\delta_{\mathbf{x},2}(\mathbf{t}_2) = -3.80$	0.930	0.564	0.450	0.310	0.942	0.578	0.448	0.262	0.938	0.590	0.480	0.280
$\zeta_{\mathbf{x},1}(\mathbf{t}_0) = 6.95$	0.946	0.594	1.000	0.718	0.932	0.586	1.000	0.736	0.946	0.596	1.000	0.698
$\zeta_{\mathbf{x},1}(\mathbf{t}_1) = 4.20$	0.918	0.812	0.974	0.718	0.918	0.838	0.960	0.736	0.954	0.830	0.994	0.698
$\zeta_{\mathbf{x},2}(\mathbf{t}_0) = 4.64$	0.956	0.000	1.000	1.000	0.934	0.000	1.000	1.000	0.940	0.000	1.000	1.000
$\zeta_{\mathbf{x},2}(\mathbf{t}_2) = 2.39$	0.932	0.000	0.782	1.000	0.922	0.000	0.774	1.000	0.932	0.000	0.780	1.000

Table 4.1: Empirical coverage probabilities and powers of our method and Guo et al. (2022) with $\boldsymbol{\Sigma}_G$ generated by `glasso`

4.6 When Heredity Structure Presents

As pointed out in Section 4.1, for a better interpretation of the main effects and interactions, following the spirit of C. Wang et al. (2020); Qiu, Tao, and Zhou (2021), we assume that $\theta^{\star}_{G,h} = 0 \implies \theta^{\star}_{D_j,h} = 0$ for $j = 1, \ldots, q$ and $h = 1, \ldots, p$.

The existing approaches to achieve $\widehat{\theta}^{\star}_{G,h} = 0 \implies \widehat{\theta}^{\star}_{D_j,h} = 0$ can be roughly divided into the following categories: (1) multi-step procedures (Shi, Wang, & Ding, 2023; R. Li, Zhu, Lee, & Initiative, 2024); (2) Bayesian approaches that impose heredity through priors (Chipman, 1996); (3) formulating the heredity constraint as a non-convex optimization problem (Choi, Li, & Zhu, 2010); (4) hierarchical group-lasso regularization (Lim & Hastie, 2015); (5) overlapping group Lasso (Radchenko & James, 2010; Jenatton, Audibert, & Bach, 2011; Yan & Bien, 2017); and (6) other regularized regression methods (Yuan, Joseph, & Zou, 2009; Jacob, Obozinski, & Vert, 2009; Bien, Taylor, & Tibshirani, 2013). Among these methods, only (R. Li et al., 2024) considers the high-dimensional meditation setting but lacks theoretical guarantees. In this section, we propose a heredity-inducing procedure for valid inference of the causal effects with high-dimensional mediators.

4.6.1 Overlapped Group SCAD for Heredity

Motivated by Jenatton et al. (2011); Yan and Bien (2017), we propose the following objective function with the overlapping group SCAD penalty to promote heredity:

$$\widehat{\boldsymbol{\theta}}^{\mathrm{H}}_W = \underset{\boldsymbol{\theta}_W \in \mathbb{R}^{1+q+r+p+pq}}{\arg\min} \frac{1}{2n}\|\mathbf{Y} - \mathbf{W}\boldsymbol{\theta}_W\|_2^2 + \sum_{h=1}^{p} p_{\lambda_n}(\|\boldsymbol{\eta}_h\|_2) + \sum_{h=1}^{p}\sum_{j=1}^{q} p_{\lambda_n}(|\theta_{D_j,h}|), \tag{4.6.1}$$

where $p_{\lambda_n}(\cdot)$ is the SCAD penalty (Fan & Li, 2001) and $\boldsymbol{\eta}_h = [\theta_{G,h}, \theta_{D_1,h}, \ldots, \theta_{D_q,h}]^{\mathrm{T}} \in \mathbb{R}^{q+1}$ is the collection of the main effect and interactions of the hth mediator for $h = 1, \ldots, p$. The overlapped penalty on the interaction $|\theta_{D_j,h}|$ tends to shrink $|\theta_{D_j,h}|$ more towards zero compared to the main effect $|\theta_{G,h}|$. If $\theta^{\star}_{G,h}$ and $\theta^{\star}_{D_j,h}$, $j = 1, \ldots, q$, satisfy the heredity principle, Theorem 4.6.1 shows that (4.6.1) has the oracle property, which is similar to Theorem 4.4.1:

Theorem 4.6.1. *Suppose that $p_{\lambda_n}(\cdot)$ is the SCAD penalty, and the true parameter $\boldsymbol{\theta}_W^\star$ satisfies that $\theta_{G,h}^\star = 0 \implies \theta_{D_j,h}^\star = 0,\ j = 1, \ldots, q$ for $h = 1, \ldots, p$. Under Conditions A.1.2–A.1.5, if $s = o(\sqrt{n})$ and $r \leq C \cdot s$ for an absolute constant $C > 0$, then with probability tending to 1, we have $\widehat{\boldsymbol{\theta}}_{W,\mathcal{A}^c}^{\mathrm{H}} = \mathbf{0}_{(1+q)p-s}$ and $\|\widehat{\boldsymbol{\theta}}_{W,\mathcal{A}}^{\mathrm{H}} - \boldsymbol{\theta}_{W,\mathcal{A}}^\star\|_2 = O_p(\sqrt{s/n})$.*

Remark 4.6.1. *Theorem 4.6.1 still holds if the penalty $p_{\lambda_n}(\cdot)$ is changed to MCP (C.-H. Zhang, 2010). The asymptotic normality of the estimators of NIEs and NDEs is the same as Theorem 4.4.2.*

4.6.2 Algorithm to Solve the Overlapping Group SCAD

Although (4.6.1) enjoys the oracle property, the optimization problem is computationally challenging to solve with the concave penalty function. Motivated by the idea from Zou and Li (2008) and Fan, Xue, and Zou (2014), we approximate the concave penalties in (4.6.1) locally using a linear function, *i.e.* , local linear approximation (LLA). The objective function in (4.6.1) can be approximated by the following overlapping group Lasso:

$$\begin{aligned}
\widehat{\boldsymbol{\theta}}_W^{\mathrm{LLA}} &= \underset{\boldsymbol{\theta}_W \in \mathbb{R}^{1+q+r+p+pq}}{\arg\min} \frac{1}{2n}\|\mathbf{Y} - \mathbf{W}\boldsymbol{\theta}_W\|_2^2 \\
&+ \sum_{h=1}^{p} p'_{\lambda_n}(\|\boldsymbol{\eta}_h^{(0)}\|_2)\|\boldsymbol{\eta}_h\|_2 + \sum_{h=1}^{p}\sum_{j=1}^{q} p'_{\lambda_n}(|\theta_{D_j,h}^{(0)}|)|\theta_{D_j,h}|, \qquad (4.6.2)
\end{aligned}$$

where $\boldsymbol{\theta}_W^{(0)} = [\theta_0^{(0)}\ \boldsymbol{\theta}_T^{(0)\mathrm{T}}\ \boldsymbol{\theta}_X^{(0)\mathrm{T}}\ \boldsymbol{\theta}_G^{(0)\mathrm{T}}\ \boldsymbol{\theta}_{D_1}^{(0)\mathrm{T}}\ \ldots\ \boldsymbol{\theta}_{D_q}^{(0)\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{1+q+r+p+pq}$ is the initial value with elements $\boldsymbol{\eta}_h^{(0)} = [\theta_{G,h}^{(0)}\ \theta_{D_1,h}^{(0)}\ \ldots\ \theta_{D_q,h}^{(0)}]^{\mathrm{T}}$ for $h = 1, \ldots, p$. The following lemma shows that if $\boldsymbol{\theta}_W^{(0)}$ is close enough to $\boldsymbol{\theta}_W^\star$, with an additional regularity condition, (4.6.2) yields an optimal estimator $\widehat{\boldsymbol{\theta}}_W^{\mathrm{LLA}}$ with heredity property.

Lemma 4.6.1. *Suppose that* $\min_{h\in\mathcal{A}_G}\{|\theta^\star_{G,h}|\} \wedge \min_{j=1,\ldots,q,h\in\mathcal{A}_{D_j}}\{|\theta^\star_{D_j,h}|\} > (1+\gamma)\lambda_n > 0$, *with some* $\gamma > 2$, *and that* $\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\mathbf{W}_{\mathcal{A}}$ *is nonsingular. Define*

$$\begin{aligned}
\mathcal{E}_{\text{initial}} &= \left\{\|\boldsymbol{\theta}_W^{(0)} - \boldsymbol{\theta}^\star_W\|_\infty \leq \lambda_n\right\} \cap \left\{\|\boldsymbol{\eta}_h^{(0)}\|_2 \leq \lambda_n,\ \forall\ h \in \{1,\ldots,p\}\setminus\mathcal{A}_G\right\}, \text{ and}\\
\mathcal{E}_{\text{regularity}} &= \left\{\|\mathbf{W}_{\mathcal{A}^c}^{\mathrm{T}}\left[\mathbf{I}_n - \mathbf{W}_{\mathcal{A}}\left(\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\mathbf{W}_{\mathcal{A}}\right)^{-1}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\right]\boldsymbol{\varepsilon}_Y\|_\infty < n\lambda_n\right\}.
\end{aligned}$$

Then on the event $\mathcal{E}_{\text{initial}}$, (4.6.2) *becomes*

$$\underset{\boldsymbol{\theta}_W\in\mathbb{R}^{1+q+r+p+pq}}{\arg\min}\ \frac{1}{2n}\|\mathbf{Y}-\mathbf{W}\boldsymbol{\theta}_W\|_2^2 + \sum_{h\notin\mathcal{A}_G}\lambda_n\|\boldsymbol{\eta}_h\|_2 + \sum_{j=1}^{q}\sum_{h\notin\mathcal{A}_{D_j}}\lambda_n|\theta_{D_j,h}|. \tag{4.6.3}$$

Further, on the event $\mathcal{E}_{\text{initial}} \cap \mathcal{E}_{\text{regularity}}$, (4.6.2) *has a unique solution* $\widehat{\boldsymbol{\theta}}_W^{\mathrm{LLA}} \in \mathbb{R}^{1+q+r+p+pq}$ *such that* $\widehat{\boldsymbol{\theta}}_{W,\mathcal{A}}^{\mathrm{LLA}} = (\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\mathbf{W}_{\mathcal{A}})^{-1}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\mathbf{Y}$ *and* $\widehat{\boldsymbol{\theta}}_{W,\mathcal{A}^c}^{\mathrm{LLA}} = \mathbf{0}_{(1+q)p-s}$.

Jenatton et al. (2011) shows that the solution of (4.6.3) induces heredity if $\mathbf{Y}$, given $\mathbf{W}$, has an absolutely continuous distribution. Thus, the event $\mathcal{E}_{\text{initial}}$ guarantees the desired heredity of $\widehat{\boldsymbol{\theta}}_W^{\mathrm{LLA}}$. Additionally, the event $\mathcal{E}_{\text{initial}} \cap \mathcal{E}_{\text{regularity}}$ guarantees that $\widehat{\boldsymbol{\theta}}_W^{\mathrm{LLA}}$ performs as well as knowing the true support $\mathcal{A}$. As shown in Theorem 4.6.2 below, if the initial value $\boldsymbol{\theta}_W^{(0)}$ is $\widehat{\boldsymbol{\theta}}^\star_W$ from (4.3.1), then $\mathbb{P}\{\mathcal{E}_{\text{initial}} \cap \mathcal{E}_{\text{regularity}}\} \to 1$ so that $\widehat{\boldsymbol{\theta}}_W^{\mathrm{LLA}}$ of (4.6.2) also enjoys the oracle property.

Theorem 4.6.2. *Suppose that the true parameter* $\boldsymbol{\theta}^\star_W$ *satisfies that* $\theta^\star_{G,h} = 0 \implies \theta^\star_{D_j,h} = 0$, *for* $j = 1,\ldots,q$ *and* $h = 1,\ldots,p$. *Under Conditions A.1.2–A.1.5, if* $s = o(\sqrt{n})$, *and* $r \leq C\cdot s$ *for an absolute constant* $C > 0$, *and* $\boldsymbol{\theta}_W^{(0)} = \widehat{\boldsymbol{\theta}}^\star_W$ *from* (4.3.1), *then* $\mathbb{P}\{\mathcal{E}_{\text{initial}} \cap \mathcal{E}_{\text{regularity}}\} \to 1$, *i.e. , with probability tending to* 1, $\widehat{\boldsymbol{\theta}}_{W,\mathcal{A}^c}^{\mathrm{LLA}} = \mathbf{0}_{(1+q)p-s}$ *and* $\|\widehat{\boldsymbol{\theta}}_{W,\mathcal{A}}^{\mathrm{LLA}} - \boldsymbol{\theta}^\star_{W,\mathcal{A}}\|_2 = O_p(\sqrt{s/n})$.

Following Theorem 4.1 in Tseng (2001), under the event $\mathcal{E}_{\text{initial}}$, the Block Coordinate Descent (BCD) method converges to a global minimum of (4.6.2). We relegate the corresponding algorithm to Section A.3.4.

4.7 Real Data Analysis

The dataset is from `https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-GEOD-77445`, which was previously analyzed by Houtepen et al. (2016); van Kesteren and Oberski (2019); Guo et al. (2022). However, the potential interactions between childhood trauma and DNA methylation were ignored in previous research. In this study, we re-analyze the same dataset to incorporate such interactions, allowing us to detect the potentially different effects of DNA methylation on cortisol stress reactivity under different levels of childhood trauma.

In this section, we perform a genome-wide analysis of blood DNA methylation in 85 healthy individuals. The original exposure S is a score obtained through a questionnaire on childhood trauma. This score ranges from 24 to 63, and we categorize it according to the 0.33 quantile, 27, and the 0.66 quantile, 32. That is, $T_1 = \mathcal{I}\{S \in [27, 32)\}$ and $T_2 = \mathcal{I}\{S \in [32, 63]\}$. In our case, $q = 2$. The response variable Y is the increased area under the curve (iAUC) in cortisol after a stress test. The mediators $\mathbf{G}$ analyzed is of dimension $p = 1008$, which contain a subset of the 385,882 DNA methylation loci in the blood. Specifically, $\mathbf{G}$ consists of the top 1000 mediators with the largest absolute values of the products of the following two correlations, namely, the correlation between the DNA methylation locus and S, and the correlation between the DNA methylation locus and Y; together with `cg27512205` (G_{1001}), `cg05608730` (G_{1002}), and `cg26179948` (G_{1003}) discovered in Houtepen et al. (2016), and `cg02309301` (G_{1004}), `cg12500973` (G_{1005}), `cg16657538` (G_{1006}), `cg25626453` (G_{1007}), and `cg13136721` (G_{1008}) discovered in van Kesteren and Oberski (2019). Note that `cg19230917` (G_{27}), `cg06422529` (G_{785}), and `cg03199124` (G_{976}) discovered by Guo et al. (2022) are also included in $\mathbf{G}$. The covariates $\mathbf{X}$ include age(X_1), sex(X_2), B cell proportion(X_3), CD4 T cell proportion(X_4), CD8 T cell proportion(X_5), Monocytes cell proportion(X_6), Granulocytes cell proportion(X_7), and Natural Killer cell proportion(X_8). We standardize the variables in $\mathbf{X}$ except the binary variable X_2 to have zero means and

ℓ_2 norm $\sqrt{n}$. The tuning parameters are selected by high-dimensional BIC (L. Wang et al., 2013).

Following *Step 1* in Section 4.3, we apply the SCAD penalty to estimate $\boldsymbol{\theta}_W^\star \in \mathbb{R}^{1+q+r+p+pq}$ in the outcome model, that is,

$$\widehat{\boldsymbol{\theta}}_W^\star = \underset{\boldsymbol{\theta}_W \in \mathbb{R}^{3035}}{\arg\min} \frac{1}{2n}\|\mathbf{Y} - \mathbf{W}\boldsymbol{\theta}_W\|_2^2 + \sum_{h=1}^{1008} p_{\lambda_{\text{main}}}(|\theta_{G,h}|) + \sum_{j=1}^{2}\sum_{h=1}^{1008} p_{\lambda_{\text{int}}}(|\theta_{D_j,h}|).$$

Note that we adopt different values for λ in $p_\lambda(\cdot)$ for the main effects and the interactions, *i.e.* , λ_{main} and λ_{int}, respectively. By minimizing the high-dimensional BIC, λ_{main} is selected to be 32.72 and $\lambda_{\text{int}} = 0.50\lambda_{\text{main}}$.

We further adjust $\widehat{\boldsymbol{\theta}}_W^\star$ to promote heredity using the LLA procedure in (4.6.2):

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_W^{\text{LLA}} \;&=\; \underset{\boldsymbol{\theta}_W \in \mathbb{R}^{3035}}{\arg\min} \frac{1}{2n}\|\mathbf{Y} - \mathbf{W}\boldsymbol{\theta}_W\|_2^2 \\ &+\; \sum_{h=1}^{1008} p'_{\lambda_{\text{main}}}(\|\boldsymbol{\eta}_h^{(0)}\|_2)\|\boldsymbol{\eta}_h\|_2 + \sum_{h=1}^{1008}\sum_{j=1}^{2} p'_{\lambda_{\text{int}}}(|\theta_{D_j,h}^{(0)}|)|\theta_{D_j,h}|,\end{aligned}$$

where the initial values are from $\widehat{\boldsymbol{\theta}}_W^\star$. The tuning parameters are chosen to be $\lambda_{\text{main}} =$ 39.02 and $\lambda_{\text{int}} = 0.75\lambda_{\text{main}}$ by high-dimensional BIC. The optimization is performed by the proposed block coordinate descent algorithm, from which the minimizer, denoted as $\widehat{\boldsymbol{\theta}}_W^{\text{LLA}} = [\widehat{\theta}_0^{\text{LLA}}, \widehat{\boldsymbol{\theta}}_T^{\text{LLA},\text{T}}, \widehat{\boldsymbol{\theta}}_X^{\text{LLA},\text{T}}, \widehat{\boldsymbol{\theta}}_G^{\text{LLA},\text{T}}, \widehat{\boldsymbol{\theta}}_{D_1}^{\text{LLA},\text{T}}, \widehat{\boldsymbol{\theta}}_{D_2}^{\text{LLA},\text{T}}]^{\text{T}} \in \mathbb{R}^{3035}$, follows heredity structure.

The estimated support of the main effects $\widehat{\boldsymbol{\theta}}_G^{\text{LLA}}$, denoted as $\widehat{\mathcal{A}}_G^{\text{LLA}}$, is {20, 27, 55, 106, 146, 149, 268, 497, 589, 641, 807, 879, 1003} with $|\widehat{\mathcal{A}}_G^{\text{LLA}}| = 13$. The estimated supports of the interactions $\widehat{\boldsymbol{\theta}}_{D_1}^{\text{LLA}}$ and $\widehat{\boldsymbol{\theta}}_{D_2}^{\text{LLA}}$ are both {106, 807, 879}, a subset of $\widehat{\mathcal{A}}_G^{\text{LLA}}$, which manifests heredity. The mediators with nonzero main effects are presented in Table 4.2. The corresponding genes are presented in Section A.5 of the Appendix.

In Table 4.2, `cg19230917` (G_{27}) and `cg26179948` (G_{1003}) are also detected by (Guo et al., 2022). Yet, we detect many new mediators that are not identified by Guo et al. (2022). This is probably due to that we allow the interactions $\mathbf{T} : \mathbf{G}$. For example, T_1G_{106}, T_1G_{879},

Table 4.2: Individual Coefficients in the Outcome Model with Heredity Adjustment

Effect	Estimated Value	%95 CI	Test Statistic	p value
intercept	-201.833	(-1865.312, 1461.647)	-0.238	0.812
T_1	162.091	(-41.362, 365.544)	1.561	0.118
T_2	-66.991	(-277.356, 143.373)	-0.624	0.533
X_1	10.478	(-34.191, 55.147)	0.460	0.646
X_2	327.959	(220.162, 435.757)	5.963	0.000 ***
X_3	-85.494	(-182.682, 11.694)	-1.724	0.085 .
X_4	-163.554	(-406.959, 79.852)	-1.317	0.188
X_5	-68.896	(-211.340, 73.548)	-0.948	0.343
X_6	-82.106	(-194.798, 30.586)	-1.428	0.153
X_7	-265.370	(-621.623, 90.883)	-1.460	0.144
X_8	-106.238	(-277.947, 65.471)	-1.213	0.225
cg06719334 (G_{20})	-327.588	(-523.529, -131.648)	-3.277	0.001 **
cg19230917 (G_{27})	225.222	(48.720, 401.724)	2.501	0.012 *
cg18634806 (G_{55})	232.015	(84.012, 380.018)	3.073	0.002 **
cg02860705 (G_{106})	20.986	(-129.725, 171.697)	0.273	0.785
cg11032707 (G_{146})	-132.129	(-256.900, -7.358)	-2.076	0.038 *
cg05781698 (G_{149})	-249.047	(-450.879, -47.216)	-2.418	0.016 *
cg14063008 (G_{268})	427.849	(238.036, 617.663)	4.418	0.000 ***
cg07439316 (G_{497})	-223.606	(-484.882, 37.671)	-1.677	0.093 .
cg06957003 (G_{589})	361.628	(168.760, 554.496)	3.675	0.000 ***
cg13662644 (G_{641})	456.226	(315.528, 596.924)	6.355	0.000 ***
cg16142488 (G_{807})	-62.824	(-259.801, 134.153)	-0.625	0.532
cg03025830 (G_{879})	25.722	(-72.643, 124.088)	0.513	0.608
cg26179948 (G_{1003})	-329.274	(-524.897, -133.650)	-3.299	0.001 **
T_1*cg02860705 (T_1G_{106})	533.339	(327.015, 739.663)	5.066	0.000 ***
T_1*cg03025830 (T_1G_{879})	285.611	(164.926, 406.296)	4.638	0.000 ***
T_2*cg16142488 (T_2G_{807})	334.846	(78.397, 591.295)	2.559	0.010 *

and T_2G_{807} are significant interactions as presented in Table 4.2, which are missed by the model in Guo et al. (2022). Figure 4.3 demonstrates the different behaviors of these three mediators under different levels of $\mathbf{T}$, which further justifies the existence of interactions. Additionally, based on our method, $\{G_{20}, G_{27}, G_{55}, G_{146}, G_{149}, G_{286}, G_{589}, G_{641}, G_{1003}, T_1G_{106}, T_1G_{879}, T_2G_{807}\}$ are significant as presented in Table 4.2. If we fit OLS with these mediators, *i.e.* , $Y \sim T_1 + T_2 + \mathbf{X} + G_{20} + G_{27} + G_{55} + G_{146} + G_{149} + G_{286} + G_{589} + G_{641} + G_{1003} + T_1G_{106} + T_1G_{879} + T_2G_{807}$, R^2 is 0.904 with these 22 variables. In contrast, if we fit $Y \sim T_1 + T_2 + \mathbf{X} + \mathbf{G}_{\mathcal{A}_{11}} + T_1\mathbf{G}_{\mathcal{A}_{11}} + T_2\mathbf{G}_{\mathcal{A}_{11}}$, where $\mathcal{A}_{11}$ are those eleven mediators selected by Guo et al. (2022), there are no significant interactions and R^2 is 0.834 with 43 variables. We believe our selected mediators are more appropriate with categorized exposure and with interactions.

Recall that for $j = 1$ and 2, the NIEs are $\delta_{\mathbf{x},j}(\mathbf{t}_j) = \boldsymbol{\beta}_j^{\mathrm{T}}(\boldsymbol{\theta}_G^\star + \boldsymbol{\theta}_{D_j}^\star)$ and $\delta_{\mathbf{x},j}(\mathbf{t}_0) = \boldsymbol{\beta}_j^{\mathrm{T}}\boldsymbol{\theta}_G^\star$. The NDEs are $\zeta_{\mathbf{x},j}(\mathbf{t}_j) = \theta_{T,j}^\star + (\boldsymbol{\beta}_0 + \boldsymbol{\beta}_j + \mathbf{A}\mathbf{x})^{\mathrm{T}}\boldsymbol{\theta}_{D_j}^\star$ and $\zeta_{\mathbf{x},j}(\mathbf{t}_0) = \theta_{T,j}^\star + (\boldsymbol{\beta}_0 + \mathbf{A}\mathbf{x})^{\mathrm{T}}\boldsymbol{\theta}_{D_j}^\star$.

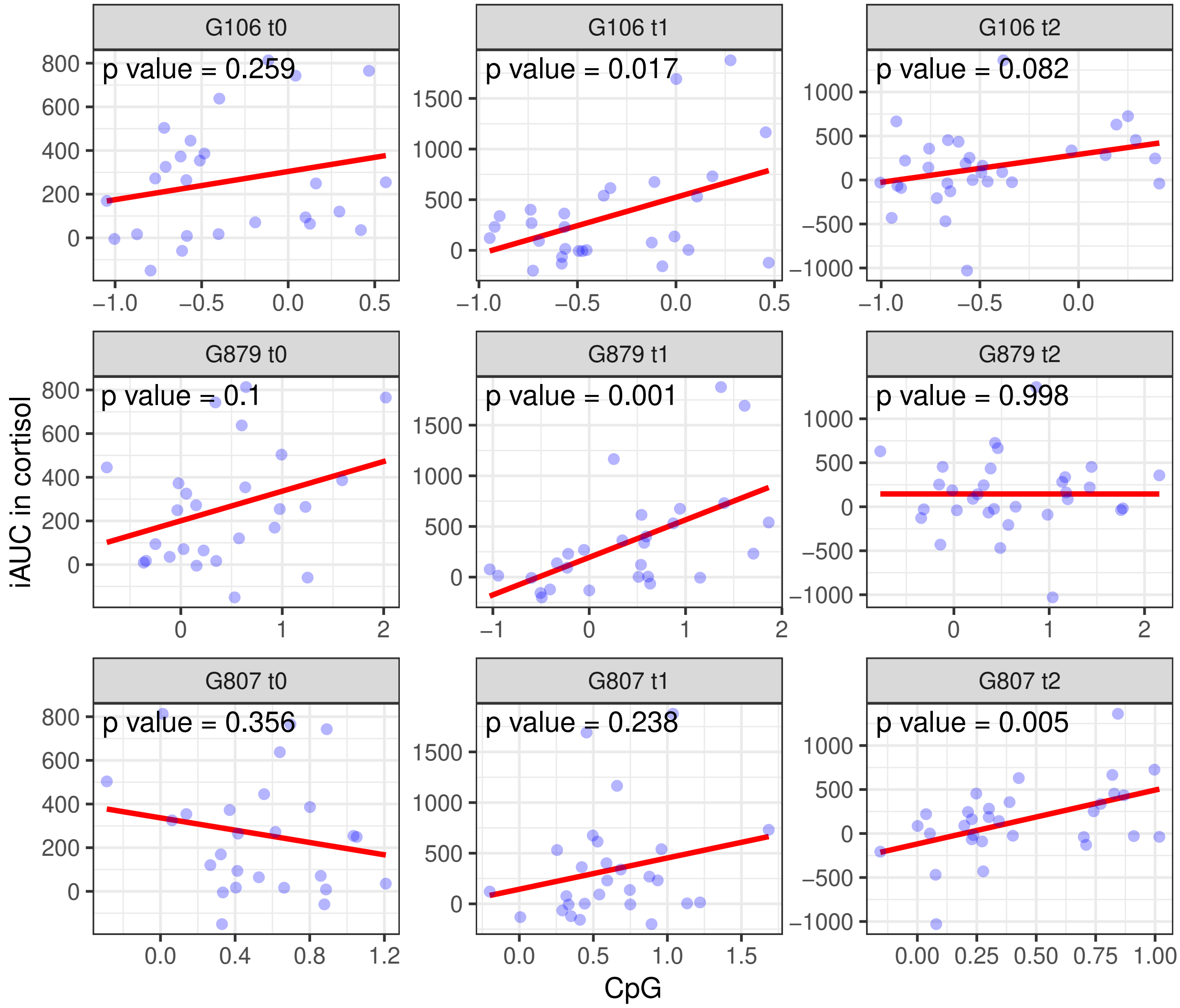


Figure 4.3: Scatter plots of Y and G_{106}, G_{879}, G_{807} under different treatment levels. The p-values of G_{106} and G_{876} under $\mathbf{t}_1$ and G_{807} under $\mathbf{t}_2$ are all significant, which is consistent with Table 4.2.

Following the *Step 2* in Section 4.3 and setting $\mathbf{x}$ to be the empirical average of $\mathbf{X}$, Table 4.3 presents the results for the inference of those NIEs and NDEs.

Similar to Guo et al. (2022), the estimated NDEs are positive while the estimated NIEs are negative. When comparing $\mathbf{t}_1$ and $\mathbf{t}_0$, *i.e.* comparing $S < 27$ with $S \in [27, 32)$, we have $\delta_{\mathbf{x},1}(\mathbf{t}_1) - \delta_{\mathbf{x},1}(\mathbf{t}_0) = \zeta_{\mathbf{x},1}(\mathbf{t}_1) - \zeta_{\mathbf{x},1}(\mathbf{t}_0) = \widehat{\boldsymbol{\beta}}_1^{\mathrm{T}} \widehat{\boldsymbol{\theta}}_{D_1}^{\mathrm{LLA}} = -39.47$. When comparing $\mathbf{t}_2$ and $\mathbf{t}_0$, *i.e.* comparing $S < 27$ with $S \geq 32$, we have $\delta_{\mathbf{x},2}(\mathbf{t}_2) - \delta_{\mathbf{x},2}(\mathbf{t}_0) = \zeta_{\mathbf{x},2}(\mathbf{t}_2) - \zeta_{\mathbf{x},2}(\mathbf{t}_0) = \widehat{\boldsymbol{\beta}}_2^{\mathrm{T}} \widehat{\boldsymbol{\theta}}_{D_2}^{\mathrm{LLA}} = -53.54$. Hence the increment in the score of childhood trauma reduces the cortisol stress reactivity in both NIEs and NDEs.

Table 4.3: NIE and NDE

Effect	Estimated Value	Standard Error	%95 CI	Test Statistic	p value
$\delta_{\mathbf{x},1}(\mathbf{t}_0)$	-80.899	75.299	(-228.483, 66.685)	-1.074	0.283
$\delta_{\mathbf{x},1}(\mathbf{t}_1)$	-120.369	120.990	(-357.506, 116.768)	-0.995	0.320
$\delta_{\mathbf{x},2}(\mathbf{t}_0)$	-168.056	77.954	(-320.843, -15.269)	-2.156	0.031 *
$\delta_{\mathbf{x},2}(\mathbf{t}_2)$	-221.595	85.343	(-388.865, -54.325)	-2.597	0.009 *
$\zeta_{\mathbf{x},1}(\mathbf{t}_0)$	128.285	68.742	(-6.448, 263.017)	1.866	0.062 .
$\zeta_{\mathbf{x},1}(\mathbf{t}_1)$	88.815	66.240	(-41.013, 218.642)	1.341	0.180
$\zeta_{\mathbf{x},2}(\mathbf{t}_0)$	124.289	48.618	(28.999, 219.579)	2.556	0.011 *
$\zeta_{\mathbf{x},2}(\mathbf{t}_2)$	70.750	48.883	(-25.060, 166.560)	1.447	0.148

4.8 Discussion

In this work, we propose an approach for causal mediation analysis that accommodates high-dimensional mediators in the presence of interactions between the treatment and mediators in their effects on the outcome. Under linear models, we establish the asymptotic normality of the proposed estimators for NIEs and NDEs by leveraging the oracle property of the SCAD or MCP, and demonstrate the importance of incorporating the interactions through simulation studies and real data application. We further extend our approach to impose heredity through overlapped group SCAD.

Several limitations of the work are to be addressed in the future research. First, our approach is developed within the framework of linear structural equation models where both the mediators and the outcome are continuous. The development of high-dimensional mediation approaches for nonlinear models is still at an early stage in the literature. In the setup of continuous mediators and survival outcome, preliminary dimension reduction on the mediators through sure independence screening (Fan & Lv, 2008) combined with variable selection in the outcome model have been proposed to detect active mediators (Cui, Luo, Luo, & Yu, 2021; H. Zhang, Zheng, Hou, Zheng, & Liu, 2021; Chi, Flowers, Li, Huang, & Wei, 2024). Nevertheless, these procedures do not allow for interactions between the treatment and the mediators, nor do they aim to estimate the NIEs and NDEs. How to extend these procedures to allow for interactions and inference of the NIEs and NDEs is an important topic to explore in future research. In addition, nonlinear model for mediators imposes further challenges on inference. Second, we focus on the joint effects of the mediators as a group. In practice, the causal pathways through each individual mediator may also be of interest. When the mediators are not parallel, causally antecedent mediators may become post-treatment confounders between the relationship of the decedent mediators and the outcome, rendering many path-specific natural effects unidentifiable (T. VanderWeele & Vansteelandt, 2014; Daniel, De Stavola, Cousens, & Vansteelandt, 2015). One potential avenue for further investigation is to shift the focus to the interventional direct and indirect effects (T. J. VanderWeele, Vansteelandt, & Robins, 2014; Vansteelandt & Daniel, 2017) which can be identified in the presence of post-treatment confounders. Third, in our heredity-inducing algorithm, we approximate the overlapping group SCAD penalty using local linear approximation, *i.e.* , the overlapping group Lasso. However, the overlapping group Lasso can be time-consuming in practice. To reduce the computational time, Qi and Li (2024) recently introduced a non-overlapping penalty to approximate the overlapping group Lasso penalty, which comes with theoretical guarantees. It is of future research interest to employ this new approach to enhance the computational efficiency of our heredity-inducing algorithm.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this dissertation, we develop a panoramic understanding of the least squares inference in both large and small sample regimes, under unknown yet arbitrarily correlated errors and random design. We then extend the linear modeling framework to high-dimensional mediation analysis with treatment-by-mediator interaction. Specifically, in Chapter 2, we establish the robustness of the least squares inference in the large sample regime against correlated error with unknown structure. Leveraging the random design and a self-normalization argument, we derive Berry–Esseen bounds that control the Type I error rate. Additionally, we develop a local power approximation to conceptually illustrate the power phase transition under varying signal strengths. In Chapter 3, we shift focus to the small sample regime and delineate the relationship between the F distribution and the spherical distribution. Through a geometric perspective, we visualize the symmetric roles of the design matrix and the error vector in the linear model. This symmetry guarantees the validity of F- and t- tests against arbitrary errors, provided the design matrix has spherical symmetry. In Chapter 4, wwe address high-dimensional mediation analysis in the presence of treatment-by-mediator interaction. We propose a two-step estimation procedure for the natural indirect and direct effects. The first step applies a nonconvex penalized least squares approach to the outcome model, achieving the oracle property and reducing the high-dimensional mediators to their low-dimensional support. The second step applies least squares estimation to the mediator model, restricted to the directions of the identified active mediators, and establishes asymptotic normality of the resulting causal effect estimators. To promote interpretability via the heredity principle, we further introduce an overlapping group SCAD penalty.

5.2 Future Work

In the development of the Berry-Esseen bounds in Chapter 2, the randomness in $\mathbf{X}$ dominates the correlation in $\boldsymbol{\varepsilon}$ which leads to the ultimate valid inference. Such perspective of randomness is similar to the complete randomization setting in Shi and Ding (2022) traced back to the potential outcome framework in Neyman (1923). Regarding the key role of random design in this paper, addressing correlated data under arbitrarily fixed design remains an open question. Recently, different permutation schemes have been proposed under exchangeable errors, such as the Cyclic Permutation Test in Lei and Bickel (2021) and the Permutation-Augmented Linear Model Regression Test in Guan (2023), as well as the residual randomization in Toulis (2019) under invariance assumption. For more general correlation structures and fixed designs, the exploration of valid inference is a challenging but intriguing future work.

Another key assumption in this paper is the independence across the rows in $\mathbf{X}$. For the projection estimator, a violation of this assumption has complicated influences. If $\mathbf{X} = \mathbf{\Phi}^{1/2}\mathbf{Z}\mathbf{\Sigma}^{1/2}$ where the row dependence is captured by $\mathbf{\Phi} \in \mathbb{R}^{n\times n}$, the bias $\widehat{\boldsymbol{\beta}}_{\text{OLS}} - \boldsymbol{\beta} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\boldsymbol{\varepsilon} = \mathbf{\Sigma}^{-1/2}(\mathbf{Z}^{\mathsf{T}}\mathbf{\Phi}\mathbf{Z})^{-1}\mathbf{Z}^{\mathsf{T}}\mathbf{\Phi}^{1/2}\boldsymbol{\varepsilon}$ and $\mathbf{\Phi}^{1/2}\boldsymbol{\varepsilon}$ will yield diverse interactions between $\mathbf{X}$ and $\mathbb{C}\mathrm{ov}(\boldsymbol{\varepsilon})$, such as the spurious correlation from a network structure (Y. Lee & Ogburn, 2021) and the cluster effects on the standard error (Abadie et al., 2017). In practice, similar correlation structures in $\mathbf{\Phi}$ and $\mathbb{C}\mathrm{ov}(\boldsymbol{\varepsilon})$ are typical when $\mathbf{x}_{(i)}$ and y_i are sampled from a common structure. Theoretical properties of projection estimators with different types of $\mathbf{\Phi}$ and $\mathbb{C}\mathrm{ov}(\boldsymbol{\varepsilon})$ remain to be investigated in the future.

References

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. (2017). *When should you adjust standard errors for clustering?* (Tech. Rep.). National Bureau of Economic Research.

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. M. (2020). Sampling-based versus design-based uncertainty in regression analysis. *Econometrica*, *88*(1), 265–296.

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. M. (2023). When should you adjust standard errors for clustering? *Q. J. Econ.*, *138*(1), 1–35.

Aitken, A. C. (1936). IV.-On least squares and linear combination of observations. *Proceedings of the Royal Society of Edinburgh*, *55*, 42–48.

Aldrich, J. (1978). An alternative derivation of Durbin's h statistic. *Econometrica*, *46*(6), 1493.

Ali, M. M., & Ponnapalli, R. (1990). An optimal property of the Gauss-Markoff estimator. *Journal of Multivariate Analysis*, *32*(2), 171–176.

Anderson, M. J. (2001). Permutation tests for univariate or multivariate analysis of variance and regression. *Canadian journal of fisheries and aquatic sciences*, *58*(3), 626–639.

Anderson, R., & Bancroft, T. (1952). *Statistical theory in research.* New York: McGraw-Hill.

Anderson, T. W. (2011). *The statistical analysis of time series.* John Wiley & Sons.

Andrews, D. W. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 817–858.

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion.* Princeton University Press.

Anselin, L. (1988). *Spatial econometrics: Methods and models* (Vol. 4). Springer Science & Business Media.

Anselin, L. (2022). Spatial econometrics. *Handbook of Spatial Analysis in the Social Sciences*, 101–122.

Arellano, M., et al. (1987). Computing robust standard errors for within-groups estimators.

Oxford Bulletin of Economics and Statistics, *49*(4), 431–434.

Arnold, S. F. (1979). Linear models with exchangeably distributed errors. *Journal of the American statistical Association*, *74*(365), 194–199.

Atiqullah, M. (1964). The robustness of the covariance analysis of a one-way classification. *Biometrika*, *51*(3-4), 365–372.

Bancroft, T. A. (1944). On biases in estimation due to the use of preliminary tests of significance. *The Annals of Mathematical Statistics*, *15*(2), 190–204.

Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*. CRC press.

Banerjee, S., Gelfand, A. E., Finley, A. O., & Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B*, *70*(4), 825–848.

Banks, C. A., Boanca, G., Lee, Z. T., Eubanks, C. G., Hattem, G. L., Peak, A., ... Washburn, M. P. (2016). TNIP2 is a hub protein in the NF-κB network with both protein and RNA mediated interactions. *Molecular & Cellular Proteomics*, *15*(11), 3435–3449.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173.

Barrios, T., Diamond, R., Imbens, G. W., & Kolesár, M. (2012). Clustering, spatial correlations, and randomization inference. *Journal of the American Statistical Association*, *107*(498), 578–591.

Beach, C. M., & MacKinnon, J. G. (1978). A maximum likelihood procedure for regression with autocorrelated errors. *Econometrica*, 51–58.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B*, *36*(2), 192–225.

Bester, C. A., Conley, T. G., & Hansen, C. B. (2011). Inference with dependent data using cluster covariance estimators. *Journal of Econometrics*, *165*(2), 137–151.

Bickel, P. J., & Freedman, D. (1982). Bootstrapping regression models with many parameters. In *A festschrift for erich l. lehmann* (pp. 28–48). Belmont, Calif.: Wadsworth Statist./Probab. Ser.

Bien, J., Taylor, J., & Tibshirani, R. (2013). A lasso for hierarchical interactions. *The Annals of Statistics*, *41*(3), 1111.

Billingsley, P. (1995). *Probability and measure* (3rd ed.). John Wiley & Sons.

Bloomfield, P., & Watson, G. S. (1975). The inefficiency of least squares. *Biometrika*, *62*(1), 121–128.

Bohrer, R. (1973). An optimality property of scheffé bounds. *The Annals of Statistics*, *1*(4), 766–772.

Box, G. E. (1953). Non-normality and tests on variances. *Biometrika*, *40*(3/4), 318–335.

Box, G. E. (1954a). Some theorems on quadratic forms applied in the study of analysis of variance problems, i. effect of inequality of variance in the one-way classification. *The annals of mathematical statistics*, *25*, 290–302.

Box, G. E. (1954b). Some theorems on quadratic forms applied in the study of analysis of variance problems, ii. effects of inequality of variance and of correlation between errors in the two-way classification. *The Annals of Mathematical Statistics*, *25*(3), 484–498.

Box, G. E., Hunter, W. H., Hunter, S., et al. (1978). *Statistics for experimenters: An introduction to design, data analysis, and model building*. New York: John Wiley and Sons.

Box, G. E., & Watson, G. S. (1962). Robustness to non-normality of regression tests. *Biometrika*, *49*(1-2), 93–106.

Breusch, T. S., & Pagan, A. R. (1980). The Lagrange multiplier test and its applications to model specification in econometrics. *The Review of Economic Studies*, *47*(1), 239–253.

Brown, M. B., & Forsythe, A. (1974). The anova and multiple comparisons for data with heterogeneous variances. *Biometrics*, *30*, 719–724.

Brunner, E., Dette, H., & Munk, A. (1997). Box-type approximations in nonparametric factorial designs. *Journal of the American Statistical Association*, *92*(440), 1494–1502.

Burridge, P. (1980). On the Cliff-Ord test for spatial correlation. *Journal of the Royal Statistical Society: Series B*, *42*(1), 107–108.

Cai, T., Liu, W., & Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, *106*(494), 594–607.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2011). Robust inference with multiway clustering. *Journal of Business & Economic Statistics*, *29*(2), 238–249.

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and applications*. Cambridge University Press.

Campbell, C. (1977). *Properties of ordinary and weighted least squares estimators of regression coefficients for two-stage samples*. Southern Methodist University.

Carlstein, E. (1986). Simultaneous confidence regions for predictions. *The American Statistician*, *40*(4), 277–279.

Case, A. C. (1991). Spatial patterns in household demand. *Econometrica*, 953–965.

Casella, G., & Berger, R. (2024). *Statistical inference*. CRC Press.

Chén, O. Y., Crainiceanu, C., Ogburn, E. L., Caffo, B. S., Wager, T. D., & Lindquist, M. A. (2018). High-dimensional multivariate mediation with application to neuroimaging data. *Biostatistics*, *19*(2), 121–136.

Chetverikov, D., Hahn, J., Liao, Z., & Santos, A. (2023). Standard errors when a regressor is randomly assigned. *arXiv:2303.10306*.

Chi, S., Flowers, C. R., Li, Z., Huang, X., & Wei, P. (2024). MASH: Mediation analysis of survival outcome and high-dimensional omics mediators with application to complex diseases. *The Annals of Applied Statistics*, *18*(2), 1360 – 1377. Retrieved from `https://doi.org/10.1214/23-AOAS1838` doi: 10.1214/23-AOAS1838

Chipman, H. (1996). Bayesian variable selection with related predictors. *Canadian Journal of Statistics*, *24*(1), 17–36.

Chivukula, R. R., Montoro, D. T., Leung, H. M., Yang, J., Shamseldin, H. E., Taylor, M. S., ... others (2020). A human ciliopathy reveals essential functions for NEK10 in airway mucociliary clearance. *Nature Medicine*, *26*(2), 244–251.

Choi, N. H., Li, W., & Zhu, J. (2010). Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association*, *105*(489), 354–364.

Clark-Boucher, D., Zhou, X., Du, J., Liu, Y., Needham, B. L., Smith, J. A., & Mukherjee, B. (2023). Methods for mediation analysis with high-dimensional DNA methylation data: Possible choices and comparisons. *PLoS Genetics*, *19*(11), e1011022.

Cliff, A., & Ord, K. (1972). Testing for spatial autocorrelation among regression residuals. *Geographical Analysis*, *4*(3), 267–284.

Cochran, W. (1934). The distribution of quadratic forms in a normal system with applications to the analysis of covariance. *Mathematical Proceedings of the Cambridge Philosophical Society*, *30*, 178–191.

Cochrane, D., & Orcutt, G. H. (1949). Application of least squares regression to relationships containing auto-correlated error terms. *Journal of the American statistical association*, *44*(245), 32–61.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge.

Conley, T. G. (1999). GMM estimation with cross sectional dependence. *Journal of Econometrics*, *92*(1), 1–45.

Cressie, N. (2015). *Statistics for spatial data*. John Wiley & Sons.

Cui, Y., Luo, C., Luo, L., & Yu, Z. (2021). High-dimensional mediation analysis based on additive hazards model for survival data. *Frontiers in Genetics*, *12*, 771932.

D'Agostino, R. B. (2017). *Goodness-of-fit-techniques*. Routledge.

Dai, J. Y., Stanford, J. L., & LeBlanc, M. (2022). A multiple-testing procedure for high-dimensional mediation hypotheses. *Journal of the American Statistical Association*, *117*(537), 198–213.

Daniel, R. M., De Stavola, B. L., Cousens, S. N., & Vansteelandt, S. (2015). Causal mediation analysis with multiple mediators. *Biometrics*, *71*(1), 1–14.

David, F. N., & Johnson, N. (1951). The effect of non-normality on the power function of the f-test in the analysis of variance. *Biometrika*, *38*(1/2), 43–57.

David, F. N., & Neyman, J. (1938). Extension of the markoff theorem on least squares. *Statistical Research Memoirs*, *2*, 105–117.

Davies, M. N., Verdi, S., Burri, A., Trzaskowski, M., Lee, M., Hettema, J. M., ... Spector, T. D. (2015). Generalised anxiety disorder–a twin study of genetic architecture, genome-wide association and differential gene expression. *PLoS One*, *10*(8), e0134865.

Davis, C. S. (2002). *Statistical methods for the analysis of repeated measurements* (Tech. Rep.). Springer.

Derkach, A., Pfeiffer, R. M., Chen, T.-H., & Sampson, J. N. (2019). High dimensional mediation analysis with latent variables. *Biometrics*, *75*(3), 745–756.

Diggle, P. J. (1988). An approach to the analysis of repeated measurements. *Biometrics*, 959–971.

Ding, P. (2024). Linear Model and Extensions.
(*arXiv*: 2401.00649)

Ding, P., & Dasgupta, T. (2018). A randomization-based perspective on analysis of variance: a test statistic robust to treatment effect heterogeneity. *Biometrika*, *105*(1), 45–56.

Draper, N., Guttman, I., & Kanemasu, H. (1971). The distribution of certain regression statistics. *Biometrika*, *58*(2), 295–298.

Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (Vol. 326). John Wiley & Sons.

Durbin, J. (1970). Testing for serial correlation in least-squares regression when some of the regressors are lagged dependent variables. *Econometrica*, 410–421.

Durbin, J., & Watson, G. S. (1950). Testing for serial correlation in least squares regression: I. *Biometrika*, *37*(3/4), 409–428.

Durbin, J., & Watson, G. S. (1951). Testing for serial correlation in least squares regression: II. *Biometrika*, *38*(1/2), 159–177.

Eaton, M. L., & Efron, B. (1970). Hotelling's t2 test under symmetry conditions. *Journal of the American Statistical Association*, *65*(330), 702–711.

Efron, B. (1969). Student's t-test under symmetry conditions. *Journal of the American Statistical Association*, *64*(328), 1278–1302.

Efroymson, M. A. (1960). Multiple regression analysis. In A. Ralston & H. S. Wilf (Eds.), *Mathematical methods for digital computers* (Vol. I, pp. 191–203). New York: Wiley.

Eicker, F. (1967). Limit theorems for regressions with unequal and dependent errors. *Proc. 5th Berkeley Symp. Math. Stat. Probab.*, *1*(1), 59–82.

Epstein, B. (1960a). Tests for the validity of the assumption that the underlying distribution of life is exponential part i. *Technometrics*, *2*(1), 83–101.

Epstein, B. (1960b). Tests for the validity of the assumption that the underlying distribution of life is exponential: Part ii. *Technometrics*, *2*(2), 167–183.

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

Fan, J., & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *70*(5), 849–911.

Fan, J., & Lv, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Transactions on Information Theory*, *57*(8), 5467–5484.

Fan, J., Xue, L., & Zou, H. (2014). Strong oracle optimality of folded concave penalized estimation. *The Annals of Statistics*, *42*(3), 819.

Fang, K. W. (2018). *Symmetric multivariate and related distributions.* Chapman and Hall/CRC.

Favelyukis, S., Till, J. H., Hubbard, S. R., & Miller, W. T. (2001). Structure and autoregulation of the insulin-like growth factor 1 receptor kinase. *Nature Structural Biology*, *8*(12), 1058–1063.

Fisher, R. A. (1924). On a distribution yielding the error functions of several well-known statistics. *Proceedings of the International Congress of Mathematicians*, *2*, 805-–813.

Fisher, R. A. (1928). *Statistical methods for research workers* (No. 5). Oliver and Boyd.

Fisher, R. A. (1935). *The design of experiments.* Edinburgh: Oliver and Boyd. (Later editions: 1937, 1942, 1947, 1949, 1951, 1960, 1966. The eighth edition (1966) is reprinted as part of *Statistical Methods, Experimental Design, and Scientific Inference*, Oxford University Press, Oxford (1990).)

Freedman, D., & Lane, D. (1983). A nonstochastic interpretation of reported significance levels. *Journal of Business & Economic Statistics*, *1*(4), 292–298.

Frisch, R., & Waugh, F. V. (1933). Partial time regressions as compared with individual trends. *Econometrica*, *1*, 387–401.

Gao, Y., Yang, H., Fang, R., Zhang, Y., Goode, E. L., & Cui, Y. (2019). Testing mediation effects in high-dimensional epigenetic studies. *Frontiers in Genetics*, *10*, 1195.

Gauss, C. F. (1855). *Méthode des moindres carriés (trans. j. bertrand).* Paris.

Gelfand, A. E., Diggle, P., Guttorp, P., & Fuentes, M. (2010). *Handbook of spatial statistics.* CRC press.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of educational research*, *42*(3), 237–288.

Godfrey, L. G. (1978). Testing for higher order serial correlation in regression equations when the regressors include lagged dependent variables. *Econometrica*, 1303–1310.

Gorroochurn, P. (2016). *Classic topics on the history of modern mathematical statistics: From laplace to more recent times.* John Wiley & Sons.

Greenwald, B. C. (1983). A general analysis of bias in the estimated standard errors of least squares coefficients. *Journal of Econometrics*, *22*(3), 323–338.

Griffiths, W., & Beesley, P. (1984). The small-sample properties of some preliminary test estimators in a linear model with autocorrelated errors. *Journal of Econometrics*, *25*(1-2), 49–61.

Gross, M. D., Eggen, M. A., Simon, A. M., & Van Pilsum, J. F. (1986). The purification and characterization of human kidney L-arginine: glycine amidinotransferase. *Archives of Biochemistry and Biophysics*, *251*(2), 747–755.

Guan, L. (2023). A conformal test of linear models via permutation-augmented regressions. *arXiv preprint arXiv:2309.05482*.

Guo, X., Li, R., Liu, J., & Zeng, M. (2022). High-dimensional mediation analysis for selecting DNA methylation loci mediating childhood trauma and cortisol stress reactivity. *Journal of the American Statistical Association*, *117*(539), 1110–1121.

Guo, X., Li, R., Liu, J., & Zeng, M. (2023). Statistical inference for linear mediation models with high-dimensional mediators and application to studying stock reaction to COVID-19 pandemic. *Journal of Econometrics*, *235*(1).

Hald, A. (2008). *A history of parametric statistical inference from bernoulli to fisher, 1713-1935*. Springer Science & Business Media.

Halperin, M. (1951). Normal regression theory in the presence of intra-class correlation. *The Annals of Mathematical Statistics*, 573–580.

Hamada, M., & Wu, C. J. (1992). Analysis of designed experiments with complex aliasing. *Journal of Quality Technology*, *24*(3), 130–137.

Hansen, C. B. (2007). Generalized least squares inference in panel and multilevel models with serial correlation and fixed effects. *Journal of Econometrics*, *140*(2), 670–694.

Hayashi, F. (2011). *Econometrics*. Princeton University Press.

Hofmeister, W., Nilsson, D., Topa, A., Anderlid, B.-M., Darki, F., Matsson, H., … others (2015). CTNND2—a candidate gene for reading problems and mild intellectual disability. *Journal of Medical Genetics*, *52*(2), 111–122.

Holland, P. W. (1988). Causal inference, path analysis and recursive structural equations models. *ETS Research Report Series*, *1988*(1), i–50.

Hong, G., Deutsch, J., & Hill, H. D. (2015). Ratio-of-mediator-probability weighting for causal mediation analysis in the presence of treatment-by-mediator interaction. *Journal of Educational and Behavioral Statistics*, *40*(3), 307–340.

Houtepen, L. C., Vinkers, C. H., Carrillo-Roa, T., Hiemstra, M., Van Lier, P. A., Meeus, W., … others (2016). Genome-wide DNA methylation levels and altered cortisol stress reactivity following childhood trauma in humans. *Nature Communications*, *7*(1), 10967.

Huang, Y.-T., & Pan, W.-C. (2016). Hypothesis test of mediation effect in causal mediation model with high-dimensional continuous mediators. *Biometrics*, *72*(2), 402–413.

Imai, K., Keele, L., & Yamamoto, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, *25*(1), 51–71. doi: 10.1214/10-STS321

Irwin, J. (1931). Mathematical theorems involved in the analysis of variance. *Journal of the Royal Statistical Society*, *94*, 284–300.

Irwin, J. (1934). On the independence of the constituent items in the analysis of variance. *Supplement to the Journal of the Royal Statistical Society*, *1*, 236–251.

Jacob, L., Obozinski, G., & Vert, J.-P. (2009). Group lasso with overlap and graph lasso. In *Proceedings of the 26th annual international conference on machine learning* (pp. 433–440).

Javanmard, A., & Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, *15*(1), 2869–2909.

Jenatton, R., Audibert, J.-Y., & Bach, F. (2011). Structured variable selection with sparsity-inducing norms. *The Journal of Machine Learning Research*, *12*, 2777–2824.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). *Continuous univariate distributions* (2nd ed., Vol. 1). New York: John wiley & sons.

Kariya, T., & Kurata, H. (2004). *Generalized least squares*. John Wiley & Sons.

Kelejian, H. H., & Prucha, I. R. (1999). A generalized moments estimator for the autoregressive parameter in a spatial model. *International Economic Review*, *40*(2), 509–533.

Kelker, D. (1970). Distribution theory of spherical distributions and a location-scale parameter generalization. *Sankhyā: The Indian Journal of Statistics, Series A*, *32*, 419–430.

Kiefer, N. M., Vogelsang, T. J., & Bunzel, H. (2000). Simple robust testing of regression hypotheses. *Econometrica*, *68*(3), 695–714.

Kleinbaum, D. G., Kupper, L. L., Nizam, A., & Rosenberg, E. S. (2013). *Applied regression analysis and other multivariable methods*. Cengage Learning.

Kloek, T. (1981). OLS estimation in a model where a microvariable is explained by aggregates and contemporaneous disturbances are equicorrelated. *Econometrica*, 205–207.

Koreisha, S. G., & Fang, Y. (2001). Generalized least squares with misspecified serial correlation structures. *Journal of the Royal Statistical Society: Series B*, *63*(3), 515–531.

Krämer, W., & Donninger, C. (1987). Spatial autocorrelation among errors and the relative efficiency of OLS in the linear regression model. *Journal of the American Statistical Association*, *82*(398), 577–579.

Kruskal, W. (1968). When are Gauss-Markov and least squares estimators identical? A coordinate-free approach. *The Annals of Mathematical Statistics*, *39*(1), 70–75.

Kucherenko, M. M., & Shcherbata, H. R. (2018). Stress-dependent miR-980 regulation of Rbfox1/A2bp1 promotes ribonucleoprotein granule formation and cell survival. *Nature Communications*, *9*(1), 312.

Kuchibhotla, A. K., Rinaldo, A., & Wasserman, L. (2020). Berry-Esseen bounds for projection parameters and partial correlations with increasing dimension. *arXiv preprint arXiv:2007.09751*.

Lane, T. P., & DuMouchel, W. H. (1994). Simultaneous confidence intervals in multiple regression. *The American Statistician*, *48*(4), 315–321.

Lee, S., & Xing, E. P. (2012). Leveraging input and output structures for joint mapping of epistatic and marginal eqtls. *Bioinformatics*, *28*(12), i137–i146.

Lee, Y., & Ogburn, E. L. (2021). Network dependence can lead to spurious associations and invalid inference. *Journal of the American Statistical Association*, *116*(535), 1060–1074.

Lehmann, E. L., & Romano, J. P. (2005). *Testing statistical hypotheses* (3rd ed.). New York, NY, USA: Springer.

Lei, L., & Bickel, P. J. (2021). An assumption-free exact test for fixed-design linear models with exchangeable errors. *Biometrika*, *108*(2), 397–412.

Li, R., Zhu, X., Lee, S., & Initiative, A. D. N. (2024). Model selection for exposure-mediator interaction. *Data Science in Science*, *3*(1), 2360892.

Li, X., Sudarsanam, N., & Frey, D. D. (2006). Regularities in data from factorial experiments. *Complexity*, *11*(5), 32–45.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13–22.

Liang, K.-Y., & Zeger, S. L. (1993). Regression analysis for correlated data. *Annual Review of Public Health*, *14*(1), 43–68.

Liao, Z.-Z., Wang, Y.-D., Qi, X.-Y., & Xiao, X.-H. (2019). JAZF1, a relevant metabolic regulator in type 2 diabetes. *Diabetes/Metabolism Research and Reviews*, *35*(5), e3148.

Lieberman, G., Miller Jr, R., & Hamilton, M. A. (1967). Unlimited simultaneous discrimination intervals in regression. *Biometrika*, *54*(1-2), 133–145.

Lim, M., & Hastie, T. (2015). Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics*, *24*(3), 627–654.

Limam, M. M., & Thomas, D. R. (1988). Simultaneous tolerance intervals for the linear regression model. *Journal of the American Statistical Association*, *83*(403), 801–804.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *Ann. Appl. Stat.*, *7*(1), 295–318.

Lin, Z., & Bai, Z. (2010). *Probability inequalities.* Springer.

Liu, Z., Shen, J., Barfield, R., Schwartz, J., Baccarelli, A. A., & Lin, X. (2022). Large-scale hypothesis testing for causal mediation effects with applications in genome-wide epigenetic studies. *Journal of the American Statistical Association*, *117*(537), 67–81.

Lobato, I. N. (2001). Testing that a dependent process is uncorrelated. *Journal of the American Statistical Association*, *96*(455), 1066–1076.

Lovell, M. C. (1963). Seasonal adjustment of economic time series and multiple regression analysis. *Journal of the American Statistical Association*, *58*(304), 993–1010.

MacKinnon, D. P. (2012). *Introduction to statistical mediation analysis.* Routledge.

MacKinnon, D. P., Valente, M. J., & Gonzalez, O. (2020). The correspondence between causal and traditional mediation analysis: The link is the mediator by treatment interaction. *Prevention Science*, *21*, 147–157.

Mahalanobis, P. (1932). Auxiliary tables for fisher's z-test in analysis of variance. *Indian Journal of Agricultural Science*, *2*, 679–693.

Mallya, M., Campbell, R. D., & Aguado, B. (2002). Transcriptional analysis of a novel cluster of LY-6 family members in the human and mouse major histocompatibility complex: five genes with many splice forms. *Genomics*, *80*(1), 113–123.

Mardia, K., & Watkins, A. (1989). On multimodality of the likelihood in the spatial linear model. *Biometrika*, *76*(2), 289–295.

Mardia, K. V., & Marshall, R. J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, *71*(1), 135–146.

Markov, A. A. (1912). *Wahrscheinlichkeitsrechnung.* BG Teubner.

McElroy, F. (1967). A necessary and sufficient condition that ordinary least-squares estimators be best linear unbiased. *Journal of the American Statistical Association*, *62*(320), 1302–1304.

Meng, W., Mushika, Y., Ichii, T., & Takeichi, M. (2008). Anchorage of microtubule minus ends to adherens junctions regulates epithelial cell-cell contacts. *Cell*, *135*(5), 948–959.

Møller, J. (2013). *Spatial statistics and computational methods* (Vol. 173). Springer Science & Business Media.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, *37*(1/2), 17–23.

Moulton, B. R. (1986). Random group effects and the precision of regression estimates. *Journal of Econometrics*, *32*(3), 385–397.

Moulton, B. R. (1990). An illustration of a pitfall in estimating the effects of aggregate variables on micro units. *The Review of Economics and Statistics*, 334–338.

Muirhead, R. J. (2009). *Aspects of multivariate statistical theory*. New York: John Wiley & Sons.

Müller, U. K., & Watson, M. W. (2022a). Spatial correlation robust inference. *Econometrica*, *90*(6), 2901–2935.

Müller, U. K., & Watson, M. W. (2022b). Spatial correlation robust inference in linear regression and panel models. *Journal of Business & Economic Statistics*, 1–15.

Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, *55*, 703–708.

Newey, W. K., & West, K. D. (1994). Automatic lag selection in covariance matrix estimation. *The Review of Economic Studies*, *61*(4), 631–653.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. essay on principles. *Annals of Agricultural Sciences*, 1–51.

Olshen, R. A. (1973). The conditional level of the f—test. *Journal of the American Statistical Association*, *68*(343), 692–698.

Ord, K. (1975). Estimation methods for models of spatial interaction. *Journal of the American Statistical Association*, *70*(349), 120–126.

Pearl, J. (2001). Direct and indirect effects. *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, 411–420.

Peixoto, J. L. (1987). Hierarchical variable selection in polynomial regression models. *The American Statistician*, *41*(4), 311–313.

Pitman, E. J. (1937). Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society*, *4*(1), 119–130.

Pitman, E. J. G. (1938). Significance tests which may be applied to samples from any populations: III. The analysis of variance test. *Biometrika*, *29*(3/4), 322–335.

Plackett, R. L. (1949). A historical note on the method of least squares. *Biometrika*, *36*(3/4), 458–460.

Pope, P. T., & Webster, J. T. (1972). The use of an f-statistic in stepwise regression procedures. *Technometrics*, *14*(2), 327–340.

Powers, D. E., & Swinton, S. S. (1984). Effects of self-study for coachable test item types. *Journal of Educational Psychology*, *76*(2), 266.

Prais, S. J., & Winsten, C. B. (1954). *Trend estimators and serial correlation* (Tech. Rep.). Cowles Commission Discussion Paper, Chicago.

Preacher, K. J., Rucker, D. D., & Hayes, A. F. (2007). Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral Research*, *42*(1), 185–227.

Qi, M., & Li, T. (2024). The non-overlapping statistical approximation to overlapping group lasso. *Journal of Machine Learning Research*, *25*(115), 1–70.

Qiu, Y., Tao, J., & Zhou, X.-H. (2021). Inference of heterogeneous treatment effects using observational data with high-dimensional covariates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *83*(5), 1016–1043.

Radchenko, P., & James, G. M. (2010). Variable selection using adaptive nonlinear interaction structures in high dimensions. *Journal of the American Statistical Association*, *105*(492), 1541–1553.

Ramsay, J. O., & Silverman, B. W. (1997). *Functional data analysis.* New York: Springer.

Rao, C. R., Rao, C. R., Statistiker, M., Rao, C. R., & Rao, C. R. (1973). *Linear statistical inference and its applications* (Vol. 2). Wiley New York.

Ravikumar, B., Imarisio, S., Sarkar, S., O'Kane, C. J., & Rubinsztein, D. C. (2008). Rab5 modulates aggregation and toxicity of mutant huntingtin through macroautophagy in cell and fly models of Huntington disease. *Journal of Cell Science*, *121*(10), 1649–1660.

Robins, J. M., & Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, *3*(2), 143–155.

Rudelson, M., & Vershynin, R. (2015). Delocalization of eigenvectors of random matrices with independent entries. *Duke Mathematical Journal*, *164*(13), 2507–2538.

Safran, M., Dalah, I., Alexander, J., Rosen, N., Iny Stein, T., Shmoish, M., ... others (2010). GeneCards Version 3: the human gene integrator. *Database*, *2010*.

Schabenberger, O., & Gotway, C. A. (2017). *Statistical methods for spatial data analysis.* CRC press.

Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika*, *40*(1-2), 87–110.

Scheffé, H. (1959). *The analysis of variance.* New York: Wiley.

Schochet, P. Z., Pashley, N. E., Miratrix, L. W., & Kautz, T. (2022). Design-based ratio estimators and central limit theorems for clustered, blocked RCTs. *Journal of the American Statistical Association*, *117*(540), 2135–2146.

Scott, A. J., & Holt, D. (1982). The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association*, *77*(380), 848–854.

Seal, H. L. (1967). Studies in the history of probability and statistics. XV The historical development of the Gauss linear model. *Biometrika*, *54*(1-2), 1–24.

Seber, G. A., & Lee, A. J. (2003). *Linear regression analysis.* John Wiley & Sons.

Shen, Q., & Faraway, J. (2004). An f test for linear models with functional responses. *Statistica Sinica*, *14*, 1239–1257.

Shi, L., & Ding, P. (2022). Berry-Esseen bounds for design-based causal inference with possibly diverging treatment levels and varying group sizes. *arXiv preprint arXiv:2209.12345*.

Shi, L., Wang, J., & Ding, P. (2023). Forward screening and post-screening inference in factorial designs. *arXiv preprint arXiv:2301.12045*.

Snedecor, G. W. (1934). *Calculation and interpretation of analysis of variance and covariance.* Ames, Iowa: Collegiate Press.

Snedecor, G. W., & Cochran, W. G. (1980). *Statistical methods* (7th ed.). Ames, Iowa: Iowa State University Press.

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological Methodology*, *13*, 290–312.

Song, Y., Zhou, X., Kang, J., Aung, M. T., Zhang, M., Zhao, W., … Bhramar, M. (2021). Bayesian sparse mediation analysis with targeted penalization of natural indirect effects. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *70*(5), 1391–1412.

Song, Y., Zhou, X., Zhang, M., Zhao, W., Liu, Y., Kardia, S. L., … Mukherjee, B. (2020). Bayesian shrinkage estimation of high dimensional causal mediation effects in omics studies. *Biometrics*, *76*(3), 700–710.

Stigler, S. M. (2004). Discussion of d. denis. *Journal de la société française de statistique*, *145*(4), 63–64.

Stigler, S. M. (2016). *The seven pillars of statistical wisdom.* Harvard University Press.

Su, F., & Ding, P. (2021). Model-assisted analyses of cluster-randomized experiments. *Journal of the Royal Statistical Society Series B*, *83*(5), 994–1015.

Taylor, W. E. (1977). Small sample properties of a class of two stage Aitken estimators. *Econometrica*, 497–508.

Toulis, P. (2019). Invariant inference via residual randomization. *arXiv preprint arXiv:1908.04218*.

Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, *109*, 475–494.

Turner, T. N., Sharma, K., Oh, E. C., Liu, Y. P., Collins, R. L., Sosa, M. X., … Chakravarti1, A. (2015). Loss of δ-catenin function in severe autism. *Nature*, *520*(7545), 51–56.

Valeri, L., & VanderWeele, T. J. (2013). Mediation analysis allowing for exposure–mediator interactions and causal interpretation: Theoretical assumptions and implementation with SAS and SPSS macros. *Psychological Methods*, *18*(2), 137.

van de Geer, S., Bühlmann, P., Ritov, Y., & Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 1166–1202.

VanderWeele, T. (2015). *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press.

VanderWeele, T., & Vansteelandt, S. (2014). Mediation analysis with multiple mediators. *Epidemiologic Methods*, *2*(1), 95–115.

VanderWeele, T. J. (2016). Mediation analysis: a practitioner's guide. *Annual Review of Public Health*, *37*, 17–32.

VanderWeele, T. J., Asomaning, K., Tchetgen Tchetgen, E. J., Han, Y., Spitz, M. R., Shete, S., … Lin, X. (2012). Genetic variants on 15q25.1, smoking, and lung cancer: an assessment of mediation and interaction. *American Journal of Epidemiology*, *175*(10), 1013–1020.

VanderWeele, T. J., & Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and its Interface*, *2*(4), 457–468.

VanderWeele, T. J., & Vansteelandt, S. (2010). Odds ratios for mediation analysis for a dichotomous outcome. *American Journal of Epidemiology*, *172*(12), 1339–1348.

VanderWeele, T. J., Vansteelandt, S., & Robins, J. M. (2014). Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology*, *25*(2), 300–306.

van Kesteren, E.-J., & Oberski, D. L. (2019). Exploratory mediation analysis with many potential mediators. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(5), 710–723.

Vansteelandt, S., & Daniel, R. M. (2017). Interventional effects for mediation analysis with multiple mediators. *Epidemiology*, *28*(2), 258–265.

Vershynin, R. (2012). *Introduction to the non-asymptotic analysis of random matrices.* Cambridge: Cambridge University Press.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint* (Vol. 48). Cambridge University Press.

Wang, C., Hu, J., Blaser, M. J., & Li, H. (2020). Estimating and testing the microbial causal mediation effect with high-dimensional and compositional microbiome data. *Bioinformatics*, *36*(2), 347–355.

Wang, L., Kim, Y., & Li, R. (2013). Calibrating non-convex penalized regression in ultra-high dimension. *The Annals of Statistics*, *41*(5), 2505.

Wang, Y. S., Lee, S. K., Toulis, P., & Kolar, M. (2021). Robust inference for high-dimensional linear models via residual randomization. In *International conference on machine learning* (pp. 10805–10815).

Warnes, J. J., & Ripley, B. D. (1987). Problems with likelihood estimation of covariance functions of spatial Gaussian processes. *Biometrika*, *74*(3), 640–642.

Welch, B. L. (1951). On the comparison of several mean values: an alternative approach. *Biometrika*, *38*(3/4), 330–336.

Wen, K., Wang, T., & Wang, Y. (2022). Residual permutation test for high-dimensional regression coefficient testing. *arXiv preprint arXiv:2211.16182*.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 817–838.

Wickens, T. D., & Keppel, G. (2004). *Design and analysis: A researcher's handbook.* New Jersey: Prentice-Hall.

Wilkinson, G. N. (1970). A general recursive procedure for analysis of variance. *Biometrika*, *27*, 19–46.

Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., & Nichols, T. E. (2014). Permutation inference for the general linear model. *Neuroimage*, *92*, 381–397.

Wooldridge, J. M. (2012). *Introductory econometrics: A modern approach.* Cengage Learning.

Wu, C. J., & Hamada, M. S. (2011). *Experiments: planning, analysis, and optimization.* John Wiley & Sons.

Yan, X., & Bien, J. (2017). Hierarchical sparse modeling: A choice of two group lasso formulations. *Statistical Science*, *32*(4), 531–560. doi: 10.1214/17-STS622

Yang, C., Wan, X., Yang, Q., Xue, H., & Yu, W. (2010). Identifying main effects and epistatic interactions from large-scale snp data via adaptive group lasso. *BMC bioinformatics*, *11*, 1–11.

Yang, F., Eckardt, S., Leu, N. A., McLaughlin, K. J., & Wang, P. J. (2008). Mouse TEX15 is essential for DNA double-strand break repair and chromosomal synapsis during male meiosis. *The Journal of Cell Biology*, *180*(4), 673–679.

Yuan, M., Joseph, V. R., & Zou, H. (2009). Structured variable selection and estimation. *The Annals of Applied Statistics*, 1738–1757.

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

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American statistical Association*, *57*(298), 348–368.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, *38*(2), 894–942. doi: 10.1214/09-AOS729

Zhang, C.-H., & Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *76*(1), 217–242.

Zhang, H., Zheng, Y., Hou, L., Zheng, C., & Liu, L. (2021). Mediation analysis for survival data with high-dimensional mediators. *Bioinformatics*, *37*(21), 3815–3821.

Zhang, H., Zheng, Y., Zhang, Z., Gao, T., Joyce, B., Yoon, G., ... Liu, L. (2016). Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics*, *32*(20), 3150–3154.

Zhang, Q. (2021). High-dimensional mediation analysis with applications to causal gene identification. *Statistics in biosciences*, 1–20.

Zhang, R., He, X., Liu, W., Lu, M., Hsieh, J.-T., & Min, W. (2003). AIP1 mediates TNF-α–induced ASK1 activation by facilitating dissociation of ASK1 from its inhibitor 14-3-3. *The Journal of Clinical Investigation*, *111*(12), 1933–1943.

Zhang, X., Ibrahimi, O. A., Olsen, S. K., Umemori, H., Mohammadi, M., & Ornitz, D. M. (2006). Receptor specificity of the fibroblast growth factor family: the complete mammalian FGF family. *Journal of Biological Chemistry*, *281*(23), 15694–15700.

Zhao, A., & Ding, P. (2021). Covariate-adjusted fisher randomization tests for the average treatment effect. *Journal of Econometrics*, *225*(2), 278–294.

Zhao, Y., Lindquist, M. A., & Caffo, B. S. (2020). Sparse principal component based high-dimensional mediation analysis. *Computational Statistics & Data Analysis*, *142*, 106835.

Zhao, Y., & Luo, X. (2022). Pathway lasso: pathway estimation and selection with high-dimensional mediators. *Statistics and its interface*, *15*(1), 39.

Zhou, R. R., Wang, L., & Zhao, S. D. (2020). Estimation and inference for the indirect effect in high-dimensional linear mediation models. *Biometrika*, *107*(3), 573–589.

Zou, H., & Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, *36*(4), 1509.

Zyskind, G., & Martin, F. B. (1969). On best linear estimation and general Gauss-Markov theorem in linear models with arbitrary nonnegative covariance structure. *SIAM Journal on Applied Mathematics*, *17*(6), 1190–1202.

Appendix A

Appendix

In this Appendix, we provide technical proofs and additional simulation results for Chapter 4. Section A.1 presents the proofs for Theorem 4.4.1, with Section A.1.1 for the expression of $\mathbf{\Sigma}$, Section A.1.2 for the standard conditions for the proofs, and Section A.1.3 for the proofs of Theorem 4.4.1. Section A.2 presents the proofs for Theorem 4.4.2. Section A.3 presents the proofs for all the results in Section 4.6, with Section A.3.1 for Theorem 4.6.1, Section A.3.2 for Lemma 4.6.1, Section A.3.3 for Theorem 4.6.2, and Section A.3.4 for the algorithm to solve the overlapping group Lasso in (4.6.2). Section A.4 presents additional numerical results. Section A.5 presents the detected DNA methylation loci in Section 4.7 of the main paper together with the genes to which they belong.

Notation: Throughout this article, for sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \lesssim b_n$ and $a_n = O(b_n)$ if $a_n \leq Cb_n$ for sufficiently large n and C is independent of n. We write $a_n \ll b_n$ and $a_n = o(b_n)$ if $a_n/b_n \to 0$. Given an $n \times p$ matrix $\mathbf{M}$, let $\|\mathbf{M}\|$ and $\|\mathbf{M}\|_{\mathrm{F}}$ denote the operator norm and the Frobenius norm of $\mathbf{M}$, respectively. Let $\|\mathbf{M}\|_{2,\infty} = \sup_{\boldsymbol{\nu}:\|\boldsymbol{\nu}\|_2=1} \|\mathbf{M}\boldsymbol{\nu}\|_\infty$. For a square matrix $\mathbf{M} \in \mathbb{R}^{p\times p}$, let $\mathrm{diag}(\mathbf{M}) = [M_{11}, \ldots, M_{pp}]^{\mathrm{T}} \in \mathbb{R}^p$ denote the vector of all the diagonal entries in $\mathbf{M}$. Let $\mathbf{I}_p \in \mathbb{R}^{p\times p}$ denote the identity matrix. Let $\mathrm{tr}(\mathbf{M})$ denote the trace of $\mathbf{M}$. For matrices $\mathbf{A}_j \in \mathbb{R}^{m_j\times n_j}$ with $j = 1, \ldots, q$, let $\mathrm{diag}\{\mathbf{A}_j\}_{j=1}^q = \mathrm{diag}\{\mathbf{A}_1, \ldots, \mathbf{A}_q\} \in \mathbb{R}^{(m_1+\ldots m_q)\times(n_1+\cdots+n_q)}$ denote the block diagonal matrix with $\mathbf{A}_1, \ldots, \mathbf{A}_q$ lying along the diagonal and all the other entries are 0's. If $m_1 = \cdots = m_q = m$, let $\mathcal{R}\{\mathbf{A}_j\}_{j=1}^q \in \mathbb{R}^{m\times(n_1+\cdots+n_q)}$ denote the block matrix by placing $\mathbf{A}_j$, $j = 1, \ldots, q$ in one row. If $n_1 = \cdots = n_q = n$, let $\mathcal{C}\{\mathbf{A}_j\}_{j=1}^q \in \mathbb{R}^{(m_1+\cdots+m_q)\times n}$ denote the block matrix by placing $\mathbf{A}_j$, $j = 1, \ldots, q$ in one column. For matrices $\mathbf{A}_{kl} \in \mathbb{R}^{m_k\times n_l}$ with $k = 1, \ldots, p$ and $l = 1, \ldots, q$, let $\{\mathbf{A}_{kl}\}_{k=1,l=1}^{p,q} \in \mathbb{R}^{(m_1+\ldots m_p)\times(n_1+\cdots+n_q)}$ denote the block matrix by placing $\mathbf{A}_{kl}$ in the (k, l) position. We use C, C_1, C_2, $\ldots$ to denote positive and finite absolute constants that unless otherwise indicated can change from line to line.

A.1 Proofs of Theorem 4.4.1

A.1.1 The Expression of $\mathbf{\Sigma}$

For the ease of notation, we organize the treatment indicator matrix $\mathbf{T}$ as

$$\mathbf{T} = \left[\operatorname{diag}\left\{\mathbf{1}_{n_1}, \ldots, \mathbf{1}_{n_q}\right\}^{\mathrm{T}}, \mathbf{O}_{n_0 \times q}^{\mathrm{T}}\right]^{\mathrm{T}} \in \mathbb{R}^{n \times q}, \tag{A.1.1}$$

that is, if we denote $\mathcal{N}_j,\ j = 0, 1, \ldots, q$, the index sets of units receiving treatment level $\mathbf{t}_j$, then under the treatment groups in (A.1.1), $\mathcal{N}_1 = \{1, \ldots, n_1\}$, $\mathcal{N}_2 = \{n_1 + 1, \ldots, n_1 + n_2\}$, $\ldots$, $\mathcal{N}_q = \{\sum_{j=1}^{q-1} n_j + 1, \ldots, \sum_{j=1}^{q} n_j\}$, and $\mathcal{N}_0 = \{\sum_{j=1}^{q} n_j + 1, \ldots, n\}$, such that $|\mathcal{N}_j| = n_j$.

Recalling $\mathbf{W}_{\mathcal{A}} = [\mathbf{1}_n, \mathbf{T}, \mathbf{X}, \mathbf{G}_{\mathcal{A}_G}, \mathbf{T} : \mathbf{G}_{\mathcal{A}_D}] \in \mathbb{R}^{n \times (1+q+r+s)}$ defined in Section 4.4.1. We can split $\mathbf{W}_{\mathcal{A}}$ into four blocks: $[\mathbf{1}_n, \mathbf{T}]$, $\mathbf{X}$, $\mathbf{G}_{\mathcal{A}_G}$, and $\mathbf{T} : \mathbf{G}_{\mathcal{A}_D}$, where

$$\mathbf{T} : \mathbf{G}_{\mathcal{A}_D} = \left[\operatorname{diag}\left\{\mathbf{G}_{\mathcal{N}_1 \times \mathcal{A}_{D_1}}, \ldots, \mathbf{G}_{\mathcal{N}_q \times \mathcal{A}_{D_q}}\right\}^{\mathrm{T}}, \mathbf{O}_{n_0 \times (s_{D_1} + \cdots + s_{D_q})}^{\mathrm{T}}\right]^{\mathrm{T}} \in \mathbb{R}^{n \times (s_{D_1} + \cdots + s_{D_q})},$$

$\mathbf{G}_{\mathcal{N}_j \times \mathcal{A}_{D_j}} \in \mathbb{R}^{n_j \times s_{D_j}}, j = 1, \ldots, q$ are sub-matrices of $\mathbf{G}$ which consists of all the rows receiving treatment level $\mathbf{t}_j$ and all the columns in $\mathcal{A}_{D_j}$. Then both $n^{-1}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\mathbf{W}_{\mathcal{A}}$ and its limit $\mathbf{\Sigma} \in \mathbb{R}^{(1+q+r+s) \times (1+q+r+s)}$ have 4×4 blocks, *i.e.* ,

$$\frac{1}{n}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\mathbf{W}_{\mathcal{A}} = \left\{\widehat{\mathbf{\Sigma}}_{kl}\right\}_{k=1,l=1}^{4,4} \xrightarrow{P} \mathbf{\Sigma} = \left\{\mathbf{\Sigma}_{kl}\right\}_{k=1,l=1}^{4,4}. \tag{A.1.2}$$

The specific forms are listed in Table A.3.

By applying the paired row and column operations which uses the diagonal blocks of $\mathbf{\Sigma}$ to remove the off-diagonal blocks, we have the following result:

Remark A.1.1. *Given $p_0 + p_1 + \cdots + p_q = 1$, if $\mathbf{\Sigma}$ from Table A.3 is positive definite for arbitrary $\boldsymbol{\mu}_X \in \mathbb{R}^r$ and $\mathbf{A} \in \mathbb{R}^{p\times r}$, it is equivalent to*

$$\begin{cases} p_0, p_1, \ldots, p_q > 0; \\ \mathbf{\Sigma}_X \in \mathbb{R}^{r\times r} \text{ is positive definite;} \\ \mathbf{\Sigma}_{\mathbf{G}(\mathcal{A}_G\times\mathcal{A}_G)} - \sum_{j=1}^{q} p_j \mathbf{\Sigma}_{\mathbf{G}(\mathcal{A}_G\times\mathcal{A}_{D_j})} \mathbf{\Sigma}^{-1}_{\mathbf{G}(\mathcal{A}_{D_j}\times\mathcal{A}_{D_j})} \mathbf{\Sigma}_{\mathbf{G}(\mathcal{A}_{D_j}\times\mathcal{A}_G)} \text{ is positive definite.} \end{cases} \tag{A.1.3}$$

Remark A.1.2. *Notice that if $\mathbf{X}_i$'s are not presented in model* (4.2.1) *and* (4.2.2)*, then*

$$\frac{\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\mathbf{W}_{\mathcal{A}}}{n} \xrightarrow{P} \mathbf{\Sigma} = \{\mathbf{\Sigma}_{kl}(\mathbf{A} = \mathbf{O})\}_{k=1,3,4;l=1,3,4} \in \mathbb{R}^{(1+q+s)\times(1+q+s)},$$

where $\mathbf{\Sigma}$ above is removing $\mathbf{\Sigma}_{j2}$ and $\mathbf{\Sigma}_{2j}$ for $j = 1, \ldots, 4$ in Table A.3, and $\mathbf{\Sigma}_{kl}(\mathbf{A} = \mathbf{O})$ is $\mathbf{\Sigma}_{kl}$ when $\mathbf{A} = \mathbf{O}$ for $k, l \in \{1, 3, 4\}$.

A.1.2 Conditions

Condition A.1.1. *For simplicity, in this condition, we use λ instead of λ_n in $p_{\lambda_n}()$. Assume that $p_\lambda(t_0)$ is increasing and concave in $t_0 \in [0, \infty)$ and the derivative $p'_\lambda(t_0) = \partial p_\lambda(t_0)/\partial t_0$ is continuous on $t_0 \in [0, \infty)$ with $p'_\lambda(0+) > 0$. Define $\rho(t_0, \lambda) = p_\lambda(t_0)/\lambda$ for $\lambda > 0$. Then $\rho'(t_0, \lambda) = \partial\rho(t_0, \lambda)/\partial t_0 = \lambda^{-1}p'_\lambda(t_0)$ is increasing in $\lambda \in (0, \infty)$ for any $t_0 \in [0, \infty)$ and $\rho'(0+, \lambda)$ does not depend on λ.*

Condition A.1.2. *Suppose that q is fixed. Let n_j for $j = 0, 1, \ldots, q$ be the number of units receiving treatment $\mathbf{t}_j$. The proportions of treatments satisfy $|n_j/n - p_j| \le 1/\sqrt{n}$ with $p_j > 0$ for $j = 0, 1, \ldots, q$ such that $\sum_{j=0}^{q} p_j = 1$. Assume that the dimension of the mediators has $p > 1$ and p is allowed to increase with n or to be greater than n.*

Condition A.1.3. *Assume that $\lambda_{\min}(\mathbf{\Sigma}) > C_\Sigma > 0$ for some absolute constant $C_\Sigma > 0$ and $\|\mathbf{W}_{\mathcal{A}^c}^{\mathrm{T}}\mathbf{W}_{\mathcal{A}}\|_{2,\infty} = O_p(n)$.*

Condition A.1.4. *Let* $d_n = \left(\min_{h\in\mathcal{A}_G}\{|\theta^\star_{G,h}|\} \wedge \min_{j=1,\ldots,q,h\in\mathcal{A}_{D_j}}\{|\theta^\star_{D_j,h}|\}\right)/2$ *and* $d_n \gg \lambda_n \gg \max\left\{\sqrt{s/n}, \sqrt{\log p/n}\right\}$*. Suppose* $\lambda_n\kappa_0 = o(1)$ *where* $\kappa_0 = \max_{\boldsymbol{\nu}^{\mathrm{T}}\in\mathcal{V}_0}\kappa(\rho,\boldsymbol{\nu},\lambda_n)$ *and*

$$\mathcal{V}_0 = \left\{\boldsymbol{\nu}\in\mathbb{R}^s : \|\boldsymbol{\nu} - [\boldsymbol{\theta}^{\star T}_{G,\mathcal{A}_G}\ \boldsymbol{\theta}^{\star T}_{D_1,\mathcal{A}_{D_1}}\ \ldots\ \boldsymbol{\theta}^{\star T}_{D_q,\mathcal{A}_{D_q}}]^{\mathrm{T}}\|_2 \le d_n\right\},$$
$$\kappa(\rho,\boldsymbol{\nu},\lambda_n) = \lim_{\epsilon\to 0+}\max_{1\le k\le s}\sup_{t_1<t_2\in(|\nu_k|-\epsilon,|\nu_k|+\epsilon)} -\frac{\rho'(t_2,\lambda_n)-\rho'(t_1,\lambda_n)}{t_2-t_1}.$$

Condition A.1.5. *Suppose* $\mathbb{E}(X^4_{im}) \le C_X$ *for* $i = 1,\ldots,n$ *and* $m = 1,\ldots,r$*, where* $C_X > 0$ *is an absolute constant. Suppose* $\mathbb{E}[(\beta_{0,h} + \beta_{j,h} + \mathbf{A}_{(h)}\mathbf{X}_i + \epsilon_{G_i,h})^4] \le C_G$ *and* $\mathbb{E}[(\beta_{0,h} + \mathbf{A}_{(h)}\mathbf{X}_i + \epsilon_{G_i,h})^4] \le C_G$ *for* $j = 1,\ldots,q$ *and* $h = 1,\ldots,p$*, where* $\beta_{0,h}$*,* $\beta_{j,h}$*, and* $\epsilon_{G_i,h}$ *are the* h*-th coordinate of* $\boldsymbol{\beta}_0$*,* $\boldsymbol{\beta}_j$ *and* $\boldsymbol{\varepsilon}_{G_i}$*, respectively,* $\mathbf{A}_{(h)}\in\mathbb{R}^{1\times r}$ *is the* h*-th row of* $\mathbf{A}$*, and* $C_G > 0$ *is an absolute constant. For some* $\varpi > 2$*, there exists a positive sequence* $\{K_n\}_{n=1}^\infty$ *such that* $K_n^2\log p/n^{1-2/\varpi-\varphi} \longrightarrow 0$ *for some* $\varphi > 0$*. Further, suppose* $\mathbb{E}\|\mathbf{G}_{i,\mathcal{A}^c_\cap}\epsilon_{Y_i}\|^\varpi_\infty \le K_n^\varpi$ *for any* $i = 1,\ldots,n$*, where* $\mathcal{A}^c_\cap = (\{1,\ldots,p\}\setminus\mathcal{A}_G)\cup\left(\cup_{j=1}^q\{1,\ldots,p\}\setminus\mathcal{A}_{D_j}\right)$*,* $\mathbf{G}_{i,\mathcal{A}^c_\cap}$ *is sub-vector of* $\mathbf{G}_i$*, whose coordinates are in* $\mathcal{A}^c_\cap$*. Suppose* $\mathbb{E}\epsilon^4_{Y_i} < C_{EY}$ *for an absolute constant* $C_{EY} > 0$*.*

Condition A.1.1 is the standard condition for the penalty function (Fan & Lv, 2011; Guo et al., 2023). Condition A.1.2 highlights the relatively balanced group sizes across different treatment levels and the focus on the high-dimensional regime. Condition A.1.3 is similar to the lower eigenvalue and mutual incoherence conditions commonly assumed in Lasso (Wainwright, 2019). Condition A.1.4 imposes the minimal signal condition. Condition A.1.5 imposes the moment conditions necessary to bound $\lambda_{\max}(\boldsymbol{\Sigma})$ and $\mathrm{tr}(\boldsymbol{\Sigma})$, to achieve the $O_p(\sqrt{s/n})$ convergence rate, and to establish sparsity.

A.1.3 Proof of Theorem 4.4.1

Follow the general idea of the proof of Theorem 1 in Guo et al. (2023), we divide the proof into three steps. In **Step 1**, we show the existence of the local minimizer $\overline{\boldsymbol{\theta}}_{W,\mathcal{A}}\in\mathbb{R}^{1+q+r+s}$ for $\boldsymbol{\theta}_{W,\mathcal{A}} = [\theta_0, \boldsymbol{\theta}^{\mathrm{T}}_T, \boldsymbol{\theta}^{\mathrm{T}}_X, \boldsymbol{\theta}^{\mathrm{T}}_{G,\mathcal{A}_G}, \boldsymbol{\theta}^{\mathrm{T}}_{D_1,\mathcal{A}_{D_1}}, \ldots, \boldsymbol{\theta}^{\mathrm{T}}_{D_q,\mathcal{A}_{D_q}}]^{\mathrm{T}}$ such that $\|\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} - \boldsymbol{\theta}^\star_{W,\mathcal{A}}\|_2 = O_p(\sqrt{s/n})$, when the global objective function (4.3.1) is constrained with $\boldsymbol{\theta}_{W,\mathcal{A}^c} = \mathbf{0}$. In **Step 2**, we show that extending $\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} \in \mathbb{R}^{1+q+r+s}$ from **Step 1** with $(q+1)p-s$ zeros in the location of $\mathcal{A}^c$ forms a local minimizer of the global objective function (4.3.1), *i.e.* , $\widehat{\boldsymbol{\theta}}^{\star}_{W}$ in (4.3.1) has $\widehat{\boldsymbol{\theta}}^{\star}_{W,\mathcal{A}} = \overline{\boldsymbol{\theta}}_{W,\mathcal{A}}$ and $\widehat{\boldsymbol{\theta}}^{\star}_{W,\mathcal{A}^c} = \mathbf{0}_{(q+1)p-s}$. In **Step 3**, we develop the asymptotic expansion $\sqrt{n}(\widehat{\boldsymbol{\theta}}^{\star}_{W,\mathcal{A}} - \boldsymbol{\theta}^{\star}_{W,\mathcal{A}}) = (1/\sqrt{n})\boldsymbol{\Sigma}^{-1}\mathbf{W}^{\mathrm{T}}_{\mathcal{A}}\boldsymbol{\varepsilon}_Y + o_p(\mathbf{1})$.

Step 1. $\|\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} - \boldsymbol{\theta}^{\star}_{W,\mathcal{A}}\|_2 = O_p(\sqrt{s/n})$: In this step, we constrain the loss function (4.3.1) on the $(1+q+r+s)$-dimensional support: $\{\boldsymbol{\theta}_W \in \mathbb{R}^{1+q+r+p+pq} : \boldsymbol{\theta}_{\mathcal{A}^c} = \mathbf{0}_{(q+1)p-s}\}$. So we focus on the minimization of the following constrained partially penalized least squares function

$$\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} = \underset{\boldsymbol{\theta}_{W,\mathcal{A}} \in \mathbb{R}^{1+q+r+s}}{\arg\min}\, \overline{Q}(\boldsymbol{\theta}_{W,\mathcal{A}}), \tag{A.1.4}$$

where

$$\overline{Q}(\boldsymbol{\theta}_{W,\mathcal{A}}) = \frac{1}{2n}\|\mathbf{Y} - \mathbf{W}_{\mathcal{A}}\boldsymbol{\theta}_{W,\mathcal{A}}\|_2^2 + \sum_{h\in\mathcal{A}_G} p_{\lambda_n}(|\theta_{G,h}|) + \sum_{j=1}^{q}\sum_{h\in\mathcal{A}_{D_j}} p_{\lambda_n}(|\theta_{D_j,h}|).$$

We will show that there exists a local minimizer $\overline{\boldsymbol{\theta}}_{W,\mathcal{A}}$ such that $\|\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} - \boldsymbol{\theta}^{\star}_{W,\mathcal{A}}\|_2 = O_p(\sqrt{s/n})$, *i.e.* , for any given $\tau > 0$, we will show that $\mathbb{P}\{\forall\overline{\boldsymbol{\theta}}_{W,\mathcal{A}},\ \|\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} - \boldsymbol{\theta}^{\star}_{W,\mathcal{A}}\|_2 \geq \tau\sqrt{s/n}\} \leq C/\tau^2 + o(1)$ with an absolute constant C. Since $\{\forall\overline{\boldsymbol{\theta}}_{W,\mathcal{A}},\ \|\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} - \boldsymbol{\theta}^{\star}_{W,\mathcal{A}}\|_2 \geq \tau\sqrt{s/n}\} \subset H_n$, where

$$H_n = \left\{\min_{\boldsymbol{\theta}_{W,\mathcal{A}}} \overline{Q}(\boldsymbol{\theta}_{W,\mathcal{A}}) \leq \overline{Q}(\boldsymbol{\theta}^{\star}_{W,\mathcal{A}}) \text{ where } \|\boldsymbol{\theta}_{W,\mathcal{A}} - \boldsymbol{\theta}^{\star}_{W,\mathcal{A}}\|_2 = \tau\sqrt{s/n}\right\}, \tag{A.1.5}$$

in the following parts of **Step 1**, we will bound $\mathbb{P}\{H_n\}$.

After applying the Taylor expansion of $\overline{Q}(\boldsymbol{\theta}_{W,\mathcal{A}})$ at $\boldsymbol{\theta}^{\star}_{W,\mathcal{A}}$, we have

$$\overline{Q}(\boldsymbol{\theta}_{W,\mathcal{A}}) - \overline{Q}(\boldsymbol{\theta}^{\star}_{W,\mathcal{A}}) = -\boldsymbol{\nu}^{\mathrm{T}}\left(\boldsymbol{\theta}_{W,\mathcal{A}} - \boldsymbol{\theta}^{\star}_{W,\mathcal{A}}\right) + \frac{1}{2}\left(\boldsymbol{\theta}_{W,\mathcal{A}} - \boldsymbol{\theta}^{\star}_{W,\mathcal{A}}\right)^{\mathrm{T}}\overline{\mathbf{D}}\left(\boldsymbol{\theta}_{W,\mathcal{A}} - \boldsymbol{\theta}^{\star}_{W,\mathcal{A}}\right), \tag{A.1.6}$$

where

$$\boldsymbol{\nu} \;=\; -\left.\frac{\partial \overline{Q}}{\partial \boldsymbol{\theta}_{W,\mathcal{A}}}\right|_{\boldsymbol{\theta}^{\star}_{W,\mathcal{A}}} = \boldsymbol{\nu}_1 - \boldsymbol{\nu}_2,$$

$$\boldsymbol{\nu}_1 \;=\; \frac{1}{n}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\left(\mathbf{Y} - \mathbf{W}_{\mathcal{A}}\boldsymbol{\theta}^{\star}_{W,\mathcal{A}}\right), \quad \boldsymbol{\nu}_2 = \begin{bmatrix}\mathbf{0}^{\mathrm{T}}_{1+q+r} & \boldsymbol{\nu}_3^{\mathrm{T}}(\boldsymbol{\theta}^{\star}_{\mathcal{A}_{G,D}})\end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^{1+q+r+s}. \quad \text{(A.1.7)}$$

For $\boldsymbol{\nu}_3(\boldsymbol{\theta}^{\star}_{\mathcal{A}_{G,D}})$ in (A.1.7), we first define

$$\boldsymbol{\theta}_{\mathcal{A}_{G,D}} = [\boldsymbol{\theta}^{\mathrm{T}}_{G,\mathcal{A}_G} \; \boldsymbol{\theta}^{\mathrm{T}}_{D_1,\mathcal{A}_{D_1}} \; \ldots \; \boldsymbol{\theta}^{\mathrm{T}}_{D_q,\mathcal{A}_{D_q}}]^{\mathrm{T}} \in \mathbb{R}^{s}, \quad \text{(A.1.8)}$$

which is the vector of all penalized quantities in $\overline{Q}(\boldsymbol{\theta}_{W,\mathcal{A}})$, and $\boldsymbol{\theta}^{\star}_{\mathcal{A}_{G,D}}$ is the corresponding true value. The vector-valued function $\boldsymbol{\nu}_3(\boldsymbol{\theta}_{\mathcal{A}_{G,D}}) \in \mathbb{R}^s$ has entries $\partial p_{\lambda_n}(|t|)/\partial t$ where t is the entry in $\boldsymbol{\theta}_{\mathcal{A}_{G,D}}$ and $\boldsymbol{\nu}_3(\boldsymbol{\theta}^{\star}_{\mathcal{A}_{G,D}}) \in \mathbb{R}^s$ in (A.1.7) is $\boldsymbol{\nu}_3(\boldsymbol{\theta}_{\mathcal{A}_{G,D}})$ evaluated at the true value $\boldsymbol{\theta}^{\star}_{\mathcal{A}_{G,D}}$.

In the second term in the Taylor expansion (A.1.6),

$$\overline{\mathbf{D}} = \frac{1}{n}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\mathbf{W}_{\mathcal{A}} + \begin{bmatrix}\mathbf{O}_{(1+q+r)\times(1+q+r)} & \mathbf{O}_{(1+q+r)\times s}\\ \mathbf{O}_{s\times(1+q+r)} & \overline{\mathbf{D}}_1\end{bmatrix},$$

where $\overline{\mathbf{D}}_1 \in \mathbb{R}^{s\times s}$ is a diagonal matrix whose diagonal entries are $2\{p_{\lambda_n}(|t|) - p_{\lambda_n}(|t^\star|) - [\partial p_{\lambda_n}(|t|)/\partial t|_{t=t^\star}](t-t^\star)\}/(t-t^\star)^2$, t is the entry in $\boldsymbol{\theta}_{\mathcal{A}_{G,D}}$ in (A.1.8), and $t^\star$ is the corresponding true values of t, *i.e.* , the values in $\boldsymbol{\theta}^{\star}_{\mathcal{A}_{G,D}}$. If $\partial^2 p_{\lambda_n}(|t|)/\partial t^2$ exists in $[t, t^\star]$ or $[t^\star, t]$, then the diagonal entry of $\overline{\mathbf{D}}_1$, *i.e.* , $2\{p_{\lambda_n}(|t|) - p_{\lambda_n}(|t^\star|) - [\partial p_{\lambda_n}(|t|)/\partial t|_{t=t^\star}](t-t^\star)\}/(t-t^\star)^2$ is just $\partial^2 p_{\lambda_n}(|t|)/\partial t^2$ evaluated at some value between t and $t^\star$ by the Taylor expansion.

The following proofs consist of (a) bounding $\lambda_{\min}(\overline{\mathbf{D}})$ away from zero; (b) showing that $\mathbb{E}\|\boldsymbol{\nu}\|_2^2 = O(s/n)$; (c) combing (a) and (b) to bound $\mathbb{P}\{H_n\}$ from (A.1.5)

(a) Bound $\lambda_{\min}(\overline{\mathbf{D}})$ in (A.1.6):

By (A.1.2), we have $\mathbb{E}\|n^{-1}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\mathbf{W}_{\mathcal{A}} - \boldsymbol{\Sigma}\|_{\mathrm{F}}^2 = \sum_{k=1}^{4}\sum_{l=1}^{4}\mathbb{E}\|\widehat{\boldsymbol{\Sigma}}_{kl} - \boldsymbol{\Sigma}_{kl}\|_{\mathrm{F}}^2$, with $k, l \in \{1,2,3,4\}$. By $|n_j/n - p_j| \le n^{-1/2}$ and fixed q in Condition A.1.2, we have $\mathbb{E}\|\widehat{\boldsymbol{\Sigma}}_{11} - \boldsymbol{\Sigma}_{11}\|_{\mathrm{F}}^2 \le O(1/n)$. By the bounded fourth moments of the entries in $\mathbf{X}_i$ and $\mathbf{G}_i$ in

Condition A.1.5 and $r \leq C \cdot s$ in Theorem 4.4.1, by Lemma A.1.1, we have $\mathbb{E}\|\widehat{\boldsymbol{\Sigma}}_{1l} - \boldsymbol{\Sigma}_{1l}\|_{\mathrm{F}}^2 \leq O(s/n)$ for $l = 2, 3, 4$, and $\mathbb{E}\|\widehat{\boldsymbol{\Sigma}}_{kl} - \boldsymbol{\Sigma}_{kl}\|_{\mathrm{F}}^2 \leq O(s^2/n)$ for $k, l \in \{2, 3, 4\}$. Then we have

$$\mathbb{E}\|n^{-1}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\mathbf{W}_{\mathcal{A}} - \boldsymbol{\Sigma}\|_{\mathrm{F}}^2 \leq O(s^2/n), \quad \|n^{-1}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\mathbf{W}_{\mathcal{A}} - \boldsymbol{\Sigma}\|_{\mathrm{F}} = O_p(s/\sqrt{n}). \tag{A.1.9}$$

For the diagonal entries in $\overline{\mathbf{D}}_1$, from Condition A.1.4, for any $\boldsymbol{\theta}_{W,\mathcal{A}}$ in event H_n, it satisfies $\|\boldsymbol{\theta}_{\mathcal{A}_{G,D}} - \boldsymbol{\theta}^{\star}_{\mathcal{A}_{G,D}}\|_2 \leq \|\boldsymbol{\theta}_{W,\mathcal{A}} - \boldsymbol{\theta}^{\star}_{W,\mathcal{A}}\|_2 = \tau\sqrt{s/n} \ll \lambda_n \ll d_n$ where $\tau > 0$ is given in (A.1.5). Hence we have $\boldsymbol{\theta}_{\mathcal{A}_{G,D}} \in \mathcal{V}_0$ defined in Condition A.1.4. By $\lambda_n\kappa_0 = o(1)$, the diagonal entries of $\overline{\mathbf{D}}_1$ have absolute values of order $o(1)$. Since $\overline{\mathbf{D}}_1$ is a diagonal matrix, we have

$$|\lambda_{\min}(\overline{\mathbf{D}}_1)| = o(1). \tag{A.1.10}$$

Therefore, we have $\lambda_{\min}(\overline{\mathbf{D}}) \geq C_\Sigma/4$ with overwhelming probability, where C_Σ is from Condition A.1.3. Specifically, define

$$\mathcal{E}_\Sigma = \{|\lambda_{\min}(n^{-1}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\mathbf{W}_{\mathcal{A}}) - \lambda_{\min}(\boldsymbol{\Sigma})| \leq C_{\boldsymbol{\Sigma}}/2\}, \tag{A.1.11}$$

so we have

$$\begin{aligned}
\mathbb{P}\{\mathcal{E}_\Sigma^c\} &\leq \mathbb{P}\left\{\|n^{-1}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\mathbf{W}_{\mathcal{A}} - \boldsymbol{\Sigma}\| > C_\Sigma/2\right\} \leq \mathbb{P}\left\{\|n^{-1}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\mathbf{W}_{\mathcal{A}} - \boldsymbol{\Sigma}\|_{\mathrm{F}} > C_\Sigma/2\right\} \\
&\leq 4\mathbb{E}\|n^{-1}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\mathbf{W}_{\mathcal{A}} - \boldsymbol{\Sigma}\|_{\mathrm{F}}^2/C_\Sigma^2 \leq O(s^2/n), \qquad (A.1.12)
\end{aligned}$$

where the first inequality follows from the Weyl's inequality and the last inequality follows from (A.1.9). Therefore, we have

$$\mathbb{P}\left\{\lambda_{\min}\left(\overline{\mathbf{D}}\right) < \frac{C_\Sigma}{4}\right\} \leq \mathbb{P}\left\{\lambda_{\min}\left(\overline{\mathbf{D}}\right) < \frac{C_\Sigma}{4}, \mathcal{E}_\Sigma\right\} + \mathbb{P}\left\{\lambda_{\min}\left(\overline{\mathbf{D}}\right) < \frac{C_\Sigma}{4}, \mathcal{E}_\Sigma^c\right\}$$

$$
\begin{aligned}
&\le \mathbb{P}\left\{\lambda_{\min}\left(\frac{\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\mathbf{W}_{\mathcal{A}}}{n}\right)+\lambda_{\min}\left(\begin{bmatrix}\mathbf{O} & \mathbf{O}\\ \mathbf{O} & \overline{\mathbf{D}}_1\end{bmatrix}\right)<\frac{C_\Sigma}{4},\mathcal{E}_\Sigma\right\}+\mathbb{P}\left\{\mathcal{E}_\Sigma^c\right\}\\
&\le \mathbb{P}\left\{\lambda_{\min}\left(\mathbf{\Sigma}\right)-\frac{C_\Sigma}{2}+\lambda_{\min}\left(\begin{bmatrix}\mathbf{O} & \mathbf{O}\\ \mathbf{O} & \overline{\mathbf{D}}_1\end{bmatrix}\right)<\frac{C_\Sigma}{4}\right\}+O\left(\frac{s^2}{n}\right)\\
&\le \mathbb{P}\left\{\frac{C_\Sigma}{2}+\lambda_{\min}\left(\overline{\mathbf{D}}_1\right)<\frac{C_\Sigma}{4}\right\}+O\left(\frac{s^2}{n}\right) \qquad \text{(A.1.13)}\\
&\le O\left(\frac{s^2}{n}\right), \qquad \text{(A.1.14)}
\end{aligned}
$$

where (A.1.13) follows from $\lambda_{\min}(\mathbf{\Sigma}) \ge C_\Sigma$ in Condition A.1.3. In (A.1.14), if n is sufficiently large, by (A.1.10), we have $\mathbb{P}\{C_\Sigma/2+\lambda_{\min}(\overline{\mathbf{D}}_1) < C_{\mathbf{\Sigma}}/4\} = 0$ so that (A.1.14) holds. By $s = o(\sqrt{n})$ assumed in Theorem 4.4.1, we have $\mathbb{P}\{\lambda_{\min}(\overline{\mathbf{D}}) \ge C_\Sigma/4\} \ge 1 - O(s^2/n) \to 1$.

(b) Show that the $\mathbb{E}\|\boldsymbol{\nu}\|_2^2 = O(s/n)$ with $\boldsymbol{\nu}$ in (A.1.6). Since $\|\boldsymbol{\nu}\|_2 = \|\boldsymbol{\nu}_1 - \boldsymbol{\nu}_2\|_2 \le \|\boldsymbol{\nu}_1\|_2 + \|\boldsymbol{\nu}_2\|_2$, we bound $\boldsymbol{\nu}_1$ and $\boldsymbol{\nu}_2$ separately.

For $\boldsymbol{\nu}_1 = n^{-1}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\varepsilon}_Y$, we have

$$
\begin{aligned}
\mathbb{E}\|\boldsymbol{\nu}_1\|_2^2 &= \frac{1}{n^2}\mathbb{E}\|\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\varepsilon}_Y\|_2^2 = \frac{1}{n^2}\mathbb{E}\left[\operatorname{tr}(\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\varepsilon}_Y\boldsymbol{\varepsilon}_Y^{\mathrm{T}}\mathbf{W}_{\mathcal{A}})\right] = \frac{1}{n}\sigma_Y^2\mathbb{E}\left[\operatorname{tr}\left(\frac{\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\mathbf{W}_{\mathcal{A}}}{n}\right)\right]\\
&= \frac{1}{n}\sigma_Y^2\{\operatorname{tr}[\mathbb{E}(\widehat{\mathbf{\Sigma}}_{11})]+\operatorname{tr}[\mathbb{E}(\widehat{\mathbf{\Sigma}}_{22})]+\operatorname{tr}[\mathbb{E}(\widehat{\mathbf{\Sigma}}_{33})]+\operatorname{tr}[\mathbb{E}(\widehat{\mathbf{\Sigma}}_{44})]\}\\
&\le C\frac{s}{n}, \qquad \text{(A.1.15)}
\end{aligned}
$$

where the last step follows from the bounded moment conditions in Condition A.1.5 and $r \le C \cdot s$ in Theorem 4.4.1. The term σ_Y^2 is also absorbed by C.

For $\boldsymbol{\nu}_2$ defined in (A.1.7), by Condition A.1.1, $p'_{\lambda_n}(|t|)$ is non-negative and is decreasing as $|t|$ increases. So we have $\|\boldsymbol{\nu}_2\|_2 \le \sqrt{sp'^2_{\lambda_n}(d_n)} = \sqrt{s}p'_{\lambda_n}(d_n)$. By $p'_{\lambda_n}(d_n) = o(1/\sqrt{ns})$ assumed in Theorem 4.4.1, we have $\|\boldsymbol{\nu}_2\|_2 \le o(1/\sqrt{n})$.

Combining the bounds of $\|\boldsymbol{\nu}_1\|_2$ and $\|\boldsymbol{\nu}_2\|_2$, we have

$$\mathbb{E}\|\boldsymbol{\nu}\|_2^2 \leq \mathbb{E}\|\boldsymbol{\nu}_1\|_2^2 + \|\boldsymbol{\nu}_2\|_2^2 + 2\|\boldsymbol{\nu}_2\|_2\mathbb{E}\|\boldsymbol{\nu}_1\|_2 = O\left(\frac{s}{n}\right). \tag{A.1.16}$$

(c) Collecting (a) and (b) together, for any $\tau > 0$ in (A.1.5), we have

$$\begin{aligned}
& \mathbb{P}\left\{\forall\overline{\boldsymbol{\theta}}_{W,\mathcal{A}},\ \|\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} - \boldsymbol{\theta}^{\star}_{W,\mathcal{A}}\|_2 \geq \tau\sqrt{\frac{s}{n}}\right\} \leq \mathbb{P}\left\{H_n\right\} && \text{(A.1.17)}\\
\leq\ & \mathbb{P}\left\{H_n,\ \lambda_{\min}\left(\overline{\mathbf{D}}\right) \geq \frac{C_\Sigma}{4}\right\} + \mathbb{P}\left\{H_n,\ \lambda_{\min}\left(\overline{\mathbf{D}}\right) < \frac{C_\Sigma}{4}\right\} && \\
\leq\ & \mathbb{P}\left\{\|\boldsymbol{\nu}\|_2 \geq \frac{C_\Sigma}{8}\tau\sqrt{\frac{s}{n}}\right\} + \mathbb{P}\left\{\lambda_{\min}\left(\overline{\mathbf{D}}\right) < \frac{C_\Sigma}{4}\right\} && \text{(A.1.18)}\\
\leq\ & \frac{64n\mathbb{E}\|\boldsymbol{\nu}\|_2^2}{C_\Sigma^2\tau^2 s} + O\left(\frac{s^2}{n}\right) && \text{(A.1.19)}\\
\leq\ & \frac{C}{\tau^2} + o_n(1), && \text{(A.1.20)}
\end{aligned}$$

where (A.1.17) follows from the definition of H_n in (A.1.5). For (A.1.18), if $\|\boldsymbol{\theta}_{W,\mathcal{A}} - \boldsymbol{\theta}^{\star}_{W,\mathcal{A}}\|_2 = \tau\sqrt{s/n}$, by the Taylor expansion (A.1.6) and $\lambda_{\min}(\overline{\mathbf{D}}) \geq C_\Sigma/4$, we have

$$\begin{aligned}
\overline{Q}(\boldsymbol{\theta}_{W,\mathcal{A}}) - \overline{Q}(\boldsymbol{\theta}^{\star}_{W,\mathcal{A}}) &\geq -\|\boldsymbol{\nu}\|_2\|\boldsymbol{\theta}_{W,\mathcal{A}} - \boldsymbol{\theta}^{\star}_{W,\mathcal{A}}\|_2 + (C_\Sigma/8)\|\boldsymbol{\theta}_{W,\mathcal{A}} - \boldsymbol{\theta}^{\star}_{W,\mathcal{A}}\|_2^2\\
&= -\|\boldsymbol{\nu}\|_2\tau\sqrt{s/n} + (C_\Sigma\tau^2 s)/(8n).
\end{aligned}$$

Under event H_n, for $\|\boldsymbol{\theta}_{W,\mathcal{A}} - \boldsymbol{\theta}^{\star}_{W,\mathcal{A}}\|_2 = \tau\sqrt{s/n}$, we have $\min \overline{Q}(\boldsymbol{\theta}_{W,\mathcal{A}}) - \overline{Q}(\boldsymbol{\theta}^{\star}_{W,\mathcal{A}}) \leq 0$ so that $-\|\boldsymbol{\nu}\|_2\tau\sqrt{s/n} + (C_\Sigma\tau^2 s)/(8n) \leq 0$, *i.e.* , $\|\boldsymbol{\nu}\|_2 \geq (C_\Sigma/8)\tau\sqrt{s/n}$. (A.1.19) follows from Markov's inequality and (A.1.14) for sufficiently large n. (A.1.20) follows from (A.1.16) with C_Σ absorbed by C.

Therefore, we obtain $\|\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} - \boldsymbol{\theta}^{\star}_{W,\mathcal{A}}\|_2 = O_p(\sqrt{s/n})$, where $\overline{\boldsymbol{\theta}}_{W,\mathcal{A}}$ could be the local minimizer of $\overline{Q}(\boldsymbol{\theta}_{W,\mathcal{A}})$ closest to $\boldsymbol{\theta}^{\star}_{W,\mathcal{A}}$.

Step 2: Sparsity. In this step, we show that extending $\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} \in \mathbb{R}^{1+q+r+s}$ from **Step 1** with $(q+1)p - s$ zeros in the location of $\mathcal{A}^c$ forms a local minimizer $\overline{\boldsymbol{\theta}}_W \in \mathbb{R}^{1+q+r+p+pq}$ of (4.3.1). To this end, we will verify equation (7), (8), and (9) of Theorem 1 in Fan and Lv (2011).

By the fact that $\overline{\boldsymbol{\theta}}_{W,\mathcal{A}}$ is a local minimizer of $\overline{Q}(\boldsymbol{\theta}_{W,\mathcal{A}})$, equation (7) exists with $\lim_{n\to\infty}$ $\mathbb{P}\{\mathbf{W}_{\mathcal{A}}(\mathbf{Y}-\mathbf{W}\overline{\boldsymbol{\theta}}_W)-n[\mathbf{0}_{1+q+r}^{\mathrm{T}}\ \boldsymbol{\nu}_3^{\mathrm{T}}(\overline{\boldsymbol{\theta}}_{\mathcal{A}_{G,D}})]^{\mathrm{T}}=\mathbf{0}\}=1$. By Condition A.1.3 and (A.1.12), equation (9) exists with $\lim_{n\to\infty}\mathbb{P}\{\lambda_{\min}(\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\mathbf{W}_{\mathcal{A}})>n\lambda_n\kappa(\rho,\overline{\boldsymbol{\theta}}_{\mathcal{A}_{G,D}},\lambda_n)\}=1$.

For equation (8) of Theorem 1 in Fan and Lv (2011), we need to show that $\lim_{n\to\infty}$ $\mathbb{P}\{(n\lambda_n)^{-1}\|\mathbf{W}_{\mathcal{A}^c}^{\mathrm{T}}(\mathbf{Y}-\mathbf{W}\overline{\boldsymbol{\theta}}_W)\|_\infty\geq\rho'(0+,\lambda_n)\}=0$, where $\rho'(0+,\lambda_n)$ is from Condition A.1.1.

We have the following decomposition

$$\begin{aligned}
&\mathbf{W}_{\mathcal{A}^c}^{\mathrm{T}}(\mathbf{Y}-\mathbf{W}\overline{\boldsymbol{\theta}}_W)\\
=\;&\mathbf{W}_{\mathcal{A}^c}^{\mathrm{T}}(\mathbf{Y}-\mathbf{W}_{\mathcal{A}}\overline{\boldsymbol{\theta}}_{W,\mathcal{A}})=\mathbf{W}_{\mathcal{A}^c}^{\mathrm{T}}[(\mathbf{Y}-\mathbf{W}_{\mathcal{A}}\boldsymbol{\theta}_{W,\mathcal{A}}^{\star})-\mathbf{W}_{\mathcal{A}}(\overline{\boldsymbol{\theta}}_{W,\mathcal{A}}-\boldsymbol{\theta}_{W,\mathcal{A}}^{\star})]\\
=\;&\mathbf{W}_{\mathcal{A}^c}^{\mathrm{T}}\boldsymbol{\varepsilon}_Y-\mathbf{W}_{\mathcal{A}^c}^{\mathrm{T}}\mathbf{W}_{\mathcal{A}}(\overline{\boldsymbol{\theta}}_{W,\mathcal{A}}-\boldsymbol{\theta}_{W,\mathcal{A}}^{\star}).
\end{aligned}\tag{A.1.21}$$

For the second term, we have

$$\begin{aligned}
\|\mathbf{W}_{\mathcal{A}^c}^{\mathrm{T}}\mathbf{W}_{\mathcal{A}}(\overline{\boldsymbol{\theta}}_{W,\mathcal{A}}-\boldsymbol{\theta}_{W,\mathcal{A}}^{\star})\|_\infty &\leq \|\mathbf{W}_{\mathcal{A}^c}^{\mathrm{T}}\mathbf{W}_{\mathcal{A}}\|_{2,\infty}\|\overline{\boldsymbol{\theta}}_{W,\mathcal{A}}-\boldsymbol{\theta}_{W,\mathcal{A}}^{\star}\|_2\\
&\overset{\text{Condition }A.1.3}{=} O_p(\sqrt{sn}).
\end{aligned}\tag{A.1.22}$$

For the first term, denote

$$\mathbf{W}_{\mathcal{A}^c}=\begin{bmatrix}\mathcal{C}\left\{\mathbf{G}_{\mathcal{N}_j\times\mathcal{A}_G^c}\right\}_{j=1}^q & \operatorname{diag}\left\{\mathbf{G}_{\mathcal{N}_j\times\mathcal{A}_{D_j}^c}\right\}\\ \mathbf{G}_{\mathcal{N}_0\times\mathcal{A}_G^c} & \mathbf{O}_{n_0\times(qp-s_{D_1}-\cdots-s_{D_q})}\end{bmatrix},\quad \boldsymbol{\varepsilon}_Y=\begin{bmatrix}\mathcal{C}\left\{\boldsymbol{\varepsilon}_{Y,\mathcal{N}_j}\right\}_{j=1}^q\\ \boldsymbol{\varepsilon}_{Y,\mathcal{N}_0}\end{bmatrix},$$

where $\mathcal{N}_j$ is defined in (A.1.1). Therefore,

$$\|\mathbf{W}_{\mathcal{A}^c}^{\mathrm{T}}\boldsymbol{\varepsilon}_Y\|_\infty=\max\left\{\|\sum_{j=0}^q\mathbf{G}_{\mathcal{N}_j\times\mathcal{A}_G^c}^{\mathrm{T}}\boldsymbol{\varepsilon}_{Y,\mathcal{N}_j}\|_\infty,\|\mathbf{G}_{\mathcal{N}_1\times\mathcal{A}_{D_1}^c}^{\mathrm{T}}\boldsymbol{\varepsilon}_{Y,\mathcal{N}_1}\|_\infty,\ldots,\|\mathbf{G}_{\mathcal{N}_q\times\mathcal{A}_{D_q}^c}^{\mathrm{T}}\boldsymbol{\varepsilon}_{Y,\mathcal{N}_q}\|_\infty\right\}.\tag{A.1.23}$$

We first show that $\|\mathbf{G}^{\mathrm{T}}_{\mathcal{N}_j\times\mathcal{A}^c_{D_j}}\boldsymbol{\varepsilon}_{Y,\mathcal{N}_j}\|_\infty = O_p(\sqrt{n\log p})$, given $j=1,\ldots,q$. Specifically, denote the hth coordinate in $\mathbf{G}^{\mathrm{T}}_{\mathcal{N}_j\times\mathcal{A}^c_{D_j}}\boldsymbol{\varepsilon}_{Y,\mathcal{N}_j}\in\mathbb{R}^{p-s_{D_j}},\ h\in\{1,\ldots,p\}\setminus\mathcal{A}_{D_j}$ as $\sum_{i\in\mathcal{N}_j} g_{i,h}\epsilon_{Y_i}$, where $g_{i,h}\epsilon_{Y_i}$ are i.i.d. for $i\in\mathcal{N}_j$.

Let $a_n = n^{1/\varpi+\varphi/2}K_n$, $b=\sqrt{Mn\log p}$, where $M>0$ will be determined later. Write

$$
\begin{aligned}
g_{i,h}\epsilon_{Y_i} &= g_{i,h}\epsilon_{Y_i}\mathcal{I}\left\{|g_{i,h}\epsilon_{Y_i}|\le a_n\right\} - \mathbb{E}\left(g_{i,h}\epsilon_{Y_i}\mathcal{I}\left\{|g_{i,h}\epsilon_{Y_i}|\le a_n\right\}\right)\\
&+ g_{i,h}\epsilon_{Y_i}\mathcal{I}\left\{|g_{i,h}\epsilon_{Y_i}|> a_n\right\} - \mathbb{E}\left(g_{i,h}\epsilon_{Y_i}\mathcal{I}\left\{|g_{i,h}\epsilon_{Y_i}|> a_n\right\}\right)\\
&:= \epsilon_{i,h,1}+\epsilon_{i,h,2},
\end{aligned}
$$

so that

$$
\begin{aligned}
&\mathbb{P}\left\{\|\mathbf{G}^{\mathrm{T}}_{\mathcal{N}_j\times\mathcal{A}^c_{D_j}}\boldsymbol{\varepsilon}_{Y,\mathcal{N}_j}\|_\infty > b\right\} = \mathbb{P}\left\{\exists h\in\{1,\ldots,p\}\setminus\mathcal{A}_{D_j},\ \left|\sum_{i\in\mathcal{N}_j} g_{i,h}\epsilon_{Y_i}\right| > b\right\}\\
\le\ & \mathbb{P}\left\{\exists h\in\{1,\ldots,p\}\setminus\mathcal{A}_{D_j},\ \left|\sum_{i\in\mathcal{N}_j}\epsilon_{i,h,1}\right| + \left|\sum_{i\in\mathcal{N}_j}\epsilon_{i,h,2}\right| > b\right\}\\
\le\ & \mathbb{P}\left\{\exists h\in\{1,\ldots,p\}\setminus\mathcal{A}_{D_j},\ \left|\sum_{i\in\mathcal{N}_j}\epsilon_{i,h,1}\right| > \frac{b}{2}\right\}\\
+\ & \mathbb{P}\left\{\exists h\in\{1,\ldots,p\}\setminus\mathcal{A}_{D_j},\ \left|\sum_{i\in\mathcal{N}_j}\epsilon_{i,h,2}\right| > \frac{b}{2}\right\}\\
:=\ & \mathbb{P}_1+\mathbb{P}_2.
\end{aligned}
$$

For $\mathbb{P}_1$, given h, since $\epsilon_{i,h,1}$ are i.i.d. with zero mean and bounded by $2a_n$, we have

$$
\begin{aligned}
\mathbb{P}_1 &\le \sum_{h\in\{1,\ldots,p\}\setminus\mathcal{A}_{D_j}} \mathbb{P}\left\{\left|\sum_{i\in\mathcal{N}_j}\epsilon_{i,h,1}\right| > \frac{b}{2}\right\} && \text{(A.1.24)}\\
&\le 2(p-s_{D_j})\max_{h\in\{1,\ldots,p\}\setminus\mathcal{A}_{D_j}}\exp\left[\frac{-b^2/8}{\sum_{i\in\mathcal{N}_j}\mathbb{E}(\epsilon^2_{i,h,1})+2a_nb/6}\right] && \text{(A.1.25)}\\
&\le 2p\exp\left(\frac{-b^2/8}{n_j\sqrt{C_GC_{EY}}+a_nb/3}\right) && \text{(A.1.26)}\\
&\le \frac{2}{p^{M/\left(16\sqrt{C_GC_{EY}}\right)-1}}, && \text{(A.1.27)}
\end{aligned}
$$

where (A.1.24) follows from the union bound, (A.1.25) follows from the Bernstein's inequality, (A.1.26) follows from $\mathbb{E}(\epsilon_{i,h,1}^2) \leq \mathbb{E}g_{i,h}^2\epsilon_{Y_i}^2\mathcal{I}\{|g_{i,h}\epsilon_{Y_i}| \leq a_n\} \leq \mathbb{E}(g_{i,h}^2\epsilon_{Y_i}^2) \leq [\mathbb{E}(g_{i,h}^4)\mathbb{E}(\epsilon_{Y_i}^4)]^{1/2} \leq \sqrt{C_G C_{EY}}$. For (A.1.27), we have

$$\frac{b^2/8}{n_j\sqrt{C_G C_{EY}} + (a_n b)/3} = \frac{M\log p/8}{(n_j/n)\sqrt{C_G C_{EY}} + \frac{1}{3}\sqrt{M}\sqrt{(K_n^2\log p)/n^{1-2/\varpi-\varphi}}} \geq \frac{M\log p/8}{2\sqrt{C_G C_{EY}}},$$

where the last inequality holds when n is large enough such that $K_n^2\log p/n^{1-2/\varpi-\varphi} \longrightarrow 0$ by Condition A.1.5 and $n_j/n \leq 1$. Thus (A.1.27) can be controlled by M if $p > 1$.

For $\mathbb{P}_2$, for any $h \in \{1,\ldots,p\} \setminus \mathcal{A}_{D_j}$, we have

$$\begin{aligned}\left|\sum_{i\in\mathcal{N}_j}\epsilon_{i,h,2}\right| &\leq \sum_{i\in\mathcal{N}_j}\max_{h\in\{1,\ldots,p\}\setminus\mathcal{A}_{D_j}}\left(|g_{i,h}\epsilon_{Y_i}|\,\mathcal{I}\left\{|g_{i,h}\epsilon_{Y_i}| > a_n\right\}\right) \\ &+ n_j\max_{h\in\{1,\ldots,p\}\setminus\mathcal{A}_{D_j}}\mathbb{E}\left|g_{i,h}\epsilon_{Y_i}\right|\mathcal{I}\left\{|g_{i,h}\epsilon_{Y_i}| > a_n\right\}. \end{aligned} \tag{A.1.28}$$

In the last term, for any $h \in \{1,\ldots,p\} \setminus \mathcal{A}_{D_j}$, we have $\mathbb{E}|g_{i,h}\epsilon_{Y_i}|\mathcal{I}\{|g_{i,h}\epsilon_{Y_i}| > a_n\} \leq \sqrt{\mathbb{E}g_{i,h}^2\epsilon_{Y_i}^2}\sqrt{\mathbb{E}\mathcal{I}\{|g_{i,h}\epsilon_{Y_i}| > a_n\}} \leq \sqrt[4]{\mathbb{E}g_{i,h}^4\mathbb{E}\epsilon_{Y_i}^4}\sqrt{\mathbb{P}\{|g_{i,h}\epsilon_{Y_i}| > a_n\}} \leq \sqrt[4]{C_G C_{EY}}\sqrt{\mathbb{E}|g_{i,h}\epsilon_{Y_i}|^{\varpi}/a_n^{\varpi}} \leq \sqrt[4]{C_G C_{EY}}\sqrt{\mathbb{E}\max_{h\in\{1,\ldots,p\}\setminus\mathcal{A}_{D_j}}|g_{i,h}\epsilon_{Y_i}|^{\varpi}/a_n^{\varpi}} \leq \sqrt[4]{C_G C_{EY}}\sqrt{K_n^{\varpi}/a_n^{\varpi}}$, where the last step follows from Condition A.1.5. Recalling $a_n = n^{1/\varpi+\varphi/2}K_n$, we have $\mathbb{E}|g_{i,h}\epsilon_{Y_i}|\mathcal{I}\{|g_{i,h}\epsilon_{Y_i}| > a_n\} \leq \sqrt[4]{C_G C_{EY}}n^{-1/2}n^{-\varphi\varpi/4}$. Thus the last term in (A.1.28) satisfies

$$n_j\max_{h\in\{1,\ldots,p\}\setminus\mathcal{A}_{D_j}}\mathbb{E}\left|g_{i,h}\epsilon_{Y_i}\right|\mathcal{I}\left\{|g_{i,h}\epsilon_{Y_i}| > a_n\right\} \leq \frac{n_j}{n}\sqrt[4]{C_G C_{EY}}\sqrt{n}\frac{1}{n^{\varpi\varphi/4}} = o\left(\sqrt{n}\right). \tag{A.1.29}$$

Then for some $M > 0$ given in b, we have

$$\begin{aligned}\mathbb{P}_2 &\leq \mathbb{P}\left\{\sum_{i\in\mathcal{N}_j}\max_{h\in\{1,\ldots,p\}\setminus\mathcal{A}_{D_j}}\left(|g_{i,h}\epsilon_{Y_i}|\,\mathcal{I}\left\{|g_{i,h}\epsilon_{Y_i}| > a_n\right\}\right) > \frac{b}{4}\right\} \\ &+ \mathbb{P}\left\{n_j\max_{h\in\{1,\ldots,p\}\setminus\mathcal{A}_{D_j}}\mathbb{E}\left|g_{i,h}\epsilon_{Y_i}\right|\mathcal{I}\left\{|g_{i,h}\epsilon_{Y_i}| > a_n\right\} > \frac{b}{4}\right\} \\ &\overset{(A.1.29)}{\leq} \mathbb{P}\left\{\exists i\in\mathcal{N}_j, \max_{h\in\{1,\ldots,p\}\setminus\mathcal{A}_{D_j}}|g_{i,h}\epsilon_{Y_i}| > a_n\right\} + \mathbb{P}\left\{o\left(\sqrt{n}\right) > \frac{1}{4}\sqrt{Mn\log p}\right\}\end{aligned}$$

$$
\begin{aligned}
&\le \quad n_j \frac{\mathbb{E}\left(\max_{h\in\{1,\dots,p\}\backslash\mathcal{A}_{D_j}} |g_{i,h}\epsilon_{Y_i}|\right)^{\varpi}}{a_n^{\varpi}} + \mathbb{P}\left\{o\left(1\right) > \frac{1}{4}\sqrt{M\log p}\right\} \\
&\le \quad \frac{n_j}{n}\frac{K_n^{\varpi}}{K_n^{\varpi} n^{\varphi\varpi/2}} + 0 \overset{n\to\infty}{\longrightarrow} 0,
\end{aligned}
$$

where in the last inequality, the first probability follows from Condition A.1.5, and the second probability is zero if n is large enough. Combining $\mathbb{P}_1$ and $\mathbb{P}_2$, we have for $j = 1, \dots, q$, $\|\mathbf{G}^{\mathrm{T}}_{\mathcal{N}_j\times\mathcal{A}^c_{D_j}}\boldsymbol{\varepsilon}_{Y,\mathcal{N}_j}\|_\infty = O_p(\sqrt{n\log p})$.

For the first element in (A.1.23), we have

$$
\begin{aligned}
\mathbb{P}\left\{\|\sum_{j=0}^{q}\mathbf{G}^{\mathrm{T}}_{\mathcal{N}_j\times\mathcal{A}^c_G}\boldsymbol{\varepsilon}_{Y,\mathcal{N}_j}\|_\infty > b\right\} &\le \mathbb{P}\left\{\sum_{j=0}^{q}\|\mathbf{G}^{\mathrm{T}}_{\mathcal{N}_j\times\mathcal{A}^c_G}\boldsymbol{\varepsilon}_{Y,\mathcal{N}_j}\|_\infty > b\right\} \\
&\le \sum_{j=0}^{q}\mathbb{P}\left\{\|\mathbf{G}^{\mathrm{T}}_{\mathcal{N}_j\times\mathcal{A}^c_G}\boldsymbol{\varepsilon}_{Y,\mathcal{N}_j}\|_\infty > \frac{b}{q+1}\right\}
\end{aligned}
$$

and in the last term, for any $j = 0, 1, \dots, q$, by the same approach of $\mathbb{P}_1$ and $\mathbb{P}_2$ decomposition, we have $\|\mathbf{G}^{\mathrm{T}}_{\mathcal{N}_j\times\mathcal{A}^c_G}\boldsymbol{\varepsilon}_{Y,\mathcal{N}_j}\|_\infty = O_p(\sqrt{n\log p})$.

Since q is fixed, we have $\|\mathbf{W}_{\mathcal{A}^c}\boldsymbol{\varepsilon}_Y\|_\infty = O_p(\sqrt{n\log p})$, together with $\|\mathbf{W}_{\mathcal{A}^c}\mathbf{W}_{\mathcal{A}}(\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} - \boldsymbol{\theta}^\star_{W,\mathcal{A}})\|_\infty = O_p(\sqrt{sn})$ from (A.1.22), we have

$$
\begin{aligned}
& \mathbb{P}\left\{\frac{1}{n\lambda_n}\|\mathbf{W}_{\mathcal{A}^c}(\mathbf{Y} - \mathbf{W}\overline{\boldsymbol{\theta}}_W)\|_\infty \ge \rho'(0+, \lambda_n)\right\} \\
\overset{(\text{A.1.21})}{\le} \quad & \mathbb{P}\left\{\frac{1}{n\lambda_n}\|\mathbf{W}_{\mathcal{A}^c}\boldsymbol{\varepsilon}_Y\|_\infty \ge \frac{\rho'(0+, \lambda_n)}{2}\right\} \\
+ \quad & \mathbb{P}\left\{\frac{1}{n\lambda_n}\|\mathbf{W}_{\mathcal{A}^c}\mathbf{W}_{\mathcal{A}}(\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} - \boldsymbol{\theta}^\star_{W,\mathcal{A}})\|_\infty \ge \frac{\rho'(0+, \lambda_n)}{2}\right\} \\
\overset{n\to\infty}{\longrightarrow} \quad & 0,
\end{aligned} \tag{A.1.30}
$$

where the last step follows from the constant $\rho'(0+, \lambda_n)$ in Condition A.1.1 and $\lambda_n \gg \max\{\sqrt{s/n}, \sqrt{\log p/n}\}$ in Condition A.1.4. So we have obtained equation (8) of Theorem 1 in (Fan & Lv, 2011) and **Step 2** is finished.

Step 3: Asymptotic expansion. In this step, we will show that $\sqrt{n}(\widehat{\boldsymbol{\theta}}^\star_{W,\mathcal{A}} - \boldsymbol{\theta}^\star_{W,\mathcal{A}}) = (1/\sqrt{n})\boldsymbol{\Sigma}^{-1}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\varepsilon}_Y + o_p(\mathbf{1})$, where $\widehat{\boldsymbol{\theta}}^\star_{W,\mathcal{A}} = \overline{\boldsymbol{\theta}}_{W,\mathcal{A}}$. From the outcome model (4.2.1), we have

$$\frac{1}{\sqrt{n}}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\varepsilon}_Y = \frac{1}{\sqrt{n}}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\left(\mathbf{Y} - \mathbf{W}_{\mathcal{A}}\boldsymbol{\theta}^\star_{W,\mathcal{A}}\right). \tag{A.1.31}$$

Denote

$$\overline{\boldsymbol{\theta}}_{\mathcal{A}_{G,D}} = \left[\overline{\boldsymbol{\theta}}^{\mathrm{T}}_{G,\mathcal{A}_G}\overline{\boldsymbol{\theta}}^{\mathrm{T}}_{D_1,\mathcal{A}_{D_1}}\dots\overline{\boldsymbol{\theta}}^{\mathrm{T}}_{D_q,\mathcal{A}_{D_q}}\right]^{\mathrm{T}} \in \mathbb{R}^s, \tag{A.1.32}$$

the sub-vector of $\overline{\boldsymbol{\theta}}_{W,\mathcal{A}}$. From $\partial\overline{Q}/\partial\boldsymbol{\theta}_{W,\mathcal{A}}|_{\overline{\boldsymbol{\theta}}_{W,\mathcal{A}}} = \mathbf{0}$, the definition of $\boldsymbol{\nu}_3(\boldsymbol{\theta}_{\mathcal{A}_{G,D}})$ in (A.1.7), and $\overline{\boldsymbol{\theta}}_{\mathcal{A}_{G,D}}$ in (A.1.32), we have

$$\sqrt{n}\left[\mathbf{0}^{\mathrm{T}}_{1+q+r}\ \boldsymbol{\nu}_3^{\mathrm{T}}(\overline{\boldsymbol{\theta}}_{\mathcal{A}_{G,D}})\right]^{\mathrm{T}} = \frac{1}{\sqrt{n}}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\left(\mathbf{Y} - \mathbf{W}_{\mathcal{A}}\overline{\boldsymbol{\theta}}_{W,\mathcal{A}}\right). \tag{A.1.33}$$

From (A.1.31) - (A.1.33), we have

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\varepsilon}_Y - \sqrt{n}\left[\mathbf{0}^{\mathrm{T}}_{1+q+r}\ \boldsymbol{\nu}_3^{\mathrm{T}}(\overline{\boldsymbol{\theta}}_{\mathcal{A}_{G,D}})\right]^{\mathrm{T}} \\
= &\ \sqrt{n}\left(\frac{\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\mathbf{W}_{\mathcal{A}}}{n} - \boldsymbol{\Sigma}\right)\left(\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} - \boldsymbol{\theta}^\star_{W,\mathcal{A}}\right) + \sqrt{n}\boldsymbol{\Sigma}\left(\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} - \boldsymbol{\theta}^\star_{W,\mathcal{A}}\right),
\end{aligned}$$

which implies that

$$\begin{aligned}
\sqrt{n}\left(\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} - \boldsymbol{\theta}^\star_{W,\mathcal{A}}\right) &= \boldsymbol{\Sigma}^{-1}\frac{1}{\sqrt{n}}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\varepsilon}_Y - \boldsymbol{\Sigma}^{-1}\sqrt{n}\left[\mathbf{0}^{\mathrm{T}}_{1+q+r}, \boldsymbol{\nu}_3^{\mathrm{T}}(\overline{\boldsymbol{\theta}}_{\mathcal{A}_{G,D}})\right]^{\mathrm{T}} \\
&- \boldsymbol{\Sigma}^{-1}\sqrt{n}\left(\frac{\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\mathbf{W}_{\mathcal{A}}}{n} - \boldsymbol{\Sigma}\right)\left(\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} - \boldsymbol{\theta}^\star_{W,\mathcal{A}}\right).
\end{aligned} \tag{A.1.34}$$

We will show the last two terms on the right hand side of (A.1.34) are $o_p(\mathbf{1}) \in \mathbb{R}^{1+q+r+s}$.

Suppose that for $h = 1, \dots, s$, $\overline{\theta}_{\mathcal{A}_{G,D},h}$ is an entry in $\overline{\boldsymbol{\theta}}_{\mathcal{A}_{G,D}}$ defined in (A.1.32), and $\theta^\star_{\mathcal{A}_{G,D},h}$ is the corresponding true value. If $\|\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} - \boldsymbol{\theta}^\star_{W,\mathcal{A}}\|_2 \le d_n$, then for any entry in $\overline{\boldsymbol{\theta}}_{\mathcal{A}_{G,D}}$, since they are also entries in $\overline{\boldsymbol{\theta}}_{W,\mathcal{A}}$, we have $d_n \ge |\theta^\star_{\mathcal{A}_{G,D},h} - \overline{\theta}_{\mathcal{A}_{G,D},h}| \ge |\theta^\star_{\mathcal{A}_{G,D},h}| - |\overline{\theta}_{\mathcal{A}_{G,D},h}| \ge (\min_{l\in\mathcal{A}_G}\{|\theta^\star_{G,l}|\} \wedge \min_{j=1,\dots,q,l\in\mathcal{A}_{D_j}}\{|\theta^\star_{D_j,l}|\}) - |\overline{\theta}_{\mathcal{A}_{G,D},h}| = 2d_n - |\overline{\theta}_{\mathcal{A}_{G,D},h}|$, *i.e.* , $|\overline{\theta}_{\mathcal{A}_{G,D},h}| \ge d_n$

for any $\bar{\theta}_{\mathcal{A}_{G,D},h}$ in $\overline{\boldsymbol{\theta}}_{\mathcal{A}_{G,D}}$. That is, $\|\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} - \boldsymbol{\theta}^{\star}_{W,\mathcal{A}}\|_2 \le d_n$ so that $\min_{h=1,\ldots,s} |\bar{\theta}_{\mathcal{A}_{G,D},h}| \ge d_n$. Define

$$\mathcal{E}_{\min \overline{\boldsymbol{\theta}}_{\mathcal{A}_{G,D}}} = \left\{ \min_{h=1,\ldots,s} |\bar{\theta}_{\mathcal{A}_{G,D},h}| \ge d_n \right\}.$$

We have

$$1 \ge \mathbb{P}\left\{ \mathcal{E}_{\min \overline{\boldsymbol{\theta}}_{\mathcal{A}_{G,D}}} \right\} \ge \mathbb{P}\left\{ \|\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} - \boldsymbol{\theta}^{\star}_{W,\mathcal{A}}\|_2 \le d_n \right\} \to 1, \tag{A.1.35}$$

where the last step follows from the conclusion $\|\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} - \boldsymbol{\theta}^{\star}_{W,\mathcal{A}}\|_2 = O_p(\sqrt{s/n})$ in **Step 1** and $d_n \gg \lambda_n \gg \sqrt{s/n}$ in Condition A.1.4. Hence, we can claim that

$$\|\boldsymbol{\Sigma}^{-1}\sqrt{n}\left[\mathbf{0}^{\mathrm{T}}_{1+q+r}, \boldsymbol{\nu}_3^{\mathrm{T}}(\overline{\boldsymbol{\theta}}_{\mathcal{A}_{G,D}})\right]^{\mathrm{T}}\|_2 = O_p\left(\sqrt{ns}p'_{\lambda_n}(d_n)\right), \tag{A.1.36}$$

because for any given $\tau > 0$, we have

$$\begin{aligned}
& \mathbb{P}\left\{ \frac{\|\boldsymbol{\Sigma}^{-1}\sqrt{n}\left[\mathbf{0}^{\mathrm{T}}_{1+q+r}, \boldsymbol{\nu}_3^{\mathrm{T}}(\overline{\boldsymbol{\theta}}_{\mathcal{A}_{G,D}})\right]^{\mathrm{T}}\|_2}{\sqrt{ns}p'_{\lambda_n}(d_n)} > \tau \right\} \le \mathbb{P}\left\{ \frac{\sqrt{n}\lambda_{\max}(\boldsymbol{\Sigma}^{-1})\|\boldsymbol{\nu}_3(\overline{\boldsymbol{\theta}}_{\mathcal{A}_{G,D}})\|_2}{\sqrt{ns}p'_{\lambda_n}(d_n)} > \tau \right\} \\
\le \;& \mathbb{P}\left\{ \frac{\|\boldsymbol{\nu}_3(\overline{\boldsymbol{\theta}}_{\mathcal{A}_{G,D}})\|_2}{C_\Sigma\sqrt{s}p'_{\lambda_n}(d_n)} > \tau, \mathcal{E}_{\min \overline{\boldsymbol{\theta}}_{\mathcal{A}_{G,D}}} \right\} + \mathbb{P}\left\{ \frac{\|\boldsymbol{\nu}_3(\overline{\boldsymbol{\theta}}_{\mathcal{A}_{G,D}})\|_2}{C_\Sigma\sqrt{s}p'_{\lambda_n}(d_n)} > \tau, \mathcal{E}^c_{\min \overline{\boldsymbol{\theta}}_{\mathcal{A}_{G,D}}} \right\} \\
\le \;& \mathbb{P}\left\{ \frac{\sqrt{sp'^2_{\lambda_n}(d_n)}}{C_\Sigma\sqrt{s}p'_{\lambda_n}(d_n)} > \tau \right\} + \mathbb{P}\left\{ \mathcal{E}^c_{\min \overline{\boldsymbol{\theta}}_{\mathcal{A}_{G,D}}} \right\} \qquad \text{(A.1.37)} \\
= \;& \mathbb{P}\left\{1/C_\Sigma > \tau\right\} + o(1) \to 0, \qquad \text{(A.1.38)}
\end{aligned}$$

where (A.1.37) follows from the fact that $p'_{\lambda_n}(t)$ is decreasing on $[0, \infty)$. For (A.1.38), if $\tau > 1/C_\Sigma$, the first probability is zero and the second probability follows from (A.1.35). Hence if $p'_{\lambda_n}(d_n) = o(1/\sqrt{ns})$, we have $\|\boldsymbol{\Sigma}^{-1}\sqrt{n}\left[\mathbf{0}^{\mathrm{T}}_{1+q+r}\ \boldsymbol{\nu}_3^{\mathrm{T}}(\overline{\boldsymbol{\theta}}_{\mathcal{A}_{G,D}})\right]^{\mathrm{T}}\|_2 = o_p(1)$.

We also have

$$\begin{aligned}
& \|\boldsymbol{\Sigma}^{-1}\sqrt{n}(n^{-1}\mathbf{W}^{\mathrm{T}}_{\mathcal{A}}\mathbf{W}_{\mathcal{A}} - \boldsymbol{\Sigma})(\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} - \boldsymbol{\theta}^{\star}_{W,\mathcal{A}})\|_2 \\
\le \;& \sqrt{n}\lambda_{\max}(\boldsymbol{\Sigma}^{-1})\|n^{-1}\mathbf{W}^{\mathrm{T}}_{\mathcal{A}}\mathbf{W}_{\mathcal{A}} - \boldsymbol{\Sigma}\|\|\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} - \boldsymbol{\theta}^{\star}_{W,\mathcal{A}}\|_2 \\
\le \;& \sqrt{n}C_\Sigma^{-1}\|n^{-1}\mathbf{W}^{\mathrm{T}}_{\mathcal{A}}\mathbf{W}_{\mathcal{A}} - \boldsymbol{\Sigma}\|_{\mathrm{F}}\|\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} - \boldsymbol{\theta}^{\star}_{W,\mathcal{A}}\|_2
\end{aligned}$$

$$
\begin{aligned}
&= C_\Sigma^{-1}\sqrt{n}O_p(s/\sqrt{n})O_p(\sqrt{s/n}) \\
&= O_p(\sqrt{s^3/n}). \qquad \text{(A.1.39)}
\end{aligned}
$$

As long as $s = o(n^{1/3})$, then $\|\mathbf{\Sigma}^{-1}\sqrt{n}(n^{-1}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\mathbf{W}_{\mathcal{A}} - \mathbf{\Sigma})(\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} - \boldsymbol{\theta}^{\star}_{W,\mathcal{A}})\|_2 = o_p(1)$.

Replacing $\overline{\boldsymbol{\theta}}_{W,\mathcal{A}}$ with $\widehat{\boldsymbol{\theta}}^{\star}_{W,\mathcal{A}}$ in (A.1.34), we complete the proof. $\square$

Lemma A.1.1. *Define objects $\mathbf{M}_j = n_j\widehat{\mathbf{b}}_j/n - p_j\boldsymbol{\mu}_{\gamma_j}$ for $j = 0, 1, \ldots q$, where $\widehat{\mathbf{b}}_j = n_j^{-1}\sum_{i=1}^{n_j}\boldsymbol{\gamma}_{i,j}$ and $n_0 + n_1 + \cdots + n_q = n$. Assume that $\boldsymbol{\gamma}_{i,j}$ are mutually independent for all i and j. For given $j = 0, 1, \ldots, q$, $\boldsymbol{\gamma}_{i,j}$ are identically distributed which can either be vectors in $\mathbb{R}^s$ or matrices in $\mathbb{R}^{r\times s}$ with mean $\boldsymbol{\mu}_{\gamma_j}$. Suppose the elements in $\boldsymbol{\gamma}_{i,j}$ given all i, j have the second moments bounded by an absolute constant and $|n_j/n - p_j| \le n^{-1/2}$. Then $\mathbb{E}\|\sum_{j=0}^q \mathbf{M}_j\|_2^2 \le O(q^2 s/n)$ if $\boldsymbol{\gamma}_{i,j} \in \mathbb{R}^s$, and $\mathbb{E}\|\sum_{j=0}^q \mathbf{M}_j\|_{\mathrm{F}}^2 \le O(q^2 rs/n)$ if $\boldsymbol{\gamma}_{i,j} \in \mathbb{R}^{r\times s}$.*

Proof. Let $\|\|_*$ be either the ℓ_2 norm or the Frobenius norm. Then we have

$$
\begin{aligned}
\mathbb{E}\|\sum_{j=0}^q \mathbf{M}_j\|_*^2 &\le \mathbb{E}\left(\sum_{j=0}^q \|\mathbf{M}_j\|_*\right)^2 \\
&= \mathbb{E}\left(\sum_{j=0}^q \|\mathbf{M}_j\|_*^2 + \sum_{j=0}^q\sum_{k\neq j}\|\mathbf{M}_j\|_*\|\mathbf{M}_k\|_*\right) \\
&\le \sum_{j=0}^q\left[\sqrt{\mathbb{E}\left(\|\mathbf{M}_j\|_*^2\right)}\right]^2 + \sum_{j=0}^q\sum_{k\neq j}\sqrt{\mathbb{E}\left(\|\mathbf{M}_j\|_*^2\right)}\sqrt{\mathbb{E}\left(\|\mathbf{M}_k\|_*^2\right)} \qquad \text{(A.1.40)} \\
&= \left[\sum_{j=0}^q\sqrt{\mathbb{E}\left(\|\mathbf{M}_j\|_*^2\right)}\right]^2, \qquad \text{(A.1.41)}
\end{aligned}
$$

where (A.1.40) follows from Hölder's inequality. Since $\|\mathbf{M}_j\|_* = \|n_j\widehat{\mathbf{b}}_j/n - p_j\boldsymbol{\mu}_{\gamma_j}\|_* = \|n_j\widehat{\mathbf{b}}_j/n - n_j\boldsymbol{\mu}_{\gamma_j}/n + n_j\boldsymbol{\mu}_{\gamma_j}/n - p_j\boldsymbol{\mu}_{\gamma_j}\|_* \le |n_j/n|\|\widehat{\mathbf{b}}_j - \boldsymbol{\mu}_{\gamma_j}\|_* + |n_j/n - p_j|\|\boldsymbol{\mu}_{\gamma_j}\|_*$, we have $\mathbb{E}(\|\mathbf{M}_j\|_*^2) \le |n_j/n|^2\mathbb{E}\|\widehat{\mathbf{b}}_j - \boldsymbol{\mu}_{\gamma_j}\|_*^2 + |n_j/n - p_j|^2\|\boldsymbol{\mu}_{\gamma_j}\|_*^2 + 2|n_j/n||n_j/n - p_j|\|\boldsymbol{\mu}_{\gamma_j}\|_*\mathbb{E}\|\widehat{\mathbf{b}}_j - \boldsymbol{\mu}_{\gamma_j}\|_*$. Since $\mathbb{E}\|\widehat{\mathbf{b}}_j - \boldsymbol{\mu}_{\gamma_j}\|_* \le \mathbb{E}^{1/2}(\|\widehat{\mathbf{b}}_j - \boldsymbol{\mu}_{\gamma_j}\|_*^2)$, we have

$$
\begin{aligned}
\mathbb{E}\left(\|\mathbf{M}_j\|_*^2\right) &\le \left|\frac{n_j}{n}\right|^2\mathbb{E}\|\widehat{\mathbf{b}}_j - \boldsymbol{\mu}_{\gamma_j}\|_*^2 + \left|\frac{n_j}{n} - p_j\right|^2\|\boldsymbol{\mu}_{\gamma_j}\|_*^2 \\
&+ 2\left|\frac{n_j}{n}\right|\left|\frac{n_j}{n} - p_j\right|\|\boldsymbol{\mu}_{\gamma_j}\|_*\sqrt{\mathbb{E}\|\widehat{\mathbf{b}}_j - \boldsymbol{\mu}_{\gamma_j}\|_*^2}
\end{aligned}
$$

$$= \left(\left| \frac{n_j}{n} \right| \sqrt{\mathbb{E}\|\widehat{\mathbf{b}}_j - \boldsymbol{\mu}_{\gamma_j}\|_*^2} + \left| \frac{n_j}{n} - p_j \right| \|\boldsymbol{\mu}_{\gamma_j}\|_* \right)^2. \qquad \text{(A.1.42)}$$

Suppose $\|\|_*$ is the ℓ_2 norm and $\boldsymbol{\gamma}_{i,j} = [\gamma_{i,j,1}, \ldots, \gamma_{i,j,s}]^{\mathrm{T}} \in \mathbb{R}^s$ with mean $\boldsymbol{\mu}_{\gamma_j} = [\mu_{\gamma_j,1}, \ldots, \mu_{\gamma_j,s}]^{\mathrm{T}} \in \mathbb{R}^s$ for simplicity. Then

$$\begin{aligned}
\mathbb{E}\|\widehat{\mathbf{b}}_j - \boldsymbol{\mu}_{\gamma_j}\|_2^2 &= \mathbb{E}\|n_j^{-1} \sum_{i=1}^{n_j} (\boldsymbol{\gamma}_{i,j} - \boldsymbol{\mu}_{\gamma_j})\|_2^2 \\
&= n_j^{-2} \mathbb{E}\| \sum_{i=1}^{n_j} (\boldsymbol{\gamma}_{i,j} - \boldsymbol{\mu}_{\gamma_j})\|_2^2 \\
&= n_j^{-2} \mathbb{E} \left\| \sum_{i=1}^{n_j} \left[(\gamma_{i,j,1} - \mu_{\gamma_j,1}), \ldots, (\gamma_{i,j,1} - \mu_{\gamma_j,1}) \right]^{\mathrm{T}} \right\|_2^2 \\
&= n_j^{-2} \mathbb{E} \left\| \left[\sum_{i=1}^{n_j} (\gamma_{i,j,1} - \mu_{\gamma_j,1}), \ldots, \sum_{i=1}^{n_j} (\gamma_{i,j,1} - \mu_{\gamma_j,1}) \right]^{\mathrm{T}} \right\|_2^2 \\
&= n_j^{-2} \left\{ \mathbb{E} \left[\sum_{i=1}^{n_j} (\gamma_{i,j,1} - \mu_{\gamma_j,1}) \right]^2 + \cdots + \mathbb{E} \left[\sum_{i=1}^{n_j} (\gamma_{i,j,s} - \mu_{\gamma_j,s}) \right]^2 \right\}.
\end{aligned}$$

By the i.i.d. property of $\gamma_{i,j,1},\ i = 1, \ldots, n_j$, we have $\mathbb{E}[\sum_{i=1}^{n_j} (\gamma_{i,j,1} - \mu_{\gamma_j,1})]^2 = \sum_{i=1}^{n_j} \mathbb{E}(\gamma_{i,j,1} - \mu_{\gamma_j,1})^2 = n_j \mathrm{Var}(\gamma_{i,j,1}) \le C n_j$, where C is obtained by the bounded moment condition of $\gamma_{i,j,1}$. Similarly, $\mathbb{E}[\sum_{i=1}^{n_j} (\gamma_{i,j,s} - \mu_{\gamma_j,s})]^2 \le C n_j$. Hence $\mathbb{E}\|\widehat{\mathbf{b}}_j - \boldsymbol{\mu}_{\gamma_j}\|_2^2 \le Cs/n_j$.

Since $\|\boldsymbol{\mu}_{\gamma_j}\|_2 \le C\sqrt{s}$ and $|n_j/n - p_j| \le n^{-1/2}$, by (A.1.42), we have $\mathbb{E}(\|\mathbf{M}_j\|_2^2) \le Cs/n$. By (A.1.41), we have $\mathbb{E}\| \sum_{j=1}^q \mathbf{M}_j\|_2^2 \le Cq^2 s/n$.

If $\|\|_*$ is the Frobenius norm, notice that $\mathbb{E}\|\widehat{\mathbf{b}}_j - \boldsymbol{\mu}_{\gamma_j}\|_{\mathrm{F}}^2 = \mathbb{E}\|\mathrm{vec}(\widehat{\mathbf{b}}_j) - \mathrm{vec}(\boldsymbol{\mu}_{\gamma_j})\|_2^2 \le Crs/n_j$ and $\mathrm{vec}(\widehat{\mathbf{b}}_j),\ \mathrm{vec}(\boldsymbol{\mu}_{\gamma_j}) \in \mathbb{R}^{rs}$. Hence $\mathbb{E}\| \sum_{j=1}^q \mathbf{M}_j\|_{\mathrm{F}}^2 \le Cq^2 rs/n$. $\square$

A.2 Proof of Theorem 4.4.2

A.2.1 Proof of $\widehat{\delta}_{\mathbf{x},j}(\mathbf{t}_0)$

We first show the asymptotic normality of $\widehat{\delta}_{\mathbf{x},j}(\mathbf{t}_0)$, *i.e.* , $\sqrt{n}[\widehat{\delta}_{\mathbf{x},j}(\mathbf{t}_0) - \delta_{\mathbf{x},j}(\mathbf{t}_0)] \xrightarrow{L} N(0, \mathbf{V}_{\delta_{\mathbf{x},j}(\mathbf{t}_0)})$, where $\delta_{\mathbf{x},j}(\mathbf{t}_0) = \boldsymbol{\beta}_j^{\mathrm{T}} \boldsymbol{\theta}_G^\star$. By (4.3.3), $\widehat{\delta}_{\mathbf{x},j}(\mathbf{t}_0)$ has the form $\widehat{\delta}_{\mathbf{x},j}(\mathbf{t}_0) = \widehat{\boldsymbol{\beta}_j^{\mathrm{T}} \boldsymbol{\theta}_G^\star} = \mathbf{e}_{j+1}^{\mathrm{T}} (\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-1} \mathbf{H}^{\mathrm{T}} \mathbf{G} \widehat{\boldsymbol{\theta}}_G$, where $\widehat{\boldsymbol{\theta}}_G^\star$ is from (4.3.1) such that all the entries in $\{1, \ldots, p\} \setminus \mathcal{A}_G$

are zeros and all the entries in $\mathcal{A}_G$ are from $\overline{\boldsymbol{\theta}}_{G,\mathcal{A}_G}$, a sub-vector of $\overline{\boldsymbol{\theta}}_{W,\mathcal{A}}$ in (A.1.4). By some calculation, we have

$$\begin{aligned}
\sqrt{n}\left[\widehat{\delta}_{\mathbf{x},j}(\mathbf{t}_0)-\delta_{\mathbf{x},j}(\mathbf{t}_0)\right] &= \sqrt{n}\left[\mathbf{e}_{j+1}^{\mathrm{T}}(\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{G}\widehat{\boldsymbol{\theta}}_G^{\star}-\boldsymbol{\beta}_j^{\mathrm{T}}\boldsymbol{\theta}_G^{\star}\right]\\
&= \boldsymbol{\beta}_{j,\mathcal{A}_G}^{\mathrm{T}}\tilde{\mathbf{I}}_{\mathcal{A}_G}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}n^{-1/2}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\varepsilon}_Y+\mathbf{e}_{j+1}^{\mathrm{T}}\mathbf{J}^{-1}\mathbf{H}^{\mathrm{T}}n^{-1/2}\boldsymbol{\varepsilon}_{G,\mathcal{A}_G}\boldsymbol{\theta}_{G,\mathcal{A}_G}^{\star}\\
&+ \ Error, \qquad\qquad (\mathrm{A.2.1})
\end{aligned}$$

where $\boldsymbol{\varepsilon}_{G,\mathcal{A}_G}\in\mathbb{R}^{n\times s_G}$ is a sub-matrix of $\boldsymbol{\varepsilon}_G$ consisting of all the columns in $\mathcal{A}_G$, $\tilde{\mathbf{I}}_{\mathcal{A}_G}\in\mathbb{R}^{(1+q+r+s)\times s_G}$ is defined before Theorem 4.4.2, $\mathbf{e}_{j+1}$ is the $(j+1)$th standard basis in $\mathbb{R}^{1+q+r}$, $\mathbf{J}\in\mathbb{R}^{(1+q+r)\times(1+q+r)}$ is the upper-left $(1+q+r)\times(1+q+r)$ sub-matrix of $\boldsymbol{\Sigma}$, and

$$\begin{aligned}
Error &= -\boldsymbol{\beta}_{j,\mathcal{A}_G}^{\mathrm{T}}\tilde{\mathbf{I}}_{\mathcal{A}_G}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\sqrt{n}\left[\mathbf{0}_{1+q+r}^{\mathrm{T}},\boldsymbol{\nu}_3^{\mathrm{T}}(\overline{\boldsymbol{\theta}}_{\mathcal{A}_{G,D}})\right]^{\mathrm{T}}\\
&- \ \boldsymbol{\beta}_{j,\mathcal{A}_G}^{\mathrm{T}}\tilde{\mathbf{I}}_{\mathcal{A}_G}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\sqrt{n}\left(\frac{\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\mathbf{W}_{\mathcal{A}}}{n}-\boldsymbol{\Sigma}\right)\left(\overline{\boldsymbol{\theta}}_{W,\mathcal{A}}-\boldsymbol{\theta}_{W,\mathcal{A}}^{\star}\right)\\
&+ \ \mathbf{e}_{j+1}^{\mathrm{T}}(\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\boldsymbol{\varepsilon}_{G,\mathcal{A}_G}\sqrt{n}\tilde{\mathbf{I}}_{\mathcal{A}_G}^{\mathrm{T}}\left(\overline{\boldsymbol{\theta}}_{W,\mathcal{A}}-\boldsymbol{\theta}_{W,\mathcal{A}}^{\star}\right)\\
&+ \ \mathbf{e}_{j+1}^{\mathrm{T}}\left[\left(\frac{\mathbf{H}^{\mathrm{T}}\mathbf{H}}{n}\right)^{-1}-\mathbf{J}^{-1}\right]n^{-1/2}\mathbf{H}^{\mathrm{T}}\boldsymbol{\varepsilon}_{G,\mathcal{A}_G}\boldsymbol{\theta}_{G,\mathcal{A}_G}^{\star}.
\end{aligned}$$

We will show that $Error=o_p(1)$ first.

For the first term in $Error$, by (A.1.36), we have $\|\boldsymbol{\Sigma}^{-1}\sqrt{n}\left[\mathbf{0}_{1+q+r}^{\mathrm{T}},\boldsymbol{\nu}_3^{\mathrm{T}}(\overline{\boldsymbol{\theta}}_{\mathcal{A}_{G,D}})\right]^{\mathrm{T}}\|_2 = O_p(\sqrt{n}sp'_{\lambda_n}(d_n))$. By Condition A.1.5, we can find an absolute constant C_β such that $\beta_{j,h}\le C_\beta$ for $j=1,\dots,q$ and $h=1,\dots,p$ so that $\|\boldsymbol{\beta}_{j,\mathcal{A}_G}\|_2=O(\sqrt{s})$. Hence we have

$$\begin{aligned}
|\boldsymbol{\beta}_{j,\mathcal{A}_G}^{\mathrm{T}}\tilde{\mathbf{I}}_{\mathcal{A}_G}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\sqrt{n}[\mathbf{0}_{1+q+r}^{\mathrm{T}},\boldsymbol{\nu}_3^{\mathrm{T}}(\overline{\boldsymbol{\theta}}_{\mathcal{A}_{G,D}})]^{\mathrm{T}}| &\le \|\boldsymbol{\beta}_{j,\mathcal{A}_G}\|_2\|\boldsymbol{\Sigma}^{-1}\sqrt{n}[\mathbf{0}_{1+q+r}^{\mathrm{T}}\ \boldsymbol{\nu}_3^{\mathrm{T}}(\overline{\boldsymbol{\theta}}_{\mathcal{A}_{G,D}})]^{\mathrm{T}}\|_2\\
&= O_p(\sqrt{n}sp'_{\lambda_n}(d_n)).
\end{aligned}$$

If we assume that $p'_{\lambda_n}(d_n)=o(1/[\sqrt{n}s])$, the first term in $Error$ is $o_p(1)$.

For the second term in *Error*, we have

$$\begin{aligned}
& |\boldsymbol{\beta}_{j,\mathcal{A}_G}^{\mathrm{T}} \tilde{\mathbf{I}}_{\mathcal{A}_G}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \sqrt{n}(n^{-1}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\mathbf{W}_{\mathcal{A}} - \boldsymbol{\Sigma})(\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} - \boldsymbol{\theta}_{W,\mathcal{A}}^{\star})| \\
\leq \; & \|\boldsymbol{\beta}_{j,\mathcal{A}_G}\|_2 \|\boldsymbol{\Sigma}^{-1} \sqrt{n}(n^{-1}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\mathbf{W}_{\mathcal{A}} - \boldsymbol{\Sigma})(\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} - \boldsymbol{\theta}_{W,\mathcal{A}}^{\star})\|_2 \\
\leq \; & O(\sqrt{s}) O_p(\sqrt{s^3/n}) = O_p(\sqrt{s^4/n}),
\end{aligned}$$

where the second inequality follows from (A.1.39). Hence if $s = o(n^{1/4})$, the second term of *Error* is $o_p(1)$.

For the third term in *Error*, we have

$$\begin{aligned}
& |\mathbf{e}_{j+1}^{\mathrm{T}} (\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-1} \sqrt{n} \mathbf{H}^{\mathrm{T}} \boldsymbol{\varepsilon}_{G,\mathcal{A}_G} \tilde{\mathbf{I}}_{\mathcal{A}_G}^{\mathrm{T}} (\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} - \boldsymbol{\theta}_{W,\mathcal{A}}^{\star})| \\
\leq \; & \|(n^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-1} n^{-1/2} \mathbf{H}^{\mathrm{T}} \boldsymbol{\varepsilon}_{G,\mathcal{A}_G} \tilde{\mathbf{I}}_{\mathcal{A}_G}^{\mathrm{T}} (\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} - \boldsymbol{\theta}_{W,\mathcal{A}}^{\star})\|_2 \\
\leq \; & \|(n^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-1}\| \|n^{-1/2} \mathbf{H}^{\mathrm{T}} \boldsymbol{\varepsilon}_{G,\mathcal{A}_G} \tilde{\mathbf{I}}_{\mathcal{A}_G}^{\mathrm{T}}\| \|\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} - \boldsymbol{\theta}_{W,\mathcal{A}}^{\star}\|_2 \\
\leq \; & \lambda_{\min}^{-1}(n^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{H}) \|n^{-1/2} \mathbf{H}^{\mathrm{T}} \boldsymbol{\varepsilon}_{G,\mathcal{A}_G}\|_{\mathrm{F}} \|\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} - \boldsymbol{\theta}_{W,\mathcal{A}}^{\star}\|_2 \\
\leq \; & [C_\Sigma - o_p(1)]^{-1} O_p(s) O_p(\sqrt{s/n}) = O_p(\sqrt{s^3/n}). \qquad \text{(A.2.2)}
\end{aligned}$$

For the last inequality in (A.2.2), since $\lambda_{\max}(n^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{H} - \mathbf{J}) \geq \lambda_{\min}(n^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{H}) - \lambda_{\min}(\mathbf{J}) \geq \lambda_{\min}(n^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{H} - \mathbf{J})$, we have $\lambda_{\min}(n^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{H}) = \lambda_{\min}(\mathbf{J}) - [\lambda_{\min}(\mathbf{J}) - \lambda_{\min}(n^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{H})] \geq \lambda_{\min}(\mathbf{J}) - |\lambda_{\min}(n^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{H}) - \lambda_{\min}(\mathbf{J})| \geq \lambda_{\min}(\boldsymbol{\Sigma}) - \|n^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{H} - \mathbf{J}\| \geq C_\Sigma - \|n^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{H} - \mathbf{J}\|_{\mathrm{F}} \geq C_\Sigma - \|n^{-1}\mathbf{W}\mathcal{A}^{\mathrm{T}}\mathbf{W}_{\mathcal{A}} - \boldsymbol{\Sigma}\|_{\mathrm{F}} = C_\Sigma - o_p(1)$. By similar arguments of (A.1.9), we have $\|n^{-1/2}\mathbf{H}^{\mathrm{T}}\boldsymbol{\varepsilon}_{G,\mathcal{A}_G}\|_{\mathrm{F}} = O_p(s)$. From the proof of **Step 1** in Section A.1.3, we have $\|\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} - \boldsymbol{\theta}_{W,\mathcal{A}}^{\star}\|_2 = O_p(\sqrt{s/n})$. Hence if $s = o(n^{1/3})$, the third term in *Error* is $o_p(1)$.

For the fourth term in *Error*, we have

$$\begin{aligned}
& \left| \mathbf{e}_{j+1}^{\mathrm{T}} \left[\left(\frac{\mathbf{H}^{\mathrm{T}}\mathbf{H}}{n} \right)^{-1} - \mathbf{J}^{-1} \right] \frac{\mathbf{H}^{\mathrm{T}} \; \boldsymbol{\varepsilon}_{G,\mathcal{A}_G}}{\sqrt{n}} \boldsymbol{\theta}_{G,\mathcal{A}_G}^{\star} \right| \\
\leq \; & \left\| \left[\left(\frac{\mathbf{H}^{\mathrm{T}}\mathbf{H}}{n} \right)^{-1} - \mathbf{J}^{-1} \right] \frac{\mathbf{H}^{\mathrm{T}} \; \boldsymbol{\varepsilon}_{G,\mathcal{A}_G}}{\sqrt{n}} \boldsymbol{\theta}_{G,\mathcal{A}_G}^{\star} \right\|_2
\end{aligned}$$

$$
\begin{aligned}
&\leq \left\| \left(\frac{\mathbf{H}^{\mathrm{T}}\mathbf{H}}{n}\right)^{-1} - \mathbf{J}^{-1} \right\| \left\| \frac{\mathbf{H}^{\mathrm{T}}\ \boldsymbol{\varepsilon}_{G,\mathcal{A}_G}}{\sqrt{n}} \right\| \left\| \boldsymbol{\theta}^{\star}_{G,\mathcal{A}_G} \right\|_2 \\
&\overset{(i)}{\leq} \frac{\left\| \left(\frac{\mathbf{H}^{\mathrm{T}}\mathbf{H}}{n}\right) - \mathbf{J} \right\|_{\mathrm{F}}}{\lambda_{\min}\left(\frac{\mathbf{H}^{\mathrm{T}}\mathbf{H}}{n}\right)\lambda_{\min}\left(\mathbf{J}\right)} O_p(s)\sqrt{s}C_\theta \\
&\overset{(ii)}{\leq} C_\theta \frac{O_p(1/\sqrt{n})}{C_\Sigma\left[C_{\boldsymbol{\Sigma}} - O_p(1/\sqrt{n})\right]} O_p(s)\sqrt{s} \\
&= O_p(\sqrt{s^3/n}).
\end{aligned}
$$

In step (i), we use the fact that for two positive definite matrices $\mathbf{M}_1$ and $\mathbf{M}_2$, $\mathbf{M}_1^{-1} - \mathbf{M}_2^{-1} = \mathbf{M}_2^{-1}(\mathbf{M}_2 - \mathbf{M}_1)\mathbf{M}_1^{-1}$, so that $\|\mathbf{M}_1^{-1} - \mathbf{M}_2^{-1}\| \leq \|\mathbf{M}_2 - \mathbf{M}_1\|\lambda_{\min}^{-1}(\mathbf{M}_1)\lambda_{\min}^{-1}(\mathbf{M}_2) \leq \|\mathbf{M}_2 - \mathbf{M}_1\|_{\mathrm{F}}\lambda_{\min}^{-1}(\mathbf{M}_1)\lambda_{\min}^{-1}(\mathbf{M}_2)$. Since we assumed the bound for the entries of $\boldsymbol{\theta}^{\star}_{G,\mathcal{A}_G}$ is C_θ in Theorem 4.4.2, we have $\|\boldsymbol{\theta}^{\star}_{G,\mathcal{A}_G}\|_2 \leq \sqrt{s}C_\theta$. Also recall $\|n^{-1/2}\mathbf{H}^{\mathrm{T}}\boldsymbol{\varepsilon}_{G,\mathcal{A}_G}\|_{\mathrm{F}} = O_p(s)$ from the third term in *Error*. In step (ii), since we assumed r is fixed, by Lemma A.1.1, we have $\|n^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{H} - \mathbf{J}\|_{\mathrm{F}} = O_p(1/\sqrt{n})$. Therefore, if $s = o(n^{1/3})$, the fourth in *Error* term is $o_p(1)$.

Combining the four terms in *Error* together, we have shown that if $s = o(n^{1/4})$, $p'_{\lambda_n}(d_n) = o(1/[\sqrt{n}s])$, $\|\boldsymbol{\theta}^{\star}_{G,\mathcal{A}_G}\|_\infty \leq C_\theta$, and r is fixed, then $Error = o_p(1)$ and (A.2.1) becomes

$$
\sqrt{n}\left[\widehat{\delta}_{\mathbf{x},j}(\mathbf{t}_0) - \delta_{\mathbf{x},j}(\mathbf{t}_0)\right] = \boldsymbol{\varrho}^{\mathrm{T}}\frac{1}{\sqrt{n}}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\varepsilon}_Y + \boldsymbol{\varkappa}^{\mathrm{T}}\frac{1}{\sqrt{n}}\mathbf{H}^{\mathrm{T}}\boldsymbol{\varepsilon}_{G,\mathcal{A}_G}\boldsymbol{\theta}^{\star}_{G,\mathcal{A}_G} + o_p(1), \tag{A.2.3}
$$

where

$$
\begin{aligned}
\boldsymbol{\varrho} &= \boldsymbol{\Sigma}^{-1}\tilde{\mathbf{I}}_{\mathcal{A}_G}\boldsymbol{\beta}_{j,\mathcal{A}_G} := \begin{bmatrix} \underset{1\times 1}{\varrho_0} & \underset{1\times 1}{\varrho_{T_1}} & \cdots & \underset{1\times 1}{\varrho_{T_q}} & \underset{1\times r}{\boldsymbol{\varrho}_X^{\mathrm{T}}} & \underset{1\times s_G}{\boldsymbol{\varrho}_{\mathcal{A}_G}^{\mathrm{T}}} & \underset{1\times s_{D_1}}{\boldsymbol{\varrho}_{\mathcal{A}_{D_1}}^{\mathrm{T}}} & \cdots & \underset{1\times s_{D_q}}{\boldsymbol{\varrho}_{\mathcal{A}_{D_q}}^{\mathrm{T}}} \end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^{1+q+r+s}, \\
\boldsymbol{\varkappa} &= \mathbf{J}^{-1}\mathbf{e}_{j+1} := \begin{bmatrix} \underset{1\times 1}{\varkappa_0} & \underset{1\times 1}{\varkappa_{T_1}} & \cdots & \underset{1\times 1}{\varkappa_{T_q}} & \underset{1\times r}{\boldsymbol{\varkappa}_X^{\mathrm{T}}} \end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^{1+q+r}.
\end{aligned}
$$

The entries in $\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\varepsilon}_Y$ and $\mathbf{H}^{\mathrm{T}}\boldsymbol{\varepsilon}_{G,\mathcal{A}_G}\boldsymbol{\theta}^{\star}_{G,\mathcal{A}_G}$ are sum of independent random variables with different means in terms of different treatment levels. To develop the asymptotic normality, we first obtain the asymptotic normality for each treatment level and then sum them up.

Specifically,

$$\boldsymbol{\varrho}^{\mathrm{T}} \frac{1}{\sqrt{n}} \mathbf{W}_{\mathcal{A}}^{\mathrm{T}} \boldsymbol{\varepsilon}_Y + \boldsymbol{\varkappa}^{\mathrm{T}} \frac{1}{\sqrt{n}} \mathbf{H}^{\mathrm{T}} \boldsymbol{\varepsilon}_{G,\mathcal{A}_G} \boldsymbol{\theta}^{\star}_{G,\mathcal{A}_G} = \sum_{j=0}^{q} \sqrt{\frac{n_j}{n}} \mathbf{1}_5^{\mathrm{T}} \digamma_j, \tag{A.2.4}$$

where for $j = 1, \ldots, q$,

$$\digamma_j = \frac{1}{\sqrt{n_j}} \sum_{i \in \mathcal{N}_j} \begin{bmatrix} \left[\varrho_0 + \varrho_{T_j} + \boldsymbol{\varrho}_{\mathcal{A}_G}^{\mathrm{T}} \left(\boldsymbol{\beta}_{0,\mathcal{A}_G} + \boldsymbol{\beta}_{j,\mathcal{A}_G}\right) + \boldsymbol{\varrho}_{\mathcal{A}_{D_j}}^{\mathrm{T}} \left(\boldsymbol{\beta}_{0,\mathcal{A}_{D_j}} + \boldsymbol{\beta}_{j,\mathcal{A}_{D_j}}\right)\right] \epsilon_{Y_i} \\ \left(\boldsymbol{\varrho}_X^{\mathrm{T}} + \boldsymbol{\varrho}_{\mathcal{A}_G}^{\mathrm{T}} \mathbf{A}_{\mathcal{A}_G \times [r]} + \boldsymbol{\varrho}_{\mathcal{A}_{D_j}}^{\mathrm{T}} \mathbf{A}_{\mathcal{A}_{D_j} \times [r]}\right)^{\mathrm{T}} \mathbf{X}_i \epsilon_{Y_i} \\ \begin{bmatrix} \boldsymbol{\varrho}_{\mathcal{A}_G}^{\mathrm{T}} & \boldsymbol{\varrho}_{\mathcal{A}_{D_j}}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \boldsymbol{\varepsilon}_{G_i,\mathcal{A}_G} \\ \boldsymbol{\varepsilon}_{G_i,\mathcal{A}_{D_j}} \end{bmatrix} \epsilon_{Y_i} \\ \left(\varkappa_0 + \varkappa_{T_j}\right) \boldsymbol{\varepsilon}_{G_i,\mathcal{A}_G}^{\mathrm{T}} \boldsymbol{\theta}^{\star}_{G,\mathcal{A}_G} \\ \boldsymbol{\varkappa}_X^{\mathrm{T}} \mathbf{X}_i \boldsymbol{\varepsilon}_{G_i,\mathcal{A}_G}^{\mathrm{T}} \boldsymbol{\theta}^{\star}_{G,\mathcal{A}_G} \end{bmatrix},$$

and

$$\digamma_0 = \frac{1}{\sqrt{n_0}} \sum_{i \in \mathcal{N}_0} \begin{bmatrix} \left(\varrho_0 + \boldsymbol{\varrho}_{\mathcal{A}_G}^{\mathrm{T}} \boldsymbol{\beta}_{\beta_{0,\mathcal{A}_G}}\right) \epsilon_{Y_i} \\ \left(\boldsymbol{\varrho}_X^{\mathrm{T}} + \boldsymbol{\varrho}_{\mathcal{A}_G}^{\mathrm{T}} \mathbf{A}_{\mathcal{A}_G \times [r]}\right)^{\mathrm{T}} \mathbf{X}_i \epsilon_{Y_i} \\ \boldsymbol{\varrho}_{\mathcal{A}_G}^{\mathrm{T}} \boldsymbol{\varepsilon}_{G_i,\mathcal{A}_G} \epsilon_{Y_i} \\ \varkappa_0 \boldsymbol{\varepsilon}_{G_i,\mathcal{A}_G}^{\mathrm{T}} \boldsymbol{\theta}^{\star}_{G,\mathcal{A}_G} \\ \boldsymbol{\varkappa}_X^{\mathrm{T}} \mathbf{X}_i \boldsymbol{\varepsilon}_{G_i,\mathcal{A}_G}^{\mathrm{T}} \boldsymbol{\theta}^{\star}_{G,\mathcal{A}_G} \end{bmatrix}.$$

Then for $j = 1, \ldots, q$, we have $\digamma_j \xrightarrow{L} \mathcal{N}(\mathbf{0}_5, \boldsymbol{\Psi}_j)$, where the entries in $\boldsymbol{\Psi}_j$ are $\Psi_{j,kt}$, for $k = 1, \ldots, 5$ and $t = 1, \ldots, 5$. Specifically, we have $\Psi_{j,11} = \sigma_Y^2[\varrho_0 + \varrho_{T_j} + \boldsymbol{\varrho}_{\mathcal{A}_G}^{\mathrm{T}}(\boldsymbol{\beta}_{0,\mathcal{A}_G} + \boldsymbol{\beta}_{j,\mathcal{A}_G}) + \boldsymbol{\varrho}_{\mathcal{A}_{D_j}}^{\mathrm{T}}(\boldsymbol{\beta}_{0,\mathcal{A}_{D_j}} + \boldsymbol{\beta}_{j,\mathcal{A}_{D_j}})]^2$, $\Psi_{j,12} = \sigma_Y^2[\varrho_0 + \varrho_{T_j} + \boldsymbol{\varrho}_{\mathcal{A}_G}^{\mathrm{T}}(\boldsymbol{\beta}_{0,\mathcal{A}_G} + \boldsymbol{\beta}_{j,\mathcal{A}_G}) + \boldsymbol{\varrho}_{\mathcal{A}_{D_j}}^{\mathrm{T}}(\boldsymbol{\beta}_{0,\mathcal{A}_{D_j}} + \boldsymbol{\beta}_{j,\mathcal{A}_{D_j}})]\boldsymbol{\mu}_X^{\mathrm{T}}(\boldsymbol{\varrho}_X + \mathbf{A}_{\mathcal{A}_G \times [r]}^{\mathrm{T}} \boldsymbol{\varrho}_{\mathcal{A}_G} + \mathbf{A}_{\mathcal{A}_{D_j} \times [r]}^{\mathrm{T}} \boldsymbol{\varrho}_{\mathcal{A}_{D_j}})$, $\Psi_{j,13} = \Psi_{j,14} = \Psi_{j,15} = 0$, $\Psi_{j,22} = \sigma_Y^2(\boldsymbol{\varrho}_X^{\mathrm{T}} + \boldsymbol{\varrho}_{\mathcal{A}_G}^{\mathrm{T}} \mathbf{A}_{\mathcal{A}_G \times [r]} + \boldsymbol{\varrho}_{\mathcal{A}_{D_j}}^{\mathrm{T}} \mathbf{A}_{\mathcal{A}_{D_j} \times [r]})(\boldsymbol{\Sigma}_X + \boldsymbol{\mu}_X \boldsymbol{\mu}_X^{\mathrm{T}})(\boldsymbol{\varrho}_X + \mathbf{A}_{\mathcal{A}_G \times [r]}^{\mathrm{T}} \boldsymbol{\varrho}_{\mathcal{A}_G} + \mathbf{A}_{\mathcal{A}_{D_j} \times [r]}^{\mathrm{T}} \boldsymbol{\varrho}_{\mathcal{A}_{D_j}})$, $\Psi_{j,23} = \Psi_{j,24} = \Psi_{j,25} = 0$, $\Psi_{j,33} = \sigma_Y^2[\boldsymbol{\varrho}_{\mathcal{A}_G}^{\mathrm{T}} \boldsymbol{\Sigma}_{G(\mathcal{A}_G \times \mathcal{A}_G)} \boldsymbol{\varrho}_{\mathcal{A}_G} + \boldsymbol{\varrho}_{\mathcal{A}_G}^{\mathrm{T}} \boldsymbol{\Sigma}_{G(\mathcal{A}_G \times \mathcal{A}_{D_j})} \boldsymbol{\varrho}_{\mathcal{A}_{D_j}} + \boldsymbol{\varrho}_{\mathcal{A}_{D_j}}^{\mathrm{T}} \boldsymbol{\Sigma}_{G(\mathcal{A}_{D_j} \times \mathcal{A}_G)} \boldsymbol{\varrho}_{\mathcal{A}_G} + \boldsymbol{\varrho}_{\mathcal{A}_{D_j}}^{\mathrm{T}} \boldsymbol{\Sigma}_{G(\mathcal{A}_{D_j} \times \mathcal{A}_{D_j})} \boldsymbol{\varrho}_{\mathcal{A}_{D_j}}]$, $\Psi_{j,34} = \Psi_{j,35} = 0$, $\Psi_{j,44} = (\varkappa_0 + \varkappa_{T_j})^2 \boldsymbol{\theta}^{\star\mathrm{T}}_{G,\mathcal{A}_G} \boldsymbol{\Sigma}_{G(\mathcal{A}_G \times \mathcal{A}_G)} \boldsymbol{\theta}^{\star}_{G,\mathcal{A}_G}$, $\Psi_{j,45} = (\varkappa_0 + \varkappa_{T_j}) \boldsymbol{\theta}^{\star\mathrm{T}}_{G,\mathcal{A}_G} \boldsymbol{\Sigma}_{G(\mathcal{A}_G \times \mathcal{A}_G)} \boldsymbol{\theta}^{\star}_{G,\mathcal{A}_G} \boldsymbol{\varkappa}_X^{\mathrm{T}} \boldsymbol{\mu}_X$, $\Psi_{j,55} = \boldsymbol{\varkappa}_X^{\mathrm{T}}(\boldsymbol{\mu}_X \boldsymbol{\mu}_X^{\mathrm{T}} + \boldsymbol{\Sigma}_X) \boldsymbol{\varkappa}_X \boldsymbol{\theta}^{\star\mathrm{T}}_{G,\mathcal{A}_G} \boldsymbol{\Sigma}_{G(\mathcal{A}_G \times \mathcal{A}_G)} \boldsymbol{\theta}^{\star}_{G,\mathcal{A}_G}$.

Similarly, we have $\digamma_0 \xrightarrow{L} \mathcal{N}(\mathbf{0}_5, \boldsymbol{\Psi}_0)$, where $\boldsymbol{\Psi}_0 \in \mathbb{R}^{5\times 5}$ has entries $\Psi_{0,11} = \sigma_Y^2(\varrho_0 + \boldsymbol{\varrho}_{\mathcal{A}_G}^{\mathrm{T}}\boldsymbol{\beta}_{0,\mathcal{A}_G})^2$, $\Psi_{0,12} = \sigma_Y^2(\varrho_0 + \boldsymbol{\varrho}_{\mathcal{A}_G}^{\mathrm{T}}\boldsymbol{\beta}_{0,\mathcal{A}_G})\boldsymbol{\mu}_X^{\mathrm{T}}(\boldsymbol{\varrho}_X + \mathbf{A}_{\mathcal{A}_G\times[r]}^{\mathrm{T}}\boldsymbol{\varrho}_{\mathcal{A}_G})$, $\Psi_{0,13} = \Psi_{0,14} = \Psi_{0,15} = 0$, $\Psi_{0,22} = \sigma_Y^2(\boldsymbol{\varrho}_X^{\mathrm{T}} + \boldsymbol{\varrho}_{\mathcal{A}_G}^{\mathrm{T}}\mathbf{A}_{\mathcal{A}_G\times[r]})(\boldsymbol{\Sigma}_X + \boldsymbol{\mu}_X\boldsymbol{\mu}_X^{\mathrm{T}})(\boldsymbol{\varrho}_X + \mathbf{A}_{\mathcal{A}_G\times[r]}^{\mathrm{T}}\boldsymbol{\varrho}_{\mathcal{A}_G})$, $\Psi_{0,23} = \Psi_{0,24} = \Psi_{0,25} = 0$, $\Psi_{0,33} = \sigma_Y^2\boldsymbol{\varrho}_{\mathcal{A}_G}^{\mathrm{T}}\boldsymbol{\Sigma}_{G(\mathcal{A}_G\times\mathcal{A}_G)}\boldsymbol{\varrho}_{\mathcal{A}_G}$, $\Psi_{0,34} = \Psi_{0,35} = 0$, $\Psi_{0,44} = \varkappa_0^2\boldsymbol{\theta}_{G,\mathcal{A}_G}^{\star\mathrm{T}}\boldsymbol{\Sigma}_{G(\mathcal{A}_G\times\mathcal{A}_G)}\boldsymbol{\theta}_{G,\mathcal{A}_G}^{\star}$, $\Psi_{0,45} = \varkappa_0\boldsymbol{\theta}_{G,\mathcal{A}_G}^{\star\mathrm{T}}\boldsymbol{\Sigma}_{G(\mathcal{A}_G\times\mathcal{A}_G)}\boldsymbol{\theta}_{G,\mathcal{A}_G}^{\star}\boldsymbol{\varkappa}_X^{\mathrm{T}}\boldsymbol{\mu}_X$, $\Psi_{0,55} = \boldsymbol{\varkappa}_X^{\mathrm{T}}(\boldsymbol{\mu}_X\boldsymbol{\mu}_X^{\mathrm{T}} + \boldsymbol{\Sigma}_X)\boldsymbol{\varkappa}_X\boldsymbol{\theta}_{G,\mathcal{A}_G}^{\star\mathrm{T}}\boldsymbol{\Sigma}_{G(\mathcal{A}_G\times\mathcal{A}_G)}\boldsymbol{\theta}_{G,\mathcal{A}_G}^{\star}$.

By Slutsky's theorem, $\sqrt{n_j/n}\mathbf{1}_5^{\mathrm{T}}\digamma_j \xrightarrow{L} N(0, p_j\mathbf{1}_5^{\mathrm{T}}\boldsymbol{\Psi}_j\mathbf{1}_5)$ and $\digamma_j$, $j = 0, 1\ldots, q$ are independent. Thus $\sum_{j=0}^{q}\sqrt{n_j/n}\mathbf{1}_5\digamma_j \xrightarrow{L} N(0, \sum_{j=0}^{q}p_j\mathbf{1}_5^{\mathrm{T}}\boldsymbol{\Psi}_j\mathbf{1}_5) = N(0, \mathbf{V}_{\delta_{\mathbf{x},j}(\mathbf{t}_0)})$ which concludes the proof from (A.2.3) and (A.2.4). $\square$

A.2.2 Proof of $\widehat{\delta}_{\mathbf{x},j}(\mathbf{t}_j)$

Next we show the asymptotic normality of $\widehat{\delta}_{\mathbf{x},j}(\mathbf{t}_j)$, *i.e.*, $\sqrt{n}[\widehat{\delta}_{\mathbf{x},j}(\mathbf{t}_j) - \delta_{\mathbf{x},j}(\mathbf{t}_j)] \xrightarrow{L} N(0, \mathbf{V}_{\delta_{\mathbf{x},j}(\mathbf{t}_j)})$, where $\delta_{\mathbf{x},j}(\mathbf{t}_j) = \boldsymbol{\beta}_j^{\mathrm{T}}(\boldsymbol{\theta}_G^{\star} + \boldsymbol{\theta}_{D_j}^{\star})$. From (4.3.3), we have $\widehat{\delta}_{\mathbf{x},j}(\mathbf{t}_j) = \widehat{\boldsymbol{\beta}_j^{\mathrm{T}}\boldsymbol{\theta}_G^{\star}} + \widehat{\boldsymbol{\beta}_j^{\mathrm{T}}\boldsymbol{\theta}_{D_j}^{\star}} = \mathbf{e}_{j+1}^{\mathrm{T}}(\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-1}\ \mathbf{H}^{\mathrm{T}}\mathbf{G}(\widehat{\boldsymbol{\theta}}_G^{\star} + \widehat{\boldsymbol{\theta}}_{D_j}^{\star})$. We have shown $\mathbf{e}_{j+1}^{\mathrm{T}}(\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{G}\widehat{\boldsymbol{\theta}}_G^{\star}$ in the last section. For $\mathbf{e}_{j+1}^{\mathrm{T}}(\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{G}\widehat{\boldsymbol{\theta}}_{D_j}^{\star}$, $\widehat{\boldsymbol{\theta}}_{D_j}^{\star}$ is also from (4.3.1) such that all the entries in $\{1, \ldots, p\} \setminus \mathcal{A}_{D_j}$ are zeros, and all the entries in $\mathcal{A}_{D_j}$ are from $\overline{\boldsymbol{\theta}}_{D_j,\mathcal{A}_{D_j}}$, a sub-vector of $\overline{\boldsymbol{\theta}}_{W,\mathcal{A}}$ in (A.1.4). By similar calculation, we have

$$\begin{aligned}
\sqrt{n}\left(\widehat{\boldsymbol{\beta}_j^{\mathrm{T}}\boldsymbol{\theta}_{D_j}^{\star}} - \boldsymbol{\beta}_j^{\mathrm{T}}\boldsymbol{\theta}_{D_j}^{\star}\right) &= \sqrt{n}\left[\mathbf{e}_{j+1}^{\mathrm{T}}(\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{G}\widehat{\boldsymbol{\theta}}_{D_j}^{\star} - \boldsymbol{\beta}_j^{\mathrm{T}}\boldsymbol{\theta}_{D_j}^{\star}\right] \\
&= \boldsymbol{\beta}_{j,\mathcal{A}_{D_j}}^{\mathrm{T}}\tilde{\mathbf{I}}_{\mathcal{A}_{D_j}}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}n^{-1/2}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\varepsilon}_Y + \mathbf{e}_{j+1}^{\mathrm{T}}\mathbf{J}^{-1}n^{-1/2}\mathbf{H}^{\mathrm{T}}\boldsymbol{\varepsilon}_{G,\mathcal{A}_{D_j}}\boldsymbol{\theta}_{D_j,\mathcal{A}_{D_j}}^{\star} \\
&+ \mathit{Error}, \qquad \text{(A.2.5)}
\end{aligned}$$

where $\boldsymbol{\varepsilon}_{G,\mathcal{A}_{D_j}} \in \mathbb{R}^{n\times s_{D_j}}$ is a sub-matrix of $\boldsymbol{\varepsilon}_G$ consisting of all the columns in $\mathcal{A}_{D_j}$, $\tilde{\mathbf{I}}_{\mathcal{A}_{D_j}} \in \mathbb{R}^{(1+q+r+s)\times s_{D_j}}$ is defined before Theorem 4.4.2, and the $\mathit{Error} = o_p(1)$ by the same arguments as (A.2.1).

Hence combining the equation above and (A.2.1), we have

$$\sqrt{n}\left[\widehat{\delta}_{\mathbf{x},j}(\mathbf{t}_j)-\delta_{\mathbf{x},j}(\mathbf{t}_j)\right]=\boldsymbol{\varrho}^{\mathrm{T}}\frac{1}{\sqrt{n}}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\varepsilon}_Y+\boldsymbol{\varkappa}^{\mathrm{T}}\frac{1}{\sqrt{n}}\mathbf{H}^{\mathrm{T}}\begin{bmatrix}\boldsymbol{\varepsilon}_{G,\mathcal{A}_G} & \boldsymbol{\varepsilon}_{G,\mathcal{A}_{D_j}}\end{bmatrix}\begin{bmatrix}\boldsymbol{\theta}^{\star}_{G,\mathcal{A}_G}\\ \boldsymbol{\theta}^{\star}_{D_j,\mathcal{A}_{D_j}}\end{bmatrix}+o_p(1) \tag{A.2.6}$$

where with a slight abuse of notation, we define

$$\boldsymbol{\varrho}=\boldsymbol{\Sigma}^{-1}\left(\tilde{\mathbf{I}}_{\mathcal{A}_G}\boldsymbol{\beta}_{j,\mathcal{A}_G}+\tilde{\mathbf{I}}_{\mathcal{A}_{D_j}}\boldsymbol{\beta}_{j,\mathcal{A}_{D_j}}\right)\in\mathbb{R}^{1+q+r+s},$$

and $\boldsymbol{\varkappa}=\mathbf{J}^{-1}\mathbf{e}_{j+1}\in\mathbb{R}^{1+q+r}$ is the same as Section A.2.1. Since the only difference between (A.2.6) and (A.2.4) is that (A.2.6) extends $\boldsymbol{\varepsilon}_{G,\mathcal{A}_G}\boldsymbol{\theta}^{\star}_{G,\mathcal{A}_G}$ to $[\boldsymbol{\varepsilon}_{G,\mathcal{A}_G}\ \boldsymbol{\varepsilon}_{G,\mathcal{A}_{D_j}}][\boldsymbol{\theta}^{\star,\mathrm{T}}_{G,\mathcal{A}_G}\ \boldsymbol{\theta}^{\star,\mathrm{T}}_{D_j,\mathcal{A}_{D_j}}]^{\mathrm{T}}$, following the same method in (A.2.4), we obtain the desired result. □

A.2.3 Proof of Remark 4.4.1

For $k=1,\dots,1+q+r+s$, by (A.1.34), we have

$$\begin{aligned}
&\sqrt{n}\left(\widehat{\theta}^{\star}_{\mathcal{A},k}-\theta^{\star}_{\mathcal{A},k}\right)=\tilde{\mathbf{e}}_k^{\mathrm{T}}\sqrt{n}\left(\overline{\boldsymbol{\theta}}_{W,\mathcal{A}}-\boldsymbol{\theta}^{\star}_{W,\mathcal{A}}\right)\\
=\;&\tilde{\mathbf{e}}_k^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\frac{1}{\sqrt{n}}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\varepsilon}_Y \qquad\qquad (\text{A.2.7})\\
-\;&\tilde{\mathbf{e}}_k^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\sqrt{n}\left[\mathbf{0}^{\mathrm{T}}_{1+q+r},\boldsymbol{\nu}_3^{\mathrm{T}}(\overline{\boldsymbol{\theta}}_{\mathcal{A}_{G,D}})\right]^{\mathrm{T}}-\tilde{\mathbf{e}}_k^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\sqrt{n}\left(\frac{\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\mathbf{W}_{\mathcal{A}}}{n}-\boldsymbol{\Sigma}\right)\left(\overline{\boldsymbol{\theta}}_{W,\mathcal{A}}-\boldsymbol{\theta}^{\star}_{W,\mathcal{A}}\right)\\
=\;&\boldsymbol{\varrho}^{\mathrm{T}}\frac{1}{\sqrt{n}}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\varepsilon}_Y+o_p(1). \qquad\qquad (\text{A.2.8})
\end{aligned}$$

In (A.2.7), under the conditions in Remark 4.4.1, the last two terms are $o_p(1)$. In (A.2.8), with a slight abuse of notation, set $\boldsymbol{\varrho}=\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{e}}_k\in\mathbb{R}^{1+q+r+s}$. Following the same method in Section A.2.1, *i.e.*, obtaining the asymptotic normality of each treatment set $\mathcal{N}_j$ for $j=0,1,\dots,q$ and then adding them up, we have the desired result. □

A.2.4 Proof of $\widehat{\zeta}_{\mathbf{x},j}(\mathbf{t}_0)$ and $\widehat{\zeta}_{\mathbf{x},j}(\mathbf{t}_j)$

Since $\widehat{\zeta}_{\mathbf{x},j}(\mathbf{t}_0) = \theta^\star_{T,j} + \boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{\theta}^\star_{D_j} + \mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\boldsymbol{\theta}^\star_{D_j}$, from (4.3.3), we have $\widehat{\zeta}_{\mathbf{x},j}(\mathbf{t}_0) = \widehat{\theta}^\star_{T,j} + \widehat{\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{\theta}^\star_{D_j}} + \mathbf{x}^{\mathrm{T}}\widehat{\mathbf{A}^{\mathrm{T}}\boldsymbol{\theta}^\star_{D_j}}$. From the proof of Remark 4.4.1, by setting $\tilde{\mathbf{e}}_k$ with $\tilde{\mathbf{e}}_{j+1}$, we have

$$\sqrt{n}\left(\widehat{\theta}^\star_{T,j} - \theta^\star_{T,j}\right) = \tilde{\mathbf{e}}_{j+1}^{\mathrm{T}}\sqrt{n}\left(\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} - \boldsymbol{\theta}^\star_{W,\mathcal{A}}\right) = \tilde{\mathbf{e}}_{j+1}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\frac{1}{\sqrt{n}}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\varepsilon}_Y + o_p(1).$$

From (4.3.3), $\widehat{\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{\theta}^\star_{D_j}} = \mathbf{e}_1^{\mathrm{T}}(\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{G}\widehat{\boldsymbol{\theta}}^\star_{D_j}$, where $\widehat{\boldsymbol{\theta}}^\star_{D_j}$ is from (4.3.1) such that all the entries in $\{1,\dots,p\}\setminus\mathcal{A}_{D_j}$ are zeros and all the entries in $\mathcal{A}_{D_j}$ are from $\overline{\boldsymbol{\theta}}_{D_j,\mathcal{A}_{D_j}}$, a sub-vector of $\overline{\boldsymbol{\theta}}_{W,\mathcal{A}}$ in (A.1.4). By some calculation, we have

$$\begin{aligned}\sqrt{n}\left(\widehat{\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{\theta}^\star_{D_j}} - \boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{\theta}^\star_{D_j}\right) &= \boldsymbol{\beta}_{0,\mathcal{A}_{D_j}}^{\mathrm{T}}\tilde{\mathbf{I}}_{\mathcal{A}_{D_j}}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}n^{-1/2}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\varepsilon}_Y \\ &+ \mathbf{e}_1^{\mathrm{T}}\mathbf{J}^{-1}n^{-1/2}\mathbf{H}^{\mathrm{T}}\boldsymbol{\varepsilon}_{G,\mathcal{A}_{D_j}}\tilde{\mathbf{I}}_{\mathcal{A}_{D_j}}^{\mathrm{T}}\boldsymbol{\theta}^\star_{W,\mathcal{A}} + Error,\end{aligned}$$

where

$$\begin{aligned}Error &= -\boldsymbol{\beta}_{0,\mathcal{A}_{D_j}}^{\mathrm{T}}\tilde{\mathbf{I}}_{\mathcal{A}_{D_j}}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\sqrt{n}\left[\mathbf{0}_{1+q+r}^{\mathrm{T}}, \boldsymbol{\nu}_3^{\mathrm{T}}(\overline{\boldsymbol{\theta}}_{\mathcal{A}_{G,D}})\right]^{\mathrm{T}} \\ &- \boldsymbol{\beta}_{0,\mathcal{A}_{D_j}}^{\mathrm{T}}\tilde{\mathbf{I}}_{\mathcal{A}_{D_j}}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\sqrt{n}\left(\frac{\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\mathbf{W}_{\mathcal{A}}}{n} - \boldsymbol{\Sigma}\right)\left(\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} - \boldsymbol{\theta}^\star_{W,\mathcal{A}}\right) \\ &+ \mathbf{e}_1^{\mathrm{T}}(\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\boldsymbol{\varepsilon}_{G,\mathcal{A}_{D_j}}\sqrt{n}\tilde{\mathbf{I}}_{\mathcal{A}_{D_j}}^{\mathrm{T}}\left(\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} - \boldsymbol{\theta}^\star_{W,\mathcal{A}}\right) \\ &+ \mathbf{e}_1^{\mathrm{T}}\left[\left(\frac{\mathbf{H}^{\mathrm{T}}\mathbf{H}}{n}\right)^{-1} - \mathbf{J}^{-1}\right]n^{-1/2}\mathbf{H}^{\mathrm{T}}\boldsymbol{\varepsilon}_{G,\mathcal{A}_{D_j}}\boldsymbol{\theta}^\star_{D_j,\mathcal{A}_{D_j}}.\end{aligned}$$

Follow the same arguments in the Section A.2.1, we have $Error = o_p(1)$.

From (4.3.3), we have $\widehat{\mathbf{A}^{\mathrm{T}}\boldsymbol{\theta}^\star_{D_j}} = \mathbf{I}_X^{\mathrm{T}}(\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{G}\widehat{\boldsymbol{\theta}}^\star_{D_j}$, so that

$$\begin{aligned}&\sqrt{n}\left(\mathbf{x}^{\mathrm{T}}\widehat{\mathbf{A}^{\mathrm{T}}\boldsymbol{\theta}^\star_{D_j}} - \mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\boldsymbol{\theta}^\star_{D_j}\right) \\ = \;& \mathbf{x}^{\mathrm{T}}\mathbf{A}_{\mathcal{A}_{D_j}\times[r]}^{\mathrm{T}}\tilde{\mathbf{I}}_{\mathcal{A}_{D_j}}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\frac{\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\varepsilon}_Y}{\sqrt{n}} + \mathbf{x}^{\mathrm{T}}\mathbf{I}_X^{\mathrm{T}}\mathbf{J}^{-1}\frac{\mathbf{H}^{\mathrm{T}}\boldsymbol{\varepsilon}_{G,\mathcal{A}_{D_j}}}{\sqrt{n}}\tilde{\mathbf{I}}_{\mathcal{A}_{D_j}}^{\mathrm{T}}\boldsymbol{\theta}^\star_{W,\mathcal{A}} + Error,\end{aligned}$$

where

$$
\begin{aligned}
Error &= -\mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}_{\mathcal{A}_{D_j}\times[r]}\tilde{\mathbf{I}}^{\mathrm{T}}_{\mathcal{A}_{D_j}}\boldsymbol{\Sigma}^{-1}\sqrt{n}\left[\mathbf{0}^{\mathrm{T}}_{1+q+r}, \boldsymbol{\nu}_3^{\mathrm{T}}(\overline{\boldsymbol{\theta}}_{\mathcal{A}_{G,D}})\right]^{\mathrm{T}} \\
&- \mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}_{\mathcal{A}_{D_j}\times[r]}\tilde{\mathbf{I}}^{\mathrm{T}}_{\mathcal{A}_{D_j}}\boldsymbol{\Sigma}^{-1}\sqrt{n}\left(\frac{\mathbf{W}^{\mathrm{T}}_{\mathcal{A}}\mathbf{W}_{\mathcal{A}}}{n} - \boldsymbol{\Sigma}\right)\left(\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} - \boldsymbol{\theta}^{\star}_{W,\mathcal{A}}\right) \\
&+ \mathbf{x}^{\mathrm{T}}\mathbf{I}^{\mathrm{T}}_{X}(\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\boldsymbol{\varepsilon}_{G,\mathcal{A}_{D_j}}\sqrt{n}\tilde{\mathbf{I}}^{\mathrm{T}}_{\mathcal{A}_{D_j}}\left(\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} - \boldsymbol{\theta}^{\star}_{W,\mathcal{A}}\right) \\
&+ \mathbf{x}^{\mathrm{T}}\mathbf{I}^{\mathrm{T}}_{X}\left[\left(\frac{\mathbf{H}^{\mathrm{T}}\mathbf{H}}{n}\right)^{-1} - \mathbf{J}^{-1}\right]n^{-1/2}\mathbf{H}^{\mathrm{T}}\boldsymbol{\varepsilon}_{G,\mathcal{A}_{D_j}}\boldsymbol{\theta}^{\star}_{D_j,\mathcal{A}_{D_j}}.
\end{aligned}
$$

Follow the same arguments in Section A.2.1, we have $Error = o_p(1)$.

Therefore, we have

$$
\begin{aligned}
&\sqrt{n}\left[\widehat{\zeta}_{\mathbf{x},j}(\mathbf{t}_0) - \zeta_{\mathbf{x},j}(\mathbf{t}_0)\right] = \sqrt{n}\left(\widehat{\theta}^{\star}_{T,j} - \theta^{\star}_{T,j}\right) + \sqrt{n}\left(\widehat{\boldsymbol{\beta}^{\mathrm{T}}_0\boldsymbol{\theta}^{\star}_{D_j}} - \boldsymbol{\beta}^{\mathrm{T}}_0\boldsymbol{\theta}^{\star}_{D_j}\right) \\
+ \;& \sqrt{n}\left(\mathbf{x}^{\mathrm{T}}\widehat{\mathbf{A}^{\mathrm{T}}\boldsymbol{\theta}^{\star}_{D_j}} - \mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\boldsymbol{\theta}^{\star}_{D_j}\right) = \boldsymbol{\varrho}^{\mathrm{T}}\frac{1}{\sqrt{n}}\mathbf{W}^{\mathrm{T}}_{\mathcal{A}}\boldsymbol{\varepsilon}_Y + \boldsymbol{\varkappa}^{\mathrm{T}}\frac{1}{\sqrt{n}}\mathbf{H}^{\mathrm{T}}\boldsymbol{\varepsilon}_{G,\mathcal{A}_{D_j}}\boldsymbol{\theta}^{\star}_{G,\mathcal{A}_{D_j}} + o_p(1),
\end{aligned}
$$

where with a slight abuse of notation, we denote

$$
\boldsymbol{\varrho} = \boldsymbol{\Sigma}^{-1}\left(\tilde{\mathbf{e}}_{j+1} + \tilde{\mathbf{I}}_{\mathcal{A}_{D_j}}\boldsymbol{\beta}_{0,\mathcal{A}_{D_j}} + \tilde{\mathbf{I}}_{\mathcal{A}_{D_j}}\mathbf{A}_{\mathcal{A}_{D_j}\times[r]}\mathbf{x}\right), \quad \boldsymbol{\varkappa} = \mathbf{J}^{-1}\left(\mathbf{e}_1 + \mathbf{I}_X\mathbf{x}\right).
$$

Following the same arguments in Section A.2.1 and only replacing $\boldsymbol{\varepsilon}_{G,\mathcal{A}_G}\boldsymbol{\theta}^{\star}_{G,\mathcal{A}_G}$ in (A.2.4) with $\boldsymbol{\varepsilon}_{G,\mathcal{A}_{D_j}}\boldsymbol{\theta}^{\star}_{G,\mathcal{A}_{D_j}}$, we have the desired result.

In the case of $\widehat{\zeta}_{\mathbf{x},j}(\mathbf{t}_j)$, since $\widehat{\zeta}_{\mathbf{x},j}(\mathbf{t}_j) = \widehat{\zeta}_{\mathbf{x},j}(\mathbf{t}_0) + \widehat{\boldsymbol{\beta}^{\mathrm{T}}_j\boldsymbol{\theta}^{\star}_{D_j}}$ and $\zeta_{\mathbf{x},j}(\mathbf{t}_j) = \zeta_{\mathbf{x},j}(\mathbf{t}_0) + \boldsymbol{\beta}^{\mathrm{T}}_j\boldsymbol{\theta}^{\star}_{D_j}$, combining the results above and (A.2.5), we have

$$
\begin{aligned}
\sqrt{n}\left[\widehat{\zeta}_{\mathbf{x},j}(\mathbf{t}_j) - \zeta_{\mathbf{x},j}(\mathbf{t}_j)\right] &= \sqrt{n}\left(\widehat{\theta}^{\star}_{T,j} - \theta^{\star}_{T_j}\right) + \sqrt{n}\left(\widehat{\boldsymbol{\beta}^{\mathrm{T}}_0\boldsymbol{\theta}^{\star}_{D_j}} - \boldsymbol{\beta}^{\mathrm{T}}_0\boldsymbol{\theta}^{\star}_{D_j}\right) \\
&+ \sqrt{n}\left(\mathbf{x}^{\mathrm{T}}\widehat{\mathbf{A}^{\mathrm{T}}\boldsymbol{\theta}^{\star}_{D_j}} - \mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\boldsymbol{\theta}^{\star}_{D_j}\right) + \sqrt{n}\left(\widehat{\boldsymbol{\beta}^{\mathrm{T}}_j\boldsymbol{\theta}^{\star}_{D_j}} - \boldsymbol{\beta}^{\mathrm{T}}_j\boldsymbol{\theta}^{\star}_{D_j}\right) \\
&= \boldsymbol{\varrho}^{\mathrm{T}}\frac{1}{\sqrt{n}}\mathbf{W}^{\mathrm{T}}_{\mathcal{A}}\boldsymbol{\varepsilon}_Y + \boldsymbol{\varkappa}^{\mathrm{T}}\frac{1}{\sqrt{n}}\mathbf{H}^{\mathrm{T}}\boldsymbol{\varepsilon}_{G,\mathcal{A}_{D_j}}\boldsymbol{\theta}^{\star}_{G,\mathcal{A}_{D_j}} + o_p(1),
\end{aligned}
$$

where with a slight abuse of notation, we denote

$$\begin{aligned}
\boldsymbol{\varrho} &= \boldsymbol{\Sigma}^{-1}\left(\tilde{\mathbf{e}}_{j+1} + \tilde{\mathbf{I}}_{\mathcal{A}_{D_j}}\boldsymbol{\beta}_{0,\mathcal{A}_{D_j}} + \tilde{\mathbf{I}}_{\mathcal{A}_{D_j}}\mathbf{A}_{\mathcal{A}_{D_j}\times[r]}\mathbf{x} + \tilde{\mathbf{I}}_{\mathcal{A}_{D_j}}\boldsymbol{\beta}_{j,\mathcal{A}_{D_j}}\right), \\
\boldsymbol{\varkappa} &= \mathbf{J}^{-1}\left(\mathbf{e}_1 + \mathbf{I}_X\mathbf{x} + \mathbf{e}_{j+1}\right).
\end{aligned}$$

Following the same arguments, we have the desired result. □

A.3 Proofs of Theorems in Section 4.6

A.3.1 Proof of Theorem 4.6.1

The proof follows the same idea of the proof in Theorem 4.4.1 with the objective function (4.3.1) replaced by the overlapping group SCAD penalty (A.3.1) below. Specifically, **Step 1** shows the existence of the local minimizer $\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} \in \mathbb{R}^{1+q+r+s}$ when the global objective function (4.6.1) is constrained on the space $\mathcal{B} = \{\boldsymbol{\theta}_W \in \mathbb{R}^{1+q+r+p+pq} : \boldsymbol{\theta}_{\mathcal{A}^c} = \mathbf{0}_{(q+1)p-s}\}$ such that $\|\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} - \boldsymbol{\theta}^\star_{W,\mathcal{A}}\|_2 = O_p(\sqrt{s/n})$. Then **Step 2** shows that extending $\overline{\boldsymbol{\theta}}_{W,\mathcal{A}}$ with $(q+1)p-s$ zeros in the location of $\mathcal{A}^c$ forms a local minimizer $\overline{\boldsymbol{\theta}}_W \in \mathbb{R}^{1+q+r+p+pq}$ of (4.6.1).

Step 1. We only list the differences compared with the **Step 1** in the proof of Theorem 4.4.1. Denote the objective function in (4.6.1) as $Q_{\mathrm{H}}(\boldsymbol{\theta}_W)$ for $\boldsymbol{\theta}_W \in \mathbb{R}^{1+q+r+p+pq}$, *i.e.* ,

$$Q_{\mathrm{H}}(\boldsymbol{\theta}_W) = \frac{1}{2n}\|\mathbf{Y} - \mathbf{W}\boldsymbol{\theta}_W\|_2^2 + \sum_{h=1}^{p} p_\lambda(\|\boldsymbol{\eta}_h\|_2) + \sum_{j=1}^{q}\sum_{h=1}^{p} p_\lambda(|\theta_{D_j,h}|), \tag{A.3.1}$$

where the subscript $_{\mathrm{H}}$ indicates heredity. Constraining $Q_{\mathrm{H}}(\boldsymbol{\theta}_W)$ in $\mathcal{B}$, we have

$$\overline{Q}_{\mathrm{H}}(\boldsymbol{\theta}_{W,\mathcal{A}}) = \frac{1}{2n}\|\mathbf{Y} - \mathbf{W}_{\mathcal{A}}\boldsymbol{\theta}_{W,\mathcal{A}}\|_2^2 + P_{\mathrm{H}}(\boldsymbol{\theta}_{\mathcal{A}_{G,D}}), \tag{A.3.2}$$

with the local minimizer denoted as $\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} \in \mathbb{R}^{1+q+r+s}$, where

$$P_{\mathrm{H}}(\boldsymbol{\theta}_{\mathcal{A}_{G,D}}) = \sum_{h\in\mathcal{A}_G} p_\lambda(\|\boldsymbol{\eta}_h\|_2) + \sum_{j=1}^{q}\sum_{h\in\mathcal{A}_{D_j}} p_\lambda(|\theta_{D_j,h}|),$$

and $\boldsymbol{\theta}_{\mathcal{A}_{G,D}}$ is defined in (A.1.8). In $P_{\mathrm{H}}(\boldsymbol{\theta}_{\mathcal{A}_{G,D}})$, given $h \in \mathcal{A}_G$, $\boldsymbol{\eta}_h$ only contains $\theta_{G,h}$ and those $\theta_{D_j,h}$ with $\theta^{\star}_{D_j,h} \neq 0$ for $j = 1, \ldots, q$, rather than $\boldsymbol{\eta}_h = [\theta_{G,h}, \theta_{D_1,h}, \ldots, \theta_{D_q,h}]^{\mathrm{T}} \in \mathbb{R}^{1+q}$ in $Q_{\mathrm{H}}(\boldsymbol{\theta}_W)$. Applying Taylor expansion to $\overline{Q}_{\mathrm{H}}(\boldsymbol{\theta}_{W,\mathcal{A}})$ in (A.3.2), we have $\overline{Q}_{\mathrm{H}}(\boldsymbol{\theta}_{W,\mathcal{A}}) - \overline{Q}_{\mathrm{H}}(\boldsymbol{\theta}^{\star}_{W,\mathcal{A}}) = -\boldsymbol{\nu}_{\mathrm{H}}^{\mathrm{T}}(\boldsymbol{\theta}_{W,\mathcal{A}} - \boldsymbol{\theta}^{\star}_{W,\mathcal{A}}) + 2^{-1}(\boldsymbol{\theta}_{W,\mathcal{A}} - \boldsymbol{\theta}^{\star}_{W,\mathcal{A}})^{\mathrm{T}}\overline{\mathbf{D}}_{\mathrm{H}}(\boldsymbol{\theta}_{W,\mathcal{A}} - \boldsymbol{\theta}^{\star}_{W,\mathcal{A}})$, where

$$\begin{aligned}
\boldsymbol{\nu}_{\mathrm{H}} &= -\left.\frac{\partial \overline{Q}_{\mathrm{H}}}{\partial \boldsymbol{\theta}_{W,\mathcal{A}}}\right|_{\boldsymbol{\theta}^{\star}_{W,\mathcal{A}}} = \boldsymbol{\nu}_{\mathrm{H},1} - \boldsymbol{\nu}_{\mathrm{H},2},\ \boldsymbol{\nu}_{\mathrm{H},1} = \frac{1}{n}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\left(\mathbf{Y} - \mathbf{W}_{\mathcal{A}}\boldsymbol{\theta}^{\star}_{W,\mathcal{A}}\right), \\
\boldsymbol{\nu}_{\mathrm{H},2} &= \begin{bmatrix} \mathbf{0}^{\mathrm{T}}_{1+q+r} & \boldsymbol{\nu}^{\mathrm{T}}_{\mathrm{H},3}(\boldsymbol{\theta}^{\star}_{\mathcal{A}_{G,D}}) \end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^{1+q+r+s}, \\
\overline{\mathbf{D}}_{\mathrm{H}} &= \frac{1}{n}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\mathbf{W}_{\mathcal{A}} + \begin{bmatrix} \mathbf{O}_{(1+q+r)\times(1+q+r)} & \mathbf{O}_{(1+q+r)\times s} \\ \mathbf{O}_{s\times(1+q+r)} & \overline{\mathbf{D}}_{\mathrm{H},1} \end{bmatrix}.
\end{aligned} \tag{A.3.3}$$

Since $\boldsymbol{\nu}_{\mathrm{H},3}(\boldsymbol{\theta}^{\star}_{\mathcal{A}_{G,D}}) = \partial P_{\mathrm{H}}(\boldsymbol{\theta}_{\mathcal{A}_{G,D}})/\partial \boldsymbol{\theta}_{\mathcal{A}_{G,D}} \in \mathbb{R}^s$ evaluated at $\boldsymbol{\theta}^{\star}_{\mathcal{A}_{G,D}}$, the first s_G entries in $\boldsymbol{\nu}_{\mathrm{H},3}(\boldsymbol{\theta}^{\star}_{\mathcal{A}_{G,D}})$ are

$$\left.\frac{\partial P_{\mathrm{H}}}{\partial \theta_{G,h}}\right|_{\theta^{\star}_{G,h}} = p'_{\lambda}(\|\boldsymbol{\eta}^{\star}_h\|_2)\frac{\theta^{\star}_{G,h}}{\|\boldsymbol{\eta}^{\star}_h\|_2} \text{ for } h \in \mathcal{A}_G, \tag{A.3.4}$$

where $\boldsymbol{\eta}^{\star}_h$ is a vector containing $\theta^{\star}_{G,h}$ and all the nonzero $\theta^{\star}_{D_j,h}$, $j = 1, \ldots, q$. The rest $s_{D_1} + \cdots + s_{D_q}$ entries are

$$\left.\frac{\partial P_{\mathrm{H}}}{\partial \theta_{D_j,h}}\right|_{\theta^{\star}_{D_j,h}} = p'_{\lambda}(\|\boldsymbol{\eta}^{\star}_h\|_2)\frac{\theta^{\star}_{D_j,h}}{\|\boldsymbol{\eta}^{\star}_h\|_2} + p'_{\lambda}(|\theta^{\star}_{D_j,h}|)\mathrm{sgn}(\theta^{\star}_{D_j,h}) \text{ for } j = 1, \ldots, q,\ h \in \mathcal{A}_{D_j}. \tag{A.3.5}$$

We also have $\overline{\mathbf{D}}_{\mathrm{H},1} = \partial^2 P_{\mathrm{H}}(\boldsymbol{\theta}_{\mathcal{A}_{G,D}})/\partial \boldsymbol{\theta}_{\mathcal{A}_{G,D}}\partial \boldsymbol{\theta}^{\mathrm{T}}_{\mathcal{A}_{G,D}} \in \mathbb{R}^{s\times s}$ evaluated at $\widetilde{\boldsymbol{\theta}}_{\mathcal{A}_{G,D}}$, which is between $\boldsymbol{\theta}_{\mathcal{A}_{G,D}}$ and $\boldsymbol{\theta}^{\star}_{\mathcal{A}_{G,D}}$. For $\partial P_{\mathrm{H}}/\partial \theta_{G,h} = p'_{\lambda}(\|\boldsymbol{\eta}_h\|_2)\theta_{G,h}\|\boldsymbol{\eta}\|_2^{-1}$ given $h \in \mathcal{A}_G$, which is a function of $\boldsymbol{\eta}_h$, *i.e.* , $\theta_{G,h}$ and those $\theta_{D_j,h}$ with $\theta^{\star}_{D_j,h} \neq 0$, we have

$$\begin{aligned}
\frac{\partial^2 P_{\mathrm{H}}}{\partial \theta^2_{G,h}} &= p''_{\lambda}(\|\boldsymbol{\eta}_h\|_2)\frac{\theta^2_{G,h}}{\|\boldsymbol{\eta}_h\|_2^2} + \frac{p'_{\lambda}(\|\boldsymbol{\eta}_h\|_2)}{\|\boldsymbol{\eta}_h\|_2}\left[1 - \frac{\theta^2_{G,h}}{\|\boldsymbol{\eta}_h\|_2^2}\right], \\
\frac{\partial^2 P_{\mathrm{H}}}{\partial \theta_{G,h}\partial \theta_{D_j,h}} &= p''_{\lambda}(\|\boldsymbol{\eta}_h\|_2)\frac{\theta_{G,h}\theta_{D_j,h}}{\|\boldsymbol{\eta}_h\|_2^2} - p'_{\lambda}(\|\boldsymbol{\eta}_h\|_2)\frac{\theta_{G,h}\theta_{D_j,h}}{\|\boldsymbol{\eta}_h\|_2^3},\ \text{if } \theta^{\star}_{D_j,h} \neq 0 \text{ given } h \in \mathcal{A}_G,
\end{aligned}$$

and other partial derivatives of $\partial P_{\mathrm{H}}/\partial\theta_{G,h}$ are all zeros. For $j = 1, \ldots, q$ and $h \in \mathcal{A}_{D_j}$, since $\partial P_{\mathrm{H}}/\partial\theta_{D_j,h} = p'_\lambda(\|\boldsymbol{\eta}\|_2)\theta_{D_j,h}\|\boldsymbol{\eta}\|_2^{-1} + p'_\lambda(|\theta_{D_j,h}|)\mathrm{sgn}(\theta_{D_j,h})$ which is a function of $\boldsymbol{\eta}_h$, *i.e.* , $\theta_{G,h}$, $\theta_{D_j,h}$, and other $\theta_{D_k,h}$ with $\theta^\star_{D_k,h} \neq 0$, we have

$$
\begin{aligned}
\frac{\partial^2 P_{\mathrm{H}}}{\partial\theta^2_{D_j,h}} &= p''_\lambda(\|\boldsymbol{\eta}_h\|_2)\frac{\theta^2_{D_j,h}}{\|\boldsymbol{\eta}_h\|_2^2} + \frac{p'_\lambda(\|\boldsymbol{\eta}_h\|_2)}{\|\boldsymbol{\eta}_h\|_2}\left[1 - \frac{\theta^2_{D_j,h}}{\|\boldsymbol{\eta}_h\|_2^2}\right] + p''_\lambda(|\theta_{D_j,h}|),\\
\frac{\partial^2 P_{\mathrm{H}}}{\partial\theta_{D_j,h}\partial\theta_{G,h}} &= p''_\lambda(\|\boldsymbol{\eta}_h\|_2)\frac{\theta_{G,h}\theta_{D_j,h}}{\|\boldsymbol{\eta}_h\|_2^2} - p'_\lambda(\|\boldsymbol{\eta}_h\|_2)\frac{\theta_{G,h}\theta_{D_j,h}}{\|\boldsymbol{\eta}_h\|_2^3},\\
\frac{\partial^2 P_{\mathrm{H}}}{\partial\theta_{D_j,h}\partial\theta_{D_k,h}} &= p''_\lambda(\|\boldsymbol{\eta}_h\|_2)\frac{\theta_{D_j,h}\theta_{D_k,h}}{\|\boldsymbol{\eta}_h\|_2^2} - p'_\lambda(\|\boldsymbol{\eta}_h\|_2)\frac{\theta_{D_j,h}\theta_{D_k,h}}{\|\boldsymbol{\eta}_h\|_2^3},\ \theta^\star_{D_k,h} \neq 0,
\end{aligned}
$$

and other partial derivatives of $\partial P_{\mathrm{H}}/\partial\theta_{D_j,h}$ are zeros.

Since we consider SCAD or MCP penalty, $p'_\lambda(|t|) = 0$ if $|t| > \gamma\lambda$ with constant γ. For $\boldsymbol{\nu}_{\mathrm{H}}$ in (A.3.3), we claim that $\boldsymbol{\nu}_{\mathrm{H}} = \boldsymbol{\nu}_{\mathrm{H},1}$ if n is sufficiently large. This follows from the fact that $\|\boldsymbol{\eta}^\star_h\|_2 \geq \theta^\star_{D_j,h} \gg \gamma\lambda$ under Condition A.1.4. Thus by (A.3.4) and (A.3.5), we have $\boldsymbol{\nu}_{\mathrm{H},3}(\boldsymbol{\theta}^\star_{\mathcal{A}_{G,D}}) = \mathbf{0}_s$ and $\boldsymbol{\nu}_{\mathrm{H},2} = \mathbf{0}_{1+q+r+s}$. For $\overline{\mathbf{D}}_{\mathrm{H}}$ in (A.3.3), we claim that $\overline{\mathbf{D}}_{\mathrm{H}} = n^{-1}\mathbf{W}^{\mathrm{T}}_{\mathcal{A}}\mathbf{W}_{\mathcal{A}}$ if n is sufficiently large. This follows from the fact that $\widetilde{\boldsymbol{\theta}}_{\mathcal{A}_{G,D}}$ is between $\boldsymbol{\theta}_{\mathcal{A}_{G,D}}$ and $\boldsymbol{\theta}^\star_{\mathcal{A}_{G,D}}$. For $\widetilde{\theta}_l$ as an entry of $\widetilde{\boldsymbol{\theta}}_{\mathcal{A}_{G,D}}$, we have $|\widetilde{\theta}_l| \geq |\theta^\star_l| - |\theta^\star_l - \widetilde{\theta}_l| \geq d_n - \|\boldsymbol{\theta}^\star_{W,\mathcal{A}} - \boldsymbol{\theta}_{W,\mathcal{A}}\|_2 = d_n - \tau\sqrt{s/n} \geq d_n/2 \gg \gamma\lambda$ when $\boldsymbol{\theta}_{W,\mathcal{A}}$ is from H_n. Thus for SCAD or MCP penalties with large n, we have $p''(\|\widetilde{\boldsymbol{\eta}}_h\|_2) = p'(\|\widetilde{\boldsymbol{\eta}}_h\|_2) = p'(\|\widetilde{\theta}_{D_j,h}\|_2) = 0$, which implies that $\overline{\mathbf{D}}_{\mathrm{H},1} = \mathbf{O}_{s\times s}$.

Following similar arguments in (A.1.17)–(A.1.20), we have $\|\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} - \boldsymbol{\theta}^\star_{W,\mathcal{A}}\|_2 = O_p(\sqrt{s/n})$.

Step 2. Let $\overline{\boldsymbol{\theta}}_W \in \mathbb{R}^{1+q+r+p+pq}$ be the vector constructed by extending $\overline{\boldsymbol{\theta}}_{W,\mathcal{A}}$ in **Step 1** with $(1+q)p - s$ zeros in the locations of $\mathcal{A}^c$. By Lemma A.3.1, we only need to show that $\mathbb{P}\{\|\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} - \boldsymbol{\theta}^\star_{W,\mathcal{A}}\|_2 \geq d_n/8\} \to 0$, $\mathbb{P}\{\lambda_{\min}(n^{-1}\mathbf{W}^{\mathrm{T}}_{\mathcal{A}}\mathbf{W}_{\mathcal{A}}) < C_\Sigma/4\} \to 0$, and $\mathbb{P}\{\|n^{-1}\mathbf{W}^{\mathrm{T}}_{\mathcal{A}^c}(\mathbf{Y} - \mathbf{W}\overline{\boldsymbol{\theta}}_W)\|_\infty \geq p'_\lambda(0+)\} \to 0$, which all exist by $\|\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} - \boldsymbol{\theta}^\star_{W,\mathcal{A}}\|_2 = O_p(\sqrt{s/n})$ in **Step 1**, (A.1.12), $\lambda_{\min}(\boldsymbol{\Sigma}) \geq C_\Sigma$, and (A.1.30). □

Lemma A.3.1. *Suppose $\boldsymbol{\theta}^\star_W$ satisfies $\theta^\star_{G,h} = 0 \implies \theta^\star_{D_j,h} = 0,\ j = 1, \ldots, q$ for $h = 1, \ldots, p$. Given $\mathbf{W} \in \mathbb{R}^{n\times(1+q+r+p+pq)}$, $\mathbf{Y} \in \mathbb{R}^n$, $Q_{\mathrm{H}}(\boldsymbol{\theta}_W)$ in* (A.3.1)*, and $\overline{Q}_{\mathrm{H}}(\boldsymbol{\theta}_{W,\mathcal{A}})$ in* (A.3.2) *with $p_\lambda()$ in the form of SCAD or MCP, suppose that $\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} \in \mathbb{R}^{1+q+r+s}$ is a local minimizer of $\overline{Q}_{\mathrm{H}}(\boldsymbol{\theta}_{W,\mathcal{A}})$ such that $\|\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} - \boldsymbol{\theta}^\star_{W,\mathcal{A}}\|_2 < d_n/8$, where $\boldsymbol{\theta}^\star_W$ and d_n satisfy Condition A.1.4.*

Suppose that $\lambda_{\min}(n^{-1}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\mathbf{W}_{\mathcal{A}}) \geq C_{\Sigma}/4$, where $C_{\Sigma} \leq \lambda_{\min}(\mathbf{\Sigma})$ from Condition A.1.3. Let $\overline{\boldsymbol{\theta}}_W \in \mathbb{R}^{1+q+r+p+pq}$ denote the vector extending $\overline{\boldsymbol{\theta}}_{W,\mathcal{A}}$ with $(q+1)p-s$ zeros in the location of $\boldsymbol{\theta}_{\mathcal{A}^c}$. If $\|n^{-1}\mathbf{W}_{\mathcal{A}^c}^{\mathrm{T}}(\mathbf{Y}-\mathbf{W}\overline{\boldsymbol{\theta}}_W)\|_\infty < p'_\lambda(0+)$, then $\overline{\boldsymbol{\theta}}_W$ is a local minimizer of $Q_{\mathrm{H}}(\boldsymbol{\theta}_W)$ with sufficiently large n.

Proof of Lemma A.3.1. The proof follows the idea similar to the proof for Theorem 1 in (Fan & Lv, 2011).

Since $\overline{\boldsymbol{\theta}}_{W,\mathcal{A}} \in \mathbb{R}^{1+q+r+s}$ is a local minimizer of $\overline{Q}_{\mathrm{H}}(\boldsymbol{\theta}_{W,\mathcal{A}})$, there exists $\delta_1 \in (0, d_n/8)$ such that for all $\boldsymbol{\theta}_{W,\mathcal{A}}$ satisfying $0 < \|\boldsymbol{\theta}_{W,\mathcal{A}} - \overline{\boldsymbol{\theta}}_{W,\mathcal{A}}\|_2 < \delta_1$, we have $\overline{Q}_{\mathrm{H}}(\boldsymbol{\theta}_{W,\mathcal{A}}) - \overline{Q}_{\mathrm{H}}(\overline{\boldsymbol{\theta}}_{W,\mathcal{A}}) > 0$ for sufficiently large n. This is from the fact that for a specific entry of $\boldsymbol{\theta}_{\mathcal{A}_{G,D}}$ denoted as θ_l, we have $|\overline{\theta}_l| \geq |\theta_l^\star| - |\theta_l^\star - \overline{\theta}_l| \geq d_n - d_n/8 \gg \gamma\lambda > 0$ so that $p_\lambda()$ is differentiable around $\overline{\theta}_l$ when n is large. Thus $\overline{Q}_{\mathrm{H}}(\boldsymbol{\theta}_{W,\mathcal{A}})$ is also differentiable around $\overline{\boldsymbol{\theta}}_{W,\mathcal{A}}$ and with zero gradient at $\overline{\boldsymbol{\theta}}_{W,\mathcal{A}}$. By the Taylor expansion, we have $\overline{Q}_{\mathrm{H}}(\boldsymbol{\theta}_{W,\mathcal{A}}) - \overline{Q}_{\mathrm{H}}(\overline{\boldsymbol{\theta}}_{W,\mathcal{A}}) = 2^{-1}(\boldsymbol{\theta}_{W,\mathcal{A}} - \overline{\boldsymbol{\theta}}_{W,\mathcal{A}})^{\mathrm{T}}(\partial^2\overline{Q}_{\mathrm{H}}/\partial\boldsymbol{\theta}_{W,\mathcal{A}}\partial\boldsymbol{\theta}_{W,\mathcal{A}}^{\mathrm{T}}|_{\widetilde{\boldsymbol{\theta}}_{W,\mathcal{A}}})(\boldsymbol{\theta}_{W,\mathcal{A}} - \overline{\boldsymbol{\theta}}_{W,\mathcal{A}})$ where $\widetilde{\boldsymbol{\theta}}_{W,\mathcal{A}}$ is between $\boldsymbol{\theta}_{W,\mathcal{A}}$ and $\overline{\boldsymbol{\theta}}_{W,\mathcal{A}}$. Since $\widetilde{\boldsymbol{\theta}}_{W,\mathcal{A}}$ is close enough to $\overline{\boldsymbol{\theta}}_{W,\mathcal{A}}$ controlled by δ_1, by the similar arguments in the proof of **Step 1**, we have $|\widetilde{\theta}_l| \gg \gamma\lambda$ where $\widetilde{\theta}_l$ is an entry of $\widetilde{\boldsymbol{\theta}}_{\mathcal{A}_{G,D}}$, so that $\partial^2\overline{Q}_{\mathrm{H}}/\partial\boldsymbol{\theta}_{W,\mathcal{A}}\partial\boldsymbol{\theta}_{W,\mathcal{A}}^{\mathrm{T}}|_{\widetilde{\boldsymbol{\theta}}_{W,\mathcal{A}}} = n^{-1}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\mathbf{W}_{\mathcal{A}}$ and $\overline{Q}_{\mathrm{H}}(\boldsymbol{\theta}_{W,\mathcal{A}}) - \overline{Q}_{\mathrm{H}}(\overline{\boldsymbol{\theta}}_{W,\mathcal{A}}) \geq 2^{-1}\lambda_{\min}(n^{-1}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\mathbf{W}_{\mathcal{A}})\|\boldsymbol{\theta}_{W,\mathcal{A}} - \overline{\boldsymbol{\theta}}_{W,\mathcal{A}}\|_2^2 > 8^{-1}C_\Sigma\|\boldsymbol{\theta}_{W,\mathcal{A}} - \overline{\boldsymbol{\theta}}_{W,\mathcal{A}}\|_2^2 > 0$.

By the continuity of $n^{-1}\mathbf{W}_{\mathcal{A}^c}^{\mathrm{T}}(\mathbf{Y}-\mathbf{W}\boldsymbol{\theta}_W)$ in terms of $\boldsymbol{\theta}_W$ and $p'_\lambda()$ around zero, there exits $\delta_2 > 0$ such that for arbitrary $\boldsymbol{\theta}_W$ such $\|\boldsymbol{\theta}_W - \overline{\boldsymbol{\theta}}_W\|_2 \leq \delta_2$, $\|n^{-1}\mathbf{W}_{\mathcal{A}^c}^{\mathrm{T}}(\mathbf{Y}-\mathbf{W}\boldsymbol{\theta}_W)\|_\infty < p'_\lambda(\delta_2)$.

Let $\delta = \min(\delta_1, \delta_2)$, for arbitrary $\boldsymbol{\theta}_W$ such that $0 < \|\boldsymbol{\theta}_W - \overline{\boldsymbol{\theta}}_W\|_2 < \delta$, we will show that $Q_{\mathrm{H}}(\boldsymbol{\theta}_W) - Q_{\mathrm{H}}(\overline{\boldsymbol{\theta}}_W) > 0$. The entries in $\boldsymbol{\theta}_W$ can be partitioned into $\boldsymbol{\theta}_{W,\mathcal{A}} \in \mathbb{R}^{1+q+r+s}$ and $\boldsymbol{\theta}_{W,\mathcal{A}^c} \in \mathbb{R}^{(q+1)p-s}$. We denote $\boldsymbol{\theta}_W^{\mathcal{B}} \in \mathbb{R}^{1+q+r+p+pq}$ the vector of $\boldsymbol{\theta}_W$ projected onto $\mathcal{B} = \{\boldsymbol{\theta}_W \in \mathbb{R}^{1+q+r+p+pq} : \boldsymbol{\theta}_{W,\mathcal{A}^c} = \mathbf{0}_{(q+1)p-s}\}$, *i.e.* , $\boldsymbol{\theta}_W^{\mathcal{B}}$ replaces the entries of $\boldsymbol{\theta}_W$ in $\mathcal{A}^c$ with zeros. Since $\|\boldsymbol{\theta}_W^{\mathcal{B}} - \overline{\boldsymbol{\theta}}_W\|_2 = \|\boldsymbol{\theta}_{W,\mathcal{A}} - \overline{\boldsymbol{\theta}}_{W,\mathcal{A}}\|_2 \leq \|\boldsymbol{\theta}_W - \overline{\boldsymbol{\theta}}_W\|_2 < \delta$, we have $Q_{\mathrm{H}}(\boldsymbol{\theta}_W^{\mathcal{B}}) - Q_{\mathrm{H}}(\overline{\boldsymbol{\theta}}_W) = \overline{Q}_{\mathrm{H}}(\boldsymbol{\theta}_{W,\mathcal{A}}) - \overline{Q}_{\mathrm{H}}(\overline{\boldsymbol{\theta}}_{W,\mathcal{A}}) > 0$. So the remaining part is to show that $Q_{\mathrm{H}}(\boldsymbol{\theta}_W) - Q_{\mathrm{H}}(\boldsymbol{\theta}_W^{\mathcal{B}}) > 0$.

Without loss of generality, we assume all the entries in $\boldsymbol{\theta}_{W,\mathcal{A}^c}$ are nonzero for $\boldsymbol{\theta}_W$. For $t \in [0,1]$, define $\tilde{Q}_{\mathrm{H}}(t) = \tilde{L}_{\mathrm{H}}(t) + \tilde{P}_{\mathrm{H}}(t)$, where

$$
\begin{aligned}
\tilde{L}_{\mathrm{H}}(t) &= \frac{1}{2n}\|\mathbf{Y} - \mathbf{W}_{\mathcal{A}}\boldsymbol{\theta}_{W,\mathcal{A}} - \mathbf{W}_{\mathcal{A}^c}(t\boldsymbol{\theta}_{W,\mathcal{A}^c})\|_2^2, \\
\tilde{P}_{\mathrm{H}}(t) &= \sum_{h\in\mathcal{A}_G} p_\lambda\left(\|\boldsymbol{\eta}_h(t)\|_2\right) + \sum_{h\notin\mathcal{A}_G} p_\lambda\left(\|t\boldsymbol{\eta}_h\|_2\right) \\
&+ \sum_{j=1}^{q}\left[\sum_{h\in\mathcal{A}_{D_j}} p_\lambda\left(|\theta_{D_j,h}|\right) + \sum_{h\notin\mathcal{A}_{D_j}} p_\lambda\left(|t\theta_{D_j,h}|\right)\right],
\end{aligned}
$$

$\boldsymbol{\eta}_h(t) \in \mathbb{R}^{1+q}$, $h \in \mathcal{A}_G$ is a vector with entries $\theta_{G,h}$, $\theta_{D_j,h}$ if $\mathcal{A}_{D_j}$ contains h, and $t\theta_{D_j,h}$ if $\mathcal{A}_{D_j}$ does not contain h, for $j = 1, \ldots, q$. Hence $\tilde{Q}_{\mathrm{H}}(t = 1) = Q_{\mathrm{H}}(\boldsymbol{\theta}_W)$ and $\tilde{Q}_H(t = 0) = Q_{\mathrm{H}}(\boldsymbol{\theta}_W^{\mathcal{B}})$. By Lagrange's mean value theorem, $Q_{\mathrm{H}}(\boldsymbol{\theta}_W) - Q_{\mathrm{H}}(\boldsymbol{\theta}_W^{\mathcal{B}}) = \tilde{Q}_{\mathrm{H}}(t = 1) - \tilde{Q}_{\mathrm{H}}(t = 0) = \partial\tilde{Q}_{\mathrm{H}}/\partial t|_{t=t^*}$ with $t^* \in (0,1)$. Note that $\partial\tilde{Q}_{\mathrm{H}}/\partial t = \partial\tilde{L}_{\mathrm{H}}/\partial t + \partial\tilde{P}_{\mathrm{H}}/\partial t$, where

$$
\begin{aligned}
\frac{\partial\tilde{L}_{\mathrm{H}}}{\partial t} &= \left(-\frac{1}{n}\right)\boldsymbol{\theta}_{W,\mathcal{A}^c}^{\mathrm{T}}\mathbf{W}_{\mathcal{A}^c}^{\mathrm{T}}\left[\mathbf{Y} - \mathbf{W}_{\mathcal{A}}\boldsymbol{\theta}_{W,\mathcal{A}} - \mathbf{W}_{\mathcal{A}^c}\left(t\boldsymbol{\theta}_{W,\mathcal{A}^c}\right)\right], \\
\frac{\partial\tilde{P}_{\mathrm{H}}}{\partial t} &= \sum_{h\in\mathcal{A}_G} p'_\lambda\left(\|\boldsymbol{\eta}_h(t)\|_2\right)\frac{t\sum_{j:h\notin\mathcal{A}_{D_j}}\theta_{D_j,h}^2}{\|\boldsymbol{\eta}_h(t)\|_2} + \sum_{h\notin\mathcal{A}_G} p'_\lambda(t\|\boldsymbol{\eta}_h\|_2)\|\boldsymbol{\eta}_h\|_2 \\
&+ \sum_{j=1}^{q}\sum_{h\notin\mathcal{A}_{D_j}} p'_\lambda(t|\theta_{D_j,h}|)|\theta_{D_j,h}|.
\end{aligned}
$$

Note that

$$
-\left.\frac{\partial\tilde{L}_{\mathrm{H}}}{\partial t}\right|_{t=t^*} \le \|n^{-1}\mathbf{W}_{\mathcal{A}^c}[\mathbf{Y} - \mathbf{W}_{\mathcal{A}}\boldsymbol{\theta}_{W,\mathcal{A}} - \mathbf{W}_{\mathcal{A}^c}(t^*\boldsymbol{\theta}_{W,\mathcal{A}^c})]\|_\infty\|\boldsymbol{\theta}_{W,\mathcal{A}^c}\|_1 < p'_\lambda(\delta)\|\boldsymbol{\theta}_{W,\mathcal{A}^c}\|_1,
$$

where the last step is from the fact that the vector consisting of $\boldsymbol{\theta}_{W,\mathcal{A}}$ and $t^*\boldsymbol{\theta}_{W,\mathcal{A}^c}$, denoted as $\boldsymbol{\theta}_W^* \in \mathbb{R}^{1+q+r+p+pq}$, satisfies $\|\boldsymbol{\theta}_W^* - \overline{\boldsymbol{\theta}}_W\|_2 \le \|\boldsymbol{\theta}_W - \overline{\boldsymbol{\theta}}_W\|_2 < \delta \le \delta_2$ so that $\|n^{-1}\mathbf{W}_{\mathcal{A}^c}[\mathbf{Y} - \mathbf{W}_{\mathcal{A}}\boldsymbol{\theta}_{W,\mathcal{A}} - \mathbf{W}_{\mathcal{A}^c}(t^*\boldsymbol{\theta}_{W,\mathcal{A}^c})]\|_\infty < p'_\lambda(\delta_2) \le p'_\lambda(\delta)$. Also observe that for $h \notin \mathcal{A}_G$, $t^*\|\boldsymbol{\eta}_h\|_2 \le \|\boldsymbol{\theta}_W^* - \overline{\boldsymbol{\theta}}_W\|_2 < \delta$, and for $h \notin \mathcal{A}_{D_j}, j = 1, \ldots, q$, $t^*|\theta_{D_j,h}| \le \|\boldsymbol{\theta}_W^* - \overline{\boldsymbol{\theta}}_W\|_2 < \delta$. Hence after removing the first term in $\partial\tilde{P}_{\mathrm{H}}/\partial t|_{t=t^*}$ and using the fact that $p'_\lambda()$ is decreasing on $[0,\infty)$, we have $\partial \tilde{P}_{\mathrm{H}}/\partial t|_{t=t^*} \geq p'_\lambda(\delta)\|\boldsymbol{\theta}_{W,\mathcal{A}^c}\|_1$. Thus we conclude that $-\partial \tilde{L}_{\mathrm{H}}/\partial t|_{t=t^*} < p'_\lambda(\delta)\|\boldsymbol{\theta}_{W,\mathcal{A}^c}\|_1 \leq \partial \tilde{P}_{\mathrm{H}}/\partial t|_{t=t^*}$, *i.e.* , $Q_{\mathrm{H}}(\boldsymbol{\theta}_W) - Q_{\mathrm{H}}(\boldsymbol{\theta}_W^{\mathcal{B}}) = \partial \tilde{Q}_{\mathrm{H}}/\partial t|_{t=t^*} = \partial \tilde{L}_{\mathrm{H}}/\partial t|_{t=t^*} + \partial \tilde{P}_{\mathrm{H}}/\partial t|_{t=t^*} > 0$.

□

A.3.2 Proof of Lemma 4.6.1

The proof uses techniques similar to those in (Fan et al., 2014). Under $\mathcal{E}_{\text{initial}}$, for any $h \in \mathcal{A}_G$, $\|\boldsymbol{\eta}_h^{(0)}\|_2 \geq |\theta_{G,h}^{(0)}| \geq |\theta^\star_{G,h}| - |\theta^\star_{G,h} - \theta_{G,h}^{(0)}| > (1+\gamma)\lambda_n - \|\boldsymbol{\theta}_W^{(0)} - \boldsymbol{\theta}_W^\star\|_\infty \geq (1+\gamma)\lambda_n - \lambda_n = \gamma\lambda_n$ so that $p'_{\lambda_n}(\|\boldsymbol{\eta}_h^{(0)}\|_2) = 0$. Similarly, for any $j = 1, \ldots, q$ and any $h \in \mathcal{A}_{D_j}$, we have $p'_{\lambda_n}(|\theta_{D_j,h}^{(0)}|) = 0$. For any $h \in \{1, \ldots, p\} \setminus \mathcal{A}_G$, $\|\boldsymbol{\eta}_h^{(0)}\|_2 \leq \lambda_n$ so that $p'_{\lambda_n}(\|\boldsymbol{\eta}_h^{(0)}\|_2) = \lambda_n$. Similarly, for any $j = 1, \ldots, q$ and any $h \in \{1, \ldots, p\} \setminus \mathcal{A}_{D_j}$, $|\theta_{D_j,h}^{(0)}| = |\theta_{D_j,h}^{(0)} - \theta^\star_{D_j,h}| \leq \|\boldsymbol{\theta}_W^{(0)} - \boldsymbol{\theta}_W^\star\|_\infty \leq \lambda_n$ so that $p'_{\lambda_n}(|\theta_{D_j,h}^{(0)}|) = \lambda_n$. Thus under $\mathcal{E}_{\text{initial}}$, (4.6.2) becomes (4.6.3).

Denote $\ell(\boldsymbol{\theta}_W) = (2n)^{-1}\|\mathbf{Y} - \mathbf{W}\boldsymbol{\theta}_W\|_2^2$ so that $\partial\ell/\partial\boldsymbol{\theta}_W|_{\widehat{\boldsymbol{\theta}}_W^{\mathrm{LLA}}} = -n^{-1}\mathbf{W}^{\mathrm{T}}(\mathbf{Y} - \mathbf{W}\widehat{\boldsymbol{\theta}}_W^{\mathrm{LLA}}) = -n^{-1}\mathbf{W}^{\mathrm{T}}(\mathbf{Y} - \mathbf{W}_{\mathcal{A}}\widehat{\boldsymbol{\theta}}_{W,\mathcal{A}}^{\mathrm{LLA}}) = -n^{-1}\mathbf{W}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P}_{\mathbf{W}_{\mathcal{A}}})\boldsymbol{\varepsilon}_Y$, with $\mathbf{P}_{\mathbf{W}_{\mathcal{A}}} = \mathbf{W}_{\mathcal{A}}(\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\mathbf{W}_{\mathcal{A}})^{-1}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}$. Since $\boldsymbol{\theta}_W \in \mathbb{R}^{1+q+r+p+pq}$ can be split into sub-vectors $\boldsymbol{\theta}_{W,\mathcal{A}} \in \mathbb{R}^{1+q+r+s}$ and $\boldsymbol{\theta}_{W,\mathcal{A}^c} \in \mathbb{R}^{(q+1)p-s}$, we have $\partial\ell/\partial\boldsymbol{\theta}_W|_{\widehat{\boldsymbol{\theta}}_W^{\mathrm{LLA}}}$ split into $\partial\ell/\partial\boldsymbol{\theta}_{W,\mathcal{A}}|_{\widehat{\boldsymbol{\theta}}_W^{\mathrm{LLA}}} = -n^{-1}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P}_{\mathbf{W}_{\mathcal{A}}})\boldsymbol{\varepsilon}_Y = \mathbf{0}_{1+q+r+s}$ and $\partial\ell/\partial\boldsymbol{\theta}_{W,\mathcal{A}^c}|_{\widehat{\boldsymbol{\theta}}_W^{\mathrm{LLA}}} = -n^{-1}\mathbf{W}_{\mathcal{A}^c}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P}_{\mathbf{W}_{\mathcal{A}}})\boldsymbol{\varepsilon}_Y$. By the convexity of $\ell(\boldsymbol{\theta}_W)$ we have

$$\begin{aligned}
\ell(\boldsymbol{\theta}_W) - \ell(\widehat{\boldsymbol{\theta}}_W^{\mathrm{LLA}}) &\geq \left.\frac{\partial\ell}{\partial\boldsymbol{\theta}_W}\right|_{\widehat{\boldsymbol{\theta}}_W^{\mathrm{LLA}}}^{\mathrm{T}} \left(\boldsymbol{\theta}_W - \widehat{\boldsymbol{\theta}}_W^{\mathrm{LLA}}\right) \\
&= \sum_{h\notin\mathcal{A}_G} \left.\frac{\partial\ell}{\partial\theta_{G,h}}\right|_{\widehat{\boldsymbol{\theta}}_W^{\mathrm{LLA}}} \theta_{G,h} + \sum_{j=1}^{q}\sum_{h\notin\mathcal{A}_{D_j}} \left.\frac{\partial\ell}{\partial\theta_{D_j,h}}\right|_{\widehat{\boldsymbol{\theta}}_W^{\mathrm{LLA}}} \theta_{D_j,h}, \qquad \text{(A.3.6)}
\end{aligned}$$

where the last equality follows from $\partial\ell/\partial\boldsymbol{\theta}_{W,\mathcal{A}}|_{\widehat{\boldsymbol{\theta}}_W^{\mathrm{LLA}}} = \mathbf{0}_{1+q+r+s}$, $\widehat{\theta}_{G,h}^{\mathrm{LLA}} = 0$ for $h \notin \mathcal{A}_G$, and $\widehat{\theta}_{D_j,h}^{\mathrm{LLA}} = 0$ for $h \notin \mathcal{A}_{D_j}$, $j = 1, \ldots, q$.

Now we can show that under event $\mathcal{E}_{\text{initial}} \cap \mathcal{E}_{\text{rejularity}}$, $\widehat{\boldsymbol{\theta}}_W^{\mathrm{LLA}}$ is the unique global solution to (4.6.2). To see this, for any $\boldsymbol{\theta}_W \in \mathbb{R}^{1+q+r+s+p+pq}$, we have

$$\ell(\boldsymbol{\theta}_W) + \sum_{h=1}^{p} p'_{\lambda_n}(\|\boldsymbol{\eta}_h^{(0)}\|_2)\|\boldsymbol{\eta}_h\|_2 + \sum_{j=1}^{q}\sum_{h=1}^{p} p'_{\lambda_n}(|\theta_{D_j,h}^{(0)}|)|\theta_{D_j,h}| - \ell(\widehat{\boldsymbol{\theta}}_W^{\mathrm{LLA}})$$

$$
-\sum_{h=1}^{p} p'_{\lambda_n}(\|\boldsymbol{\eta}_h^{(0)}\|_2)\|\widehat{\boldsymbol{\eta}}_h^{\mathrm{LLA}}\|_2 - \sum_{j=1}^{q}\sum_{h=1}^{p} p'_{\lambda_n}(|\theta_{D_j,h}^{(0)}|)|\widehat{\theta}_{D_j,h}^{\mathrm{LLA}}| \overset{(4.6.3)}{=} \ell(\boldsymbol{\theta}_W) - \ell(\widehat{\boldsymbol{\theta}}_W^{\mathrm{LLA}})
$$

$$
+\sum_{h\notin\mathcal{A}_G}\lambda_n\|\boldsymbol{\eta}_h\|_2 + \sum_{j=1}^{q}\sum_{h\notin\mathcal{A}_{D_j}}\lambda_n|\theta_{D_j,h}| \overset{(\mathrm{A}.3.6)}{\geq} \sum_{h\notin\mathcal{A}_G}\left(\frac{\partial\ell}{\partial\theta_{G,h}}\bigg|_{\widehat{\boldsymbol{\theta}}_W^{\mathrm{LLA}}}\theta_{G,h} + \lambda_n\|\boldsymbol{\eta}_h\|_2\right)
$$

$$
+\sum_{j=1}^{q}\sum_{h\notin\mathcal{A}_{D_j}}\left(\frac{\partial\ell}{\partial\theta_{D_j,h}}\bigg|_{\widehat{\boldsymbol{\theta}}_W^{\mathrm{LLA}}}\theta_{D_j,h} + \lambda_n|\theta_{D_j,h}|\right) \geq \sum_{h\notin\mathcal{A}_G}\left[\frac{\partial\ell}{\partial\theta_{G,h}}\bigg|_{\widehat{\boldsymbol{\theta}}_W^{\mathrm{LLA}}}\mathrm{sgn}(\theta_{G,h}) + \lambda_n\right]|\theta_{G,h}|
$$

$$
+\sum_{j=1}^{q}\sum_{h\notin\mathcal{A}_{D_j}}\left[\frac{\partial\ell}{\partial\theta_{D_j,h}}\bigg|_{\widehat{\boldsymbol{\theta}}_W^{\mathrm{LLA}}}\mathrm{sgn}(\theta_{D_j,h}) + \lambda_n\right]|\theta_{D_j,h}| \overset{\mathcal{E}_{\mathrm{rejularity}}}{\geq} 0.
$$

The strict > 0 holds if $\theta_{G,h} \neq 0$ for $h \notin \mathcal{A}_G$ or $\theta_{D_j,h} \neq 0$ for $h \notin \mathcal{A}_{D_j}$, $j = 1, \ldots, q$. Additionally, $\widehat{\boldsymbol{\theta}}_{W,\mathcal{A}}^{\mathrm{LLA}}$ is the unique minimizer of $(2n)^{-1}\|\mathbf{Y} - \mathbf{W}_{\mathcal{A}}\boldsymbol{\theta}_{W,\mathcal{A}}\|_2^2$. Hence $\widehat{\boldsymbol{\theta}}_W^{\mathrm{LLA}}$ is the unique global minimizer of (4.6.2). ☐

A.3.3 Proof of Theorem 4.6.2

To prove Theorem 4.6.2, we only need to show that $\mathbb{P}\{\mathcal{E}_{\mathrm{initial}}^c\} \to 0$, $\mathbb{P}\{\mathcal{E}_{\mathrm{regularity}}^c\} \to 0$, and $(\widehat{\boldsymbol{\theta}}_{W,\mathcal{A}}^{\mathrm{LLA}} - \boldsymbol{\theta}_{W,\mathcal{A}}^{\star}) = O_p(\sqrt{s/n})$.

Given the initial value $\boldsymbol{\theta}_W^{(0)} = \widehat{\boldsymbol{\theta}}_W^{\star}$ from (4.3.1), we have $\mathbb{P}\{\|\boldsymbol{\theta}_W^{(0)} - \boldsymbol{\theta}_W^{\star}\|_\infty > \lambda_n\} \leq \mathbb{P}\{\|\boldsymbol{\theta}_{W,\mathcal{A}}^{(0)} - \boldsymbol{\theta}_{W,\mathcal{A}}^{\star}\|_\infty > \lambda_n\} + \mathbb{P}\{\|\boldsymbol{\theta}_{W,\mathcal{A}^c}^{(0)}\|_\infty > \lambda_n\} \leq \mathbb{P}\{\|\boldsymbol{\theta}_{W,\mathcal{A}}^{(0)} - \boldsymbol{\theta}_{W,\mathcal{A}}^{\star}\|_2 > \sqrt{s/n}(\sqrt{s/n}^{-1}\lambda_n)\} + \mathbb{P}\{\|\mathbf{0}_{(1+q)p-s}\|_\infty > \lambda_n\} \to 0$ by the oracle property of $\boldsymbol{\theta}_W^{(0)}$ and $\lambda_n \gg \sqrt{s/n}$. In addition, $\mathbb{P}\{\|\boldsymbol{\eta}_h^{(0)}\|_2 > \lambda_n,\ \exists h \in \{1, \ldots, p\} \setminus \mathcal{A}_G\} = \mathbb{P}\{\|\mathbf{0}_{1+q}\|_2 > \lambda_n,\ \exists h \in \{1, \ldots, p\} \setminus \mathcal{A}_G\} = 0$ by the heredity structure of $\boldsymbol{\theta}_W^{\star}$ so that $\boldsymbol{\eta}_h^{\star} = \mathbf{0}_{1+q}$ for $h \in \{1, \ldots, p\} \setminus \mathcal{A}_G$ and the oracle property of $\boldsymbol{\eta}_h^{(0)}$. Hence $\mathbb{P}\{\mathcal{E}_{\mathrm{initial}}^c\} \to 0$.

Next we show that $(\widehat{\boldsymbol{\theta}}_{W,\mathcal{A}}^{\mathrm{LLA}} - \boldsymbol{\theta}_{W,\mathcal{A}}) = (n^{-1}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\mathbf{W}_{\mathcal{A}})^{-1}(n^{-1}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\varepsilon}_Y) = O_p(\sqrt{s/n})$. For any $\tau > 0$, we have

$$
\begin{aligned}
& \mathbb{P}\{\|(n^{-1}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\mathbf{W}_{\mathcal{A}})^{-1}(n^{-1}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\varepsilon}_Y)\|_2 \geq \tau\sqrt{s/n}\} \\
\leq\ & \mathbb{P}\{\lambda_{\max}(n^{-1}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\mathbf{W}_{\mathcal{A}})^{-1}\|(n^{-1}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\varepsilon}_Y)\|_2 \geq \tau\sqrt{s/n}\} \\
\leq\ & \mathbb{P}\{\mathcal{E}_\Sigma^c\} + \mathbb{P}\{\mathcal{E}_\Sigma,\ \lambda_{\min}^{-1}(n^{-1}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\mathbf{W}_{\mathcal{A}})^{-1}\|(n^{-1}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\varepsilon}_Y)\|_2 \geq \tau\sqrt{s/n}\} \qquad (\mathrm{A}.3.7) \\
\leq\ & O(s^2/n) + \mathbb{P}\{\|(n^{-1}\mathbf{W}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\varepsilon}_Y)\|_2 \geq 2^{-1}C_\Sigma\tau\sqrt{s/n}\} \qquad (\mathrm{A}.3.8)
\end{aligned}
$$

$$\leq \quad O(s^2/n) + 4C/(C_\Sigma^2\tau^2), \tag{A.3.9}$$

where (A.3.7) follows from (A.1.11), (A.3.8) follows from (A.1.12) and the fact that under $\mathcal{E}_\Sigma$ and Condition A.1.3, we have $\lambda_{\min}(n^{-1}\mathbf{W}_\mathcal{A}^{\mathrm{T}}\mathbf{W}_\mathcal{A}) > \lambda_{\min}(\boldsymbol{\Sigma}) - C_\Sigma/2 \geq C_\Sigma/2$, and (A.3.9) follows from (A.1.15). The we have $(\widehat{\boldsymbol{\theta}}_{W,\mathcal{A}}^{\mathrm{LLA}} - \boldsymbol{\theta}_{W,\mathcal{A}}^\star) = O_p(\sqrt{s/n})$.

For $\mathbb{P}\{\mathcal{E}_{\text{regularity}}^c\}$, note that $\mathbf{W}_{\mathcal{A}^c}^{\mathrm{T}}\left[\mathbf{I}_n - \mathbf{W}_\mathcal{A}\left(\mathbf{W}_\mathcal{A}^{\mathrm{T}}\mathbf{W}_\mathcal{A}\right)^{-1}\mathbf{W}_\mathcal{A}^{\mathrm{T}}\right]\boldsymbol{\varepsilon}_Y = \mathbf{W}_{\mathcal{A}^c}^{\mathrm{T}}\boldsymbol{\varepsilon}_Y - \mathbf{W}_{\mathcal{A}^c}^{\mathrm{T}}\mathbf{W}_\mathcal{A}(\widehat{\boldsymbol{\theta}}_{W,\mathcal{A}}^{\mathrm{LLA}} - \boldsymbol{\theta}_{W,\mathcal{A}}^\star)$. Following the same arguments for (A.1.22) and replacing $\overline{\boldsymbol{\theta}}_{W,\mathcal{A}}$ with $\widehat{\boldsymbol{\theta}}_{W,\mathcal{A}}^{\mathrm{LLA}}$, we can show that $\mathbb{P}\{\mathcal{E}_{\text{regularity}}^c\} \to 0$. □

A.3.4 Algorithm to Solve (4.6.2)

Recall that $\boldsymbol{\eta}_h = [\theta_{G,h}, \theta_{D_1,h}, \ldots, \theta_{D_q,h}]^{\mathrm{T}} \in \mathbb{R}^{1+q}$ for $h = 1, \ldots, p$, and denote $\boldsymbol{\eta}_0 = [\theta_0, \boldsymbol{\theta}_T^{\mathrm{T}}, \boldsymbol{\theta}_X^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{1+q+r}$. The block coordinate decent method cyclically optimizes (4.6.2) with respect to each $\boldsymbol{\eta}_h$ for $h = 0, 1, \ldots, p$, given the current $\boldsymbol{\eta}_l$'s, $l \neq h$. We repeat such a procedure iteratively until convergence to find the minimizer of (4.6.2).

Within each iteration of BCD, when we minimize (4.6.2) with respect to $\boldsymbol{\eta}_0$ given the current $\boldsymbol{\eta}_h$'s, $h = 1, \ldots, p$, it is a least squares problem because $\boldsymbol{\eta}_0$ is not penalized. Hence the minimizer $\widehat{\boldsymbol{\eta}}_0$ is $(\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{r}_0$, where $\mathbf{H}$ is defined in (4.2.2) and $\mathbf{r}_0$ is the partial residual of $\mathbf{Y}$ removing all the terms not related to $\boldsymbol{\eta}_0$. Specifically, in the outcome model (4.2.1), define $\boldsymbol{G}_h \in \mathbb{R}^{n\times(1+q)}$ which consists of columns corresponding to $\theta_{G,h}$, $\theta_{D_1,h}$, $\ldots$, $\theta_{D_q,h}$ for $h = 1, \ldots, p$, then $\mathbb{E}\mathbf{Y} = \mathbf{H}\boldsymbol{\eta}_0^\star + \sum_{h=1}^p \boldsymbol{G}_h\boldsymbol{\eta}_h^\star$ where $\boldsymbol{\eta}_h^\star$ is the true value of $\boldsymbol{\eta}_h$ for $h = 0, 1, \ldots, p$. The partial residual $\mathbf{r}_0 = \mathbf{Y} - \sum_{h=1}^p \mathbf{G}_h\boldsymbol{\eta}_h$.

When we minimize (4.6.2) with respect to each $\boldsymbol{\eta}_h$, $h \in \{1, \ldots, p\}$ given the current $\boldsymbol{\eta}_l$'s, $l \in \{0, 1, \ldots, p\} \setminus \{h\}$, the problem becomes

$$\widehat{\boldsymbol{\eta}}_h = \underset{\boldsymbol{\eta}_h \in \mathbb{R}^{1+q}}{\arg\min} \frac{1}{2n}\|\mathbf{r}_h - \boldsymbol{G}_h\boldsymbol{\eta}_h\|_2^2 + w_{\mathrm{grp},h}\|\boldsymbol{\eta}_h\|_2 + \sum_{j=1}^q w_{j,h}|\theta_{D_j,h}|, \tag{A.3.10}$$

where $\mathbf{r}_h = \mathbf{Y} - \mathbf{H}\boldsymbol{\eta}_0 - \sum_{l\neq h,l=1}^{p} \boldsymbol{G}_l \boldsymbol{\eta}_l$ is the partial residual removing all the terms not related to $\boldsymbol{\eta}_h$, $w_{\mathrm{grp},h} = p'_\lambda(\|\boldsymbol{\eta}_h^{(0)}\|_2)$, and $w_{j,h} = p'_\lambda(|\theta_{D_j,h}^{(0)}|)$, $j = 1, \ldots, q$. Since the derivative of $p_{\lambda_n}(\cdot)$ is non-negative and decreasing on $[0, \infty)$, we have that $w_{j,h} \geq w_{\mathrm{grp},h} \geq 0$.

A common approach to optimizing (A.3.10) involves pre-standardizing $\boldsymbol{G}_h$ to $\boldsymbol{G}_{h,\mathrm{T}}$ such that $n^{-1}\boldsymbol{G}_{h,\mathrm{T}}^{\mathrm{T}}\boldsymbol{G}_{h,\mathrm{T}} = \mathbf{I}_{1+q}$ (Yuan & Lin, 2006; Yan & Bien, 2017). If $\boldsymbol{G}_h$ is replaced by $\boldsymbol{G}_{h,\mathrm{T}}$, the solution $\widehat{\boldsymbol{\eta}}_{h,\mathrm{T}}$ to (A.3.10) will have a simple closed form. However, the heredity will be achieved in $\widehat{\boldsymbol{\eta}}_{h,\mathrm{T}}$ instead of the original $\widehat{\boldsymbol{\eta}}_h$. And therefore, to maintain the heredity in $\widehat{\boldsymbol{\eta}}_h$ using $\boldsymbol{G}_{h,\mathrm{T}}$, one needs to change the penalty function in (A.3.10), which complicates the procedure. Thus we choose not to standardize $\boldsymbol{G}_h$ in analyzing (A.3.10).

For ease of notation, we suppress the index h below in $\boldsymbol{\eta}_h$, $w_{\mathrm{grp},h}$, and $w_{j,h}$ for each j in (A.3.10). By setting the first order derivative of (A.3.10) to zero, we have

$$\mathbf{D}\boldsymbol{\eta} - \mathbf{u} + w_{\mathrm{grp}} \frac{\partial \|\boldsymbol{\eta}\|_2}{\partial \boldsymbol{\eta}} + \sum_{j=1}^{q} w_j \frac{\partial |\eta_j|}{\partial \boldsymbol{\eta}} = 0, \tag{A.3.11}$$

where $\boldsymbol{\eta} = [\eta_0, \eta_1, \ldots, \eta_q]^{\mathrm{T}}$, $\mathbf{u} = n^{-1}\boldsymbol{G}_h^{\mathrm{T}}\mathbf{r}_h = [u_0, u_1, \ldots, u_q]^{\mathrm{T}} \in \mathbb{R}^{q+1}$. The arrowhead matrix $\mathbf{D} = n^{-1}\boldsymbol{G}_h^{\mathrm{T}}\boldsymbol{G}_h \in \mathbb{R}^{(q+1)\times(q+1)}$ has the form

$$\mathbf{D} = \begin{bmatrix} d_t & d_1 & \ldots & d_q \\ d_1 & d_1 & & \\ \vdots & & \ddots & \\ d_q & & & d_q \end{bmatrix},$$

where $d_j = n^{-1}\sum_{i\in\mathcal{N}_j} g_{i,h}^2$, $g_{i,h}$ is the (i,h)th entry of $\mathbf{G}$ in (4.2.1), $\mathcal{N}_j$ is the index set of units receiving treatment level $\mathbf{t}_j$ for $j = 0, 1, \ldots, q$, $d_t = \sum_{j=0}^{q} d_j$, and all other empty parts in $\mathbf{D}$ are zeros.

To solve (A.3.11), for a parameter $\mu \geq 0$, we define the following soft-threshold function of a vector $\boldsymbol{\xi}$ as $S(\boldsymbol{\xi}, \mu) = \mathcal{I}\{\|\boldsymbol{\xi}\|_2 > \mu\}(1 - \|\boldsymbol{\xi}\|_2^{-1}\mu)\boldsymbol{\xi}$. Recall that the first column of $\boldsymbol{G}_h$ corresponds to the main effect $\theta_{G,h}^{\star}$ and the other columns of $\boldsymbol{G}_h$ correspond to the interactions $\theta^\star_{D_j,h}$ for $j = 1, \ldots, q$. Since $\mathbf{u} = n^{-1}\boldsymbol{G}_h^{\mathrm{T}}\mathbf{r}_h$, the first entry u_0 contains the effect of $\theta^\star_{G,h}$ while $u_1, \ldots, u_q$ reflect the effects of $\theta^\star_{D_1,h}, \ldots, \theta_{D_q,h}$. If $[u_0^2 + \sum_{j=1}^q S^2(u_j, w_j)]^{1/2} \leq w_{\mathrm{grp}}$, that is, if the adjusted scale of all the effects is too small, we set $\widehat{\boldsymbol{\eta}} = \mathbf{0}_{1+q}$. Otherwise, if the adjusted scale of all the effects is larger than w_{grp}, but such large scale is mainly contributed by $|u_0|$, *i.e.* , $|u_0| > w_{\mathrm{grp}}$ while the scales of $|u_j|$, $j = 1, \ldots, q$ are all small, we set $\widehat{\eta}_0 = S(u_0, w_{\mathrm{grp}})/d_t$ and $\widehat{\eta}_j = 0$ for $j = 1, \ldots, q$. If the adjusted scale of all the effects is larger than w_{grp} and such large scale is contributed by both $|u_0|$ and some of $|u_j|$'s, $j \in \{1, \ldots, q\}$, then by solving (A.3.11), we can obtain $\widehat{\eta}_0 = x$ and $\|\widehat{\boldsymbol{\eta}}\|_2 = y$ from

$$1 = \frac{x^2}{y^2} + \sum_{j=1}^{q} \frac{S^2(u_j - d_j x, w_j)}{(d_j y + w_{\mathrm{grp}})^2}, \; u_0 - d_t x - w_{\mathrm{grp}}\frac{x}{y} = \sum_{j=1}^{q} f_j(y, w_{\mathrm{grp}}) S(u_j - d_j x, w_j), \quad \text{(A.3.12)}$$

where $f_j(y, w_{\mathrm{grp}}) = [(d_j y)/(d_j y + w_{\mathrm{grp}})]$ for $y \in (0, \infty)$ and $w_{\mathrm{grp}} \geq 0$ with $j = 1, \ldots, q$. Then for $j = 1, \ldots, q$, $\widehat{\eta}_j$ is solved by

$$\widehat{\eta}_j = \begin{cases} \frac{u_j - d_j\widehat{\eta}_0}{d_j + \frac{w_{\mathrm{grp}}}{\|\widehat{\boldsymbol{\eta}}\|_2} + \frac{w_j}{|\widehat{\eta}_j|}}, \text{ if } |u_j - d_j\widehat{\eta}_0| > w_j \text{ with } |\widehat{\eta}_j| = \frac{|u_j - d_j\widehat{\eta}_0| - w_j}{d_j + \frac{w_{\mathrm{grp}}}{\|\widehat{\boldsymbol{\eta}}\|_2}}; \\ 0, \text{ if } |u_j - d_j\widehat{\eta}_0| \leq w_j. \end{cases} \quad \text{(A.3.13)}$$

This procedure is summarized in Algorithm 1.

Algorithm 1 Optimization for One Block

Input: The group penalty weight $w_{\text{grp}} \in \mathbb{R}$, individual penalty weights $w_j \in \mathbb{R}$, $j = 1, \ldots, q$, the arrowhead matrix $\mathbf{D} \in \mathbb{R}^{(q+1)\times(q+1)}$, the $(q+1)$-dimensional vector $\mathbf{u}$.

if $[u_0^2 + \sum_{j=1}^{q} S^2(u_j, w_j)]^{1/2} \le w_{\text{grp}}$ **then**
    $\widehat{\boldsymbol{\eta}} = \mathbf{0}_{q+1}$ and stop
else if $|u_0| > w_{\text{grp}}$ and $\forall j = 1, \ldots, q,\ |u_j - d_j S(u_0, w_{\text{grp}})/d_t| \le w_j$ **then**
    $\widehat{\boldsymbol{\eta}} = [S(u_0, w_{\text{grp}})/d_t,\ \mathbf{0}_q^{\mathrm{T}}]^{\mathrm{T}}$ and stop.
else
    Solves (A.3.12) and (A.3.13).
end if

Output: A global minimzer $\widehat{\boldsymbol{\eta}}$ for (A.3.10).

A.4 Extra Simulation Results

In this section, we present extra simulation results mentioned in the main paper. From Section 4.5.1, the bias of Lasso method is shown in Figure A.1. The Lasso method, denoted as M5, changes the penalty in (4.3.1) to the ℓ_1 penalty and other procedures are the same as our method M1.

For both Setting 1 and Setting 2, additional results are presented for cases where the entries in $\mathbf{w}_i$ independently follow $t(4)/\sqrt{2}$ in $\boldsymbol{\varepsilon}_{G_i} = \boldsymbol{\Sigma}_G^{1/2}\mathbf{w}_i$ and $p = 1000$, as shown in Figures A.2-A.7. In these figures, "C" and "P" in the titles denote empirical coverage probability and power, respectively, while "0" and "1" refer to treatments t_0 and t_1. For instance, "NIE C0" represents the empirical coverage probability of the NIE $\delta(t_0)$, and "NDE P1" represents the empirical power of NDE $\zeta(t_1)$. Each empirical coverage probability and power are calculated over 500 replications. We let M0, M1, M2, M3, M4 denote the oracle method, our method, the difference method in (Guo et al., 2023), the debiased product method in Zhou et al. (2020), and the PCMA method in Huang and Pan (2016), respectively.

In Figure 4.1 of the main paper, under Setting 1, both our proposed method and the oracle method exhibit asymmetric power curves for $\zeta(t_0)$. This asymmetry arises from the quadratic form $\boldsymbol{\theta}^{\star\mathrm{T}}_{D_j,\mathcal{A}_{D_j}}\ \boldsymbol{\Sigma}_{G(\mathcal{A}_{D_j}\times\mathcal{A}_{D_j})}\ \boldsymbol{\theta}^{\star}_{D_j,\mathcal{A}_{D_j}}$ in the standard error from Theorem 4.4.2. As the off-diagonal entries of $\boldsymbol{\Sigma}_G$ are 0.5's, this quadratic form increases dramatically as the magnitudes of $\boldsymbol{\theta}_{D_j,\mathcal{A}_{D_j}}$, *i.e.* , c_D, increases. In Figure A.8, we present the results when $\boldsymbol{\Sigma}_G = \mathbf{I}_p$, where the quadratic form in the standard error increases slower and the power curves of M0 and M1 for $\zeta(t_0)$ are more symmetrical.

From Section 4.5.2, Table A.1 presents the coverage probabilities and powers of our method (M1) and the method in Guo et al. (2022) (M1) when $\boldsymbol{\Sigma}_G$ is generated by autoregressive structure whose (i,j)-element is $0.5^{|i-j|}$. If the coverage probability is lower than 0.8, the corresponding power cell is marked in gray.

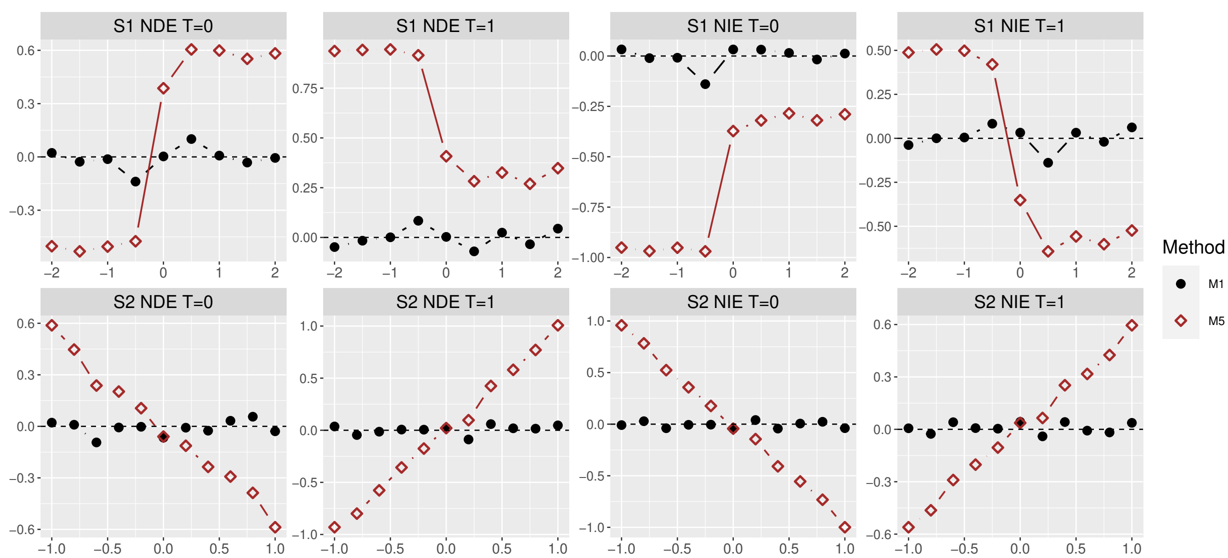


Figure A.1: The bias of Lasso (M5) and our method (M1) when $p = 500$ and $\boldsymbol{\varepsilon}_{G_i} \sim \mathcal{N}(\mathbf{0}_p, \boldsymbol{\Sigma}_G)$, $\epsilon_{Y_i} \sim N(0, \sigma_Y^2)$. The first row are the results from Setting 1. The second row are the results from Setting 2.

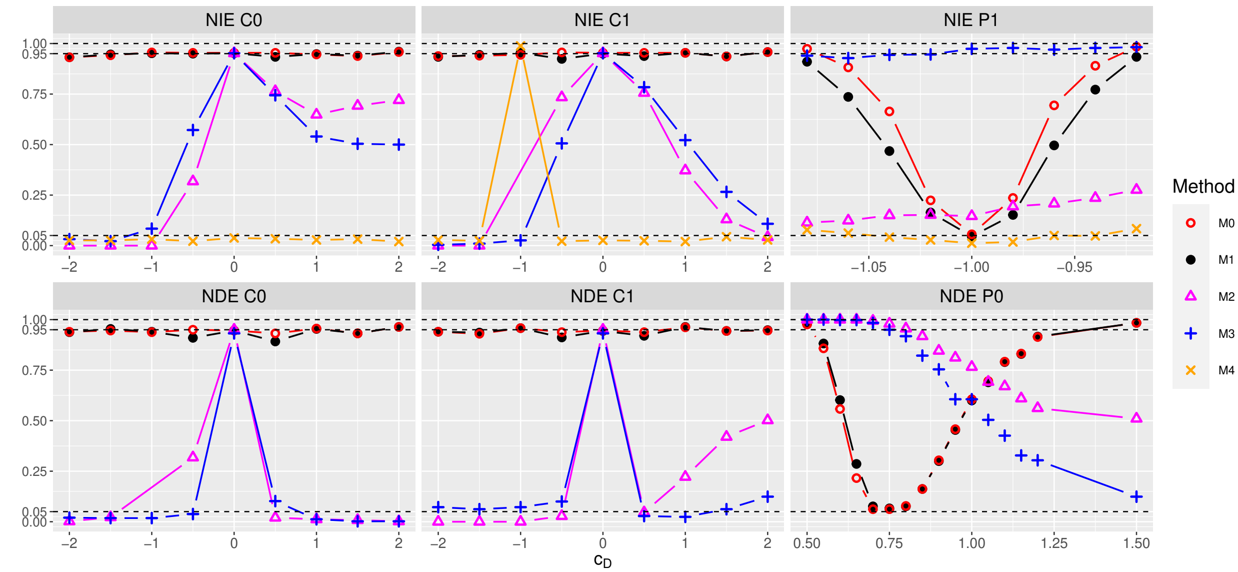


Figure A.2: The comparison of different methods under Setting 1 when $p = 500$ and $\boldsymbol{\varepsilon}_{G_i} = \boldsymbol{\Sigma}_G^{1/2}\mathbf{w}_i$, $\mathbf{w}_i$ has i.i.d. $t(4)/\sqrt{2}$ entries, $\epsilon_{Y_i} \sim N(0, \sigma_Y^2)$. Under Setting 1, $\delta(t_0) = 4.2$, $\delta(t_1) = 4.2(1 + c_D)$, $\zeta(t_0) = 3.15 - 4.2c_D$, and $\zeta(t_1) = 3.15$.

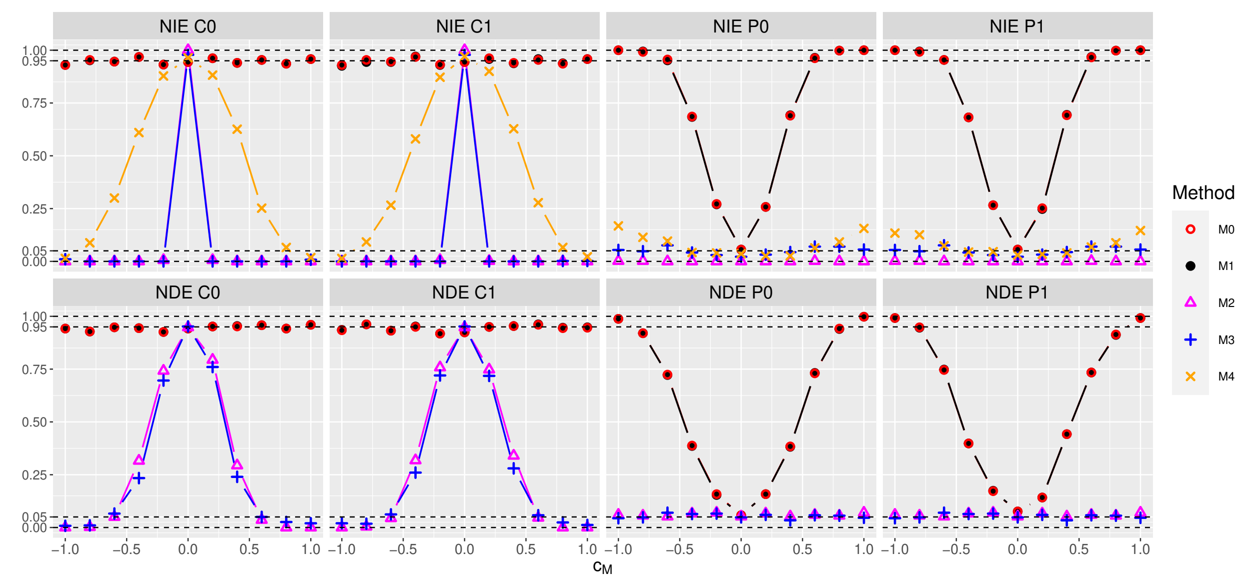


Figure A.3: The comparison of different methods under Setting 2 when $p = 500$ and $\boldsymbol{\varepsilon}_{G_i} = \boldsymbol{\Sigma}_G^{1/2}\mathbf{w}_i$, $\mathbf{w}_i$ has i.i.d. $t(4)/\sqrt{2}$ entries, $\epsilon_{Y_i} \sim N(0, \sigma_Y^2)$. Under Setting 2, $\delta(t_0) = \zeta(t_0) = 4.2c_M$, $\delta(t_1) = \zeta(t_1) = -4.2c_M$.

	$p = 1000$											
Effect	$\boldsymbol{\varepsilon}_{G_i}$: Normal; ϵ_{Y_i}: Normal				$\boldsymbol{\varepsilon}_{G_i}$: t(8); ϵ_{Y_i}: Normal				$\boldsymbol{\varepsilon}_{G_i}$: Poisson; ϵ_{Y_i}: t(8)			
	Coverage		Power		Coverage		Power		Coverage		Power	
	M1	M2	M1	M2	M1	M2	M1	M2	M1	M2	M1	M2
$\delta_{\mathbf{x},1}(\mathbf{t}_0) = -1.55$	0.960	0.854	0.156	0.294	0.944	0.846	0.164	0.332	0.950	0.872	0.168	0.306

$\delta_{\mathbf{x},1}(\mathbf{t}_1) = -4.30$	0.932	0.746	0.464	0.294	0.942	0.750	0.494	0.332	0.954	0.762	0.498	0.306
$\delta_{\mathbf{x},2}(\mathbf{t}_0) = -1.55$	0.946	0.850	0.178	0.320	0.952	0.862	0.164	0.330	0.938	0.858	0.148	0.304
$\delta_{\mathbf{x},2}(\mathbf{t}_2) = -3.80$	0.938	0.790	0.482	0.320	0.946	0.796	0.464	0.330	0.936	0.804	0.444	0.304
$\zeta_{\mathbf{x},1}(\mathbf{t}_0) = 6.56$	0.952	0.658	1.000	0.838	0.936	0.656	1.000	0.848	0.940	0.670	1.000	0.880
$\zeta_{\mathbf{x},1}(\mathbf{t}_1) = 3.81$	0.930	0.696	0.940	0.838	0.948	0.710	0.958	0.848	0.936	0.720	0.962	0.880
$\zeta_{\mathbf{x},2}(\mathbf{t}_0) = 3.58$	0.930	0.000	0.984	1.000	0.942	0.000	0.992	1.000	0.942	0.000	0.994	1.000
$\zeta_{\mathbf{x},2}(\mathbf{t}_2) = 1.33$	0.944	0.000	0.354	1.000	0.932	0.000	0.372	1.000	0.946	0.000	0.322	1.000

$p = 2000$

Effect	$\boldsymbol{\varepsilon}_{G_i}$: Normal; ϵ_{Y_i}: Normal				$\boldsymbol{\varepsilon}_{G_i}$: t(8); ϵ_{Y_i}: Normal				$\boldsymbol{\varepsilon}_{G_i}$: Poisson; ϵ_{Y_i}: t(8)			
	Coverage		Power		Coverage		Power		Coverage		Power	
	M1	M2	M1	M2	M1	M2	M1	M2	M1	M2	M1	M2
$\delta_{\mathbf{x},1}(\mathbf{t}_0) = -1.55$	00.932	0.756	0.156	0.298	0.928	0.756	0.156	0.290	0.922	0.758	0.176	0.304
$\delta_{\mathbf{x},1}(\mathbf{t}_1) = -4.30$	0.940	0.486	0.458	0.298	0.948	0.528	0.450	0.290	0.946	0.498	0.484	0.304
$\delta_{\mathbf{x},2}(\mathbf{t}_0) = -1.55$	0.928	0.758	0.154	0.310	0.938	0.748	0.142	0.298	0.932	0.746	0.194	0.324
$\delta_{\mathbf{x},2}(\mathbf{t}_2) = -3.80$	0.932	0.562	0.460	0.310	0.950	0.566	0.450	0.298	0.942	0.546	0.476	0.324
$\zeta_{\mathbf{x},1}(\mathbf{t}_0) = 6.95$	0.942	0.612	1.000	0.706	0.918	0.602	1.000	0.708	0.944	0.568	1.000	0.696
$\zeta_{\mathbf{x},1}(\mathbf{t}_1) = 4.20$	0.934	0.830	0.970	0.706	0.930	0.794	0.966	0.708	0.940	0.838	0.986	0.696
$\zeta_{\mathbf{x},2}(\mathbf{t}_0) = 4.64$	0.958	0.000	1.000	1.000	0.958	0.000	1.000	1.000	0.946	0.000	1.000	1.000
$\zeta_{\mathbf{x},2}(\mathbf{t}_2) = 2.39$	0.926	0.000	0.768	1.000	0.922	0.000	0.780	1.000	0.932	0.000	0.802	1.000

Table A.1: Empirical coverage probabilities and powers of our method and Guo et al. (2022) with $\boldsymbol{\Sigma}_G$ generated by auto-regressive structure

A.5 The Genes Corresponding to the Detected DNA Methylation Loci

We present the 13 DNA methylation loci in $\widehat{\mathcal{A}}_G^{\mathrm{LLA}}$ from Section 4.7 along with the genes to which they belong in Table A.2 below. The descriptions were summarized from the GeneCards database (Safran et al., 2010).

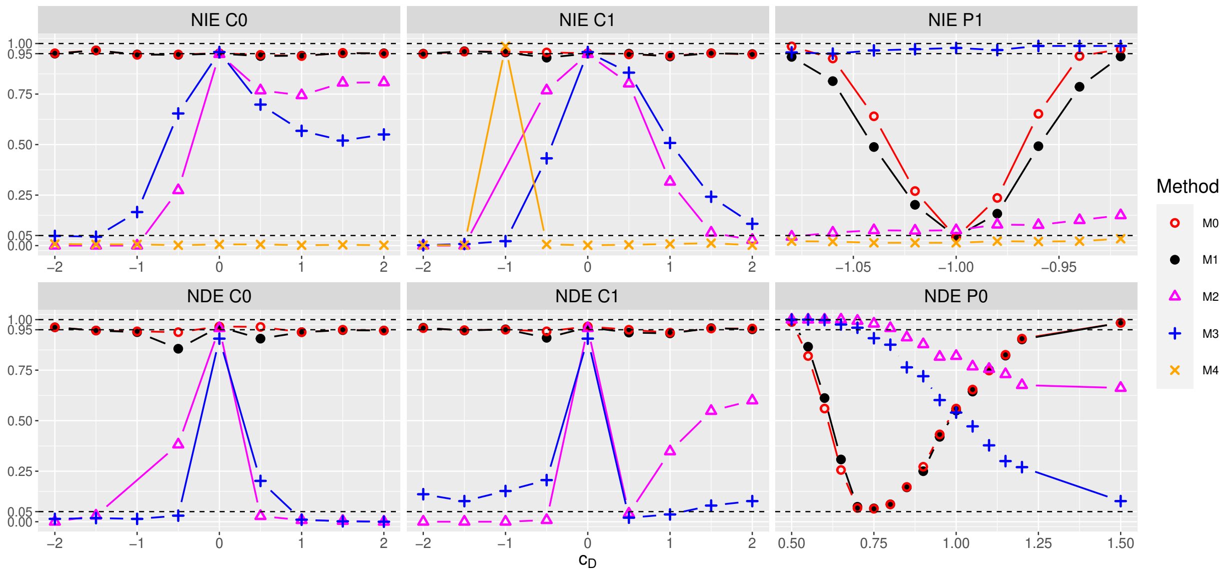


Figure A.4: The comparison of different methods under Setting 1 when $p = 1000$ and $\boldsymbol{\varepsilon}_{G_i} = \mathcal{N}(\mathbf{0}_p, \boldsymbol{\Sigma}_G)$, $\epsilon_{Y_i} \sim N(0, \sigma_Y^2)$. Under Setting 1, $\delta(t_0) = 4.2$, $\delta(t_1) = 4.2(1 + c_D)$, $\zeta(t_0) = 3.15 - 4.2c_D$, and $\zeta(t_1) = 3.15$.

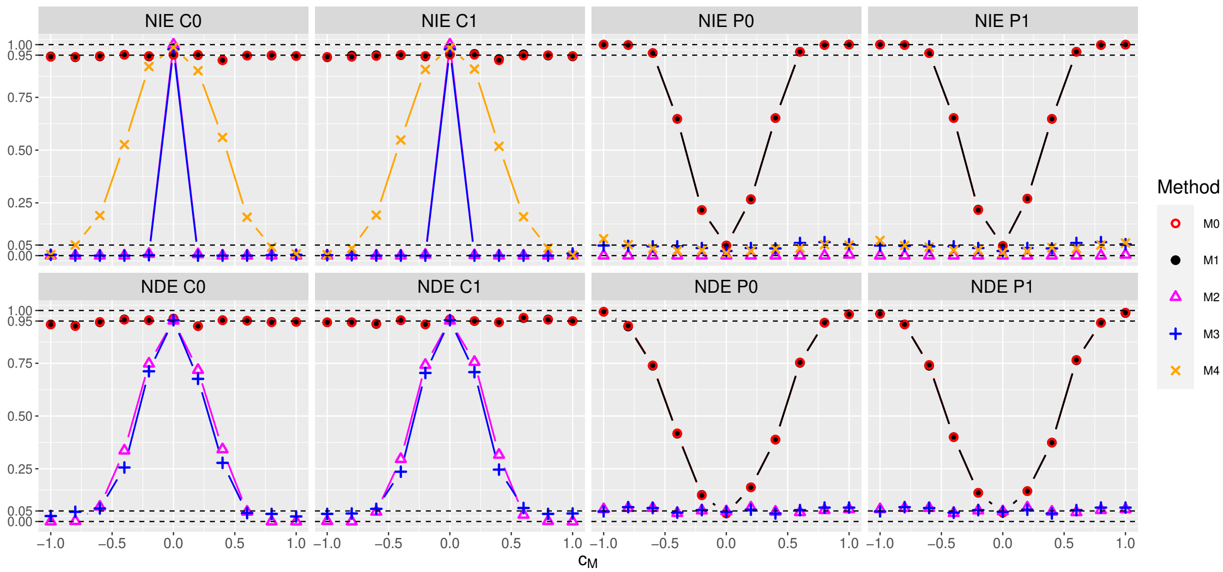


Figure A.5: The comparison of different methods under Setting 2 when $p = 1000$ and $\boldsymbol{\varepsilon}_{G_i} = \mathcal{N}(\mathbf{0}_p, \boldsymbol{\Sigma}_G)$, $\epsilon_{Y_i} \sim N(0, \sigma_Y^2)$. Under Setting 2, $\delta(t_0) = \zeta(t_0) = 4.2c_M$, $\delta(t_1) = \zeta(t_1) = -4.2c_M$.

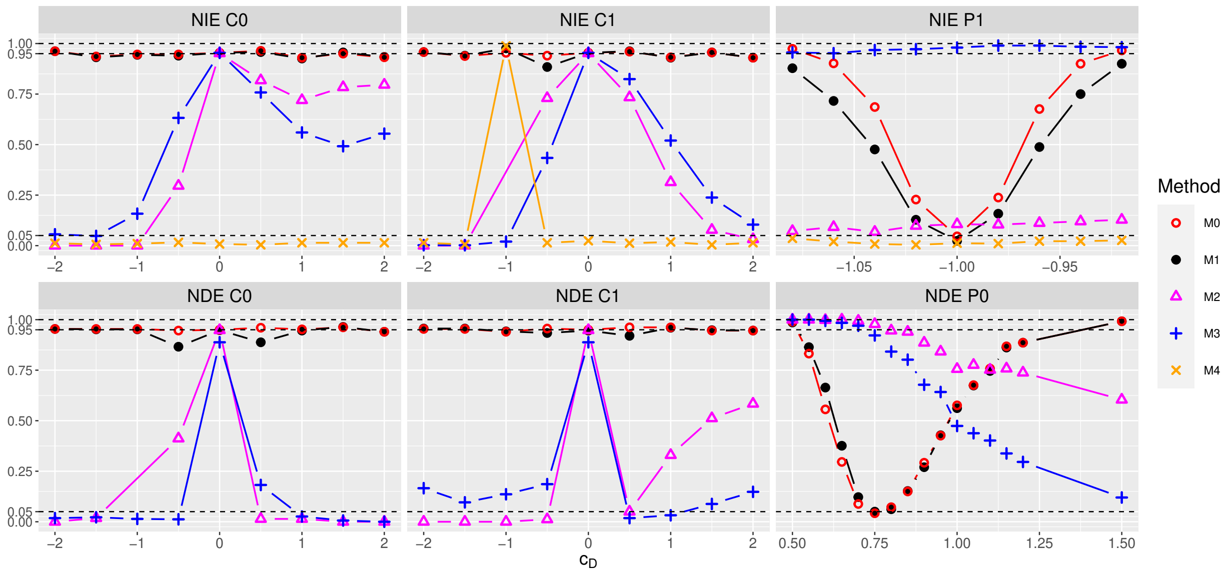


Figure A.6: The comparison of different methods under Setting 1 when $p = 1000$ and $\boldsymbol{\varepsilon}_{G_i} = \boldsymbol{\Sigma}_G^{1/2}\mathbf{w}_i$, $\mathbf{w}_i$ has i.i.d. $t(4)/\sqrt{2}$ entries, $\epsilon_{Y_i} \sim N(0, \sigma_Y^2)$. Under Setting 1, $\delta(t_0) = 4.2$, $\delta(t_1) = 4.2(1 + c_D)$, $\zeta(t_0) = 3.15 - 4.2c_D$, and $\zeta(t_1) = 3.15$.

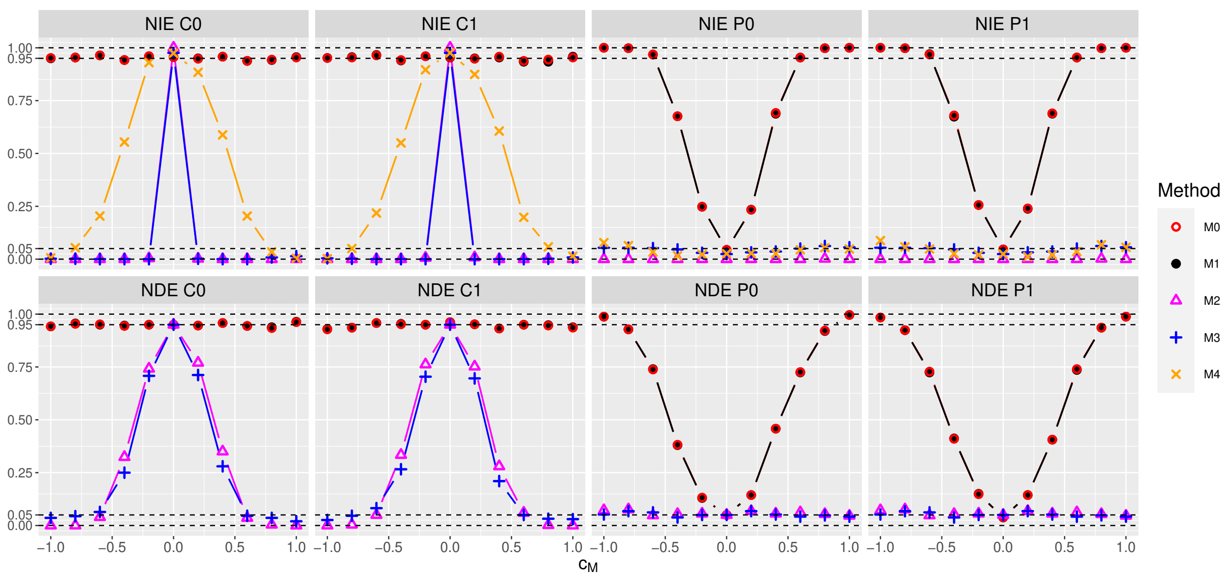


Figure A.7: The comparison of different methods under Setting 2 when $p = 1000$ and $\boldsymbol{\varepsilon}_{G_i} = \boldsymbol{\Sigma}_G^{1/2}\mathbf{w}_i$, $\mathbf{w}_i$ has i.i.d. $t(4)/\sqrt{2}$ entries, $\epsilon_{Y_i} \sim N(0, \sigma_Y^2)$. Under Setting 2, $\delta(t_0) = \zeta(t_0) = 4.2c_M$, $\delta(t_1) = \zeta(t_1) = -4.2c_M$.

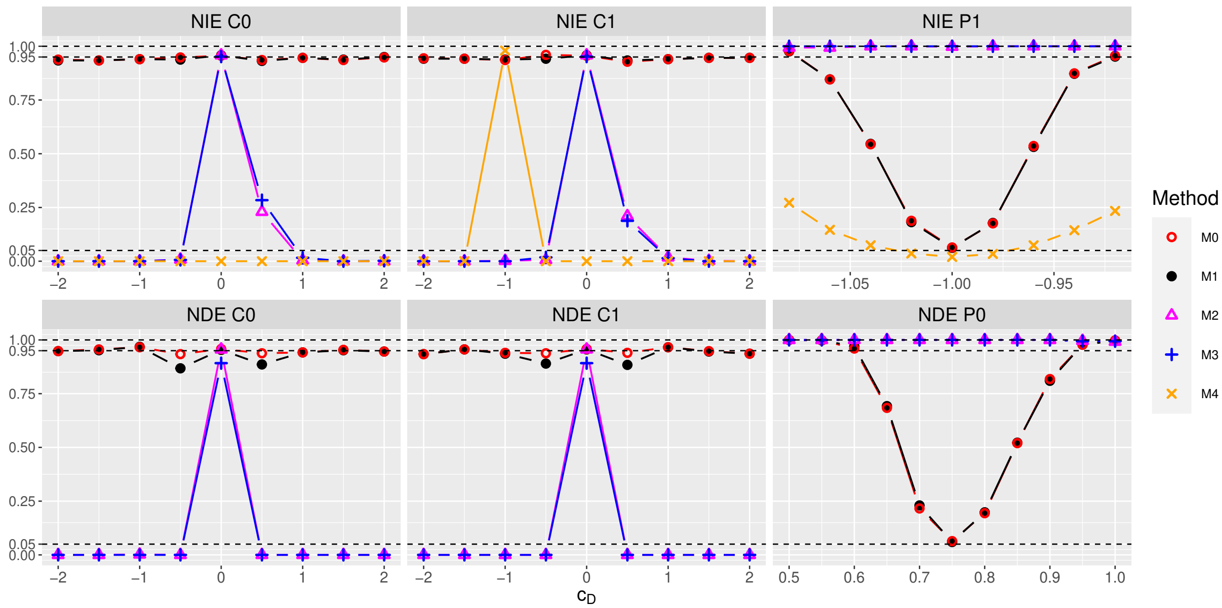


Figure A.8: The comparison of different methods under Setting 1 when $p = 500$, $\boldsymbol{\varepsilon}_{G_i} \sim \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p)$, and $\epsilon_{Y_i} \sim N(0, \sigma_Y^2)$. Under Setting 1, $\delta(t_0) = 4.2$, $\delta(t_1) = 4.2(1 + c_D)$, $\zeta(t_0) = 3.15 - 4.2c_D$, and $\zeta(t_1) = 3.15$.

Table A.2: Annotation of the Included Mediators in $\widehat{\mathcal{A}}_G^{\mathrm{BCD}}$

Locus	Gene	Description
cg06719334 (G_{20})	RBFOX1	Involved in the regulation of alternative splicing of pre-mRNA. (Davies et al., 2015; Kucherenko & Shcherbata, 2018)
cg19230917 (G_{27})	RAB5IF	Involved in the regulation of endocytic trafficking (Ravikumar, Imarisio, Sarkar, O'Kane, & Rubinsztein, 2008)
cg18634806 (G_{55})	OBI1	Involved in the regulation of the cell cycle and stress response (Banks et al., 2016)
cg02860705 (G_{106})	NEK10	Involved in the regulation of mucociliary transport (Chivukula et al., 2020)
cg11032707 (G_{146})	GATM	Important in the biosynthesis of creatine, a molecule essential for energy storage in muscles and the brain. (Gross, Eggen, Simon, & Van Pilsum, 1986)
cg05781698 (G_{149})	IGF1R	Important in mediating actions of insulin-like growth factor 1 (Favelyukis, Till, Hubbard, & Miller, 2001)
cg14063008 (G_{268})	DAB2IP	Involved in various signaling pathways that regulate cell growth and apoptosis (R. Zhang et al., 2003)
cg07439316 (G_{497})	LY6G5C	Involved in signal transduction and immune responses (Mallya, Campbell, & Aguado, 2002)
cg06957003 (G_{589})	TEX15	Important for DNA double-strand break repair during meiosis (F. Yang, Eckardt, Leu, McLaughlin, & Wang, 2008)
cg13662644 (G_{641})	CAMSAP3	Involved in the organization of the cytoskeleton and cell polarity (Meng, Mushika, Ichii, & Takeichi, 2008)
cg16142488 (G_{807})	CTNND2	Involved in cell adhesion and signaling, and mutations in this gene have been associated with cognitive disorders (Hofmeister et al., 2015; Turner et al., 2015)
cg03025830 (G_{879})	FGF17	Involved in normal brain development (X. Zhang et al., 2006)
cg26179948 (G_{1003})	JAZF1	Involved in lipid metabolism and glucose homeostasis (Liao, Wang, Qi, & Xiao, 2019)

Table A.3: Blocks in $\widehat{\boldsymbol{\Sigma}}$ and $\boldsymbol{\Sigma}$

Block	Expression	Block	Expression
$\widehat{\boldsymbol{\Sigma}}_{11} \in \mathbb{R}^{(1+q)\times(1+q)}$	$\frac{1}{n}\begin{bmatrix} n & \mathcal{R}\{n_j\}_{j=1}^q \\ \mathcal{C}\{n_j\}_{j=1}^q & \operatorname{diag}\{n_j\}_{j=1}^q \end{bmatrix}$	$\boldsymbol{\Sigma}_{11} \in \mathbb{R}^{(1+q)\times(1+q)}$	$\begin{bmatrix} 1 & \mathcal{R}\{p_j\}_{j=1}^q \\ \mathcal{C}\{p_j\}_{j=1}^q & \operatorname{diag}\{p_j\}_{j=1}^q \end{bmatrix}$
$\widehat{\boldsymbol{\Sigma}}_{12} \in \mathbb{R}^{(1+q)\times r}$	$\left[\frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{1}_n,\ \mathcal{R}\left\{\frac{n_j}{n}\mathbf{X}^{\mathrm{T}}_{\mathcal{N}_j\times[r]}\mathbf{1}_{n_j}\right\}_{j=1}^q\right]^{\mathrm{T}}$	$\boldsymbol{\Sigma}_{12} \in \mathbb{R}^{(1+q)\times r}$	$\left[\boldsymbol{\mu}_X,\ \mathcal{R}\left\{p_j\boldsymbol{\mu}_X\right\}_{j=1}^q\right]^{\mathrm{T}}$
$\widehat{\boldsymbol{\Sigma}}_{13} \in \mathbb{R}^{(1+q)\times s_G}$	$\left[\frac{1}{n}\mathbf{G}^{\mathrm{T}}_{\mathcal{A}_G}\mathbf{1}_n,\ \mathcal{R}\left\{\frac{1}{n_j}\mathbf{G}^{\mathrm{T}}_{\mathcal{N}_j\times\mathcal{A}_G}\mathbf{1}_{n_j}\right\}\right]^{\mathrm{T}}$	$\boldsymbol{\Sigma}_{13} \in \mathbb{R}^{(1+q)\times s_G}$	$\begin{bmatrix} \sum_{j=1}^q p_j\left(\boldsymbol{\beta}_{0,\mathcal{A}_G}+\boldsymbol{\beta}_{j,\mathcal{A}_G}+\mathbf{A}_{\mathcal{A}_G\times[r]}\boldsymbol{\mu}_X\right)^{\mathrm{T}}+p_0\left(\boldsymbol{\beta}_{0,\mathcal{A}_G}+\mathbf{A}_{\mathcal{A}_G\times[r]}\boldsymbol{\mu}_X\right)^{\mathrm{T}} \\ \mathcal{R}\left\{p_j\left(\boldsymbol{\beta}_{0,\mathcal{A}_G}+\boldsymbol{\beta}_{j,\mathcal{A}_G}+\mathbf{A}_{\mathcal{A}_G\times[r]}\boldsymbol{\mu}_X\right)\right\}_{j=1}^{q,\mathrm{T}} \end{bmatrix}$
$\widehat{\boldsymbol{\Sigma}}_{14} \in \mathbb{R}^{(1+q)\times\sum_{j=1}^q s_{D_j}}$	$\begin{bmatrix} \mathcal{R}\left\{\frac{n_j}{n}\frac{1}{n_j}\mathbf{1}^{\mathrm{T}}_{n_j}\mathbf{G}_{\mathcal{N}_j\times\mathcal{A}_{D_j}}\right\}_{j=1}^q \\ \operatorname{diag}\left\{\frac{n_j}{n}\frac{1}{n_j}\mathbf{1}^{\mathrm{T}}_{n_j}\mathbf{G}_{\mathcal{N}_j\times\mathcal{A}_{D_j}}\right\}_{j=1}^q \end{bmatrix}$	$\boldsymbol{\Sigma}_{14} \in \mathbb{R}^{(1+q)\times\sum_{j=1}^q s_{D_j}}$	$\begin{bmatrix} \mathcal{R}\left\{p_j\left(\boldsymbol{\beta}_{0,\mathcal{A}_{D_j}}+\boldsymbol{\beta}_{j,\mathcal{A}_{D_j}}+\mathbf{A}_{\mathcal{A}_{D_j}\times[r]}\boldsymbol{\mu}_X\right)^{\mathrm{T}}\right\}_{j=1}^q \\ \operatorname{diag}\left\{p_j\left(\boldsymbol{\beta}_{0,\mathcal{A}_{D_j}}+\boldsymbol{\beta}_{j,\mathcal{A}_{D_j}}+\mathbf{A}_{\mathcal{A}_{D_j}\times[r]}\boldsymbol{\mu}_X\right)^{\mathrm{T}}\right\}_{j=1}^q \end{bmatrix}$
$\widehat{\boldsymbol{\Sigma}}_{22} \in \mathbb{R}^{r\times r}$	$\frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{X}$	$\boldsymbol{\Sigma}_{22} \in \mathbb{R}^{r\times r}$	$\boldsymbol{\Sigma}_X+\boldsymbol{\mu}_X\boldsymbol{\mu}_X^{\mathrm{T}}$

Block	Expression
$\widehat{\boldsymbol{\Sigma}}_{23} \in \mathbb{R}^{r\times s_G}$	$\sum_{j=1}^q \frac{n_j}{n}\frac{1}{n_j}\mathbf{X}^{\mathrm{T}}_{\mathcal{N}_j\times[r]}\mathbf{G}_{\mathcal{N}_j\times\mathcal{A}_G}+\frac{n_0}{n}\frac{1}{n_0}\mathbf{X}^{\mathrm{T}}_{\mathcal{N}_0\times[r]}\mathbf{G}_{\mathcal{N}_0\times\mathcal{A}_G}$
$\boldsymbol{\Sigma}_{23} \in \mathbb{R}^{r\times s_G}$	$\sum_{j=1}^q p_j\left\{\boldsymbol{\mu}_X\left(\boldsymbol{\beta}_{0,\mathcal{A}_G}+\boldsymbol{\beta}_{j,\mathcal{A}_G}+\mathbf{A}_{\mathcal{A}_G\times[r]}\boldsymbol{\mu}_X\right)^{\mathrm{T}}+\boldsymbol{\Sigma}_X\mathbf{A}^{\mathrm{T}}_{\mathcal{A}_G\times[r]}\right\}+p_0\left\{\boldsymbol{\mu}_X\left(\boldsymbol{\beta}_{0,\mathcal{A}_G}+\mathbf{A}_{\mathcal{A}_G\times[r]}\boldsymbol{\mu}_X\right)^{\mathrm{T}}+\boldsymbol{\Sigma}_X\mathbf{A}^{\mathrm{T}}_{\mathcal{A}_G\times[r]}\right\}$
$\widehat{\boldsymbol{\Sigma}}_{24} \in \mathbb{R}^{r\times\sum_{j=1}^q s_{D_j}}$	$\mathcal{R}\left\{\frac{n_j}{n}\frac{1}{n_j}\mathbf{X}^{\mathrm{T}}_{\mathcal{N}_j\times[r]}\mathbf{G}_{\mathcal{N}_j\times\mathcal{A}_{D_j}}\right\}_{j=1}^q$
$\boldsymbol{\Sigma}_{24} \in \mathbb{R}^{r\times\sum_{j=1}^q s_{D_j}}$	$\mathcal{R}\left\{p_j\left[\boldsymbol{\mu}_X\left(\boldsymbol{\beta}_{0,\mathcal{A}_{D_j}}+\boldsymbol{\beta}_{j,\mathcal{A}_{D_j}}+\mathbf{A}_{\mathcal{A}_{D_j}\times[r]}\boldsymbol{\mu}_X\right)^{\mathrm{T}}+\boldsymbol{\Sigma}_X\mathbf{A}^{\mathrm{T}}_{\mathcal{A}_{D_j}\times[r]}\right]\right\}_{j=1}^q$
$\widehat{\boldsymbol{\Sigma}}_{33} \in \mathbb{R}^{s_G\times s_G}$	$\sum_{j=1}^q \frac{n_j}{n}\frac{1}{n_j}\mathbf{G}^{\mathrm{T}}_{\mathcal{N}_j\times\mathcal{A}_G}\mathbf{G}_{\mathcal{N}_j\times\mathcal{A}_G}+\frac{n_0}{n}\frac{1}{n_0}\mathbf{G}^{\mathrm{T}}_{\mathcal{N}_0\times\mathcal{A}_G}\mathbf{G}_{\mathcal{N}_0\times\mathcal{A}_G}$
$\boldsymbol{\Sigma}_{33} \in \mathbb{R}^{s_G\times s_G}$	$\sum_{j=1}^q p_j\{(\boldsymbol{\beta}_{0,\mathcal{A}_G}+\boldsymbol{\beta}_{j,\mathcal{A}_G}+\mathbf{A}_{\mathcal{A}_G\times[r]}\boldsymbol{\mu}_X)(\boldsymbol{\beta}_{0,\mathcal{A}_G}+\boldsymbol{\beta}_{j,\mathcal{A}_G}+\mathbf{A}_{\mathcal{A}_G\times[r]}\boldsymbol{\mu}_X)^{\mathrm{T}}+\mathbf{A}_{\mathcal{A}_G\times[r]}\boldsymbol{\Sigma}_X\mathbf{A}^{\mathrm{T}}_{\mathcal{A}_G\times[r]}+\boldsymbol{\Sigma}_{G(\mathcal{A}_G\times\mathcal{A}_G)}\}$ $+p_0\{(\boldsymbol{\beta}_{0,\mathcal{A}_G}+\mathbf{A}_{\mathcal{A}_G\times[r]}\boldsymbol{\mu}_X)(\boldsymbol{\beta}_{0,\mathcal{A}_G}+\mathbf{A}_{\mathcal{A}_G\times[r]}\boldsymbol{\mu}_X)^{\mathrm{T}}+\mathbf{A}_{\mathcal{A}_G\times[r]}\boldsymbol{\Sigma}_X\mathbf{A}^{\mathrm{T}}_{\mathcal{A}_G\times[r]}+\boldsymbol{\Sigma}_{G(\mathcal{A}_G\times\mathcal{A}_G)}\}$
$\widehat{\boldsymbol{\Sigma}}_{34} \in \mathbb{R}^{s_G\times\sum_{j=1}^q s_{D_j}}$	$\mathcal{R}\left\{\frac{n_1}{n}\frac{1}{n_1}\mathbf{G}^{\mathrm{T}}_{\mathcal{N}_j\times\mathcal{A}_G}\mathbf{G}_{\mathcal{N}_j\times\mathcal{A}_{D_j}}\right\}_{j=1}^q$
$\boldsymbol{\Sigma}_{34} \in \mathbb{R}^{s_G\times\sum_{j=1}^q s_{D_j}}$	$\mathcal{R}\left\{p_j\left[\left(\boldsymbol{\beta}_{0,\mathcal{A}_G}+\boldsymbol{\beta}_{j,\mathcal{A}_G}+\mathbf{A}_{\mathcal{A}_G\times[r]}\boldsymbol{\mu}_X\right)\left(\boldsymbol{\beta}_{0,\mathcal{A}_{D_j}}+\boldsymbol{\beta}_{j,\mathcal{A}_{D_j}}+\mathbf{A}_{\mathcal{A}_{D_j}\times[r]}\boldsymbol{\mu}_X\right)^{\mathrm{T}}+\mathbf{A}_{\mathcal{A}_G\times[r]}\boldsymbol{\Sigma}_X\mathbf{A}^{\mathrm{T}}_{\mathcal{A}_{D_j}\times[r]}+\boldsymbol{\Sigma}_{G(\mathcal{A}_G\times\mathcal{A}_{D_j})}\right]\right\}_{j=1}^q$
$\widehat{\boldsymbol{\Sigma}}_{44} \in \mathbb{R}^{\sum_{j=1}^q s_{D_j}\times\sum_{j=1}^q s_{D_j}}$	$\operatorname{diag}\left\{\frac{n_j}{n}\frac{1}{n_j}\mathbf{G}^{\mathrm{T}}_{\mathcal{N}_j\times\mathcal{A}_{D_j}}\mathbf{G}_{\mathcal{N}_j\times\mathcal{A}_{D_j}}\right\}_{j=1}^q$
$\boldsymbol{\Sigma}_{44} \in \mathbb{R}^{\sum_{j=1}^q s_{D_j}\times\sum_{j=1}^q s_{D_j}}$	$\operatorname{diag}\left\{p_j\left[\left(\boldsymbol{\beta}_{0,\mathcal{A}_{D_j}}+\boldsymbol{\beta}_{j,\mathcal{A}_{D_j}}+\mathbf{A}_{\mathcal{A}_{D_j}\times[r]}\boldsymbol{\mu}_X\right)\left(\boldsymbol{\beta}_{0,\mathcal{A}_{D_j}}+\boldsymbol{\beta}_{j,\mathcal{A}_{D_j}}+\mathbf{A}_{\mathcal{A}_{D_j}\times[r]}\boldsymbol{\mu}_X\right)^{\mathrm{T}}+\mathbf{A}_{\mathcal{A}_{D_j}\times[r]}\boldsymbol{\Sigma}_X\mathbf{A}^{\mathrm{T}}_{\mathcal{A}_{D_j}\times[r]}+\boldsymbol{\Sigma}_{G(\mathcal{A}_{D_j}\times\mathcal{A}_{D_j})}\right]\right\}_{j=1}^q$