THE UNIVERSITY OF CHICAGO

LATENT VARIABLES IN ‘OMIC’ DATA

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY
CHRISTOPHER GORDON MCKENNAN

CHICAGO, ILLINOIS
AUGUST 2019

To Katherine, Oie, AJ, Molly and Mom. Everything good in my life is because of you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I first and foremost would like to thank my advisor, Dan Nicolae, for his support, encouragement and guidance over the past four years. I always left his office smiling, excited and ready for the next challenge. None of my work would have been possible without him, and I can only hope that I have a similar impact on others as he has had on me.

I am deeply indebted to Carole Ober, who graciously opened up her lab to me and allowed me to collaborate with her and other members of her lab. The motivation for each chapter in this thesis can be traced back to her and her talented group. I also thank James Gern, Hans Bisgaard and other researchers involved in the Children's Respiratory and Environmental Workgroup (CREW) consortium and Copenhagen Prospective Studies on Asthma in Childhood (COPSAC) cohorts for letting me learn from their data.

I would also like to thank Matthew Stephens for letting me attend his group meetings, for always finding time to meet with me and for sharing his experience as a researcher. I now know that I am in good company whenever I get lost in the "cloud". I also thank Chao Gao for his useful feedback on Chapters 2 and 3 of my thesis, and for his general advice on navigating the statistics landscape as a young faculty member.

Last, but certainly not least, I wish to acknowledge Matt Bonakdarpour, Vivak Patel, Jon Eskreis-Winkler, Lei Sun, Yongrui Chen, Youngseok Kim, Petr Panov and the rest of my fellow graduate students for making graduate school feel a little less like graduate school.

ABSTRACT

Nearly all high-throughput ‘omic’ data are influenced by technical and biological factors unknown to the researcher, which, if unaccounted for, can severely obfuscate estimation of and inference on the effects of interest. While the importance of this problem has precipitated the development of many methods that attempt to correct for these latent factors, most are designed for gene expression data and are not amenable for modern, complex experimental designs. In this thesis, we develop novel and provably accurate methodology to estimate and perform inference on the coefficients of interest in a multivariate linear model in the presence of latent covariates. Chapter 2 discusses this problem in the context of DNA methylation in which latent cell type typically confounds the covariate of interest. We then provide the first methods amenable to experimental designs with complex sample correlation structures in Chapters 3 and 4. Lastly, motivated by untargeted LC-MS metabolomic data, we present the first method to account for both unobserved covariates and non-random missing data in Chapter 5.

CHAPTER 1

INTRODUCTION

It is an incredibly exciting time to be involved in biology. The development of high-throughput technologies has provided biologists with a cornucopia of genetic, proteomic and metabolomic data that have the potential to elucidate the genetic components of phenotypes and the mediation of environmental exposures. While collecting these data often warrants a PhD in itself, it is only the germ of such research. The next step is to try to answer the looming question that still remains: how do we make sense of all of these data? This thesis provides novel statistical methodology to navigate this deluge of data.

While we consider many types of "omic" data in this thesis, this work is best motivated by DNA methylation data. DNA methylation is an example of an epigenetic modification in which a methyl group is added to the base cytosine in nuclear DNA. Such a cytosine, which may or may not be methylated, is called a CpG site. The presence or absence of a methyl group on a particular CpG may have consequence on the expression of certain genes in the cell, which can lead to disease or other phenotypes. Consider the ostensibly simple experiment of measuring the blood DNA methylation of $n = 100$ asthmatics and healthy controls across $p = 800,000$ CpG sites in an effort to identify asthma-dependent DNA methylation patterns. If we ignore observed nuisance covariates like the intercept, a simple model relating asthma status $\boldsymbol{X} \in \{0,1\}^n$ to blood methylation levels $\boldsymbol{y}_g \in \mathbb{R}^n$ at CpG $g = 1, \ldots, p$ is

$$\boldsymbol{y}_g = \boldsymbol{X}\beta_g + \boldsymbol{\Delta}_g \quad (g = 1, \ldots, p).$$

Our goal is to estimate and perform inference on the effect due to asthma, β_g.

In classical multivariate regression, $\mathbb{E}\left(\boldsymbol{\Delta}_g\right) = \boldsymbol{0}$, $\mathbb{V}\left(\boldsymbol{\Delta}_g\right) = \sigma_g^2 I_n$ and $\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_p$ are independent. Estimation and inference under these assumptions is well-understood (Gosset 1908, Benjamini & Hochberg 1995). However, DNA methylation data, like all omic data, are not this simple. For example, DNA methylation is cell type dependent (Titus et al. 2017), meaning $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_p$ are actually weighted averages of blood cell type-specific methylation levels, where the weights are each individual's cellular proportions. Unfortunately, cell type is typically too expensive or impossible to measure and is therefore rarely observed by the researcher. This is further complicated by the fact cell type confounds[1] asthma status, since asthmatics tend to have different blood cell proportions than healthy controls (Stein et al. 2016). Besides cell type, other examples of latent covariates (i.e. latent factors) that can influence $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_p$ include sample batch, reagent quality and machine performance, to name a few. If we let $\boldsymbol{C} = (\boldsymbol{c}_1 \cdots \boldsymbol{c}_n) \in \mathbb{R}^{n \times K}$ contain the K latent factors, a more reasonable model for $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_p$ would therefore be

$$\begin{aligned} &\boldsymbol{y}_g = \boldsymbol{X}\beta_g + \boldsymbol{C}\boldsymbol{\ell}_g + \boldsymbol{e}_g, \quad \mathbb{E}\left(\boldsymbol{e}_g\right) = \boldsymbol{0}, \quad \mathbb{V}\left(\boldsymbol{e}_g\right) = \sigma_g^2 I_n \quad (g = 1, \ldots, p) \\ &\boldsymbol{e}_1, \ldots, \boldsymbol{e}_p \text{ are independent.} \end{aligned} \tag{1.1}$$

In this model,

$$\begin{aligned} &\mathbb{E}\left(\boldsymbol{y}_g\right) = \boldsymbol{X}\beta_g + \mathbb{E}\left(\boldsymbol{C}\right)\boldsymbol{\ell}_g \quad (g = 1, \ldots, p) \\ &\operatorname{Cov}\left(\left[\boldsymbol{y}_g\right]_i, \left[\boldsymbol{y}_h\right]_i\right) = \boldsymbol{\ell}_g^{\mathrm{T}}\, \mathbb{V}\left(\boldsymbol{c}_i\right)\boldsymbol{\ell}_h + \sigma_g^2 I\left(g = h\right) \quad (g, h = 1, \ldots, p; i = 1, \ldots, n), \end{aligned}$$

where $\left[\boldsymbol{y}_g\right]_i$ is the ith element of $\boldsymbol{y}_g$. Consequently, accounting for $\boldsymbol{C}$ when estimating and performing inference on β_g mitigates biases and reduces the correlation between test statistics for different CpG sites. In general, the methods we develop in this thesis are designed to estimate β_g in Model (1.1)[2] as accurately as when $\boldsymbol{C}$ is observed.

The problem of latent covariates in high throughput data is well known to be one of the most critical impediments to reproducible research (Leek et al. 2010, Jaffe & Irizarry 2014).

1. We use the term "confounder" throughout this thesis to mean any variable that is associated with both the independent and dependent variables. It does not imply a causal relationship.

2. The assumption on $\mathbb{V}\left(\boldsymbol{e}_g\right)$ in Model (1.1) is not necessary. We consider the model $\mathbb{V}\left(\boldsymbol{e}_g\right) = v_{g,1}\boldsymbol{B}_1 + \cdots + v_{g,b}\boldsymbol{B}_b$ for known $\boldsymbol{B}_1, \ldots, \boldsymbol{B}_b$ in Chapter 3.

Evidently, recovering $\boldsymbol{C}$ from Model (1.1) is hopeless if the number of genomic units $p = 1$, because it would be impossible to distinguish variation due to $\boldsymbol{X}$ from that due to $\boldsymbol{C}$. The key observation is that $\boldsymbol{C}$ in Model (1.1) is shared across all genomic units $g = 1, \ldots, p >> 1$, since all p measurements from a single sample $i \in \{1, \ldots, n\}$ are influenced by sample i's latent cell type, batch number, etc. Therefore, one can potentially recover $\boldsymbol{C}$ (or more precisely, the column space of $\boldsymbol{C}$) if p is suitably large, which is generally the case in the omic data we consider.

Given the importance of this problem, an impressive number of methods have been developed to attempt to solve it (Leek & Storey 2007, 2008, Gagnon-Bartsch & Speed 2012, Sun et al. 2012, Gagnon-Bartsch et al. 2013, Wang et al. 2017, Lee et al. 2017, Fan & Han 2017, Wang et al. 2017, Gerard & Stephens 2018). However, nearly all of these methods are motivated by gene expression data with independent samples $i = 1, \ldots, n$, and therefore tend to fail when applied to other types of omic data or experimental designs. In this thesis, each chapter contains novel methodology to estimate and perform inference on β_g from Model (1.1)[2] in the presence of the latent matrix $\boldsymbol{C}$, which is applicable to gene expression, DNA methylation, proteomic and metabolomic data with potentially correlated samples.

Chapter 2 is motivated by DNA methylation data from Nicodemus-Johnson et al. (2016) in which latent cell type appears to confound the effect of asthma status. We prove that previously proposed methods tend to underestimate the extent of this confounding, which biases their estimates for β_g and inflates their test statistics. We then propose the method BCconf, which we prove corrects said bias and inflation, and also returns estimates for β_g are asymptotically equivalent to the ordinary least squares estimator when $\boldsymbol{C}$ in Model (1.1) is observed. We demonstrate the utility of our method by showing BCconf better accounts for latent cell type in the aforementioned data from Nicodemus-Johnson et al. (2016) than existing methods.

Chapter 3 is motivated by DNA methylation data from McKennan et al. (2018), in which methylation levels from different samples were correlated. We therefore developed CBCV and CorrConf: provably accurate and computationally efficient methods to choose K and recover $\boldsymbol{C}$ in omic data when $\mathbb{V}\left(\boldsymbol{e}_g\right)$ from Model (1.1) is not a multiple of I_n. To the best of our knowledge, these are the first methods to choose K or recover $\boldsymbol{C}$ in such data, and has application in longitudinal data, multi-tissue data, data with multiple treatment conditions and data with related individuals.

In Chapter 4, we analyze longitudinal DNA methylation methylation data to show that race/ethnicity-associated blood methylation patterns are primarily genetically determined. These, along with other results, suggest that blood DNA methylation patterns are robust to many environmental exposures during the first seven years of life. In this chapter, we also extend the false sign rate paradigm introduced in Stephens (2016) and developed a novel inference paradigm for longitudinal DNA methylation data.

Lastly, in Chapter 5, we propose (to the best of our knowledge) the first method to estimate and perform inference on β_g in Model (1.1) when entries of $\boldsymbol{y}_g$ are missing not at random, which has application to untargeted liquid chromatography-mass spectrometry (LC-MS) metabolomic data.

1.1 Notation used throughout this thesis

Let $n > 0$ be an integer. We define $\boldsymbol{1}_n, \boldsymbol{0}_n \in \mathbb{R}^n$ to be the vectors of all ones and zeros, $I_n \in \mathbb{R}^{n \times n}$ to be the identity matrix, $[n] = \{1, \ldots, n\}$ and $[\boldsymbol{x}]_i$ to be the ith element of $\boldsymbol{x} \in \mathbb{R}^n$. For $\boldsymbol{M} \in \mathbb{R}^{n \times m}$, we let $[\boldsymbol{M}]_{ij}$ be the (i, j) element of $\boldsymbol{M}$, and define P_M and $P_M^{\perp}$ to be the orthogonal projections matrices onto $\operatorname{Im}(\boldsymbol{M}) = \{\boldsymbol{M}\boldsymbol{v} : \boldsymbol{v} \in \mathbb{R}^m\}$ and $\boldsymbol{M}^{\perp} = \left\{\boldsymbol{v} \in \mathbb{R}^n : \boldsymbol{M}^{\mathrm{T}}\boldsymbol{v} = \boldsymbol{0}\right\}$. If $d = \dim\left(\boldsymbol{M}^{\perp}\right) > 0$, we define $\boldsymbol{Q}_M \in \mathbb{R}^{n \times d}$ to be a matrix whose columns form an orthonormal basis for $\boldsymbol{M}^{\perp}$, i.e. $P_M^{\perp} = \boldsymbol{Q}_M \boldsymbol{Q}_M^{\mathrm{T}}$. We use the relation $\boldsymbol{X} \overset{\mathcal{D}}{=} \boldsymbol{Y}$ if the two random variables, vectors or matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ have the same distribution, $\boldsymbol{X}_n \overset{\mathcal{D}}{\to} \boldsymbol{X}$ if the sequence $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ converges in distribution to $\boldsymbol{X}$ and $\boldsymbol{X} \dot{\sim} (\boldsymbol{\mu}, \boldsymbol{G})$ if $\mathbb{E}(\boldsymbol{X}) = \boldsymbol{\mu}$ and $\mathbb{V}(\boldsymbol{X}) = \boldsymbol{G}$.

CHAPTER 2

ACCOUNTING FOR UNOBSERVED COVARIATES WITH VARYING DEGREES OF ESTIMABILITY IN HIGH DIMENSIONAL BIOLOGICAL DATA

2.1 Introduction

Suppose we observe data $\boldsymbol{Y} \in \mathbb{R}^{p\times n}$, where the number of genomic units, p, is on the order of or larger than the sample size, n. For example, in most DNA methylation data, the number of studied methylation sites, p, is between 10^4 and 10^6 and n is tens to hundreds. Assume the true model for $\boldsymbol{Y}$ is

$$\boldsymbol{Y}_{p\times n} = \boldsymbol{\beta}_{p\times d}\boldsymbol{X}_{n\times d}^{\mathrm{T}} + \boldsymbol{L}_{p\times K}\boldsymbol{C}_{n\times K}^{\mathrm{T}} + \boldsymbol{E}_{p\times n} \tag{2.1a}$$

$$\boldsymbol{E}_{p\times n} \sim MN_{p\times n}\left(0, \boldsymbol{\Sigma}_{p\times p}, I_n\right), \quad \boldsymbol{\Sigma} = \operatorname{diag}\left(\sigma_1^2, \ldots, \sigma_p^2\right) \tag{2.1b}$$

where $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1 \cdots \boldsymbol{\beta}_p\right)^{\mathrm{T}} \in \mathbb{R}^{p\times d}$, $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ contains the covariates of interest and $\boldsymbol{C} \in \mathbb{R}^{n\times K}$ contains the K unobserved covariates. Our goal is to estimate and perform inference on the coefficients of interest, $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p \in \mathbb{R}^d$.

Under model (2.1), the naive ordinary least squares estimate of $\boldsymbol{\beta}$,

$$\boldsymbol{Y}_1 = \boldsymbol{Y}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1} = \boldsymbol{\beta} + \boldsymbol{L}\left\{\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{C}\right\}^{\mathrm{T}} + \boldsymbol{E}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1} = \boldsymbol{\beta} + \boldsymbol{L}\boldsymbol{\Omega}^{\mathrm{T}} + \boldsymbol{E}_1,$$

is biased by $\boldsymbol{L}\boldsymbol{\Omega}^{\mathrm{T}}$, where $\boldsymbol{\Omega} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{C}$ is the ordinary least squares coefficient estimate for the regression of $\boldsymbol{C}$ on to $\boldsymbol{X}$. The bias induced by $\boldsymbol{L}$ and $\boldsymbol{\Omega}$ is often consequential in biological data. For example, in DNA methylation studies where disease status is the covariate of interest, DNA methylation $\boldsymbol{Y}$ depends on the latent cellular heterogeneity of the n samples (Jaffe & Irizarry 2014), and cellular heterogeneity often depends on disease status $\boldsymbol{X}$ (Fahy 2002, Stein et al. 2016). Ignoring unobserved covariates $\boldsymbol{C}$ when analyzing these types of data can therefore drastically affect the interpretation of results.

There have been a number of methods proposed to estimate and correct for the latent factors $\boldsymbol{C}$ in model (2.1) (Leek & Storey 2008, Gagnon-Bartsch & Speed 2012, Gagnon-Bartsch et al. 2013, Sun et al. 2012, Houseman et al. 2014, Lee et al. 2017). While these methods perform well on selected datasets, they either do not have the requisite theory to justify downstream inference on $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$ (Leek & Storey 2008, Sun et al. 2012, Houseman et al. 2014, Lee et al. 2017), or they require the practitioner to have prior knowledge regarding which coefficients $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$ are zero (Gagnon-Bartsch & Speed 2012, Gagnon-Bartsch et al. 2013).

Recently, Fan & Han (2017), Wang et al. (2017) proposed methods that first compute $\hat{\boldsymbol{L}}$, an estimate of $\boldsymbol{L}$, from $\boldsymbol{Y} P_X^\perp = \boldsymbol{L} \left(P_X^\perp \boldsymbol{C} \right)^{\mathrm{T}} + \boldsymbol{E} P_X^\perp$. They then estimate $\boldsymbol{\Omega}$ by regressing $\boldsymbol{Y}_1$ on to $\hat{\boldsymbol{L}}$, and finally estimate $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$ by subtracting the estimated bias $\hat{\boldsymbol{L}} \hat{\boldsymbol{\Omega}}^{\mathrm{T}}$ from $\boldsymbol{Y}_1$. The advantage of this estimation paradigm is obvious: it decouples the estimation of $\boldsymbol{L}$ and $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$ without requiring the practitioner to have prior knowledge regarding which coefficients $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$ are zero. These articles are quite remarkable because when their assumptions hold, the authors prove that they can perform inference on $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$ that is as accurate as when $\boldsymbol{C}$ is known. However, it has been observed that these methods tend to inflate test statistics and beget anti-conservative inference in both simulated and real data (van Iterson et al. 2017).

One source of the discrepancy between theory and practice is that the aforementioned articles assume that all K of the non-zero eigenvalues of $\boldsymbol{\mathcal{I}} = P_X^\perp \boldsymbol{C} \left(p^{-1} \boldsymbol{L}^{\mathrm{T}} \boldsymbol{L} \right) \boldsymbol{C}^{\mathrm{T}} P_X^\perp$ are on the order of the number of samples, n, and are overtly larger than the average residual variance $p^{-1} \left(\sigma_1^2 + \cdots + \sigma_p^2 \right)$. If these assumptions were valid, there would be an unambiguous gap between the Kth and $[K+1]$st eigenvalues of $p^{-1} P_X^\perp \boldsymbol{Y}^{\mathrm{T}} \boldsymbol{Y} P_X^\perp$. However, this rarely occurs in practice (Cangelosi & Goriely 2007, Owen & Wang 2016, Wang et al. 2017). When these eigenvalue assumptions are violated, we show that previous methods' techniques to estimate $\boldsymbol{\Omega}$ from the regression of $\boldsymbol{Y}_1$ on to $\hat{\boldsymbol{L}}$ are sensitive to the error in the estimated design matrix $\hat{\boldsymbol{L}}$, which begets inaccurate estimates of $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$. In practice, some of the non-zero eigenvalues of $\boldsymbol{\mathcal{I}}$ will not be large if the sample size is not sufficiently large, if some of the K latent covariates do not influence the response of every genomic unit, or if some of the latent covariates are correlated with the covariate of interest $\boldsymbol{X}$, since this will dampen $P_X^{\perp}\boldsymbol{C}$. The latter is common in DNA methylation data because unobserved cellular heterogeneity is often correlated with $\boldsymbol{X}$ (Jaffe & Irizarry 2014).

The purpose of this chapter is to fill the described gap in the literature by studying the unobserved covariate problem when some or all of the K non-zero eigenvalues of $\boldsymbol{\mathcal{I}}$ are not exceedingly large. We prove that when the eigenvalues fall below a certain threshold, then for fixed $g \in \{1, \ldots, p\}$, previous methods have a propensity to inflate type I error when testing the null hypothesis $H_{0,g} : \boldsymbol{\beta}_g = 0$, and even tend to falsely reject $H_{0,g}$ with probability tending to 1 when using the conservative Bonferroni correction. We then provide alternative estimators for $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$ and prove that when $\boldsymbol{\beta}$ is suitably sparse, our estimators are asymptotically equivalent to the ordinary least squares estimators obtained using the design matrix $(\boldsymbol{X}\, \boldsymbol{C})$, regardless of the size of the eigenvalues of $\boldsymbol{\mathcal{I}}$. We lastly use simulated data and real DNA methylation data from Nicodemus-Johnson et al. (2016) to show that latent covariates with ostensibly "small" effects can be detrimental to inference if not properly accounted for, and that our method can better account for latent covariates than the leading competitors. An R package called BCconf that implements our method is freely available for download from GitHub. The proofs of all results can be found in Section 2.7.

2.2 The model, our estimation procedure and intuition

2.2.1 Notation

In addition to the notation defined in Section 1.1, we define $[\boldsymbol{G}]_{r*} \in \mathbb{R}^m$ and $[\boldsymbol{G}]_{*s} \in \mathbb{R}^n$ to be the rth row and sth column of the matrix $\boldsymbol{G} \in \mathbb{R}^{n \times m}$ in this chapter.

2.2.2 A model for the data

Let $\boldsymbol{Y} \in \mathbb{R}^{p \times n}$ be the observed data, where $[\boldsymbol{Y}]_{gi}$ is an observation at genomic unit $g \in [p]$ in sample $i \in [n]$. Let $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ be an observed, full rank matrix containing the covariates of interest and define $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1 \cdots \boldsymbol{\beta}_p\right)^{\mathrm{T}} \in \mathbb{R}^{p \times d}$ to be their corresponding coefficients across all p genomic units. We also define an additional covariate matrix $\boldsymbol{C} \in \mathbb{R}^{n \times K}$ and let $\boldsymbol{L} \in \mathbb{R}^{p \times K}$ be its corresponding coefficient. We assume that $\boldsymbol{C}$ is unobserved but K is known. Evidently, K is rarely known in true data applications. While we acknowledge that estimating K is a challenging problem, there is a large body of work devoted to estimating it (Leek & Storey 2008, Onatski 2010, Gagnon-Bartsch & Speed 2012, Owen & Wang 2016, McKennan & Nicolae 2018*b*). We discuss how different values of K affect our downstream estimates in Section 2.4. We assume (2.1) is the true model for $\boldsymbol{Y}$, and we define

$$\rho = p^{-1}\left(\sigma_1^2 + \cdots + \sigma_p^2\right). \tag{2.2}$$

We also define

$$\boldsymbol{\Omega} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{C} \in \mathbb{R}^{d \times K}, \qquad \boldsymbol{C}_2 = P_X^{\perp}\boldsymbol{C} \in \mathbb{R}^{n \times K} \tag{2.3}$$

to be the ordinary least squares coefficient estimates and residuals from the regression of $\boldsymbol{C}$ on to $\boldsymbol{X}$, respectively. We have not assumed an explicit relationship between $\boldsymbol{C}$ and $\boldsymbol{X}$, because one can always decompose $\boldsymbol{C}$ as

$$\boldsymbol{C} = P_X\boldsymbol{C} + P_X^{\perp}\boldsymbol{C} = X\boldsymbol{\Omega} + \boldsymbol{C}_2.$$

A more general model for $\boldsymbol{Y}$ would be $\boldsymbol{Y} = \boldsymbol{\beta}\boldsymbol{X}^{\mathrm{T}} + \boldsymbol{M}\boldsymbol{Z}^{\mathrm{T}} + \boldsymbol{L}\boldsymbol{C}^{\mathrm{T}} + \boldsymbol{E}$, where $\boldsymbol{Z} \in \mathbb{R}^{n \times r}$ contains observed nuisance covariates, like the intercept or technical covariates, whose coefficients $[\boldsymbol{M}]_{1*}, \ldots, [\boldsymbol{M}]_{p*}$ are not of interest. We can get back to model (2.1) by multiplying $\boldsymbol{Y}$ on the right by a matrix whose columns form an orthonormal basis for the null space of $\boldsymbol{Z}^{\mathrm{T}}$. Therefore, we work exclusively with model (2.1) and assume any observed nuisance factors have already been rotated out, as they would be in ordinary least squares.

2.2.3 Estimating $\boldsymbol{\beta}$ when $\boldsymbol{C}$ is unobserved

We break $\boldsymbol{Y}$ into two independent pieces using a technique proposed in Sun et al. (2012):

$$\boldsymbol{Y}_1 = \boldsymbol{Y}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1} = \boldsymbol{\beta} + \boldsymbol{L}\boldsymbol{\Omega}^{\mathrm{T}} + \boldsymbol{E}_1 \tag{2.4}$$

$$\boldsymbol{Y}_2 = \boldsymbol{Y}P_X^{\perp} = \boldsymbol{L}\boldsymbol{C}_2^{\mathrm{T}} + \boldsymbol{E}_2 \tag{2.5}$$

where $\boldsymbol{E}_1 = \boldsymbol{E}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}$ and $\boldsymbol{E}_2 = \boldsymbol{E}P_X^{\perp}$ are independent because $\boldsymbol{X}^{\mathrm{T}}P_X^{\perp} = 0$ and $\boldsymbol{E} \sim MN_{p\times n}\left(0, \boldsymbol{\Sigma}, I_n\right)$. The matrix $\boldsymbol{Y}_1$ is the ordinary least squares estimate of $\boldsymbol{\beta}$ that ignores $\boldsymbol{C}$, and the rows of $\boldsymbol{E}_2$ lie on an $n-d$ dimensional linear subspace of $\mathbb{R}^n$. We now describe how to use $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ to derive the ordinary least squares estimates of $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_p$ when $\boldsymbol{C}$ is observed. This will provide a template for estimating $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_p$ when $\boldsymbol{C}$ is unobserved.

***Algorithm* 2.1** (Ordinary least squares when $\boldsymbol{C}$ is observed). *Let $\boldsymbol{Y}_1 \in \mathbb{R}^{p\times d}$, $\boldsymbol{Y}_2 \in \mathbb{R}^{p\times n}$, $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ and $\boldsymbol{C} \in \mathbb{R}^{n\times K}$ be given. Our goal is to use ordinary least squares to estimate and perform inference on $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_p$, the rows of $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$.*

(a) Set $\boldsymbol{C}_2 = P_X^{\perp}\boldsymbol{C}$. Use $\boldsymbol{Y}_2$ to estimate $\boldsymbol{\Sigma} = \operatorname{diag}\left(\sigma_1^2, \dots, \sigma_p^2\right)$ and L as

$$\left(\hat{\sigma}_g^{\mathrm{OLS}}\right)^2 = (n-d-K)^{-1}\left[\boldsymbol{Y}_2\right]_{g*}^{\mathrm{T}} P_{C_2}^{\perp}\left[\boldsymbol{Y}_2\right]_{g*} \quad (g = 1, \dots, p)$$
$$\hat{\boldsymbol{L}}^{\mathrm{OLS}} = \boldsymbol{Y}_2\boldsymbol{C}_2\left(\boldsymbol{C}_2^{\mathrm{T}}\boldsymbol{C}_2\right)^{-1}$$

where $\left[\boldsymbol{Y}_2\right]_{g*} \in \mathbb{R}^n$ is the gth row of $\boldsymbol{Y}_2$.

(b) Set $\boldsymbol{\Omega} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{C}$.

(c) Define the ordinary least squares estimate of $\boldsymbol{\beta}_g$ *to be*

$$\hat{\boldsymbol{\beta}}_g^{\mathrm{OLS}} = [\boldsymbol{Y}_1]_{g*} - \boldsymbol{\Omega}\left[\hat{\boldsymbol{L}}^{\mathrm{OLS}}\right]_{g*} \quad (g = 1, \ldots, p) \tag{2.6}$$

where $[\boldsymbol{Y}_1]_{g*} \in \mathbb{R}^d$ *and* $\left[\hat{\boldsymbol{L}}^{\mathrm{OLS}}\right]_{g*} \in \mathbb{R}^K$ *are the gth rows of* $\boldsymbol{Y}_1$ *and* $\hat{\boldsymbol{L}}^{\mathrm{OLS}}$*, respectively.*

It is straightforward to derive the asymptotic properties of the estimators defined in Algorithm 2.1. In Step (a),

$$\hat{\sigma}_g^{\mathrm{OLS}} = \sigma_g + o_{\mathrm{p}}(1) \text{ as } n \to \infty, \quad \hat{\boldsymbol{L}}^{\mathrm{OLS}} \sim MN_{p\times K}\left\{\boldsymbol{L}, \boldsymbol{\Sigma}, \left(\boldsymbol{C}_2^{\mathrm{T}}\boldsymbol{C}_2\right)^{-1}\right\}.$$

Since $\boldsymbol{E}_2$ is independent of $\boldsymbol{E}_1$, both of these estimates are independent of $\boldsymbol{Y}_1$. This implies that for all $g \in [p]$, the asymptotic distribution of $\hat{\boldsymbol{\beta}}_g^{\mathrm{OLS}}$ is

$$\left(\hat{\sigma}_g^{\mathrm{OLS}}\right)^{-1}\left\{\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1} + \boldsymbol{\Omega}\left(\boldsymbol{C}_2^{\mathrm{T}}\boldsymbol{C}_2\right)^{-1}\boldsymbol{\Omega}^{\mathrm{T}}\right\}^{-1/2}\left(\hat{\boldsymbol{\beta}}_g^{\mathrm{OLS}} - \boldsymbol{\beta}_g\right) \overset{\mathcal{D}}{=} Z + o_{\mathrm{p}}(1)$$

as $n \to \infty$, where $\boldsymbol{Z} \sim N_d\left(0, I_d\right)$.

A property of the ordinary least squares estimate $\hat{\boldsymbol{\beta}}_g^{\mathrm{OLS}}$ is

$$\hat{\boldsymbol{\beta}}_g^{\mathrm{OLS}} = [\boldsymbol{Y}_1]_{g*} - \boldsymbol{\Omega}\hat{\boldsymbol{L}}_{g*}^{\mathrm{OLS}} = \left(\boldsymbol{X}^{\mathrm{T}}P_C^{\perp}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}P_C^{\perp}\left[\boldsymbol{Y}\right]_{g*} \quad (g = 1, \ldots, p).$$

That is, $\hat{\boldsymbol{\beta}}_g^{\mathrm{OLS}}$ depends only on the column space $\boldsymbol{C}$, meaning we may replace $\boldsymbol{C}$ with $\boldsymbol{C}\boldsymbol{\Psi}$ as input in Algorithm 2.1 for any invertible matrix $\boldsymbol{\Psi} \in \mathbb{R}^{K\times K}$. In particular, we may choose $\boldsymbol{\Psi}$ so that $n^{-1}\boldsymbol{C}_2^{\mathrm{T}}\boldsymbol{C}_2 = n^{-1}\boldsymbol{C}^{\mathrm{T}}P_X^{\perp}\boldsymbol{C} = I_K$. This parametrization of $\boldsymbol{C}$, and therefore $\boldsymbol{C}_2$, is convenient because it suggests that a reasonable estimate of $\boldsymbol{C}_2$ when $\boldsymbol{C}$ is unobserved is a scalar multiple of the first K right singular vectors of $\boldsymbol{Y}_2$. Using this intuition, we now present our method to estimate and perform inference on $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$ when $\boldsymbol{C}$ is unobserved. This is described in Algorithm 2.2, which mimics the three steps of Algorithm 2.1.

***Algorithm* 2.2** (Estimation and inference when $\boldsymbol{C}$ is unobserved)**.** *Let* $\boldsymbol{Y}_1 \in \mathbb{R}^{p\times d}$, $\boldsymbol{Y}_2 \in$

$\mathbb{R}^{p\times n}$, $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ *and* $K \geq 1$ *be given. Our goal is to estimate and perform inference on* $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$, *the rows of* $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$.

(a) Let $\boldsymbol{Y}_2 = \boldsymbol{U} \operatorname{diag}(\tau_1, \ldots, \tau_n) \boldsymbol{V}^{\mathrm{T}}$ *be the singular value decomposition of* $\boldsymbol{Y}_2$ *where* $\tau_1 \geq \cdots \geq \tau_n \geq 0$ *and* $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} = I_n$. *Define* $\hat{\boldsymbol{C}}_2 = n^{1/2}([\boldsymbol{V}]_{*1} \cdots [\boldsymbol{V}]_{*K})$, *where* $[\boldsymbol{V}]_{*k}$ *is the kth column of* $\boldsymbol{V} \in \mathbb{R}^{n\times n}$. *Estimate* $\boldsymbol{\Sigma} = \operatorname{diag}\left(\sigma_1^2, \ldots, \sigma_p^2\right)$ *and* $\boldsymbol{L}$ *as*

$$\hat{\sigma}_g^2 = (n-d-K)^{-1} [\boldsymbol{Y}_2]_{g*}^{\mathrm{T}} P_{\hat{\boldsymbol{C}}_2}^{\perp} [\boldsymbol{Y}_2]_{g*} \quad (g = 1, \ldots, p) \tag{2.7}$$

$$\hat{\boldsymbol{L}} = \boldsymbol{Y}_2 \hat{\boldsymbol{C}}_2 \left(\hat{\boldsymbol{C}}_2^{\mathrm{T}} \hat{\boldsymbol{C}}_2\right)^{-1}. \tag{2.8}$$

(b) Define $\hat{\rho} = p^{-1}\left(\hat{\sigma}_1^2 + \cdots + \hat{\sigma}_p^2\right)$ *and*

$$\hat{\lambda}_k = np^{-1} \left[\hat{\boldsymbol{L}}\right]_{*k}^{\mathrm{T}} \left[\hat{\boldsymbol{L}}\right]_{*k} \quad (k = 1, \ldots, K), \tag{2.9}$$

where $\left[\hat{\boldsymbol{L}}\right]_{*k}$ *is the kth column of* $\hat{\boldsymbol{L}} \in \mathbb{R}^{p\times K}$. *Estimate* $\boldsymbol{\Omega}$ *as*

$$\hat{\boldsymbol{\Omega}} = \boldsymbol{Y}_1^{\mathrm{T}} \hat{\boldsymbol{L}} \left(\hat{\boldsymbol{L}}^{\mathrm{T}} \hat{\boldsymbol{L}}\right)^{-1} \operatorname{diag}\left\{\hat{\lambda}_1 / \left(\hat{\lambda}_1 - \hat{\rho}\right), \ldots, \hat{\lambda}_K / \left(\hat{\lambda}_K - \hat{\rho}\right)\right\}. \tag{2.10}$$

(c) Estimate $\boldsymbol{\beta}_g$ *as*

$$\hat{\boldsymbol{\beta}}_g = [\boldsymbol{Y}_1]_{g*} - \hat{\boldsymbol{\Omega}} \left[\hat{\boldsymbol{L}}\right]_{g*} \quad (g = 1, \ldots, p). \tag{2.11}$$

Just like the estimates $\left(\hat{\sigma}_g^{\mathrm{OLS}}\right)^2$ and $\hat{\boldsymbol{L}}^{\mathrm{OLS}}$, $\hat{\sigma}_g^2$ and $\hat{\boldsymbol{L}}$ defined in (2.7) and (2.8) and are independent of $\boldsymbol{Y}_1$. To perform inference on $\boldsymbol{\beta}_g$, we assume

$$\hat{\sigma}_g^{-1}\left\{\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}+\hat{\boldsymbol{\Omega}}\left(\hat{\boldsymbol{C}}_2^{\mathrm{T}}\hat{\boldsymbol{C}}_2\right)^{-1}\hat{\boldsymbol{\Omega}}^{\mathrm{T}}\right\}^{-1/2}\left(\hat{\boldsymbol{\beta}}_g-\boldsymbol{\beta}_g\right)\sim N_d\left(0,I_d\right)\quad(g=1,\ldots,p).$$

2.2.4 Intuition regarding Step (b) of Algorithm 2.2

The estimates of σ_g^2 $(g = 1,\ldots,p)$ and L in Step (a) of Algorithm 2.2 are similar to those used in Sun et al. (2012), Gagnon-Bartsch et al. (2013), Lee et al. (2017), Wang et al. (2017). However, the estimate of $\boldsymbol{\Omega}$ in Step (b) is different from those used in previous methods. Recall from (2.4) that $\boldsymbol{Y}_1 = \boldsymbol{\beta} + \boldsymbol{L}\boldsymbol{\Omega}^{\mathrm{T}} + \boldsymbol{E}_1$. If $\boldsymbol{\beta}$ is sufficiently sparse, Sun et al. (2012), Gagnon-Bartsch et al. (2013), Lee et al. (2017), Wang et al. (2017) propose using variations of the following estimator to recover $\boldsymbol{\Omega}$:

$$\hat{\boldsymbol{\Omega}}^{\text{shrunk}} = \boldsymbol{Y}_1^{\mathrm{T}}\hat{\boldsymbol{L}}\left(\hat{\boldsymbol{L}}^{\mathrm{T}}\hat{\boldsymbol{L}}\right)^{-1}. \tag{2.12}$$

That is, they ignore the uncertainty in $\hat{\boldsymbol{L}}$ when regressing $\boldsymbol{Y}_1$ on to $\hat{\boldsymbol{L}}$. To see why this is imprudent, let $\hat{\boldsymbol{R}} = \hat{\boldsymbol{L}} - \boldsymbol{L}$ be the residual and suppose for sake of argument that $\boldsymbol{L} \approx \boldsymbol{0}$. Then the regression coefficients from the regression $\boldsymbol{Y}_1 \sim \hat{\boldsymbol{L}}$ should be very close to $\boldsymbol{0}$, since $\hat{\boldsymbol{L}} \approx \hat{\boldsymbol{R}}$ is independent of $\boldsymbol{Y}_1$. In other words, existing estimates of $\boldsymbol{\Omega}$ are shrunk towards 0. We quantify the shrinkage exactly in Section 2.3.3 and use that result to derive an inflation term, $\hat{\boldsymbol{\Gamma}} = \operatorname{diag}\left\{\hat{\lambda}_1/(\hat{\lambda}_1-\hat{\rho}),\ldots,\hat{\lambda}_K/(\hat{\lambda}_K-\hat{\rho})\right\}$. We then use $\hat{\boldsymbol{\Gamma}}$ to inflate the shrunken estimate $\hat{\Omega}^{\text{shrunk}}$, which allows us to better estimate $\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_p$ in Step (c) of Algorithm 2.2.

The importance of the inflation term $\hat{\boldsymbol{\Gamma}}$ in (2.10) is related to how informative the data are for $\boldsymbol{C}$. The estimate $\hat{\lambda}_k$ $(k = 1,\ldots,K)$ defined in (2.9) is the kth largest eigenvalue of $p^{-1}\boldsymbol{Y}_2^{\mathrm{T}}\boldsymbol{Y}_2$, and can therefore be viewed as an estimate of λ_k, the kth largest eigenvalue of $\boldsymbol{\mathcal{I}} = p^{-1}\mathbb{E}\left(\boldsymbol{Y}_2\right)^{\mathrm{T}}\mathbb{E}\left(\boldsymbol{Y}_2\right) = P_X^{\perp}\boldsymbol{C}\left(p^{-1}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\right)\boldsymbol{C}^{\mathrm{T}}P_X^{\perp}$. The eigenvalue λ_k is also the kth largest eigenvalue of $\left(np^{-1}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\right)\left(n^{-1}\boldsymbol{C}^{\mathrm{T}}P_X^{\perp}\boldsymbol{C}\right)$. When λ_k is sufficiently large for all $k = 1,\ldots,K$, we say that the data are strongly informative for the latent factors $\boldsymbol{C}$. Under this regime,

$\hat{\lambda}_1, \ldots, \hat{\lambda}_K$ will tend to dominate $\hat{\rho}$, an estimate of the constant ρ defined in (2.2), meaning $\hat{\boldsymbol{\Gamma}}$ will be negligible. In this case it suffices to use $\hat{\boldsymbol{\Omega}}^{\text{shrunk}}$ or other previously proposed estimates of Ω in place of $\hat{\boldsymbol{\Omega}}$ in (2.11). On the other hand, we say the data are only moderately informative for $\boldsymbol{C}$ if one or more of $\lambda_1, \ldots, \lambda_K$ is not large. This can occur if the sample size n is not large enough, if some of columns of $\boldsymbol{C}$ do not affect the expression or methylation of all p genomic units, or if $\boldsymbol{X}$ is correlated with the columns of $\boldsymbol{C}$, since this will dampen $P_X^{\perp}\boldsymbol{C}$. In these cases, $\left[\hat{\boldsymbol{\Gamma}}\right]_{11}, \ldots, \left[\hat{\boldsymbol{\Gamma}}\right]_{KK}$ will be moderate to large. In fact, we prove in Section 2.3.3 and show with simulation and a real data example in Section 2.4 that existing methods that ignore the shrinkage in their estimates of $\boldsymbol{\Omega}$ are not amenable to inference. We define the informativeness of the data for $\boldsymbol{C}$ precisely in Definition 2.1 in Section 2.3.3.

2.3 Theoretical results

2.3.1 Assumptions

In all of our assumptions and theoretical results, we assume model (2.1) holds, $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ are as defined in (2.4) and (2.5), and $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1 \cdots \boldsymbol{\beta}_p\right)^{\mathrm{T}} \in \mathbb{R}^{p \times d}$.

***Assumption* 2.1.** *(a) $\boldsymbol{X}$ is an observed, non-random matrix such that $\lim_{n \to \infty} n^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = \boldsymbol{\Sigma}_X \succ 0$.*

(b) $\boldsymbol{L}\boldsymbol{C}^{\mathrm{T}} \in \mathbb{R}^{p \times n}$ is an unobserved, non-random matrix with K non-zero singular values, where $K \geq 1$ is a known constant.

(c) For some constant $c_1 > 1$, $\sigma_g^2 \in \left[c_1^{-1}, c_1\right]$ for all $g = 1, \ldots, p$.

Under (a) and (b), $\mathbb{E}\left(\boldsymbol{Y}_2\right) = \boldsymbol{L}\left(P_X^{\perp}\boldsymbol{C}\right)^{\mathrm{T}} = \boldsymbol{L}\boldsymbol{C}_2^{\mathrm{T}}$, $\mathbb{E}\left(\boldsymbol{Y}_1\right) = \boldsymbol{\beta} + \boldsymbol{L}\boldsymbol{\Omega}^{\mathrm{T}}$ and $\mathbb{V}\left(\left[\boldsymbol{Y}\right]_{g1}\right) = \sigma_g^2$ $(g = 1, \ldots, p)$ are identifiable. The choice to treat $\boldsymbol{L}\boldsymbol{C}^{\mathrm{T}}$ as non-random is to illustrate that ignoring this term tends to bias estimates of $\boldsymbol{\beta}$. However, all of our results in Sections 2.3.2, 2.3.3 and 2.3.4 can be extended to the case when $\boldsymbol{L}\boldsymbol{C}^{\mathrm{T}}$ is a random variable using results from Proposition 2.7, Corollary 2.3 and Lemma 2.7 in Section 2.7.9. Item (c) is a standard assumption in the high dimensional factor analysis literature (Bai & Li 2012, Wang et al. 2017). We next place assumptions on $\boldsymbol{L}\boldsymbol{C}_2^{\mathrm{T}}$.

***Assumption* 2.2.** *Let $\boldsymbol{\mathcal{I}} = \boldsymbol{C}_2\left(p^{-1}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\right)\boldsymbol{C}_2^{\mathrm{T}} \in \mathbb{R}^{n\times n}$ and $c_2 > 1$ be a constant.*

(a) *$[\boldsymbol{L}]_{g*}^{\mathrm{T}}\left(n^{-1}\boldsymbol{C}_2^{\mathrm{T}}\boldsymbol{C}_2\right)[\boldsymbol{L}]_{g*} \leq c_2^2$ for all $g = 1,\ldots,p$.*

(b) *$\boldsymbol{\mathcal{I}}$ has K non-zero eigenvalues $\lambda_1 > \cdots > \lambda_K > 0$ such that $\lambda_k \in \left[c_2^{-1}, c_2 n\right]$ and $\left(\lambda_k - \lambda_{k+1}\right)/\lambda_k \geq c_2^{-1}$ for all $k = 1,\ldots,K$, where $\lambda_{K+1} = 0$.*

(c) *p is a non-decreasing function of n such that $n/p \leq c_2$ and $n^{3/2}/\left(p\lambda_K\right) \to 0$ as $n \to \infty$.*

The quantity $[\boldsymbol{L}]_{g*}^{\mathrm{T}}\left(n^{-1}\boldsymbol{C}_2^{\mathrm{T}}\boldsymbol{C}_2\right)[\boldsymbol{L}]_{g*}$ is identifiable because $\boldsymbol{L}\boldsymbol{C}_2^{\mathrm{T}}$ is identifiable, and Item (a) is equivalent to $\|[\boldsymbol{L}]_{g*}\|_2 \leq c_2$ for all $g = 1,\ldots,p$ if $n^{-1}\boldsymbol{C}_2^{\mathrm{T}}\boldsymbol{C}_2 = I_K$. We comment on this further after we state Proposition 2.1 below. The assumptions on $\lambda_1,\ldots,\lambda_K$ in (b) are weaker than those considered in previous work that provide inferential guarantees, which focused on the case when $\lambda_1 \asymp \lambda_K \asymp n$ (Bai & Li 2012, Fan & Han 2017, Wang et al. 2017). The authors of Lee et al. (2017) do allow $\lambda_K = o(n)$, provided $\lambda_1 \asymp \lambda_K$ and $\lambda_K \to \infty$ as $n \to \infty$. However, they only prove the consistency of their estimates of $\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_p$. In fact, we show in Section 2.3.3 that inference with their method, as well as other existing methods, is fallacious if $\lambda_K \in \left[c^{-1}, cn^{1/2}\right]$ for some $c > 1$. The assumptions on n, p in (c) are the same as those used in Wang et al. (2017), which only considers the case $\lambda_1 \asymp \lambda_K \asymp n$. We next place assumptions on the parameters of $\mathbb{E}\left(\boldsymbol{Y}_1\right) = \boldsymbol{\beta} + \boldsymbol{L}\boldsymbol{\Omega}^{\mathrm{T}}$.

***Assumption* 2.3.** *Let $c_3 > 0$ be a constant.*

(a) *$p^{-1}\left\{I\left([\boldsymbol{\beta}]_{1r} \neq 0\right) + \cdots + I\left([\boldsymbol{\beta}]_{pr} \neq 0\right)\right\} = o\left(n^{-3/2}\lambda_K\right)$ for all $r \in [d]$ as $n, p \to \infty$.*

(b) *$|[\boldsymbol{\beta}]_{gr}| \leq c_3$ for all $g \in [p]$ and $r \in [d]$.*

(c) *Let $\boldsymbol{C} \in \mathbb{R}^{n\times K}$ be any matrix such that $\mathbb{E}\left(\boldsymbol{Y}\right) = \boldsymbol{\beta}\boldsymbol{X}^{\mathrm{T}} + \boldsymbol{L}\boldsymbol{C}^{\mathrm{T}}$ for some $\boldsymbol{L} \in \mathbb{R}^{p\times K}$. Then for $\boldsymbol{\Omega} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{C}$ and $\boldsymbol{C}_2 = P_X^{\perp}\boldsymbol{C}$, $\|\boldsymbol{\Omega}\left(n^{-1}\boldsymbol{C}_2^{\mathrm{T}}\boldsymbol{C}_2\right)^{-1}\boldsymbol{\Omega}^{\mathrm{T}}\|_2 \leq c_3$.*

Item (a) is the same sparsity as assumed in Wang et al. (2017). Item (c) is justifiable because we prove that $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}\left(n^{-1}\boldsymbol{C}_2^{\mathrm{T}}\boldsymbol{C}_2\right)^{-1}\boldsymbol{\Omega}^{\mathrm{T}}$ are identifiable under Assumptions 2.1, 2.2 and 2.3(a) in Proposition 2.1 below and Proposition 2.3 in Section 2.7.1.

In DNA methylation data with $p \approx 3 \times 10^5 - 8 \times 10^5$, $n \approx 10^2$ and in the previously unexplored regime $\lambda_K \in \left[c^{-1}, cn^{1/2}\right]$, Assumption 2.3(a) restricts the number of genomic units with non-zero coefficient of interest to be on the order of hundreds to thousands, which is common in many studies (Morales et al. 2016, Liu et al. 2016, Yang et al. 2017, Zhang et al. 2018). We also show through simulations that we can egregiously violate Assumption 2.3(a) and still do accurate inference on $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$. We now state a proposition regarding the identifiability of $\boldsymbol{L}$ and $\boldsymbol{C}$.

***Proposition* 2.1.** *Let* $\mathcal{G} = \left\{\operatorname{diag}\left(a_1, \ldots, a_K\right) : a_1, \ldots, a_K \in \{-1, 1\}\right\}$, *suppose Assumptions 2.1 and 2.2 hold and define the parameter space*

$$\begin{aligned}\Theta_{(0)} = \Big\{ &(\boldsymbol{L}, \boldsymbol{C}) \in \mathbb{R}^{p\times K} \times \mathbb{R}^{n\times K} : \mathbb{E}\left(\boldsymbol{Y}_2\right) = \boldsymbol{L}\boldsymbol{C}_2^{\mathrm{T}}, n^{-1}\boldsymbol{C}_2^{\mathrm{T}}\boldsymbol{C}_2 = I_K, \\ &np^{-1}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L} = \operatorname{diag}\left(\lambda_1, \ldots, \lambda_K\right) \text{ for } \boldsymbol{C}_2 = P_X^{\perp}\boldsymbol{C} \Big\}.\end{aligned} \tag{2.13}$$

Then $\Theta_{(0)}$ *is non-empty and if* $\left\{\boldsymbol{L}^{(a)}, \boldsymbol{C}^{(a)}\right\}, \left\{\boldsymbol{L}^{(b)}, \boldsymbol{C}^{(b)}\right\} \in \Theta_{(0)}$, *then* $\boldsymbol{L}^{(b)} = \boldsymbol{L}^{(a)}\boldsymbol{\Pi}$ *and* $\boldsymbol{C}_2^{(b)} = \boldsymbol{C}_2^{(a)}\boldsymbol{\Pi}$ *for some* $\boldsymbol{\Pi} \in \mathcal{G}$. *If Assumptions 2.1, 2.2 and 2.3(a) hold, then there exists a constant* $c_4 > 0$ *such that* $\boldsymbol{\beta}$ *is identifiable and*

$$\Theta_{(1)} = \Theta_{(0)} \cap \left\{(\boldsymbol{L}, \boldsymbol{C}) \in \mathbb{R}^{p\times K} \times \mathbb{R}^{n\times K} : \mathbb{E}\left(\boldsymbol{Y}_1\right) = \boldsymbol{\beta} + \boldsymbol{L}\boldsymbol{\Omega}^{\mathrm{T}} \text{ for } \boldsymbol{\Omega} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{C}\right\} \tag{2.14}$$

is non-empty for all $n \geq c_4$. *Further, if* $\left\{\boldsymbol{L}^{(a)}, \boldsymbol{C}^{(a)}\right\}, \left\{\boldsymbol{L}^{(b)}, \boldsymbol{C}^{(b)}\right\} \in \Theta_{(1)}$, *then* $\boldsymbol{L}^{(b)} = \boldsymbol{L}^{(a)}\boldsymbol{\Pi}$ *and* $\boldsymbol{C}^{(b)} = \boldsymbol{C}^{(a)}\boldsymbol{\Pi}$ *for some* $\boldsymbol{\Pi} \in \mathcal{G}$ *for all* $n \geq c_4$.

The condition that $(\boldsymbol{L}, \boldsymbol{C}) \in \Theta_{(0)}$ is a classic constraint to identify the components of factor models (Bai & Li 2012). If $(\boldsymbol{L}, \boldsymbol{C}) \in \Theta_{(0)}$, Assumption 2.2(a) becomes $\|[\boldsymbol{L}]_{g*}\|_2 \leq c_2$

for all $g = 1, \ldots, p$ and if $(\boldsymbol{L}, \boldsymbol{C}) \in \Theta_{(1)}$, $\boldsymbol{\Omega} \left(n^{-1}\boldsymbol{C}_2^{\mathrm{T}}\boldsymbol{C}_2\right)^{-1} \boldsymbol{\Omega}^{\mathrm{T}} = \boldsymbol{\Omega}\boldsymbol{\Omega}^{\mathrm{T}}$. While we prove it is unnecessary to assume a particular parametrization of $\boldsymbol{L}$ and $\boldsymbol{C}$ to estimate and perform inference on $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$ using Algorithm 2.2, we use the parameter spaces $\Theta_{(0)}$ and $\Theta_{(1)}$ in the statements of theoretical results regarding the accuracy of estimates of $\boldsymbol{L}$ and $\boldsymbol{\Omega}$, respectively, in Sections 2.3.2, 2.3.3 and 2.3.4.

2.3.2 Asymptotic properties of the estimates from Step (a) of Algorithm 2.2

We start by illustrating the asymptotic properties of $\hat{\sigma}_g^2$ $(g = 1, \ldots, p)$ and $\hat{\boldsymbol{L}}$ defined in (2.7) and (2.8).

***Lemma* 2.1.** *Suppose Assumptions 2.1 and 2.2 hold and $n \to \infty$. Then for ρ defined in* (2.2)*,*

$$\hat{\sigma}_g^2 = \sigma_g^2 + o_{\mathrm{p}}(1) \quad (g = 1, \ldots, p) \tag{2.15}$$

$$\hat{\rho} = p^{-1}\left(\hat{\sigma}_1^2 + \cdots + \hat{\sigma}_p^2\right) = \rho + o_{\mathrm{p}}\left(n^{-1/2}\right). \tag{2.16}$$

***Lemma* 2.2.** *Suppose Assumptions 2.1 and 2.2 hold and $n \to \infty$. Then for $\hat{\lambda}_1, \ldots, \hat{\lambda}_K$ defined in* (2.9)*,*

$$\hat{\lambda}_k/\lambda_k = 1 + \rho/\lambda_k + o_{\mathrm{p}}\left(n^{-1/2}\right) \quad (k = 1, \ldots, K). \tag{2.17}$$

Let $\hat{\boldsymbol{C}}_2$ be as defined in Step (a) of Algorithm 2.2. If we also assume that $(\boldsymbol{L}, \boldsymbol{C}) \in \Theta_{(0)}$ and the K diagonal elements of $\hat{\boldsymbol{C}}_2^{\mathrm{T}}\boldsymbol{C}_2$ are non-negative, then for $\boldsymbol{W} \sim N_K(0, I_K)$,

$$n^{1/2}\hat{\sigma}_g^{-1}\left(\left[\hat{\boldsymbol{L}}\right]_{g*} - [\boldsymbol{L}]_{g*}\right) \overset{\mathcal{D}}{=} \boldsymbol{W} + o_{\mathrm{p}}(1) \quad (g = 1, \ldots, p). \tag{2.18}$$

***Remark* 2.1.** *The identifiability constraints, that $(\boldsymbol{L}, \boldsymbol{C}) \in \Theta_{(0)}$ and $\hat{\boldsymbol{C}}_2^{\mathrm{T}}\boldsymbol{C}_2$ has non-negative* *diagonal elements, are equivalent to the IC3 constraint used in Bai & Li (2012) to identify the components of factor models.*

***Remark* 2.2.** *When $\boldsymbol{C}$ is observed and $(\boldsymbol{L}, \boldsymbol{C}) \in \Theta_{(0)}$,* (2.17) *and* (2.18) *hold for the ordinary least squares estimator $\hat{\boldsymbol{L}}^{\text{OLS}}$ defined in Step (a) of Algorithm 2.1.*

***Remark* 2.3.** *Equation* (2.18) *shows that the asymptotic distribution for $\left[\hat{\boldsymbol{L}}\right]_{g*}^{\mathrm{T}} \left[\hat{\boldsymbol{L}}\right]_{g*}$ under the null hypothesis $[\boldsymbol{L}]_{g*} = \boldsymbol{0}$ is $n\hat{\sigma}_g^{-2} \left[\hat{\boldsymbol{L}}\right]_{g*}^{\mathrm{T}} \left[\hat{\boldsymbol{L}}\right]_{g*} \overset{\mathcal{D}}{=} z^2 + o_P(1)$ as $n \to \infty$, where $z^2 \sim \chi_K^2$. Further, the asymptotic distribution does not depend on the parametrization of $\boldsymbol{L}$ or $\boldsymbol{C}$.*

Lemmas 2.1 and 2.2 show that $\left[\hat{\boldsymbol{L}}\right]_{g*}$ and $\hat{\sigma}_g^2$ have the same asymptotic properties as $\left[\hat{\boldsymbol{L}}^{\text{OLS}}\right]_{g*}$ and $\left(\hat{\sigma}_g^{\text{OLS}}\right)^2$, the ordinary least squares estimates of $[\boldsymbol{L}]_{g*}$ and σ_g^2 defined in Algorithm 2.1. However, (2.17) states that the estimates of λ_k are biased by ρ, which we show below is the primary reason why previously proposed methods often return inflated test statistics.

2.3.3 Previous estimates of $\boldsymbol{\Omega}$ in Step (b) of Algorithm 2.2 inflate test statistics

As noted in Section 2.2.4, existing methods that use the estimation paradigm outlined in Algorithm 2.2 ignore the uncertainty in $\hat{\boldsymbol{L}}$, and use variations of $\hat{\boldsymbol{\Omega}}^{\text{shrunk}}$ to estimate $\boldsymbol{\Omega}$. We show that these methods tend to underestimate $\boldsymbol{\Omega}$, which can lead to spurious inference on $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$ in Proposition 2.2 and Corollary 2.1 below.

***Proposition* 2.2.** *Suppose Assumptions 2.1, 2.2 and 2.3 hold with $\boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_p = \boldsymbol{0}$, $n \to \infty$ and $(\boldsymbol{L}, \boldsymbol{C}) \in \Theta_{(1)}$, where $\Theta_{(1)}$ was defined in* (2.14). *In addition, suppose the diagonal elements of $\hat{\boldsymbol{C}}_2^{\mathrm{T}} \boldsymbol{C}_2$ are non-negative and $\lambda_1/\lambda_K \le c_5$ for some constant $c_5 > 1$. If we estimate $\boldsymbol{\Omega}$ as $\hat{\boldsymbol{\Omega}}^{\text{shrunk}}$ defined in* (2.12), *then*

$$n^{1/2} \|\hat{\boldsymbol{\Omega}}^{\text{shrunk}} - \boldsymbol{\Omega}\, \mathrm{diag}\left\{\lambda_1/(\rho + \lambda_1), \ldots, \lambda_K/(\rho + \lambda_K)\right\}\|_2 = o_{\mathrm{p}}\left(1\right). \tag{2.19}$$

***Corollary* 2.1.** *Fix some $g \in [p]$ and let $c_6 > 0$ be a small constant. In addition to the assumptions of Proposition 2.2, suppose $d = 1$ and the following hold:*

(i) *We replace $\hat{\boldsymbol{\Omega}}$ with $\hat{\boldsymbol{\Omega}}^{\text{shrunk}}$ in* (2.11) *and estimate $\boldsymbol{\beta}_g = \beta_g = 0 \in \mathbb{R}$ as $\hat{\beta}_g^{\text{shrunk}} = [\boldsymbol{Y}_1]_g - \hat{\boldsymbol{\Omega}}^{\text{shrunk}} \left[\hat{\boldsymbol{L}}\right]_{g*}$.*

(ii) $|\sum_{k=1}^{K} [\boldsymbol{\Omega}]_{1k} [\boldsymbol{L}]_{gk} \{(\lambda_K + \rho) / (\lambda_k + \rho)\}| \geq \epsilon$ *for some constant $\epsilon > 0$.*

Define $z_g = \hat{\sigma}_g^{-1} \left(\|\boldsymbol{X}\|_2^{-2} + n^{-1} \|\hat{\boldsymbol{\Omega}}^{\text{shrunk}}\|_2^2\right)^{-1/2} \hat{\beta}_g^{\text{shrunk}}$ to be the gth z-score and let $\alpha \in (0, 1)$ be any significance level. Then for $q_{1-\alpha/2}$ the $1 - \alpha/2$ quantile of the standard normal distribution, there exists a constant $\delta > 0$ such that as $n \to \infty$,

$$\begin{cases} \mathbb{P}\left(|z_g| > q_{1-\alpha/2}\right) = \alpha + o(1) & \textit{if } \lambda_K^{-1} n^{1/2} \to 0 \\ \mathbb{P}\left(|z_g| > q_{1-\alpha/2}\right) \geq \alpha + \delta + o(1) & \textit{if } \lambda_K^{-1} n^{1/2} \geq c_6 \\ \mathbb{P}\left(|z_g| > q_{1-\alpha/2}\right) = 1 + o(1) & \textit{if } \lambda_K^{-1} n^{1/2} \to \infty \end{cases}.$$

Further, if $n^{-r} p \to 0$ for some constant $r > 0$ and $\lambda_K^{-1} n^{1/2 - c_6} \to \infty$ as $n \to \infty$, then

$$\mathbb{P}\left(|z_g| > q_{1-(p^{-1}\alpha)/2}\right) = 1 + o(1)$$

as $n \to \infty$, where $p^{-1}\alpha$ is the Bonferroni threshold at a level α.

***Remark* 2.4.** *The authors of Gagnon-Bartsch et al. (2013) use $\hat{\boldsymbol{\Omega}}^{\text{shrunk}}$ to estimate $\boldsymbol{\Omega}$, but the authors of Wang et al. (2017), Lee et al. (2017) use slightly different estimators. We prove analogous versions of Proposition 2.2 and Corollary 2.1 for the estimators used in Wang et al. (2017), Lee et al. (2017) in Propositions 2.5 and 2.6 and Corollary 2.2 in Section 2.7.8.*

***Remark* 2.5.** *The assumption that $\lambda_1 / \lambda_K \leq c_5$ made in Proposition 2.2 and Corollary 2.1 requires the eigenvalues be on the same order of magnitude. It is a standard assumption* *made by previous authors who use versions of Algorithm 2.2 to estimate $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$ (Wang et al. 2017, Lee et al. 2017). We discuss how to extend it to allow λ_1/λ_K to diverge in Remark 2.7 after the statement of Theorem 2.2.*

When Condition (ii) in the statement of Corollary 2.1 does not hold, it implies the bias $\boldsymbol{\Omega}\left[\boldsymbol{L}\right]_{g*}$ in $\left[\boldsymbol{Y}_1\right]_{g*}$ is minor, or the largest components of $\boldsymbol{\Omega}$ load on to the columns of $\boldsymbol{L}$ corresponding to the largest eigenvalues λ_k, which are the components least affected by the shrinkage in Proposition 2.2. The shrinkage in $\hat{\boldsymbol{\Omega}}^{\text{shrunk}}$ will have less of an impact on inference in these cases. If $K = 1$, Item (ii) can be replaced with $\left|\boldsymbol{\Omega}\left[\boldsymbol{L}\right]_{g*}\right| \geq \epsilon$ for some constant $\epsilon > 0$.

The results of Proposition 2.2 and Corollary 2.1 show that ignoring the uncertainty in $\hat{\boldsymbol{L}}$ when estimating $\boldsymbol{\Omega}$ can lead to inflated test statistics and type I errors if $\lambda_K^{-1} n^{1/2}$ is not small enough, even if one uses the conservative Bonferroni threshold. We therefore define the informativeness of the data for $\boldsymbol{C}$ in terms of the magnitude of λ_K in relation to $n^{1/2}$.

***Definition* 2.1** (Informativeness of the data for $\boldsymbol{C}$)**.** *The data $\boldsymbol{Y}$ are strongly informative for $\boldsymbol{C}$ if $\lambda_K^{-1} n^{1/2} \to 0$ as $n \to \infty$ and moderately informative for $\boldsymbol{C}$ if there exists a constant $c_7 > 1$ such that $\lambda_K \in \left[c_7^{-1}, n^{1/2} c_7\right]$ for all n.*

Corollary 2.1 shows that existing methods risk performing anti-conservative inference when the data are only moderately informative for $\boldsymbol{C}$. We next show that our shrinkage-corrected estimate of $\boldsymbol{\Omega}$ in (2.10) begets estimates of $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$ that are asymptotically equivalent to the corresponding ordinary least squares estimates obtained when $\boldsymbol{C}$ is known, even when the data are only moderately informative for $\boldsymbol{C}$.

2.3.4 The estimates of $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$ from Algorithms 2.1 and 2.2 are asymptotically equivalent

We first prove that our shrinkage-corrected estimate of $\boldsymbol{\Omega}$, $\hat{\boldsymbol{\Omega}}$, corrects the aforementioned shrinkage present in existing methods' estimates of $\boldsymbol{\Omega}$.

***Lemma* 2.3.** *Suppose Assumptions 2.1, 2.2 and 2.3 hold and* $(\boldsymbol{L}, \boldsymbol{C}) \in \Theta_{(1)}$. *Further, assume the diagonal entries of* $\hat{\boldsymbol{C}}_2^{\mathrm{T}} \boldsymbol{C}_2$ *are non-negative and* $\lambda_1/\lambda_K \leq c_5$*, where* $c_5 > 1$ *was defined in the statement of Proposition 2.2. If* $\hat{\boldsymbol{\Omega}}$ *is defined as* (2.10) *and* $n \to \infty$*, then*

$$n^{1/2} \|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}\|_2 = o_{\mathrm{p}}(1)\,. \tag{2.20}$$

We use this result to prove that inference with $\hat{\boldsymbol{\beta}}_g$ $(g = 1, \ldots, p)$ is asymptotically equivalent to the ordinary least squares estimator obtained when $\boldsymbol{C}$ is known.

***Theorem* 2.1.** *Let* $g \in [p]$ *and suppose Assumptions 2.1, 2.2 and 2.3 hold with* $\lambda_1/\lambda_K \leq c_5$ *and* $n \to \infty$*. Then inference with* $\hat{\boldsymbol{\beta}}_g$ *is asymptotically equivalent to inference with* $\hat{\boldsymbol{\beta}}_g^{\mathrm{OLS}}$ *in the following sense:*

$$n^{1/2} \|\hat{\boldsymbol{\beta}}_g - \hat{\boldsymbol{\beta}}_g^{\mathrm{OLS}}\|_2 = o_{\mathrm{p}}(1) \tag{2.21}$$

$$\hat{\sigma}_g^{-1} \left\{ \left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{X}\right)^{-1} + n^{-1} \hat{\boldsymbol{\Omega}} \hat{\boldsymbol{\Omega}}^{\mathrm{T}} \right\}^{-1/2} \left(\hat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_g\right) \overset{\mathcal{D}}{=} \boldsymbol{Z} + o_{\mathrm{p}}(1). \tag{2.22}$$

The estimates $\hat{\boldsymbol{\beta}}_g^{\mathrm{OLS}}$, $\hat{\boldsymbol{\Omega}}$ *and* $\hat{\boldsymbol{\beta}}_g$ *are defined in* (2.6), (2.10) *and* (2.11), *and* $\boldsymbol{Z} \sim N_d(0, I_d)$.

In some real experimental data, the largest eigenvalue λ_1 may be substantially larger than the smallest eigenvalue λ_K. We therefore extend Theorem 2.1 to relax the assumption that the λ_k's be the same order of magnitude in the following theorem.

***Theorem* 2.2.** *Let* $g \in [p]$*, suppose Assumptions 2.1, 2.2 and 2.3 hold and assume* $n \to \infty$*. Define* $\boldsymbol{m}_k \in \mathbb{R}^p$ *to be the* k*th left singular vector of* $\boldsymbol{L}\boldsymbol{C}_2^{\mathrm{T}}$ $(k = 1, \ldots, K)$*. If*

$$(\lambda_r \lambda_s)^{1/2} |\boldsymbol{m}_r^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{m}_s| \leq c_8 \lambda_{\max(r,s)}$$

for some constant $c_8 > 0$ *for all* $r, s \in [K]$*, then* (2.21) *and* (2.22) *hold.*

***Remark* 2.6.** $(\lambda_r \lambda_s)^{1/2} |\boldsymbol{m}_r^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{m}_s|$ *is identifiable for all* $r, s \in [K]$ *under Assumptions 2.1 and 2.2. If Assumptions 2.1 and 2.2 hold and* $(\boldsymbol{L}, \boldsymbol{C}) \in \Theta_{(0)}$, $(\lambda_r \lambda_s)^{1/2} |\boldsymbol{m}_r^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{m}_s| =$

$|np^{-1}[\boldsymbol{L}]_{*r}^{\mathrm{T}}\boldsymbol{\Sigma}[\boldsymbol{L}]_{*s}|$ *for all* $r, s \in [K]$.

***Remark* 2.7.** *Proposition 2.2 and Corollary 2.1 can be extended to accommodate data where* λ_1/λ_K *diverges by replacing the condition that* $\lambda_1/\lambda_K \leq c_5$ *with* $(\lambda_r\lambda_s)^{1/2}|\boldsymbol{m}_r^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{m}_s| \leq c_8\lambda_{\max(r,s)}$ *for all* $r, s \in [K]$.

The condition on $\boldsymbol{m}_r^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{m}_s$ $(r, s = 1, \ldots, K)$ is quite general, as it can be shown to hold in probability when $[\boldsymbol{L}]_{g*} \sim F_\ell$ and $\sigma_g^2 \sim F_{\sigma^2}$ $(g = 1, \ldots, p)$ for any distributions F_ℓ and F_{σ^2} with compact support, such that $np^{-1}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}$ has eigenvalues bounded away from 0 with high probability. We refer the reader to Proposition 2.4 in Section 2.7.2 for more detail.

2.3.5 *Inference on the relationship between* $\boldsymbol{C}$ *and* $\boldsymbol{X}$

One may be interested in understanding the origin of $\boldsymbol{C}$. For example, if components of $\hat{\boldsymbol{\Omega}}$ were large, it would be informative to know if this were due to random experimental variation, or if some of the columns of $\boldsymbol{C}$ truly depended on $\boldsymbol{X}$. To incorporate this type of inference, we state the following theorem that allows $\boldsymbol{C}$, and therefore $\boldsymbol{\Omega}$, to be treated as a random variable.

***Theorem* 2.3.** *Let* $c_9 > 1$ *be a constant. In addition to Assumptions 2.1(a), 2.1(c) and 2.3(b), suppose the following hold*

(i) $\|\boldsymbol{X}\|_\infty \leq c_9$ *and* $\boldsymbol{L} \in \mathbb{R}^{p \times K}$ *is a non-random matrix such that* $np^{-1}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L} = \mathrm{diag}(\lambda_1, \ldots, \lambda_K)$*, where* $K \geq 1$ *is known.*

(ii) *Let* $\lambda_{K+1} = 0$*. Then* $\lambda_k \in \left[c_9^{-1}, c_9 n\right]$ *and* $(\lambda_k - \lambda_{k+1})/\lambda_k \geq c_9^{-1}$ *for all* $k \in [K]$*,* $\|[\boldsymbol{L}]_{g*}\|_2 \leq c_9$ *for all* $g \in [p]$ *and* $|np^{-1}[\boldsymbol{L}]_{*r}^{\mathrm{T}}\boldsymbol{\Sigma}[\boldsymbol{L}]_{*s}| \leq c_9\lambda_{\max(r,s)}$ *for all* $r, s \in [K]$.

(iii) p *is a non-decreasing function of* n *such that* $n/p \leq c_9$*,* $n^{3/2}/(\lambda_K p) \to 0$ *as* $n \to \infty$ *and* $p^{-1}\left\{I([\boldsymbol{\beta}]_{1r} \neq 0) + \cdots + I\left([\boldsymbol{\beta}]_{pr} \neq 0\right)\right\} = o\left(n^{-3/2}\lambda_K\right)$ *for all* $r \in [d]$.

(iv) $\boldsymbol{C} = \boldsymbol{X}\boldsymbol{A} + \boldsymbol{\Xi} \in \mathbb{R}^{n \times K}$ *where* $A \in \mathbb{R}^{d \times K}$ *is non-random and* $\boldsymbol{\Xi} \in \mathbb{R}^{n \times K}$ *has independent and identically distributed rows* $[\boldsymbol{\Xi}]_{1*}, \ldots, [\boldsymbol{\Xi}]_{n*} \in \mathbb{R}^K$ *that are independent of* $\boldsymbol{E}$

such that $\mathbb{E}\left([\boldsymbol{\Xi}]_{i*}\right) = \mathbf{0}$, $\mathbb{E}\left([\boldsymbol{\Xi}]_{i*}[\boldsymbol{\Xi}]_{i*}^{\mathrm{T}}\right) = I_K$ *and* $\mathbb{E}\left([\boldsymbol{\Xi}]_{ik}^4\right) < \infty$ *for all* $i \in [n]$ *and* $k \in [K]$.

Let $\mathcal{W}_d\left(I_d, K\right)$ *be the standard Wishart distribution in* d *dimensions with* K *degrees of freedom. If the null hypothesis* $A = 0$ *is true and* $n \to \infty$, *then*

$$\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{1/2}\hat{\boldsymbol{\Omega}}\hat{\boldsymbol{\Omega}}^{\mathrm{T}}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{1/2} \overset{\mathcal{D}}{=} \boldsymbol{V} + o_{\mathrm{p}}(1),$$

where $\hat{\boldsymbol{\Omega}}$ *is defined in* (2.10) *and* $\boldsymbol{V} \sim \mathcal{W}_d\left(I_d, K\right)$. *If* $d = 1$, $\boldsymbol{V} = V \sim \chi_K^2$.

***Remark* 2.8.** *Under the definition of* $\boldsymbol{C}$ *in Item* (iv), $\boldsymbol{\Omega} = \boldsymbol{A} + \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Xi}$ *and* $\mathbb{E}\left(\boldsymbol{\Omega}\right) = \boldsymbol{A}$.

2.4 Simulations and data analysis

2.4.1 Simulation study

In this section, we use simulations to compare the performance of our shrinkage-corrected method defined by Algorithm 2.2 to that of methods proposed in Wang et al. (2017), Lee et al. (2017), Leek & Storey (2008), Gagnon-Bartsch & Speed (2012), Gagnon-Bartsch et al. (2013), as well as the ordinary least squares estimator when $\boldsymbol{C}$ is known and when it is ignored. We do not include results from Fan & Han (2017) or Houseman et al. (2014), because these methods perform similarly to those proposed in Wang et al. (2017) and Lee et al. (2017), respectively. In all of our simulations, we set $n = 100, p = 10^5$ and $K = 10$ to mimic DNA methylation data where p ranges from 3×10^5 to 8×10^5, although our results are nearly identical for p on the order of gene expression data ($p \approx 10^4$). We set $d = 1$ and assigned 50 samples to the treatment group and the rest to the control group so that $\boldsymbol{X} = \left(\mathbf{1}_{n/2}^{\mathrm{T}}, \mathbf{0}_{n/2}^{\mathrm{T}}\right)^{\mathrm{T}} \in \mathbb{R}^n$. We then set the eigenvalues $\lambda_1, \ldots, \lambda_K$ so that $\lambda_1 = n - 2$,

$\lambda_K = 1$ and the remaining to be

$$\lambda_k = \begin{cases} (n-2)^{(K-k)/(K-1)}, & k \leq K/2 \\ \{(n-2)/5\}^{(K-k)/(K-1)}, & k > K/2 \end{cases}.$$

For a predefined value of $\boldsymbol{A} \in \mathbb{R}^{1\times K}$, we simulated $\boldsymbol{\beta} = \left(\beta_1 \cdots \beta_p\right)^{\mathrm{T}} \in \mathbb{R}^{p\times 1}, \boldsymbol{L} \in \mathbb{R}^{p\times K}, \boldsymbol{C} \in \mathbb{R}^{n\times K}, \boldsymbol{\Sigma} = \operatorname{diag}\left(\sigma_1^2, \dots, \sigma_p^2\right)$ and $\boldsymbol{E} \in \mathbb{R}^{p\times n}$ according to

$$\begin{aligned} \beta_g &\sim 0.95\delta_0 + 0.05N\left(0, 0.4^2\right) \quad (g = 1, \dots, p) \\ \tau_k^2 &= \max\left\{\lambda_k/(n-2), 0.5^2\right\} \quad (k = 1, \dots, K) \\ [\boldsymbol{L}]_{gk} &\sim \pi_k\delta_0 + (1-\pi_k)\, N\left(0, \tau_k^2\right) \quad (g = 1, \dots, p;\ k = 1, \dots, K) \\ \boldsymbol{C} &\sim MN_{n\times K}\left(\boldsymbol{X}\boldsymbol{A}, I_n, I_K\right) \\ \sigma_g^2 &\sim \text{Gamma}\left(1/0.5^2, 1/0.5^2\right) \quad (g = 1, \dots, p) \\ \boldsymbol{E}_{gi} &\sim 2^{-1/2}\sigma_g T_4 \quad (g = 1, \dots, p;\ i = 1, \dots, n) \end{aligned} \tag{2.23}$$

where π_k was chosen so that $\mathbb{E}\left([\boldsymbol{L}]_{*k}^{\mathrm{T}} [\boldsymbol{L}]_{*k}\right) = \lambda_k$ and T_4 is the t-distribution with 4 degrees of freedom. We then set the observed data to be $\boldsymbol{Y} = \boldsymbol{\beta}\boldsymbol{X}^{\mathrm{T}} + \boldsymbol{L}\boldsymbol{C}^{\mathrm{T}} + \boldsymbol{E} \in \mathbb{R}^{p\times n}$. Although our theory from Section 2.3 assumes the residuals $\boldsymbol{E}$ are normally distributed, we simulated t-distributed data to mimic real data with heavy tails. The values used for τ_k and π_k $(k = 1, \dots, 10)$ are given in Table 2.1. We also include additional simulation results in Section 2.6 where we simulate $\boldsymbol{\beta}$ according $\beta_g \sim 0.80\delta_0 + 0.20N\left(0, 0.4^2\right)$.

Table 2.1: *The τ_k and π_k values $(k = 1, \dots, 10)$ used to simulate $\boldsymbol{L}$*

Factor number (k)	1	2	3	4	5	6	7	8	9	10
τ_k	1	0.78	0.60	0.5	0.5	0.5	0.5	0.5	0.5	0.5
π_k	0	0	0	0.13	0.48	0.85	0.89	0.92	0.94	0.96
λ_k	98.0	58.9	35.4	21.3	12.8	3.8	2.7	1.9	1.4	1.0

We set the parameter $\boldsymbol{A}$ used to simulate $\boldsymbol{C}$, where $\boldsymbol{A} = \mathbb{E}(\boldsymbol{\Omega})$ in (2.23), to be one of two values:

$$\boldsymbol{A}_1 = \alpha\left(\mathbf{1}_5^{\mathrm{T}}, \mathbf{0}_5^{\mathrm{T}}\right) \qquad \boldsymbol{A}_2 = \alpha\left(\mathbf{0}_5^{\mathrm{T}}, \mathbf{1}_5^{\mathrm{T}}\right)$$

with the scalar α chosen so that $\boldsymbol{C}$ explained 30% of the variability in group status $\boldsymbol{X}$, on average. The choice of 30% was not arbitrary, as we estimated that over 30% of the variance in group status was explained by $\boldsymbol{C}$ in our data application in Section 2.4.2.

As simulated, the eigenvalues $\lambda_1, \ldots, \lambda_5$ are large enough so that the shrinkage terms $\lambda_k / (\lambda_k + \rho)$ $(k = 1, \ldots, 5)$ from (2.19) in Proposition 2.2 are negligible. This implies that when $\boldsymbol{A} = \boldsymbol{A}_1$, $\hat{\boldsymbol{\Omega}}^{\text{shrunk}}$ will likely be a suitable estimate of $\boldsymbol{\Omega} \in \mathbb{R}^{1\times 10}$, since $\hat{\boldsymbol{\Omega}}^{\text{shrunk}}$ will correctly estimate the largest and most important components of $\boldsymbol{\Omega}$, $\boldsymbol{\Omega}_{*1}, \ldots, \boldsymbol{\Omega}_{*5}$. The anti-conservative nature of $\hat{\boldsymbol{\Omega}}^{\text{shrunk}}$ implied by Corollary 2.1 does not apply when $\boldsymbol{A} = \boldsymbol{A}_1$ because Item (ii) in the statement of Corollary 2.1 will generally not hold. We would therefore expect our shrinkage-corrected method defined by Algorithm 2.2 to perform similarly to previous methods that ignore the shrinkage in their estimates of $\boldsymbol{\Omega}$ in this simulation scenario. However, when $\boldsymbol{A} = \boldsymbol{A}_2$, $\hat{\boldsymbol{\Omega}}^{\text{shrunk}}$ will not recover the largest and most consequential components of $\boldsymbol{\Omega}$, $[\boldsymbol{\Omega}]_{*6}, \ldots, [\boldsymbol{\Omega}]_{*10}$, because of the substantial shrinkage caused by the relatively small eigenvalues $\lambda_6, \ldots, \lambda_{10}$. In this case, Corollary 2.1 and Remark 2.7 suggest that ignoring the shrinkage will lead to anti-conservative inference on $\beta_1, \ldots, \beta_p$, whereas Theorems 2.1 and 2.2 imply our shrinkage-corrected method will be asymptotically equivalent to ordinary least squares when $\boldsymbol{C}$ is observed.

We simulated 100 datasets with $\boldsymbol{A} = \boldsymbol{A}_1$ and another 100 with $\boldsymbol{A} = \boldsymbol{A}_2$. We found that we could do the best inference on $\beta_1, \ldots, \beta_p$ with each method by performing ordinary least squares with the design matrix $\left(\mathbf{1}_n \, \boldsymbol{X} \, \hat{\boldsymbol{C}}\right)$, where $\hat{\boldsymbol{C}}$ was $\boldsymbol{C}$ if $\boldsymbol{C}$ was known, or was estimated with any one of the six methods described above. Our shrinkage-corrected estimate of $\boldsymbol{C}$ was $\hat{\boldsymbol{C}}_2 + \boldsymbol{X}\hat{\boldsymbol{\Omega}}$, where $\hat{\boldsymbol{C}}_2$ was defined in Step (a) of Algorithm 2.2. We describe how the other five methods estimate $\boldsymbol{C}$ below. We compared the ordinary least squares t-statistics from all methods to a t-distribution with $n-2-K$ degrees of freedom to compute P values for the null hypotheses $H_{0,g} : \beta_g = 0$ $(g = 1, \ldots, p)$. We then judged the performance of each method by comparing their true false discovery proportion at a nominal 20% false discovery rate, estimated using q-values (Storey 2001), because this is the inference method popular among biologists.

Figure 2.1 provides the simulation results. We see that our shrinkage-corrected method, BC, is able to control the false discovery rate both when K is known to be 10 and when we drastically overestimate it to be 20. Further, our method's power to detect units with non-zero β_g at this nominal 20% false discovery rate threshold was 13.6% when $\hat{K} = 10$ and 12.8% when $\hat{K} = 20$, which is compared to 13.6% when $\boldsymbol{C}$ was known. The power of all three methods was the same for both values of $\boldsymbol{A}$. This is exactly what one would expect from Theorems 2.1 and 2.2, which prove that inference with our shrinkage-corrected estimator is asymptotically equivalent to that with ordinary least squares when $\boldsymbol{C}$ is known. This equivalence was also manifested when we overtly violated Assumption 2.3(a) and simulated $\beta_1, \ldots, \beta_p \sim 0.80\delta_0 + 0.20N\left(0, 0.4^2\right)$ in Section 2.6.

It is also informative to study the performance of the other five methods, as this can be important to practitioners deciding which method to apply to their data.

CATE-RR (Wang et al. 2017) *and dSVA* (Lee et al. 2017): These methods estimate $\boldsymbol{C}$ as $\hat{\boldsymbol{C}}_2^{\text{cate}} + \boldsymbol{X}\hat{\boldsymbol{\Omega}}^{\text{cate}}$ and $\hat{\boldsymbol{C}}_2^{\text{dSVA}} + \boldsymbol{X}\hat{\boldsymbol{\Omega}}^{\text{dSVA}}$, respectively, where their estimates of $\boldsymbol{C}_2$, $\hat{\boldsymbol{C}}_2^{\text{cate}}$ and $\hat{\boldsymbol{C}}_2^{\text{dSVA}}$, are nearly identical to $\hat{\boldsymbol{C}}_2$ defined in Step (a) of Algorithm 2.2. However, their estimates of $\boldsymbol{\Omega}$, $\hat{\boldsymbol{\Omega}}^{\text{cate}}$ and $\hat{\boldsymbol{\Omega}}^{\text{dSVA}}$, ignore the shrinkage described in Proposition 2.2. We would therefore expect them to introduce more type I errors when $\boldsymbol{A} = \boldsymbol{A}_2$. Both CATE-RR and dSVA's false discovery proportion estimates were closer to nominal values when $\beta_1, \ldots, \beta_p \sim 0.80\delta_0 + 0.20N\left(0, 0.4^2\right)$, since any rejection region was likely to have more genomic units with non-zero coefficients of interest.

IRW-SVA (Leek & Storey 2007): IRW-SVA estimates $\boldsymbol{C}$ by performing a factor analysis on

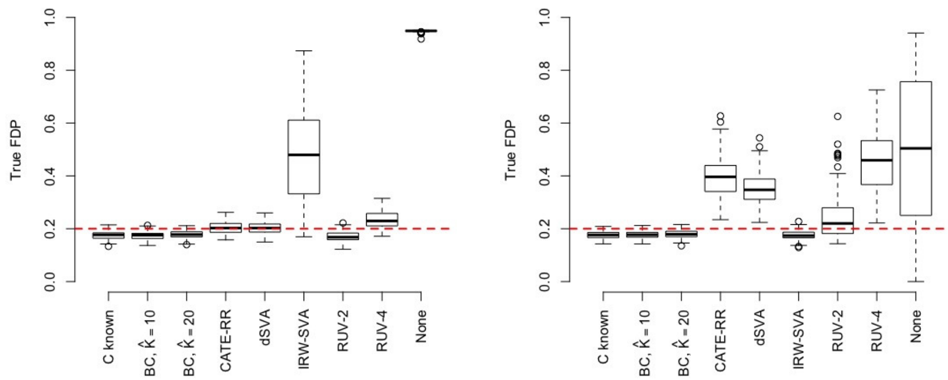


Figure 2.1: The false discovery proportion (FDP) for each method at a q-value threshold of 0.2 in simulations when $\boldsymbol{A} = \boldsymbol{A}_1$ (left) and $\boldsymbol{A} = \boldsymbol{A}_2$ (right). "BC" is our shrinkage-corrected method enumerated in Algorithm 2.2 and $\hat{K}$ is the number of factors used to estimate $\boldsymbol{C}$. CATE-RR, dSVA, IRW-SVA, RUV-2 and RUV-4 are the methods proposed in Wang et al. (2017), Lee et al. (2017), Leek & Storey (2007), Gagnon-Bartsch & Speed (2012) and Gagnon-Bartsch et al. (2013), respectively. These five methods were all applied with $\hat{K} = K = 10$. Inference with "None" was performed using the design matrix $(\boldsymbol{1}_n \, \boldsymbol{X})$.

diag $\left(\hat{\pi}_1, \ldots, \hat{\pi}_p\right) \boldsymbol{Y}$, where $\hat{\pi}_g$ is an estimate of $\mathbb{P}\left([\boldsymbol{L}]_{g*} \neq \boldsymbol{0}, \beta_g = 0 \mid \boldsymbol{Y}\right)$ $(g = 1, \ldots, p)$, by iteratively estimating $\boldsymbol{C}$ and $\mathbb{P}\left([\boldsymbol{L}]_{g*} \neq \boldsymbol{0}, \beta_g = 0 \mid \boldsymbol{Y}\right)$. Since the first iteration assumes $\hat{\boldsymbol{C}} = P_X^{\perp} \hat{\boldsymbol{C}}$, $\hat{\pi}_g$ tends to be small if the marginal correlation between $[\boldsymbol{Y}]_{g*}$ and $\boldsymbol{X}$ is large, which occurs if $|\boldsymbol{\Omega}\, [\boldsymbol{L}]_{g*}|$ is large. Therefore, the latent factors that influence diag $\left(\hat{\pi}_1, \ldots, \hat{\pi}_p\right) \boldsymbol{Y}$ will be different than those of $\boldsymbol{Y}$ if the latent factors with the largest effects are also correlated with $\boldsymbol{X}$. This explains why IRW-SVA performs poorly when $\boldsymbol{A} = \boldsymbol{A}_1$. Unfortunately, there is no theory that states when IRW-SVA is expected to accurately recover $\boldsymbol{C}$.

RUV-2 (Gagnon-Bartsch & Speed 2012) *and RUV-4* (Gagnon-Bartsch et al. 2013): Both RUV-2 and RUV-4 assume the practitioner has prior knowledge of a subset $\mathcal{S} \subseteq [p]$ of control genomic units where $\beta_g = 0$ for all $g \in \mathcal{S}$. We selected $|\mathcal{S}| = 600 = 20 \times 30$ control units uniformly at random from the set of all genomic units with $\beta_g = 0$ across all simulations, because simulations in Wang et al. (2017) use 30 control units when $p = 5{,}000 = 10^5/20$. RUV-2 estimates $\boldsymbol{C}$ via factor analysis using only data from genomic units in $\mathcal{S}$, whereas

RUV-4 first estimates $\boldsymbol{C}_2$ and $\boldsymbol{L}$ as $\hat{\boldsymbol{C}}_2$ and $\hat{\boldsymbol{L}}$ defined in Step (a) of Algorithm 2.2, and then estimates $\boldsymbol{\Omega}$ as $\hat{\boldsymbol{\Omega}}^{\text{RUV-4}} = [\boldsymbol{Y}_1]_{\mathcal{S}*}^{\text{T}} \left[\hat{\boldsymbol{L}}\right]_{\mathcal{S}*} \left(\left[\hat{\boldsymbol{L}}\right]_{\mathcal{S}*}^{\text{T}} \left[\hat{\boldsymbol{L}}\right]_{\mathcal{S}*}\right)^{-1}$. Here $[\boldsymbol{Y}_1]_{\mathcal{S}*}$ and $\left[\hat{\boldsymbol{L}}\right]_{\mathcal{S}*}$ are the submatrices of $\boldsymbol{Y}_1$ and $\hat{\boldsymbol{L}}$ restricted to the rows in $\mathcal{S}$. RUV-4's estimate of $\boldsymbol{C}$ is then $\hat{\boldsymbol{C}}_2 + \boldsymbol{X}\hat{\boldsymbol{\Omega}}^{\text{RUV-4}}$. The obvious caveat of RUV-2 and RUV-4 is the practitioner must have a list of units whose coefficients of interest are zero and whose expression or methylation carries the latent factor signature, i.e. the first K eigenvalues of $\boldsymbol{C}\left(|\mathcal{S}|^{-1}\boldsymbol{L}_{\mathcal{S}*}^{\text{T}}\boldsymbol{L}_{\mathcal{S}*}\right)\boldsymbol{C}^{\text{T}}$ must be suitably large. For example, the large variability in RUV-2's false discovery proportion when $\boldsymbol{A} = \boldsymbol{A}_2$ is because the $|\mathcal{S}| = 600$ control units were not sufficient to capture the latent factor signature in many simulations.

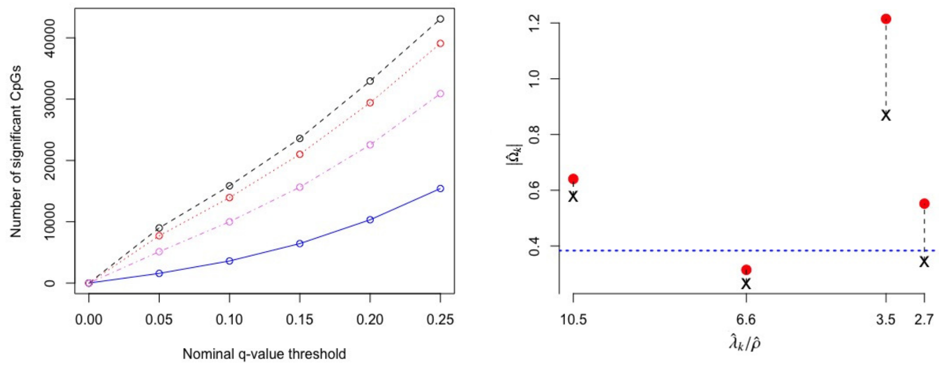
2.4.2 Data application

In order to demonstrate the importance of using our shrinkage-corrected estimator, we applied our method to re-analyze data from Nicodemus-Johnson et al. (2016), which studied the correlation between adult asthma and DNA methylation in lung epithelial cells. The authors collected endobronchial brushings from 74 adult patients with a current doctor's diagnosis of asthma and 41 healthy adults and quantified their DNA methylation at $p = 327,271$ methylation sites, also referred to as CpGs, using the Infinium Human Methylation 450K Bead Chip (Dedeurwaerder et al. 2011). The authors then used ordinary least squares to regress the methylation at each of the p sites on to the mean model subspace that included asthma status, age, ethnicity, sex and smoking status to estimate the effect due to asthma, $\left(\beta_1 \cdots \beta_p\right)^{\text{T}} \in \mathbb{R}^{p\times 1}$. They found 40,892 CpGs that were differentially methylated between asthmatics and healthy patients at a nominal false discovery rate of 5%.

We investigated whether or not the strong association between DNA methylation and asthma status was in part due to unobserved covariates. In particular, lung cell composition may differ between asthmatics and non-asthmatics, with asthmatic patients generally having a greater proportion of airway goblet cells that excrete mucus (Bai & Knight 2005, Rogers 2002). We therefore re-analyzed these data to account for latent covariates with our shrinkage-corrected method defined by Algorithm 2.2, and compared the results to those obtained using the methods proposed in Leek & Storey (2007), Wang et al. (2017) and Lee et al. (2017). We could not apply the methods proposed in Gagnon-Bartsch & Speed (2012), Gagnon-Bartsch et al. (2013) because we did not have access to control CpGs. We first used bi-cross validation (Owen & Wang 2016) to estimate that there were $K = 4$ latent factors in these data, and subsequently estimated $\boldsymbol{C} \in \mathbb{R}^{115\times 4}$ using the four different methods. We then computed P values for the null hypotheses $H_{0,g} : \beta_g = 0$ $(g = 1, \dots, p)$ using ordinary least squares with the design matrix $\left(\boldsymbol{X}\,\boldsymbol{Z}\,\hat{\boldsymbol{C}}\right)$, where $\boldsymbol{X} \in \{0, 1\}^n$ was asthma status and $\boldsymbol{Z}$ contained the observed nuisance covariates age, ethnicity, sex and smoking status. The total number of asthma-related CpGs returned by each method as a function of q-value cutoffs (Storey et al. 2015), as well the uncorrected and shrinkage-corrected estimates of $\boldsymbol{\Omega} \in \mathbb{R}^{1\times 4}$, are given in Fig. 2.2. At a q-value threshold of 20%, our method identifies 10,324 asthma-related CpGs, while the methods proposed in Leek & Storey (2007), Wang et al. (2017) and Lee et al. (2017) ostensibly identify 32,952, 29,415 and 22,545 asthma-related CpGs, respectively. These numbers changed only slightly when we let K be as high as seven.

We estimated that approximately 36% of the variance in asthma status was explained by $\boldsymbol{C}$, which using Theorem 2.3 corresponded to a P value for the null hypothesis $\mathbb{E}\left(\boldsymbol{C} \mid \boldsymbol{X}\right) = \boldsymbol{0}$ of 3.2×10^{-12}. Moreover, assuming $(\boldsymbol{L}, \boldsymbol{C}) \in \Theta_{(1)}$, the largest component of $\boldsymbol{\Omega}$ appeared to load on to the third column of $\boldsymbol{L} \in \mathbb{R}^{p\times 4}$, where $\lambda_3/\rho \approx 2.5$. Since this was much smaller than $n^{1/2} = 10.7$ and we estimated $[\boldsymbol{L}]_{g3} \neq 0$ at over 40% of the studied CpGs $g \in [p]$, Proposition 2.2, Corollary 2.1 and simulations connote that the methods proposed in Wang et al. (2017) and Lee et al. (2017) are likely underestimating the fraction of CpGs with $\beta_g = 0$ at any nominal q-value threshold. It is likely the case that λ_3, the third largest eigenvalue of $\boldsymbol{\mathcal{I}} = P_X^{\perp}\boldsymbol{C}\left(p^{-1}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\right)\boldsymbol{C}^{\mathrm{T}}P_X^{\perp}$, was small even though the third factor explained a significant portion of the variability in methylation levels because its strong correlation with asthma status dampened $P_X^{\perp}\boldsymbol{C}$.

We next sought to determine if differences in lung cell composition between asthmatic



(a) (b)

Figure 2.2: Results from our analysis of lung DNA methylation data from Nicodemus-Johnson et al. (2016). (a): The number of asthma-related CpGs at a given q-value cutoff using our shrinkage-corrected estimator (blue, solid line), as well as the estimators proposed in Lee et al. (2017) (pink, dot-dashed line), Wang et al. (2017) (red, dotted line) and Leek & Storey (2007) (black, dashed line). (b): The $K = 4$ components of $\hat{\mathbf{\Omega}}^{\text{shrunk}}$ (black "x") and $\hat{\mathbf{\Omega}}$ (red "•") as a function of $\hat{\lambda}_1, \dots, \hat{\lambda}_4$. The dashed blue line is the 0.95 quantile of the $\tilde{n}^{-1/2}\chi_1$ distribution, where $\tilde{n}$ is defined such that $\tilde{n}\hat{\mathbf{\Omega}}\hat{\mathbf{\Omega}}^{\mathrm{T}}$ converges to a chi-squared random variable with $K = 4$ degrees of freedom under the null hypothesis from Theorem 2.3.

and healthy patients were responsible for some of the correlation between asthma status and the latent factors, since understanding the origin of the latent covariates could help practitioners determine which method is most appropriate for their data. To do so, we fit a topic model with $r = 7$ topics on the same individuals' gene expression data, which has been shown to cluster bulk RNA-seq samples by tissue and cell type (Taddy 2012, Dey et al. 2017). We then used the n-dimensional factor whose corresponding loading was the largest on the *MUC5AC* gene as a proxy for the proportion of goblet cells in each sample, as *MUC5AC* is a unique identifier for goblet cells (Zuhdi Alimam et al. 2000). Just as one would expect, asthmatics tended to have a higher proportion of estimated goblet cells than healthy controls, and we rejected the null hypothesis that asthmatics and healthy controls had the same mean estimated goblet cell proportion at the significance level of $\alpha = 0.01$. This indicates that cell composition is presumably driving much of the observed correlation between methylation levels and asthma status in Nicodemus-Johnson et al. (2016), as well as the results from the re-analysis with the methods proposed in Wang et al. (2017) and Lee et al. (2017).

These conclusions also help to explain why the method proposed in Leek & Storey (2007) is likely underestimating the number of false discoveries. We used Lemma 2.2 and Remark 2.3 to estimate that

$$p^{-1}\left\{I\left([\boldsymbol{L}]_{1*} \neq 0\right) + \cdots + I\left([\boldsymbol{L}]_{p*} \neq 0\right)\right\} \approx 0.90$$

in these data, which is precisely what one would expect if cellular heterogeneity were among the unobserved factors, since changes in methylation help drive cellular differentiation. And since we have already shown that $\boldsymbol{X}$ is correlated with $\boldsymbol{C}$, the method proposed in Leek & Storey (2007) would not be expected to control the false discovery rate, as simulations in Section 2.4.1 showed exactly this when $|\boldsymbol{\Omega}\,[\boldsymbol{L}]_{g*}|$ was large for many genomic units $g \in [p]$.

2.5 Discussion

The prevalence of unobserved covariates in high throughput 'omic' data has precipitated the development of methods that account for unobserved factors $\boldsymbol{C}$ in downstream inference. While these methods perform well when the data are strongly informative for $\boldsymbol{C}$, they are not amenable to inference when the data are only moderately informative for $\boldsymbol{C}$. On the other hand, we prove that inference using estimates from our shrinkage-corrected method in Algorithm 2.2 is asymptotically equivalent to ordinary least squares when $\boldsymbol{C}$ is observed.

Our method is not a cure-all for inference with unobserved covariates. For example, Assumption 2.3(a) restricts the number of units with non-zero main effect in DNA methylation data to be on the order of hundreds to thousands when the data are only moderately informative $\boldsymbol{C}$. Even though simulations show we can potentially relax this number substantially to tens, or even hundreds of thousands in practice, it begs the question as to whether or not practitioners should spend time and money to measure nuisance variables like cellular heterogeneity, or estimate them directly from the data. If the practitioner is concerned that $\boldsymbol{C}$ is correlated with $\boldsymbol{X}$ but has reason to believe $\boldsymbol{\beta}$ is sparse, our theory suggests the effort should be spent collecting more samples. However, if $\boldsymbol{C}$ is correlated with $\boldsymbol{X}$ and $\boldsymbol{\beta}$ is dense, it may be worthwhile to attempt to measure some of the latent factors with other technologies. We are currently working with the authors of Nicodemus-Johnson et al. (2016) to use external sources of information to potentially better account for cellular heterogeneity in their data.

2.6 Additional simulations

We first include an empirical verification of the surprising results from Proposition 2.2 and Lemma 2.2 that we can accurately estimate $\boldsymbol{L}$, but the naive estimate for $\boldsymbol{\Omega}$ in (2.12) is biased. The data were simulated as follows:

$$
\begin{aligned}
& d = K = 1,\ n = 100,\ p = 10^5 \\
& \boldsymbol{X} \sim N_n\left(0, I_n\right) \\
& \boldsymbol{\beta} = \boldsymbol{0} \\
& [\boldsymbol{L}]_g \sim \left(1 - n^{-1}\right)\delta_0 + n^{-1} N_1(0,1) \ \ (g = 1, \ldots, p) \\
& \boldsymbol{X} = \left(\boldsymbol{1}_{n/2}^{\mathrm{T}}, \boldsymbol{0}_{n/2}^{\mathrm{T}}\right)^{\mathrm{T}} \\
& \boldsymbol{C}_2 \sim N_{n-d}\left(0, I_{n-d}\right) \\
& \boldsymbol{C} = \boldsymbol{X} + \boldsymbol{Q}_X \boldsymbol{C}_2 \\
& \boldsymbol{E} \sim MN_{p\times n}\left(0, I_p, I_n\right).
\end{aligned}
$$

where $\boldsymbol{Q}_X \in \mathbb{R}^{n\times(n-d)}$ is a matrix whose columns form an orthonormal basis for the null sapce of $\boldsymbol{X}$. Therefore, $\Omega = 1$ for every simulation. $\boldsymbol{L}$ and $\boldsymbol{\Sigma}$ were simulated so that $\lambda_1 \approx 1$ and $\rho = 1$. If (2.18) from Lemma 2.2 was correct, then $n^{1/2}\left(\left[\hat{\boldsymbol{L}}\right]_g - [\boldsymbol{L}]_g\right) \approx N(0,1)$ for each $g \in [p]$. If we let $\boldsymbol{W} = n^{1/2}\left(\hat{\boldsymbol{L}} - \boldsymbol{L}\right) \in \mathbb{R}^p$, we validate Lemma 2.2 by partitioning the components of $\boldsymbol{W}$ by whether or not the corresponding component of $\boldsymbol{L}$ was non-zero. If Lemma 2.2 were true, both histograms in Fig. 2.3 should look as if they were sampled from a $N(0,1)$ random variable, which they clearly do.

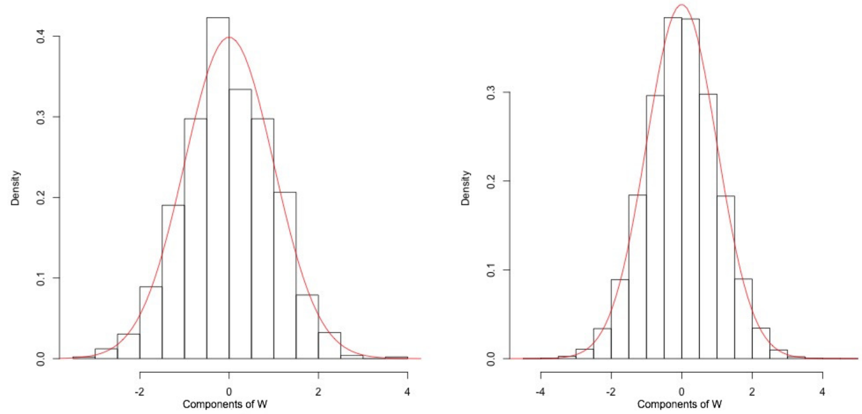


Figure 2.3: $\boldsymbol{W} = n^{1/2}\left(\hat{\boldsymbol{L}} - \boldsymbol{L}\right) \in \mathbb{R}^p$ for one simulation for components of $\boldsymbol{L}$ that are non-zero (left) or 0 (right). The overlaid red curve is the density of a $N(0,1)$ random variable.

We next empirically verified (2.19) from Proposition 2.2 using 20 simulations. Figure 2.4 contains the results of the 20 simulations, which clearly shows that $n^{1/2}\|\hat{\Omega}^{\text{shrunk}} - \Omega\lambda_1(\lambda_1 + \rho)^{-1}\|_2 \approx 0$.

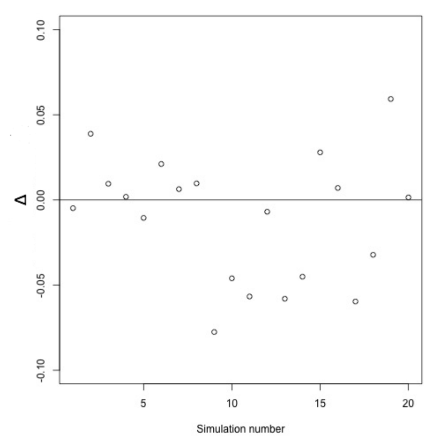


Figure 2.4: $\Delta = n^{1/2}\left\{\hat{\Omega}^{\text{shrunk}} - \Omega\lambda_1(\lambda_1 + \rho)^{-1}\right\}$ for 20 simulations.

Lastly, Fig. 2.5 gives the simulation results from Section 4.1, with $\beta_g \sim 0.80\delta_0 + 0.20N\left(0, 0.4^2\right)$. The average power for the simulations on the left panel for $\boldsymbol{C}$ known, BC

$\hat{K} = 10$, BC $\hat{K} = 20$ was $23.3\%, 23.3\%, 22.1\%$ and $23.0\%, 23.0\%, 21.9\%$ for the simulations on the right panel.

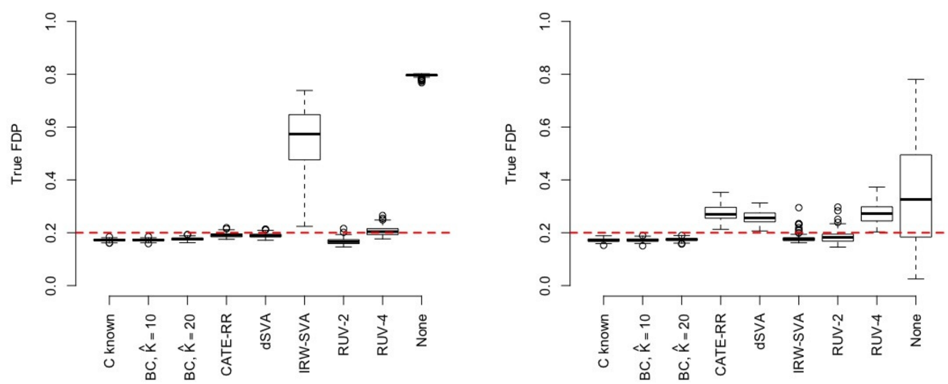


Figure 2.5: Simulations with $\beta_g \sim 0.80\delta_0 + 0.20N\left(0, 0.4^2\right)$ for $\boldsymbol{A} = \boldsymbol{A}_1$ (left) and $\boldsymbol{A} = \boldsymbol{A}_2$ (right). All other parameters are the same as the simulations in Section 4.1.

2.7 Proofs of all propositions, lemmas, and theorems

2.7.1 Proof of Proposition 2.1 and the identifiability of $\boldsymbol{\Omega}\left(n^{-1}\boldsymbol{C}_2^T\boldsymbol{C}_2\right)^{-1}\boldsymbol{\Omega}^T$

For the remainder of the chapter, we define $[\boldsymbol{L}]_{g*} = \boldsymbol{\ell}_g$ for all $g \in [p]$. Let $\boldsymbol{X}$ be a matrix, vector or scalar and let $\|\boldsymbol{X}\|_2$ be the spectral norm, Euclidean norm or magnitude of $\boldsymbol{X}$. We use the notation that $\boldsymbol{X} = O_{\mathrm{p}}(a_n)$ if for some sequence a_n, $\|\boldsymbol{X}\|_2/a_n = O_{\mathrm{p}}(1)$. Similarly, $\boldsymbol{X} = o_{\mathrm{p}}(a_n)$ if $\|\boldsymbol{X}\|_2/a_n = o_{\mathrm{p}}(1)$. Lastly, for any vector $\boldsymbol{v} \in \mathbb{R}^m$, we define $\boldsymbol{v}_j$ to be the jth element of $\boldsymbol{v}$ for all $j = 1, \ldots, m$. If the vector has a subscript r, then we define $\boldsymbol{v}_{rj}$ to be the jth elements of $\boldsymbol{v}_r \in \mathbb{R}^m$ for all $j = 1, \ldots, m$.

We first prove Proposition 2.1.

Proof of Proposition 2.1. Under Assumptions 2.1 and 2.2, we can find an $\boldsymbol{L} \in \mathbb{R}^{p \times K}$, $\boldsymbol{C} \in \mathbb{R}^{n \times K}$ such that for $\boldsymbol{C}_2 = P_X^{\perp}\boldsymbol{C}$, $\mathbb{E}(\boldsymbol{Y}_2) = \boldsymbol{L}\boldsymbol{C}_2^{\mathrm{T}}$ and

$$np^{-1}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L} = \operatorname{diag}(\lambda_1, \ldots, \lambda_K), \quad n^{-1}\boldsymbol{C}_2^{\mathrm{T}}\boldsymbol{C}_2 = I_K \tag{2.24}$$

by taking the singular value decomposition of $\mathbb{E}(\boldsymbol{Y}_2)$. The columns of $\boldsymbol{L}$ and $\boldsymbol{C}_2$ are unique up to sign by the uniqueness of the singular value decomposition, since $\lambda_k > \lambda_{k+1}$ for all $k = 1, \ldots, K$ (where $\lambda_{K+1} = 0$). That is, if $\tilde{\boldsymbol{L}}$ and $\tilde{\boldsymbol{C}}$ also satisfy (2.24), then $\tilde{\boldsymbol{L}} = \boldsymbol{L}\boldsymbol{\Pi}$ and $\tilde{\boldsymbol{C}}_2 = \boldsymbol{C}_2\boldsymbol{\Pi}$ where $\boldsymbol{\Pi} = \operatorname{diag}(a_1, \ldots, a_K)$ and $a_k \in \{-1, 1\}$ for all $k \in [K]$.

Next, suppose Assumption 2.3(a) holds and let $\boldsymbol{\beta}^{(a)}, \boldsymbol{L}^{(a)}, \boldsymbol{C}^{(a)}$ and $\boldsymbol{\beta}^{(b)}, \boldsymbol{L}^{(b)}, \boldsymbol{C}^{(b)}$ be such that $\boldsymbol{L}^{(a)} \left\{\boldsymbol{C}_2^{(a)}\right\}^{\mathrm{T}} = \mathbb{E}(\boldsymbol{Y}_2) = \boldsymbol{L}^{(b)} \left\{\boldsymbol{C}_2^{(b)}\right\}^{\mathrm{T}}$ and

$$\boldsymbol{\beta}^{(a)} + \boldsymbol{L}^{(a)} \left\{\boldsymbol{\Omega}^{(a)}\right\}^{\mathrm{T}} = \mathbb{E}(\boldsymbol{Y}_1) = \boldsymbol{\beta}^{(b)} + \boldsymbol{L}^{(b)} \left\{\boldsymbol{\Omega}^{(b)}\right\}^{\mathrm{T}}.$$

We can find invertible matrices $\boldsymbol{R}^{(a)}, \boldsymbol{R}^{(b)} \in \mathbb{R}^{K \times K}$ such that $\boldsymbol{L}^{(a)}\boldsymbol{R}^{(a)}, \boldsymbol{C}^{(a)} \left\{\boldsymbol{R}^{(a)}\right\}^{-\mathrm{T}}$ and $\boldsymbol{L}^{(b)}\boldsymbol{R}^{(b)}, \boldsymbol{C}^{(b)} \left\{\boldsymbol{R}^{(b)}\right\}^{-\mathrm{T}}$ satisfy (2.24), where

$$\boldsymbol{L}^{(i)} \left\{\boldsymbol{\Omega}^{(i)}\right\}^{\mathrm{T}} = \left\{\boldsymbol{L}^{(i)}\boldsymbol{R}^{(i)}\right\} \left[\boldsymbol{\Omega}^{(i)} \left\{\boldsymbol{R}^{(i)}\right\}^{-\mathrm{T}}\right]^{\mathrm{T}} \quad (i = a, b).$$

Therefore, to prove the identifiability of $\boldsymbol{\beta}$, it suffices to assume $\boldsymbol{L}^{(a)}, \boldsymbol{C}^{(a)}$ and $\boldsymbol{L}^{(b)}, \boldsymbol{C}^{(b)}$ satisfy (2.24), meaning $\boldsymbol{L}^{(b)} = \boldsymbol{L}^{(a)}\boldsymbol{\Pi}$ and $\boldsymbol{C}_2^{(b)} = \boldsymbol{C}_2^{(a)}\boldsymbol{\Pi}$ for some $\boldsymbol{\Pi} = \operatorname{diag}(a_1, \ldots, a_K)$ where $a_k \in \{-1, 1\}$ for all $k \in [K]$. Define $\boldsymbol{L} = \boldsymbol{L}^{(a)}$ (for notational convenience), $\mathcal{S} = \left\{g \in [p] : \left[\boldsymbol{\beta}^{(a)}\right]_{g*} = \left[\boldsymbol{\beta}^{(b)}\right]_{g*} = 0\right\}$ and $\boldsymbol{L}_{\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}| \times K}$ to be the submatrix of $\boldsymbol{L}$ restricted to the rows $g \in \mathcal{S}$. To prove that $\boldsymbol{\beta}^{(a)} = \boldsymbol{\beta}^{(b)}$ and $\boldsymbol{C}^{(b)} = \boldsymbol{C}^{(a)}\boldsymbol{\Pi}$, it suffices to show that $\boldsymbol{L}_{\mathcal{S}}^{\mathrm{T}}\boldsymbol{L}_{\mathcal{S}} \succ 0$, i.e. $\boldsymbol{L}_{\mathcal{S}}$ has full column rank.

$$n(p\lambda_K)^{-1} \boldsymbol{L}_{\mathcal{S}}^{\mathrm{T}}\boldsymbol{L}_{\mathcal{S}} = \operatorname{diag}\left(\lambda_1\lambda_K^{-1}, \cdots, \lambda_{K-1}\lambda_K^{-1}, 1\right) - n(p\lambda_K)^{-1} \sum_{g \in [p] \setminus \mathcal{S}} \boldsymbol{\ell}_g \boldsymbol{\ell}_g^{\mathrm{T}}$$

where for any $r, s \in [K]$,

$$
\begin{aligned}
|n\left(p\lambda_K\right)^{-1} \sum_{g\in[p]\setminus\mathcal{S}} \boldsymbol{\ell}_{gr}\boldsymbol{\ell}_{gs}| \leq & n\left(p\lambda_K\right)^{-1} \sum_{g\in[p]\setminus\mathcal{S}} |\boldsymbol{\ell}_{gr}||\boldsymbol{\ell}_{gs}| \\
& \leq c_2^2 n\lambda_K^{-1} \left[p^{-1}\sum_{g=1}^{p} I\left\{\left[\boldsymbol{\beta}^{(a)}\right]_{g*} \neq 0\right\}\right. \\
& \left. +p^{-1}\sum_{g=1}^{p} I\left\{\left[\boldsymbol{\beta}^{(b)}\right]_{g*} \neq 0\right\}\right] = o\left(n^{-1/2}\right),
\end{aligned}
$$

which completes the proof. $\square$

***Proposition* 2.3.** *Suppose Assumptions 2.1, 2.2 and 2.3(a) hold. Then for c_4 is defined in the statement of Proposition 2.1, $\boldsymbol{\Omega}\left(n^{-1}\boldsymbol{C}_2^{\mathrm{T}}\boldsymbol{C}_2\right)^{-1}\boldsymbol{\Omega}^{\mathrm{T}}$ is identifiable for all $n \geq c_4$.*

Proof. Under these assumptions, Proposition 2.1 proves that $\boldsymbol{\beta}$ is identifiable for all $n \geq c_4$, meaning $\mathbb{E}(\boldsymbol{Y}) - \boldsymbol{\beta}\boldsymbol{X}^{\mathrm{T}} = \boldsymbol{L}\boldsymbol{C}^{\mathrm{T}}$ is identifiable for all $n \geq c_4$. Suppose $\boldsymbol{C}_{(a)}, \boldsymbol{C}_{(b)} \in \mathbb{R}^{n\times K}$ and $\boldsymbol{L}_{(a)}, \boldsymbol{L}_{(b)} \in \mathbb{R}^{p\times K}$ are such that

$$
\boldsymbol{L}_{(a)}\boldsymbol{C}_{(a)}^{\mathrm{T}} = \mathbb{E}(\boldsymbol{Y}) - \boldsymbol{\beta}\boldsymbol{X}^{\mathrm{T}} = \boldsymbol{L}_{(b)}\boldsymbol{C}_{(b)}^{\mathrm{T}}.
$$

Under Assumptions 2.1 and 2.2, $\boldsymbol{L}_{(a)}$ and $\boldsymbol{L}_{(b)}$ have full column rank, meaning we may define

$$
\boldsymbol{R} = \boldsymbol{L}_{(a)}^{\mathrm{T}}\boldsymbol{L}_{(b)}\left\{\boldsymbol{L}_{(b)}^{\mathrm{T}}\boldsymbol{L}_{(b)}\right\}^{-1},
$$

where $\boldsymbol{C}_{(b)} = \boldsymbol{C}_{(a)}\boldsymbol{R}$. Since $\boldsymbol{C}_{(a)}$ and $\boldsymbol{C}_{(b)}$ have full column rank by Assumptions 2.1 and 2.2, $\boldsymbol{R}$ must be invertible. Therefore, for $\boldsymbol{\Omega}_{(i)} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{C}_{(i)}$ $(i = a, b)$,

$$
\begin{aligned}
\boldsymbol{\Omega}_{(b)}\left\{n^{-1}\boldsymbol{C}_{(b)}^{\mathrm{T}}P_X^{\perp}\boldsymbol{C}_{(b)}\right\}^{-1}\boldsymbol{\Omega}_{(b)}^{\mathrm{T}} &= \boldsymbol{\Omega}_{(a)}\boldsymbol{R}\boldsymbol{R}^{-1}\left\{n^{-1}\boldsymbol{C}_{(a)}^{\mathrm{T}}P_X^{\perp}\boldsymbol{C}_{(a)}\right\}^{-1}\boldsymbol{R}^{-\mathrm{T}}\boldsymbol{R}^{\mathrm{T}}\boldsymbol{\Omega}_{(a)}^{\mathrm{T}} \\
&= \boldsymbol{\Omega}_{(a)}\left\{n^{-1}\boldsymbol{C}_{(a)}^{\mathrm{T}}P_X^{\perp}\boldsymbol{C}_{(a)}\right\}^{-1}\boldsymbol{\Omega}_{(a)}^{\mathrm{T}},
\end{aligned}
$$

which completes the proof. $\square$

2.7.2 The behavior of the off-diagonal elements of $np^{-1}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{L}$

Let $\boldsymbol{m}_k \in \mathbb{R}^p$ be the kth left singular vector of $\boldsymbol{L}\boldsymbol{C}_2^{\mathrm{T}}$ $(k = 1, \ldots, K)$. In this section, we state and prove a proposition regarding the generality of the condition that $(\lambda_r \lambda_s)^{1/2} |\boldsymbol{m}_r^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{m}_s| \leq c_8 \lambda_{\max(r,s)}$ for all $r, s \in [K]$, which is used in the statements of Theorems 2.2 and 2.3. To do so, we note that $(\lambda_r \lambda_s)^{1/2} |\boldsymbol{m}_r^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{m}_s| = |np^{-1} [\boldsymbol{L}]_{*r}^{\mathrm{T}} \boldsymbol{\Sigma} [\boldsymbol{L}]_{*s}|$ for some $\boldsymbol{L}$ such that $(\boldsymbol{L}, \boldsymbol{C}) \in \Theta_{(0)}$.

Proposition 2.4. *Let $\bar{\boldsymbol{L}} = \left[\bar{\boldsymbol{\ell}}_1 \cdots \bar{\boldsymbol{\ell}}_p\right]^{\mathrm{T}} \in \mathbb{R}^{p \times K}$ and $\boldsymbol{\Sigma} = \mathrm{diag}\left(\sigma_1^2, \ldots, \sigma_p^2\right)$ where for each $g \in [p]$,*

$$\bar{\boldsymbol{\ell}}_g \sim F_{\bar{\boldsymbol{\ell}}}$$
$$\sigma_g^2 \sim F_{\sigma^2}$$

where the distributions $F_{\bar{\boldsymbol{\ell}}}$ and F_{σ^2} have compact support. Suppose $\boldsymbol{C} \in \mathbb{R}^{n \times K}$ and $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ are non-random matrices and define $\boldsymbol{R} \in \mathbb{R}^{K \times K}$ such that $\boldsymbol{R}^2 = n^{-1}\boldsymbol{C}^{\mathrm{T}} P_X^{\perp} \boldsymbol{C}$. In addition, let

(a) *$\gamma_K \leq \cdots \leq \gamma_1$ be the eigenvalues of $np^{-1}\bar{\boldsymbol{L}}^{\mathrm{T}}\bar{\boldsymbol{L}}$*

(b) *$\lambda_K \leq \cdots \leq \lambda_1$ be the first K eigenvalues of $P_X^{\perp}\boldsymbol{C}\left(p^{-1}\bar{\boldsymbol{L}}^{\mathrm{T}}\bar{\boldsymbol{L}}\right)\boldsymbol{C}^{\mathrm{T}}P_X^{\perp}$.*

(c) *$\boldsymbol{L} = \bar{\boldsymbol{L}}\boldsymbol{R}\boldsymbol{U}$, where $\boldsymbol{U} \in \mathbb{R}^{K \times K}$ is a unitary matrix such that $np^{-1}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L} = \mathrm{diag}(\lambda_1, \ldots, \lambda_K)$.*

Suppose the following assumptions hold:

(i) *$\|n^{-1}\boldsymbol{C}^{\mathrm{T}}P_X^{\perp}\boldsymbol{C}\|_2, \|\left(n^{-1}\boldsymbol{C}^{\mathrm{T}}P_X^{\perp}\boldsymbol{C}\right)^{-1}\|_2 \leq c^2$ for some constant $c \geq 1$.*

(ii) *For any $\epsilon > 0$, there exists a $\delta_\epsilon > 0$ such that $\mathbb{P}\left(\gamma_K p/n \geq \delta_\epsilon\right) \geq 1 - \epsilon$ for all n, p.*

Then for any $r, s \in [K]$,

$$n\left\{p\lambda_{\max(r,s)}\right\}^{-1} [\boldsymbol{L}]_{*r}^{\mathrm{T}} \boldsymbol{\Sigma} [\boldsymbol{L}]_{*s} = O_{\mathrm{p}}(1)$$

as $n, p \to \infty$.

Proof. First, $n(\gamma_K p)^{-1} = O_{\mathrm{p}}(1)$ by Item (ii). Next, by the sampling mechanism used to draw $\bar{\boldsymbol{L}}$ and $\boldsymbol{\Sigma}$, $\bar{\boldsymbol{L}}$ and $\boldsymbol{\Sigma}$ are independent. Suppose $r \leq s$ and define $\boldsymbol{\ell}_g = [\boldsymbol{L}]_{g*}$ for all $g \in [p]$. First,

$$\begin{aligned}
\mathbb{E}\left[n\left\{p\lambda_{\max(r,s)}\right\}^{-1}[\boldsymbol{L}]_{*r}^{\mathrm{T}}\,\boldsymbol{\Sigma}\,[\boldsymbol{L}]_{*s}\mid\bar{\boldsymbol{L}}\right] &= n\left\{p\lambda_{\max(r,s)}\right\}^{-1}\sum_{g=1}^{p}\boldsymbol{\ell}_{gr}\boldsymbol{\ell}_{gs}\mathbb{E}\left(\sigma_g^2\mid\bar{\boldsymbol{L}}\right)\\
&= \mathbb{E}\left(\sigma_1^2\right)n\left\{p\lambda_{\max(r,s)}\right\}^{-1}\sum_{g=1}^{p}\boldsymbol{\ell}_{gr}\boldsymbol{\ell}_{gs} = 0.
\end{aligned}$$

Next, $\|\boldsymbol{\ell}_g\|_2 = \|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{R}\bar{\boldsymbol{\ell}}_g\|_2 \leq c\|\bar{\boldsymbol{\ell}}_g\|_2$, meaning $\|\boldsymbol{\ell}_g\|_2^2, \sigma_g^2 \leq a$ for all $g \in [p]$ for some constant $a > 0$ not dependent on n or p (since $F_{\bar{\boldsymbol{\ell}}}$ and F_{σ^2} have compact support). Let $\bar{\boldsymbol{\Psi}} = np^{-1}\bar{\boldsymbol{L}}^T\bar{\boldsymbol{L}}$ and $\boldsymbol{\Psi} = np^{-1}\boldsymbol{L}^T\boldsymbol{L}$. Then

$$\lambda_K^{-1} = \|\boldsymbol{\Psi}^{-1}\|_2 \leq c^2\|\bar{\boldsymbol{\Psi}}^{-1}\|_2 = c^2\gamma_K^{-1}$$

and

$$\gamma_K^{-1} = \|\bar{\boldsymbol{\Psi}}^{-1}\|_2 \leq c^2\|\boldsymbol{\Psi}^{-1}\|_2 = c^2\lambda_K^{-1},$$

which implies $c^{-2}\gamma_K \leq \lambda_K \leq c^2\gamma_K$. Therefore, $n(p\lambda_K)^{-1} = O_{\mathrm{p}}(1)$ and

$$\begin{aligned}
\mathbb{V}\left[n\left\{p\lambda_{\max(r,s)}\right\}^{-1}\boldsymbol{L}_{*r}^{\mathrm{T}}\boldsymbol{\Sigma}\,[\boldsymbol{L}]_{*s}\mid\bar{\boldsymbol{L}}\right] &= n^2(p\lambda_s)^{-2}\sum_{g=1}^{p}\boldsymbol{\ell}_{gr}^2\boldsymbol{\ell}_{gs}^2\,\mathbb{V}\left(\sigma_g^2\mid\bar{\boldsymbol{L}}\right)\\
&= \mathbb{V}\left(\sigma_1^2\right)n^2(p\lambda_s)^{-2}\sum_{g=1}^{p}\boldsymbol{\ell}_{gr}^2\boldsymbol{\ell}_{gs}^2\\
&\leq a^2n^2(p\lambda_s)^{-2}\sum_{g=1}^{p}\boldsymbol{\ell}_{gs}^2 = a^2n(p\lambda_s)^{-1}\\
&\leq a^2n(p\lambda_K)^{-1} = O_{\mathrm{p}}(1).
\end{aligned}$$

This proves the claim. $\square$

***Remark* 2.9.** *Item (ii) is a weak assumption because $p\gamma_K/n$ is the smallest eigenvalue of $\bar{\boldsymbol{L}}^{\mathrm{T}}\bar{\boldsymbol{L}}$. Further, we assume $p\lambda_K/n \to \infty$ as $n \to \infty$ (where $\gamma_K \asymp \lambda_K$) in Assumption 2.2.*

***Remark* 2.10.** *We can extend to Proposition 2.4 to include the case that $\boldsymbol{C}$ is random if we assume $\boldsymbol{C}$ is independent of $\boldsymbol{\Sigma}$ and if for all $\epsilon > 0$, there exists an $M > 0$ such that $\mathbb{P}\left(\|n^{-1}\boldsymbol{C}^{\mathrm{T}}P_X^{\perp}\boldsymbol{C}\|_2 \leq M\right), \mathbb{P}\left\{\|\left(n^{-1}\boldsymbol{C}^{\mathrm{T}}P_X^{\perp}\boldsymbol{C}\right)^{-1}\|_2 \leq M\right\} \geq 1-\epsilon$. To prove the claim, we would simply condition on $\bar{\boldsymbol{L}}$ and $\boldsymbol{C}$ instead of just $\bar{\boldsymbol{L}}$.*

2.7.3 A re-definition of $\boldsymbol{C}_2$ and $\boldsymbol{Y}_2$

Let $\boldsymbol{Q}_X \in \mathbb{R}^{n\times(n-d)}$ be a matrix whose columns form an orthonormal basis for the null space of $\boldsymbol{X}^{\mathrm{T}}$. For the remainder of the chapter, we re-define $\boldsymbol{C}_2$ to be

$$\boldsymbol{C}_2 = \boldsymbol{Q}_X^{\mathrm{T}}\boldsymbol{C} \in \mathbb{R}^{(n-d)\times K}. \tag{2.25}$$

Note that since $P_X^{\perp} = \boldsymbol{Q}_X\boldsymbol{Q}_X^{\mathrm{T}}$, $\boldsymbol{C}_2^{\mathrm{T}}\boldsymbol{C}_2 = \boldsymbol{C}^{\mathrm{T}}P_X^{\perp}\boldsymbol{C}$ and the $\boldsymbol{C}_2$ defined in (2.3) is simply $P_X^{\perp}\boldsymbol{C} = \boldsymbol{Q}_X\left(\boldsymbol{Q}_X^{\mathrm{T}}\boldsymbol{C}\right) = \boldsymbol{Q}_X\boldsymbol{C}_2$. This implies that the first $n-d$ singular values and left singular vectors of $\boldsymbol{Y}P_X^{\perp}$ and $\boldsymbol{Y}\boldsymbol{Q}_X$ are the same and if we let $\boldsymbol{V}_1 \in \mathbb{R}^{n\times(n-d)}$ and $\boldsymbol{V}_2 \in \mathbb{R}^{(n-d)\times(n-d)}$ be the right singular vectors of $\boldsymbol{Y}P_X^{\perp}$ and $\boldsymbol{Y}\boldsymbol{Q}_X$ corresponding to the non-zero singular values, then $\boldsymbol{V}_1 = \boldsymbol{Q}_X\boldsymbol{V}_2$. We therefore replace $\boldsymbol{C}_2$ defined in (2.3) with that defined in (2.25) in the statements of all remaining propositions, lemmas and theorems, as well as the proofs of all propositions, lemmas and theorems stated in the main text.

Using this definition of $\boldsymbol{C}_2$, we define $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ from to be

$$\boldsymbol{Y}_1 = \boldsymbol{\beta} + \boldsymbol{L}\boldsymbol{\Omega}^{\mathrm{T}} + \boldsymbol{E}_1 \tag{2.26}$$

$$\boldsymbol{Y}_2 = \boldsymbol{Y}\boldsymbol{Q}_X = \boldsymbol{L}\boldsymbol{C}_2^{\mathrm{T}} + \boldsymbol{E}_2 \tag{2.27}$$

where $\boldsymbol{E}_1 \sim MN_{p\times d}\left\{0, \boldsymbol{\Sigma}, \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\right\}$ and $\boldsymbol{E}_2 \sim MN_{p\times(n-d)}\left(0, \boldsymbol{\Sigma}, I_{n-d}\right)$ are independent. Note that this $\boldsymbol{Y}_1$ is the same as the one defined in the main text. To get back to the $\boldsymbol{Y}_2$ defined in the main text, we simply multiply the $\boldsymbol{Y}_2$ defined in (2.27) on the right by $\boldsymbol{Q}_X^{\mathrm{T}}$. It is easy to see that because $\boldsymbol{Q}_X$ has orthonormal columns and $\boldsymbol{Q}_X\boldsymbol{Q}_X^{\mathrm{T}} = P_X^{\perp}$, Assumptions 2.1, 2.2 and 2.3 are equivalent with this redefinition of $\boldsymbol{C}_2$ and $\boldsymbol{Y}_2$. Further, Propositions 2.1 and 2.3 hold with $\boldsymbol{C}_2 = \boldsymbol{Q}_X^{\mathrm{T}}\boldsymbol{C}$.

We also define $\boldsymbol{y}_{2_g}$ and $\boldsymbol{y}_{1_g}$ to be the gth rows of $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ defined in (2.26) and (2.27), respectively. If $\boldsymbol{V} \in \mathbb{R}^{(n-d)\times K}$ are the first K right singular vectors of $\boldsymbol{Y}_2$ defined in (2.27), then $\hat{\boldsymbol{C}}_2 = n^{1/2}\boldsymbol{V}$, $\hat{\boldsymbol{L}} = \boldsymbol{Y}_2\hat{\boldsymbol{C}}_2\left(\hat{\boldsymbol{C}}_2^{\mathrm{T}}\hat{\boldsymbol{C}}_2\right)^{-1}$ and $\hat{\sigma}_g^2 = (n-d-K)^{-1}\boldsymbol{y}_{g2}^{\mathrm{T}}P_{\hat{C}_2}^{\perp}\boldsymbol{y}_{g2}$. Therefore, none of our estimators for $\boldsymbol{L}$, $\boldsymbol{\Sigma}$, $\boldsymbol{\Omega}$ or $\boldsymbol{\beta}$ change when we use this definition of $\boldsymbol{C}_2$ and $\boldsymbol{Y}_2$.

Since $\boldsymbol{\Sigma}$, $\boldsymbol{L}\boldsymbol{C}_2^{\mathrm{T}}$ and $\lambda_1, \ldots, \lambda_K$ are identifiable under Assumptions 2.1(a) and 2.1(b) and $\boldsymbol{\beta}, \boldsymbol{L}\boldsymbol{\Omega}^{\mathrm{T}}$ and $\boldsymbol{\Omega}^{\mathrm{T}}\left(n^{-1}\boldsymbol{C}_2^{\mathrm{T}}\boldsymbol{C}_2\right)^{-1}\boldsymbol{\Omega}$ are identifiable under Assumptions 2.1, 2.2 and 2.3(a) (see Propositions 2.1 and 2.3), then Lemma 2.1, (2.17) in Lemma 2.2 and Theorems 2.1 and 2.2 hold regardless of the parametrization of $\boldsymbol{L}$ and $\boldsymbol{C}$. Therefore, we will assume $(\boldsymbol{L}, \boldsymbol{C}) \in \Theta_{(0)}$ when Assumptions 2.1 and 2.2 hold, and will assume $(\boldsymbol{L}, \boldsymbol{C}) \in \Theta_{(1)}$ when 2.1, 2.2 and 2.3(a) hold (again, where $\boldsymbol{C}_2 = \boldsymbol{Q}_X^{\mathrm{T}}\boldsymbol{C}$). The first goal is to understand the asymptotic properties of $\hat{\boldsymbol{L}}$ and $\hat{\boldsymbol{C}}_2$, which are essential to all of the proofs that follow.

2.7.4 Understanding the behavior of $\hat{\boldsymbol{C}}_2$ and $\hat{\boldsymbol{L}}$

We start by stating and proving Lemmas 2.4 and 2.5 and use their results to prove theoretical statements made in the main text. For ease of notation, we assume for the statements and proofs in this subsection (Section 2.7.4) that

$$Y_{p\times n} = L_{p\times K}C_{K\times n}^{\mathrm{T}} + \boldsymbol{E}_{p\times n}, \quad \boldsymbol{E} \sim MN_{p\times n}\left(0, \boldsymbol{\Sigma}, I_n\right) \tag{2.28}$$

where $n^{-1}C^{\mathrm{T}}C = I_K$. We also define

$$\tilde{\boldsymbol{C}} = n^{-1/2}C \tag{2.29}$$

$$\tilde{\boldsymbol{L}} = n^{1/2}p^{-1/2}L. \tag{2.30}$$

We will lastly define a matrix $Q \in \mathbb{R}^{n\times n-K}$ such that $Q^{\mathrm{T}}Q = I_{n-K}$ and $\boldsymbol{Q}_T\tilde{\boldsymbol{C}} = 0_{(n-K)\times K}$. We use a technique developed in Paul (2007) to define the rotated matrix $F_{n\times n}$ to be

$$\begin{aligned} F &= \begin{pmatrix} \tilde{\boldsymbol{C}}^{\mathrm{T}} \\ Q^{\mathrm{T}} \end{pmatrix} p^{-1}Y^{\mathrm{T}}Y \begin{pmatrix} \tilde{\boldsymbol{C}} & Q \end{pmatrix} \\ &= \begin{bmatrix} \left(\tilde{\boldsymbol{L}} + p^{-1/2}\tilde{\boldsymbol{E}}_1\right)^{\mathrm{T}} \left(\tilde{\boldsymbol{L}} + p^{-1/2}\tilde{\boldsymbol{E}}_1\right) & \left(\tilde{\boldsymbol{L}} + p^{-1/2}\tilde{\boldsymbol{E}}_1\right)^{\mathrm{T}} p^{-1/2}\tilde{\boldsymbol{E}}_2 \\ p^{-1/2}\tilde{\boldsymbol{E}}_2^{\mathrm{T}} \left(\tilde{\boldsymbol{L}} + p^{-1/2}\tilde{\boldsymbol{E}}_1\right) & p^{-1}\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\tilde{\boldsymbol{E}}_2 \end{bmatrix} \end{aligned} \tag{2.31}$$

where $\tilde{\boldsymbol{E}}_1 = \boldsymbol{E}\tilde{\boldsymbol{C}}$ and $\tilde{\boldsymbol{E}}_2 = \boldsymbol{E}Q$ are independent. Since $\begin{pmatrix} \tilde{\boldsymbol{C}} & Q \end{pmatrix}$ is a unitary matrix, the eigenvalues of F are also the eigenvalues of $p^{-1}Y^{\mathrm{T}}Y$. For the remainder of the chapter, we assume

$$\begin{pmatrix} \hat{V}_{K\times K} \\ \hat{Z}_{(n-K)\times K} \end{pmatrix}$$

are the first K eigenvectors of F, meaning $\tilde{\boldsymbol{C}}\hat{V} + Q\hat{Z}$ are the first K eigenvectors of $p^{-1}Y^{\mathrm{T}}Y$. Further, since $\tilde{\boldsymbol{E}}_1$ and $\tilde{\boldsymbol{E}}_2$ are independent, the upper left block of F is independent of $\tilde{\boldsymbol{E}}_2$. We exploit this by first studying the eigen-structure of the upper left block in Lemma 2.4, and then using those results to enumerate the asymptotic properties of the first K eigenvalues and eigenvectors of F in Lemma 2.5. In order to avoid confusing subscripts and superscripts, we define the scalar $\boldsymbol{v}[s]$ to be the sth component of the vector $\boldsymbol{v}$.

***Lemma* 2.4.** *Let* $\tilde{\boldsymbol{L}} \in \mathbb{R}^{p\times K}$, $\tilde{\boldsymbol{E}}_1 \sim MN_{p\times K}(0, \boldsymbol{\Sigma}, I_K)$ *and* $\tilde{\boldsymbol{N}} = \tilde{\boldsymbol{L}} + p^{-1/2}\tilde{\boldsymbol{E}}_1$. *Assume* $\tilde{\boldsymbol{L}}^{\mathrm{T}}\tilde{\boldsymbol{L}} = diag(\lambda_1, \ldots, \lambda_K)$ *where the* λ_k*'s are the same as those given in Assumption 2.2 and*

$\boldsymbol{\Sigma} = \mathrm{diag}\left(\sigma_1^2, \ldots, \sigma_p^2\right)$ *follows Assumption 2.1(c). If* $d_k^2 = \lambda_k\left(\tilde{\boldsymbol{N}}^{\mathrm{T}}\tilde{\boldsymbol{N}}\right)$ *and* $\boldsymbol{v}_k$ *are the* k^{th} *eigenvalue and eigenvector of* $\tilde{\boldsymbol{N}}^{\mathrm{T}}\tilde{\boldsymbol{N}}$*, then*

$$d_k^2\lambda_k^{-1} = 1 + \rho\lambda_k^{-1} + O_{\mathrm{p}}\left\{(\lambda_k p)^{-1/2}\right\} \tag{2.32}$$

and

$$\begin{aligned}\boldsymbol{v}_k = &\left[1 + O_{\mathrm{p}}\left\{(\lambda_k p)^{-1}\right\}\right]\boldsymbol{e}_k + O_{\mathrm{p}}\left\{(\lambda_1 p)^{-1/2}\right\}\boldsymbol{e}_1 + \cdots + O_{\mathrm{p}}\left\{(\lambda_{k-1} p)^{-1/2}\right\}\boldsymbol{e}_{k-1} \\ &+ O_{\mathrm{p}}\left\{(\lambda_k p)^{-1/2}\right\}\boldsymbol{e}_{k+1} + \cdots + O_{\mathrm{p}}\left\{(\lambda_k p)^{-1/2}\right\}\boldsymbol{e}_K\end{aligned} \tag{2.33}$$

where $\boldsymbol{e}_k$*,* $k = 1, \ldots, K$*, are the standard basis vectors in* $\mathbb{R}^K$*.*

Proof. First, $\tilde{\boldsymbol{N}}^{\mathrm{T}}\tilde{\boldsymbol{N}} = \tilde{\boldsymbol{L}}^{\mathrm{T}}\tilde{\boldsymbol{L}} + \rho I_K + p^{-1/2}\tilde{\boldsymbol{L}}^{\mathrm{T}}\tilde{\boldsymbol{E}}_1 + p^{-1/2}\tilde{\boldsymbol{E}}_1^{\mathrm{T}}\tilde{\boldsymbol{L}} + \boldsymbol{B}$ where the entries of $\boldsymbol{B}$ are $O_{\mathrm{p}}\left(p^{-1/2}\right)$. Let $\boldsymbol{R}\boldsymbol{R}^{\mathrm{T}} = \tilde{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{\Sigma}\tilde{\boldsymbol{L}}$ where $\boldsymbol{R}$ is a lower triangular matrix. By Cauchy-Schwartz, the kth row of $\boldsymbol{R}$ is $\boldsymbol{R}_k^{\mathrm{T}} = O\left(\lambda_k^{1/2}\right)$. We also note that $p^{-1/2}\tilde{\boldsymbol{L}}^{\mathrm{T}}\tilde{\boldsymbol{E}}_1 \sim \boldsymbol{R}\boldsymbol{M}$ where the entries of $\boldsymbol{M} \in \mathbb{R}^{K\times K}$ are $O_{\mathrm{p}}\left(p^{-1/2}\right)$. If we let the columns of $\boldsymbol{M}$ be $\boldsymbol{M}_s$ ($s \in [K]$), then $[\boldsymbol{R}\boldsymbol{M}]_{ks} = \boldsymbol{R}_k^{\mathrm{T}}\boldsymbol{M}_s = O_{\mathrm{p}}\left\{\left(\lambda_k p^{-1}\right)^{1/2}\right\}$ ($k, s \in [K]$). Next, define the matrix $\boldsymbol{A}^{(1)} \in \mathbb{R}^{K\times K}$ to be

$$\boldsymbol{A}^{(1)} = \lambda_1^{-1}\tilde{\boldsymbol{N}}^{\mathrm{T}}\tilde{\boldsymbol{N}} = \begin{pmatrix} \mu_1 & a_{12} & \cdots & a_{1K} \\ a_{21} & \mu_2 & \cdots & a_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ a_{K1} & a_{K2} & \cdots & \mu_K \end{pmatrix}$$

where

$$\begin{aligned}\mu_k &= (\lambda_k + \rho)\lambda_1^{-1} + 2\lambda_1^{-1}\boldsymbol{R}_k^{\mathrm{T}}\boldsymbol{M}_k + \lambda_1^{-1}\boldsymbol{B}_{kk} \\ a_{ks} &= \lambda_1^{-1}\boldsymbol{R}_k^{\mathrm{T}}\boldsymbol{M}_s + \lambda_1^{-1}\boldsymbol{R}_s^{\mathrm{T}}\boldsymbol{M}_k + \lambda_1^{-1}\boldsymbol{B}_{sk} = O_{\mathrm{p}}\left(\lambda_k^{1/2}\lambda_1^{-1}p^{-1/2}\right)\end{aligned}$$

for $k < s$. Our goal is to break $\boldsymbol{A}^{(1)}$ into K rank one pieces, each of which are approximately orthogonal. The procedure is enumerated in four steps:

(1) Define $\boldsymbol{A}_1 = \boldsymbol{A}_1^{(1)}$, $\boldsymbol{A}_2 = \left(0, \left[\boldsymbol{A}^{(1)}\right]_{22}, \ldots, \left[\boldsymbol{A}^{(1)}\right]_{2K}\right), \ldots,$

$$\boldsymbol{A}_K = \left(\underbrace{0, \ldots, 0}_{K-1 \text{ 0's}}, \left[\boldsymbol{A}^{(1)}\right]_{KK}\right)^{\mathrm{T}}.$$

(2) We wish to first modify $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ so that they are orthogonal. To do this, we will add ϵ_2 to $\boldsymbol{A}_2[1]$ and remove ϵ_2 from $\boldsymbol{A}_1[2]$. That is, we define $\boldsymbol{A}_{1_2} = \boldsymbol{A}_1 + \epsilon_2 \boldsymbol{e}_2$ and $\boldsymbol{A}_{2_2} = \boldsymbol{A}_2 - \epsilon_2 \boldsymbol{e}_1$ such that

$$0 = \boldsymbol{A}_{1_2}^{\mathrm{T}} \boldsymbol{A}_{2_2} = \boldsymbol{A}_1^{\mathrm{T}} \boldsymbol{A}_2 + \epsilon_2 \mu_2 - \epsilon_2 \mu_1 = a_{12}\mu_2 + \epsilon_2\mu_2 - \epsilon_2\mu_1 + O_{\mathrm{p}}\left(\lambda_2^{1/2}\lambda_1^{-3/2}p^{-1}\right)$$

meaning $\epsilon_2 = a_{12}\mu_2 \left(\mu_1 - \mu_2\right)^{-1} + O_{\mathrm{p}}\left(\lambda_2^{1/2}\lambda_1^{-3/2}p^{-1}\right) = O_{\mathrm{p}}\left(\lambda_2\lambda_1^{-3/2}p^{-1/2}\right)$. We now have $\boldsymbol{A}_{1_2}^{\mathrm{T}} \boldsymbol{A}_{2_2} = 0$.

(3) Define $\boldsymbol{A}_{1_k} = \boldsymbol{A}_{1_{k-1}} + \epsilon_k \boldsymbol{e}_k$ and $\boldsymbol{A}_{k_2} = \boldsymbol{A}_k - \epsilon_k \boldsymbol{e}_1$ inductively:

$$\begin{aligned} 0 = &\left(\boldsymbol{A}_{1_{k-1}} + \epsilon_k \boldsymbol{e}_k\right)^{\mathrm{T}} \left(\boldsymbol{A}_k - \epsilon_k \boldsymbol{e}_1\right) = \boldsymbol{A}_{1_{k-1}}^{\mathrm{T}} \boldsymbol{A}_k + \epsilon_k\mu_k - \epsilon_k\mu_1 \\ = &a_{1k}\mu_k + \epsilon_k\mu_k - \epsilon_k\mu_1 + O_{\mathrm{p}}\left(\lambda_k^{1/2}\lambda_1^{-3/2}p^{-1}\right) \end{aligned}$$

meaning $\epsilon_k = a_{1k}\mu_k \left(\mu_1 - \mu_k\right)^{-1} + O_{\mathrm{p}}\left(\lambda_k^{1/2}\lambda_1^{-3/2}p^{-1}\right) = O_{\mathrm{p}}\left(\lambda_k\lambda_1^{-3/2}p^{-1/2}\right)$.

(4) After we complete this process $K-1$ times to get $\boldsymbol{A}_{1_K}$, we now have for $s < K$

$$
\begin{aligned}
\boldsymbol{A}_{1_K}^{\mathrm{T}} \boldsymbol{A}_{s_2} &= (\boldsymbol{A}_1 + \epsilon_2 \boldsymbol{e}_2 + \cdots + \epsilon_K \boldsymbol{e}_K)^{\mathrm{T}} (\boldsymbol{A}_s - \epsilon_s \boldsymbol{e}_1) \\
&= (\boldsymbol{A}_1 + \epsilon_2 \boldsymbol{e}_2 + \cdots + \epsilon_s \boldsymbol{e}_s)^{\mathrm{T}} (\boldsymbol{A}_s - \epsilon_s \boldsymbol{e}_1) \\
&\quad + (\epsilon_{s+1} \boldsymbol{e}_{s+1} + \cdots + \epsilon_K \boldsymbol{e}_K)^{\mathrm{T}} (\boldsymbol{A}_s - \epsilon_s \boldsymbol{e}_1) \\
&= 0 + \epsilon_{s+1} a_{s,s+1} + \cdots \epsilon_K a_{s,K} \\
&= O_{\mathrm{p}} \left(\lambda_{s+1} \lambda_1^{-3/2} p^{-1/2} \lambda_s^{1/2} \lambda_1^{-1} p^{-1/2} \right) \\
&= O_{\mathrm{p}} \left\{ \left(\lambda_s \lambda_1^{-1} \right)^{3/2} (\lambda_1 p)^{-1} \right\}
\end{aligned}
$$

and $\boldsymbol{A}_{1_K}^{\mathrm{T}} \boldsymbol{A}_{1_K} = \mu_1^2 + O_{\mathrm{p}} \left\{ (\lambda_1 p)^{-1} \right\}$, meaning $\|\boldsymbol{A}_{1_K}\|_2 = \mu_1 + O_{\mathrm{p}} \left\{ (\lambda_1 p)^{-1} \right\}$.

We now have

$$
\begin{aligned}
\boldsymbol{A}^{(1)} &= \underbrace{\begin{pmatrix} \boldsymbol{A}_{1_K} & \rightarrow \\ \downarrow & \mathbf{0}_{(K-1)\times(K-1)} \end{pmatrix}}_{\boldsymbol{B}^{(1)}} + \underbrace{\begin{pmatrix} 0 & \uparrow & \mathbf{0}_{1\times(K-2)} \\ \leftarrow & \boldsymbol{A}_{2_2} & \rightarrow \\ \mathbf{0}_{(K-2)\times 1} & \downarrow & \mathbf{0}_{(K-2)\times(K-2)} \end{pmatrix}}_{\boldsymbol{B}^{(2)}} \\
&\quad + \cdots + \underbrace{\begin{pmatrix} \mathbf{0}_{(K-1)\times(K-1)} & \uparrow \\ \leftarrow & \boldsymbol{A}_{K_2} \end{pmatrix}}_{\boldsymbol{B}^{(K)}} \\
&= \begin{pmatrix} \mu_1 & a_{12}+\epsilon_2 & \cdots & a_{1K}+\epsilon_K \\ a_{12}+\epsilon_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{1K}+\epsilon_K & 0 & \cdots & 0 \end{pmatrix} + \begin{pmatrix} 0 & -\epsilon_2 & 0 & \cdots & 0 \\ -\epsilon_2 & \mu_2 & a_{23} & \cdots & a_{2K} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & a_{2K} & 0 & \cdots & 0 \end{pmatrix} \\
&\quad + \cdots + \begin{pmatrix} 0 & \cdots & 0 & -\epsilon_K \\ \vdots & \cdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 \\ -\epsilon_K & \cdots & 0 & \mu_K \end{pmatrix}
\end{aligned}
$$

Define

$$
\boldsymbol{u}_{1_K} = \|\boldsymbol{A}_{1_K}\|_2^{-1}\boldsymbol{A}_{1_K} = \left\{1, (a_{12}+\epsilon_2)\,\mu_1^{-1}, \ldots, (a_{1K}+\epsilon_K)\,\mu_1^{-1}\right\}^{\mathrm{T}} + O_{\mathrm{p}}\left\{(\lambda_1 p)^{-1}\right\}.
$$

Then $\boldsymbol{B}^{(1)} = \mu_1 \boldsymbol{u}_{1_K} \boldsymbol{u}_{1_K}^{\mathrm{T}} + O_{\mathrm{p}}\left\{(\lambda_1 p)^{-1}\right\}$. Further, for $s \in [K]$,

$$\|\boldsymbol{B}^{(s)} \boldsymbol{u}_{1_K}\|_2 = \left\| \begin{pmatrix} -\epsilon_s \left(a_{1s} + \epsilon_s\right) \|\boldsymbol{A}_{1_K}\|_2^{-1} \\ 0 \\ \vdots \\ 0 \\ \|\boldsymbol{A}_{1_K}\|_2^{-1} \boldsymbol{A}_{s_2}^{\mathrm{T}} \boldsymbol{A}_{1_K} \\ \|\boldsymbol{A}_{1_K}\|_2^{-1} a_{s,s+1} \left(a_{1s} + \epsilon_s\right) \\ \vdots \\ \|\boldsymbol{A}_{1_K}\|_2^{-1} a_{s,K} \left(a_{1s} + \epsilon_s\right) \end{pmatrix} \right\|_2 = O_{\mathrm{p}}\left\{(\lambda_1 p)^{-1}\right\}$$

which means $\boldsymbol{A}^{(1)} \boldsymbol{u}_{1_K} = \mu_1 \boldsymbol{u}_{1_K} + O_{\mathrm{p}}\left\{(\lambda_1 p)^{-1}\right\}$. We then define

$$\begin{aligned} \delta &= \boldsymbol{u}_{1_K}^{\mathrm{T}} \boldsymbol{A}^{(1)} \boldsymbol{u}_{1_K} = \mu_1 + O_{\mathrm{p}}\left\{(\lambda_1 p)^{-1}\right\} \\ \gamma &= \|\boldsymbol{A}^{(1)} \boldsymbol{u}_{1_K} - \delta \boldsymbol{u}_{1_K}\|_2 = O_{\mathrm{p}}\left\{(\lambda_1 p)^{-1}\right\}. \end{aligned}$$

By Weyl's Theorem, the eigenvalues of $\boldsymbol{A}^{(1)}$ are $\mu_k + O_{\mathrm{p}}\left\{(\lambda_1 p)^{-1/2}\right\}$, so if ξ is the second largest eigenvalue of $\boldsymbol{A}^{(1)}$, $\xi = \mu_2 + O_{\mathrm{p}}\left\{(\lambda_1 p)^{-1/2}\right\}$, meaning $f = \delta - \xi = (\lambda_1 - \lambda_2)\, \lambda_1^{-1} + O_{\mathrm{p}}\left\{(\lambda_1 p)^{-1/2}\right\}$. By Theorem 3.6 in Auffinger & Tang (2015), we have

(1) There exists an eigenvalue λ_γ of $\boldsymbol{A}^{(1)}$ s.t. $\lambda_\gamma \in [\delta - \gamma, \delta + \gamma]$, i.e. $\lambda_\gamma = \mu_1 + O_{\mathrm{p}}\left\{(\lambda_1 p)^{-1}\right\}$.

(2) If λ_γ is the only eigenvalue in $[\delta - \gamma, \delta + \gamma]$ and $\boldsymbol{v}_\gamma$ is the eigenvector corresponding to λ_γ and $f > \gamma$,

$$\|\boldsymbol{v}_\gamma - \boldsymbol{u}_{1_K}^{\mathrm{T}} \boldsymbol{v}_\gamma \boldsymbol{u}_{1_K}\|_2 \leq 2\gamma \left(f - \gamma\right)^{-1} = O_{\mathrm{p}}\left\{(\lambda_1 p)^{-1}\right\}.$$

Let $G_{\lambda_\gamma,n,p} = \left\{\lambda_\gamma \text{ is the maximum eigenvalue of } \boldsymbol{A}^{(1)}\right\}$. Then

$$\begin{aligned}\mathbb{P}\left[\left|\lambda_1\left\{\boldsymbol{A}^{(1)}\right\} - \mu_1\right| \geq M\right] &\leq \mathbb{P}\left(|\lambda_\gamma - \mu_1| \geq M, G_{\lambda_\gamma,n,p}\right) + \mathbb{P}\left(G^c_{\lambda_\gamma,n,p}\right)\\ &\leq \mathbb{P}\left(|\delta - \mu_1| \geq M\right) + \mathbb{P}\left(G^c_{\lambda_\gamma,n,p}\right)\end{aligned}$$

Since $\mathbb{P}\left(G^c_{\lambda_\gamma,n,p}\right) \to 0$ and $|\lambda_\gamma - \mu_1| = O_{\mathrm{p}}\left\{(\lambda_1 p)^{-1}\right\}$, $d_1^2\lambda_1^{-1} = \lambda_1\left\{\boldsymbol{A}^{(1)}\right\} = \mu_1 + O_{\mathrm{p}}\left\{(\lambda_1 p)^{-1}\right\}$. We can apply an identical procedure to show that $\|\boldsymbol{v}_1 - \boldsymbol{u}_{1_K}^{\mathrm{T}}\boldsymbol{v}_1\boldsymbol{u}_{1_K}\|_2 = O_{\mathrm{p}}\left\{(\lambda_1 p)^{-1}\right\}$ since on the event that λ_γ is the largest eigenvalue of $\boldsymbol{A}^{(1)}$, $\lambda_\gamma - \xi > c + o_{\mathrm{p}}(1)$, where c is a constant that does not depend on n or p (i.e. λ_γ is the only eigenvalue in $[\delta - \gamma, \delta + \gamma]$ and $f > \delta$ with probability tending to 1). Since $\boldsymbol{v}_1$ and $\boldsymbol{u}_{1_K}$ are unit vectors, we must have $\boldsymbol{u}_{1_K}^{\mathrm{T}}\boldsymbol{v}_1 = \pm 1 + O_{\mathrm{p}}\left\{(\lambda_1 p)^{-2}\right\}$. That is, we know $\boldsymbol{v}_1$ up to sign parity.

We then have

$$\boldsymbol{A}^{(2)} = \lambda_2^{-1}\left(\lambda_1\boldsymbol{A}^{(1)} - d_1^2\boldsymbol{v}_1\boldsymbol{v}_1^{\mathrm{T}}\right) = \lambda_1\lambda_2^{-1}\boldsymbol{B}^{(2)} + \cdots + \lambda_1\lambda_2^{-1}\boldsymbol{B}^{(K)} + O_{\mathrm{p}}\left\{(\lambda_2 p)^{-1}\right\}.$$

Since $\epsilon_k\lambda_1\lambda_2^{-1} = O_{\mathrm{p}}\left\{\lambda_k\lambda_2^{-1}(\lambda_1 p)^{-1/2}\right\}$, all off-diagonal entries of the above matrix at most $O_{\mathrm{p}}\left\{(\lambda_2 p)^{-1/2}\right\}$. We can then apply the exact same procedure as we did above to show that for all $k \in [K]$,

$$d_k^2\lambda_k^{-1} = 1 + \rho\lambda_k^{-1} + O_{\mathrm{p}}\left\{(\lambda_k p)^{-1/2}\right\}$$

and

$$
\boldsymbol{v}_k = \begin{bmatrix} O_{\mathrm{p}}\left\{(\lambda_k p)^{-1/2}\right\} \\ \vdots \\ 1 + O_{\mathrm{p}}\left\{(\lambda_k p)^{-1}\right\} \\ \vdots \\ O_{\mathrm{p}}\left\{(\lambda_k p)^{-1/2}\right\}. \end{bmatrix}
$$

Lastly, for $s < k$,

$$
0 = \boldsymbol{v}_s^{\mathrm{T}} \boldsymbol{v}_k = \boldsymbol{v}_k[s]\boldsymbol{v}_s[s] + O_{\mathrm{p}}\left\{(\lambda_k p)^{-1}\right\} + \boldsymbol{v}_s[k]\boldsymbol{v}_k[k] = \boldsymbol{v}_k[s] + O_{\mathrm{p}}\left\{(\lambda_s p)^{-1/2}\right\}
$$

meaning $\boldsymbol{v}_k[s] = O_{\mathrm{p}}\left\{(\lambda_s p)^{-1/2}\right\}$ since $\lambda_s^{1/2}\lambda_k^{-1}p^{-1/2} \to 0$ by assumption. This completes the proof.

□

We use $\tilde{\boldsymbol{E}}_1$, $\tilde{\boldsymbol{E}}_2$, $\tilde{\boldsymbol{N}}$, d_k and $\boldsymbol{v}_k$ defined in Lemma 2.4 in the remainder of the paper. We also define

$$
\boldsymbol{R} = p^{-1}\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\tilde{\boldsymbol{E}}_2 - \rho I_{n-K} \tag{2.34}
$$

and let $\boldsymbol{V} = \begin{bmatrix} \boldsymbol{v}_1 & \cdots & \boldsymbol{v}_K \end{bmatrix}$, $\boldsymbol{U} = \begin{bmatrix} \boldsymbol{u}_1 \cdots \boldsymbol{u}_K \end{bmatrix}$ be the first K right and left singular values of $\tilde{\boldsymbol{N}}$. That is

$$
\tilde{\boldsymbol{N}} = \tilde{\boldsymbol{L}} + p^{-1/2}\tilde{\boldsymbol{E}}_1 = \boldsymbol{U}\boldsymbol{D}V^{\mathrm{T}} \tag{2.35}
$$

is the singular value decomposition of $\tilde{\boldsymbol{N}}$, where $\boldsymbol{D} = \operatorname{diag}(d_1, \ldots, d_K)$. By Theorem 5.39 in Eldar & Kutyniok (2012), $\|\boldsymbol{R}\|_2 = O_{\mathrm{p}}\left\{\left(np^{-1}\right)^{1/2}\right\}$ under Assumptions 2.1(c) and 2.2(c).

The next lemma uses what we have established in Lemma 2.4 to prove convergence properties of the first K eigenvalues and eigenvectors of $\boldsymbol{F}$ (see (2.31)).

***Lemma* 2.5.** *Suppose the probability model for $\boldsymbol{Y}$ is given by (2.28) and that Assumptions 2.1 and 2.2 hold for $d = 0$ (d is the number of columns in $\boldsymbol{X}$). Then*

$$\hat{\lambda}_k = \lambda_k\left(\boldsymbol{F}\right) = d_k^2 + O_{\mathrm{p}}\left(np^{-1}\right). \tag{2.36}$$

Define $\begin{bmatrix} \hat{\boldsymbol{v}}_k \\ \hat{\boldsymbol{z}}_k \end{bmatrix}$, *$\hat{\boldsymbol{v}}_k \in \mathbb{R}^K$ and $\hat{\boldsymbol{z}}_k \in \mathbb{R}^{n-K}$ to be the k^{th} eigenvector of $\boldsymbol{F}$. Then*

$$\hat{\boldsymbol{v}}_k = \boldsymbol{v}_k + \boldsymbol{\epsilon}_k, \quad \|\boldsymbol{\epsilon}_k\|_2 = O_{\mathrm{p}}\left\{n\left(\lambda_k p\right)^{-1}\right\}. \tag{2.37}$$

and

$$\hat{\boldsymbol{z}}_k = d_k\lambda_k^{-1}p^{-1/2}\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\boldsymbol{u}_k + d_k\lambda_k^{-1}p^{-1/2}\boldsymbol{R}\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\boldsymbol{u}_k + O_{\mathrm{p}}\left\{n^{3/2}\left(\lambda_k p\right)^{-3/2} + n^{1/2}\left(p\lambda_k\right)^{-1}\right\} \tag{2.38}$$

where d_k and $\boldsymbol{v}_k$ are defined (2.32) and (2.33) and $\boldsymbol{u}_k$ is the k^{th} left singular vector of $\boldsymbol{Y}\tilde{\boldsymbol{C}}$. Further, if $np^{-1}\left[\boldsymbol{L}\right]_{*k}^{\mathrm{T}}\boldsymbol{\Sigma}\left[\boldsymbol{L}\right]_{*s} \leq c\lambda_{\max(k,s)}$ for all $k, s \in [K]$, then for any $s < k$,

$$\boldsymbol{\epsilon}_k[s] = o_{\mathrm{p}}\left(\lambda_k\lambda_s^{-1}n^{-1/2}\right). \tag{2.39}$$

Proof. First, define

$$\boldsymbol{F}^{(1)} = \boldsymbol{F} = \lambda_1\begin{bmatrix} \hat{\boldsymbol{A}}_1 & \boldsymbol{H}_1 \\ \boldsymbol{H}_1^{\mathrm{T}} & \boldsymbol{J}_1 \end{bmatrix}.$$

We immediately observe from the expression for $\boldsymbol{F}$ in (2.31) that

$$\hat{\lambda}_1\lambda_1^{-1} = d_1^2\lambda_1^{-1} + O_{\mathrm{p}}\left\{n^{1/2}\left(\lambda_1 p\right)^{-1/2}\right\} = \left(\lambda_1 + \rho\right)\lambda_1^{-1} + O_{\mathrm{p}}\left\{n^{1/2}\left(\lambda_1 p\right)^{-1/2}\right\}$$

by Weyl's Theorem. The eigenvalue equations for $\boldsymbol{F}^{(1)}$ are

$$\hat{\lambda}_1\lambda_1^{-1}\hat{\boldsymbol{v}}_1 = \hat{\boldsymbol{A}}_1\hat{\boldsymbol{v}}_1 + \boldsymbol{H}_1\hat{\boldsymbol{z}}_1$$
$$\hat{\lambda}_1\lambda_1^{-1}\hat{\boldsymbol{z}}_1 = \boldsymbol{H}_1^{\mathrm{T}}\hat{\boldsymbol{v}}_1 + \boldsymbol{J}_1\hat{\boldsymbol{z}}_1$$

which then implies

$$\hat{\boldsymbol{z}}_1 = \left(\hat{\lambda}_1\lambda_1^{-1}I_{n-K} - \boldsymbol{J}_1\right)^{-1}\boldsymbol{H}_1^{\mathrm{T}}\hat{\boldsymbol{v}}_1$$
$$\hat{\lambda}_1\lambda_1^{-1}\hat{\boldsymbol{v}}_1 = \hat{\boldsymbol{A}}_1\hat{\boldsymbol{v}}_1 + \boldsymbol{H}_1\left(\hat{\lambda}_1\lambda_1^{-1}I_{n-K} - \boldsymbol{J}_1\right)^{-1}\boldsymbol{H}_1^{\mathrm{T}}\hat{\boldsymbol{v}}_1$$

where

$$\boldsymbol{H}_1 = \lambda_1^{-1}p^{-1/2}\left(\tilde{\boldsymbol{L}} + p^{-1/2}\tilde{\boldsymbol{E}}_1\right)^{\mathrm{T}}\tilde{\boldsymbol{E}}_2$$
$$\hat{\lambda}_1\lambda_1^{-1}I_{n-K} - \boldsymbol{J}_1 = \left(\hat{\lambda}_1 - \rho\right)\lambda_1^{-1}I_{n-K} - \lambda_1^{-1}\boldsymbol{R}.$$

The latter is invertible with eigenvalues that are uniformly bounded away from 0 with high probability, since

$$\hat{\lambda}_1\lambda_1^{-1} = (\lambda_1 + \rho)\lambda_1^{-1} + O_{\mathrm{p}}\left\{n^{1/2}(\lambda_1 p)^{-1/2}\right\}$$

and $\|\boldsymbol{R}\|_2 = O_{\mathrm{p}}\left(n^{1/2}p^{-1/2}\right)$. Therefore,

$$\|\boldsymbol{H}_1\left(\hat{\lambda}_1\lambda_1^{-1}I_{n-K} - \boldsymbol{J}_1\right)^{-1}\boldsymbol{H}_1^{\mathrm{T}}\|_2 = O_{\mathrm{p}}\left\{n(\lambda_1 p)^{-1}\right\}.$$

Since $\hat{\boldsymbol{A}}_1 = \boldsymbol{A}^{(1)}$ (see Lemma 2.4),

$$\hat{\lambda}_1\lambda_1^{-1} = \lambda_1\left\{\boldsymbol{A}^{(1)}\right\} + O_{\mathrm{p}}\left\{n(\lambda_1 p)^{-1}\right\} = d_1^2\lambda_1^{-1} + O_{\mathrm{p}}\left\{n(\lambda_1 p)^{-1}\right\}$$

by Weyl's Theorem. To determine the behavior of $\hat{\boldsymbol{v}}_1$, we first notice that since $\hat{\boldsymbol{z}}_1^{\mathrm{T}}\hat{\boldsymbol{z}}_1 =$

$O_{\mathrm{p}}\left\{n\left(\lambda_1 p\right)^{-1}\right\}$ and $\|\hat{\boldsymbol{v}}_1\|_2^2+\|\hat{\boldsymbol{z}}_1\|_2^2=1$, $\|\hat{\boldsymbol{v}}_1\|_2=1-O_{\mathrm{p}}\left\{n\left(\lambda_1 p\right)^{-1}\right\}$. This shows that,

$$\hat{\boldsymbol{v}}_1=\boldsymbol{v}_1+O_{\mathrm{p}}\left\{n\left(\lambda_1 p\right)^{-1}\right\}.$$

Recall from (2.35) that $\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}}=\tilde{\boldsymbol{L}}+p^{-1/2}\tilde{\boldsymbol{E}}_1$ is the singular value decomposition of $\tilde{\boldsymbol{L}}+p^{-1/2}\tilde{\boldsymbol{E}}_1$. Using these above relations and the fact that

$$\left(\hat{\lambda}_1-\rho\right)\lambda_1^{-1}=1+O_{\mathrm{p}}\left\{\left(\lambda_1 p\right)^{-1/2}+n\left(\lambda_1 p\right)^{-1}\right\},$$

we can get an expression for $\hat{\boldsymbol{z}}_1$:

$$\begin{aligned}
\hat{\boldsymbol{z}}_1&=\left(\hat{\lambda}_1\lambda_1^{-1}I_{n-K}-\boldsymbol{J}_1\right)^{-1}\boldsymbol{H}_1^{\mathrm{T}}\hat{\boldsymbol{v}}_1\\
&=\lambda_1^{-1}p^{-1/2}\left(\hat{\lambda}_1\lambda_1^{-1}I_{n-K}-\left(\lambda_1 p\right)^{-1}\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\tilde{\boldsymbol{E}}_2\right)^{-1}\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\left(\tilde{\boldsymbol{L}}+p^{-1/2}\tilde{\boldsymbol{E}}_1\right)\hat{\boldsymbol{v}}_1\\
&=\lambda_1^{-1}p^{-1/2}\left\{\left(\hat{\lambda}_1-\rho\right)\lambda_1^{-1}I_{n-K}-\lambda_1^{-1}\boldsymbol{R}\right\}^{-1}\tilde{\boldsymbol{E}}_2^{\mathrm{T}}UDV^{\mathrm{T}}\boldsymbol{v}_1+O_{\mathrm{p}}\left\{n^{3/2}\left(\lambda_1 p\right)^{-3/2}\right\}\\
&=d_1\lambda_1^{-1}p^{-1/2}\left(I_{n-K}-\lambda_1^{-1}\boldsymbol{R}\right)^{-1}\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\boldsymbol{u}_1+O_{\mathrm{p}}\left\{n^{3/2}\left(\lambda_1 p\right)^{-3/2}+n^{1/2}\left(p\lambda_1\right)^{-1}\right\}\\
&=d_1\lambda_1^{-1}p^{-1/2}\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\boldsymbol{u}_1+d_1\lambda_1^{-2}p^{-1/2}\boldsymbol{R}\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\boldsymbol{u}_1+O_{\mathrm{p}}\left\{n^{3/2}\left(\lambda_1 p\right)^{-3/2}+n^{1/2}\left(p\lambda_1\right)^{-1}\right\}
\end{aligned}$$

since

$$\|\left(I_{n-K}-\lambda_1^{-1}\boldsymbol{R}\right)^{-1}-\left(I_{n-K}+\lambda_1^{-1}\boldsymbol{R}\right)\|_2=O\left(\|\lambda_1^{-2}\boldsymbol{R}^2\|_2\right)=O_{\mathrm{p}}\left\{n\left(\lambda_1^2 p\right)^{-1}\right\}.$$

We can then find expressions for $\hat{\lambda}_k$, $\hat{\boldsymbol{v}}_k$ and $\hat{\boldsymbol{z}}_k$ by induction. First, assume the following three conditions hold for all $s \le k$, where $k < K$.

$$\hat{\lambda}_s = d_s^2 + O_{\mathrm{p}}\left(np^{-1}\right) \tag{2.40a}$$

$$\hat{\boldsymbol{v}}_s = \boldsymbol{v}_s + O_{\mathrm{p}}\left\{n\left(\lambda_s p\right)^{-1}\right\} \tag{2.40b}$$

$$\hat{\boldsymbol{z}}_s = d_s \lambda_s^{-1} p^{-1/2} \tilde{\boldsymbol{E}}_2^{\mathrm{T}} \boldsymbol{u}_s + d_s \lambda_s^{-2} p^{-1/2} \boldsymbol{R} \tilde{\boldsymbol{E}}_2^{\mathrm{T}} \boldsymbol{u}_s + O_{\mathrm{p}}\left\{n^{3/2}\left(\lambda_s p\right)^{-3/2} + n^{1/2}\left(p\lambda_s\right)^{-1}\right\} \tag{2.40c}$$

$$\begin{aligned}
\lambda_s \boldsymbol{H}_s^{\mathrm{T}} =& p^{-1/2} \tilde{\boldsymbol{E}}_2^{\mathrm{T}} \tilde{\boldsymbol{N}} - \hat{\lambda}_1 \hat{\boldsymbol{z}}_1 \hat{\boldsymbol{v}}_1^{\mathrm{T}} - \cdots - \hat{\lambda}_{s-1} \hat{\boldsymbol{z}}_{s-1} \hat{\boldsymbol{v}}_{s-1}^{\mathrm{T}} \\
=& O_{\mathrm{p}}\left\{n^{1/2}\left(\lambda_1 p\right)^{-1/2}\right\} \boldsymbol{v}_1^{\mathrm{T}} + \cdots + O_{\mathrm{p}}\left\{n^{1/2}\left(\lambda_{s-1} p\right)^{-1/2}\right\} \boldsymbol{v}_{s-1}^{\mathrm{T}} \\
&+ p^{-1/2} \tilde{\boldsymbol{E}}_2^{\mathrm{T}} \sum_{\ell=s}^{K} d_\ell \boldsymbol{u}_\ell \boldsymbol{v}_\ell^{\mathrm{T}} + O_{\mathrm{p}}\left\{\lambda_{s-1}^{-1/2}\left(np^{-1}\right)^{3/2} + n^{1/2} p^{-1}\right\}.
\end{aligned} \tag{2.40d}$$

If we can show that these hold for $k+1$, this would prove (2.36), (2.37) and (2.38). To show that the above hold for $k+1$, we first show that (2.40d) holds, and then use the result to show that (2.40a), (2.40b) and then (2.40c) hold. Due to the lengthy calculations, we break the proof into four steps for convenience.

(1)

$$\begin{aligned}
\lambda_{k+1} \boldsymbol{H}_{k+1}^{\mathrm{T}} =& p^{-1/2} \tilde{\boldsymbol{E}}_2^{\mathrm{T}} \tilde{\boldsymbol{N}} - \hat{\lambda}_1 \hat{\boldsymbol{z}}_1 \hat{\boldsymbol{v}}_1^{\mathrm{T}} - \cdots - \hat{\lambda}_k \hat{\boldsymbol{z}}_k \hat{\boldsymbol{v}}_k^{\mathrm{T}} = \lambda_k \boldsymbol{H}_k - \hat{\lambda}_k \hat{\boldsymbol{z}}_k \hat{\boldsymbol{v}}_k^{\mathrm{T}} \\
=& O_{\mathrm{p}}\left\{n^{1/2}\left(\lambda_1 p\right)^{-1/2}\right\} \boldsymbol{v}_1^{\mathrm{T}} + \cdots + O_{\mathrm{p}}\left\{n^{1/2}\left(\lambda_{k-1} p\right)^{-1/2}\right\} \boldsymbol{v}_{k-1}^{\mathrm{T}} \\
&+ d_k p^{-1/2} \tilde{\boldsymbol{E}}_2^{\mathrm{T}} \boldsymbol{u}_k \boldsymbol{v}_k^{\mathrm{T}} + p^{-1/2} \tilde{\boldsymbol{E}}_2^{\mathrm{T}} \sum_{\ell=k+1}^{K} d_\ell \boldsymbol{u}_\ell \boldsymbol{v}_\ell^{\mathrm{T}} \\
&+ O_{\mathrm{p}}\left\{\lambda_{k-1}^{-1/2}\left(np^{-1}\right)^{3/2} + n^{1/2} p^{-1}\right\} - \left(\hat{\lambda}_k \lambda_k^{-1}\right) d_k p^{-1/2} \tilde{\boldsymbol{E}}_2^{\mathrm{T}} \boldsymbol{u}_k \hat{\boldsymbol{v}}_k^{\mathrm{T}} \\
&- \left(\hat{\lambda}_k \lambda_k^{-1}\right) \boldsymbol{R} d_k \lambda_k^{-1} p^{-1/2} \tilde{\boldsymbol{E}}_2^{\mathrm{T}} \boldsymbol{u}_k \hat{\boldsymbol{v}}_k^{\mathrm{T}} \\
=& O_{\mathrm{p}}\left\{n^{1/2}\left(\lambda_1 p\right)^{-1/2}\right\} \boldsymbol{v}_1^{\mathrm{T}} + \cdots + O_{\mathrm{p}}\left\{n^{1/2}\left(\lambda_k p\right)^{-1/2}\right\} \boldsymbol{v}_k^{\mathrm{T}} \\
&+ O_{\mathrm{p}}\left\{\lambda_k^{-1/2}\left(np^{-1}\right)^{3/2} + n^{1/2} p^{-1}\right\} + p^{-1/2} \tilde{\boldsymbol{E}}_2^{\mathrm{T}} \sum_{\ell=k+1}^{K} d_\ell \boldsymbol{u}_\ell \boldsymbol{v}_\ell^{\mathrm{T}}.
\end{aligned}$$

where third equality follows because

$$
\begin{aligned}
&\left(\hat{\lambda}_k \lambda_k^{-1}\right) = 1 + \rho \lambda_k^{-1} + O_{\mathrm{p}}\left\{(\lambda_k p)^{-1/2} + n\left(p\lambda_k\right)^{-1}\right\} \\
&d_k p^{-1/2} \tilde{\boldsymbol{E}}_2^{\mathrm{T}} \boldsymbol{u}_k \hat{\boldsymbol{v}}_k^{\mathrm{T}} = d_k p^{-1/2} \tilde{\boldsymbol{E}}_2^{\mathrm{T}} \boldsymbol{u}_k \boldsymbol{v}_k^{\mathrm{T}} + O_{\mathrm{p}}\left\{\lambda_k^{-1/2}\left(np^{-1}\right)^{3/2}\right\} \\
&\left(\hat{\lambda}_k \lambda_k^{-1}\right) \boldsymbol{R} d_k \lambda_k^{-1} p^{-1/2} \tilde{\boldsymbol{E}}_2^{\mathrm{T}} \boldsymbol{u}_k \hat{\boldsymbol{v}}_k^{\mathrm{T}} = O_{\mathrm{p}}\left(np^{-1}\lambda_k^{-1/2}\right) \boldsymbol{v}_k^{\mathrm{T}} + O_{\mathrm{p}}\left\{\lambda_k^{-1/2}\left(np^{-1}\right)^{3/2}\right\}.
\end{aligned}
$$

This shows (2.40d) in the inductive hypothesis also holds for $k+1$, and shows that $\|\boldsymbol{H}_{k+1}\|_2 = O_{\mathrm{p}}\left\{n^{1/2}\left(\lambda_{k+1} p\right)^{-1/2}\right\}$.

(2) We next see that

$$
\begin{aligned}
\lambda_{k+1} \hat{\boldsymbol{A}}_{k+1} = &\tilde{\boldsymbol{N}}^{\mathrm{T}} \tilde{\boldsymbol{N}} - \hat{\lambda}_1 \hat{\boldsymbol{v}}_1 \hat{\boldsymbol{v}}_1^{\mathrm{T}} - \cdots - \hat{\lambda}_k \hat{\boldsymbol{v}}_k \hat{\boldsymbol{v}}_k^{\mathrm{T}} = \tilde{\boldsymbol{N}}^{\mathrm{T}} \tilde{\boldsymbol{N}} - d_1^2 \boldsymbol{v}_1 \boldsymbol{v}_1^{\mathrm{T}} - \cdots - d_k^2 \boldsymbol{v}_k \boldsymbol{v}_k^{\mathrm{T}} \\
&+ O_{\mathrm{p}}\left(np^{-1}\right) = \lambda_{k+1} \boldsymbol{A}^{(k+1)} + O_{\mathrm{p}}\left(np^{-1}\right)
\end{aligned}
$$

(3) Lastly,

$$
\lambda_{k+1} \boldsymbol{J}_{k+1} = p^{-1} \tilde{\boldsymbol{E}}_2^{\mathrm{T}} \tilde{\boldsymbol{E}}_2 - \hat{\lambda}_1 \hat{\boldsymbol{z}}_1 \hat{\boldsymbol{z}}_1^{\mathrm{T}} - \cdots - \hat{\lambda}_k \hat{\boldsymbol{z}}_k \hat{\boldsymbol{z}}_k^{\mathrm{T}} = p^{-1} \tilde{\boldsymbol{E}}_2^{\mathrm{T}} \tilde{\boldsymbol{E}}_2 + O_{\mathrm{p}}\left(np^{-1}\right).
$$

By the above expressions for $\hat{\boldsymbol{A}}_{k+1}$, $\boldsymbol{H}_{k+1}$ and $\boldsymbol{J}_{k+1}$,

$$
\begin{aligned}
\left(\hat{\lambda}_{k+1} - \rho\right) \lambda_{k+1}^{-1} &= \left(d_{k+1}^2 - \rho\right) \lambda_{k+1}^{-1} + O_{\mathrm{p}}\left\{n^{1/2}\left(\lambda_{k+1} p\right)^{-1/2}\right\} \\
&= 1 + O_{\mathrm{p}}\left\{n^{1/2}\left(\lambda_{k+1} p\right)^{-1/2}\right\}
\end{aligned}
$$

by Weyl's Theorem. Therefore,

$$
\hat{\lambda}_{k+1} \lambda_{k+1}^{-1} I_{n-K} - \boldsymbol{J}_{k+1} = \left(\hat{\lambda}_{k+1} - \rho\right) \lambda_{k+1}^{-1} I_{n-K} - \lambda_{k+1}^{-1} \boldsymbol{R} + O_{\mathrm{p}}\left\{n\left(\lambda_{k+1} p\right)^{-1}\right\}
$$

is invertible with high probability.

(4) We can then put parts 1, 2 and 3 together to find expressions for the eigenvalue $\hat{\lambda}_{k+1}$ and components of the eigenvector $\hat{\boldsymbol{v}}_{k+1}$, $\hat{\boldsymbol{z}}_{k+1}$.

(a)

$$\begin{aligned}\hat{\lambda}_{k+1}\lambda_{k+1}^{-1}\hat{\boldsymbol{v}}_{k+1} &= \hat{\boldsymbol{A}}_{k+1}\hat{\boldsymbol{v}}_{k+1} + \boldsymbol{H}_{k+1}^{\mathrm{T}}\left(\hat{\lambda}_{k+1}\lambda_{k+1}^{-1}I_{n-K} - \boldsymbol{J}_{k+1}\right)^{-1}\boldsymbol{H}_{k+1}\hat{\boldsymbol{v}}_{k+1}\\ &= \boldsymbol{A}^{(k+1)}\hat{\boldsymbol{v}}_{k+1} + O_{\mathrm{p}}\left\{n\left(\lambda_{k+1}p\right)^{-1}\right\}\end{aligned}$$

(b)

$$\begin{aligned}\hat{\boldsymbol{z}}_{k+1} &= \left(\hat{\lambda}_{k+1}\lambda_{k+1}^{-1}I_{n-K} - \boldsymbol{J}_{k+1}\right)\boldsymbol{H}_{k+1}\hat{\boldsymbol{v}}_{k+1}\\ &= \left[\left(\hat{\lambda}_{k+1} - \rho\right)\lambda_{k+1}^{-1}I_{n-K} - \lambda_{k+1}^{-1}\boldsymbol{R} + O_{\mathrm{p}}\left\{n\left(\lambda_{k+1}p\right)^{-1}\right\}\right]^{-1}\lambda_{k+1}^{-1}p^{-1/2}\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\times\\ &\quad\times\sum_{\ell=k+1}^{K} d_\ell\boldsymbol{u}_\ell\boldsymbol{v}_\ell^{\mathrm{T}}\hat{\boldsymbol{v}}_{k+1} + O_{\mathrm{p}}\left(\lambda_{k+1}^{-1}\lambda_1^{-1/2}n^{1/2}p^{-1/2}\right)\boldsymbol{v}_1^{\mathrm{T}}\hat{\boldsymbol{v}}_{k+1} + \cdots\\ &\quad + O_{\mathrm{p}}\left(\lambda_{k+1}^{-1}\lambda_k^{-1/2}n^{1/2}p^{-1/2}\right)\boldsymbol{v}_k^{\mathrm{T}}\hat{\boldsymbol{v}}_{k+1} + O_{\mathrm{p}}\left\{\lambda_k^{-1/2}\lambda_{k+1}^{-1}\left(np^{-1}\right)^{3/2}\right\}\\ &\quad + O_{\mathrm{p}}\left\{n^{1/2}\left(\lambda_{k+1}p\right)^{-1}\right\}\end{aligned}$$

Therefore

$$\|\hat{\boldsymbol{z}}_{k+1}\|_2 = O_{\mathrm{p}}\left\{n^{1/2}\left(\lambda_{k+1}p\right)^{-1/2}\right\},$$

meaning

$$\|\hat{\boldsymbol{v}}_{k+1}\|_2 = 1 - O_{\mathrm{p}}\left\{n\left(\lambda_{k+1}p\right)^{-1}\right\}.$$

We can then use this and what we showed in part a. to get that

$$\hat{\boldsymbol{v}}_{k+1} = \boldsymbol{v}_{k+1} + O_{\mathrm{p}}\left\{n\left(\lambda_{k+1}p\right)^{-1}\right\}$$
$$\hat{\lambda}_{k+1} = d_{k+1}^2 + O_{\mathrm{p}}\left(np^{-1}\right)$$

which means the 1. of the inductive hypothesis applies for $k+1$. Using the fact that for any $s \leq k$

$$\boldsymbol{v}_s^{\mathrm{T}}\hat{\boldsymbol{v}}_{k+1} = O_{\mathrm{p}}\left\{n\left(p\lambda_{k+1}\right)^{-1}\right\}$$
$$\left(\hat{\lambda}_{k+1} - \rho\right)\lambda_{k+1}^{-1} = 1 + O_{\mathrm{p}}\left\{\left(\lambda_{k+1}p\right)^{-1/2} + n\left(\lambda_{k+1}p\right)^{-1}\right\},$$

we can then modify our expression for $\hat{\boldsymbol{z}}_{k+1}$ to get:

(c)

$$\begin{aligned}
\hat{\boldsymbol{z}}_{k+1} =& \lambda_{k+1}^{-1}p^{-1/2}\left\{\left(\hat{\lambda}_{k+1} - \rho\right)\lambda_{k+1}^{-1}I_{n-K} - \lambda_{k+1}^{-1}\boldsymbol{R}\right\}^{-1}\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\sum_{\ell=k+1}^{K} d_\ell \boldsymbol{u}_\ell \boldsymbol{v}_\ell^{\mathrm{T}}\hat{\boldsymbol{v}}_{k+1} \\
&+ O_{\mathrm{p}}\left\{\left(\lambda_{k+1}^2\lambda_1^{1/2}\right)^{-1}\left(np^{-1}\right)^{3/2}\right\} + \cdots + O_{\mathrm{p}}\left\{\left(\lambda_{k+1}^2\lambda_k^{1/2}\right)^{-1}\left(np^{-1}\right)^{3/2}\right\} \\
&+ O_{\mathrm{p}}\left\{n^{3/2}\left(p\lambda_{k+1}\right)^{-3/2} + n^{1/2}\left(\lambda_{k+1}p\right)^{-1}\right\} \\
=& \lambda_{k+1}^{-1}p^{-1/2}\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\sum_{\ell=k+1}^{K} d_\ell \boldsymbol{u}_\ell \boldsymbol{v}_\ell^{\mathrm{T}}\hat{\boldsymbol{v}}_{k+1} + \lambda_{k+1}^{-2}p^{-1/2}\boldsymbol{R}\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\sum_{\ell=k+1}^{K} d_\ell \boldsymbol{u}_\ell \boldsymbol{v}_\ell^{\mathrm{T}}\hat{\boldsymbol{v}}_{k+1} \\
&+ O_{\mathrm{p}}\left\{n^{3/2}\left(p\lambda_{k+1}\right)^{-3/2} + n^{1/2}\left(\lambda_{k+1}p\right)^{-1}\right\} \\
=& d_{k+1}\lambda_{k+1}^{-1}p^{-1/2}\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\boldsymbol{u}_{k+1} + d_{k+1}\lambda_{k+1}^{-2}p^{-1/2}\boldsymbol{R}\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\boldsymbol{u}_{k+1} \\
&+ O_{\mathrm{p}}\left\{n^{3/2}\left(p\lambda_{k+1}\right)^{-3/2} + n^{1/2}\left(\lambda_{k+1}p\right)^{-1}\right\}.
\end{aligned}$$

This completes the proof by induction and therefore proves (2.36), (2.37) and (2.38). It remains to show (2.39).

Since $\boldsymbol{F}$ is symmetric with distinct eigenvalues (with probability 1), for $s < k$ (i.e.

$\lambda_s > \lambda_k$),

$$0 = \hat{\boldsymbol{v}}_s^{\mathrm{T}} \hat{\boldsymbol{v}}_k + \hat{\boldsymbol{z}}_s^{\mathrm{T}} \hat{\boldsymbol{z}}_k = \left(\boldsymbol{v}_s + \boldsymbol{\epsilon}_s\right)^{\mathrm{T}} \left(\boldsymbol{v}_k + \boldsymbol{\epsilon}_k\right) + \hat{\boldsymbol{z}}_s^{\mathrm{T}} \hat{\boldsymbol{z}}_k = 0 + \boldsymbol{\epsilon}_s^{\mathrm{T}} \hat{\boldsymbol{v}}_k + \boldsymbol{v}_s^{\mathrm{T}} \boldsymbol{\epsilon}_k + \hat{\boldsymbol{z}}_s^{\mathrm{T}} \hat{\boldsymbol{z}}_k.$$

where

$$\begin{aligned}
\boldsymbol{\epsilon}_s^{\mathrm{T}} \hat{\boldsymbol{v}}_k &= O_{\mathrm{p}} \left\{ n \left(p\lambda_s\right)^{-1} \right\} = o_{\mathrm{p}} \left(\lambda_k \lambda_s^{-1} n^{-1/2} \right) \\
\boldsymbol{v}_s^{\mathrm{T}} \boldsymbol{\epsilon}_k &= \boldsymbol{\epsilon}_k[s] + O_{\mathrm{p}} \left\{ \left(\lambda_s p\right)^{-1/2} n \left(p\lambda_k\right)^{-1} + n \left(p^2 \lambda_s \lambda_k\right)^{-1} \right\} = \boldsymbol{\epsilon}_k[s] + o_{\mathrm{p}} \left(\lambda_k \lambda_s^{-1} n^{-1/2} \right)
\end{aligned}$$

Therefore, if $\hat{\boldsymbol{z}}_s^{\mathrm{T}} \hat{\boldsymbol{z}}_k = o_{\mathrm{p}} \left(\lambda_k \lambda_s^{-1} n^{-1/2} \right)$, we must have $\boldsymbol{\epsilon}_k[s] = o_{\mathrm{p}} \left(\lambda_k \lambda_s^{-1} n^{-1/2} \right)$. By our above expression for $\hat{\boldsymbol{z}}_k$,

$$\begin{aligned}
\hat{\boldsymbol{z}}_s^{\mathrm{T}} \hat{\boldsymbol{z}}_k = &\left[d_s \lambda_s^{-1} p^{-1/2} \tilde{\boldsymbol{E}}_2^{\mathrm{T}} \boldsymbol{u}_s + d_s \lambda_s^{-2} p^{-1/2} \boldsymbol{R} \tilde{\boldsymbol{E}}_2^{\mathrm{T}} \boldsymbol{u}_s + O_{\mathrm{p}} \left\{ n^{3/2} \left(p\lambda_s\right)^{-3/2} + n^{1/2} \left(\lambda_s p\right)^{-1} \right\} \right]^{\mathrm{T}} \\
&\times \left[d_k \lambda_k^{-1} p^{-1/2} \tilde{\boldsymbol{E}}_2^{\mathrm{T}} \boldsymbol{u}_k + d_k \lambda_k^{-2} p^{-1/2} \boldsymbol{R} \tilde{\boldsymbol{E}}_2^{\mathrm{T}} \boldsymbol{u}_k + O_{\mathrm{p}} \left\{ n^{3/2} \left(p\lambda_k\right)^{-3/2} + n^{1/2} \left(\lambda_k p\right)^{-1} \right\} \right]
\end{aligned}$$

We see that

$$\begin{aligned}
&O_{\mathrm{p}} \left\{ n^{3/2} \left(p\lambda_k\right)^{-3/2} + n^{1/2} \left(\lambda_k p\right)^{-1} \right\} \|\hat{\boldsymbol{z}}_s\|_2 \\
=&O_{\mathrm{p}} \left\{ \left(np^{-1}\right)^2 \lambda_k^{-3/2} \lambda_s^{-1/2} + n \left(p\lambda_k\right)^{-1} \left(p\lambda_s\right)^{-1/2} \right\} = o_{\mathrm{p}} \left(\lambda_k \lambda_s^{-1} n^{-1/2} \right)
\end{aligned}$$

$$\begin{aligned}
\| d_s \lambda_s^{-2} p^{-1/2} \boldsymbol{R} \tilde{\boldsymbol{E}}_2^{\mathrm{T}} \boldsymbol{u}_s + O_{\mathrm{p}} \left\{ n^{3/2} \left(p\lambda_s\right)^{-3/2} + n^{1/2} \left(\lambda_s p\right)^{-1} \right\} \|_2 \|\hat{\boldsymbol{z}}_k\|_2 &= O_{\mathrm{p}} \left\{ n \left(p\lambda_s\right)^{-1} \right\} \\
&= o_{\mathrm{p}} \left(\lambda_k \lambda_s^{-1} n^{-1/2} \right)
\end{aligned}$$

Therefore,

$$\hat{\boldsymbol{z}}_s^{\mathrm{T}} \hat{\boldsymbol{z}}_k = \left(d_s \lambda_s^{-1} p^{-1/2} \tilde{\boldsymbol{E}}_2^{\mathrm{T}} \boldsymbol{u}_s\right)^{\mathrm{T}} \left(d_k \lambda_k^{-1} p^{-1/2} \tilde{\boldsymbol{E}}_2^{\mathrm{T}} \boldsymbol{u}_k + d_k \lambda_k^{-2} p^{-1/2} \boldsymbol{R} \tilde{\boldsymbol{E}}_2^{\mathrm{T}} \boldsymbol{u}_k\right)$$
$$+ o_{\mathrm{p}}\left(\lambda_k \lambda_s^{-1} n^{-1/2}\right) = d_s d_k \left(\lambda_s \lambda_k p\right)^{-1} \boldsymbol{u}_s^{\mathrm{T}} \tilde{\boldsymbol{E}}_2 \tilde{\boldsymbol{E}}_2^{\mathrm{T}} \boldsymbol{u}_k$$
$$+ d_s d_k \left(\lambda_s \lambda_k^2 p\right)^{-1} \boldsymbol{u}_s^{\mathrm{T}} \tilde{\boldsymbol{E}}_2 \boldsymbol{R} \tilde{\boldsymbol{E}}_2^{\mathrm{T}} \boldsymbol{u}_k + o_{\mathrm{p}}\left(\lambda_k \lambda_s^{-1} n^{-1/2}\right).$$

We analyze the two terms in the above equation in 1. and 2. below.

(1) Define

$$\boldsymbol{U}_{s,k} = \left(\boldsymbol{u}_s\, \boldsymbol{u}_k\right), \quad \boldsymbol{W} = \left(\boldsymbol{U}_{s,k}^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{U}_{s,k}\right)^{1/2}, \quad \boldsymbol{M} \sim MN_{(n-K)\times 2}\left(0, I_{n-K}, I_2\right).$$

Then

$$\begin{aligned} d_s d_k \left(\lambda_s \lambda_k p\right)^{-1} \boldsymbol{u}_s^{\mathrm{T}} \tilde{\boldsymbol{E}}_2 \tilde{\boldsymbol{E}}_2^{\mathrm{T}} \boldsymbol{u}_k &= \left[\boldsymbol{U}_{s,k}^{\mathrm{T}} \tilde{\boldsymbol{E}}_2 \tilde{\boldsymbol{E}}_2^{\mathrm{T}} \boldsymbol{U}_{s,k}\right]_{1,2} \\ &\overset{\mathcal{D}}{=} \left[d_s d_k n \left(\lambda_s \lambda_k p\right)^{-1} \boldsymbol{W} \left(n^{-1} \boldsymbol{M}^{\mathrm{T}} \boldsymbol{M}\right) \boldsymbol{W}\right]_{1,2} \\ &= d_s d_k n \left(\lambda_s \lambda_k p\right)^{-1} \left[\boldsymbol{W}^2 + O_{\mathrm{p}}\left(n^{-1/2}\right)\right]_{1,2} \\ &= d_s d_k n \left(\lambda_s \lambda_k p\right)^{-1} \boldsymbol{u}_s^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{u}_k + O_{\mathrm{p}}\left(n^{1/2} \lambda_s^{-1/2} \lambda_k^{-1/2} p^{-1}\right) \\ &= d_s d_k n \left(\lambda_s \lambda_k p\right)^{-1} \boldsymbol{u}_s^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{u}_k + o_{\mathrm{p}}\left(\lambda_k \lambda_s^{-1} n^{-1/2}\right). \end{aligned}$$

If $\boldsymbol{\Sigma} = \sigma^2 I_p$, we would be done. However, if $\boldsymbol{\Sigma}$ were arbitrary then under no assumptions $\boldsymbol{u}_s^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{u}_k = O_{\mathrm{p}}(1)$, meaning $d_s d_k n \left(\lambda_s \lambda_k p\right)^{-1} \boldsymbol{u}_s^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{u}_k = O_{\mathrm{p}}\left(\lambda_s^{-1/2} \lambda_k^{-1/2} n p^{-1}\right)$ which is not necessarily $o_{\mathrm{p}}\left(\lambda_k \lambda_s^{-1} n^{-1/2}\right)$. To see this, if $\lambda_s = n$ and $\lambda_k = 1$ then $O_{\mathrm{p}}\left(\lambda_s^{-1/2} \lambda_k^{-1/2} n p^{-1}\right) = O_{\mathrm{p}}\left(n^{1/2} p^{-1}\right)$, which is not $o_{\mathrm{p}}\left(\lambda_k \lambda_s^{-1} n^{-1/2}\right)$. We will use the assumption that $n p^{-1} \left[\boldsymbol{L}\right]_{*k}^{\mathrm{T}} \boldsymbol{\Sigma} \left[\boldsymbol{L}\right]_{*s} = O_{\mathrm{p}}\left\{\lambda_{\max(k,s)}\right\}$ in the statement of the lemma to show that $\boldsymbol{u}_s^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{u}_k = O_{\mathrm{p}}\left(\lambda_k^{1/2} \lambda_s^{-1/2}\right)$. If this were the case, we would have $d_s d_k n \left(\lambda_s \lambda_k p\right)^{-1} \boldsymbol{u}_s^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{u}_k = O_{\mathrm{p}}\left\{n \left(\lambda_s p\right)^{-1}\right\} = o_{\mathrm{p}}\left(\lambda_k \lambda_s^{-1} n^{-1/2}\right)$. Lemma

2.8 in Section 2.7.10 proves $\boldsymbol{u}_s^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{u}_k = O_{\mathrm{p}}\left(\lambda_k^{1/2}\lambda_s^{-1/2}\right)$ under the assumption that $np^{-1}\left[\boldsymbol{L}\right]_{*k}^{\mathrm{T}}\boldsymbol{\Sigma}\left[\boldsymbol{L}\right]_{*s} = O_{\mathrm{p}}\left\{\lambda_{\max(k,s)}\right\}$.

(2) Recall that $\boldsymbol{R} = p^{-1}\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\tilde{\boldsymbol{E}}_2 - \rho I_{n-K}$. In Lemma 2.9 in Section 2.7.10, we prove

$$p^{-1}\boldsymbol{u}_s^{\mathrm{T}}\tilde{\boldsymbol{E}}_2\boldsymbol{R}\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\boldsymbol{u}_k = O_{\mathrm{p}}\left\{\left(np^{-1}\right)^2 + np^{-3/2}\right\}.$$

This will then imply

$$d_s d_k\left(\lambda_s\lambda_k^2 p\right)^{-1}\boldsymbol{u}_s^{\mathrm{T}}\tilde{\boldsymbol{E}}_2\boldsymbol{R}\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\boldsymbol{u}_k = o_{\mathrm{p}}\left(\lambda_k\lambda_s^{-1}n^{-1/2}\right).$$

This proves (2.39) and completes the proof.

□

2.7.5 Proof of Lemmas 2.1 and 2.2

In this section, we prove Lemmas 2.1 and 2.2. To do so, we first prove a modified version of Lemma 2.2 in which we modify (2.18) to be

$$n^{1/2}\left(\hat{\boldsymbol{\ell}}_g - \boldsymbol{\ell}_g\right) \stackrel{\mathcal{D}}{=} \sigma_g\boldsymbol{W} + o_{\mathrm{p}}(1),$$

where $\boldsymbol{W} \sim N_K\left(0, I_K\right)$. We then prove Lemma 2.1. (2.18) from Lemma 2.2 then follows. For ease of notation, we use the definition of $\boldsymbol{Y}$ from Section 2.7.4 defined in (2.28).

Proof of Lemma 2.2. We first note that (2.17) is a direct consequence of (2.32) in Lemma 2.4 and (2.36) in Lemma 2.5. It therefore remains to prove (2.18), the asymptotic distribution of $\hat{\boldsymbol{\ell}}_g$. Define $\boldsymbol{y}_g$ and $\tilde{\boldsymbol{e}}_{i,g}$ to be the g^{th} rows of $\boldsymbol{Y}$ and $\tilde{\boldsymbol{E}}_i$ $(i = 1, 2)$.

$$n^{1/2}\hat{\boldsymbol{\ell}}_g = \hat{\tilde{\boldsymbol{C}}}^{\mathrm{T}}\boldsymbol{y}_g = \left(\hat{\boldsymbol{V}}^{\mathrm{T}}\tilde{\boldsymbol{C}}^{\mathrm{T}} + \hat{\boldsymbol{Z}}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\right)\boldsymbol{y}_g = n^{1/2}\hat{\boldsymbol{V}}^{\mathrm{T}}\boldsymbol{\ell}_g + \hat{\boldsymbol{V}}^{\mathrm{T}}\tilde{\boldsymbol{e}}_{1,g} + \hat{\boldsymbol{Z}}^{\mathrm{T}}\tilde{\boldsymbol{e}}_{2,g}.$$

We then have

$$\begin{aligned}
n^{1/2}\hat{\boldsymbol{V}}^{\mathrm{T}}\boldsymbol{\ell}_g =& n^{1/2}\boldsymbol{\ell}_g + n^{1/2}O_{\mathrm{p}}\left\{(p\lambda_K)^{-1/2} + n\,(p\lambda_K)^{-1}\right\}\\
\hat{\boldsymbol{V}}^{\mathrm{T}}\tilde{\boldsymbol{e}}_{1,g} \sim& N\left(0,\sigma_g^2 I_K\right) + O_{\mathrm{p}}\left\{(p\lambda_K)^{-1/2} + n\,(p\lambda_K)^{-1}\right\}\\
\hat{\boldsymbol{z}}_k^{\mathrm{T}}\tilde{\boldsymbol{e}}_{2,g} =& d_k\lambda_k^{-1}p^{-1/2}\boldsymbol{u}_k[g]\tilde{\boldsymbol{e}}_{2,g}^{\mathrm{T}}\tilde{\boldsymbol{e}}_{2,g} + d_k\lambda_k^{-1}p^{-1/2}\boldsymbol{u}_k[-g]^{\mathrm{T}}\tilde{\boldsymbol{E}}_2[-g,]\tilde{\boldsymbol{e}}_{2,g}\\
&+ O_{\mathrm{p}}\left\{n^{3/2}\,(p\lambda_k)^{-1}\right\}
\end{aligned}$$

where

$$\boldsymbol{u}_k[g] = n^{1/2}d_k^{-1}p^{-1/2}\left(\boldsymbol{\ell}_g + n^{-1/2}\tilde{\boldsymbol{e}}_{1,g}\right)^{\mathrm{T}}\boldsymbol{v}_k = O_{\mathrm{p}}\left(n^{1/2}\lambda_k^{-1/2}p^{-1/2}\right).$$

Therefore, $d_k\lambda_k^{-1}p^{-1/2}\boldsymbol{u}_k[g]\tilde{\boldsymbol{e}}_{2,g}^{\mathrm{T}}\tilde{\boldsymbol{e}}_{2,g} = O_{\mathrm{p}}\left\{n^{3/2}\,(p\lambda_k)^{-1}\right\}$. Lastly,

$$\boldsymbol{u}_k[-g]^{\mathrm{T}}\tilde{\boldsymbol{E}}_2[-g,]\tilde{\boldsymbol{e}}_{2,g} \sim N\left(0,\boldsymbol{u}_k[-g]^{\mathrm{T}}\boldsymbol{\Sigma}[-g]\boldsymbol{u}_k[-g]\tilde{\boldsymbol{e}}_{2,g}^{\mathrm{T}}\tilde{\boldsymbol{e}}_{2,g}\right) = O_{\mathrm{p}}\left(n^{1/2}\right).$$

Therefore, $\hat{\boldsymbol{Z}}^{\mathrm{T}}\tilde{\boldsymbol{e}}_{2,g} = O_{\mathrm{p}}\left\{n^{3/2}\,(p\lambda_K)^{-1} + n^{1/2}\,(p\lambda_K)^{-1/2}\right\}$, which means $n^{1/2}\left(\hat{\boldsymbol{\ell}}_g - \boldsymbol{\ell}_g\right) \xrightarrow{\mathcal{D}} N_K\left(0,\sigma_g^2 I_K\right)$.

We also note that this also shows that

$$n^{1/2}\|\hat{\boldsymbol{\ell}}_g^{\mathrm{OLS}} - \hat{\boldsymbol{\ell}}_g\|_2 = o_{\mathrm{p}}(1)$$

where $\hat{\boldsymbol{\ell}}_g^{\mathrm{OLS}} = \boldsymbol{\ell}_g + n^{-1/2}\tilde{\boldsymbol{e}}_{1,g}$ is the ordinary least squares estimate for $\boldsymbol{\ell}_g$ when $\boldsymbol{C}$ is known, since

$$n^{1/2}\|\hat{\boldsymbol{V}}^{\mathrm{T}}\boldsymbol{\ell}_g - \boldsymbol{\ell}_g\|_2, \|\hat{\boldsymbol{V}}^{\mathrm{T}}\tilde{\boldsymbol{e}}_{1,g} - \tilde{\boldsymbol{e}}_{1,g}\|_2 = o_{\mathrm{p}}(1).$$

□

Proof of Lemma 2.1. Once we estimate $\boldsymbol{C}$ by singular value decomposition, we let

$$\hat{\sigma}_g^2 = (n-K)^{-1}\boldsymbol{y}_g^{\mathrm{T}} P_{\hat{C}}^{\perp} \boldsymbol{y}_g$$

for each site $g = 1, \ldots, p$. We will prove (2.15) and (2.16) by showing the following:

(a) $\hat{\sigma}_g^2 = \sigma_g^2 + O_{\mathrm{p}}\left\{n^{-1/2} + n^{1/2}(p\lambda_K)^{-1/2}\right\} = \sigma_g^2 + o_{\mathrm{p}}(1)$.

(b) $\hat{\rho} = p^{-1}\sum_{g=1}^{p}\hat{\sigma}_g^2 = \rho + O_{\mathrm{p}}\left\{(p\lambda_K)^{-1/2} + n\,(p\lambda_K)^{-1}\right\} = \rho + o_{\mathrm{p}}\left(n^{-1/2}\right)$.

We first define the estimated scaled covariates $\hat{\boldsymbol{W}} = n^{-1/2}\hat{\boldsymbol{C}} = \tilde{\boldsymbol{C}}\hat{\boldsymbol{V}} + \boldsymbol{Q}\hat{\boldsymbol{Z}} \in \mathbb{R}^{n\times K}$, where $\hat{\boldsymbol{V}}$, $\hat{\boldsymbol{Z}}$, $\tilde{\boldsymbol{C}}^{\mathrm{T}}$ and $\boldsymbol{Q}^{\mathrm{T}}$ are given in Lemmas 2.4 and 2.5. Also, define $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1 \cdots \boldsymbol{\epsilon}_K)$, where $\boldsymbol{\epsilon}_k \in \mathbb{R}^K$ is as defined in (2.37) of Lemma 2.5. We then see that

$$\begin{aligned}
(n-K)\,\hat{\sigma}_g^2 &= \boldsymbol{y}_g^{\mathrm{T}}\boldsymbol{y}_g - \boldsymbol{y}_g^{\mathrm{T}}P_{\hat{\boldsymbol{W}}}\boldsymbol{y}_g = \boldsymbol{y}_g^{\mathrm{T}}\boldsymbol{y}_g - \boldsymbol{y}_g^{\mathrm{T}}\hat{\boldsymbol{W}}\hat{\boldsymbol{W}}^{\mathrm{T}}\boldsymbol{y}_g \\
&= \boldsymbol{y}_g^{\mathrm{T}}\boldsymbol{y}_g - \boldsymbol{y}_g^{\mathrm{T}}\left(\tilde{\boldsymbol{C}}\hat{\boldsymbol{V}} + \boldsymbol{Q}\hat{\boldsymbol{Z}}\right)\left(\hat{\boldsymbol{V}}^{\mathrm{T}}\tilde{\boldsymbol{C}}^{\mathrm{T}} + \hat{\boldsymbol{Z}}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\right)\boldsymbol{y}_g \\
&= \underbrace{\left(\boldsymbol{y}_g^{\mathrm{T}}\boldsymbol{y}_g - \boldsymbol{y}_g^{\mathrm{T}}\tilde{\boldsymbol{C}}\hat{\boldsymbol{V}}\hat{\boldsymbol{V}}^{\mathrm{T}}\tilde{\boldsymbol{C}}^{\mathrm{T}}\boldsymbol{y}_g\right)}_{(1)} - 2\underbrace{\boldsymbol{y}_g^{\mathrm{T}}\tilde{\boldsymbol{C}}\hat{\boldsymbol{V}}\hat{\boldsymbol{Z}}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{y}_g}_{(2)} - \underbrace{\boldsymbol{y}_g^{\mathrm{T}}\boldsymbol{Q}\hat{\boldsymbol{Z}}\hat{\boldsymbol{Z}}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{y}_g}_{(3)}
\end{aligned}$$

We define $\tilde{\boldsymbol{e}}_{1,g}$ and $\tilde{\boldsymbol{e}}_{2,g}$ to be the gth rows of $\tilde{\boldsymbol{E}}_1$ and $\tilde{\boldsymbol{E}}_2$, respectively, and derive the asymptotic properties of (1), (2) and (3) to show (a) and (b) above.

(1)

$$\boldsymbol{y}_g^{\mathrm{T}}\boldsymbol{y}_g - \boldsymbol{y}_g^{\mathrm{T}}\tilde{\boldsymbol{C}}\hat{\boldsymbol{V}}\hat{\boldsymbol{V}}^{\mathrm{T}}\tilde{\boldsymbol{C}}^{\mathrm{T}}\boldsymbol{y}_g = \underbrace{\boldsymbol{y}_g^{\mathrm{T}}\boldsymbol{y}_g - \boldsymbol{y}_g^{\mathrm{T}}\tilde{\boldsymbol{C}}\tilde{\boldsymbol{C}}^{\mathrm{T}}\boldsymbol{y}_g}_{\text{(i)}} + 2\underbrace{\boldsymbol{y}_g^{\mathrm{T}}\tilde{\boldsymbol{C}}\boldsymbol{\delta}^{\mathrm{T}}\tilde{\boldsymbol{C}}^{\mathrm{T}}\boldsymbol{y}_g}_{\text{(ii)}} + \underbrace{\boldsymbol{y}_g^{\mathrm{T}}\tilde{\boldsymbol{C}}\boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{\delta}\tilde{\boldsymbol{C}}^{\mathrm{T}}\boldsymbol{y}_g}_{\text{(iii)}}$$

where $\boldsymbol{\delta} = \hat{\boldsymbol{V}} - I_K$.

(a)

$$(n-K)^{-1}\left(\boldsymbol{y}_g^{\mathrm{T}}\boldsymbol{y}_g - \boldsymbol{y}_g^{\mathrm{T}}\tilde{\boldsymbol{C}}\hat{\boldsymbol{V}}\hat{\boldsymbol{V}}^{\mathrm{T}}\tilde{\boldsymbol{C}}^{\mathrm{T}}\boldsymbol{y}_g\right) = \hat{\sigma}_{g,\mathrm{OLS}}^2 + O_{\mathrm{p}}\left\{(\lambda_k p)^{-1/2} + n\left(\lambda_k p\right)^{-1}\right\}$$
$$= \sigma_g^2 + O_{\mathrm{p}}\left(n^{-1/2}\right)$$

(b) (i)

$$(n-K)^{-1}p^{-1}\sum_{g=1}^{p}\left(\boldsymbol{y}_g^{\mathrm{T}}\boldsymbol{y}_g - \boldsymbol{y}_g^{\mathrm{T}}\tilde{\boldsymbol{C}}\tilde{\boldsymbol{C}}^{\mathrm{T}}\boldsymbol{y}_g\right) = p^{-1}\sum_{g=1}^{p}\hat{\sigma}_{g,\mathrm{OLS}}^2$$
$$= \rho + O_{\mathrm{p}}\left\{(np)^{-1/2}\right\}$$

(ii)

$$|(np)^{-1}\sum_{g=1}^{p}\boldsymbol{y}_g^{\mathrm{T}}\tilde{\boldsymbol{C}}\boldsymbol{\delta}^{\mathrm{T}}\tilde{\boldsymbol{C}}^{\mathrm{T}}\boldsymbol{y}_g| \leq \|\boldsymbol{\delta}\|_2 p^{-1}\sum_{g=1}^{p}\left(\boldsymbol{\ell}_g + n^{-1/2}\tilde{\boldsymbol{e}}_{1,g}\right)^{\mathrm{T}}\left(\boldsymbol{\ell}_g + n^{-1/2}\tilde{\boldsymbol{e}}_{1,g}\right)$$
$$= O_{\mathrm{p}}\left\{(\lambda_k p)^{-1/2} + n\left(\lambda_k p\right)^{-1}\right\}$$

(iii)

$$(np)^{-1}\sum_{g=1}^{p}\boldsymbol{y}_g^{\mathrm{T}}\tilde{\boldsymbol{C}}\boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{\delta}\tilde{\boldsymbol{C}}^{\mathrm{T}}\boldsymbol{y}_g = o_{\mathrm{p}}\left\{(\lambda_k p)^{-1/2} + n\left(\lambda_k p\right)^{-1}\right\}$$

(2)

$$(n-K)^{-1}\boldsymbol{y}_g^{\mathrm{T}}\boldsymbol{Q}\hat{\boldsymbol{Z}}\hat{\boldsymbol{Z}}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{y}_g = (n-K)^{-1}\tilde{\boldsymbol{e}}_{2,g}^{\mathrm{T}}\hat{\boldsymbol{Z}}\hat{\boldsymbol{Z}}^{\mathrm{T}}\tilde{\boldsymbol{e}}_{2,g} \leq \|\hat{\boldsymbol{Z}}\|_2^2(n-K)^{-1}\tilde{\boldsymbol{e}}_{2,g}^{\mathrm{T}}\tilde{\boldsymbol{e}}_{2,g}$$

where $\|\hat{\boldsymbol{Z}}\|_2^2 = O_{\mathrm{p}}\left\{n\left(\lambda_K p\right)^{-1}\right\}$ and $(n-K)^{-1}\tilde{\boldsymbol{e}}_{2,g}^{\mathrm{T}}\tilde{\boldsymbol{e}}_{2,g} = O_{\mathrm{p}}(1)$.

(a)

$$(n-K)^{-1}\boldsymbol{y}_g^{\mathrm{T}}\boldsymbol{Q}\hat{\boldsymbol{Z}}\hat{\boldsymbol{Z}}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{y}_g = O_{\mathrm{p}}\left\{n\left(\lambda_K p\right)^{-1}\right\}.$$

(b)

$$
\begin{aligned}
(n-K)^{-1}p^{-1}\sum_{g=1}^{p}\boldsymbol{y}_g^{\mathrm{T}}\boldsymbol{Q}\hat{\boldsymbol{Z}}\hat{\boldsymbol{Z}}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{y}_g &\leq \|\hat{\boldsymbol{Z}}\|_2^2 p^{-1}\sum_{g=1}^{p}(n-K)^{-1}\tilde{\boldsymbol{e}}_{2,g}^{\mathrm{T}}\tilde{\boldsymbol{e}}_{2,g} \\
&= O_{\mathrm{p}}\left\{n\left(\lambda_K p\right)^{-1}\right\}.
\end{aligned}
$$

(3)

$$
\begin{aligned}
n^{-1}\boldsymbol{y}_g^{\mathrm{T}}\tilde{\boldsymbol{C}}\hat{\boldsymbol{V}}\hat{\boldsymbol{Z}}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{y}_g &= \left(\boldsymbol{\ell}_g + n^{-1/2}\tilde{\boldsymbol{e}}_{1,g}\right)^{\mathrm{T}}\hat{\boldsymbol{V}}\hat{\boldsymbol{Z}}^{\mathrm{T}}n^{-1/2}\tilde{\boldsymbol{e}}_{2,g} \\
&= \underbrace{\left(\boldsymbol{\ell}_g + n^{-1/2}\tilde{\boldsymbol{e}}_{1,g}\right)^{\mathrm{T}} V\hat{\boldsymbol{Z}}^{\mathrm{T}}n^{-1/2}\tilde{\boldsymbol{e}}_{2,g}}_{\text{(i)}} \\
&\quad + \underbrace{\left(\boldsymbol{\ell}_g + n^{-1/2}\tilde{\boldsymbol{e}}_{1,g}\right)^{\mathrm{T}}\boldsymbol{\epsilon}\hat{\boldsymbol{Z}}^{\mathrm{T}}n^{-1/2}\tilde{\boldsymbol{e}}_{2,g}}_{\text{(ii)}}.
\end{aligned}
$$

(a)

$$
\begin{aligned}
|\left(\boldsymbol{\ell}_g + n^{-1/2}\tilde{\boldsymbol{e}}_{1,g}\right)^{\mathrm{T}}\hat{\boldsymbol{V}}\hat{\boldsymbol{Z}}^{\mathrm{T}}n^{-1/2}\tilde{\boldsymbol{e}}_{2,g}| &\leq \|\left(\boldsymbol{\ell}_g + n^{-1/2}\tilde{\boldsymbol{e}}_{1,g}\right)^{\mathrm{T}}\hat{\boldsymbol{V}}\|_2\|\hat{\boldsymbol{Z}}^{\mathrm{T}}\|_2\|n^{-1/2}\tilde{\boldsymbol{e}}_{2,g}\|_2 \\
&= O_{\mathrm{p}}\left\{n^{1/2}\left(p\lambda_K\right)^{-1/2}\right\}
\end{aligned}
$$

(b) (i)

$$
\begin{aligned}
&|p^{-1}\sum_{g=1}^{p}\left(\boldsymbol{\ell}_g+n^{-1/2}\tilde{\boldsymbol{e}}_{1,g}\right)^{\mathrm{T}}\boldsymbol{\epsilon}\hat{\boldsymbol{Z}}^{\mathrm{T}}n^{-1/2}\tilde{\boldsymbol{e}}_{2,g}| \\
&\leq \underbrace{\left\{p^{-1}\sum_{g=1}^{p}\left(\boldsymbol{\ell}_g+n^{-1/2}\tilde{\boldsymbol{e}}_{1,g}\right)^{\mathrm{T}}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}}\left(\boldsymbol{\ell}_g+n^{-1/2}\tilde{\boldsymbol{e}}_{1,g}\right)\right\}^{1/2}}_{O_{\mathrm{p}}\left\{n(p\lambda_K)^{-1}\right\}} \\
&\times \underbrace{\left(p^{-1}\sum_{g=1}^{p}n^{-1}\tilde{\boldsymbol{e}}_{2,g}^{\mathrm{T}}\hat{\boldsymbol{Z}}\hat{\boldsymbol{Z}}^{\mathrm{T}}\tilde{\boldsymbol{e}}_{2,g}\right)^{1/2}}_{O_{\mathrm{p}}\left\{n^{1/2}(p\lambda_K)^{-1/2}\right\}} = o_{\mathrm{p}}\left\{n\left(p\lambda_K\right)^{-1}\right\}
\end{aligned}
$$

(ii)

$$
\begin{aligned}
\left(\boldsymbol{\ell}_g+n^{-1/2}\tilde{\boldsymbol{e}}_{1,g}\right)^{\mathrm{T}}V\hat{\boldsymbol{Z}}^{\mathrm{T}}n^{-1/2}\tilde{\boldsymbol{e}}_{2,g} =& p^{1/2}n^{-1/2}\begin{pmatrix} d_1\boldsymbol{u}_1[g] & \cdots & d_K\boldsymbol{u}_K[g]\end{pmatrix} \\
&\times\begin{pmatrix}\hat{\boldsymbol{z}}_1^{\mathrm{T}}n^{-1/2}\tilde{\boldsymbol{e}}_{2,g}\\ \vdots \\ \hat{\boldsymbol{z}}_K^{\mathrm{T}}n^{-1/2}\tilde{\boldsymbol{e}}_{2,g}\end{pmatrix} \\
=& p^{1/2}n^{-1/2}d_1\boldsymbol{u}_1[g]\hat{\boldsymbol{z}}_1^{\mathrm{T}}n^{-1/2}\tilde{\boldsymbol{e}}_{2,g}+\cdots \\
&+p^{1/2}n^{-1/2}d_K\boldsymbol{u}_K[g]\hat{\boldsymbol{z}}_K^{\mathrm{T}}n^{-1/2}\tilde{\boldsymbol{e}}_{2,g}
\end{aligned}
$$

and for any $k\in[K]$,

$$
\begin{aligned}
p^{-1}\sum_{g=1}^{p}p^{1/2}n^{-1/2}d_k\boldsymbol{u}_k[g]\hat{\boldsymbol{z}}_k^{\mathrm{T}}n^{-1/2}\tilde{\boldsymbol{e}}_{2,g} &= (np)^{-1/2}\,d_k\hat{\boldsymbol{z}}_k^{\mathrm{T}}\sum_{g=1}^{p}\boldsymbol{u}_k[g]n^{-1/2}\tilde{\boldsymbol{e}}_{2,g} \\
&= O_{\mathrm{p}}\left(p^{-1}\right).
\end{aligned}
$$

The second equality follows because

$$
\begin{aligned}
&(np)^{-1/2}\, d_k \hat{\boldsymbol{z}}_k^{\mathrm{T}} = O_{\mathrm{p}}\left(p^{-1}\right)\\
&\sum_{g=1}^{p} \boldsymbol{u}_k[g] n^{-1/2} \tilde{\boldsymbol{e}}_{2,g} \sim n^{-1/2} N_n\left(0, \boldsymbol{u}_k^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{u}_k I_n\right) = O_{\mathrm{p}}\left(1\right).
\end{aligned}
$$

This completes the proof.

□

2.7.6 Proof of (2.20) *from Lemma 2.3 under the conditions of Theorem 2.2*

In this section, and for the remainder of the chapter, we return to assuming $\boldsymbol{Y}$ is distributed according (2.1a). However, we continue to use $\tilde{\boldsymbol{E}}_1$, $\tilde{\boldsymbol{E}}_2 \in \mathbb{R}^{p\times(n-d-K)}$, $\tilde{\boldsymbol{N}}$, $\boldsymbol{v}_k$, $\boldsymbol{V}$, $\hat{\boldsymbol{v}}_k$, $\hat{\boldsymbol{V}}$, $\hat{\boldsymbol{z}}_k \in \mathbb{R}^{n-d-K}$, $\hat{\boldsymbol{Z}} \in \mathbb{R}^{(n-d-K)\times K}$ and $\boldsymbol{R} \in \mathbb{R}^{(n-d-K)\times(n-d-K)}$ defined in Lemmas 2.4 and 2.5 in Section 2.7.4 in what follows.

We now prove Lemma 2.6, which will be useful in the proof of Theorems 2.1 and 2.2, and also acts as a proof of Lemma 2.3.

***Lemma* 2.6.** *Suppose the conditions of Theorem 2.2 hold and the diagonal elements of* $\hat{\boldsymbol{C}}_2^{\mathrm{T}} \boldsymbol{C}_2$ *are non-negative. Then*

$$
n^{1/2} \|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}\|_2 = o_{\mathrm{p}}\left(1\right)
$$

where $\hat{\boldsymbol{\Omega}}$ *is defined in* (2.10).

Proof. Recall

$$\hat{\boldsymbol{\Omega}}^{\mathrm{T}} = \operatorname{diag}\left\{\hat{\lambda}_1\left(\hat{\lambda}_1 - \hat{\rho}\right)^{-1}, \ldots, \hat{\lambda}_K\left(\hat{\lambda}_K - \hat{\rho}\right)^{-1}\right\}\left(\hat{\boldsymbol{L}}^{\mathrm{T}}\hat{\boldsymbol{L}}\right)^{-1}\hat{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{Y}_1 =$$

$$\begin{pmatrix} \frac{\hat{\lambda}_1}{\hat{\lambda}_1 - \hat{\rho}} & & \\ & \ddots & \\ & & \frac{\hat{\lambda}_K}{\hat{\lambda}_K - \hat{\rho}} \end{pmatrix}\begin{pmatrix} \frac{\lambda_1}{\hat{\lambda}_1} & & \\ & \ddots & \\ & & \frac{\lambda_K}{\hat{\lambda}_K} \end{pmatrix}\left\{\underbrace{\left(\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\right)^{-1}\hat{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{\beta}}_{\text{(a)}} + \underbrace{\left(\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\right)^{-1}\hat{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{L}}_{\text{(b)}}\boldsymbol{\Omega}^{\mathrm{T}}\right.$$

$$\left. + \underbrace{\left(\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\right)^{-1}\hat{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{E}_1}_{\text{(c)}}\right\}.$$

We will go through each one of these terms to prove $\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}\|_2 = o_p\left(n^{-1/2}\right)$.

(a) $\boldsymbol{M}_a = \left(\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\right)^{-1}\hat{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{\beta} = n^{1/2}p^{-1/2}\left(\tilde{\boldsymbol{L}}^{\mathrm{T}}\tilde{\boldsymbol{L}}\right)^{-1}\hat{\tilde{\boldsymbol{L}}}^{\mathrm{T}}\boldsymbol{\beta}$. Define $n^{1/2}p^{-1/2}\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}$ and let $\boldsymbol{M}_a[k,]$ be the kth row of $\boldsymbol{M}_a$.

$$\boldsymbol{M}_a[k,] = \lambda_k^{-1}\hat{\boldsymbol{v}}_k^{\mathrm{T}}\left(\tilde{\boldsymbol{L}} + p^{-1/2}\tilde{\boldsymbol{E}}_1\right)^{\mathrm{T}}\tilde{\boldsymbol{\beta}} + \lambda_k^{-1}p^{-1/2}\hat{\boldsymbol{z}}_k^{\mathrm{T}}\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\tilde{\boldsymbol{\beta}} = \lambda_k^{-1}\hat{\boldsymbol{v}}_k^{\mathrm{T}}\tilde{\boldsymbol{L}}^{\mathrm{T}}\tilde{\boldsymbol{\beta}} + o_{\mathrm{p}}\left(n^{-1/2}\right)$$

where the second equality follows because $[\boldsymbol{\beta}]_{*j}^{\mathrm{T}}[\boldsymbol{\beta}]_{*j} = o\left(n^{-3/2}p\lambda_K\right)$ for all $j = 1, \ldots, d$ by Assumption 2.3 and

$$\begin{aligned}
&\lambda_k^{-1}p^{-1/2}\hat{\boldsymbol{z}}_k^{\mathrm{T}}\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\tilde{\boldsymbol{\beta}} = O_{\mathrm{p}}\left\{n\left(\lambda_k p\right)^{-1}\|n\left(\lambda_k p\right)^{-1}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}\|_2^{1/2}\right\} = o_{\mathrm{p}}\left(n^{-1/2}\right)\\
&\hat{\boldsymbol{v}}_k^{\mathrm{T}}\left(\lambda_k p\right)^{-1/2}\tilde{\boldsymbol{E}}_1^{\mathrm{T}}\left(\lambda_k^{-1/2}\tilde{\boldsymbol{\beta}}\right) = O_{\mathrm{p}}\left\{\left(\lambda_k p\right)^{-1/2}\|n\left(\lambda_k p\right)^{-1}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}\|_2^{1/2}\right\} = o_{\mathrm{p}}\left(n^{-1/2}\right).
\end{aligned}$$

Lastly, the s, j element of of $\lambda_k^{-1}\tilde{\boldsymbol{L}}^{\mathrm{T}}\tilde{\boldsymbol{\beta}} \in \mathbb{R}^{K \times d}$ is such that

$$\begin{aligned}
|n\left(\lambda_k p\right)^{-1}\sum_{g=1}^{p}\boldsymbol{\ell}_{gs}\boldsymbol{\beta}_{gj}| &\leq n\left(\lambda_k p\right)^{-1}\left\{c + o_{\mathrm{p}}(1)\right\}\sum_{g=1}^{p} I\left(\boldsymbol{\beta}_{gj} \neq 0\right) = n\lambda_k^{-1}\left\{c + o_{\mathrm{p}}(1)\right\}\delta_j\\
&= o_{\mathrm{p}}\left(n^{-1/2}\right)
\end{aligned}$$

by Assumption 2.3, where $\delta_j = p^{-1} \sum_{g=1}^{p} I\left([\boldsymbol{\beta}]_{gj} \neq 0\right)$ and $c > 0$ is a constant that does not depend on n or p. The first inequality above is because the magnitude of the entries of $\boldsymbol{\beta}$ and $\boldsymbol{L}$ are bounded by a constant by Assumptions 2.2 and 2.3. Therefore, $\|\lambda_k^{-1}\hat{\boldsymbol{v}}_k^{\mathrm{T}}\tilde{\boldsymbol{L}}^{\mathrm{T}}\tilde{\boldsymbol{\beta}}\|_2 = o_{\mathrm{p}}\left(n^{-1/2}\right)$ for all $k = 1, \ldots, K$.

(b) $\left(\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\right)^{-1}\hat{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{L} = \left(\tilde{\boldsymbol{L}}^{\mathrm{T}}\tilde{\boldsymbol{L}}\right)^{-1}\hat{\tilde{\boldsymbol{L}}}^{\mathrm{T}}\tilde{\boldsymbol{L}} = \mathrm{diag}\left(\lambda_1^{-1}, \ldots, \lambda_K^{-1}\right)\hat{\tilde{\boldsymbol{L}}}^{\mathrm{T}}\tilde{\boldsymbol{L}}$ where

$$\begin{aligned}\hat{\tilde{\boldsymbol{L}}}^{\mathrm{T}}\tilde{\boldsymbol{L}} =& \hat{\boldsymbol{V}}^{\mathrm{T}}\left(\tilde{\boldsymbol{L}} + p^{-1/2}\tilde{\boldsymbol{E}}_1\right)^{\mathrm{T}}\tilde{\boldsymbol{L}} + \hat{\boldsymbol{Z}}^{\mathrm{T}}p^{-1/2}\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\tilde{\boldsymbol{L}} \\ =& \underbrace{\boldsymbol{\epsilon}^{\mathrm{T}}\left(\tilde{\boldsymbol{L}} + p^{-1/2}\tilde{\boldsymbol{E}}_1\right)^{\mathrm{T}}\tilde{\boldsymbol{L}}}_{\text{(i)}} + \underbrace{V^{\mathrm{T}}\left(\tilde{\boldsymbol{L}} + p^{-1/2}\tilde{\boldsymbol{E}}_1\right)^{\mathrm{T}}\tilde{\boldsymbol{L}}}_{\text{(ii)}} + \underbrace{\hat{\boldsymbol{Z}}^{\mathrm{T}}p^{-1/2}\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\left(\tilde{\boldsymbol{L}} + p^{-1/2}\tilde{\boldsymbol{E}}_1\right)}_{\text{(iii)}} \\ &+ O_{\mathrm{p}}\left(np^{-1}\right)\end{aligned}$$

(i) Suppose $\boldsymbol{\epsilon} = \begin{pmatrix}\boldsymbol{\epsilon}_1 & \cdots & \boldsymbol{\epsilon}_K\end{pmatrix}$ where $\boldsymbol{\epsilon}_k \in \mathbb{R}^K$ was defined in Lemma 2.5 as $\hat{\boldsymbol{v}}_k - v_k$. Since $\boldsymbol{\epsilon} = O_{\mathrm{p}}\left\{n\left(\lambda_K p\right)^{-1}\right\}$ and $p^{-1/2}\tilde{\boldsymbol{E}}_1^{\mathrm{T}}\tilde{\boldsymbol{L}} = O_{\mathrm{p}}\left(\lambda_1^{1/2}p^{-1/2}\right)$, then

$$\|\left(\tilde{\boldsymbol{L}}^{\mathrm{T}}\tilde{\boldsymbol{L}}\right)^{-1}\boldsymbol{\epsilon}^{\mathrm{T}}p^{-1/2}\tilde{\boldsymbol{E}}_1^{\mathrm{T}}\tilde{\boldsymbol{L}}\|_2 = \lambda_K^{-1/2}O_{\mathrm{p}}\left\{\frac{n}{p\lambda_K}\left(\frac{\lambda_1}{\lambda_K p}\right)^{1/2}\right\} = o_{\mathrm{p}}\left(n^{-1/2}\right).$$

Next, by Lemma 2.5,

$$\left(\tilde{\boldsymbol{L}}^{\mathrm{T}}\tilde{\boldsymbol{L}}\right)^{-1}\boldsymbol{\epsilon}^{\mathrm{T}}\tilde{\boldsymbol{L}}^{\mathrm{T}}\tilde{\boldsymbol{L}} = \begin{pmatrix}\boldsymbol{\epsilon}_1[1] & \frac{\lambda_2}{\lambda_1}\boldsymbol{\epsilon}_1[2] & \cdots & \frac{\lambda_K}{\lambda_1}\boldsymbol{\epsilon}_1[K] \\ \vdots & \ddots & \cdots & \vdots \\ \frac{\lambda_1}{\lambda_K}\boldsymbol{\epsilon}_K[1] & \frac{\lambda_2}{\lambda_K}\boldsymbol{\epsilon}_K[2] & \cdots & \boldsymbol{\epsilon}_K[K]\end{pmatrix} = o_{\mathrm{p}}\left(n^{-1/2}\right).$$

Therefore, $\boldsymbol{\epsilon}^{\mathrm{T}}\left(\tilde{\boldsymbol{L}} + p^{-1/2}\tilde{\boldsymbol{E}}_1\right)^{\mathrm{T}}\tilde{\boldsymbol{L}} = o_{\mathrm{p}}\left(n^{-1/2}\right)$.

(ii) $\left(\tilde{\boldsymbol{L}}^{\mathrm{T}}\tilde{\boldsymbol{L}}\right)^{-1} V^{\mathrm{T}}\left(\tilde{\boldsymbol{L}}+p^{-1/2}\tilde{\boldsymbol{E}}_1\right)^{\mathrm{T}}\tilde{\boldsymbol{L}}$

$$
\begin{aligned}
V^{\mathrm{T}}\left(\tilde{\boldsymbol{L}}+p^{-1/2}\tilde{\boldsymbol{E}}_1\right)^{\mathrm{T}}\tilde{\boldsymbol{L}} =& V^{\mathrm{T}}\left(\tilde{\boldsymbol{L}}+p^{-1/2}\tilde{\boldsymbol{E}}_1\right)^{\mathrm{T}}\left(\tilde{\boldsymbol{L}}+p^{-1/2}\tilde{\boldsymbol{E}}_1\right) - V^{\mathrm{T}}\tilde{\boldsymbol{L}}^{\mathrm{T}}p^{-1/2}\tilde{\boldsymbol{E}}_1 \\
& - \rho V^{\mathrm{T}} + O_{\mathrm{p}}\left(p^{-1/2}\right) \\
=& \operatorname{diag}\left(d_1^2-\rho,\ldots,d_K^2-\rho\right)V^{\mathrm{T}} - V^{\mathrm{T}}\tilde{\boldsymbol{L}}^{\mathrm{T}}p^{-1/2}\tilde{\boldsymbol{E}}_1 \\
& + O_{\mathrm{p}}\left(p^{-1/2}\right) \\
=& \operatorname{diag}\left(d_1^2-\rho,\ldots,d_K^2-\rho\right)V^{\mathrm{T}} \\
& - V^{\mathrm{T}}\begin{bmatrix} \leftarrow & O_{\mathrm{p}}\left(\frac{\lambda_1^{1/2}}{p^{1/2}}\right) & \rightarrow \\ \vdots & \ddots & \vdots \\ \leftarrow & O_{\mathrm{p}}\left(\frac{\lambda_K^{1/2}}{p^{1/2}}\right) & \rightarrow \end{bmatrix} + O_{\mathrm{p}}\left(p^{-1/2}\right) \\
=& \operatorname{diag}\left(\lambda_1,\ldots,\lambda_K\right) \\
& + \operatorname{diag}\left\{O_{\mathrm{p}}\left(\frac{\lambda_1^{1/2}}{p^{1/2}}\right),\ldots,O_{\mathrm{p}}\left(\frac{\lambda_K^{1/2}}{p^{1/2}}\right)\right\} \\
& - \begin{bmatrix} \leftarrow & O_{\mathrm{p}}\left(\frac{\lambda_1^{1/2}}{p^{1/2}}\right) & \rightarrow \\ \vdots & \ddots & \vdots \\ \leftarrow & O_{\mathrm{p}}\left(\frac{\lambda_K^{1/2}}{p^{1/2}}\right) & \rightarrow \end{bmatrix} + O_{\mathrm{p}}\left(p^{-1/2}\right)
\end{aligned}
$$

Therefore,

$$
\left(\tilde{\boldsymbol{L}}^{\mathrm{T}}\tilde{\boldsymbol{L}}\right)^{-1} V^{\mathrm{T}}\left(\tilde{\boldsymbol{L}}+p^{-1/2}\tilde{\boldsymbol{E}}_1\right)^{\mathrm{T}}\tilde{\boldsymbol{L}} = I_K + O_{\mathrm{p}}\left\{(\lambda_K p)^{-1/2}\right\}.
$$

(iii) $\left(\tilde{\boldsymbol{L}}^{\mathrm{T}}\tilde{\boldsymbol{L}}\right)^{-1}\hat{\boldsymbol{Z}}^{\mathrm{T}}p^{-1/2}\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\left(\tilde{\boldsymbol{L}}+p^{-1/2}\tilde{\boldsymbol{E}}_1\right)$

$$\left(\tilde{\boldsymbol{L}}^{\mathrm{T}}\tilde{\boldsymbol{L}}\right)^{-1}\hat{\boldsymbol{Z}}^{\mathrm{T}}p^{-1/2}\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\left(\tilde{\boldsymbol{L}}+p^{-1/2}\tilde{\boldsymbol{E}}_1\right)=\begin{pmatrix}\lambda_1^{-1}p^{-1/2}\hat{\boldsymbol{z}}_1^{\mathrm{T}}\tilde{\boldsymbol{E}}_2\sum_{k=1}^{K}d_k\boldsymbol{u}_k v_k^{\mathrm{T}}\\ \vdots \\ \lambda_K^{-1}p^{-1/2}\hat{\boldsymbol{z}}_K^{\mathrm{T}}\tilde{\boldsymbol{E}}_2\sum_{k=1}^{K}d_k\boldsymbol{u}_k v_k^{\mathrm{T}}\end{pmatrix}$$

The largest row (in magnitude) in the above matrix will obviously be the K^{th} row, so we need only focus on that row. By the expression for $\hat{\boldsymbol{z}}_K$ given in (2.38) and Lemmas 2.8 and 2.9,

$$\begin{aligned}\frac{d_1}{\lambda_K p^{1/2}}\hat{\boldsymbol{z}}_K^{\mathrm{T}}\tilde{\boldsymbol{E}}_2\boldsymbol{u}_1 =&\frac{d_1 d_K}{\lambda_K^2 p}\boldsymbol{u}_K^{\mathrm{T}}\tilde{\boldsymbol{E}}_2\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\boldsymbol{u}_1+\frac{d_1 d_K}{\lambda_K^3 p}\boldsymbol{u}_K^{\mathrm{T}}\tilde{\boldsymbol{E}}_2 R\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\boldsymbol{u}_1\\ &+O_{\mathrm{p}}\left[\frac{\lambda_1^{1/2}}{\lambda_K^{3/2}}\frac{n^{1/2}}{p^{1/2}}\left\{\left(\frac{n}{\lambda_K p}\right)^{3/2}+\frac{n^{1/2}}{\lambda_K p}\right\}\right]\end{aligned}$$

where

$$\begin{aligned}&\frac{d_1 d_K}{\lambda_K^2 p}\boldsymbol{u}_K^{\mathrm{T}}\tilde{\boldsymbol{E}}_2\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\boldsymbol{u}_1=O_{\mathrm{p}}\left(\frac{n}{p\lambda_K}+\frac{n^{1/2}d_1}{\lambda_K^{3/2}p}\right)=o_{\mathrm{p}}\left(n^{-1/2}\right)\\ &\frac{d_1 d_K}{\lambda_K^3 p}\boldsymbol{u}_K^{\mathrm{T}}\tilde{\boldsymbol{E}}_2 R\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\boldsymbol{u}_1=O_{\mathrm{p}}\left(\frac{d_1 n}{\lambda_K p}\frac{n}{p\lambda_K^{3/2}}+\frac{d_1}{\lambda_K^{1/2}p^{1/2}}\frac{n}{p\lambda_K^2}\right)=o_{\mathrm{p}}\left(n^{-1/2}\right)\\ &\frac{\lambda_1^{1/2}}{\lambda_K^{3/2}}\frac{n^{1/2}}{p^{1/2}}\left\{\left(\frac{n}{\lambda_K p}\right)^{3/2}+\frac{n^{1/2}}{\lambda_K p}\right\}=\frac{n}{\lambda_K^2 p}\frac{\lambda_1^{1/2}n}{p}+\frac{\lambda_1^{1/2}}{\lambda_K^{1/2}p^{1/2}}\frac{n}{\lambda_K^2 p}=o\left(n^{-1/2}\right)\end{aligned}$$

Second,

$$\frac{d_K}{\lambda_K p^{1/2}}\hat{\boldsymbol{z}}_K^{\mathrm{T}}\tilde{\boldsymbol{E}}_2\boldsymbol{u}_K=O_{\mathrm{p}}\left(\frac{n}{\lambda_K p}\right)=o_{\mathrm{p}}\left(n^{-1/2}\right)$$

Therefore, $\left(\tilde{\boldsymbol{L}}^{\mathrm{T}}\tilde{\boldsymbol{L}}\right)^{-1}\hat{\boldsymbol{Z}}^{\mathrm{T}}p^{-1/2}\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\left(\tilde{\boldsymbol{L}}+p^{-1/2}\tilde{\boldsymbol{E}}_1\right)=o_{\mathrm{p}}\left(n^{-1/2}\right)$.

We have shown that $\left(\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\right)^{-1}\hat{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{L}=I_K+o_{\mathrm{p}}\left(n^{-1/2}\right)$.

(c) Recall that $\boldsymbol{Y}_1 = \boldsymbol{Y}\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{X})^{-1}$ and $\boldsymbol{Y}_2 = \boldsymbol{Y}\boldsymbol{Q}_X$ where $\boldsymbol{Q}_X^{\mathrm{T}}\boldsymbol{X} = 0_{(n-d)\times d}$. Since the residuals $\boldsymbol{E} \sim MN_{p\times n}(0, \boldsymbol{\Sigma}, I_n)$, $\boldsymbol{E}_1 = \boldsymbol{E}\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{X})^{-1}$ and $\boldsymbol{E}_2 = \boldsymbol{E}\boldsymbol{Q}_X$ are independent. And since we use $\boldsymbol{Y}_2$ to estimate $\hat{\boldsymbol{L}}$, $\hat{\boldsymbol{L}}$ and $\boldsymbol{E}_1$ are independent. (We abuse notation here. $\tilde{\boldsymbol{E}}_1$ and $\boldsymbol{E}_1$ are different. $\tilde{\boldsymbol{E}}_1$ is defined using the second set of data in part 1). Therefore,

$$
\begin{aligned}
\left(\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\right)^{-1}\hat{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{E}_1 \sim\, & p^{-1/2}\mathrm{diag}\left(\lambda_1^{-1/2},\ldots,\lambda_K^{-1/2}\right) MN_{K\times d}\left\{0, \mathrm{diag}\left(\lambda_1^{-1/2},\ldots,\lambda_K^{-1/2}\right)\right. \\
& \left.\times\hat{\hat{\boldsymbol{L}}}^{\mathrm{T}}\boldsymbol{\Sigma}\hat{\hat{\boldsymbol{L}}}\mathrm{diag}\left(\lambda_1^{-1/2},\ldots,\lambda_K^{-1/2}\right), \left(n^{-1}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right)^{-1}\right\} \\
=\, & O_{\mathrm{p}}\left(\lambda_K^{-1/2}p^{-1/2}\right) = o_{\mathrm{p}}\left(n^{-1/2}\right).
\end{aligned}
$$

The above work shows that

$$
\left(\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\right)^{-1}\hat{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{\beta} + \left(\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\right)^{-1}\hat{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{\Omega} + \left(\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\right)^{-1}\hat{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{E}_1 = \boldsymbol{\Omega} + o_{\mathrm{p}}\left(n^{-1/2}\right).
$$

Our last task is to understand $\left\{\hat{\lambda}_k\left(\hat{\lambda}_k - \hat{\rho}\right)^{-1}\right\}\left(\lambda_k\hat{\lambda}_k^{-1}\right)$ for $k \in [K]$. By Lemmas 2.1, 2.4 and 2.5,

$$
\begin{aligned}
\frac{\hat{\lambda}_k}{\hat{\lambda}_k - \hat{\rho}}\frac{\lambda_k}{\hat{\lambda}_k} &= \left(\frac{\hat{\lambda}_k - \hat{\rho}}{\lambda_k}\right)^{-1} = \left[1 + (\rho - \hat{\rho})\lambda_k^{-1} + O_{\mathrm{p}}\left\{\lambda_K^{-1/2}p^{-1/2} + n(\lambda_k p)^{-1}\right\}\right]^{-1} \\
&= \left\{1 + o_{\mathrm{p}}\left(n^{-1/2}\right)\right\}^{-1} = 1 + o_{\mathrm{p}}\left(n^{-1/2}\right).
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\hat{\boldsymbol{\Omega}}^{\mathrm{T}} = &\begin{pmatrix} \frac{\hat{\lambda}_1}{\hat{\lambda}_1-\hat{\rho}} & & \\ & \ddots & \\ & & \frac{\hat{\lambda}_K}{\hat{\lambda}_K-\hat{\rho}} \end{pmatrix} \begin{pmatrix} \frac{\lambda_1}{\hat{\lambda}_1} & & \\ & \ddots & \\ & & \frac{\lambda_K}{\hat{\lambda}_K} \end{pmatrix} \left\{ \left(\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\right)^{-1} \hat{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{\beta} + \left(\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\right)^{-1} \hat{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{\Omega}^{\mathrm{T}} \right. \\
&\left. + \left(\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\right)^{-1} \hat{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{E}_1 \right\} \\
&= \boldsymbol{\Omega}^{\mathrm{T}} + o_{\mathrm{p}}\left(n^{-1/2}\right).
\end{aligned}
$$

□

2.7.7 Proof of the remaining theory from Sections 2.3.2, 2.3.3 and 2.3.4

In this section, we prove Theorem 2.2, Proposition 2.2 and Corollary 2.1 (in that order). We need not prove Theorem 2.1, since Theorem 2.1 is a special case of Theorem 2.2.

Proof of Theorem 2.2. Define $\boldsymbol{e}_{1,g}$ to be the g^{th} row of $\boldsymbol{E}_1$. Then

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_g &= \boldsymbol{\Omega}\left(\boldsymbol{\ell}_g - \hat{\boldsymbol{\ell}}_g\right) + \boldsymbol{e}_{1,g} + \left(\boldsymbol{\Omega} - \hat{\boldsymbol{\Omega}}\right)\hat{\boldsymbol{\ell}}_g \\
\hat{\boldsymbol{\beta}}_g^{\mathrm{OLS}} - \boldsymbol{\beta}_g &= \boldsymbol{\Omega}\left(\boldsymbol{\ell}_g - \hat{\boldsymbol{\ell}}_g^{\mathrm{OLS}}\right) + \boldsymbol{e}_{1,g}
\end{aligned}
$$

where $\hat{\boldsymbol{\ell}}_g^{\mathrm{OLS}}$ is the ordinary least squares estimate for $\boldsymbol{\ell}_g$, assuming $\boldsymbol{C}$ was known. By the proof of Lemma 2.2, $n^{1/2}\|\boldsymbol{\Omega}\hat{\boldsymbol{\ell}}_g^{\mathrm{OLS}} - \boldsymbol{\Omega}\hat{\boldsymbol{\ell}}_g\|_2 = o_{\mathrm{p}}(1)$. Equation (2.21) and its equivalent in the statement of Theorem 2.2 follow because $n^{1/2}\|\boldsymbol{\Omega} - \hat{\boldsymbol{\Omega}}\|_2 = o_{\mathrm{p}}(1)$. Equation (2.22) and its equivalent in the statement of Theorem 2.2 then follows because $\hat{\sigma}_g = \sigma_g + o_{\mathrm{p}}(1)$ and

$$
n^{1/2}\sigma_g^{-1}\left(\hat{\boldsymbol{\beta}}_g^{\mathrm{OLS}} - \boldsymbol{\beta}_g\right) \sim N_d\left\{0, \left(n^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1} + \boldsymbol{\Omega}\boldsymbol{\Omega}^{\mathrm{T}}\right\}.
$$

□

Proof of Proposition 2.2. Define

$$\hat{\boldsymbol{\Gamma}} = \operatorname{diag}\left\{\left(\hat{\lambda}_1 - \hat{\rho}\right)/\hat{\lambda}_1, \ldots, \left(\hat{\lambda}_K - \hat{\rho}\right)/\hat{\lambda}_K\right\}$$
$$\boldsymbol{\Gamma} = \left\{\lambda_1/\left(\lambda_1 + \rho\right), \ldots, \lambda_K/\left(\lambda_K + \rho\right)\right\}.$$

By Lemmas 2.1 and 2.2,

$$\hat{\boldsymbol{\Gamma}} = \boldsymbol{\Gamma} + o_{\mathrm{p}}\left(n^{-1/2}\right).$$

And by Lemma 2.6,

$$\|\hat{\boldsymbol{\Omega}}^{\text{shrunk}} - \boldsymbol{\Omega}\boldsymbol{\Gamma}\|_2 = \|\hat{\boldsymbol{\Omega}}\hat{\boldsymbol{\Gamma}} - \boldsymbol{\Omega}\boldsymbol{\Gamma}\|_2 \leq \|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}\|_2 + o_{\mathrm{p}}\left(n^{-1/2}\right) = o_{\mathrm{p}}\left(n^{-1/2}\right).$$

□

Proof of Corollary 2.1. Define $\boldsymbol{\Omega}^{\text{shrunk}} = \boldsymbol{\Omega}\operatorname{diag}\left\{\lambda_1\left(\lambda_1+\rho\right)^{-1}, \ldots, \lambda_K\left(\lambda_K+\rho\right)^{-1}\right\}$ and let ω_k be the kth element of $\boldsymbol{\Omega} \in \mathbb{R}^{1\times K}$. Using Proposition 2.2 and the definition of $\hat{\beta}_g^{\text{shrunk}}$ from the statement of Corollary 2.1,

$$\begin{aligned}\hat{\beta}_g^{\text{shrunk}} - \beta_g =& \boldsymbol{\Omega}^{\text{shrunk}}\left(\boldsymbol{\ell}_g - \hat{\boldsymbol{\ell}}_g\right) + e_{1_g} + \rho\boldsymbol{\Omega}\operatorname{diag}\left\{\left(\lambda_1+\rho\right)^{-1}, \ldots, \left(\lambda_K+\rho\right)^{-1}\right\}\boldsymbol{\ell}_g \\ &+ o_{\mathrm{p}}\left(n^{-1/2}\right).\end{aligned} \tag{2.41}$$

where e_{1_g} is the gth row of $\boldsymbol{E}_1$. By Lemma 2.2,

$$n^{1/2}\left\{\boldsymbol{\Omega}^{\text{shrunk}}\left(\boldsymbol{\ell}_g - \hat{\boldsymbol{\ell}}_g\right) + e_{1_g}\right\} \overset{\mathcal{D}}{=} Z + o_{\mathrm{p}}(1),$$

where $\sigma_g^{-1}Z \sim N\left\{0, \left(n^{-1}\|\boldsymbol{X}\|_2^2\right)^{-1} + \|\boldsymbol{\Omega}^{\text{shrunk}}\|_2^2\right\}$. Define

$$s_g = \hat{\sigma}_g\left\{\left(n^{-1}\|\boldsymbol{X}\|_2^2\right)^{-1} + \|\hat{\boldsymbol{\Omega}}^{\text{shrunk}}\|_2^2\right\}^{1/2}.$$

If $\lambda_K^{-1}n^{1/2} \to 0$, then clearly $n^{1/2}s_g^{-1}\left(\hat{\beta}_g^{\text{shrunk}} - \beta_g\right) \overset{\mathcal{D}}{=} W + o_{\text{p}}(1)$, where $W \sim N(0,1)$. Next, we can write

$$\begin{aligned}|z_g| = s_g^{-1}|n^{1/2}\left(\hat{\beta}_g^{\text{shrunk}} - \beta_g\right)| &= s_g^{-1}|O_{\text{p}}(1) + \rho n^{1/2}\sum_{k=1}^{K}\omega_k \boldsymbol{\ell}_{gk}\left(\rho+\lambda_k\right)^{-1}| \\ &\geq s_g^{-1}\left\{\rho n^{1/2}|\sum_{k=1}^{K}\omega_k \boldsymbol{\ell}_{gk}\left(\rho+\lambda_k\right)^{-1}| - |O_{\text{p}}(1)|\right\} \end{aligned} \tag{2.42}$$

where $s_g^{-1} \geq c + o_{\text{p}}(1)$ for some constant $c > 0$. If $\lambda_K^{-1}n^{1/2} \to \infty$, then by Item (ii) in the statement of Corollary 2.1, $\mathbb{P}\left(|z_g| \geq q_{1-\alpha/2}\right) \to 1$ for any $q_{1-\alpha/2} > 0$ because

$$n^{1/2}|\sum_{k=1}^{K}\omega_k \boldsymbol{\ell}_{gk}\left(\rho+\lambda_k\right)^{-1}| \geq n^{1/2}\left(\rho+\lambda_K\right)^{-1}\epsilon \asymp n^{1/2}\lambda_K^{-1}\epsilon \to \infty.$$

Next, assume $\lambda_K^{-1}n^{1/2-c_6} \to \infty$ for some small constant $c_6 > 0$. Then

$$n^{1/2}|\sum_{k=1}^{K}\omega_k \boldsymbol{\ell}_{gk}\left(\rho+\lambda_k\right)^{-1}| \geq n^{1/2}\left(\rho+\lambda_K\right)^{-1}\epsilon \asymp n^{c_6}\left(n^{1/2-c_6}\lambda_K^{-1}\epsilon\right)$$

where for Φ the cumulative distribution function for the standard normal and $|z|_g$ large enough,

$$\log\left\{2\Phi\left(-|z_g|\right)\right\} \leq -z_g^2/2 \leq -\tilde{c}n^{2c_6}\left(n^{1/2-c_6}\lambda_K^{-1}\right)^2\left\{1+o_{\text{p}}(1)\right\}$$

for some constant $\tilde{c} > 0$. If $n^{-r}p \to 0$ for some $r > 0$ as $n, p \to \infty$, then $\exp\left(-\tilde{c}n^{2c_6}\right)p \to 0$ as $n, p \to \infty$. Therefore, for any $\alpha \in (0,1)$,

$$\mathbb{P}\left\{|z_g| \geq q_{1-(p^{-1}\alpha)/2}\right\} = \mathbb{P}\left\{2p\Phi\left(-|z_g|\right) \leq \alpha\right\} \to 1$$

as $n, p \to \infty$.

Lastly, suppose $\lambda_K^{-1}n^{1/2} \geq c_6 > 0$. By (2.42), for any $\delta > 0$, there exists an M large enough such that if $\lambda_K^{-1} n^{1/2} \geq M$, $\mathbb{P}\left(|z_g| \geq q_{1-\alpha/2}\right) \geq 1-\delta$ for all n large enough. Therefore, it suffices to assume $\lambda_K^{-1} n^{1/2}$ is bounded from above by a constant. By (2.41), this implies

$$n^{1/2} s_g^{-1}\left(\hat{\beta}_g^{\text{shrunk}} - \beta_g\right) \stackrel{\mathcal{D}}{=} W + c_{n,p} + o_{\mathrm{p}}(1)$$

where $W \sim N(0,1)$, $c_{n,p}$ is non-random and

$$\begin{aligned} |c_{n,p}| =& \sigma_g^{-1}\left\{\left(n^{-1}\|\boldsymbol{X}\|_2^2\right)^{-1} + \|\boldsymbol{\Omega}^{\text{shrunk}}\|_2^2\right\}^{-1/2} \rho \\ & \times |n^{1/2}\boldsymbol{\Omega}\,\mathrm{diag}\left\{(\lambda_1+\rho)^{-1}, \ldots, (\lambda_K+\rho)^{-1}\right\}| \geq c \end{aligned}$$

for all n, p large enough, where $c > 0$ is a constant not dependent on n or p. Since

$$\mathbb{P}\left(|W + c_{n,p}| \geq q_{1-\alpha/2}\right) \geq \mathbb{P}\left(|W + c| \geq q_{1-\alpha/2}\right) > \alpha$$

for all n, p large enough, this proves the claim. □

2.7.8 CATE-RR and dSVA inflate test statistics

We now state and prove results similar to Proposition 2.2 and Corollary 2.1, except for the estimators used in dSVA (Lee et al. 2017) and CATE-RR (Wang et al. 2017).

***Proposition* 2.5** (Estimate for $\boldsymbol{\Omega}$ used in dSVA)**.** *Suppose the assumptions of Proposition 2.2 hold but we estimate* $\boldsymbol{\Omega}$ *as*

$$\hat{\boldsymbol{\Omega}}^{\text{dSVA}} = \boldsymbol{Y}_1^{\mathrm{T}} P_{1_p}^{\perp} \hat{\boldsymbol{L}}\left(\hat{\boldsymbol{L}}^{\mathrm{T}} P_{1_p}^{\perp} \hat{\boldsymbol{L}}\right)^{-1}.$$

Then if the smallest eigenvalue of $np^{-1}\boldsymbol{L}^{\mathrm{T}} P_{1_p}^{\perp} \boldsymbol{L}$ *is greater than* $\delta\lambda_K$ *where* $\delta > 0$ *is a* *constant,*

$$\|\hat{\boldsymbol{\Omega}}^{\mathrm{dSVA}} - \boldsymbol{\Omega} \left(\boldsymbol{L}^{\mathrm{T}} P_{1_p}^{\perp} \boldsymbol{L}\right) \left(\boldsymbol{L}^{\mathrm{T}} P_{1_p}^{\perp} \boldsymbol{L} + p n^{-1} \rho I_K\right)^{-1} \|_2 = o_{\mathrm{p}} \left(n^{-1/2}\right).$$

Proof. Define $\hat{\boldsymbol{V}} = (\hat{\boldsymbol{v}}_1 \cdots \hat{\boldsymbol{v}}_K) \in \mathbb{R}^{K \times K}$ and $\hat{\boldsymbol{Z}} = (\hat{\boldsymbol{z}}_1 \cdots \hat{\boldsymbol{z}}_K) \in \mathbb{R}^{(n-d-K) \times K}$, where $\hat{\boldsymbol{v}}_1, \ldots, \hat{\boldsymbol{v}}_K$ and $\hat{\boldsymbol{z}}_1, \ldots, \hat{\boldsymbol{z}}_K$ are defined in (2.37) and (2.38) in the statement of Lemma 2.5. By Lemmas 2.4 and 2.5,

$$\hat{\boldsymbol{L}} = n^{-1/2} \boldsymbol{Y}_2 \left(\tilde{\boldsymbol{C}} \hat{\boldsymbol{V}} + \boldsymbol{Q} \hat{\boldsymbol{Z}}\right) = n^{-1/2} \boldsymbol{L} \hat{\boldsymbol{V}} + n^{-1/2} \tilde{\boldsymbol{E}}_1 \hat{\boldsymbol{V}} + n^{-1/2} \tilde{\boldsymbol{E}}_2 \hat{\boldsymbol{Z}}$$

where, $\tilde{\boldsymbol{C}} = n^{-1/2} \boldsymbol{C}_2$, $\boldsymbol{Q} = \boldsymbol{Q}_{C_2}$, $\tilde{\boldsymbol{E}}_1 = \boldsymbol{E} \boldsymbol{Q}_X \tilde{\boldsymbol{C}}$ and $\tilde{\boldsymbol{E}}_2 = \boldsymbol{E} \boldsymbol{Q}_X \boldsymbol{Q}$. Note that $\tilde{\boldsymbol{E}}_1$ and $\tilde{\boldsymbol{E}}_2$ are independent by Craig's Theorem. Therefore,

$$\begin{aligned} n (\lambda_K p)^{-1} \hat{\boldsymbol{L}}^{\mathrm{T}} P_{1_p}^{\perp} \hat{\boldsymbol{L}} =& n (\lambda_K p)^{-1} \hat{\boldsymbol{V}}^{\mathrm{T}} \boldsymbol{L}^{\mathrm{T}} P_{1_p}^{\perp} \boldsymbol{L} \hat{\boldsymbol{V}} + n^{1/2} (\lambda_K p)^{-1} \hat{\boldsymbol{V}}^{\mathrm{T}} \boldsymbol{L}^{\mathrm{T}} P_{1_p}^{\perp} \tilde{\boldsymbol{E}}_1 \hat{\boldsymbol{V}} \\ &+ \left\{n^{1/2} (\lambda_K p)^{-1} \hat{\boldsymbol{V}}^{\mathrm{T}} \boldsymbol{L}^{\mathrm{T}} P_{1_p}^{\perp} \tilde{\boldsymbol{E}}_1 \hat{\boldsymbol{V}}\right\}^{\mathrm{T}} + n^{1/2} (\lambda_K p)^{-1} \hat{\boldsymbol{V}}^{\mathrm{T}} \boldsymbol{L}^{\mathrm{T}} P_{1_p}^{\perp} \tilde{\boldsymbol{E}}_2 \hat{\boldsymbol{Z}} \\ &+ \left\{n^{1/2} (\lambda_K p)^{-1} \hat{\boldsymbol{V}}^{\mathrm{T}} \boldsymbol{L}^{\mathrm{T}} P_{1_p}^{\perp} \tilde{\boldsymbol{E}}_2 \hat{\boldsymbol{Z}}\right\}^{\mathrm{T}} + (\lambda_K p)^{-1} \hat{\boldsymbol{V}}^{\mathrm{T}} \tilde{\boldsymbol{E}}_1^{\mathrm{T}} P_{1_p}^{\perp} \tilde{\boldsymbol{E}}_1 \hat{\boldsymbol{V}} \\ &+ (\lambda_K p)^{-1} \hat{\boldsymbol{Z}}^{\mathrm{T}} \tilde{\boldsymbol{E}}_2^{\mathrm{T}} P_{1_p}^{\perp} \tilde{\boldsymbol{E}}_2 \hat{\boldsymbol{Z}} + (\lambda_K p)^{-1} \hat{\boldsymbol{Z}}^{\mathrm{T}} \tilde{\boldsymbol{E}}_2^{\mathrm{T}} P_{1_p}^{\perp} \tilde{\boldsymbol{E}}_1 \hat{\boldsymbol{V}} \\ &+ \left\{(\lambda_K p)^{-1} \hat{\boldsymbol{Z}}^{\mathrm{T}} \tilde{\boldsymbol{E}}_2^{\mathrm{T}} P_{1_p}^{\perp} \tilde{\boldsymbol{E}}_1 \hat{\boldsymbol{V}}\right\}^{\mathrm{T}}. \end{aligned}$$

By the Lemmas 2.4 and 2.5,

$$n (\lambda_K p)^{-1} \hat{\boldsymbol{V}}^{\mathrm{T}} \boldsymbol{L}^{\mathrm{T}} P_{1_p}^{\perp} \boldsymbol{L} \hat{\boldsymbol{V}} = n (\lambda_K p)^{-1} \boldsymbol{L}^{\mathrm{T}} P_{1_p}^{\perp} \boldsymbol{L} + o_{\mathrm{p}} \left(n^{-1/2}\right).$$

Next,

$$\begin{aligned} n^{1/2} (\lambda_K p)^{-1} \hat{\boldsymbol{V}}^{\mathrm{T}} \boldsymbol{L}^{\mathrm{T}} P_{1_p}^{\perp} \tilde{\boldsymbol{E}}_1 \hat{\boldsymbol{V}} &= (\lambda_K p)^{-1/2} \hat{\boldsymbol{V}}^{\mathrm{T}} \left\{n^{1/2} (\lambda_K p)^{-1/2} \boldsymbol{L}\right\}^{\mathrm{T}} P_{1_p}^{\perp} \tilde{\boldsymbol{E}}_1 \hat{\boldsymbol{V}} \\ &= O_{\mathrm{p}} \left\{(\lambda_K p)^{-1/2}\right\} = o_{\mathrm{p}} \left(n^{-1/2}\right) \end{aligned}$$

$$
\begin{aligned}
n^{1/2}(\lambda_K p)^{-1}\hat{\boldsymbol{V}}^{\mathrm{T}}\boldsymbol{L}^{\mathrm{T}}P_{1_p}^{\perp}\tilde{\boldsymbol{E}}_2\hat{\boldsymbol{Z}} &= (\lambda_K p)^{-1/2}\hat{\boldsymbol{V}}^{\mathrm{T}}\left\{n^{1/2}(\lambda_K p)^{-1/2}\boldsymbol{L}\right\}^{\mathrm{T}}P_{1_p}^{\perp}\tilde{\boldsymbol{E}}_2\hat{\boldsymbol{Z}}\\
&= O_{\mathrm{p}}\left\{n(\lambda_K p)^{-1}\right\} = o_{\mathrm{p}}\left(n^{-1/2}\right).
\end{aligned}
$$

$$
(\lambda_K p)^{-1}\hat{\boldsymbol{Z}}^{\mathrm{T}}\tilde{\boldsymbol{E}}_2^{\mathrm{T}}P_{1_p}^{\perp}\tilde{\boldsymbol{E}}_2\hat{\boldsymbol{Z}} = O_{\mathrm{p}}\left(np^{-1}\lambda_K^{-2}\right) = o_{\mathrm{p}}\left(n^{-1/2}\right)
$$

$$
\begin{aligned}
(\lambda_K p)^{-1}\hat{\boldsymbol{Z}}^{\mathrm{T}}\tilde{\boldsymbol{E}}_2^{\mathrm{T}}P_{1_p}^{\perp}\tilde{\boldsymbol{E}}_1\hat{\boldsymbol{V}} &= \lambda_K^{-1}p^{-1/2}\hat{\boldsymbol{Z}}^{\mathrm{T}}\tilde{\boldsymbol{E}}_2^{\mathrm{T}}P_{1_p}^{\perp}\left(p^{-1/2}\tilde{\boldsymbol{E}}_1\right)\hat{\boldsymbol{V}} = O_{\mathrm{p}}\left(np^{-1}\lambda_K^{-3/2}\right)\\
&= o_{\mathrm{p}}\left(n^{-1/2}\right).
\end{aligned}
$$

Lastly,

$$
(\lambda_K p)^{-1}\hat{\boldsymbol{V}}^{\mathrm{T}}\tilde{\boldsymbol{E}}_1^{\mathrm{T}}P_{1_p}^{\perp}\tilde{\boldsymbol{E}}_1\hat{\boldsymbol{V}} = \lambda_K^{-1}p^{-1}\operatorname{tr}\left(\boldsymbol{\Sigma}P_{1_p}^{\perp}\right)I_K + o_{\mathrm{p}}\left(n^{-1/2}\right) = \lambda_K^{-1}\rho I_K + o_{\mathrm{p}}\left(n^{-1/2}\right).
$$

An identical calculation shows that

$$
n(\lambda_K p)^{-1}\boldsymbol{L}^{\mathrm{T}}P_{1_p}^{\perp}\hat{\boldsymbol{L}} = n(\lambda_K p)^{-1}\boldsymbol{L}^{\mathrm{T}}P_{1_p}^{\perp}\boldsymbol{L} + o_{\mathrm{p}}\left(n^{-1/2}\right).
$$

Lastly, for $\boldsymbol{E}_1 = \boldsymbol{E}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}$,

$$
(\lambda_K p)^{-1/2}n^{1/2}\boldsymbol{E}_1^{\mathrm{T}}P_{1_p}^{\perp}\hat{\boldsymbol{L}}\left(\hat{\boldsymbol{L}}^{\mathrm{T}}P_{1_p}^{\perp}\hat{\boldsymbol{L}}\right)^{-1} = O_{\mathrm{p}}\left\{p^{-1/2}\lambda_K^{-1}\right\} = o_{\mathrm{p}}\left(n^{-1/2}\right).
$$

This completes the proof. □

***Proposition* 2.6** (Estimate for $\boldsymbol{\Omega}$ used in CATE-RR)**.** *Suppose the assumptions of Proposition 2.2 hold with $d=1$ but we estimate $\boldsymbol{\Omega}$ as*

$$
\hat{\boldsymbol{\Omega}}^{\mathrm{cate}} = \underset{\boldsymbol{\alpha}\in\mathbb{R}^{1\times K}}{\arg\min}\,\Psi\left(y_{1_g} - \boldsymbol{\alpha}\hat{\boldsymbol{\ell}}_g\right)
$$

where y_{1_g} *is the* g*th element of* $\boldsymbol{Y}_1 \in \mathbb{R}^p$ *and for some constant* $c > 0$*,*

$$\Psi(x) = \begin{cases} x^2/2 & \textit{if } |x| \leq c \\ c|x| - c^2/2 & \textit{if } |x| > c \end{cases}. \tag{2.43}$$

Note that Ψ *is Huber's loss. Suppose further that* $pn^{-r} \to 0$ *for some* $r > 0$*. Then if* $\lambda_K \to \infty$*, the results of Proposition 2.2 hold. If* $\lambda_1 = O(1)$*, then there exists a constant* $\epsilon > 0$ *such that*

$$\lim_{n,p\to\infty} \mathbb{P}\left(\|\hat{\boldsymbol{\Omega}}^{\text{cate}} - \boldsymbol{\Omega}\|_2 \geq \epsilon\|\boldsymbol{\Omega}\|_2\right) = 1.$$

Proof. Let e_{1_g} be the gth element of $\boldsymbol{E}_1 \in \mathbb{R}^p$ and $d\Psi(x)/dx = \dot{\Psi}(x)$. Since Ψ is a convex function, $\hat{\boldsymbol{\Omega}}^{\text{cate}}$ solves

$$0 = \sum_{g=1}^{p} \dot{\Psi}\left(y_{1_g} - \hat{\boldsymbol{\ell}}_g \hat{\boldsymbol{\Omega}}^{\text{cate}}\right) \hat{\boldsymbol{\ell}}_g^{\mathrm{T}} = \left(\boldsymbol{Y}_1 - \hat{\boldsymbol{L}}\hat{\boldsymbol{\Omega}}^{\text{cate}^{\mathrm{T}}}\right)^{\mathrm{T}} \hat{\boldsymbol{A}}\hat{\boldsymbol{L}}$$

where $\hat{\boldsymbol{A}} \in \mathbb{R}^{p\times p}$ is a diagonal matrix with

$$\left[\hat{\boldsymbol{A}}\right]_{gg} = \begin{cases} 1 & \text{if } |y_{1_g} - \hat{\boldsymbol{\Omega}}^{\text{cate}}\hat{\boldsymbol{\ell}}_g| \leq c \\ |y_{1_g} - \hat{\boldsymbol{\Omega}}^{\text{cate}}\hat{\boldsymbol{\ell}}_g|^{-1} & \text{if } |y_{1_g} - \hat{\boldsymbol{\Omega}}^{\text{cate}}\hat{\boldsymbol{\ell}}_g| > c \end{cases} \quad (g = 1, \ldots, p).$$

We start by assuming $\lambda_K \to \infty$. When this is true, it suffices to show that

$$\max_{g\in[p]} |y_{1_g} - \hat{\boldsymbol{\Omega}}^{\text{shrunk}}\hat{\boldsymbol{\ell}}_g| = o_{\mathrm{p}}(1).$$

We see that

$$y_{1_g} - \hat{\boldsymbol{\Omega}}^{\text{shrunk}}\hat{\boldsymbol{\ell}}_g = \hat{\boldsymbol{\Omega}}^{\text{shrunk}}\left(\boldsymbol{\ell}_g - \hat{\boldsymbol{\ell}}_g\right) + \left(\boldsymbol{\Omega} - \hat{\boldsymbol{\Omega}}^{\text{shrunk}}\right)\boldsymbol{\ell}_g + e_{1_g}$$

where

$$\max_{g \in [p]} |e_{1_g}| = O_{\mathrm{p}} \left\{ n^{-1/2} \log(p) \right\} = o_{\mathrm{p}}(1)$$

and

$$\max_{g \in [p]} |\left(\boldsymbol{\Omega} - \hat{\boldsymbol{\Omega}}^{\text{shrunk}} \right) \boldsymbol{\ell}_g| = o_{\mathrm{p}}(1)$$

because the entries of $\boldsymbol{L}$ are uniformly bounded and $\|\boldsymbol{\Omega} - \hat{\boldsymbol{\Omega}}^{\text{shrunk}}\|_2 = o_{\mathrm{p}}(1)$. To complete the proof, we need only show that

$$\|\hat{\boldsymbol{L}} - \boldsymbol{L}\|_\infty = o_{\mathrm{p}}(1).$$

By the proof of Proposition 2.5,

$$\hat{\boldsymbol{L}} - \boldsymbol{L} = \boldsymbol{L} \left(I_K - \hat{\boldsymbol{V}} \right) + n^{-1/2} \tilde{\boldsymbol{E}}_1 \hat{\boldsymbol{V}} + n^{-1/2} \tilde{\boldsymbol{E}}_2 \hat{\boldsymbol{Z}}.$$

Since $\|I_K - \hat{\boldsymbol{V}}\|_2 = o_{\mathrm{p}}(1)$, $\|\boldsymbol{L} \left(I_K - \hat{\boldsymbol{V}} \right)\|_\infty = o_{\mathrm{p}}(1)$. Next,

$$\|n^{-1/2} \tilde{\boldsymbol{E}}_1\|_\infty = O_{\mathrm{p}} \left\{ n^{-1/2} \log(p) \right\} = o_{\mathrm{p}}(1).$$

For the last term, define the random variable $Z_g = \sigma_g^{-2} \left[\tilde{\boldsymbol{E}}_2 \right]_{g*}^{\mathrm{T}} \left[\tilde{\boldsymbol{E}}_2 \right]_{g*}$. Since this is a sub-exponential random variable with parameters $\{4(n-d-K), 4\}$,

$$\mathbb{P} \left\{ Z_g \geq (n-d-K) + t\sqrt{n} \right\} \leq \exp\left(-bt^2 \right)$$

for some constant $b > 0$, provided $t = o\left(n^{1/2} \right)$. If we let $t = b' \{\log(p)\}^{1/2}$ for some constant

$b' > b^{-1}$, then

$$\mathbb{P}\left\{\max_{g\in[p]} Z_g \geq (n-d-K) + t\sqrt{n}\right\} \leq 1 - \left\{1 - \exp\left(-bt^2\right)\right\}^p$$

where

$$p \log\left\{1 - \exp\left(-bt^2\right)\right\} = -\exp\left\{\log(p)\left(1 - bb'\right)\right\}\left\{1 + o(1)\right\} = o(1).$$

Therefore, for each $k \in [K]$,

$$\max_{g\in[p]} |n^{-1/2}\left[\tilde{\boldsymbol{E}}_2\right]_{g*}^{\mathrm{T}} \hat{\boldsymbol{z}}_k| \leq O_{\mathrm{p}}\left\{n^{1/2}\left(\lambda_K p\right)^{-1/2}\right\} \max_{g\in[p]}\left(n^{-1}\left[\tilde{\boldsymbol{E}}_2\right]_{g*}^{\mathrm{T}}\left[\tilde{\boldsymbol{E}}_2\right]_{g*}\right)^{1/2} = o_{\mathrm{p}}(1)$$

which completes the proof when $\lambda_K \to \infty$.

When $\lambda_1 = O(1)$, the results $\|\boldsymbol{L} - \hat{\boldsymbol{L}}\|_\infty, \|\boldsymbol{E}_1\|_\infty = o_{\mathrm{p}}(1)$ still hold. We therefore need only understand how $\left(\boldsymbol{\Omega} - \hat{\boldsymbol{\Omega}}^{\text{shrunk}}\right)\boldsymbol{\ell}_g$ behaves. By assumption, there exists a constant $m > 0$ that does not depend on p such that

$$\max_{g\in[p]} \|\boldsymbol{\ell}_g\|_2 \leq m.$$

Define $\delta_1 = c/(2m)$, where c was defined in (2.43). If $\|\boldsymbol{\Omega}\|_2 \leq \delta_1$, then because

$$\hat{\boldsymbol{\Omega}}^{\text{shrunk}} = \boldsymbol{\Omega}\operatorname{diag}\left\{\lambda_1/\left(\lambda_1+\rho\right), \cdots, \lambda_K/\left(\lambda_K+\rho\right)\right\} + o_{\mathrm{p}}\left(n^{-1/2}\right),$$

we get that

$$\boldsymbol{\Omega} - \hat{\boldsymbol{\Omega}}^{\text{cate}} = \boldsymbol{\Omega}\operatorname{diag}\left\{\rho/\left(\lambda_1+\rho\right), \cdots, \rho/\left(\lambda_K+\rho\right)\right\} + o_{\mathrm{p}}\left(n^{-1/2}\right).$$

If $\|\boldsymbol{\Omega}\|_2 > \delta_1$, suppose we initialize the optimization problem with $\boldsymbol{\alpha}_1 \in \mathbb{R}^{1\times K}$ such that

$\|\boldsymbol{\Omega} - \boldsymbol{\alpha}_1\|_2 \leq \delta_1$. Then the next iteration will be

$$\boldsymbol{\alpha}_2 = \boldsymbol{Y}_1^{\mathrm{T}} \hat{\boldsymbol{L}} \left(\hat{\boldsymbol{L}}^{\mathrm{T}} \hat{\boldsymbol{L}} \right)^{-1} = \boldsymbol{\Omega} \operatorname{diag} \left\{ \lambda_1 / \left(\lambda_1 + \rho \right), \cdots, \lambda_K / \left(\lambda_K + \rho \right) \right\} + o_{\mathrm{p}} \left(n^{-1/2} \right)$$

with probability tending to 1, where

$$\boldsymbol{\Omega} - \boldsymbol{\alpha}_2 = \boldsymbol{\Omega} \operatorname{diag} \left\{ \rho / \left(\lambda_1 + \rho \right), \cdots, \rho / \left(\lambda_K + \rho \right) \right\} + o_{\mathrm{p}} \left(n^{-1/2} \right).$$

Therefore,

$$\|\boldsymbol{\Omega} - \boldsymbol{\alpha}_2\|_2 \geq \|\boldsymbol{\Omega}\|_2 \rho / \left(\lambda_1 + \rho \right) \left\{ 1 + o_{\mathrm{p}} \left(n^{-1/2} \right) \right\} \geq \delta_2 + o_{\mathrm{p}} \left(n^{-1/2} \right)$$

for some constant $\delta_2 > 0$ not dependent n or p, since $\lambda_1 = O(1)$ by assumption. Note that we may assume $\delta_1 > \delta_2$. Therefore,

$$\|\hat{\boldsymbol{\Omega}}^{\mathrm{cate}} - \boldsymbol{\Omega}\|_2 \geq \delta_2 + o_{\mathrm{p}} \left(1 \right)$$

which completes the proof. □

***Remark* 2.11.** *The above proof shows that the behavior of Huber's loss function is very dependent on the constant c used in* (2.43) *when $\lambda_1 = O(1)$, meaning we cannot predict its behavior. This is an additional reason why this loss function should not be used to estimate $\boldsymbol{\Omega}$ when the data are only moderately informative for $\boldsymbol{C}$.*

***Corollary* 2.2** (The results of Corollary 2.1 hold using dSVA and CATE-RR)**.** *Suppose the assumptions of Proposition 2.2 hold with $d = 1$ and for some fixed $g \in [p]$, define*

$$\hat{\beta}_g^{\mathrm{dSVA}} = y_{1_g} - \hat{\boldsymbol{\Omega}}^{\mathrm{dSVA}} \hat{\boldsymbol{\ell}}_g$$
$$\hat{\beta}_g^{\mathrm{cate}} = y_{1_g} - \hat{\boldsymbol{\Omega}}^{\mathrm{cate}} \hat{\boldsymbol{\ell}}_g.$$

In addition, suppose $K = 1$ and

(i) $n^{-r}p \to 0$ *for some* $r > 0$ *as* $n \to \infty$.

(ii) $np^{-1}\boldsymbol{L}^{\mathrm{T}} P_{1_p}^{\perp} \boldsymbol{L} \geq \delta\lambda_1$ *for some constant* $\delta > 0$

(iii) $|\boldsymbol{\Omega}\boldsymbol{\ell}_g| \geq \epsilon$ *for some constant* $\epsilon > 0$.

Then the results of Corollary 2.1 hold for the z-score

$$z_g^{\mathrm{dSVA}} = \sigma_g^{-1}\left(\|\boldsymbol{X}\|_2^{-2} + n^{-1}\|\hat{\boldsymbol{\Omega}}^{\mathrm{dSVA}}\|_2^2\right)^{-1/2} \hat{\beta}_g^{\mathrm{dSVA}}.$$

If $\lambda_1 \to \infty$, *then the results of Corollary 2.1 hold for the z-score*

$$z_g^{\mathrm{cate}} = \sigma_g^{-1}\left(\|\boldsymbol{X}\|_2^{-2} + n^{-1}\|\hat{\boldsymbol{\Omega}}^{\mathrm{cate}}\|_2^2\right)^{-1/2} \hat{\beta}_g^{\mathrm{cate}}.$$

Proof. The proof is identical to the proof of Corollary 2.1 and is omitted. □

***Remark* 2.12.** *We require* $\lambda_1 \to \infty$ *to prove Proposition 2.2 for z-scores returned by CATE-RR because the behavior of* $\hat{\boldsymbol{\Omega}}^{\mathrm{cate}}$ *depends heavily on the constant* c *chosen in* (2.43) *when* $\lambda_1 = O(1)$.

2.7.9 A framework for when $\boldsymbol{C}$ is treated as a random variable and the proof of Theorem 2.3

Next, we provide a framework to extend all of our theoretical results to the case when $\boldsymbol{C}$ is treated as a random variable. We then prove Theorem 2.3 at the end of this section. First, we prove a proposition regarding the identifiability of factor models when $\boldsymbol{C}$ is random.

***Proposition* 2.7.** *Suppose* $\boldsymbol{Y} = \boldsymbol{\beta}\boldsymbol{X}^{\mathrm{T}} + \bar{\boldsymbol{L}}\bar{\boldsymbol{C}}^{\mathrm{T}} + \boldsymbol{E}$ *where* $\boldsymbol{\beta} \in \mathbb{R}^{p \times d}$ *and* $\bar{\boldsymbol{L}} \in \mathbb{R}^{p \times K}$ *are fixed effects,* $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ *is observed and*

(i) $\boldsymbol{X}$ *has full column rank.*

(ii) $\bar{\boldsymbol{C}} \in \mathbb{R}^{n\times K}$ *is such that* $\mathbb{E}\left(\bar{\boldsymbol{C}}\right) = \boldsymbol{X}\bar{\boldsymbol{A}}$ *and* $\mathbb{V}\left\{\mathrm{vec}\left(\bar{\boldsymbol{C}}\right)\right\} = \bar{\boldsymbol{\Psi}} \otimes I_n$*, where* $\bar{\boldsymbol{A}} \in \mathbb{R}^{d\times K}$ *and* $\bar{\boldsymbol{\Psi}} \succ 0$.

(iii) $\boldsymbol{E} \in \mathbb{R}^{p\times n}$ *is independent of* $\bar{\boldsymbol{C}}$ *and* $\mathbb{V}\left\{\mathrm{vec}\left(\boldsymbol{E}\right)\right\} = I_n \otimes \boldsymbol{\Sigma}$*, where for* $\sigma_1^2, \ldots, \sigma_p^2 > 0$, $\boldsymbol{\Sigma} = \mathrm{diag}\left(\sigma_1^2, \ldots, \sigma_p^2\right)$.

(iv) *If any row is removed from* $\bar{\boldsymbol{L}}$*, there exists two sub-matrices with rank* K.

Then $\bar{\boldsymbol{L}}\bar{\boldsymbol{\Psi}}\bar{\boldsymbol{L}}^{\mathrm{T}}$ *and* $\boldsymbol{\Sigma}$ *are identifiable.*

Proof. Define $\bar{\boldsymbol{C}}_2 = \boldsymbol{Q}_X^{\mathrm{T}}\bar{\boldsymbol{C}}$ and $\boldsymbol{Y}_2 = \boldsymbol{Y}\boldsymbol{Q}_X$, where the columns of $\boldsymbol{Q}_X \in \mathbb{R}^{n\times(n-d)}$ form an orthonormal basis for the null space of $\boldsymbol{X}^{\mathrm{T}}$. Then $\mathbb{E}\left(\bar{\boldsymbol{C}}_2\right) = 0$, $\mathbb{V}\left\{\mathrm{vec}\left(\bar{\boldsymbol{C}}\right)\right\} = \bar{\boldsymbol{\Psi}} \otimes I_{n-d}$ and $\mathbb{V}\left\{\mathrm{vec}\left(\boldsymbol{Y}_2\right)\right\} = I_{n-d} \otimes \left(\boldsymbol{\Sigma} + \bar{\boldsymbol{L}}\bar{\boldsymbol{\Psi}}\bar{\boldsymbol{L}}^{\mathrm{T}}\right)$. The identifiability of $\bar{\boldsymbol{L}}\bar{\boldsymbol{\Psi}}\bar{\boldsymbol{L}}^{\mathrm{T}}$ and $\boldsymbol{\Sigma}$ then follows from Theorem 5.1 of (Anderson & Rubin 1956). □

***Corollary* 2.3.** *Let* $c > 1$ *be a constant. Suppose that in addition to the assumptions in Proposition 2.7, the following hold*

(i) p *is a non-decreasing function of* n.

(ii) $\left[\bar{\boldsymbol{L}}\right]_{g*}^{\mathrm{T}} \bar{\boldsymbol{\Psi}} \left[\bar{\boldsymbol{L}}\right]_{g*} \leq c$ *for all* $g \in [p]$.

(iii) *There are* K *non-zero eigenvalues of* $\bar{\boldsymbol{L}}\bar{\boldsymbol{\Psi}}\bar{\boldsymbol{L}}^{\mathrm{T}}$ $\gamma_1, \ldots, \gamma_K$ *such that* $c^{-1} \leq \gamma_1 \leq \cdots \leq \gamma_K \leq cn$.

(iv) *For all* $r \in [d]$, $p^{-1} \sum\limits_{g=1}^{p} I\left(\left[\boldsymbol{\beta}\right]_{gr} \neq 0\right) = o\left(n^{-1}\gamma_K\right)$.

Then $\boldsymbol{\beta}$ *is identifiable for all* $n \geq c'$*, where* $c' > 0$ *is a constant.*

Proof. Define $\boldsymbol{Y}_1 = \boldsymbol{Y}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}$, where

$$\mathbb{E}\left(\boldsymbol{Y}_1\right) = \boldsymbol{\beta} + \bar{\boldsymbol{L}}\bar{\boldsymbol{A}}^{\mathrm{T}} = \boldsymbol{\beta} + \left(\bar{\boldsymbol{L}}\bar{\boldsymbol{\Psi}}^{1/2}\right)\left(\bar{\boldsymbol{A}}\bar{\boldsymbol{\Psi}}^{-1/2}\right)^{\mathrm{T}}.$$

The identical method used to prove $\boldsymbol{\beta}$ was identifiable in Proposition 2.1 can then be used to show $\boldsymbol{\beta}$ is identifiable here. □

***Remark* 2.13.** *If Items (ii) and (iii) from the statement of Corollary 2.3 hold for some* $\bar{\boldsymbol{L}} \in \mathbb{R}^{p\times K}$ *and* $\bar{\boldsymbol{\Psi}} \succ 0$, *then Item* (iv) *from the statement of Proposition 2.7 holds for all* n, p *suitably large. Therefore, another way to express Item* (ii) *from the statement of Theorem 2.3 is to first assume Item* (iv) *from the statement of Proposition 2.7 holds to identify* $\boldsymbol{\Sigma}$ *and* $\boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}$ *(and therefore* $\lambda_1, \ldots, \lambda_K$*), and then assume* $\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}$ *is orthogonal with decreasing elements, since this will not affect the isotropic distribution assumption on* $\boldsymbol{C}$ *(or any uniform bound on the fourth moments of its entries).*

***Remark* 2.14.** *We do not need Corollary 2.3 to prove Theorem 2.3. We state it to show that our theoretical results from Sections 2.3.2, 2.3.3 and 2.3.4 can be extended to the case when* $\boldsymbol{C}$ *is treated as a random variable.*

Next, we state and prove a technical lemma to be used in the proof of Theorem 2.3. This lemma is also important because it shows that we can generalize Assumption 2.2 to the case when $\boldsymbol{C}$ is a random variable.

***Lemma* 2.7.** *Let* $a > 1$ *be a constant not dependent on* n *or* p, *suppose* $\boldsymbol{Y} = \bar{\boldsymbol{L}}\bar{\boldsymbol{C}}^{\mathrm{T}} + \boldsymbol{E}$ *where* $\bar{\boldsymbol{L}} \in \mathbb{R}^{p\times K}$, $\bar{\boldsymbol{C}} \in \mathbb{R}^{n\times K}$ *and* $\boldsymbol{E} \in \mathbb{R}^{p\times n}$ *and assume Items* (ii), (iii) *and* (iv) *from the statement of Proposition 2.7 hold. Define* $\gamma_1, \ldots, \gamma_K$ *to be the eigenvalues of* $np^{-1}\bar{\boldsymbol{\Psi}}^{1/2}\bar{\boldsymbol{L}}^{\mathrm{T}}\bar{\boldsymbol{L}}\bar{\boldsymbol{\Psi}}^{1/2}$ *with eigenvectors* $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_K \in \mathbb{R}^K$ *and assume the following hold*

(i) $\boldsymbol{E} \sim MN_{p\times n}\left(0, \boldsymbol{\Sigma}, I_n\right)$ *where* $\sigma_g^2 \in \left[a^{-1}, a\right]$ *for all* $g \in [p]$.

(ii) $\|n^{-1}\bar{\boldsymbol{C}}^{\mathrm{T}}\bar{\boldsymbol{C}} - \bar{\boldsymbol{\Psi}}\|_2 = O_{\mathrm{p}}\left(n^{-1/2}\right)$.

(iii) *The magnitude of the entries of* $\bar{\boldsymbol{L}}$ *are uniformly bounded by* a.

(iv) $a^{-1} \le \gamma_K < \cdots < \gamma_1 \le an$ *and* $\left(\gamma_k - \gamma_{k+1}\right)\gamma_k^{-1} \ge a^{-1}$ $(k = 1, \ldots, K)$ *where* γ_{K+1} *is defined to be 0.*

(v) $|\boldsymbol{u}_r^{\mathrm{T}}\left(np^{-1}\bar{\boldsymbol{\Psi}}^{1/2}\bar{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{\Sigma}\bar{\boldsymbol{L}}\bar{\boldsymbol{\Psi}}^{1/2}\right)\boldsymbol{u}_s| \le a\gamma_{\max(r,s)}$ $\quad$ *(r=1,. . . ,K; s=1,. . . ,K) .*

Then there exists an $\boldsymbol{L} \in \mathbb{R}^{p\times K}$, $\boldsymbol{C} \in \mathbb{R}^{n\times K}$ and constant $c > 1$ such that the following hold:

(1) $\bar{\boldsymbol{L}}\bar{\boldsymbol{C}}^{\mathrm{T}} = \boldsymbol{L}\boldsymbol{C}^{\mathrm{T}}$ *such that* $P_{\bar{C}} = P_C$, $n^{-1}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C} = I_K$ *and* $\sup\limits_{g\in[p],k\in[K]} |[\boldsymbol{L}]_{gk}| \leq c + o_{\mathrm{p}}(1)$.

(2) $\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}$ *is a diagonal matrix with decreasing entries* $\lambda_1, \ldots, \lambda_K$ *such that* λ_k *is the* k*th largest eigenvalue of* $\bar{\boldsymbol{C}}\left(p^{-1}\bar{\boldsymbol{L}}^{\mathrm{T}}\bar{\boldsymbol{L}}\right)\bar{\boldsymbol{C}}$ *(*$k = 1, \ldots, K$*).*

(3) $1 - \lambda_k\gamma_k^{-1} = O_{\mathrm{p}}\left(n^{-1/2}\right)$ *and* $\left(\lambda_k - \lambda_{k+1}\right)\lambda_k^{-1} \geq c^{-1} + O_{\mathrm{p}}\left(n^{-1/2}\right)$ *(*$k = 1, \ldots, K$*) where* λ_{K+1} *is defined to be 0.*

(4) $n\left\{p\lambda_{\max(r,s)}\right\}^{-1}[\boldsymbol{L}]_{*r}^{\mathrm{T}}\,\boldsymbol{\Sigma}\,[\boldsymbol{L}]_{*s} = O_{\mathrm{p}}(1)$ *(r=1,...,K; s=1,...,K).*

Proof. We first re-define $\bar{\boldsymbol{L}}$ as $\bar{\boldsymbol{L}}\bar{\boldsymbol{\Psi}}^{1/2}$ and $\bar{\boldsymbol{C}}$ as $\bar{\boldsymbol{C}}\bar{\boldsymbol{\Psi}}^{-1/2}$, meaning we now have $\|n^{-1}\bar{\boldsymbol{C}}^{\mathrm{T}}\bar{\boldsymbol{C}} - I_K\|_2 = O_{\mathrm{p}}\left(n^{-1/2}\right)$. Define $\hat{\boldsymbol{R}}$ such that $\hat{\boldsymbol{R}}^2 = n^{-1}\bar{\boldsymbol{C}}^{\mathrm{T}}\bar{\boldsymbol{C}}$ and

$$\boldsymbol{L} = \bar{\boldsymbol{L}}\hat{\boldsymbol{R}}\hat{\boldsymbol{U}}$$
$$\boldsymbol{C} = \bar{\boldsymbol{C}}\hat{\boldsymbol{R}}^{-1}\hat{\boldsymbol{U}}$$

where the columns of $\hat{\boldsymbol{U}} \in \mathbb{R}^{K\times K}$ contain the right singular vectors of $\bar{\boldsymbol{L}}\hat{\boldsymbol{R}}$. Since $n^{-1}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C} = I_K$, this proves (1) and (2).

To then prove (3) and (4), we study the eigenvalues and eigenvectors of $np^{-1}\hat{\boldsymbol{R}}^{\mathrm{T}}\bar{\boldsymbol{L}}^{\mathrm{T}}\bar{\boldsymbol{L}}^{\mathrm{T}}\hat{\boldsymbol{R}}$. We can write $np^{-1}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}$ (whose diagonal elements are the eigenvalues of $np^{-1}\hat{\boldsymbol{R}}^{\mathrm{T}}\bar{\boldsymbol{L}}^{\mathrm{T}}\bar{\boldsymbol{L}}^{\mathrm{T}}\hat{\boldsymbol{R}}$) as

$$np^{-1}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L} = \hat{\boldsymbol{U}}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\hat{\boldsymbol{R}}\boldsymbol{U}\mathrm{diag}\left(\gamma_1, \ldots, \gamma_K\right)\boldsymbol{U}^{\mathrm{T}}\hat{\boldsymbol{R}}\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\hat{\boldsymbol{U}} = \hat{\boldsymbol{U}}^{\mathrm{T}}\hat{\boldsymbol{F}}\mathrm{diag}\left(\gamma, \ldots, \gamma_K\right)\hat{\boldsymbol{F}}\hat{\boldsymbol{U}}$$

where $\boldsymbol{U} = (\boldsymbol{u}_1 \cdots \boldsymbol{u}_K) \in \mathbb{R}^{K\times K}$ and $\hat{\boldsymbol{F}} = \boldsymbol{U}^{\mathrm{T}}\hat{\boldsymbol{R}}\boldsymbol{U}$ where the diagonal entries of $\hat{\boldsymbol{F}}$ are $1 + O_{\mathrm{p}}\left(n^{-1/2}\right)$ and the off-diagonal entries are $O_{\mathrm{p}}\left(n^{-1/2}\right)$. We have also re-defined $\hat{\boldsymbol{U}}$ as $\boldsymbol{U}^{\mathrm{T}}\hat{\boldsymbol{U}}$, which is still a random unitary matrix. Define the matrix $\boldsymbol{A} = \hat{\boldsymbol{F}}\mathrm{diag}\left(\gamma_1, \ldots, \gamma_K\right)\hat{\boldsymbol{F}} \in$

$\mathbb{R}^{K\times K}$ where

$$[\boldsymbol{A}]_{kk} = \gamma_k \left\{1 + O_{\mathrm{p}}\left(n^{-1/2}\right)\right\} + \sum_{r\neq k} \gamma_r O_{\mathrm{p}}\left(n^{-1}\right) \quad (k = 1, \ldots, K)$$

$$[\boldsymbol{A}]_{rs} = (\gamma_r + \gamma_s)\, O_{\mathrm{p}}\left(n^{-1/2}\right) + \sum_{k\neq r,s} \gamma_k O_{\mathrm{p}}\left(n^{-1}\right) \quad (r = 1, \ldots, K; s = 1, \ldots, K; r \neq s).$$

Next, define $\boldsymbol{A}^{(1)} = \gamma_1^{-1}\boldsymbol{A}$ where

$$\left[\boldsymbol{A}^{(1)}\right]_{kk} = \frac{\gamma_k}{\gamma_1}\left\{1 + O_{\mathrm{p}}\left(n^{-1/2}\right)\right\} + \sum_{r\neq k} \frac{\gamma_r}{\gamma_1} O_{\mathrm{p}}\left(n^{-1}\right) \quad (k = 1, \ldots, K)$$

$$\left[\boldsymbol{A}^{(1)}\right]_{rs} = \frac{\gamma_r + \gamma_s}{\gamma_1} O_{\mathrm{p}}\left(n^{-1/2}\right) + \sum_{k\neq r,s} \frac{\gamma_k}{\gamma_1} O_{\mathrm{p}}\left(n^{-1}\right) \quad (r = 2, \ldots, K; s = 2, \ldots, K; r \neq s).$$

We first decompose $\boldsymbol{A}^{(1)}$ into K rank (approximately) 1 matrices to study the behavior of the eigenvalues and eigenvectors of $\boldsymbol{A}$. We see that

$$\boldsymbol{A}^{(1)} = \underbrace{\begin{bmatrix} \left[\boldsymbol{A}^{(1)}\right]_{11} & \left[\boldsymbol{A}^{(1)}\right]_{12} + \omega_2 & \cdots & \left[\boldsymbol{A}^{(1)}\right]_{1K} + \omega_K \\ \left[\boldsymbol{A}^{(1)}\right]_{12} + \omega_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \left[\boldsymbol{A}^{(1)}\right]_{1K} + \omega_K & 0 & \cdots & 0 \end{bmatrix}}_{=D_1}$$

$$+ \underbrace{\begin{bmatrix} 0 & -\omega_2 & 0 & \cdots & 0 \\ -\omega_2 & \left[\boldsymbol{A}^{(1)}\right]_{22} & \left[\boldsymbol{A}^{(1)}\right]_{23} & \cdots & \left[\boldsymbol{A}^{(1)}\right]_{2K} \\ 0 & \left[\boldsymbol{A}^{(1)}\right]_{23} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \left[\boldsymbol{A}^{(1)}\right]_{2K} & 0 & \cdots & 0 \end{bmatrix}}_{=D_2} + \cdots$$

$$+ \underbrace{\begin{bmatrix} 0 & 0 & \cdots & -\omega_K \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -\omega_K & 0 & \cdots & \left[\boldsymbol{A}^{(1)}\right]_{KK} \end{bmatrix}}_{=D_K}$$

where we define

$$\omega_k = \frac{\left[\boldsymbol{A}^{(1)}\right]_{kk} \left[\boldsymbol{A}^{(1)}\right]_{1k}}{\left[\boldsymbol{A}^{(1)}\right]_{11} - \left[\boldsymbol{A}^{(1)}\right]_{kk}} = O_{\mathrm{p}}\left(\frac{\gamma_k}{\gamma_1} n^{-1/2}\right) \quad (k = 2, \ldots, K).$$

Let

$$\begin{aligned}
\boldsymbol{v}_1 = & \left[\left\{ \left[\boldsymbol{A}^{(1)} \right]_{11} \right\}^2 + \left\{ \left[\boldsymbol{A}^{(1)} \right]_{12} + \omega_2 \right\}^2 + \cdots + \left\{ \left[\boldsymbol{A}^{(1)} \right]_{1K} + \omega_K \right\}^2 \right]^{-1/2} \\
& \times \left(\left[\boldsymbol{A}^{(1)} \right]_{11}, \left[\boldsymbol{A}^{(1)} \right]_{12} + \omega_2, \cdots, \left[\boldsymbol{A}^{(1)} \right]_{1K} + \omega_K \right)^{\mathrm{T}} \\
= & \left(1, \frac{\left[\boldsymbol{A}^{(1)} \right]_{12} + \omega_2}{\left[\boldsymbol{A}^{(1)} \right]_{11}}, \cdots, \frac{\left[\boldsymbol{A}^{(1)} \right]_{1K} + \omega_K}{\left[\boldsymbol{A}^{(1)} \right]_{11}} \right)^{\mathrm{T}} + O_{\mathrm{p}} \left(n^{-1} \right).
\end{aligned}$$

Then

$$\begin{aligned}
\boldsymbol{A}^{(1)} \boldsymbol{v}_1 = & \boldsymbol{D}_1 \boldsymbol{v}_1 + \boldsymbol{D}_2 \boldsymbol{v}_1 + \cdots + \boldsymbol{D}_K \boldsymbol{v}_1 = \left\{ \begin{array}{c} \left[\boldsymbol{A}^{(1)} \right]_{11} + O_{\mathrm{p}} \left(n^{-1} \right) \\ \frac{\left[\boldsymbol{A}^{(1)} \right]_{12} + \omega_2}{\left[\boldsymbol{A}^{(1)} \right]_{11}} + O_{\mathrm{p}} \left(n^{-1} \right) \\ \cdots \\ \frac{\left[\boldsymbol{A}^{(1)} \right]_{1K} + \omega_K}{\left[\boldsymbol{A}^{(1)} \right]_{11}} + O_{\mathrm{p}} \left(n^{-1} \right) \end{array} \right\} \\
& + \left\{ \begin{array}{c} O_{\mathrm{p}} \left(n^{-1} \right) \\ \underbrace{-\omega_2 + \left[\boldsymbol{A}^{(1)} \right]_{22} \frac{\left[\boldsymbol{A}^{(1)} \right]_{12} + \omega_2}{\left[\boldsymbol{A}^{(1)} \right]_{11}}}_{=0} + O_{\mathrm{p}} \left(n^{-1} \right) \\ \cdots \\ O_{\mathrm{p}} \left(n^{-1} \right) \end{array} \right\} \\
& + \cdots + \left\{ \begin{array}{c} O_{\mathrm{p}} \left(n^{-1} \right) \\ \cdots \\ O_{\mathrm{p}} \left(n^{-1} \right) \\ \underbrace{-\omega_K + \left[\boldsymbol{A}^{(1)} \right]_{KK} \frac{\left[\boldsymbol{A}^{(1)} \right]_{1K} + \omega_K}{\left[\boldsymbol{A}^{(1)} \right]_{11}}}_{=0} + O_{\mathrm{p}} \left(n^{-1} \right) \end{array} \right\}
\end{aligned}$$

and

$$
\begin{aligned}
\delta_1 &= \boldsymbol{v}_1^{\mathrm{T}} \boldsymbol{A}^{(1)} \boldsymbol{v}_1 = \left[\boldsymbol{A}^{(1)}\right]_{11} + O_{\mathrm{p}}\left(n^{-1}\right) \\
\|\boldsymbol{A}^{(1)} \boldsymbol{v}_1 - \delta_1 \boldsymbol{v}_1\|_2 &= O_{\mathrm{p}}\left(n^{-1}\right).
\end{aligned}
$$

By Weyl's Theorem and Theorem 3.6 in Auffinger & Tang (2015) the largest eigenvalue of $\boldsymbol{A}^{(1)}$ is $\hat{\mu}_1 = \left[\boldsymbol{A}^{(1)}\right]_{11} + O_{\mathrm{p}}\left(n^{-1}\right)$ with corresponding eigenvector $\hat{\boldsymbol{u}}_1$ such that $\|\hat{\boldsymbol{u}}_1 - \boldsymbol{v}_1\|_2 = O_{\mathrm{p}}\left(n^{-1}\right)$. To find the next eigenvalue and eigenvector of $\boldsymbol{A}$, we first have to remove the principal direction from $\boldsymbol{A}^{(1)}$:

$$
\boldsymbol{A}^{(1)} - \hat{\mu}_1 \hat{\boldsymbol{u}}_1 \hat{\boldsymbol{u}}_1^{\mathrm{T}} = \boldsymbol{D}_2 + \cdots + \boldsymbol{D}_K + O_{\mathrm{p}}\left(n^{-1}\right)
$$

and we define

$$
\begin{aligned}
\boldsymbol{A}^{(2)} =& \frac{\gamma_1}{\gamma_2}\left\{\boldsymbol{A}^{(1)} - \hat{\mu}_1 \hat{\boldsymbol{u}}_1 \hat{\boldsymbol{u}}_1^{\mathrm{T}}\right\} = \frac{\gamma_1}{\gamma_2}\boldsymbol{D}_2 + \cdots + \frac{\gamma_1}{\gamma_2}\boldsymbol{D}_K + O_{\mathrm{p}}\left(\frac{\gamma_1}{\gamma_2 n}\right) \\
=& \begin{bmatrix}
0 & -\frac{\gamma_1}{\gamma_2}\omega_2 & 0 & \cdots & 0 \\
-\frac{\gamma_1}{\gamma_2}\omega_2 & \left[\boldsymbol{A}^{(2)}\right]_{22} & \left[\boldsymbol{A}^{(2)}\right]_{23} & \cdot & \left[\boldsymbol{A}^{(2)}\right]_{2K} \\
0 & \left[\boldsymbol{A}^{(2)}\right]_{23} & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & \left[\boldsymbol{A}^{(2)}\right]_{2K} & 0 & \cdots & 0
\end{bmatrix} + \cdots \\
&+ \begin{bmatrix}
0 & 0 & \cdots & -\frac{\gamma_1}{\gamma_2}\omega_K \\
0 & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
-\frac{\gamma_1}{\gamma_2}\omega_K & 0 & \cdots & \left[\boldsymbol{A}^{(2)}\right]_{KK}
\end{bmatrix} + O_{\mathrm{p}}\left(\frac{\gamma_1}{\gamma_2 n}\right)
\end{aligned}
$$

where

$$
\begin{aligned}
\frac{\gamma_1}{\gamma_2}\omega_k &= O_{\mathrm{p}}\left(\frac{\gamma_k}{\gamma_2}n^{-1/2}\right) \quad (k=2,\dots,K)\\
\left[\boldsymbol{A}^{(2)}\right]_{kk} &= \frac{\gamma_k}{\gamma_2}\left\{1+O_{\mathrm{p}}\left(n^{-1/2}\right)\right\}+\sum_{r\neq k}\frac{\gamma_r}{\gamma_2}O_{\mathrm{p}}\left(n^{-1}\right) \quad (k=2,\dots,K)\\
\left[\boldsymbol{A}^{(2)}\right]_{rs} &= \frac{\gamma_r+\gamma_s}{\gamma_2}O_{\mathrm{p}}\left(n^{-1/2}\right)+\sum_{k\neq r,s}\frac{\gamma_k}{\gamma_2}O_{\mathrm{p}}\left(n^{-1}\right) \quad (r=2,\dots,K; s=2,\dots,K; r\neq s).
\end{aligned}
$$

A subsequent application of the above procedure will show that the largest eigenvalue of $\boldsymbol{A}^{(2)}$ is

$$
\hat{\mu}_2 = \left[\boldsymbol{A}^{(2)}\right]_{22} + O_{\mathrm{p}}\left(\frac{\gamma_1}{\gamma_2 n}\right)
$$

with eigenvalue $\hat{\boldsymbol{u}}_2$ such that

$$
\|\hat{\boldsymbol{u}}_2 - v_2\|_2 = O_{\mathrm{p}}\left(\frac{\gamma_1}{\gamma_2 n}\right),
$$

where v_2 is the second column of $\gamma_1\gamma_2^{-1}\boldsymbol{D}_2$. When we subsequently remove the second principal direction, we will remove $\gamma_1\gamma_2^{-1}\boldsymbol{D}_2$ and the $O_{\mathrm{p}}\left\{\gamma_1\left(\gamma_2 n\right)^{-1}\right\}$ error term will become $O_{\mathrm{p}}\left\{\gamma_1\left(\gamma_3 n\right)^{-1}\right\}$. Provided $\gamma_1\left(\gamma_k n\right)^{-1} \lesssim n^{-1/2}$, this procedure will give us estimates $\hat{\mu}_k$ and $\hat{\boldsymbol{u}}_k$ such that

$$
\lambda_k = \gamma_k\hat{\mu}_k = \gamma_k\left\{1+O_{\mathrm{p}}\left(n^{-1/2}\right)\right\} \tag{2.44}
$$

$$
\|\hat{\boldsymbol{u}}_k - \boldsymbol{e}_k\|_2 = O_{\mathrm{p}}\left(n^{-1/2}\right) \tag{2.45}
$$

where here $\boldsymbol{e}_k \in \mathbb{R}^K$ is the standard basis vector with 1 in the k^{th} position and 0's everywhere else.

We next handle the case when $\gamma_1\left(\gamma_k n\right)^{-1} \gtrsim n^{-1/2}$. Let $r \leq K$ be such that $\gamma_1\left(\gamma_k n\right)^{-1} \lesssim n^{-1/2}$ for $k \leq r$ and $\gamma_1\left(\gamma_k n\right)^{-1} \gtrsim n^{-1/2}$ for $k > r$. For these eigenvalues, we note that we can study the smallest eigenvalues and their eigenvectors of $\boldsymbol{A}$ by studying the largest eigenvalues of $\boldsymbol{A}^{-1}$. If λ_k is an eigenvalue of $\boldsymbol{A}$ with eigenvector $\hat{\boldsymbol{u}}_k$, then λ_k^{-1} is an eigenvalue of $\boldsymbol{A}^{-1}$ with the same eigenvector. We note that

$$\boldsymbol{A}^{-1} = \hat{\boldsymbol{F}}^{-1}\text{diag}\left(\gamma_1^{-1}, \dots, \gamma_K^{-1}\right)\hat{\boldsymbol{F}}^{-1} = \gamma_1^{-1}\hat{\boldsymbol{F}}^{-1}\text{diag}\left(1, \gamma_1\gamma_2^{-1}, \dots, \gamma_1\gamma_K^{-1}\right)\hat{\boldsymbol{F}}^{-1}$$

where the diagonal entries of $\hat{\boldsymbol{F}}^{-1}$ are $1 + O_{\text{p}}\left(n^{-1/2}\right)$ and the off-diagonal entries are $O_{\text{p}}\left(n^{-1/2}\right)$. If $k > r$, then $\gamma_1\gamma_k^{-1} \gtrsim n^{1/2}$, meaning $\gamma_k \lesssim n^{1/2}$, since $\gamma_1 \lesssim n$. Therefore,

$$\frac{\gamma_1\gamma_K^{-1}}{\gamma_1\gamma_k^{-1}} = \frac{\gamma_k}{\gamma_K} \lesssim n^{1/2}$$

for all $k > r$. By what we have shown above, the $K - k + 1$ eigenvalue of $\gamma_1\boldsymbol{A}^{-1}$ is $\gamma_1\gamma_k^{-1}\left\{1 + O_{\text{p}}\left(n^{-1/2}\right)\right\}$ with eigenvectors that satisfies (2.45). Therefore, the k^{th} eigenvalue of $\boldsymbol{A}$ is $\gamma_k\left\{1 + O_{\text{p}}\left(n^{-1/2}\right)\right\}$ with eigenvector that satisfies (2.45). This proves item (3).

To prove item (4),

$$np^{-1}\boldsymbol{L}^{\text{T}}\boldsymbol{\Sigma}\boldsymbol{L} = \hat{\boldsymbol{M}}^{\text{T}}\left\{np^{-1}\boldsymbol{U}^{\text{T}}\bar{\boldsymbol{L}}^{\text{T}}\boldsymbol{\Sigma}\bar{\boldsymbol{L}}\boldsymbol{U}\right\}\hat{\boldsymbol{M}} \tag{2.46}$$

where $\hat{\boldsymbol{M}} = \hat{\boldsymbol{F}}\hat{\boldsymbol{U}}$ is such that $\|\hat{\boldsymbol{M}} - I_K\|_2 = O_{\text{p}}\left(n^{-1/2}\right)$ by the analysis above. To evaluate

(2.46), we first see that

$$np^{-1}\boldsymbol{U}^{\mathrm{T}}\bar{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{\Sigma}\bar{\boldsymbol{L}}\boldsymbol{U} = \begin{bmatrix} O(\gamma_1) & O(\gamma_2) & \cdots & O(\gamma_K) \\ O(\gamma_2) & O(\gamma_2) & \cdots & O(\gamma_K) \\ \vdots & \vdots & \ddots & \vdots \\ O(\gamma_K) & O(\gamma_K) & \cdots & O(\gamma_K) \end{bmatrix} = \begin{bmatrix} O(\gamma_1) & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

$$+ \begin{bmatrix} 0 & O(\gamma_2) & \cdots & 0 \\ O(\gamma_2) & O(\gamma_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & \cdots & O(\gamma_K) \\ 0 & 0 & \cdots & O(\gamma_K) \\ \vdots & \vdots & \ddots & \vdots \\ O(\gamma_K) & O(\gamma_K) & \cdots & O(\gamma_K) \end{bmatrix}.$$

Fix some $r, s \leq K$ such that $r < s$. If $\hat{\boldsymbol{m}}_k$ is the k^{th} column of $\hat{\boldsymbol{M}}$, then

$$\hat{\boldsymbol{m}}_r^{\mathrm{T}}\left(np^{-1}\boldsymbol{U}^{\mathrm{T}}\bar{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{\Sigma}\bar{\boldsymbol{L}}\boldsymbol{U}\right)\hat{\boldsymbol{m}}_s = O_{\mathrm{p}}\left(\gamma_1 n^{-1}\right) + \cdots + O_{\mathrm{p}}\left(\gamma_{r-1} n^{-1}\right)$$

$$+\left\{O(\gamma_r)\,\hat{\boldsymbol{m}}_{r_r}\hat{\boldsymbol{m}}_{s_1} + O(\gamma_r)\,\hat{\boldsymbol{m}}_{r_r}\hat{\boldsymbol{m}}_{s_2} + \cdots + O(\gamma_r)\,\hat{\boldsymbol{m}}_{r_r}\hat{\boldsymbol{m}}_{s_r}\right\}$$

$$+O(\gamma_{r+1})\,\hat{\boldsymbol{m}}_{r_r}\hat{\boldsymbol{m}}_{s_{r+1}} + \cdots + O(\gamma_{s-1})\,\hat{\boldsymbol{m}}_{r_r}\hat{\boldsymbol{m}}_{s_{s-1}} + O_{\mathrm{p}}(\gamma_s).$$

Next, note that for any $k < s$, we have

$$0 = \hat{\boldsymbol{m}}_k^{\mathrm{T}}\mathrm{diag}(\gamma_1, \ldots, \gamma_K)\,\hat{\boldsymbol{m}}_s = \underbrace{\gamma_1\hat{\boldsymbol{m}}_{k_1}\hat{\boldsymbol{m}}_{s_1}}_{=O_{\mathrm{p}}(1)} + \cdots + \gamma_k\hat{\boldsymbol{m}}_{k_k}\hat{\boldsymbol{m}}_{s_k} + \cdots + \underbrace{\gamma_s\hat{\boldsymbol{m}}_{k_s}\hat{\boldsymbol{m}}_{s_s}}_{=O_{\mathrm{p}}(\gamma_s n^{-1/2})}$$

$$+ O_{\mathrm{p}}(1).$$

Therefore,

$$\gamma_k\hat{\boldsymbol{m}}_{s_k} = O_{\mathrm{p}}\left\{\max\left(\gamma_s n^{-1/2}, 1\right)\right\}$$

for all $k < s$. This also shows that

$$\gamma_r \hat{\boldsymbol{m}}_{r_r} \hat{\boldsymbol{m}}_{s_k} = O_{\mathrm{p}} \left\{ \max \left(\gamma_s n^{-1/2}, 1 \right) \right\}$$

for $k = 1, 2, \ldots, r$ and completes the proof.

$\square$

We now prove Theorem 2.3.

Proof of Theorem 2.3. To make notation consistent with the statement of Lemma 2.7, we first redefine $\boldsymbol{C}$, $\boldsymbol{\Omega}$, Ξ and $\boldsymbol{L}$ from the statement of Theorem 2.3 to be $\bar{\boldsymbol{C}}$, $\bar{\boldsymbol{\Omega}}$, $\bar{\boldsymbol{\Xi}}$ and $\bar{\boldsymbol{L}}$. Under the null hypothesis $\bar{\boldsymbol{\Omega}} = \boldsymbol{0}$, we define

$$\begin{aligned}
\hat{\bar{\boldsymbol{\Omega}}} &= \left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{X} \right)^{-1} \boldsymbol{X}^{\mathrm{T}} \bar{\boldsymbol{C}} = n^{-1/2} \left(n^{-1} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{X} \right)^{-1} \hat{s}_n \\
\hat{\boldsymbol{s}}_n &= n^{-1/2} \sum_{i=1}^{n} \boldsymbol{x}_i \bar{\boldsymbol{\xi}}_i^{\mathrm{T}}
\end{aligned}$$

Define $a = \mathrm{vec}\left(\mathbf{1}_d \otimes \bar{\boldsymbol{\xi}}_1 \right)$, where $\mathbf{1}_d \in \mathbb{R}^d$ is the vector of all ones, and $\varphi_a(t)$, $t \in \mathbb{R}^{dK \times dK}$, to be the characteristic function of a. Under the null hypothesis, the gradient of $\varphi_a(t)$ is 0 and the Hessian is $-\mathbf{1}_{d \times d} \otimes I_K$, where $\mathbf{1}_{d \times d} \in \mathbb{R}^{d \times d}$ is the matrix of all ones. Lastly, let $\boldsymbol{t} = \left(\boldsymbol{t}_1^{\mathrm{T}}, \ldots, \boldsymbol{t}_d^{\mathrm{T}} \right)^{\mathrm{T}}$, $\boldsymbol{t}_j \in \mathbb{R}^K$. If the magnitude of the entries of $\boldsymbol{X}$ are bounded above by x, we then have that

$$\begin{aligned}
\log \left\{ \varphi_{\mathrm{vec}(\hat{\boldsymbol{s}}_n)}(\boldsymbol{t}) \right\} &= \sum_{i=1}^{n} \log \left[\varphi_a \left\{ n^{-1/2} \begin{pmatrix} \boldsymbol{x}_i[1] \boldsymbol{t}_1 \\ \vdots \\ \boldsymbol{x}_i[d] \boldsymbol{t}_d \end{pmatrix} \right\} \right] \\
&= \sum_{i=1}^{n} \left[-(2n)^{-1} \boldsymbol{t}^{\mathrm{T}} \left\{ \left(\boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}} \right) \otimes I_K \right\} \boldsymbol{t} + o(n^{-1} x^2 \|\boldsymbol{t}\|_2^2) \right] \\
&= -2^{-1} \boldsymbol{t}^{\mathrm{T}} \left(\boldsymbol{\Sigma}_X \otimes I_K \right) \boldsymbol{t} + o(1).
\end{aligned}$$

where $\mathbf{\Sigma}_X = \lim_{n\to\infty} n^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$. Therefore,

$$\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{1/2}\hat{\bar{\mathbf{\Omega}}}\left\{n^{-1}\bar{\mathbf{\Xi}}^{\mathrm{T}}P_X^{\perp}\bar{\mathbf{\Xi}}\right\}^{-1/2} \xrightarrow{\mathcal{D}} MN_{d\times K}\left(0, I_d, I_K\right)$$

since $\|n^{-1}\bar{\mathbf{\Xi}}^{\mathrm{T}}P_X^{\perp}\bar{\mathbf{\Xi}} - I_K\|_2 = o_{\mathrm{p}}(1)$.

We next define $\mathbf{\Omega}$ be the that from the statement and proof of Lemma 2.7, i.e.

$$\mathbf{\Omega} = \hat{\bar{\mathbf{\Omega}}}\left\{n^{-1}\bar{\mathbf{\Xi}}^{\mathrm{T}}P_X^{\perp}\bar{\mathbf{\Xi}}\right\}^{-1/2}\hat{\boldsymbol{U}}$$

where $\hat{\boldsymbol{U}}$ is a unitary matrix ensuring that

$$\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L} = \hat{\boldsymbol{U}}^{\mathrm{T}}\left\{n^{-1}\bar{\mathbf{\Xi}}^{\mathrm{T}}P_X^{\perp}\bar{\mathbf{\Xi}}\right\}^{1/2}\bar{\boldsymbol{L}}^{\mathrm{T}}\bar{\boldsymbol{L}}\left\{n^{-1}\bar{\mathbf{\Xi}}^{\mathrm{T}}P_X^{\perp}\bar{\mathbf{\Xi}}\right\}^{1/2}\hat{\boldsymbol{U}}$$

is diagonal with decreasing elements. Since the assumptions of Lemma 2.7 hold with $\bar{\mathbf{\Psi}} = I_K$, it is then straightforward to adapt the proof of Lemma 2.6 to show that $n^{1/2}\|\mathbf{\Omega} - \hat{\mathbf{\Omega}}\| = o_{\mathrm{p}}(1)$ under the assumptions of Theorem 2.3. The result then follows by an application of Slutsky's Theorem.

□

2.7.10 Two technical lemmas used in the proof of Lemma 2.5

We now state and prove two technical lemmas are used in the proof of Lemma 2.5. For these two lemmas, we assume $\boldsymbol{Y}$ is distributed according to (2.28) (as it is in Lemmas 2.4 and 2.5).

Lemma 2.8. *Let $\boldsymbol{U} = (\boldsymbol{u}_1 \cdots \boldsymbol{u}_K)$, $\boldsymbol{V} = (\boldsymbol{v}_1 \cdots \boldsymbol{v}_K)$, $\boldsymbol{D} = diag(d_1, \ldots, d_K)$ and $\tilde{\boldsymbol{N}}$ be as defined in Lemmas 2.4 and 2.5 and suppose $np^{-1}\left[\boldsymbol{L}\right]_{*s}^{\mathrm{T}}\mathbf{\Sigma}\left[\boldsymbol{L}\right]_{*k} = O_{\mathrm{p}}\left(\lambda_k\right)$ for $s \le k$, where $s, k \in [K]$. Then*

$$\boldsymbol{u}_s^{\mathrm{T}}\mathbf{\Sigma}\boldsymbol{u}_k = O_{\mathrm{p}}\left(\lambda_k^{1/2}\lambda_s^{-1/2}\right).$$

Proof. We need to understand how

$$\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{U} = \boldsymbol{D}^{-1}\boldsymbol{V}^{\mathrm{T}}\tilde{\boldsymbol{N}}^{\mathrm{T}}\boldsymbol{\Sigma}\tilde{\boldsymbol{N}}\boldsymbol{V}\boldsymbol{D}^{-1}$$

behaves. First, let $\boldsymbol{R}_i\boldsymbol{R}_i^{\mathrm{T}} = \tilde{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{\Sigma}^i\tilde{\boldsymbol{L}}$ for $i = 1, 2, 3$ and define $\gamma = p^{-1}\operatorname{tr}\left(\boldsymbol{\Sigma}^2\right)$. Then

$$\boldsymbol{R}_i = \begin{bmatrix} O\left(\lambda_1^{1/2}\right) & 0 & \cdots & 0 \\ O\left(\lambda_2^{1/2}\right) & O\left(\lambda_2^{1/2}\right) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ O\left(\lambda_K^{1/2}\right) & O\left(\lambda_K^{1/2}\right) & \cdots & O\left(\lambda_K^{1/2}\right) \end{bmatrix}$$

and

$$\tilde{\boldsymbol{N}}^{\mathrm{T}}\boldsymbol{\Sigma}\tilde{\boldsymbol{N}} = \tilde{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{\Sigma}\tilde{\boldsymbol{L}} + p^{-1/2}\tilde{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{\Sigma}\tilde{\boldsymbol{E}}_1 + p^{-1/2}\tilde{\boldsymbol{E}}_1^{\mathrm{T}}\boldsymbol{\Sigma}\tilde{\boldsymbol{L}} + \gamma I_K + O_{\mathrm{p}}\left(p^{-1/2}\right)$$

The next quantity we need to determine is $\boldsymbol{V}\boldsymbol{D}^{-1}$:

$$\begin{aligned}\boldsymbol{V}\boldsymbol{D}^{-1} =& \boldsymbol{D}^{-1} + \begin{bmatrix} O_{\mathrm{p}}\left(\lambda_1^{-3/2}p^{-1}\right) & O_{\mathrm{p}}\left\{(\lambda_1\lambda_2 p)^{-1/2}\right\} & \cdots & O_{\mathrm{p}}\left\{(\lambda_1\lambda_K p)^{-1/2}\right\} \\ O_{\mathrm{p}}\left\{(\lambda_1 p)^{-1}\right\} & O_{\mathrm{p}}\left(\lambda_2^{-3/2}p^{-1}\right) & \cdots & O_{\mathrm{p}}\left\{(\lambda_2\lambda_K p)^{-1/2}\right\} \\ \vdots & \vdots & \ddots & \vdots \\ O_{\mathrm{p}}\left\{(\lambda_1 p)^{-1}\right\} & O_{\mathrm{p}}\left\{(\lambda_2 p)^{-1}\right\} & \cdots & O_{\mathrm{p}}\left(\lambda_K^{-3/2}p^{-1}\right) \end{bmatrix} = \boldsymbol{D}^{-1} \\ &+ \boldsymbol{e}\end{aligned}$$

and

$$\boldsymbol{R}_i^{\mathrm{T}}\boldsymbol{V}\boldsymbol{D}^{-1} = O_{\mathrm{p}}(1) + O_{\mathrm{p}}\left\{(\lambda_K p)^{-1/2}\right\}.$$

Then for $\boldsymbol{M} \sim MN_{K\times K}\left(0, I_K, I_K\right)$, we have

$$p^{-1/2}\tilde{\boldsymbol{E}}_1^{\mathrm{T}}\boldsymbol{\Sigma}\tilde{\boldsymbol{L}}\boldsymbol{V}\boldsymbol{D}^{-1} \sim p^{-1/2}\boldsymbol{M}\boldsymbol{R}_3^{\mathrm{T}}\boldsymbol{V}\boldsymbol{D}^{-1} = O_{\mathrm{p}}\left(p^{-1/2}\right).$$

Next

$$\boldsymbol{D}^{-1}\boldsymbol{V}^{\mathrm{T}}\left(\tilde{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{\Sigma}\tilde{\boldsymbol{L}}+\gamma I_K\right)\boldsymbol{V}\boldsymbol{D}^{-1}=\boldsymbol{D}^{-1}\left(\tilde{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{\Sigma}\tilde{\boldsymbol{L}}+\gamma I_K\right)\boldsymbol{D}^{-1}+\boldsymbol{e}^{\mathrm{T}}\left(\tilde{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{\Sigma}\tilde{\boldsymbol{L}}+\gamma I_K\right)\boldsymbol{D}^{-1}$$
$$+\boldsymbol{D}^{-1}\left(\tilde{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{\Sigma}\tilde{\boldsymbol{L}}+\gamma I_K\right)\boldsymbol{e}+\boldsymbol{e}^{\mathrm{T}}\left(\tilde{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{\Sigma}\tilde{\boldsymbol{L}}+\gamma I_K\right)\boldsymbol{e}$$

where

$$\boldsymbol{e}^{\mathrm{T}}\tilde{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{\Sigma}\tilde{\boldsymbol{L}}\boldsymbol{e}=\boldsymbol{e}^{\mathrm{T}}\boldsymbol{R}_1\boldsymbol{R}_1^{\mathrm{T}}\boldsymbol{e}=O_{\mathrm{p}}\left\{(\lambda_K p)^{-1}\right\}$$

and

$$\boldsymbol{D}^{-1}\left(\tilde{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{\Sigma}\tilde{\boldsymbol{L}}+\gamma I_K\right)\boldsymbol{e}=\boldsymbol{D}^{-1}\boldsymbol{R}_1\boldsymbol{R}_1^{\mathrm{T}}\boldsymbol{e}+O_{\mathrm{p}}\left\{(\lambda_K p)^{-1/2}\right\}=O_{\mathrm{p}}\left\{(\lambda_K p)^{-1/2}\right\}.$$

The second equality holds because $\boldsymbol{D}^{-1}\boldsymbol{R}_1=O_{\mathrm{p}}(1)$ and $\boldsymbol{R}_1^{\mathrm{T}}\boldsymbol{e}=O_{\mathrm{p}}\left\{(\lambda_K p)^{-1/2}\right\}$. Next, let $\boldsymbol{A}=\tilde{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{\Sigma}\tilde{\boldsymbol{L}}$ and $\boldsymbol{B}=\boldsymbol{D}^{-1}\left(\boldsymbol{A}+\gamma I_K\right)\boldsymbol{D}^{-1}$. Then if $s\le k$, $[\boldsymbol{A}]_{sk}=O_{\mathrm{p}}\left(\lambda_k\right)$ by assumption and

$$[\boldsymbol{B}]_{sk}=\frac{[\boldsymbol{A}]_{sk}+\gamma\delta_{sk}}{d_s d_k}=O_{\mathrm{p}}\left(\frac{\lambda_k}{d_s d_k}\right)+\frac{\gamma}{d_s d_k}\delta_{sk}=O_{\mathrm{p}}\left(\lambda_k^{1/2}\lambda_s^{-1/2}\right)$$

where $\delta_{sk}=I(s=k)$. Therefore, for $s\le k$ $(s,k\in[K])$,

$$\left[\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{U}\right]_{sk}=O_{\mathrm{p}}\left\{\lambda_k^{1/2}\lambda_s^{-1/2}+(\lambda_K p)^{-1/2}\right\}=O_{\mathrm{p}}\left(\lambda_k^{1/2}\lambda_s^{-1/2}\right).$$

□

Lemma 2.9. *Let $\boldsymbol{a}_1,\boldsymbol{a}_2\in\mathbb{R}^p$ be linearly independent unit vectors independent of $\tilde{\boldsymbol{E}}_2\sim MN_{p\times(n-K)}\left(0,\boldsymbol{\Sigma},I_{n-K}\right)$ for K is a fixed constant. Recall from* (2.34) *that $\boldsymbol{R}=p^{-1}\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\tilde{\boldsymbol{E}}_2-$*

ρI_{n-K} *where* $\rho = p^{-1}\operatorname{tr}(\boldsymbol{\Sigma})$. *Then*

$$p^{-1}\boldsymbol{a}_1^{\mathrm{T}}\tilde{\boldsymbol{E}}_2\boldsymbol{R}\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\boldsymbol{a}_2 = O_{\mathrm{p}}\left\{\left(np^{-1}\right)^2 + np^{-3/2}\right\}.$$

Proof. Since K is a constant not dependent on n or p, I will assume $\tilde{\boldsymbol{E}}_2 \sim MN_{p\times n}\left(0, \boldsymbol{\Sigma}, I_n\right)$ for notational convenience.

$$p^{-1}\boldsymbol{a}_1^{\mathrm{T}}\tilde{\boldsymbol{E}}_2\boldsymbol{R}\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\boldsymbol{a}_2 = p^{-2}\boldsymbol{a}_1^{\mathrm{T}}\tilde{\boldsymbol{E}}_2\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\tilde{\boldsymbol{E}}_2\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\boldsymbol{a}_2 - \rho p^{-1}\boldsymbol{a}_1^{\mathrm{T}}\tilde{\boldsymbol{E}}_2\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\boldsymbol{a}_2$$

We will focus our efforts on understanding $p^{-2}\boldsymbol{a}_1^{\mathrm{T}}\tilde{\boldsymbol{E}}_2\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\tilde{\boldsymbol{E}}_2\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\boldsymbol{a}_2$. Define $\boldsymbol{A} = (\boldsymbol{a}_1\,\boldsymbol{a}_2)$, $\tilde{\boldsymbol{A}} = \boldsymbol{\Sigma}\boldsymbol{A}$ and $\boldsymbol{Q} \in \mathbb{R}^{p\times(p-2)}$ s.t. $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{Q} = \boldsymbol{0}_{2\times(p-2)}$. Let $P_{\tilde{A}} = \boldsymbol{G}\boldsymbol{G}^{\mathrm{T}}$ where $\boldsymbol{G} \in \mathbb{R}^{p\times 2}$ and $P_{\tilde{A}}^{\perp} = \boldsymbol{Q}\boldsymbol{Q}^{\mathrm{T}}$. Since $P_{\tilde{A}} + P_{\tilde{A}}^{\perp} = I_p$, we have

$$\begin{aligned}p^{-2}\boldsymbol{a}_1^{\mathrm{T}}\tilde{\boldsymbol{E}}_2\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\tilde{\boldsymbol{E}}_2\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\boldsymbol{a}_2 &= p^{-2}\boldsymbol{a}_1^{\mathrm{T}}\tilde{\boldsymbol{E}}_2\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\left(P_{\tilde{A}} + P_{\tilde{A}}^{\perp}\right)\tilde{\boldsymbol{E}}_2\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\boldsymbol{a}_2\\ &= p^{-2}\boldsymbol{a}_1^{\mathrm{T}}\tilde{\boldsymbol{E}}_2\tilde{\boldsymbol{E}}_2^{\mathrm{T}}P_{\tilde{A}}\tilde{\boldsymbol{E}}_2\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\boldsymbol{a}_2 + p^{-2}\boldsymbol{a}_1^{\mathrm{T}}\tilde{\boldsymbol{E}}_2\tilde{\boldsymbol{E}}_2^{\mathrm{T}}P_{\tilde{A}}^{\perp}\tilde{\boldsymbol{E}}_2\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\boldsymbol{a}_2\end{aligned}$$

Since $\boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{a}_i \leq c$ and $\|\boldsymbol{G}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{G}\|_2 \leq c$ for some constant $c > 0$,

$$\|\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\boldsymbol{a}_i\|_2 \sim \|MN_{n\times 1}\left(0, I_n, \boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{a}_i\right)\|_2 = O_{\mathrm{p}}\left(n^{1/2}\right)$$

for $i = 1, 2$ and $\|\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\boldsymbol{G}\|_2 \sim \|MN_{n\times 2}\left(0, I_n, \boldsymbol{G}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{G}\right)\|_2 = O_{\mathrm{p}}\left(n^{1/2}\right)$. Then by Cauchy-Schwartz,

$$p^{-2}\boldsymbol{a}_1^{\mathrm{T}}\tilde{\boldsymbol{E}}_2\tilde{\boldsymbol{E}}_2^{\mathrm{T}}P_{\tilde{A}}\tilde{\boldsymbol{E}}_2\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\boldsymbol{a}_2 = p^{-2}\underbrace{\boldsymbol{a}_1^{\mathrm{T}}\tilde{\boldsymbol{E}}_2}_{1\times n}\underbrace{\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\boldsymbol{G}}_{n\times 2}\boldsymbol{G}^{\mathrm{T}}\tilde{\boldsymbol{E}}_2\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\boldsymbol{a}_2 = O_{\mathrm{p}}\left\{\left(np^{-1}\right)^2\right\}$$

By Craig's Theorem, $\tilde{\boldsymbol{E}}^{\mathrm{T}}\boldsymbol{a}_i$ and $\tilde{\boldsymbol{E}}^{\mathrm{T}}\boldsymbol{Q}$ are independent, since $\boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{Q} = 0$. We then have

$$p^{-2}\boldsymbol{a}_1^{\mathrm{T}}\tilde{\boldsymbol{E}}_2\tilde{\boldsymbol{E}}_2^{\mathrm{T}}P_{\tilde{A}}^{\perp}\tilde{\boldsymbol{E}}_2\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\boldsymbol{a}_2 = p^{-2}\boldsymbol{a}_1^{\mathrm{T}}\tilde{\boldsymbol{E}}_2\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{Q}^{\mathrm{T}}\tilde{\boldsymbol{E}}_2\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\boldsymbol{a}_2$$

Let $\boldsymbol{B} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{Q}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{\Sigma}^{1/2}$ and let $\boldsymbol{H}\boldsymbol{\Delta}\boldsymbol{H}^{\mathrm{T}}$ be its singular value decomposition. Note that $\max_{r\in[p]}\left[\boldsymbol{\Delta}\right]_{rr} \leq c$ for some constant $c > 0$ since $\boldsymbol{Q}\boldsymbol{Q}^{\mathrm{T}}$ is just a projection matrix. Therefore, $\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{Q}^{\mathrm{T}}\tilde{\boldsymbol{E}}_2 \sim \boldsymbol{J}^{\mathrm{T}}\boldsymbol{J}$, where $\boldsymbol{J} \sim MN_{p\times n}\left(0, \boldsymbol{\Delta}, I_n\right)$ and is independent of $p^{-1/2}\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\boldsymbol{a}_i = \tilde{\boldsymbol{a}}_i \in \mathbb{R}^{n\times 1}$ where $\|\tilde{\boldsymbol{a}}_i\|_2 = O_{\mathrm{p}}\left(n^{1/2}p^{-1/2}\right)$. Define $\delta = p^{-1}\operatorname{tr}\left(\boldsymbol{\Delta}\right) = \rho + O\left(p^{-1}\right)$, $\gamma = p^{-1}\operatorname{tr}\left(\boldsymbol{\Delta}^2\right)$ and $\boldsymbol{b}_i = \|\tilde{\boldsymbol{a}}_i\|_2^{-1}\tilde{\boldsymbol{a}}_i$. Then

$$
\begin{aligned}
p^{-2}\boldsymbol{a}_1^{\mathrm{T}}\tilde{\boldsymbol{E}}_2\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{Q}^{\mathrm{T}}\tilde{\boldsymbol{E}}_2\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\boldsymbol{a}_2 &\sim \|\tilde{\boldsymbol{a}}_1\|_2\|\tilde{\boldsymbol{a}}_2\|_2\boldsymbol{b}_1^{\mathrm{T}}p^{-1}\boldsymbol{J}^{\mathrm{T}}\boldsymbol{J}\boldsymbol{b}_2\\
&= \|\tilde{\boldsymbol{a}}_1\|_2\|\tilde{\boldsymbol{a}}_2\|_2\boldsymbol{b}_1^{\mathrm{T}}\begin{pmatrix} p^{-1}\boldsymbol{J}_1^{\mathrm{T}}\boldsymbol{J}_1 & \cdots & p^{-1}\boldsymbol{J}_1^{\mathrm{T}}\boldsymbol{J}_n\\ \vdots & \ddots & \vdots\\ p^{-1}\boldsymbol{J}_1^{\mathrm{T}}\boldsymbol{J}_n & \cdots & p^{-1}\boldsymbol{J}_n^{\mathrm{T}}\boldsymbol{J}_n\end{pmatrix}\boldsymbol{b}_2
\end{aligned}
$$

$$
\boldsymbol{b}_1^{\mathrm{T}}\begin{pmatrix} p^{-1}\boldsymbol{J}_1^{\mathrm{T}}\boldsymbol{J}_1 & \cdots & p^{-1}\boldsymbol{J}_1^{\mathrm{T}}\boldsymbol{J}_n\\ \vdots & \ddots & \vdots\\ p^{-1}\boldsymbol{J}_1^{\mathrm{T}}\boldsymbol{J}_n & \cdots & p^{-1}\boldsymbol{J}_n^{\mathrm{T}}\boldsymbol{J}_n\end{pmatrix}\boldsymbol{b}_2 = \sum_{i=1}^{n}\boldsymbol{b}_1[i]\boldsymbol{b}_2[i]p^{-1}\boldsymbol{J}_i^{\mathrm{T}}\boldsymbol{J}_i + \sum_{i\neq q}\boldsymbol{b}_1[i]\boldsymbol{b}_2[q]p^{-1}\boldsymbol{J}_i^{\mathrm{T}}\boldsymbol{J}_q
$$

$$
\begin{aligned}
\sum_{i=1}^{n}\boldsymbol{b}_1[i]\boldsymbol{b}_2[i]p^{-1}\boldsymbol{J}_i^{\mathrm{T}}\boldsymbol{J}_i \underbrace{=}_{X_i=\frac{1}{p}\boldsymbol{J}_i^{\mathrm{T}}\boldsymbol{J}_i-\delta} &\delta\boldsymbol{b}_1^{\mathrm{T}}\boldsymbol{b}_2 + \underbrace{\sum_{i=1}^{n}\boldsymbol{b}_1[i]\boldsymbol{b}_2[i]X_i}_{=X}\\
\mathbb{V}(X) = \sum_{i=1}^{n}\boldsymbol{b}_1[i]^2\boldsymbol{b}_2[i]^2\,\mathbb{V}\left(X_i\right) &= 2\gamma p^{-1}\sum_{i=1}^{n}\boldsymbol{b}_1[i]^2\boldsymbol{b}_2[i]^2 \leq 2\gamma p^{-1}
\end{aligned}
$$

$$
\Rightarrow \sum_{i=1}^{n}\boldsymbol{b}_1[i]\boldsymbol{b}_2[i]p^{-1}\boldsymbol{J}_i^{\mathrm{T}}\boldsymbol{J}_i = \delta\boldsymbol{b}_1^{\mathrm{T}}\boldsymbol{b}_2 + O_{\mathrm{p}}\left(p^{-1/2}\right) = \rho\boldsymbol{b}_1^{\mathrm{T}}\boldsymbol{b}_2 + O_{\mathrm{p}}\left(p^{-1/2}\right). \tag{2.47}
$$

Note that $\mathbb{E}\left(\sum_{i\neq q}\boldsymbol{b}_1[i]\boldsymbol{b}_2[q]p^{-1}\boldsymbol{J}_i^{\mathrm{T}}\boldsymbol{J}_q\right) = 0$, meaning

$$
\mathbb{V}\left(\sum_{i\neq q}\boldsymbol{b}_1[i]\boldsymbol{b}_2[q]p^{-1}\boldsymbol{J}_i^{\mathrm{T}}\boldsymbol{J}_q\right) = \mathbb{E}\left\{\left(\sum_{i\neq q}\boldsymbol{b}_1[i]\boldsymbol{b}_2[q]p^{-1}\boldsymbol{J}_i^{\mathrm{T}}\boldsymbol{J}_q\right)^2\right\}.
$$

Therefore,

$$\mathbb{V}\left(\sum_{i\neq q}\boldsymbol{b}_1[i]\boldsymbol{b}_2[q]p^{-1}\boldsymbol{J}_i^{\mathrm{T}}\boldsymbol{J}_q\right)=p^{-2}\sum_{i\neq q}\sum_{r\neq s}\boldsymbol{b}_1[i]\boldsymbol{b}_2[q]\boldsymbol{b}_1[r]\boldsymbol{b}_2[s]\mathbb{E}\left\{\left(\boldsymbol{J}_i^{\mathrm{T}}\boldsymbol{J}_q\right)\left(\boldsymbol{J}_r^{\mathrm{T}}\boldsymbol{J}_s\right)\right\}.$$

We then need to go through various scenarios to evaluate the above expression.

(1) $i\neq r,s$ and $q\neq r,s$. Then,

$$\mathbb{E}\left\{\left(\boldsymbol{J}_i^{\mathrm{T}}\boldsymbol{J}_q\right)\left(\boldsymbol{J}_r^{\mathrm{T}}\boldsymbol{J}_s\right)\right\}=0$$

(2) $i=r$.

(a) $q\neq s$

$$\begin{aligned}\mathbb{E}\left\{\left(\boldsymbol{J}_i^{\mathrm{T}}\boldsymbol{J}_q\right)\left(\boldsymbol{J}_i^{\mathrm{T}}\boldsymbol{J}_s\right)\right\}&=\mathbb{E}\left\{\boldsymbol{J}_q^{\mathrm{T}}\mathbb{E}\left(\boldsymbol{J}_i\boldsymbol{J}_i^{\mathrm{T}}\mid\boldsymbol{J}_q,\boldsymbol{J}_s\right)\boldsymbol{J}_s\right\}=\mathbb{E}\left(\boldsymbol{J}_q^{\mathrm{T}}\boldsymbol{\Delta}\boldsymbol{J}_s\right)\\&=\mathrm{tr}\left\{\boldsymbol{\Delta}\mathbb{E}\left(\boldsymbol{J}_s\boldsymbol{J}_q^{\mathrm{T}}\right)\right\}=0\end{aligned}$$

(b) $q=s$

$$\mathbb{E}\left\{\left(\boldsymbol{J}_i^{\mathrm{T}}\boldsymbol{J}_q\right)\left(\boldsymbol{J}_i^{\mathrm{T}}\boldsymbol{J}_q\right)\right\}=\mathbb{E}\left\{\boldsymbol{J}_q^{\mathrm{T}}\mathbb{E}\left(\boldsymbol{J}_i\boldsymbol{J}_i^{\mathrm{T}}\mid\boldsymbol{J}_q\right)\boldsymbol{J}_q\right\}=\mathbb{E}\left(\boldsymbol{J}_q^{\mathrm{T}}\boldsymbol{\Delta}\boldsymbol{J}_q\right)=\mathrm{tr}\left(\boldsymbol{\Delta}^2\right)=p\gamma$$

(3) $i=s$

(a) $q\neq r$

$$\mathbb{E}\left\{\left(\boldsymbol{J}_i^{\mathrm{T}}\boldsymbol{J}_q\right)\left(\boldsymbol{J}_r^{\mathrm{T}}\boldsymbol{J}_i\right)\right\}=\mathbb{E}\left\{\boldsymbol{J}_q^{\mathrm{T}}\mathbb{E}\left(\boldsymbol{J}_i\boldsymbol{J}_i^{\mathrm{T}}\mid\boldsymbol{J}_q,\boldsymbol{J}_r\right)\boldsymbol{J}_r\right\}=\mathbb{E}\left(\boldsymbol{J}_q^{\mathrm{T}}\boldsymbol{\Delta}\boldsymbol{J}_r\right)=0$$

(b) $q=r$

$$\mathbb{E}\left\{\left(\boldsymbol{J}_i^{\mathrm{T}}\boldsymbol{J}_q\right)\left(\boldsymbol{J}_q^{\mathrm{T}}\boldsymbol{J}_i\right)\right\}=p\gamma$$

(4) $q = s$, $i \neq r$ (we already have the case $q = s, i = r$ above).

$$\mathbb{E}\left\{\left(\boldsymbol{J}_i^{\mathrm{T}}\boldsymbol{J}_q\right)\left(\boldsymbol{J}_r^{\mathrm{T}}\boldsymbol{J}_q\right)\right\} = 0$$

(5) $q = r$, $i \neq s$ (we already have the case $q = r, i = s$ above).

$$\mathbb{E}\left\{\left(\boldsymbol{J}_i^{\mathrm{T}}\boldsymbol{J}_q\right)\left(\boldsymbol{J}_q^{\mathrm{T}}\boldsymbol{J}_s\right)\right\} = 0$$

Therefore,

$$\begin{aligned} p^{-2}\sum_{i\neq q}\sum_{r\neq s}\boldsymbol{b}_1[i]\boldsymbol{b}_2[q]\boldsymbol{b}_1[r]\boldsymbol{b}_2[s]\mathbb{E}\left\{\left(\boldsymbol{J}_i^{\mathrm{T}}\boldsymbol{J}_q\right)\left(\boldsymbol{J}_r^{\mathrm{T}}\boldsymbol{J}_s\right)\right\} =&\gamma p^{-1}\sum_{i\neq q}\boldsymbol{b}_1[i]^2\boldsymbol{b}_2[q]^2 \\ &+ \gamma p^{-1}\sum_{i\neq q}\boldsymbol{b}_1[i]\boldsymbol{b}_2[i]\boldsymbol{b}_1[q]\boldsymbol{b}_2[q] \end{aligned}$$

$$\sum_{i\neq q}\boldsymbol{b}_1[i]^2\boldsymbol{b}_2[q]^2 \leq \sum_{i=1}^{n}\boldsymbol{b}_1[i]^2\sum_{q=1}^{n}\boldsymbol{b}_2[q]^2 = 1$$

$$\begin{aligned} \sum_{i\neq q}\boldsymbol{b}_1[i]\boldsymbol{b}_2[i]\boldsymbol{b}_1[q]\boldsymbol{b}_2[q] =&\sum_{i=1}^{n}\boldsymbol{b}_1[i]\boldsymbol{b}_2[i]\sum_{q\neq i}^{n}\boldsymbol{b}_1[q]\boldsymbol{b}_2[q], \quad |\sum_{q\neq i}^{n}\boldsymbol{b}_1[q]\boldsymbol{b}_2[q]| \leq \|\boldsymbol{b}_{1,-i}\|_2\|\boldsymbol{b}_{2,-i}\|_2 \\ \leq& 1 \end{aligned}$$

$$\begin{aligned} \Rightarrow |\sum_{i=1}^{n}\boldsymbol{b}_1[i]\boldsymbol{b}_2[i]\sum_{q\neq i}^{n}\boldsymbol{b}_1[q]\boldsymbol{b}_2[q]| \leq& \left\{\sum_{i=1}^{n}\left(\sum_{q\neq i}^{n}\boldsymbol{b}_1[q]\boldsymbol{b}_2[q]\right)^2\boldsymbol{b}_1[i]^2\right\}^{1/2}\|\boldsymbol{b}_2\|_2 \\ \leq& \|\boldsymbol{b}_1\|_2\|\boldsymbol{b}_2\|_2 = 1 \end{aligned}$$

Therefore $\mathbb{V}\left(\sum_{i\neq q}\boldsymbol{b}_1[i]\boldsymbol{b}_2[q]p^{-1}\boldsymbol{J}_i^{\mathrm{T}}\boldsymbol{J}_q\right) \leq \gamma p^{-1}$, meaning

$$\begin{aligned} p^{-2}\boldsymbol{a}_1^{\mathrm{T}}\tilde{\boldsymbol{E}}_2\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\tilde{\boldsymbol{E}}_2\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\boldsymbol{a}_2 &= \|\tilde{\boldsymbol{a}}_1\|_2\|\tilde{\boldsymbol{a}}_2\|_2\rho\boldsymbol{b}_1^{\mathrm{T}}\boldsymbol{b}_2 + \|\tilde{\boldsymbol{a}}_1\|_2\|\tilde{\boldsymbol{a}}_2\|_2 O_{\mathrm{p}}\left(p^{-1/2}\right) + O_{\mathrm{p}}\left\{\left(np^{-1}\right)^2\right\} \\ &= \rho p^{-1}\boldsymbol{a}_1^{\mathrm{T}}\tilde{\boldsymbol{E}}_2\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\boldsymbol{a}_2 + O_{\mathrm{p}}\left(np^{-3/2}\right) + O_{\mathrm{p}}\left\{\left(np^{-1}\right)^2\right\}. \end{aligned}$$

Therefore,

$$p^{-1}\boldsymbol{a}_1^{\mathrm{T}}\tilde{\boldsymbol{E}}_2\boldsymbol{R}\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\boldsymbol{a}_2 = O_{\mathrm{p}}\left(np^{-3/2}\right) + O_{\mathrm{p}}\left\{\left(np^{-1}\right)^2\right\}.$$

□

CHAPTER 3

ESTIMATING AND ACCOUNTING FOR UNOBSERVED COVARIATES IN HIGH DIMENSIONAL CORRELATED DATA

3.1 Introduction

The development of high-throughput technologies has provided biologists with a cornucopia of genetic, proteomic and metabolomic data that can help elucidate the genetic components of phenotypes and the mediation of environmental exposures. Many of these 'omic' data have complex sample-correlation structure, which includes longitudinal data (Baumgart et al. 2016, McKennan et al. 2018), multi-tissue data (GTEx Consortium 2017), data with multiple treatment conditions (Knowles et al. 2018) and data with related individuals (Martino et al. 2013, Tung et al. 2015). Longitudinal studies have even been cited as a critical area of future DNA methylation research to assess the stability of methylation marks over time (Breton et al. 2017, Martin & Fry 2018), so it is crucial that suitable methods exist to analyze these data. An important feature of these high-throughput data is the presence of unmeasured factors that influence the measured response, which include technical factors like batch variables and biological factors like cell composition (Leek et al. 2010, Jaffe & Irizarry 2014). When unaccounted for, these factors can bias test statistics, reduce power and lead to irreproducible results (Yao et al. 2012, Peixoto et al. 2015).

There have been a number of methods developed by the statistical community to estimate and correct for latent factors in high throughput biological data (Buja & Eyuboglu 1992, Leek & Storey 2008, Gagnon-Bartsch & Speed 2012, Sun et al. 2012, Gagnon-Bartsch et al. 2013, Houseman et al. 2014, Owen & Wang 2016, Fan & Han 2017, Lee et al. 2017, Wang et al. 2017, McKennan & Nicolae 2018*a*). However, these methods make the critical assumption that conditional on both the observed and unobserved covariates, samples are independent with homogeneous residual variances, and tend to perform poorly when these assumptions are violated. The goal of this chapter is to provide a provably accurate method to both choose the number of and estimate latent factors from the measured correlated data so that downstream inference on the effects of interest is as accurate as when all latent factors are observed. To the best of our knowledge, this is the first method to estimate and correct for latent covariates in high throughput biological data with correlated samples.

We use DNA methylation quantified in $p \approx 8 \times 10^5$ methylation sites (CpGs) in $n/2 = 183$ unrelated individuals at birth and age 7 from McKennan et al. (2018) as a motivating data example, although we analyze other methylation data with a more complex covariance structure in Section 3.6. The aim of that study was to jointly model methylation at birth and age 7 to determine if the effects due to ancestry on methylation levels changed or remained constant over time. If we ignore observed nuisance variables like the intercept, a reasonable model for the methylation at birth and age 7 at CpG g in the presence of unobserved covariates $\boldsymbol{C} \in \mathbb{R}^{n \times K}$ is

$$\begin{aligned}
&\boldsymbol{y}_g = \left(\boldsymbol{y}_{g,0}^{\mathrm{T}}\, \boldsymbol{y}_{g,7}^{\mathrm{T}}\right)^{\mathrm{T}} = (\boldsymbol{A} \oplus \boldsymbol{A}) \left(\beta_{g,0}\, \beta_{g,7}\right)^{\mathrm{T}} + \boldsymbol{C}\boldsymbol{\ell}_g + \boldsymbol{e}_g, \quad \boldsymbol{e}_g \sim N_n\left(0, \boldsymbol{V}_g\right) \quad (g = 1, \ldots, p) \\
&\boldsymbol{V}_g = \phi_g^2 \boldsymbol{B}_1 + \sigma_{g,0}^2 \boldsymbol{B}_2 + \sigma_{g,7}^2 \boldsymbol{B}_3 \quad (g = 1, \ldots, p),
\end{aligned} \tag{3.1}$$

where $\boldsymbol{y}_{g,0} \in \mathbb{R}^{n/2}$ and $\boldsymbol{y}_{g,7} \in \mathbb{R}^{n/2}$ are the methylation at birth and age 7, $\boldsymbol{A} \in \mathbb{R}^{n/2}$ gives each individual's ancestry and $\beta_{g,0}, \beta_{g,7}$ are the effects of interest. $\boldsymbol{B}_1 \in \mathbb{R}^{n \times n}$ is a partition matrix that groups samples by individuals and captures the within-individual variability. $\boldsymbol{B}_2 \in \mathbb{R}^{n \times n}$ and $\boldsymbol{B}_3 \in \mathbb{R}^{n \times n}$ capture the potential difference in residual variance at birth and age 7 and are diagonal matrices with ones in the first $n/2$ and second $n/2$ diagonal entries, respectively. In order to avoid overestimating the residual variance and biasing our estimates for $\beta_{g,0}$ and $\beta_{g,7}$, we must first estimate $\boldsymbol{C}$ and its latent dimension K. One approach would be to use methods from the economics literature that allow for dependent residuals in their theoretical arguments (Bai 2009, Li et al. 2016, Lu & Su 2016, Su & Ju 2018). However, their assumptions on the homogeneity of the effects $\left(\beta_{1,0}, \beta_{1,7}\right), \ldots, \left(\beta_{p,0}, \beta_{p,7}\right)$ are too restrictive to have any application in genetic data. Another strategy would be to use existing methods designed for genetic data to estimate the factors at each age separately. However, this reduces the sample size by 50% and risks underestimating K, and subsequently biasing estimates for $\beta_{g,0}$ and $\beta_{g,7}$. To avoid these biases, a second option would be to estimate K and $\boldsymbol{C}$ using all n samples. However, naively estimating K with methods commonly applied to genetic data with independent and identically distributed residuals, like parallel analysis (Buja & Eyuboglu 1992) and bi-cross validation (Owen & Wang 2016), will typically overestimate K to be on the order of the sample size, as they will be unable to distinguish the low dimensional factors $\boldsymbol{C}$ from the high dimensional random effect. Even if K were known, the correlation among the residuals obfuscates existing estimates for $\boldsymbol{C}$.

Our proposed method uses all of the available data to estimate K and $\boldsymbol{C}$ in data whose genomic unit-specific covariance can be written as a linear combination of known matrices. This covariance structure is quite general and includes longitudinal, multi-tissue, multi-treatment and twin studies, as well as studies with individuals related through a kinship matrix. We discuss these data types in more detail in Section 3.4. While our ultimate goal is to be able to do inference on the effects of interest that is as accurate as when $\boldsymbol{C}$ is observed, our method also provides a way to perform factor analysis in data with correlated or nonexchangeable residuals, which has application in quantitative train loci studies (Knowles et al. 2018, McKennan et al. 2018).

The chapter is organized as followed. We describe the data and set up the problem in Section 3.2 and present a detailed description of our method in Section 3.3. We prove its efficacy in Section 3.4 by proving its estimate for K is consistent, its estimate for the column space of $\boldsymbol{C}$ is nearly as accurate as when samples are independent (i.e. $\boldsymbol{V}_g$ in (3.1) is a multiple of the identity) and that inference on the effects of interest is asymptotically equivalent to that when $\boldsymbol{C}$ is known. In Section 3.5 we analyze simulated multi-tissue gene expression data with a complex, gene-dependent correlation structure that demonstrates our method's superior performance in both choosing K and estimating $\boldsymbol{C}$. We lastly apply our method to a longitudinal DNA methylation dataset with measurements made on pairs of twins to identify CpGs with sex-dependent methylation levels in Section 3.6. An R package called CorrConf that implements our method to estimate K and $\boldsymbol{C}$ is freely available on GitHub. The proofs of all theorems are in Sections 3.10 – 3.14.

3.2 Problem set-up

Let $\boldsymbol{y}_g \in \mathbb{R}^n$ be the measured expression or methylation of genomic unit $g \in [p]$ in n potentially correlated samples, and let $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ be the covariates of interest. When latent factors $\boldsymbol{C} \in \mathbb{R}^{n \times K}$ influence the data $\boldsymbol{y}_g$, we assume $\boldsymbol{y}_g$ is generated as

$$\boldsymbol{y}_g = \boldsymbol{X}\boldsymbol{\beta}_g + \boldsymbol{C}\boldsymbol{\ell}_g + \boldsymbol{e}_g, \quad \boldsymbol{e}_g \sim N_n\left(\boldsymbol{0}, \boldsymbol{V}_g\right) \quad (g = 1, \dots, p) \tag{3.2a}$$

$$\boldsymbol{V}_g = v_{g,1}\boldsymbol{B}_1 + \cdots + v_{g,b}\boldsymbol{B}_b \quad (g = 1, \dots, p) \tag{3.2b}$$

$$\boldsymbol{Y}_{p \times n} = \left[\boldsymbol{y}_1 \cdots \boldsymbol{y}_p\right]^{\mathrm{T}} = \boldsymbol{\beta}_{p \times d}\boldsymbol{X}_{n \times d}^{\mathrm{T}} + \boldsymbol{L}_{p \times K}\boldsymbol{C}_{n \times K}^{\mathrm{T}} + \boldsymbol{E}_{p \times n}. \tag{3.2c}$$

The gth rows of $\boldsymbol{\beta}$, $\boldsymbol{L}$ and $\boldsymbol{E}$ are $\boldsymbol{\beta}_g$, $\boldsymbol{\ell}_g$ and $\boldsymbol{e}_g$, respectively, $\boldsymbol{e}_1, \dots, \boldsymbol{e}_p$ are independent, and our goal is to estimate and perform inference on $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_p$. We assume $\boldsymbol{C}$ is a random matrix that is independent of $\boldsymbol{E}$, and whose mean may be a function of $\boldsymbol{X}$ in our theoretical results in Section 3.4. Therefore, accounting for $\boldsymbol{C}$ can be interpreted as a way to alleviate both the dependence across genomic units and potential biases when estimating and performing inference on $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_p$. We postpone discussing the distributional assumptions on $\boldsymbol{C}$ until Section 3.4 because, other than $\boldsymbol{C}$ being independent of $\boldsymbol{E}$, they play no part in the methodology presented in Section 3.3.

The observed matrices $\boldsymbol{B}_1, \dots, \boldsymbol{B}_b \in \mathbb{R}^{n \times n}$ describe how the n samples are related, and Model (3.2b) is a typical model for the variance in mixed effect models in high throughput biological data (Martino et al. 2013, Tung et al. 2015, Chen et al. 2017, Knowles et al. 2018, McKennan et al. 2018). When $b = 1$ and $\boldsymbol{B}_1 = I_n$, Model (3.2) reduces to the model typically used to estimate $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_p$ in high throughput data with independent samples

(Leek & Storey 2008, Sun et al. 2012, Lee et al. 2017, Wang et al. 2017, Gerard & Stephens 2018, McKennan & Nicolae 2018*a*). We assume the unknown variance multipliers $\boldsymbol{v}_g = \left(v_{g,1} \cdots v_{g,b}\right)^{\mathrm{T}}$ lie in the convex polytope

$$\Theta = \left\{\boldsymbol{x} \in \mathbb{R}^b : \boldsymbol{A}_{\mathcal{E}}\boldsymbol{x} = \boldsymbol{0}, \boldsymbol{A}_{\mathcal{I}}\boldsymbol{x} \geq \boldsymbol{0}\right\}, \tag{3.3}$$

where the matrices $\boldsymbol{A}_{\mathcal{E}} \in \mathbb{R}^{N_{\mathcal{E}} \times b}$ and $\boldsymbol{A}_{\mathcal{I}} \in \mathbb{R}^{N_{\mathcal{I}} \times b}$ are the equality and inequality constraints on $\boldsymbol{v}_g$. For example, we may know that certain multipliers must be larger than others, or that the sum of two sets of multipliers must be equal to ensure the marginal variances are the same.

In many applications, there are other observed covariates $\boldsymbol{Z} \in \mathbb{R}^{n \times r}$ that may influence the response $\boldsymbol{y}_g$ but whose effects are not of interest. In that case, the model for $\boldsymbol{y}_g$ would be

$$\boldsymbol{y}_g = \boldsymbol{X}\boldsymbol{\beta}_g + \boldsymbol{Z}\boldsymbol{\omega}_g + \boldsymbol{C}\boldsymbol{\ell}_g + \boldsymbol{e}_g, \tag{3.4}$$

and we can get back to Model (3.2) by multiplying $\boldsymbol{y}_g$ by $\boldsymbol{Q}_Z^{\mathrm{T}}$, provided r does not grow with n:

$$\begin{aligned}
\boldsymbol{Q}_Z^{\mathrm{T}}\boldsymbol{y}_g &= \left(\boldsymbol{Q}_Z^{\mathrm{T}}\boldsymbol{X}\right)\boldsymbol{\beta}_g + \left(\boldsymbol{Q}_Z^{\mathrm{T}}\boldsymbol{C}\right)\boldsymbol{\ell}_g + \bar{\boldsymbol{e}}_g, \quad \bar{\boldsymbol{e}}_g \sim N_{n-r}\left(\boldsymbol{0}, \bar{\boldsymbol{V}}_g\right)\\
\bar{\boldsymbol{V}}_g &= v_{g,1}\left(\boldsymbol{Q}_Z^{\mathrm{T}}\boldsymbol{B}_1\boldsymbol{Q}_Z\right) + \cdots + v_{g,b}\left(\boldsymbol{Q}_Z^{\mathrm{T}}\boldsymbol{B}_b\boldsymbol{Q}_Z\right).
\end{aligned}$$

Therefore, we assume that the only observed covariates are contained in $\boldsymbol{X}$.

Evidently, $\boldsymbol{C}$ in Model (3.2) is not identifiable. However, we show in Section 3.4 that conditional on $\boldsymbol{C}$, $\operatorname{Im}(\boldsymbol{C})$ is identifiable when $\boldsymbol{\beta}$ satisfies a modest sparsity assumption. Therefore, our primary goal is to develop a procedure to estimate and perform inference on $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$ that only relies on an accurate estimate for $\operatorname{Im}(\boldsymbol{C})$. Our secondary goal is to simply estimate $\operatorname{Im}(\boldsymbol{C})$ when there are no covariates of interest, i.e. $d = 0$. Since accomplishing the former implies we can achieve the latter, we only consider the case when $d \geq 1$.

We estimate $\operatorname{Im}(\boldsymbol{C})$ by separately computing, and then combining, estimates for the parts of $\operatorname{Im}(\boldsymbol{C})$ in $\operatorname{Im}(\boldsymbol{X})$ and $\boldsymbol{X}^{\perp}$. Let $\boldsymbol{G} \in \mathbb{R}^{n \times n}$ be a positive definite matrix. We estimate the part of $\operatorname{Im}(\boldsymbol{C})$ in $\operatorname{Im}(\boldsymbol{X})$ using the variation in $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_p$ explainable by $\boldsymbol{X}$:

$$\boldsymbol{y}_{g_1} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{G}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{G}^{-1}\boldsymbol{y}_g = \boldsymbol{\beta}_g + \boldsymbol{\Omega}\boldsymbol{\ell}_g + \boldsymbol{e}_{g_1} \in \mathbb{R}^d \ (g = 1, \ldots, p) \tag{3.5a}$$

$$\boldsymbol{\Omega} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{G}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{G}^{-1}\boldsymbol{C} \in \mathbb{R}^{d \times K} \tag{3.5b}$$

$$\boldsymbol{Y}_1 = \left[\boldsymbol{y}_{1_1} \ \cdots \ \boldsymbol{y}_{p_1}\right]^{\mathrm{T}} = \boldsymbol{\beta} + \boldsymbol{L}\boldsymbol{\Omega}^{\mathrm{T}} + \boldsymbol{E}_1 \in \mathbb{R}^{p \times d}. \tag{3.5c}$$

Here $\boldsymbol{y}_{g_1}$ is the naive weighted least squares estimator for $\boldsymbol{\beta}_g$ that ignores $\boldsymbol{C}$ and $\boldsymbol{\Omega}$ is the weighted least squares regression coefficient for the regression of $\boldsymbol{C}$ onto $\boldsymbol{X}$. We show how to choose the appropriate $\boldsymbol{G}$ in Section 3.3.4. Next, we use the variation in $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_p$ explainable by $\boldsymbol{X}^{\perp}$ to estimate the part of $\operatorname{Im}(\boldsymbol{C})$ in $\boldsymbol{X}^{\perp}$:

$$\boldsymbol{y}_{g_2} = \boldsymbol{Q}_X^{\mathrm{T}}\boldsymbol{y}_g = \boldsymbol{C}_{\perp}\boldsymbol{\ell}_g + \boldsymbol{e}_{g_2} \in \mathbb{R}^{n-d}, \quad \boldsymbol{e}_{g_2} \sim N_{n-d}\left(\boldsymbol{0}, \boldsymbol{Q}_X^{\mathrm{T}}\boldsymbol{V}_g\boldsymbol{Q}_X\right) \ (g = 1, \ldots, p) \tag{3.6a}$$

$$\boldsymbol{C}_{\perp} = \boldsymbol{Q}_X^{\mathrm{T}}\boldsymbol{C} \in \mathbb{R}^{(n-d) \times K} \tag{3.6b}$$

$$\boldsymbol{Y}_2 = \left[\boldsymbol{y}_{1_2} \ \cdots \ \boldsymbol{y}_{p_2}\right]^{\mathrm{T}} = \boldsymbol{L}\boldsymbol{C}_{\perp}^{\mathrm{T}} + \boldsymbol{E}_2 \in \mathbb{R}^{p \times (n-d)}. \tag{3.6c}$$

Here $\boldsymbol{C}_{\perp}$ and $\boldsymbol{y}_{g_2}$ lie in the space orthogonal to $\boldsymbol{X}$, where the latter no longer depends on $\boldsymbol{\beta}_g$. Algorithm 3.1 below, which we call CBCV-CorrConf, provides a cursory description of how we use the objects defined in (3.5) and (3.6) to estimate $\operatorname{Im}(\boldsymbol{C})$ and subsequently $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$.

***Algorithm* 3.1** (CBCV-CorrConf)**.** *(a) For $k \in \{0, 1, \ldots, K_{\max}\}$, use $\boldsymbol{Y}_2$ to obtain estimates $\hat{\boldsymbol{L}} \in \mathbb{R}^{p \times k}$ and $\hat{\boldsymbol{C}}_{\perp} \in \mathbb{R}^{(n-d) \times k}$ using a factor analysis procedure we call ICaSE (**I**terative **C**orrelation **a**nd **S**ubspace **E**stimation) (Section 3.3.2).*

*(b) Use $\boldsymbol{Y}_2$ and a novel cross-validation procedure we call CBCV (**C**orrelated **B**i-**C**ross*

$\boldsymbol{V}$alidation) to define $\hat{K}$, the estimate for $\dim\{\operatorname{Im}(\boldsymbol{C})\}$ *(Section 3.3.3).*

(c) Define $\boldsymbol{G}$ and estimate $\boldsymbol{\Omega}$ using $\boldsymbol{Y}_1$ and the estimators $\hat{\boldsymbol{L}} \in \mathbb{R}^{p\times\hat{K}}$, $\hat{\boldsymbol{C}}_\perp \in \mathbb{R}^{(n-d)\times\hat{K}}$ from steps (a) and (b). Let the estimate for $\boldsymbol{C}$ be

$$\hat{\boldsymbol{C}} = \boldsymbol{X}\hat{\boldsymbol{\Omega}} + \boldsymbol{G}\boldsymbol{Q}_X\left(\boldsymbol{Q}_X^{\mathrm{T}}\boldsymbol{G}\boldsymbol{Q}_X\right)^{-1}\hat{\boldsymbol{C}}_\perp \in \mathbb{R}^{n\times\hat{K}}.$$

For $\hat{K}$ given, we call this method of estimating $\boldsymbol{C}$ CorrConf (Section 3.3.4).

(d) Use the design matrix $\left[\boldsymbol{X}\,\hat{\boldsymbol{C}}\right]$ to estimate $\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_p$ with generalized least sqaures.

The estimates for $\operatorname{Im}(\boldsymbol{C}_\perp)$ and $\operatorname{Im}(\boldsymbol{C})$ are then $\operatorname{Im}\left(\hat{\boldsymbol{C}}_\perp\right)$ and $\operatorname{Im}\left(\hat{\boldsymbol{C}}\right)$, where the estimate for $\boldsymbol{C}$ in step *(c)* follows from the fact that $\boldsymbol{C} = \boldsymbol{X}\boldsymbol{\Omega} + \boldsymbol{G}\boldsymbol{Q}_X\left(\boldsymbol{Q}_X^{\mathrm{T}}\boldsymbol{G}\boldsymbol{Q}_X\right)^{-1}\boldsymbol{C}_\perp$ for any positive definite $\boldsymbol{G}$. We show in Section 3.4 that our estimator in Step *(d)* only depends on $\operatorname{Im}\left(\hat{\boldsymbol{C}}\right)$. If $b=1$, $\boldsymbol{B}_1 = I_n$ and K were known, steps *(a)* and *(c)* with $\boldsymbol{G} = I_n$ are similar to methods used by Sun et al. (2012), Wang et al. (2017), Lee et al. (2017), McKennan & Nicolae (2018*a*).

3.3 Estimating the factor dimension K and $\operatorname{Im}(\boldsymbol{C})$

3.3.1 A computationally tractable model for the data

The generative model assumed in (3.2) is not conducive to estimating $\operatorname{Im}(\boldsymbol{C})$, since it requires jointly estimating $\operatorname{Im}(\boldsymbol{C})$ and all p covariance matrices $\boldsymbol{V}_1,\ldots,\boldsymbol{V}_p$. Instead, we use a simpler, but incorrect, model that assumes $\boldsymbol{V}_1 = \cdots = \boldsymbol{V}_p = \delta^2\boldsymbol{V}$ to estimate K, $\boldsymbol{L}$ and $\operatorname{Im}(\boldsymbol{C})$:

$$\boldsymbol{Y} = \boldsymbol{\beta}\boldsymbol{X}^{\mathrm{T}} + \boldsymbol{L}\boldsymbol{C}^{\mathrm{T}} + \boldsymbol{E}, \quad \boldsymbol{E} \sim MN_{p\times n}\left(\boldsymbol{0}, \delta^2 I_p, \boldsymbol{V}\right), \quad \det\left(\boldsymbol{Q}_X^{\mathrm{T}}\boldsymbol{V}\boldsymbol{Q}_X\right) = 1, \tag{3.7}$$

where we introduce δ^2 so that $\boldsymbol{V}$ is scale-invariant. We define δ^2 in terms of the determinant for reasons discussed in Section 3.3.3.

Let δ_*^2 and $\boldsymbol{V}_*$ be the values of δ^2 and $\boldsymbol{V}$ that, conditional on $\boldsymbol{C}$, minimize the KL-divergence between the data generating model in (3.2) and the incorrect model in (3.7). Then

$$\delta_*^2 \boldsymbol{V}_* = p^{-1}\left(\boldsymbol{V}_1 + \cdots + \boldsymbol{V}_p\right), \quad \det\left(\boldsymbol{Q}_X^{\mathrm{T}} \boldsymbol{V}_* \boldsymbol{Q}_X\right) = 1. \tag{3.8}$$

In Sections 3.3.2, 3.3.3 and 3.3.4 we show that, quite remarkably, we only need to estimate $\boldsymbol{V}_*$ and not $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_p$ to get accurate estimates for $\operatorname{Im}(\boldsymbol{C})$. Nevertheless, without any assumptions on $\boldsymbol{V}_*$, this is a challenging, if not an intractable, problem. However, since $\boldsymbol{V}_g = v_{g,1}\boldsymbol{B}_1 + \cdots + v_{g,b}\boldsymbol{B}_b$ for all $g = 1, \ldots, p$ and $\boldsymbol{B}_1, \ldots, \boldsymbol{B}_b$ are known, then $\boldsymbol{V}_* = \tau_{*_1}\boldsymbol{B}_1 + \cdots + \tau_{*_b}\boldsymbol{B}_b$ by (3.8), where

$$\begin{aligned} \boldsymbol{\tau}_* &= \left(\tau_{*_1} \cdots \tau_{*_b}\right)^{\mathrm{T}} = \left(\delta_*^2 p\right)^{-1}\left(\boldsymbol{v}_1 + \cdots + \boldsymbol{v}_p\right) \\ \delta_*^2 &= \det\left\{p^{-1}\boldsymbol{Q}_X^{\mathrm{T}}\left(\boldsymbol{V}_1 + \cdots + \boldsymbol{V}_p\right)\boldsymbol{Q}_X\right\}^{1/(n-d)}. \end{aligned}$$

Therefore, we need only estimate $\boldsymbol{\tau}_*$ to estimate $\boldsymbol{V}_*$, and subsequently $\operatorname{Im}(\boldsymbol{C})$.

3.3.2 ICaSE: an iterative factor analysis to estimate $\operatorname{Im}(\boldsymbol{C}_\perp)$

Here we present the algorithm ICaSE, a method to estimate $\operatorname{Im}(\boldsymbol{C}_\perp)$, δ_*^2 and $\boldsymbol{\tau}_*$ using $\boldsymbol{Y}_2$. Recall from (3.6) that the mean of $\boldsymbol{Y}_2$ is $\boldsymbol{L}\boldsymbol{C}_\perp^{\mathrm{T}}$ and is not dependent on $\boldsymbol{X}$. Let $\boldsymbol{W}_* = \boldsymbol{Q}_X^{\mathrm{T}}\boldsymbol{V}_*\boldsymbol{Q}_X$, which implies $\boldsymbol{W}_* = \tau_{*_1}\boldsymbol{Q}_X^{\mathrm{T}}\boldsymbol{B}_1\boldsymbol{Q}_X + \cdots + \tau_{*_b}\boldsymbol{Q}_X^{\mathrm{T}}\boldsymbol{B}_b\boldsymbol{Q}_X$ and $\delta_*^2\boldsymbol{W}_* = p^{-1}\boldsymbol{Q}_X^{\mathrm{T}}\left(\boldsymbol{V}_1 + \cdots + \boldsymbol{V}_p\right)\boldsymbol{Q}$. Therefore,

$$\boldsymbol{S}_2 = p^{-1}\boldsymbol{Y}_2^{\mathrm{T}}\boldsymbol{Y}_2 \approx \boldsymbol{C}_\perp\left(p^{-1}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\right)\boldsymbol{C}_\perp^{\mathrm{T}} + p^{-1}\boldsymbol{E}_2^{\mathrm{T}}\boldsymbol{E}_2 \approx \boldsymbol{C}_\perp\left(p^{-1}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\right)\boldsymbol{C}_\perp^{\mathrm{T}} + \delta_*^2\boldsymbol{W}_*.$$

Since $\boldsymbol{W}_*$ is not a multiple of the identity, the span of the first K eigenvectors of $\boldsymbol{S}_2$ will not be an accurate estimate for $\operatorname{Im}(\boldsymbol{C}_\perp)$, in general. If $\boldsymbol{W}_*$ were known, we could instead first estimate $\operatorname{Im}\left(\boldsymbol{W}_*^{-1/2}\boldsymbol{C}_\perp\right)$ as the span of the first K eigenvectors of $\boldsymbol{W}_*^{-1/2}\boldsymbol{S}_2\boldsymbol{W}_*^{-1/2}$, and then rescale the estimate by $\boldsymbol{W}_*^{1/2}$ to estimate $\mathrm{Im}\,(\boldsymbol{C}_\perp)$. On the other hand, if $\mathrm{Im}\,(\boldsymbol{C}_\perp)$ were known, one could easily estimate δ_*^2 and $\boldsymbol{\tau}_*$ using restricted maximum likelihood. ICaSE iterates between these two steps to estimate $\mathrm{Im}\,(\boldsymbol{C}_\perp)$, δ_*^2 and $\boldsymbol{\tau}_*$.

It remains to show how to find a starting point for δ_*^2 and $\boldsymbol{\tau}_*$ that avoids incorporating signal from the random effect into the estimate for $\mathrm{Im}\,(\boldsymbol{C}_\perp)$, as imprudent starting points often beget biased estimates for K and $\mathrm{Im}\,(\boldsymbol{C}_\perp)$. We separate the variation in $\boldsymbol{Y}_2$ due to $\boldsymbol{C}_\perp$ from the random effect by employing a "warm start" technique often used to solve penalized regression problems. First, we initialize our estimates for δ_*^2 and $\boldsymbol{\tau}_*$ assuming $K = 0$ (i.e. $\boldsymbol{L}\boldsymbol{C}_\perp^{\mathrm{T}} = \boldsymbol{0}$). We then use the estimate for $\boldsymbol{\tau}_*$ when we assume $\dim\{\mathrm{Im}\,(\boldsymbol{C}_\perp)\} = k-1$ as the starting point when we assume $\dim\{\mathrm{Im}\,(\boldsymbol{C}_\perp)\} = k$. This ensures that we attribute as much variability as possible to the random effect and only ascribe variability to the latent covariates if the observed signal is not amenable to the model for the variance. Algorithm 3.2 enumerates these steps in ICaSE.

***Algorithm* 3.2** (ICaSE)**.** *Let* $\boldsymbol{W}(\boldsymbol{x}) = \sum_{j=1}^{b} [\boldsymbol{x}]_j \boldsymbol{Q}_X^{\mathrm{T}} \boldsymbol{B}_j \boldsymbol{Q}_X$ *for* $\boldsymbol{x} \in \mathbb{R}^b$.

(0) For $k = 0$*, estimate* δ_*^2 *and* $\boldsymbol{\tau}_*$ *by maximum likelihood using the model* $\boldsymbol{Y}_2 \sim MN_{p\times(n-d)}\left\{\boldsymbol{0}, \delta^2 I_p, \boldsymbol{W}(\boldsymbol{\tau})\right\}$ *under the restriction that* $\det\{\boldsymbol{W}(\boldsymbol{\tau})\} = 1$ *and* $\delta^2\boldsymbol{\tau} \in \Theta$.

(1) Let $\hat{\boldsymbol{W}} = \boldsymbol{W}(\hat{\boldsymbol{\tau}})$*. For* $k \geq 1$ *and given estimates* $\hat{\delta}^2$ *and* $\hat{\boldsymbol{\tau}}$ *for* δ_*^2 *and* $\boldsymbol{\tau}_*$*, estimate* $\hat{\boldsymbol{W}}^{-1/2}\boldsymbol{C}_\perp$ *as the first* k *right singular vectors of* $\boldsymbol{Y}_2\hat{\boldsymbol{W}}^{-1/2}$*. Re-scale the estimate by* $(n-d)^{1/2}\hat{\boldsymbol{W}}^{1/2}$ *to get an estimate for* $\boldsymbol{C}_\perp$*,* $\hat{\boldsymbol{C}}_\perp \in \mathbb{R}^{(n-d)\times k}$.

(2) Given $\hat{\boldsymbol{C}}_\perp$*, obtain* $\hat{\delta}^2$ *and* $\hat{\boldsymbol{\tau}}$ *by restricted maximum likelihood using the model* $\boldsymbol{Y}_2 \sim MN_{p\times(n-d)}\left\{\boldsymbol{L}\hat{\boldsymbol{C}}_\perp^{\mathrm{T}}, \delta^2 I_p, \boldsymbol{W}(\boldsymbol{\tau})\right\}$ *where* $\det\{\boldsymbol{W}(\boldsymbol{\tau})\} = 1$ *and* $\delta^2\boldsymbol{\tau} \in \Theta$.

(3) Iterate between steps (1) and (2) and stop on step (1) of the second iteration. Repeat this for $k = 1, \ldots, K_{\max}$*, using* $\hat{\boldsymbol{\tau}}$*,* $\hat{\delta}^2$ *obtained when* $\dim\left\{\mathrm{Im}\left(\hat{\boldsymbol{C}}_\perp\right)\right\} = k-1$ *as the starting point for step (1) when* $\dim\left\{\mathrm{Im}\left(\hat{\boldsymbol{C}}_\perp\right)\right\} = k$.

***Remark* 3.1.** *If $b = 1$ and $\boldsymbol{B}_1 = I_n$, $\hat{\boldsymbol{C}}_\perp \in \mathbb{R}^{(n-d)\times k}$ is a scalar multiple of the first k right singular vectors of $\boldsymbol{Y}_2$ for each $k \in [K_{\max}]$.*

Evidently, iterating between steps *(1)* and *(2)* is not explicitly maximizing an objective function. However, Theorem 3.1 in Section 3.4 proves that ICaSE's estimates for $\mathrm{Im}\left(\boldsymbol{C}_\perp\right)$ are as accurate as those obtained from principal component analysis when samples are independent.

3.3.3 Correlated Bi-Cross Validation to estimate K

Here we use $\boldsymbol{Y}_2$ and a cross-validation paradigm developed in Owen & Wang (2016) to estimate $K = \dim\left\{\mathrm{Im}\left(\boldsymbol{C}\right)\right\}$. Besides having substantially shorter computation time and providing a less variable estimate for K, what differentiates our method from the one developed in Owen & Wang (2016) is we provide an approach amenable to dependent data, while their method is only valid for independent data.

Our primary concern is ensuring our estimates for K are not biased by the correlations in the data (Hastie et al. 2009). An equivalent concern is we want to avoid including factors arising from the random effect in our estimate for K. To describe our procedure, we assume for notational convenience that $\boldsymbol{Y}_{p\times n} = \boldsymbol{L}_{p\times K}\boldsymbol{C}_{n\times K}^{\mathrm{T}} + \boldsymbol{E}_{p\times n}$ where the rows $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_p$ of $\boldsymbol{E}$ are independent and $\boldsymbol{e}_g \sim N_n\left(\boldsymbol{0}, \boldsymbol{V}_g\right)$ for all $g \in [p]$. Algorithm 3.3 provides an outline of the algorithm we use to estimate K, which we call Correlated Bi-Cross Validation (CBCV).

***Algorithm* 3.3** (CBCV)**.** *Randomly partition the rows of $\boldsymbol{Y}$ (i.e. genes) into $F = O(1)$ folds and let $f \in [F]$. Without loss of generality, assume*

$$\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{Y}_{(-f)} \\ \boldsymbol{Y}_f \end{bmatrix} = \begin{bmatrix} \boldsymbol{L}_{(-f)}\boldsymbol{C}^{\mathrm{T}} \\ \boldsymbol{L}_f\boldsymbol{C}^{\mathrm{T}} \end{bmatrix} + \begin{bmatrix} \boldsymbol{E}_{(-f)} \\ \boldsymbol{E}_f \end{bmatrix}$$

where $\boldsymbol{Y}_{(-f)} \in \mathbb{R}^{p_{(-f)}\times n}$ and $\boldsymbol{Y}_f \in \mathbb{R}^{p_f\times n}$ are the training and test sets.

(a) Let $k \in \{0, 1, \ldots, K_{\max}\}$. Obtain $\hat{\boldsymbol{C}} \in \mathbb{R}^{n \times k}$ and $\hat{\boldsymbol{V}}_{(-f)}$ from $\boldsymbol{Y}_{(-f)}$ using Algorithm 3.2.

(b) Let $\bar{\boldsymbol{Y}}_f = \boldsymbol{Y}_f \hat{\boldsymbol{V}}_{(-f)}^{-1/2}$ and $\hat{\bar{\boldsymbol{C}}} = \hat{\boldsymbol{V}}_{(-f)}^{-1/2} \hat{\boldsymbol{C}}$. Define the loss for this fold, dimension pair as the leave-one-out cross validation loss:

$$\mathcal{L}_f(k) = \sum_{i=1}^{n} \|\bar{\boldsymbol{y}}_{f,i} - \hat{\boldsymbol{L}}_{f,(-i)} \hat{\bar{\boldsymbol{c}}}_i\|_2^2, \tag{3.9}$$

where $\bar{\boldsymbol{y}}_{f,i} \in \mathbb{R}^{p_f}$ and $\hat{\bar{\boldsymbol{c}}}_i \in \mathbb{R}^k$ are the ith columns of $\bar{\boldsymbol{Y}}_f$ and $\hat{\bar{\boldsymbol{C}}}^{\mathrm{T}}$, respectively. $\hat{\boldsymbol{L}}_{f,(-i)}$ is the ordinary least squares regression coefficient from the regression of $\bar{\boldsymbol{Y}}_{f,(-i)}$ onto $\hat{\bar{\boldsymbol{C}}}_{(-i)}$, where $\bar{\boldsymbol{Y}}_{f,(-i)}$ and $\hat{\bar{\boldsymbol{C}}}_{(-i)}^{\mathrm{T}}$ are submatrices of $\bar{\boldsymbol{Y}}_f$ and $\hat{\bar{\boldsymbol{C}}}^{\mathrm{T}}$ with the ith columns removed.

(c) Repeat this for folds $f = 1, \ldots, F$ and $k = 0, 1, \ldots, K_{\max}$ and set $\hat{K}$ to be

$$\hat{K} = \underset{k \in \{0,1,\ldots,K_{\max}\}}{\arg\min} \left\{ \sum_{f=1}^{F} \mathcal{L}_f(k) \right\}. \tag{3.10}$$

Since (3.9) is the leave one out cross validation squared loss using the design matrix $\hat{\bar{\boldsymbol{C}}}$ and the scaled data $\bar{\boldsymbol{Y}}_f$, it only depends on $\operatorname{Im}\left(\hat{\boldsymbol{C}}\right)$ and $\hat{\boldsymbol{V}}_{(-f)}$. Scaling by $\hat{\boldsymbol{V}}_{(-f)}^{-1/2}$ in (3.9) is sensible because it places more importance on correctly estimating the portion of $\boldsymbol{C}$ not captured by the model for the residual variance. However, unless proper care is taken, this scaled loss function would underestimate K simply because the estimated residual variance is larger for underspecified models. The restriction that $\det\left\{\hat{\boldsymbol{V}}_{(-f)}\right\} = 1$ alleviates this issue by making the loss function scale-invariant.

To understand why CBCV gives accurate estimates for K, consider the expected leave-one-out squared error for a particular fold, dimension pair, where $\bar{\boldsymbol{c}}_i$ is the ith row of $\hat{\boldsymbol{V}}_{(-f)}^{-1/2}\boldsymbol{C}$:

$$\begin{aligned}\mathbb{E}\left\{\mathcal{L}_f(k) \mid \boldsymbol{Y}_{(-f)}\right\} =& \underbrace{\sum_{i=1}^{n} \mathbb{E}\left\{\|\boldsymbol{L}_f\bar{\boldsymbol{c}}_i - \hat{\boldsymbol{L}}_{f,(-i)}\hat{\bar{\boldsymbol{c}}}_i\|_2^2 \mid \boldsymbol{Y}_{(-f)}\right\}}_{I} + \underbrace{p_f\delta_{f*}^2 \operatorname{tr}\left\{\hat{\boldsymbol{V}}_{(-f)}^{-1}\boldsymbol{V}_{f*}\right\}}_{II} \\ &+ \underbrace{\text{cross-term}}_{III} \\ \delta_{f*}\boldsymbol{V}_{f*} =& p_f^{-1}\sum_{g\,\in\,\text{fold } f}\boldsymbol{V}_g, \quad \det\left(\boldsymbol{V}_{f*}\right) = 1.\end{aligned}$$

In standard cross validation, $\hat{\boldsymbol{V}}_{(-f)} = I_n$, meaning the residual variance term (II) would be constant for all $k = 0, 1, \ldots, K_{\text{max}}$ and the cross-term (III) would be 0, since $\hat{\boldsymbol{L}}_{f,(-i)}$ would be independent of the ith column of $\boldsymbol{Y}_f$. The minimizer of the squared bias term (I) would then also minimize the expected cross validated error. However, since we must account for the correlation between samples, we now need to ensure that $I + II$ is minimized when $\hat{K} = K$ and that the cross-term does not contribute to the loss. The restriction that $\det\left\{\hat{\boldsymbol{V}}_{(-f)}\right\} = 1$ helps ensure this is the case, since $(np_f)^{-1}II \geq \delta_{f*}^2$, where the inequality holds with equality if and only if $\hat{\boldsymbol{V}}_{(-f)} = \boldsymbol{V}_{f*}$ by Jensen's Inequality. Since an accurate estimate of $\operatorname{Im}(\boldsymbol{C})$ begets an accurate estimate of $\boldsymbol{V}_{(-f)*} \approx \boldsymbol{V}_{f*}$, the minimizer of $I + II$, and also $I + II + III$ if $III \approx 0$, should be close to K. We make this rigorous and prove the consistency of $\hat{K}$ in Theorem 3.2 in Section 3.4.

3.3.4 De-biasing estimates for the main effect in correlated data

Recall from (3.5b) that for some positive definite $\boldsymbol{G} \in \mathbb{R}^{n\times n}$, $\boldsymbol{\Omega} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{G}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{G}^{-1}\boldsymbol{C}$, which helps quantify the proportion of the variability of $\boldsymbol{C}$ explainable by $\boldsymbol{X}$. While $\boldsymbol{C}_{\perp}$ enables one to estimate $\boldsymbol{\ell}_g$ and $\boldsymbol{V}_g$, $\boldsymbol{\Omega}$ allows us to estimate the effect of $\boldsymbol{X}$ on $\boldsymbol{y}_g$ while controlling for the latent factors $\boldsymbol{C}$. This helps make results reproducible.

To estimate $\boldsymbol{\Omega}$, we must first specify a suitable weighting matrix $\boldsymbol{G}$. Section 3.3.1 suggests that a reasonable choice would be $\boldsymbol{G} = \sum_{j=1}^{b} [\hat{\boldsymbol{\tau}}]_j \boldsymbol{B}_j = \hat{\boldsymbol{V}}$, where $\hat{\boldsymbol{\tau}}$ is computed using Algorithm 3.2. From here on out, we assume K is known and define $\boldsymbol{y}_{g_1}, \boldsymbol{\Omega}$ and $\boldsymbol{Y}_1$ in (3.5) by setting $\boldsymbol{G} = \hat{\boldsymbol{V}}$.

Our strategy for estimating $\boldsymbol{\Omega}$ is to regress $\boldsymbol{Y}_1$ onto the estimate for $\boldsymbol{L}$ obtainable after completing Algorithm 3.2, where for $\hat{\boldsymbol{W}} = \boldsymbol{Q}_X^{\mathrm{T}} \hat{\boldsymbol{V}} \boldsymbol{Q}_X$, $\hat{\boldsymbol{L}} = \boldsymbol{Y}_2 \hat{\boldsymbol{W}}^{-1} \hat{\boldsymbol{C}}_{\perp} \left(\hat{\boldsymbol{C}}_{\perp}^{\mathrm{T}} \hat{\boldsymbol{W}}^{-1} \hat{\boldsymbol{C}}_{\perp} \right)^{-1}$. A simple estimator for $\boldsymbol{\Omega}$ is the ordinary least squares estimate $\hat{\boldsymbol{\Omega}}^{\text{shrunk}} = \boldsymbol{Y}_1^{\mathrm{T}} \hat{\boldsymbol{L}} \left(\hat{\boldsymbol{L}}^{\mathrm{T}} \hat{\boldsymbol{L}} \right)^{-1}$. When samples are independent, this strategy is similar to those used in Sun et al. (2012), Houseman et al. (2014), Lee et al. (2017), Wang et al. (2017), Fan & Han (2017), although the exact estimators differ slightly. However, McKennan & Nicolae (2018*a*) showed that $\hat{\boldsymbol{\Omega}}^{\text{shrunk}}$ tends to underestimate $\boldsymbol{\Omega}$ when samples are independent because $\hat{\boldsymbol{L}}$ is a noisy estimate for the design matrix $\boldsymbol{L}$. Therefore, we use the following de-biased estimate of $\boldsymbol{\Omega}$, which is analogous to the estimator used in McKennan & Nicolae (2018*a*) when samples are assumed independent:

$$\hat{\boldsymbol{\Omega}} = \boldsymbol{Y}_1^{\mathrm{T}} \hat{\boldsymbol{L}} \left\{ \hat{\boldsymbol{L}}^{\mathrm{T}} \hat{\boldsymbol{L}} - p\hat{\delta}^2 \left(\hat{\boldsymbol{C}}_{\perp}^{\mathrm{T}} \hat{\boldsymbol{W}}^{-1} \hat{\boldsymbol{C}}_{\perp} \right)^{-1} \right\}^{-1}. \tag{3.11}$$

The $p\hat{\delta}^2 \left(\hat{\boldsymbol{C}}_{\perp}^{\mathrm{T}} \hat{\boldsymbol{W}}^{-1} \hat{\boldsymbol{C}}_{\perp} \right)^{-1}$ term in (3.11) removes the bias in $\hat{\boldsymbol{L}}^{\mathrm{T}} \hat{\boldsymbol{L}}$ and reduces to the bias correction used in McKennan & Nicolae (2018*a*) when $b = 1$ and $\boldsymbol{B}_1 = I_n$.

Lastly, we can express $\boldsymbol{C}$ as $\boldsymbol{C} = \boldsymbol{X\Omega} + \hat{\boldsymbol{V}} \boldsymbol{Q}_X \hat{\boldsymbol{W}}^{-1} \boldsymbol{C}_{\perp}$. Therefore, our estimate for $\boldsymbol{C}$ is

$$\hat{\boldsymbol{C}} = \boldsymbol{X} \hat{\boldsymbol{\Omega}} + \hat{\boldsymbol{V}} \boldsymbol{Q}_X \hat{\boldsymbol{W}}^{-1} \hat{\boldsymbol{C}}_{\perp}. \tag{3.12}$$

We subsequently estimate $\boldsymbol{V}_g$ and $\boldsymbol{\beta}_g$ for all $g \in [p]$ with restricted maximum likelihood followed by generalized least squares using the design matrix $\left[\boldsymbol{X} \, \hat{\boldsymbol{C}} \right]$.

3.4 Theoretical justification of CBCV-CorrConf

Here we provide a theoretical justification for the methodology presented in Sections 3.3.2, 3.3.3 and 3.3.4. For all theoretical results, we assume $d, b, K_{\max} = O(1)$ and $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_p$ are generated by (3.2), where $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_p$ are independent. The latter assumption is based off of the repeated observation in experimental genetic data that, for the purposes of marginal testing, it is sufficient to assume the dependence between genomic units is driven by a relatively small number of latent variables (Leek & Storey 2007, Maksimovic et al. 2015, Gerard & Stephens 2018). It is also a common assumption among the literature that motivates their models with high throughput biological data (Leek & Storey 2008, Sun et al. 2012, Lee et al. 2017, Wang et al. 2017, Dobriban & Owen 2019). We also provide simulation evidence in Section 3.8.3 that suggests our method is robust to dependence among $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_p$.

***Assumption* 3.1.** *$\boldsymbol{X} \in \mathbb{R}^{n \times d}$ and $\boldsymbol{B}_1, \ldots, \boldsymbol{B}_b \in \mathbb{R}^{n \times n}$ are observed, non-random matrices. Let $c_1 > 1$ be a constant. Define $\boldsymbol{M} \in \mathbb{R}^{b \times b}$ to be $[\boldsymbol{M}]_{ij} = n^{-1} \operatorname{tr}\left(\boldsymbol{B}_i \boldsymbol{B}_j\right)$ for all $i, j \in [b]$. For any matrix $\boldsymbol{D}$, let $\boldsymbol{D}_{(-i)}$ be the sub-matrix of $\boldsymbol{D}$ with the ith row removed. Assume:*

(a) $\boldsymbol{C}$ is a random matrix that is independent of $\boldsymbol{E}$, where $\mathbb{V}\left([\boldsymbol{C}]_{ij}\right) < \infty$ for all $i, j \in [n]$ and $\mathbb{V}\{\operatorname{vec}(\boldsymbol{C})\} \succ \boldsymbol{0}$. Further, $\lim_{n\to\infty} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = \boldsymbol{\Sigma}_X \succ \boldsymbol{0}$, $\boldsymbol{M} \succeq c_1^{-1} I_b$, $\boldsymbol{B}_j = \boldsymbol{B}_j^{\mathrm{T}}$, $\|\boldsymbol{B}_j\|_2 \leq c_1$ for all $j \in [b]$ and

$$\boldsymbol{L} \in \boldsymbol{\Theta}_L = \Big\{\boldsymbol{L} \in \mathbb{R}^{p \times K} : \text{For all } g \in [p],\ \boldsymbol{L}_{(-g)} \text{ contains two distinct submatrices of rank } K\Big\}.$$

(b) Let $\boldsymbol{W}_* = \boldsymbol{Q}_X^{\mathrm{T}} \boldsymbol{V}_* \boldsymbol{Q}_X$ and $\boldsymbol{\Psi} = \mathbb{E}\left(n^{-1} \boldsymbol{C}_\perp^{\mathrm{T}} \boldsymbol{W}_*^{-1} \boldsymbol{C}_\perp\right)$. Then $\boldsymbol{\ell}_g^{\mathrm{T}} \boldsymbol{\Psi} \boldsymbol{\ell}_g \leq c_1$ for all $g \in [p]$, and the first K eigenvalues $\lambda_1 > \cdots > \lambda_K > \lambda_{K+1} = 0$ of $np^{-1}\boldsymbol{L}\boldsymbol{\Psi}\boldsymbol{L}^{\mathrm{T}}$ satisfy $\lambda_1, \ldots, \lambda_K \in \left[c_1^{-1}, nc_1\right]$.

(c) $\boldsymbol{V}_g \succeq c_1^{-1} I_n$, $|v_{g,j}| \leq c_1$ *for all* $g \in [p]$ *and* $j \in [b]$. *Further,* $\|n^{-1}\boldsymbol{C}_\perp^{\mathrm{T}} \boldsymbol{W}_*^{-1} \boldsymbol{C}_\perp - \boldsymbol{\Psi}\|_2 = o_P(1)$ *as* $n, p \to \infty$ *and* $\|\boldsymbol{\Psi}\|_2, \|\boldsymbol{\Psi}^{-1}\|_2 \leq c_1$.

(d) Either $\lambda_K \geq c_1^{-1} n$, $\lambda_K \to \infty$ *but* $\lambda_K/n \to 0$, *or* $\lambda_K \leq c_1$ *as* $n, p \to \infty$. *Further,* $1 - \lambda_{k+1}/\lambda_k \geq c_1^{-1}$ *for all* $k \in [K]$ *and* $\lambda_1/\lambda_K \leq c_1$.

(e) p *is a non-decreasing function of* n *such that* n/p, $n^{3/2}/(p\lambda_K) \to 0$ *as* $n \to \infty$.

***Remark* 3.2.** *We prove* $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_p \succ \mathbf{0}$ *and* $\boldsymbol{L}\boldsymbol{\Psi}\boldsymbol{L}^{\mathrm{T}}$ *are identifiable under Assumption 3.1(a) in Proposition 3.1 in Section 3.10.1. The set* $\boldsymbol{\Theta}_L$ *is a classic way to identify parameters in factor models when* $\boldsymbol{V}_g = \sigma_g^2 I_n$ *for all* $g \in [p]$ *(Anderson & Rubin 1956, Wang et al. 2017).*

***Remark* 3.3.** *The assumptions on* $\boldsymbol{C}$ *imply*

$$\mathbb{E}\left(\boldsymbol{y}_g\right) = \boldsymbol{X}\boldsymbol{\beta}_g + \mathbb{E}\left(\boldsymbol{C}\right)\boldsymbol{\ell}_g, \quad \operatorname{Cov}\left([\boldsymbol{y}_g]_i, [\boldsymbol{y}_h]_j\right) = \boldsymbol{\ell}_g^{\mathrm{T}} \boldsymbol{\Psi}_{ij} \boldsymbol{\ell}_h + \left[\boldsymbol{V}_g\right]_{ij} I\left(g = h\right),$$

where $\boldsymbol{\Psi}_{ij}$ *is the covariance between the* i*th and* j*th rows of* $\boldsymbol{C}$. *This is a generalization of factor models commonly used in data with independent samples* $i = 1, \ldots, n$, *which generally assume* $\left[\boldsymbol{V}_g\right]_{ij} = \sigma_g^2 I\left(i = j\right)$ *and* $\boldsymbol{\Psi}_{ij} = \boldsymbol{\Psi} I\left(i = j\right) \in \mathbb{R}^{K \times K}$. *The latter assumption is typically not true in correlated data. For example, latent cell compositions from the same individual in McKennan et al. (2018) are almost surely correlated.*

***Remark* 3.4.** *Our assumptions on* $\boldsymbol{C}$ *are less stringent than that used in the existing latent factor-correction literature, which generally assume* $\boldsymbol{C} \sim MN_{n \times K}\left(\mathbf{0}, I_n, I_K\right)$ *(Owen & Wang 2016, Fan & Han 2017) or* $\boldsymbol{C} \sim MN_{n \times K}\left(\boldsymbol{X}\boldsymbol{A}, I_n, I_K\right)$ *for* $\boldsymbol{A} \in \mathbb{R}^{d \times K}$ *(Wang et al. 2017).*

The condition on $\boldsymbol{M}$ ensures $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p$ are identifiable, and one can easily verify the condition on $\|\boldsymbol{B}_j\|_2$ in data from McKennan et al. (2018) and Martino et al. (2013) discussed in Sections 3.1 and 3.6, as well as in the simulated multi-tissue data from Section 3.5. It also holds in the following general scenarios:

- If $\boldsymbol{B}_j$ is a partition matrix that divides the samples into blocks, then $\|\boldsymbol{B}_j\|_2 \leq c_1$ if the size of the blocks is at most finite.

- If $\boldsymbol{B}_j$ is a genetic relatedness or kinship matrix, it implies the sampling population is relatively homogeneous, which is generally the case in data from founder populations (Knowles et al. 2018). When data come from more than one population, there are typically a small number of large eigenvalues in the relatedness matrix corresponding to global population structure (Novembre et al. 2008). One can include their corresponding eigenvectors as covariates in $\boldsymbol{Z}$ in Model (3.4) to ensure $\|\boldsymbol{B}_j\|_2$ satisfies Assumption 3.1*(a)*.

The eigenvalues $\lambda_1, \ldots, \lambda_K$ defined in Item *(b)* of Assumption 3.1 quantify the average variation in $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_p$ due to $\boldsymbol{C}$ that can be distinguished from that due to $\boldsymbol{X}$. Item *(e)* is standard in the latent factor correction literature when $\boldsymbol{V}_g = \sigma_g^2 I_n$ for all $g \in [p]$ (Wang et al. 2017, McKennan & Nicolae 2018*a*).

***Assumption* 3.2.** *Let $c_2 > 1$ be a constant not dependent on n or p. The estimates $\hat{\delta}^2$ and $\hat{\boldsymbol{\tau}}$ from step (2) in Algorithm 3.2 are such that $\hat{\delta}^2\hat{\boldsymbol{\tau}}$ lies in the convex set*

$$\Theta_* = \Theta \cap \left\{ \boldsymbol{x} \in \boldsymbol{R}^b : \|\boldsymbol{x}\|_2 \leq bc_2c_1, \sum_{j=1}^{b} [\boldsymbol{x}]_j \, \boldsymbol{B}_j - (c_2c_1)^{-1} I_n \succ \mathbf{0} \right\},$$

where c_1 was defined in Assumption 3.1.

This makes the residual variance parameter space compact and is analogous to Assumption D in Bai & Li (2012) and Assumption 2 in Wang et al. (2017). It is also a standard assumption in likelihood theory (Wald 1949, Ferguson 1996, Douc et al. 2004). We now state Theorem 3.1.

***Theorem* 3.1** (Accuracy of ICaSE)**.** *Suppose Assumptions 3.1 and 3.2 hold and we apply Algorithm 3.2 for $k = 1, \ldots, K_{\max}$, where $K \leq K_{\max}$. Then the estimates for δ_*^2, $\boldsymbol{\tau}_*$ and*

$\boldsymbol{C}_\perp$ *satisfy*

$$
\begin{aligned}
|\hat{\delta}^2 - \delta_*^2|, \|\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}_*\|_2 &= O_P\left(n^{-1}\right), \quad k \geq K \\
\|P_{\boldsymbol{C}_\perp} - P_{\hat{\boldsymbol{C}}_\perp}\|_F^2 &= O_P\left\{n/\left(\lambda_K p\right) + \left(p\lambda_K\right)^{-1/2} + \left(n\lambda_K\right)^{-1}\right\}, \quad k = K \geq 1 \\
\|P_{\boldsymbol{C}_\perp} - P_{\hat{\boldsymbol{C}}_\perp} P_{\boldsymbol{C}_\perp}\|_F^2 &= O_P\left\{n/\left(\lambda_K p\right) + \left(p\lambda_K\right)^{-1/2} + \left(n\lambda_K\right)^{-1}\right\}, \quad k \geq K \geq 1
\end{aligned}
$$

where $\hat{\delta}^2$ *and* $\hat{\boldsymbol{\tau}}$ *depend on* k*. Further,* $\|P_{\boldsymbol{C}_\perp} - P_{\hat{\boldsymbol{C}}_\perp}\|_F^2$*,* $\|P_{\boldsymbol{C}_\perp} - P_{\hat{\boldsymbol{C}}_\perp} P_{\boldsymbol{C}_\perp}\|_F^2$ *and the estimates* $\hat{\delta}^2, \hat{\boldsymbol{\tau}}$ *are invariant of the choice of* $\boldsymbol{Q}_X$*.*

Theorem 3.1 implies that $\mathrm{Im}\left(\boldsymbol{C}_\perp\right)$ is estimated well when $k = K$ and its column space is approximately a subspace of $\mathrm{Im}\left(\hat{\boldsymbol{C}}_\perp\right)$ when we overestimate K. This result is quite remarkable because besides the additional factor $\left(n\lambda_K\right)^{-1} + \left(p\lambda_K\right)^{-1/2} << n^{-1/2}$, this is the same rate obtained from principal components analysis when the samples are independent (McKennan & Nicolae 2018*a*). Theorem 3.1 makes no assumption on the starting points for $\hat{\delta}^2$ and $\hat{\boldsymbol{\tau}}$ (other than the REML estimates are estimated in $\boldsymbol{\Theta}_*$), which proves the efficacy of our warm-start technique.

***Theorem* 3.2** (Consistency of CBCV)**.** *Let* $\hat{K}$ *be as defined in* (3.10)*. Suppose the assumptions of Theorem 3.1 hold,* $\lambda_K \geq \delta_*^2 + \epsilon$ *for some constant* $\epsilon > 0$ *and we sample* $\boldsymbol{Q}_X$ *uniformly over the space of all orthonormal bases for* $\boldsymbol{X}^\perp$*. Then* $\lim_{n,p\to\infty} \mathbb{P}\left(\hat{K} = K\right) = 1$*.*

***Remark* 3.5.** *Let* $\boldsymbol{Q}_* \in \mathbb{R}^{n\times(n-d)}$ *be a non-random matrix such that* $P_X^\perp = \boldsymbol{Q}_*\boldsymbol{Q}_*^{\mathrm{T}}$*. As defined,* $\boldsymbol{Q}_X = \boldsymbol{Q}_*\boldsymbol{M}$*, where* $\boldsymbol{M} \in \mathbb{R}^{(n-d)\times(n-d)}$ *is independent of* $\boldsymbol{C}$ *and* $\boldsymbol{E}$*, and is sampled uniformly from the space of all unitary matrices. This helps guarantee that the maximum leverage score of* $\boldsymbol{W}_*^{-1/2}\boldsymbol{C}_\perp$ *is* $o_P(1)$ *as* $n \to \infty$*, and is a common technique to uniformize the leverage scores of a matrix (Mahoney 2011).*

This theorem shows that recovering K is easier when the sample size n is larger, since $\lambda_1, \ldots, \lambda_K$ tend to grow with n. This is why it is advisable to use all of the available samples to estimate K, rather than partitioning the data to alleviate correlation, as discussed in Section 3.1.

The requirement that $\lambda_K \geq \delta_*^2 + \epsilon$ is tight, as Owen & Wang (2016) demonstrate that δ_*^2 is the lower limit of detection when $b = 1$ and $\boldsymbol{B}_1 = I_n$. This lower limit of detection is smaller than parallel analysis', which often fails to recover moderate to small factors (Owen & Wang 2016). We discuss this in more detail after the statement of Theorem 3.5 in Section 3.11. Theorem 3.2 is also, to the best of our knowledge, the first result proving that bi-cross validation-like methods are consistent. The authors of Owen & Perry (2009), Owen & Wang (2016) only showed their estimates for K minimize the expected loss when samples are independent.

***Assumption* 3.3.** *$s = p^{-1} \sum_{g=1}^{p} I\left(\boldsymbol{\beta}_g \neq 0\right) = o\left(n^{-3/2}\lambda_K\right)$, $\max_{g\in[p]} \|\boldsymbol{\beta}_g\|_2 \leq c_3$ for some constant $c_3 > 0$, and for $\boldsymbol{S} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_*^{-1}\boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_*^{-1}\boldsymbol{C}$, $\|\boldsymbol{S}\boldsymbol{\Psi}^{-1}\boldsymbol{S}^{\mathrm{T}}\|_2 = O_P(1)$ as $n, p \to \infty$.*

***Remark* 3.6.** *We prove $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$ are identifiable under Assumptions 3.1(a)-(b), 3.1(e) and the sparsity assumption on s in Proposition 3.2 in Section 3.10.2. Other than the conditions on $n^{-1}\boldsymbol{C}_\perp \boldsymbol{W}_*^{-1}\boldsymbol{C}_\perp$ and $\boldsymbol{S}$ in Assumptions 3.1(b) and 3.3, we make no assumptions about the relationship between $\boldsymbol{C}$ and $\boldsymbol{X}$.*

This provides an explicit relationship between the maximum allowable sparsity on the main effect and the informativeness of the data $\boldsymbol{Y}$ for estimating $\boldsymbol{C}$: the more signal in $\boldsymbol{L}\boldsymbol{C}_\perp^{\mathrm{T}}$, the part of $\boldsymbol{L}\boldsymbol{C}^{\mathrm{T}}$ unequivocally distinguishable from $\boldsymbol{\beta}\boldsymbol{X}^{\mathrm{T}}$, the less stringent Assumption 3.3 becomes. This is the same maximum allowable sparsity assumed in Wang et al. (2017) and McKennan & Nicolae (2018*a*), both of which assume $b = 1$ and $\boldsymbol{B}_1 = I_n$. We show through simulation in Section 3.5 that we can accurately recover $\boldsymbol{C}$ even when this assumption is egregiously violated.

***Theorem* 3.3** (Inference on $\boldsymbol{\beta}$)**.** *Let $g \in [p]$. Suppose Assumptions 3.1, 3.2 and 3.3 hold and we estimate $\boldsymbol{C}$ according to* (3.12)*. Let $\hat{\boldsymbol{v}}_g \in \Theta_*$ and $\hat{\boldsymbol{\beta}}_g$ be the restricted maximum* *likelihood (REML) and generalized least squares (GLS) estimates for $\boldsymbol{v}_g$ and $\boldsymbol{\beta}_g$ using the design matrix $\left[\boldsymbol{X}\,\hat{\boldsymbol{C}}\right]$. Then*

$$\begin{aligned}
&\|\hat{\boldsymbol{V}}_g - \boldsymbol{V}_g\|_2 = o_P(1), \quad \hat{\boldsymbol{V}}_g = \sum_{j=1}^{b} \left[\hat{\boldsymbol{v}}_g\right]_j \boldsymbol{B}_j \\
&\hat{\boldsymbol{M}}_g^{-1/2}\left(\hat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_g\right) \overset{\mathcal{D}}{=} \boldsymbol{Z} + o_P(1) \\
&\hat{\boldsymbol{M}}_g = \left(\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{V}}_g^{-1}\boldsymbol{X}\right)^{-1} + \hat{\boldsymbol{\Omega}}_g\left\{\hat{\boldsymbol{C}}_{\perp}^{\mathrm{T}}\left(\boldsymbol{Q}_X^{\mathrm{T}}\hat{\boldsymbol{V}}_g\boldsymbol{Q}_X\right)^{-1}\hat{\boldsymbol{C}}_{\perp}\right\}^{-1}\hat{\boldsymbol{\Omega}}_g^{\mathrm{T}}
\end{aligned}$$

as $n \to \infty$, where $\boldsymbol{Z} \sim N_d(0, I_d)$ and $\hat{\boldsymbol{\Omega}}_g = \left(\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{V}}_g^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{V}}_g^{-1}\hat{\boldsymbol{C}}$. Further, $\|\hat{\boldsymbol{M}}_g\boldsymbol{M}_g^{-1} - I_d\|_2 = o_P(1)$ as $n \to \infty$, where $\boldsymbol{M}_g$ is the variance of the GLS estimate for $\boldsymbol{\beta}_g$ when $\boldsymbol{C}$ and $\boldsymbol{V}_g$ are known.

***Remark* 3.7.** *The REML and GLS estimates $\hat{\boldsymbol{V}}_g$, $\hat{\boldsymbol{\beta}}_g$ and $\hat{\boldsymbol{M}}_g$ only depend on $\hat{\boldsymbol{C}}$ through* $\mathrm{Im}\left(\hat{\boldsymbol{C}}\right)$. *We prove that conditional on $\boldsymbol{C}$, $\mathrm{Im}(\boldsymbol{C})$ and $\boldsymbol{M}_g$ are identifiable for all $g \in [p]$ in Proposition 3.3 in Section 3.12.2.*

3.5 Simulated multi-tissue gene expression data analysis

3.5.1 Simulation setup and parameters

We simulated the expression of $p = 15,000$ genes from 50 individuals across three tissues with a complicated tissue-by-tissue correlation structure to compare our method against other state of the art methods designed to estimate K and $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$. We include comparisons of our method to estimate $\boldsymbol{C}_{\perp}$ in Section 3.8.1. We first randomly chose 25 individuals to be in the treatment group and set $\boldsymbol{X} \in \{0,1\}^n$ to be the treatment status for the $n = 150$ samples and $\boldsymbol{Z} = \boldsymbol{1}_{50} \otimes I_3$ to be the tissue-specific intercept. For each $g \in [p]$, we let $\boldsymbol{V}_g = I_{50} \otimes \boldsymbol{M}_g$ be the gene-specific covariance matrix for gene g and simulated $\boldsymbol{M}_g \in \mathbb{R}^{3\times 3}$ in each dataset so the average correlation matrix across all p genes was as given in Table 3.1.

Table 3.1: *The correlation matrix corresponding to the covariance matrix* $\lim_{t\to\infty} t^{-1} \sum_{g=1}^{t} \boldsymbol{M}_g$.

	Tissue 1	Tissue 2	Tissue 3
Tissue 1	1	0.72	0.58
Tissue 2	0.72	1	0.80
Tissue 3	0.58	0.80	1

Specifically, we assumed tissues two and three were more similar to one another than to the first and for each $g \in [p]$, set $\boldsymbol{M}_g$ to be

$$\begin{aligned}
&\boldsymbol{M}_g = \mathrm{Cov}\left\{\left(\epsilon_{g1}\ \epsilon_{g2}\ \epsilon_{g3}\right)^{\mathrm{T}}\right\}\\
&\epsilon_{g1} = \alpha_{g1} + \xi_{g1},\ \epsilon_{g2} = \phi_{g2}\alpha_{g1} + \alpha_{g2} + \xi_{g2},\ \epsilon_{g3} = \phi_{g3}\alpha_{g1} + \rho_{g3}\alpha_{g2} + \xi_{g3}\\
&\alpha_{g1} \sim \left(0, v_{g1}^2\right),\ \alpha_{g2} \sim \left(0, v_{g2}^2\right),\ \xi_{gj} \sim \left(0, \sigma_{gj}^2\right)\ \ (j = 1, 2, 3).
\end{aligned}$$

The constants $v_{g1}^2, \phi_{g2}, v_{g2}^2, \phi_{g3}, \rho_{g3}$ and σ_{gj}^2 were simulated from Gamma distributions with means 0.8, 1.25, 0.4, 0.75, 1 and 0.2, respectively, each with coefficient of variation equal to 0.2. We subsequently re-scaled these parameters so that $\delta_*^2 = 1$. This complex tissue-by-tissue covariance is amenable to the variance model assumed in (3.2b), since

$$\boldsymbol{V}_g = \sum_{r=1}^{3}\sum_{s=r}^{3} v_{grs}\left\{I_{50} \otimes \left(\boldsymbol{a}_{rs}\boldsymbol{a}_{rs}^{\mathrm{T}}\right)\right\} = \sum_{r=1}^{3}\sum_{s=r}^{3} v_{grs}\boldsymbol{B}_{rs} \tag{3.13}$$

where $\boldsymbol{a}_{rs} \in \mathbb{R}^3$ has a 1 in the rth and sth coordinates and 0 everywhere else and

$$v_{g12},\ v_{g13},\ v_{g23} \geq 0, \quad v_{g11} + v_{g12} + v_{g13},\ v_{g22} + v_{g12} + v_{g23},\ v_{g33} + v_{g13} + v_{g23} \geq 0.$$

We next simulated data with $K = 10$ latent factors. The parameters $\boldsymbol{X}, \boldsymbol{Z}, K$ and α, defined below, were fixed across all simulations. For $\boldsymbol{W}_* = \boldsymbol{Q}_{[X\ Z]}^{\mathrm{T}}\boldsymbol{V}_*\boldsymbol{Q}_{[X\ Z]}$, we simulated

$\boldsymbol{y}_g \in \mathbb{R}^n$, the expression of gene g in the n (individual, tissue) pairs, as follows:

$$\begin{aligned}
&\boldsymbol{y}_g = \boldsymbol{X}\beta_g + \boldsymbol{C}\boldsymbol{\ell}_g + (3/4)^{1/2}\boldsymbol{V}_g^{1/2}\tilde{\boldsymbol{e}}_g, \quad \left[\tilde{\boldsymbol{e}}_g\right]_i \sim T_8 \ \ (g=1,\ldots,p; i=1,\ldots,n)\\
&\beta_g \sim 0.8\delta_0 + 0.2N_1\left(0, 0.4^2\right) \ \ (g=1,\ldots,p)\\
&\boldsymbol{C} = \alpha\boldsymbol{X}\boldsymbol{1}_K^{\mathrm{T}} + \boldsymbol{\Xi}\left(n^{-1}\boldsymbol{\Xi}^{\mathrm{T}}\boldsymbol{Q}_{[X\,Z]}\boldsymbol{W}_*^{-1}\boldsymbol{Q}_{[X\,Z]}^{\mathrm{T}}\boldsymbol{\Xi}\right)^{-1/2}, \quad \boldsymbol{\Xi} \sim MN_{n\times K}\left(0, I_n, I_K\right)\\
&\left[\boldsymbol{\ell}_g\right]_k \sim \pi_k\delta_0 + \left(1-\pi_k\right)N_1\left(0, \eta_k^2\right) \ \ (g=1,\ldots,p; k=1,\ldots,K).
\end{aligned} \tag{3.14}$$

Here δ_0 is the point mass at 0 and T_8 is the t-distribution with eight degrees of freedom, and was chosen to emulate data with heavy tails. We chose to simulate a non-sparse main effect $\boldsymbol{\beta}$ to show that we can violate Assumption 3.3 and still do inference that is just as accurate as when $\boldsymbol{C}$ is known. The re-scaling of $\boldsymbol{\Xi}$ was simply to make the diagonal elements of $np^{-1}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}$ approximately equal to the eigenvalues $\lambda_1, \ldots, \lambda_{10}$ defined in Assumption 3.1. This re-scaling also caused $\lambda_1, \ldots, \lambda_{10}$ to shrink by a factor of 1.5, on average, thus making it harder to recover K and $\boldsymbol{C}$. The constant α was chosen so that $P_{V_*^{-1/2}Z}^{\perp}\boldsymbol{V}_*^{-1/2}\boldsymbol{C}$ explained approximately 30% of the variability in $P_{V_*^{-1/2}Z}^{\perp}\boldsymbol{V}_*^{-1/2}\boldsymbol{X}$, on average. We defined α by scaling $\boldsymbol{X}, \boldsymbol{C}$ and $\boldsymbol{Z}$ by $\boldsymbol{V}_*^{-1/2}$ because this is what one would do in generalized least squares when the covariance of the observations is $\boldsymbol{V}_*$. The values for π_k, η_k and the resulting $\lambda_1, \ldots, \lambda_{10}$ are provided in Table 3.2.

Table 3.2: *The π_k and η_k values used to simulate $\boldsymbol{L}$ and the resulting average λ_k ($k = 1, \ldots, 10$).*

Factor number (k)	1	2	3	4	5	6	7	8	9	10
π_k	0	0.45	0.60	0.71	0.79	0.85	0.90	0.92	0.94	0.96
η_k	1	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.5
λ_k	143	12.7	9.5	6.6	4.8	3.3	2.5	1.8	1.4	1.4

For each simulated dataset, we let $\boldsymbol{X}$ be the covariate of interest and the tissue specific intercepts, $\boldsymbol{Z}$, be nuisance covariates and estimated K using CBCV with $F = 3$ folds, $\boldsymbol{C}_{\perp}$ and $\boldsymbol{V}_*$ using ICaSE and $\boldsymbol{C}$ using (3.12). We then estimated $\boldsymbol{V}_g$ and $\boldsymbol{\beta}_g$ for each $g \in [p]$

using restricted maximum likelihood and generalized least squares, respectively. We include additional results for data simulated such that conditional on $\boldsymbol{C}$, expression across genes was correlated in Section 3.8.3.

3.5.2 Comparison of estimators for K and $\boldsymbol{\beta}$

Since, as far as we are aware, our methods for estimating K and $\boldsymbol{C}$ are the first methods that account for correlation among samples in high throughput biological data, we compared our estimates for K and $\boldsymbol{\beta}$ with state of the art methods designed for high throughput biological data with independent samples. We first assessed our estimates for K in 100 simulated datasets and compared them with three widely used methods: the method proposed in Bai & Ng (2002), parallel analysis (Buja & Eyuboglu 1992) and bi-cross validation (Owen & Wang 2016). The results are given in Figure 3.1. We could not compare our method to the deterministic version of parallel analysis designed for data with independent samples proposed in Dobriban & Owen (2019), since the authors did not provide R code to implement their method. However, their simulations show that their method tends to select at least as many factors as parallel analysis.

The fact that CBCV recovers K in all simulated datasets is quite remarkable given that we simulate data with heavy tails and $\lambda_K \approx \delta_*^2$ (see Theorem 3.2). We discuss the behavior of CBCV-CorrConf when λ_K is smaller than δ_*^2 in Sections 3.8.3 and 3.11.2. The method proposed in Bai & Ng (2002) severely underestimates K because it is only able to recover latent factors with overtly large effects (Owen & Wang 2016). On the other hand, bi-cross validation and parallel analysis overestimate K because both methods are treating the high dimensional random effect as part of the low dimensional effect $\boldsymbol{L}\boldsymbol{C}_{\perp}^{\mathrm{T}}$. We discuss possible adjustments to ameliorate bi-cross validation's and parallel analysis' estimates for K in these simulated data in Section 3.8.2. Suffice it to say that these adjustments either did not change the estimates, or caused them to consistently underestimate K.

We lastly estimated $\boldsymbol{\beta}$ via generalized least squares with the design matrix $\hat{\boldsymbol{M}} = \left[\boldsymbol{X}\ \boldsymbol{Z}\ \hat{\boldsymbol{C}}\right]$,

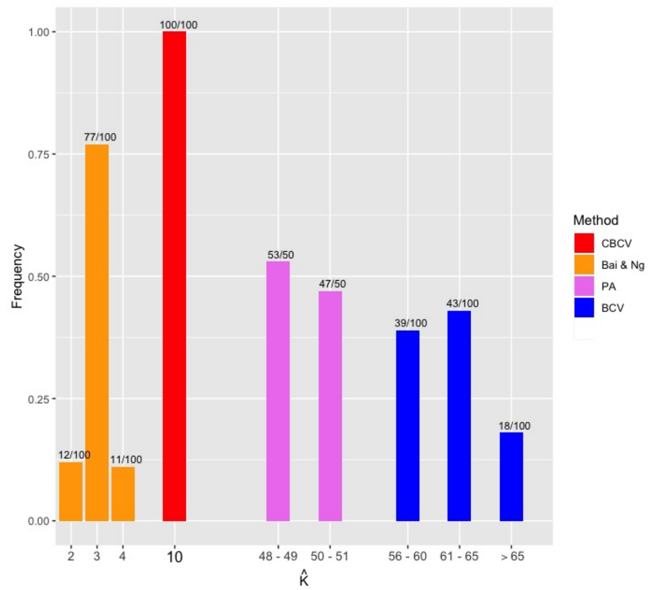


Figure 3.1: Estimates for $K = 10$ in 100 simulated datasets using our proposed method (CBCV), the method proposed in Bai & Ng (2002) (Bai & Ng), parallel analysis (PA) and bi-cross validation (BCV).

where $\hat{\boldsymbol{C}}$ was estimated with CBCV followed by CorrConf (our estimator specified in (3.12)), BCconf (McKennan & Nicolae 2018a), Cate-RR (Wang et al. 2017), dSVA (Lee et al. 2017), IRW-SVA (Leek & Storey 2008), RUV-2 (Gagnon-Bartsch & Speed 2012), RUV-4 (Gagnon-Bartsch et al. 2013) and when $\boldsymbol{C}$ was known. BCconf, Cate-RR, dSVA, IRW-SVA, RUV-2 and RUV-4 were all applied assuming $K = 10$ was known. In order to make the estimation of $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_p$ computationally tractable, we modeled the simulated residuals $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_p$ as

$$\boldsymbol{e}_g \sim N_n\left(\boldsymbol{0}, v_g \boldsymbol{V}\right) \quad (g = 1, \ldots, p), \qquad \boldsymbol{V} = \sum_{r=1}^{3} \sum_{s=r}^{3} \tau_{rs} \boldsymbol{B}_{rs},$$

where $\boldsymbol{B}_{rs}$ was defined in (3.13), and estimated $v_1, \ldots, v_p$ and τ_{rs} with restricted maximum likelihood for each of the eight methods using the estimated design matrix $\hat{\boldsymbol{M}}$. We then computed the P value for the null hypothesis $\beta_g = 0$ for all $g \in [p]$ by comparing the t-statistics to a t-distribution with $n - 4 - K = n - 14$ degrees of freedom, used these P values as input into q-value (Storey 2001) to control the false discovery rate and deemed a gene as being differentially expressed across the two treatment conditions if its q-value was no greater than 0.2. We chose q-value to control the false discovery rate because this is the software most popular among biologists. Figure 3.2 plots the true false discovery proportion in 100 simulated datasets among genes with a q-value less than or equal to 0.2 for each of the eight methods. The false discovery proportion when we completely ignored $\boldsymbol{C}$ was uniformly greater than that of IRW-SVA's.

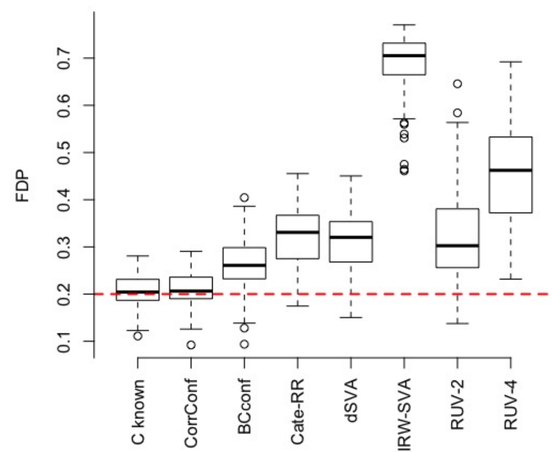


Figure 3.2: The false discovery proportion (FDP) among genes with a q-value no greater than 0.2 in 100 simulated datasets. We randomly chose 90 genes with $\beta_g = 0$ as control genes in RUV-2 and RUV-4. The FDP when we ignored $\boldsymbol{C}$ was uniformly greater than IRW-SVA's.

The performance of our method, CorrConf, as the statement of Theorem 3.3 suggests, is nearly indistinguishable from the generalized least squares estimator when $\boldsymbol{C}$ is known, with nearly identical power. On the other hand, the other six methods tend to introduce more type I errors because their estimates for $\boldsymbol{C}$ do not account for the dependence between the residuals and therefore fail to recover $\boldsymbol{\Omega}$. When K is underestimated, these six methods perform even worse than when $K = 10$ is known, and we show in Section 3.8.3 that these methods still fail to control the false discovery rate and, due to the reduction in residual degrees of freedom, exhibit a decrease in power when K is overestimated to the extent suggested by bi-cross validation and parallel analysis in Figure 3.1.

3.6 Sex-specific DNA methylation in a longitudinal twin study

We next applied our method to identify sex-specific DNA methylation patterns from a longitudinal twin study using data previously published in Martino et al. (2013). The authors measured the DNA methylation of 10 monozygotic (MZ) and 5 dizygotic (DZ) Australian twin pairs (all DZ twins were both male or both female) at birth and 18 months on the Infinium HumanMethylation450 BeadChip platform in buccal epithelium, a relatively homogeneous tissue. After probe and sample quality filtering and data-normalization, the authors were left with $p = 330,168$ methylation sites (CpGs) whose methylation was quantified in 29 male and 24 female ($n = 53$) samples as the difference between log-methylated and log-unmethylated probe intensity (see Martino et al. (2013) for all pre-processing steps). We then used our proposed method (CBCV-CorrConf), BCconf, Cate-RR, dSVA and IRW-SVA to identify CpGs whose methylation differs in males and females, which we refer to as sex-associated CpGs. We subsequently validated each method's findings using sex-associated CpGs identified at birth in previous studies with substantially larger sample sizes. We did not compare our method with RUV-2 or RUV-4, since we did not have access to control CpGs.

We first show that we can write the covariance of the 53 observations at each CpG as a linear combination of six matrices. Let $y_{m,t,a}$ be the measured DNA methylation at an arbitrary CpG for twin $t \in \{1, 2\}$, from mother $m \in [15]$ at age $a \in \{0, 18\}$, where samples with different mothers were assumed to be independent and twin 1 and twin 2 from the same mother were assumed to be exchangeable. A preliminary analysis showed that the correlation between MZ and DZ twins' methylomes was approximately the same at both ages, which is consistent with the observation that methylation patterns are in large part determined by environmental exposures (Galanter et al. 2017, Martin & Fry 2018). Therefore, the 4×4 covariance matrix for $\boldsymbol{y}_m = (y_{m,1,0} \quad y_{m,1,18} \quad y_{m,2,0} \quad y_{m,2,18})^{\mathrm{T}}$ completely determined the $n \times n$ covariance matrix for each CpG. We show in Section 3.9 that by only making assumptions on the pairwise covariances, one can write $\text{Cov}\left(\boldsymbol{y}_m\right)$ as

$$\begin{aligned}\text{Cov}\left(\boldsymbol{y}_m\right) =& v_\alpha \mathbf{1}_4\mathbf{1}_4^{\mathrm{T}} + v_\eta\left(\mathbf{1}_2\mathbf{1}_2^{\mathrm{T}} \oplus \mathbf{1}_2\mathbf{1}_2^{\mathrm{T}}\right) + v_{\phi,0}\boldsymbol{a}_0\boldsymbol{a}_0^{\mathrm{T}} + v_{\phi,18}\boldsymbol{a}_{18}\boldsymbol{a}_{18}^{\mathrm{T}} + v_0 \,\text{diag}\left(\boldsymbol{a}_0\right) \\ &+ v_{18}\,\text{diag}\left(\boldsymbol{a}_{18}\right) \end{aligned} \tag{3.15}$$

where $\boldsymbol{a}_0^{\mathrm{T}} = (1,0,1,0)$, $\boldsymbol{a}_{18}^{\mathrm{T}} = (0,1,0,1)$. The variance multipliers are such that for $t_1 \neq t_2$,

$$\begin{aligned}
&v_\alpha = \text{Cov}\left(y_{m,t_1,0}, y_{m,t_2,18}\right) \\
&v_\alpha + v_\eta = \text{Cov}\left(y_{m,t_1,0}, y_{m,t_1,18}\right), \; v_\alpha + v_{\phi,a} = \text{Cov}\left(y_{m,t_1,a}, y_{m,t_2,a}\right) \quad (a = 0, 18) \\
&v_\alpha + v_\eta + v_{\phi,a} + v_a = \mathbb{V}\left(y_{m,t_1,a}\right) \quad (a = 0, 18)
\end{aligned}$$

and lie in a convex polytope defined in (3.3). We derive these variance multiplier relationships in Section 3.9.

Since there was no evidence that the difference in methylation between males and females changed from birth to 18 months, we assumed the methylation at each CpG was a linear combination of the subject's age (birth or 18 months), sex and other unobserved factors to be estimated, where age was a nuisance covariate and sex was the phenotype of interest. We used CBCV with $F = 5$ folds and estimated K to be 2, and subsequently estimated $\boldsymbol{V}_*$ and $\boldsymbol{C}$ with ICaSE and CorrConf. Our estimates for the six average variance multipliers were all greater than 0, the average residual variance at 18 months was 25% larger than that at birth and the correlation between methylation for twins at 18 months was nearly 20% larger than that at birth. This indicated this set of twin's methylomes tended to converge over the first 18 months of life, which is consistent with previous observations that one's methylome reflects one's environmental exposure history (Galanter et al. 2017, Martin & Fry 2018).

We next computed each of the other four method's estimates for $\boldsymbol{C}$ by first using each method's default software to choose K: bi-cross validation, the default for Cate-RR and

BCconf, or parallel analysis, the default for dSVA and IRW-SVA. These methods estimated K to be 4 and 15, respectively. The fact that both estimated the latent factor dimension to be greater than CBCV's estimate of 2 is not surprising, as these methods tend to overestimate K when samples are correlated. We subsequently estimated $\boldsymbol{C}$ with each of the four methods using their software's default settings.

We lastly estimated the effect due to sex on methylation and corresponding q-values to control the false discovery rate, exactly as we did for the simulated data in Section 3.5. We deemed a CpG a sex-associated CpG if its q-value in that method was no greater than 0.2. Since we did not know the ground truth, we used sex-associated CpGs identified at birth in Yousefi et al. (2015) and Maschietto et al. (2017) as a validation set to help judge the veracity of each method's findings. Yousefi et al. (2015) and Maschietto et al. (2017) measured DNA methylation in umbilical cord blood on the Infinium HumanMethylation450 BeadChip platform in children born to 111 unrelated Brazilian and 71 unrelated Mexican American mothers, respectively. The authors of both studies measured and corrected for cord blood cellular composition and identified 2,355 and 1,928 sex-associated CpGs that were also among the 330,168 CpGs studied in Martino et al. (2013). Table 3.3 gives the fraction of sex-associated CpGs identified using $\hat{\boldsymbol{C}}$ estimated with our method (CBCV-CorrConf), along with the other four methods, that are also among the 3,532 sex-associated CpGs identified in Maschietto et al. (2017) or Yousefi et al. (2015). Since the other four methods are not designed for dependent data, we discuss two adjustments designed to alleviate potential biases in their estimates for both K and $\boldsymbol{C}$ in Section 3.9. The overlaps with the modified methods were at least as poor as those presented in Table 3.3.

Table 3.3: *The fraction of sex-associated CpGs identified using data from Martino et al. (2013) that are also one of the 3,532 sex-associated CpGs confidently identified in Yousefi et al. (2015) or Maschietto et al. (2017).*

CBCV-CorrConf ($K=2$)	BCconf ($K=4$)	Cate-RR ($K=4$)	dSVA ($K=15$)	IRW-SVA ($K=15$)
38% (278/729)	23% (424/1839)	20% (474/2404)	19% (487/2517)	28% (341/1240)

While it may be the case that most of BCconf, Cate-RR, dSVA and IRW-SVA are actual sex-associated CpGs, the results in Table 3.3 mirror the trends observed in Figure 3.2. That is, while these four methods nominally identify more sex-associated CpGs, we are less confident in their results because their estimates for $\boldsymbol{C}$ reduce the residual variance but likely do not suitably account for the variation in sex explainable by $\boldsymbol{C}$, which makes their results less reproducible.

These results also highlight the importance of the choice of K. Estimating K with CBCV, and cross-validation in general, tends to yield more reproducible results because we only include a latent factor if prediction performs suitably well on new, held-out data. When we applied all five methods with $K = 2$, BCconf, Cate-RR and dSVA performed similarly with overlaps no greater than 30%. However, 272 out of IRW-SVA's 662 sex-associated CpGs (41%) were in the validation set, which is nearly identical to CorrConf's results in Table 3.3. Similarly, when we set $K = 4$ for all methods, dSVA performed nearly identically to Cate-RR, whereas CorrConf and IRW-SVA had overlaps of 27% and 26%, respectively, and both ostensibly identified approximately 1,500 sex-associated CpGs. This similarity arises because unlike BCconf, Cate-RR and dSVA, IRW-SVA circumvents estimating $\boldsymbol{\Omega}$ with a noisy estimate for $\boldsymbol{L}$ and estimates $\boldsymbol{C}$ by performing factor analysis directly on $\boldsymbol{Y}$, restricted to the genomic units that show little marginal correlation with $\boldsymbol{X}$. In fact, we use the proof of Theorem 3.1 in Sections 3.8.4 and 3.9 to show that under certain conditions on $\boldsymbol{C}, \boldsymbol{V}_*$ and λ_K, which appear to be satisfied in these data, IRW-SVA can accurately recover $\boldsymbol{C}$ even when residuals are correlated. We believe $K = 2$ is the most appropriate choice of K for this dataset because CorrConf's estimate for $\boldsymbol{C}$ appears to explain enough of the variance in methylation to achieve reasonable power, while also accurately recovering $\boldsymbol{\Omega}$ to control for false discoveries and ensuring that the results are reproducible.

3.7 Discussion

To the best of our knowledge, we have developed the first method to account for latent factors in high throughput biological data with correlated samples. We proved that our estimate for K is consistent and that our estimate for $\mathrm{Im}\left(\boldsymbol{C}\right)$ is accurate enough so that inference on the main effects $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$ is just as accurate as when $\boldsymbol{C}$ is known. We also demonstrated our method's finite sample properties by analyzing complex, multi-tissue simulated gene expression data, and used a real longitudinal DNA methylation data from a twin study to show our method tends to give more reproducible results compared to existing methods.

Our proposed procedure is certainly not a panacea for data with arbitrary correlation structure, and relies on the residual variance $\boldsymbol{V}_g$ being a linear combination of known matrices. Data with more complex, non-linear sample correlation structure may not be amenable to (3.2b), since a linear combination of p non-linear functions will not necessarily have an apriori known functional form. This could be an interesting area of future research.

3.8 Additional simulation results

3.8.1 *Simulation results for ICaSE*

Here we compare our estimates for the column space of $\boldsymbol{C}_\perp$ to other estimators using the simulated multi-tissue data from Section 3.5. Assuming $K = 10$ was known, we compared ICaSE's estimates for $\boldsymbol{C}_\perp$ (see Algorithm 3.2) with the accuracy of K-partial SVD (Pearson 1901) and maximum likelihood (Bai & Li 2012), which first estimates $\boldsymbol{L}$ and the diagonal matrix $\boldsymbol{\Sigma}$ under the quasi likelihood model $\boldsymbol{Y}_2 \sim MN_{p\times(n-4)}\left(\boldsymbol{0}, \boldsymbol{L}\boldsymbol{L}^{\mathrm{T}} + \boldsymbol{\Sigma}, I_{n-4}\right)$, and sets $\hat{\boldsymbol{C}}_\perp = \boldsymbol{Y}_2^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{L}}\left(\hat{\boldsymbol{L}}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{L}}\right)^{-1}$. For each estimate $\hat{\boldsymbol{C}}_\perp$, we measured the angle between the column space of $\boldsymbol{C}_\perp$, $\mathrm{Im}(\boldsymbol{C}_\perp)$, and $\mathrm{Im}\left(\hat{\boldsymbol{C}}_\perp\right)$ as

$$\measuredangle\left(\boldsymbol{C}_\perp, \hat{\boldsymbol{C}}_\perp\right) = \max_{\boldsymbol{v}\in\mathrm{Im}(\boldsymbol{C}_\perp)\setminus\{\boldsymbol{0}\}} \left[\min_{\hat{\boldsymbol{v}}\in\mathrm{Im}\left(\hat{\boldsymbol{C}}_\perp\right)\setminus\{\boldsymbol{0}\}} \left\{ \cos^{-1}\left(\frac{\hat{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{v}}{\|\hat{\boldsymbol{v}}\|_2 \|\boldsymbol{v}\|_2} \right) \right\} \right],$$

which is a symmetric function, provided the dimensions of $\mathrm{Im}(\boldsymbol{C}_\perp)$ and $\mathrm{Im}\left(\hat{\boldsymbol{C}}_\perp\right)$ are the same. In order to benchmark the performance of ICaSE, we also simulated additional datasets $\bar{\boldsymbol{Y}} \in \mathbb{R}^{15{,}000\times 150}$ with independent columns, which were generated with the parameters given in (3.14) and Table 3.2, except we fixed $\boldsymbol{V}_1 = \cdots = \boldsymbol{V}_p = I_n$. The angles between the actual and estimated subspace for 50 simulated datasets $\boldsymbol{Y}$ and $\bar{\boldsymbol{Y}}$ are summarized in Figure 3.3. Just as Theorem 3.1 predicts, ICaSE accurately estimates the column space of $\boldsymbol{C}_\perp$, whereas naive singular value decomposition and maximum quasi likelihood that ignores the between-sample correlation cannot recover the latent subspace.

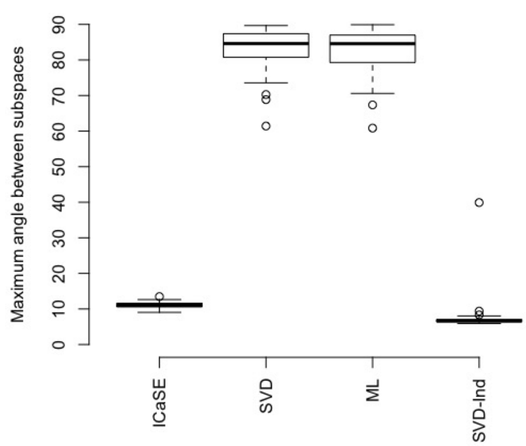


Figure 3.3: The angle between the actual and estimated $\boldsymbol{C}_\perp$ for 50 simulated datasets $\boldsymbol{Y}$, where $\hat{\boldsymbol{C}}_\perp$ was estimated using our proposed method ICaSE, the first K right singular vectors of $\boldsymbol{Y}_2$ (SVD) and maximum quasi likelihood (ML), all with $K = 10$ known. Independent SVD (SVD-Ind) estimated $\boldsymbol{C}_\perp$ as the first $K = 10$ right singular vectors of $\bar{\boldsymbol{Y}}_2$, which was generated with independent samples.

3.8.2 *Modifying existing methods that estimate K to account for sample correlation*

Given the importance of the choice of K in estimating $\boldsymbol{\beta}$ and factor analysis in general (Hayton et al. 2004, Brown 2015, Dobriban 2017), we discuss three possible adjustments one might suggest to attempt to ameliorate bi-cross validation's and parallel analysis' estimates for K in simulated data from Section 3.5. The simplest would be to merely estimate K, and subsequently $\boldsymbol{C}$, separately for each tissue, since gene expression is assumed to be independent across individuals. If the data within each tissue were sufficiently informative for K and $\boldsymbol{C}$, this procedure should estimate the within-tissue factor dimension to be 10 and $\hat{\boldsymbol{C}}_t \approx \boldsymbol{C}_t \boldsymbol{R}_t$ where $\boldsymbol{C}_t$ is $\boldsymbol{C}$ restricted to the t^{th} tissue and $\boldsymbol{R}_t$ is an invertible matrix ($t = 1, 2, 3$). The final estimate for K would then be $3 \times 10 = 30$ and $\hat{\boldsymbol{C}} \approx \boldsymbol{\Pi} \left(\boldsymbol{C}_1 \boldsymbol{R}_1 \oplus \boldsymbol{C}_2 \boldsymbol{R}_2 \oplus \boldsymbol{C}_3 \boldsymbol{R}_3 \right)$, where $\boldsymbol{\Pi}$ is a permutation matrix that reorders (individual, tissue) pairs. We call this method of estimating $\boldsymbol{C}$ the **partition method** and use it to analyze dependent data with existing methods in Section 3.8.3 below. While $\hat{\boldsymbol{C}}$ will give approximately unbiased estimates for $\boldsymbol{\beta}$, there will be a reduction in the residual degrees of freedom, and therefore power. However, analyzing each tissue separately effectively reduces the sample size (and therefore the eigenvalues $\lambda_1, \ldots, \lambda_K$) by 67% in this simulation example, which is why depending on the analyzed tissue, bi-cross validation and parallel analysis only estimate the within-tissue factor dimension to be anywhere from 1 to 4, which is a marked underestimate of K.

To discuss the remaining two adjustments, we note that in the simulation example from Section 3.5, $\boldsymbol{V}_g \approx a_g I_n + b_g \boldsymbol{B} = a_g I_n + b_g \boldsymbol{Z}_B \boldsymbol{Z}_B^{\mathrm{T}}$, where $\boldsymbol{B}$ is a partition matrix that groups the n samples into $n/3$ individuals and the columns of $\boldsymbol{Z}_B \in \mathbb{R}^{n \times n/3}$ are indicators specifying from which individual the sample originated. The second alteration would be to include $\boldsymbol{Z}_B$ in the set of nuisance covariates and restrict $\boldsymbol{Y}$ to the set of $2n/3$ within-individual contrasts. However, this effectively reduces the sample size from n to $2n/3$ and shrinks the λ_k's by at least 33%, thereby making it harder to differentiate the latent signal $\boldsymbol{L}\boldsymbol{C}_\perp^{\mathrm{T}}$ from the noise. In fact, bi-cross validation and parallel analysis had median estimates of K equal to 8 and 6, respectively, using this alteration. The third adjustment, which avoids dramatically reducing the sample size, is to rotate $\boldsymbol{Y}_2$ by the eigenvectors of $\boldsymbol{Q}_{[X\ Z]}^{\mathrm{T}} \boldsymbol{B} \boldsymbol{Q}_{[X\ Z]}$, which in this simulation example shrinks the between-sample dependence but increases the heterogeneity of the sample-specific residual variances. Since parallel analysis only compares the singular values of $\boldsymbol{Y}_2$ with singular values under a bootstrapped null model, rotating $\boldsymbol{Y}_2$ did not change parallel analysis' estimates for K. On the other hand, bi-cross validation's estimates for K should change because cross validation will not be as sensitive to correlations between samples. However, its estimates for $\boldsymbol{C}_{\perp}$ will still be inaccurate because of the heterogeneity in sample-specific residual variances, which is why its median estimate for K was only 5 in this simulation example.

3.8.3 Simulation results when expression across genes is dependent conditional on $\boldsymbol{C}$

Simulation scenarios

Here we include three additional simulation scenarios when $\boldsymbol{y}_g, \boldsymbol{y}_h$ are simulated to be dependent conditional on $\boldsymbol{C}$ for all $g, h \in [p]$. The fixed parameters $n, p, \alpha, \boldsymbol{X}$ and $\boldsymbol{Z}$ were the same as those defined in Section 3.5.1. We generated 100 synthetic data sets in each of the three simulation scenarios, where $\boldsymbol{\beta}_g$, $\boldsymbol{\ell}_g$, $\boldsymbol{C}$ and $\boldsymbol{V}_g$ were all simulated as described in Section 3.5.1. We describe each simulation scenario below, and subsequently describe five different ways we analyze each simulated data set in Section 3.8.3. Our results are given in Sections 3.8.3, 3.8.3 and 3.8.3.

(1) The purpose of this simulation was to violate the assumption that $\lambda_K > \delta_*^2$ needed to prove the consistency of CBCV in Theorem 3.2, where $\lambda_1, \ldots, \lambda_K$ are the eigenvalues defined in Assumption 3.1. Here, we simulated data with $K = 15$, where $\lambda_1, \ldots, \lambda_{10} > \delta_*^2 = 1$ and $\lambda_{11}, \ldots, \lambda_{15} < \delta_*^2 = 1$, meaning CBCV would likely fail to recover all 15 factors. Therefore, conditional on the factors that CBCV could recover, the expression

across genes was dependent in this simulation. We simulated data as

$$\boldsymbol{y}_g = \left\{ \boldsymbol{X}\boldsymbol{\beta}_g + \boldsymbol{C}\boldsymbol{\ell}_g + (3/4)^{1/2}\boldsymbol{V}_g^{1/2}\tilde{\boldsymbol{e}}_g \right\} + \tilde{\boldsymbol{C}}\boldsymbol{f}_g \quad (g = 1, \dots, p),$$

where $\tilde{\boldsymbol{e}}_g$ was were simulated according to (3.14). The parameters $\tilde{\boldsymbol{C}} \in \mathbb{R}^{n \times 5}$ and $\boldsymbol{f}_g \in \mathbb{R}^5$ were simulated as

$$\tilde{\boldsymbol{C}} \sim \boldsymbol{\Xi}\left(n^{-1}\boldsymbol{\Xi}^{\mathrm{T}}\boldsymbol{Q}_{[X\,Z]}\boldsymbol{W}_*^{-1}\boldsymbol{Q}_{[X\,Z]}^{\mathrm{T}}\boldsymbol{\Xi}\right)^{-1/2}, \quad \boldsymbol{\Xi} \sim MN_{n\times 5}\left(0, I_n, I_5\right)$$
$$\left[\boldsymbol{f}_g\right]_k \sim \pi_{(k+10)}\delta_0 + \left\{1 - \pi_{(k+10)}\right\} N_1\left\{0, \eta_{(k+10)}^2\right\} \quad (g = 1, \dots, p; k = 1, \dots, 5).$$

Note that $\left[\boldsymbol{\ell}_g\right]_k \sim \pi_k\delta_0 + (1 - \pi_k)\, N_1\left(0, \eta_k^2\right)$ for $k = 1, \dots, 10$, and just as we did for $\boldsymbol{C}$, we re-scaled $\boldsymbol{\Xi}$ by $\left(n^{-1}\boldsymbol{\Xi}^{\mathrm{T}}\boldsymbol{Q}_{[X\,Z]}\boldsymbol{W}_*^{-1}\boldsymbol{Q}_{[X\,Z]}^{\mathrm{T}}\boldsymbol{\Xi}\right)^{-1/2}$ to simply make the diagonal elements of $np^{-1}\sum\limits_{g=1}^{p}\boldsymbol{f}_g\boldsymbol{f}_g^{\mathrm{T}}$ approximately equal to the last 5 non-zero eigenvalues $\lambda_{11}, \dots, \lambda_{15}$ of

$$\mathcal{I} = \boldsymbol{W}_*^{-1/2}\left[\boldsymbol{C}_\perp\, \tilde{\boldsymbol{C}}_\perp\right]\left(p^{-1}\sum_{g=1}^{p}\begin{bmatrix}\boldsymbol{\ell}_g\\ \boldsymbol{f}_g\end{bmatrix}\begin{bmatrix}\boldsymbol{\ell}_g\\ \boldsymbol{f}_g\end{bmatrix}^{\mathrm{T}}\right)\boldsymbol{W}_*^{-1/2}\left[\boldsymbol{C}_\perp\, \tilde{\boldsymbol{C}}_\perp\right]^{\mathrm{T}}.$$

Tables 3.4 and 3.5 below give the π_k, η_k for both $\boldsymbol{\ell}_g$ and $\boldsymbol{f}_g$, as well as the resulting average $\lambda_1, \dots, \lambda_{10}$ and $\lambda_{11}, \dots, \lambda_{15}$, where $\lambda_1, \dots, \lambda_{15}$ are the non-zero eigenvalues of $\mathcal{I}$.

Table 3.4: *The* π_k *and* η_k *values used to simulate* $\boldsymbol{\ell}_g$ *(*$g = 1, \ldots, p$*) and the resulting average* λ_k *(*$k = 1, \ldots, 10$*).*

Factor number (k)	1	2	3	4	5	6	7	8	9	10
π_k	0	0.45	0.60	0.71	0.79	0.85	0.90	0.92	0.94	0.96
η_k	1	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.5
λ_k	143	12.7	9.5	6.6	4.8	3.3	2.5	1.8	1.4	1.4

Table 3.5: *The* π_k *and* η_k *values used to simulate* $\boldsymbol{f}_g$ *(*$g = 1, \ldots, p$*) and the resulting average* λ_k *(*$k = 11, \ldots, 15$*).*

Factor number (k)	11	12	13	14	15
π_k	0.98	0.987	0.993	0.997	0.998
η_k	0.5	0.5	0.5	0.5	0.5
λ_k	0.75	0.5	0.25	0.125	0.0625

(2) The purpose of this simulation was to emulate data where groups of genes interact with one another through genetic pathways. We partitioned each of the $p = 15,000$ genes into 5000 groups of 3 genes, and simulated the residuals $\boldsymbol{e}_g \in \mathbb{R}^n$ ($g = 1, \ldots, p$) such that residuals in the same group were correlated. Specifically, let $S_r = \{3r-2, 3r-1, 3r\}$ for $r = 1, \ldots, 5000$ be a partition of the 15000 genes into groups of 3. We simulated the data as

$$[\boldsymbol{y}_{3r-2}\, \boldsymbol{y}_{3r-1}\, \boldsymbol{y}_{3r}] = \boldsymbol{X}\, [\boldsymbol{\beta}_{3r-2}\, \boldsymbol{\beta}_{3r-1}\, \boldsymbol{\beta}_{3r}] + \boldsymbol{C}\, [\boldsymbol{\ell}_{3r-2}\, \boldsymbol{\ell}_{3r-1}\, \boldsymbol{\ell}_{3r}] + \boldsymbol{E}_r \quad (r = 1, \ldots, 5000)$$

Define $\tilde{\boldsymbol{E}}_r \in \mathbb{R}^{n \times 3}$ to be a random matrix whose entries are distributed as $(3/4)^{1/2}\, T_8$

and $[\cdot]_s$ to be the sth column of a matrix. We simulated $\boldsymbol{E}_r$ as follows:

$$[\boldsymbol{E}_r]_s = \boldsymbol{V}_{3r-s+1}^{1/2} \left[\tilde{\boldsymbol{E}}_r \boldsymbol{P}_r^{1/2} \right]_s \quad (s = 1, 2, 3; r = 1, \ldots, 5000)$$

$$[\boldsymbol{P}_r]_{st} = \begin{cases} 1 & \text{if } s = t \\ \left\{ \rho_{st}^{(r)} + \rho_{ts}^{(r)} \right\} /2 & \text{if } s \neq t \end{cases} \quad (s, t = 1, 2, 3; r = 1, \ldots, 5000)$$

$$\log \left\{ \frac{1 - \rho_{st}^{(r)}}{1 + \rho_{st}^{(r)}} \right\} \sim N\left(0, \sigma_\rho^2\right) \mid \boldsymbol{P}_r \succ \boldsymbol{0} \quad (s, t = 1, 2, 3; r = 1, \ldots, 5000).$$

The parameter $\sigma_\rho^2 = 0.6^2$ was chosen so that $(5000)^{-1} \sum_{r=1}^{5000} \kappa\left(\boldsymbol{P}_r\right) \approx 2$, where $\kappa(\cdot)$ is the condition number of a matrix. Define $\left[\boldsymbol{V}_g^{1/2}\right]_i$ to be the ith column of $\boldsymbol{V}_g^{1/2}$. The covariance of the entries of $\boldsymbol{E}_r$ was then

$$\text{Cov}\left([\boldsymbol{E}_r]_{is}, [\boldsymbol{E}_r]_{jt} \right) = [\boldsymbol{P}_r]_{st} \left[\boldsymbol{V}_{3r-s+1}^{1/2} \right]_i^{\mathrm{T}} \left[\boldsymbol{V}_{3r-t+1}^{1/2} \right]_j \quad (s, t = 1, 2, 3; r = 1, \ldots, 5000).$$

Note that if $\boldsymbol{V}_1 = \cdots = \boldsymbol{V}_p = \boldsymbol{V}$, this procedure would simulate the residual matrix $\boldsymbol{E} \in \mathbb{R}^{p \times n}$ as $\boldsymbol{R}\tilde{\boldsymbol{E}}\boldsymbol{V}^{1/2}$, where the entries of $\tilde{\boldsymbol{E}}$ are independent $(3/4)^{1/2}T_8$ and $\boldsymbol{R} \in \mathbb{R}^{p \times p}$ is block diagonal with rth block equal to $\boldsymbol{P}_r^{1/2} \in \mathbb{R}^{3 \times 3}$ for $r = 1, \ldots, 5000$.

(3) This simulation was the same as Simulation (2), except all 15000 genes were partitioned into 3000 groups of 5 genes. Here, $(3000)^{-1} \sum_{r=1}^{3000} \kappa\left(\boldsymbol{P}_r\right) \approx 4.3$ and

$$\text{median}\left\{ \kappa\left(\boldsymbol{P}_1\right), \ldots, \kappa\left(\boldsymbol{P}_{3000}\right) \right\} \approx 3.3.$$

Analysis techniques for existing methods designed for data with independent samples $i = 1, \dots, n$

Next, we compared our method, CBCV-CorrConf, with five different ways of using existing methods to analyze the aforementioned simulated multi-tissue gene expression data. In all five analyses, we randomly chose 90 genes with $\boldsymbol{\beta}_g = 0$ to act as control genes when estimating $\boldsymbol{C}$ with RUV-2 and RUV-4. We describe the five different analysis methods below.

(A) Just as we did in Section 3.5.2, we estimated $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_p$ by performing generalized least squares using each method's estimate for $\boldsymbol{C} \in \mathbb{R}^{n \times K}$, except we used each method's software's default algorithm to choose K. Besides our method, CorrConf, every other method uses either parallel analysis or bi-cross validation to estimate K. We computed P values by comparing t-statistics to a t-distribution with $n-4-\hat{K}$ degrees of freedom.

(B) The same as (A), except we estimated $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_p$ with ordinary least squares using existing method's estimates for $\boldsymbol{C}$.

(C) For all methods but CBCV-CorrConf, we averaged the simulated expression across tissues as a way to remove the correlation between samples $i = 1, \dots, n$. We then used existing methods designed for independent samples to estimate $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_p$ with ordinary least squares using the averaged data matrix $\bar{\boldsymbol{Y}} \in \mathbb{R}^{p \times 50}$. We used each method's software default to choose K.

(D) We used the partition method described in Section 3.8.2 to estimate $\boldsymbol{C}$ with existing methods, where we used each existing method's software default to choose K. Once we estimated $\boldsymbol{C}$, the analysis was identical to that of (A).

(E) The same as (D), except we estimated $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_p$ with ordinary least squares using existing methods' estimates for $\boldsymbol{C}$.

The methods BCconf (McKennan & Nicolae 2018*a*), Cate-RR (Wang et al. 2017), RUV-2 (Gagnon-Bartsch & Speed 2012) and RUV-4 (Gagnon-Bartsch et al. 2013) use bi-cross validation (BCV), whereas dSVA (Lee et al. 2017) and IRW-SVA (Leek & Storey 2008) use parallel analysis (PA) to estimate K. We leave out the case when $\boldsymbol{C}$ is ignored in all plots below because its false discovery rate was greater than 0.9 in each simulation scenario.

Results of Simulation (1)

The results for Analysis (A) are given in 3.4, where we exclude results for IRW-SVA and when $\boldsymbol{C}$ was ignored because their median FDPs were greater than BCV-RUV-4's. CBCV consistently estimated there were 10 factors, whereas bi-cross validation's and parallel analysis' median estimates were 61 and 49, respectively. Existing methods generally showed a reduction in power because the large estimates of K reduced the number of residual degrees of freedom. CBCV underestimated the number of factors because $\tilde{\boldsymbol{C}}$ had a small effect on gene expression, which is exactly the reason why ignoring these factors had little to no impact on the results.

All existing methods had median false discovery proportions greater than 0.5 when using Analysis (B), which is not surprising because ignoring correlation in residuals tends to inflate P values. Lastly, existing methods' false discovery rate control was uniformly worse than that shown in Figure 3.2 when using Analyses (C)-(E) because these analyses consistently estimated K to be less than 10.

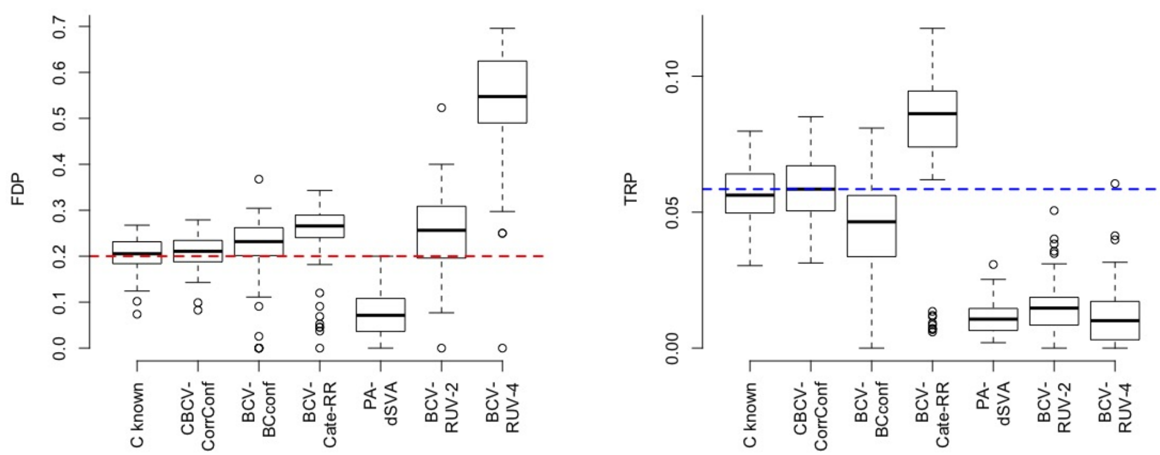


Figure 3.4: Results using Analysis (A) on Simulation (1). The false discovery proportion (FDP) and true recovery proportion (TRP) among genes with a q-value no greater than 0.2 in 100 simulated datasets, where TRP $= \frac{\text{\# of discoveries}}{\text{Total \# of genes with } \beta_g \neq 0}$. "C known" used the design matrix $\left(\boldsymbol{X}\,\boldsymbol{Z}\,\boldsymbol{C}\,\tilde{\boldsymbol{C}}\right)$. The dashed red and blue lines are FDP $= 0.2$ and TRP $=$ {Median TRP for CBCV-CorrConf}, where CBCV-CorrConf's power to identify differentially expressed genes at this q-value was 35% greater than that of BCV-BCconf's.

Results of Simulation (2)

The results for Analysis (A) are given in Figure 3.5, where we exclude results for IRW-SVA and when $\boldsymbol{C}$ was ignored because their median FDPs were greater than BCV-RUV-4's. CBCV consistently and correctly estimated that there were 10 factors, whereas bi-cross validation's and parallel analysis' median estimates were 60.5 and 49, respectively.

The behavior of existing methods when we used Analyses (B)-(E) was nearly identical to that in Simulation (1).

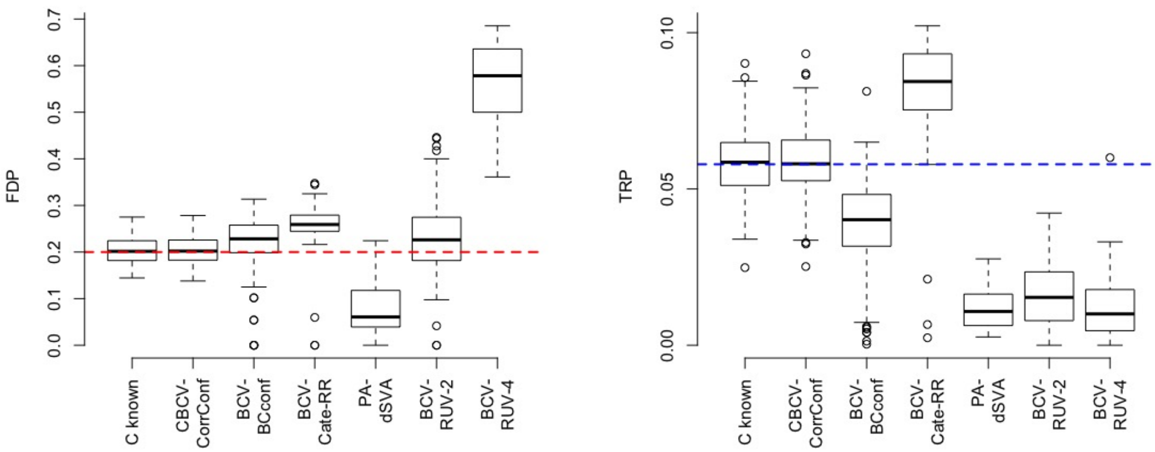


Figure 3.5: Results using Analysis (A) on Simulation (2). The false discovery proportion (FDP) and true recovery proportion (TRP) among genes with a q-value no greater than 0.2 in 100 simulated datasets, where TRP $= \frac{\text{\# of discoveries}}{\text{Total \# of genes with } \beta_g \neq 0}$. The dashed red and blue lines are FDP = 0.2 and TRP = Median TRP for CBCV-CorrConf, where CBCV-CorrConf's power to identify differentially expressed genes at this q-value was 53% greater than that of BCV-BCconf's.

Results of Simulation (3)

The results for Analysis (A) are given in Figure 3.6, where we exclude results for IRW-SVA and RUV-4 because their performance was similar to that in Simulations (1) and (2). CBCV consistently and correctly estimated that there were 10 factors, whereas bi-cross validation's and parallel analysis' median estimates were 62 and 49, respectively. Note that

$$\text{median}\,(\text{FDP for CBCV-CorrConf}) - \text{median}\,(\text{FDP when } \boldsymbol{C} \text{ is known}) = 0.002.$$

The behavior of existing methods when we used Analyses (B)-(E) was nearly identical to that in Simulation (1).

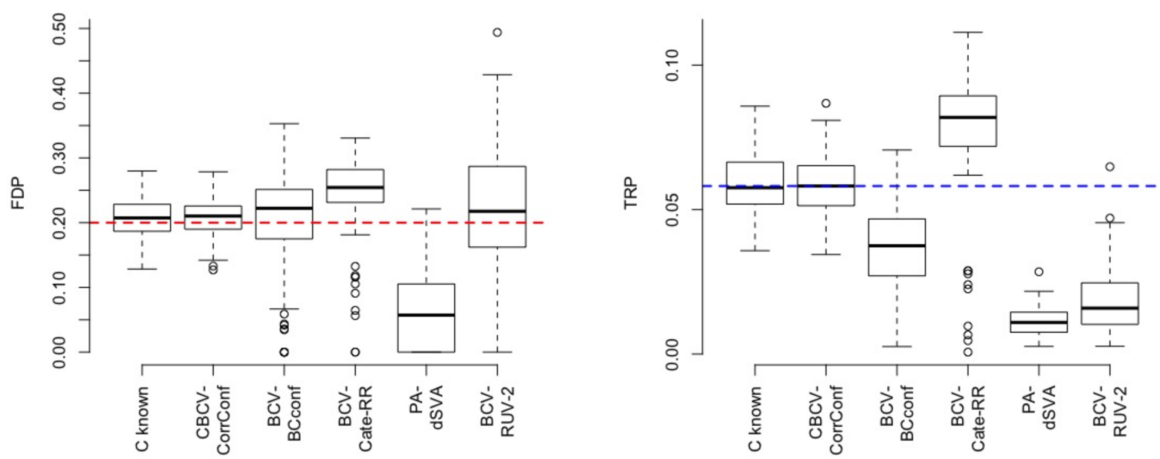


Figure 3.6: Results using Analysis (A) on Simulation (3). The false discovery proportion (FDP) and true recovery proportion (TRP) among genes with a q-value no greater than 0.2 in 100 simulated datasets, where TRP $= \frac{\#\text{ of discoveries}}{\text{Total }\#\text{ of genes with }\beta_g \neq 0}$. The dashed red and blue lines are FDP = 0.2 and TRP = Median TRP for CBCV-CorrConf, where CBCV-CorrConf's power to identify differentially expressed genes at this q-value was 66% greater than that of BCV-BCconf's.

3.8.4 The anomalous behavior of IRW-SVA

The anomalous behavior of IRW-SVA in Figure 3.2 is contingent on the size of $\boldsymbol{\Omega}$ and the eigenvalues $\lambda_1, \ldots, \lambda_K$. IRW-SVA circumvents estimating $\boldsymbol{\Omega}$ with the noisy design matrix $\hat{\boldsymbol{L}}$ by first identifying the genes with $\beta_g = 0$ and then estimating $\boldsymbol{C}$ with factor analysis on the reduced data matrix. If K is known, sufficiently many null genes are correctly identified, λ_K is sufficiently large and $n^{-1}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{V}_*\boldsymbol{C}$ is approximately a multiple of $n^{-1}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C}$, then IRW-SVA can accurately recover $\boldsymbol{C}$ when samples are correlated. In fact, if $\boldsymbol{B}=\boldsymbol{0}$ and λ_K/δ_*^2 is sufficiently larger than $\|\boldsymbol{V}_*\|_2$, one can use (3.38a) and (3.38b) in the proof of Theorem 3.10 to show that K-partial SVD of $\boldsymbol{Y}$ recovers $\boldsymbol{C}$ at approximately the rate

$$\|P_C - P_{\hat{C}}\|_F^2 = O_P\left\{n/\left(\lambda_K p\right) + \left(\lambda_K p\right)^{-1/2} + \lambda_K^{-2}\epsilon + \lambda_K^{-1}\alpha\right\}$$
$$\epsilon = \|n^{-1/2}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{V}_*\boldsymbol{Q}_C\|_2^2, \quad \alpha = \min_{c>0}\left(\|n^{-1}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{V}_*\boldsymbol{C} - cn^{-1}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C}\|_2\right).$$

This is not the case for BCconf, Cate-RR and dSVA, since in addition to using factor analysis to estimate $\boldsymbol{C}_\perp$, these methods use $\hat{\boldsymbol{L}}$ to estimate $\boldsymbol{\Omega}$, which can be inaccurate depending on the accuracy of $\hat{\boldsymbol{L}}$. However, when both λ_k and the k^{th} column of $\boldsymbol{\Omega}$ are moderate to large, IRW-SVA attributes a disproportionate amount of the variability in $\boldsymbol{C}$ as arising from the direct effect of $\boldsymbol{X}$ on $\boldsymbol{Y}$ compared to the other three methods. We refer the reader to Section 5.3 of Wang et al. (2017) for a more detailed discussion.

3.9 Additional results from Martino et al. (2013)

We first show how we derive the expression for $\mathrm{Cov}\left(\boldsymbol{y}_m\right)$ in (3.15). We avoid assuming a generative model for $\mathrm{Cov}\left(\boldsymbol{y}_m\right)$ by only making assumptions on the pairwise covariances, which averts potential biases in our estimate for $\boldsymbol{V}_*$, and therefore $\boldsymbol{C}$. First, the covariance between observations made on the same individual (or sample) should be at least as large as those made on different individuals (or samples). Second, the shared variance for twins at the same age should be as least as large as that at different ages. That is, for $a_1 \neq a_2$ and $t_1 \neq t_2$,

$$0 \leq \mathrm{Cov}\left(y_{m,t_1,a_1}, y_{m,t_2,a_2}\right) \leq \mathrm{Cov}\left(y_{m,t_1,a_1}, y_{m,t_1,a_2}\right) \leq \mathbb{V}\left(y_{m,t_1,a_i}\right) \quad (i=1,2) \tag{3.16a}$$
$$0 \leq \mathrm{Cov}\left(y_{m,t_1,a_1}, y_{m,t_2,a_2}\right) \leq \mathrm{Cov}\left(y_{m,t_1,a_i}, y_{m,t_2,a_i}\right) \leq \mathbb{V}\left(y_{m,t_1,a_i}\right) \quad (i=1,2). \tag{3.16b}$$

We can therefore write the covariance matrix for $\boldsymbol{y}_m$ as

$$\begin{aligned}\text{Cov}\left(\boldsymbol{y}_m\right) =& v_\alpha \mathbf{1}_4\mathbf{1}_4^{\mathrm{T}} + v_\eta\left(\mathbf{1}_2\mathbf{1}_2^{\mathrm{T}} \oplus \mathbf{1}_2\mathbf{1}_2^{\mathrm{T}}\right) + v_{\phi,0}\boldsymbol{a}_0\boldsymbol{a}_0^{\mathrm{T}} + v_{\phi,18}\boldsymbol{a}_{18}\boldsymbol{a}_{18}^{\mathrm{T}} + v_0 \,\text{diag}\left(\boldsymbol{a}_0\right)\\ &+ v_{18}\,\text{diag}\left(\boldsymbol{a}_{18}\right)\end{aligned}$$

where $\boldsymbol{a}_0^{\mathrm{T}} = (1,0,1,0)$, $\boldsymbol{a}_{18}^{\mathrm{T}} = (0,1,0,1)$ and

$$\begin{aligned}&v_\alpha = \text{Cov}\left(y_{m,t_1,0}, y_{m,t_2,18}\right)\\ &v_\alpha + v_\eta = \text{Cov}\left(y_{m,t_1,0}, y_{m,t_1,18}\right),\; v_\alpha + v_{\phi,a} = \text{Cov}\left(y_{m,t_1,a}, y_{m,t_2,a}\right) \quad (a=0,18)\\ &v_\alpha + v_\eta + v_{\phi,a} + v_a = \mathbb{V}\left(y_{m,t_1,a}\right) \quad (a=0,18).\end{aligned}$$

By (3.16), the variance multipliers also lie in a convex polytope that can be written in the form of Θ defined in (3.3), and are such that

$$v_\alpha \geq 0,\; v_\eta \geq 0,\; v_{\phi,a} \geq 0,\; v_\eta + v_a \geq 0,\; v_{\phi,a} + v_a \geq 0 \;\; (a=0,18).$$

Since BCconf, Cate-RR, dSVA and IRW-SVA are not designed for dependent data and given the complexity of the sample correlation structure for samples from infants with the same mother, we discuss two adjustments designed to alleviate potential biases in their estimates for K and $\boldsymbol{C}$. The first is to split the data matrix into a set of samples measured at birth and another set measured at 18 months, and subsequently rotate the two data matrices to nullify between-twin correlations. This should help to mitigate biases in bi-cross validation's estimates for K_0 and K_{18} (the number of latent factors at birth and 18 months), but leaves parallel analysis' estimates unchanged. While data splitting removes between-individual correlations, it effectively reduces the sample size by 50% when estimating $\boldsymbol{C}$ because we are forced to estimate the latent factors at birth and 18 months separately. Bi-cross validation estimated $K_0 = 3$, $K_{18} = 2$ and parallel analysis estimated $K_0 = 9$, $K_{18} = 7$. The results using this data splitting method were nearly identical as those reported in Table

3.3. Lastly, one could split the data by age and twin id into four data matrices, which would ostensibly eliminate all correlation between samples. However, since twin ids are arbitrary, estimates for K, $\boldsymbol{C}$ and subsequently $\boldsymbol{\beta}$ were heavily dependent on how twins were grouped, so we did not include comparisons with this data splitting technique.

We lastly provide an explanation as to why IRW-SVA's and CorrConf's results were similar when $K = 2$ was known (the reasoning for $K = 4$ is similar). For simplicity, we use the identifiability of $\boldsymbol{L}$ and $\boldsymbol{C}$ to assume $n^{-1}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C} = n^{-1}\hat{\boldsymbol{C}}^{\mathrm{T}}\hat{\boldsymbol{C}} = I_K$, where $\hat{\boldsymbol{C}}$ and $\hat{\boldsymbol{V}}$ were estimated with CorrConf. First, the P value for the null hypothesis that $\boldsymbol{C} \sim MN_{n\times K}\left(\boldsymbol{0}, I_n, I_K\right)$ was 0.31, indicating that $\boldsymbol{\Omega}$ was small. Therefore, since the marginal correlation between sex and methylation is relatively small, IRW-SVA is expected to do a reasonable job of identifying null genomic units (see Figure 2 in the Supplement of Wang et al. (2017)). Next, $\|1.34n^{-1}\hat{\boldsymbol{C}}^{\mathrm{T}}\hat{\boldsymbol{C}} - n^{-1}\hat{\boldsymbol{C}}^{\mathrm{T}}\hat{\boldsymbol{V}}\hat{\boldsymbol{C}}\|_2 = 0.05$ and $\|n^{-1/2}\hat{\boldsymbol{C}}^{\mathrm{T}}\hat{\boldsymbol{V}}\boldsymbol{Q}_{\hat{C}}\|_2^2 = 0.025$. This sensitivity analysis and the arguments made in Section 3.8.4 show that it is not unreasonable to expect IRW-SVA to be able to recover $\boldsymbol{C}$ in these data.

3.10 The identifiability of model parameters

3.10.1 The identifiability of $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_p$ and $\boldsymbol{L}$

For the remainder of the chapter, we define $[\boldsymbol{A}]_{i*} \in \mathbb{R}^m$ and $[\boldsymbol{A}]_{*j} \in \mathbb{R}^n$ to be the ith row and jth column of the matrix $\boldsymbol{A} \in \mathbb{R}^{n\times m}$, and define $|\boldsymbol{G}|$ to be the determinant of the matrix $\boldsymbol{G} \in \mathbb{R}^{n\times n}$.

In Section 3.10 we give the sufficient conditions to ensure that all parameters of interest in Model (3.2) are identifiable. We start by proving the identifiability of $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_p$ and $\boldsymbol{L}$.

***Proposition* 3.1** (Identifiability of $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_p$ and $\boldsymbol{L}$). *Let $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ be an observed, full rank matrix (where d is fixed) and suppose the matrices $\boldsymbol{B}_1, \ldots, \boldsymbol{B}_b \in \mathbb{R}^{n\times n}$ are observed, symmetric matrices. In addition, suppose the following hold for some constant $c > 1$:*

(i) $\|\boldsymbol{B}_j\|_2 \leq c$ for all $j \in [b]$.

(ii) Let $\boldsymbol{M} \in \mathbb{R}^{b\times b}$ *be such that* $[\boldsymbol{M}]_{ij} = n^{-1}\operatorname{tr}\left(\boldsymbol{B}_i\boldsymbol{B}_j\right)$ *for* $i, j \in [b]$*. Then* $\boldsymbol{M} \succeq c^{-1}I_b$*.*

(iii) For any matrix $\boldsymbol{D}$*, let* $\boldsymbol{D}_{(-i)}$ *be the sub-matrix of* $\boldsymbol{D}$ *with the* i*th row removed. Then*

$$\boldsymbol{L} \in \boldsymbol{\Theta}_L = \Big\{\boldsymbol{L} \in \mathbb{R}^{p\times K} : \textit{For all } g \in [p],\ \boldsymbol{L}_{(-g)} \textit{ contains two distinct submatrices of rank } K\Big\}$$

(iv) $\boldsymbol{C} \in \mathbb{R}^{n\times K}$ *is a random matrix such that* $\|\mathbb{V}\{\operatorname{vec}(\boldsymbol{C})\}\|_2 < \infty$ *and* $\mathbb{V}\{\operatorname{vec}(\boldsymbol{C})\} \succ \boldsymbol{0}$*.*

(v) $\boldsymbol{E} = \left[\boldsymbol{e}_1 \cdots \boldsymbol{e}_p\right]^{\mathrm{T}} \in \mathbb{R}^{p\times n}$ *is a random matrix that is independent of* $\boldsymbol{C}$ *such that* $\mathbb{E}\left(\boldsymbol{e}_g\right) = \boldsymbol{0}$*,* $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_p$ *are independent and*

$$\mathbb{V}\left(\boldsymbol{e}_g\right) = \boldsymbol{V}_g = \sum_{j=1}^{b} v_{g,j}\boldsymbol{B}_j \succ \boldsymbol{0} \quad (g = 1, \ldots, p).$$

Define $\boldsymbol{C}_\perp = \boldsymbol{Q}_X^{\mathrm{T}}\boldsymbol{C}$ *and* $\boldsymbol{W}_* = \boldsymbol{Q}_X^{\mathrm{T}}\left(p^{-1}\sum_{g=1}^{p}\boldsymbol{V}_g\right)\boldsymbol{Q}_X$*, where the columns of* $\boldsymbol{Q}_X$ *form an orthonormal basis for the null space of* $\boldsymbol{X}^{\mathrm{T}}$*. If we define*

$$\boldsymbol{Y} = \boldsymbol{\beta}\boldsymbol{X}^{\mathrm{T}} + \boldsymbol{L}\boldsymbol{C}^{\mathrm{T}} + \boldsymbol{E}, \tag{3.17}$$

then $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_p$*,* $\boldsymbol{L}\mathbb{E}\left(\boldsymbol{C}_\perp^{\mathrm{T}}\boldsymbol{C}_\perp\right)\boldsymbol{L}^{\mathrm{T}}$ *and* $\boldsymbol{L}\mathbb{E}\left(\boldsymbol{C}_\perp^{\mathrm{T}}\boldsymbol{W}_*^{-1}\boldsymbol{C}_\perp\right)\boldsymbol{L}^{\mathrm{T}}$ *are identifiable for all* n, p *suitably large.*

Proof. Define

$$\boldsymbol{Y}_2 = \boldsymbol{Y}\boldsymbol{Q}_X = \boldsymbol{L}\boldsymbol{C}_\perp^{\mathrm{T}} + \boldsymbol{E}_2$$

where $\boldsymbol{e}_{2_g}$, the gth row of $\boldsymbol{E}_2$, is such that $\mathbb{E}\left(\boldsymbol{e}_{2_g}\right) = \boldsymbol{0}$ and $\mathbb{V}\left(\boldsymbol{e}_{2_g}\right) = \tilde{\boldsymbol{V}}_g$, where $\tilde{\boldsymbol{V}}_g = \boldsymbol{Q}_X^{\mathrm{T}}\boldsymbol{V}_g\boldsymbol{Q}_X$. Since $\boldsymbol{Q}_X \in \mathbb{R}^{n\times(n-d)}$ has full column rank, $\mathbb{V}\{\operatorname{vec}(\boldsymbol{C}_\perp)\} \succ \boldsymbol{0}$ and

$\|\mathbb{V}\{\mathrm{vec}(\boldsymbol{C}_\perp)\}\|_2 < \infty$. Next, for $\boldsymbol{c}_i \in \mathbb{R}^K$ the ith row of $\boldsymbol{C}_\perp$, define

$$\boldsymbol{\Psi}_{ij} = \mathbb{E}\left(\boldsymbol{c}_i \boldsymbol{c}_j^{\mathrm{T}}\right) \in \mathbb{R}^{K\times K} \quad (i,j = 1,\dots,m).$$

where $m = n - d$. We first observe that for $\boldsymbol{y}_{2_g}$ the gth row of $\boldsymbol{Y}_2$,

$$\mathbb{E}\left(\boldsymbol{y}_{2_g} \boldsymbol{y}_{2_h}^{\mathrm{T}}\right) = \begin{cases} \boldsymbol{F}_{gg} + \tilde{\boldsymbol{V}}_g & \text{if } g = h \\ \boldsymbol{F}_{gh} & \text{if } g \neq h \end{cases}$$

where for $\boldsymbol{\ell}_g$ the gth row of $\boldsymbol{L}$,

$$\boldsymbol{F}_{gh} = \begin{bmatrix} \boldsymbol{\ell}_g^{\mathrm{T}} \boldsymbol{\Psi}_{11} \boldsymbol{\ell}_h & \cdots & \boldsymbol{\ell}_g^{\mathrm{T}} \boldsymbol{\Psi}_{1m} \boldsymbol{\ell}_h \\ \vdots & \ddots & \vdots \\ \boldsymbol{\ell}_g^{\mathrm{T}} \boldsymbol{\Psi}_{m1} \boldsymbol{\ell}_h & \cdots & \boldsymbol{\ell}_g^{\mathrm{T}} \boldsymbol{\Psi}_{mm} \boldsymbol{\ell}_h \end{bmatrix} \quad (g,h = 1,\dots,p).$$

Let $\boldsymbol{u} \in \mathbb{R}^m \setminus \{\boldsymbol{0}\}$. Then

$$\boldsymbol{u}^{\mathrm{T}} \boldsymbol{F}_{gh} \boldsymbol{u} = \boldsymbol{\ell}_g^{\mathrm{T}} (\boldsymbol{u} \otimes I_K)^{\mathrm{T}} \mathbb{E}\left\{\mathrm{vec}\left(\boldsymbol{C}_\perp^{\mathrm{T}}\right) \mathrm{vec}\left(\boldsymbol{C}_\perp^{\mathrm{T}}\right)^{\mathrm{T}}\right\} (\boldsymbol{u} \otimes I_K)\, \boldsymbol{\ell}_h \quad (g,h = 1,\dots,p)$$

where $(\boldsymbol{u} \otimes I_K)^{\mathrm{T}} \mathbb{E}\left\{\mathrm{vec}\left(\boldsymbol{C}_\perp^{\mathrm{T}}\right) \mathrm{vec}\left(\boldsymbol{C}_\perp^{\mathrm{T}}\right)^{\mathrm{T}}\right\} (\boldsymbol{u} \otimes I_K) \succ \boldsymbol{0}$ because $\mathbb{E}\left\{\mathrm{vec}\left(\boldsymbol{C}_\perp^{\mathrm{T}}\right) \mathrm{vec}\left(\boldsymbol{C}_\perp^{\mathrm{T}}\right)^{\mathrm{T}}\right\}$ is positive definite and $\boldsymbol{u} \otimes I_K$ has full column rank. Further, since $\boldsymbol{u}^{\mathrm{T}} \tilde{\boldsymbol{V}}_g \boldsymbol{u} > 0$ for all $g \in [p]$ and $\boldsymbol{L} \in \boldsymbol{\Theta}_L$,

$$\boldsymbol{L}\left[(\boldsymbol{u} \otimes I_K)^{\mathrm{T}} \mathbb{E}\left\{\mathrm{vec}\left(\boldsymbol{C}_\perp^{\mathrm{T}}\right) \mathrm{vec}\left(\boldsymbol{C}_\perp^{\mathrm{T}}\right)^{\mathrm{T}}\right\} (\boldsymbol{u} \otimes I_K)\right] \boldsymbol{L}^{\mathrm{T}}$$

is identifiable by Theorem 5.1 of Anderson & Rubin (1956), provided $p \geq 2K + 1$. In particular, if we let $\boldsymbol{u}$ be the canonical basis vectors in $\mathbb{R}^m$, then $\boldsymbol{L}\boldsymbol{\Psi}_{ii}\boldsymbol{L}^{\mathrm{T}}$ is identifiable for all $i = 1,\dots,m$. And lastly, if we let $\boldsymbol{u}$ have ones in the ith and jth position, for $i \neq j$, and zeros everywhere else, then $\boldsymbol{L}\boldsymbol{\Psi}_{ii}\boldsymbol{L}^{\mathrm{T}} + \boldsymbol{L}\boldsymbol{\Psi}_{jj}\boldsymbol{L}^{\mathrm{T}} + 2\boldsymbol{L}\boldsymbol{\Psi}_{ij}\boldsymbol{L}^{\mathrm{T}}$ is identifiable, meaning $\boldsymbol{L}\boldsymbol{\Psi}_{ij}\boldsymbol{L}^{\mathrm{T}}$, and therefore $\boldsymbol{F}_{gh}$, is identifiable for all $i, j \in [m]$ and $g, h \in [p]$.

To complete the proof, we first note that because $\boldsymbol{F}_{gh}$ is identifiable for all $g, h \in [p]$, $\tilde{\boldsymbol{V}}_g$, and therefore the variance multipliers $v_{g,j}$ $(g = 1, \ldots, p; j = 1, \ldots, b)$, are identifiable for all n large enough by Items (i) and (ii). The identifiability of the variance multipliers follows from the fact that Items (i) and (ii) imply that $\operatorname{vec}\left(\boldsymbol{Q}_X^{\mathrm{T}} \boldsymbol{B}_1 \boldsymbol{Q}_X\right), \ldots, \operatorname{vec}\left(\boldsymbol{Q}_X^{\mathrm{T}} \boldsymbol{B}_b \boldsymbol{Q}_X\right)$ are linearly independent for all n large enough. Since $\boldsymbol{W}_* = p^{-1} \sum_{g=1}^{p} \tilde{\boldsymbol{V}}_g$ is identifiable,

$$\boldsymbol{L} \tilde{\boldsymbol{\Psi}}_{ii} \boldsymbol{L}^{\mathrm{T}} \quad (i = 1, \ldots, m)$$

is identifiable, where for $\tilde{\boldsymbol{c}}_i \in \mathbb{R}^K$ the ith row of $\boldsymbol{W}_*^{-1/2} \boldsymbol{C}_\perp$,

$$\tilde{\boldsymbol{\Psi}}_{ii} = \mathbb{E}\left(\tilde{\boldsymbol{c}}_i \tilde{\boldsymbol{c}}_i^{\mathrm{T}}\right) \quad (i = 1, \ldots, m).$$

Therefore,

$$\begin{aligned}
\boldsymbol{L} \mathbb{E}\left(\boldsymbol{C}_\perp^{\mathrm{T}} \boldsymbol{W}_*^{-1} \boldsymbol{C}_\perp\right) \boldsymbol{L}^{\mathrm{T}} &= \boldsymbol{L}\left(\sum_{i=1}^{m} \tilde{\boldsymbol{\Psi}}_{ii}\right) \boldsymbol{L}^{\mathrm{T}} = \sum_{i=1}^{m} \boldsymbol{L} \tilde{\boldsymbol{\Psi}}_{ii} \boldsymbol{L}^{\mathrm{T}} \\
\boldsymbol{L} \mathbb{E}\left(\boldsymbol{C}_\perp^{\mathrm{T}} \boldsymbol{C}_\perp\right) \boldsymbol{L}^{\mathrm{T}} &= \boldsymbol{L}\left(\sum_{i=1}^{m} \boldsymbol{\Psi}_{ii}\right) \boldsymbol{L}^{\mathrm{T}} = \sum_{i=1}^{m} \boldsymbol{L} \boldsymbol{\Psi}_{ii} \boldsymbol{L}^{\mathrm{T}}
\end{aligned}$$

are identifiable, which completes the proof. □

***Remark* 3.8.** *$\boldsymbol{L} \mathbb{E}\left(\boldsymbol{C}_\perp^{\mathrm{T}} \boldsymbol{W}_*^{-1} \boldsymbol{C}_\perp\right) \boldsymbol{L}^{\mathrm{T}}$ and $\boldsymbol{L} \mathbb{E}\left(\boldsymbol{C}_\perp^{\mathrm{T}} \boldsymbol{C}_\perp\right) \boldsymbol{L}^{\mathrm{T}}$ are invariant of the choice of $\boldsymbol{Q}_X$.*

***Remark* 3.9.** *The parameter space $\boldsymbol{\Theta}_L$ is a classic way to identify the components of factor models when $\boldsymbol{V}_g = \sigma_g^2 I_n$ (Anderson & Rubin 1956, Wang et al. 2017).*

3.10.2 The identifiability of $\boldsymbol{\beta}$

***Proposition* 3.2** (Identifiability of $\boldsymbol{\beta}$)**.** *Let $c > 1$ be a constant and define $\lambda_1 \geq \cdots \geq \lambda_K > 0$ to be the non-zero eigenvalues of $np^{-1}\boldsymbol{L}\boldsymbol{\Psi}\boldsymbol{L}^{\mathrm{T}}$. In addition to the assumptions of Proposition 3.1, suppose the following hold:*

(i) Let $\boldsymbol{\ell}_g = [\boldsymbol{L}]_{g*}$. Then $\boldsymbol{\ell}_g^{\mathrm{T}}\boldsymbol{\Psi}\boldsymbol{\ell}_g \leq c$ for all $g \in [p]$, where $\boldsymbol{\Psi} = n^{-1}\mathbb{E}\left(\boldsymbol{C}_{\perp}^{\mathrm{T}}\boldsymbol{W}_*^{-1}\boldsymbol{C}_{\perp}\right)$.

(ii) $n/(\lambda_K p) \to 0$ as $n, p \to \infty$.

(iii) Let $\boldsymbol{\beta}$ be as defined in (3.17). Then $p^{-1}\sum_{g=1}^{p} I\left([\boldsymbol{\beta}]_{g*} \neq 0\right) = o\left(n^{-1}\lambda_K\right)$ as $n, p \to \infty$.

Then $\boldsymbol{\beta}$ is identifiable in Model (3.17) for all n, p suitably large.

Proof. For $\boldsymbol{A} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\mathbb{E}\left(\boldsymbol{C}\right)$,

$$\mathbb{E}\left\{\boldsymbol{Y}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\right\} = \boldsymbol{\beta} + \boldsymbol{L}\boldsymbol{A}^{\mathrm{T}} = \boldsymbol{\beta} + \left(\boldsymbol{L}\boldsymbol{\Psi}^{1/2}\boldsymbol{U}\right)\left(\boldsymbol{A}\boldsymbol{\Psi}^{-1/2}\boldsymbol{U}\right)^{\mathrm{T}}$$

where $\boldsymbol{U} \in \mathbb{R}^{K\times K}$ is a unitary matrix that ensures $\left(\boldsymbol{L}\boldsymbol{\Psi}^{1/2}\boldsymbol{U}\right)^{\mathrm{T}}\left(\boldsymbol{L}\boldsymbol{\Psi}^{1/2}\boldsymbol{U}\right)$ is diagonal with non-increasing elements. Therefore, it suffices to assume $\boldsymbol{\Psi} = I_K$ and $np^{-1}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L} = \mathrm{diag}\left(\lambda_1, \ldots, \lambda_K\right)$, which identifies $\boldsymbol{L}$ uniquely up to a $K \times K$ rotation matrix. Note that this implies $\boldsymbol{\ell}_g^{\mathrm{T}}\boldsymbol{\ell}_g \leq c$. Let $\boldsymbol{L}^{(1)}, \boldsymbol{\beta}^{(1)}$, $\boldsymbol{A}^{(1)}$ and $\boldsymbol{L}^{(2)}, \boldsymbol{\beta}^{(2)}$, $\boldsymbol{A}^{(2)}$ be such that $\boldsymbol{\beta}^{(1)} + \boldsymbol{L}^{(1)}\left\{\boldsymbol{A}^{(1)}\right\}^{\mathrm{T}} = \boldsymbol{\beta}^{(2)} + \boldsymbol{L}^{(2)}\left\{\boldsymbol{A}^{(2)}\right\}^{\mathrm{T}}$. Note that by the above analysis, $\boldsymbol{L}^{(1)} = \boldsymbol{L}^{(2)}\boldsymbol{Q}$ for some unitary matrix $\boldsymbol{Q} \in \mathbb{R}^{K\times K}$. To prove $\boldsymbol{\beta}$ is identifiable, it suffices to show that if $\mathcal{S} = \left\{g \in [p] : \boldsymbol{\beta}_g^{(1)} = \boldsymbol{\beta}_g^{(2)} = \boldsymbol{0}\right\}$, then $\boldsymbol{L}_{\mathcal{S}}^{(1)} \in \mathbb{R}^{|\mathcal{S}|\times K}$ has full column rank, where $\boldsymbol{L}_{\mathcal{S}}^{(1)}$ is the sub-matrix of $\boldsymbol{L}^{(1)}$ whose rows lie in $\mathcal{S}$. We see that

$$n\left(p\lambda_K\right)^{-1}\left\{\boldsymbol{L}_{\mathcal{S}}^{(1)}\right\}^{\mathrm{T}}\boldsymbol{L}_{\mathcal{S}}^{(1)} = n\left(p\lambda_K\right)^{-1}\left\{\boldsymbol{L}^{(1)}\right\}^{\mathrm{T}}\boldsymbol{L}^{(1)} - n\left(p\lambda_K\right)^{-1}\sum_{g\in\mathcal{S}^c}\boldsymbol{\ell}_g^{(1)}\left\{\boldsymbol{\ell}_g^{(1)}\right\}^{\mathrm{T}}$$

where $n\left(p\lambda_K\right)^{-1}\left\{\boldsymbol{L}^{(1)}\right\}^{\mathrm{T}}\boldsymbol{L}^{(1)} \succeq I_K$, and for any $r, t \in [K]$ and $s^{(i)} = p^{-1}\sum_{g=1}^{p} I\left\{\boldsymbol{\beta}_g^{(i)} \neq 0\right\}$

for $i = 1, 2$,

$$n\left(p\lambda_K\right)^{-1}\sum_{g\in\mathcal{S}^c}\left|\left[\boldsymbol{\ell}_g^{(1)}\right]_r\left[\boldsymbol{\ell}_g^{(1)}\right]_t\right| \leq nc\lambda_K^{-1}\left\{s^{(1)}+s^{(2)}\right\} = o\left(1\right)$$

as $n, p \to \infty$. $\square$

3.11 Assumptions and a restatement of all theory

3.11.1 Assumptions

We present the assumptions and theory stated in Section 3.4 for the reader's convenience. Recall that for all theoretical results, we assume the observed data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_p$ were generated according to model (3.2), and that $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_p$ are independent. We assume $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ $\boldsymbol{B}_1, \ldots, \boldsymbol{B}_b \in \mathbb{R}^{n\times n}$ are non-random, observed matrices, and that $d, b, K_{\max} = O(1)$.

***Assumption* 3.4.** *Let $c_1 > 1$ be a constant not dependent on n or p. Assume:*

(a) $\boldsymbol{V}_g \succeq c_1^{-1} I_n$, $|v_{g,j}| \leq c_1$, $\boldsymbol{B}_j = \boldsymbol{B}_j^{\mathrm{T}}$ *and* $\|\boldsymbol{B}_j\|_2 \leq c_1$ *for all* $g \in [p]$ *and* $j \in [b]$.

(b) Define $\boldsymbol{M} \in \mathbb{R}^{b\times b}$ *to be* $[\boldsymbol{M}]_{ij} = n^{-1}\operatorname{tr}\left(\boldsymbol{B}_i\boldsymbol{B}_j\right)$ *for all* $i, j \in [b]$. *Then* $\boldsymbol{M} \succeq c_1^{-1} I_b$.

(c) $n^{-1}\lim_{n\to\infty}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = \boldsymbol{\Sigma}_X \succ \boldsymbol{0}$.

(d) $\boldsymbol{C} \in \mathbb{R}^{n\times K}$ *is a random matrix such that* $K \leq K_{\max}$ *and*

(i) $\mathbb{V}\left\{\operatorname{vec}\left(\boldsymbol{C}\right)\right\} \succ \boldsymbol{0}$ *and* $\|\mathbb{V}\left\{\operatorname{vec}\left(\boldsymbol{C}\right)\right\}\|_2 < \infty$.

(ii) For $\boldsymbol{C}_\perp = \boldsymbol{Q}_X^{\mathrm{T}}\boldsymbol{C}$ *and* $\boldsymbol{W}_* = \boldsymbol{Q}_X^{\mathrm{T}}\left(p^{-1}\sum_{g=1}^{p}\boldsymbol{V}_g\right)\boldsymbol{Q}_X$, *define* $\boldsymbol{\Psi} = n^{-1}\mathbb{E}\left(\boldsymbol{C}_\perp^{\mathrm{T}}\boldsymbol{W}_*^{-1}\boldsymbol{C}_\perp\right)$. *Then* $\|\boldsymbol{\Psi}\|_2, \|\boldsymbol{\Psi}^{-1}\|_2 \leq c_1$ *and* $\|n^{-1}\boldsymbol{C}_\perp^{\mathrm{T}}\boldsymbol{W}_*^{-1}\boldsymbol{C}_\perp - \boldsymbol{\Psi}\|_2 = o_P(1)$ *as* $n, p \to \infty$.

(e) $\boldsymbol{L} \in \Theta_L$ *for* $\boldsymbol{\Theta}_L$ *defined in the statement of Proposition 3.1, and the matrix* $np^{-1}\boldsymbol{L}\boldsymbol{\Psi}\boldsymbol{L}^{\mathrm{T}}$ *has* $K > 0$ *non-zero eigenvalues* $\lambda_1 > \cdots > \lambda_K > 0$ *with* $\lambda_1/\lambda_K \leq c_1$. *Further*

(i) $\lambda_k \in \left[c_1^{-1}, nc_1\right]$ and $\left(\lambda_k - \lambda_{k+1}\right)/\lambda_k \geq c_1^{-1}$ for all $k \in [K]$, where $\lambda_{K+1} = 0$.

(ii) $\boldsymbol{\ell}_g^{\mathrm{T}} \boldsymbol{\Psi} \boldsymbol{\ell}_g \leq c_1$ for all $g \in [p]$.

(f) One of the following holds:

(i) $\lambda_K \geq c_1^{-1} n$ for all n, p.

(ii) $\lambda_K \to \infty$ but $\lambda_K / n \to 0$ as $n, p \to \infty$.

(iii) $\lambda_K \leq c_1$ for all n, p.

(g) p is a non-decreasing function of n such that n/p, $n^{3/2}/\left(p\lambda_K\right) \to 0$ as $n \to \infty$.

***Remark* 3.10.** *The condition that $\boldsymbol{L} \in \boldsymbol{\Theta}_L$ is the same as that assumed in Wang et al. (2017). Note that $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_p$ and $\boldsymbol{L}\boldsymbol{\Psi}\boldsymbol{L}^{\mathrm{T}}$ are identifiable under this assumption for all n, p large enough by Proposition 3.1.*

***Assumption* 3.5.** *Let $c_2 > 1$ be a constant not dependent on n or p. The estimates $\hat{\delta}^2$ and $\hat{\boldsymbol{\tau}}$ from step (2) in Algorithm 3.2 are such that $\hat{\delta}^2 \hat{\boldsymbol{\tau}}$ lies in the convex set Θ_*, where*

$$\Theta_* = \Theta \cap \left\{ \boldsymbol{x} \in \boldsymbol{R}^b : \|\boldsymbol{x}\|_2 \leq b c_2 c_1, \sum_{j=1}^{b} [\boldsymbol{x}]_j \boldsymbol{B}_j - \left(c_2 c_1\right)^{-1} I_n \succ \boldsymbol{0} \right\}$$

where c_1 was defined in Assumption 3.4.

***Assumption* 3.6.** *Let $c_3 > 0$ be a constant not dependent on n or p. Assume:*

(a) $p^{-1} \sum_{g=1}^{p} I\left(\boldsymbol{\beta}_g \neq 0\right) = o\left(n^{-3/2}\lambda_K\right)$ as $n, p \to \infty$.

(b) $\|\boldsymbol{\beta}_g\|_2 \leq c_3$ for all $g \in [p]$.

(c) Define $\boldsymbol{S} = \left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{V}_*^{-1} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{V}_*^{-1} \boldsymbol{C}$. Then $\|\boldsymbol{S}\boldsymbol{\Psi}^{-1}\boldsymbol{S}^{\mathrm{T}}\|_2 = O_P(1)$ as $n, p \to \infty$.

***Remark* 3.11.** *$\boldsymbol{\beta}_g$ is identifiable for all $g \in [p]$ under Assumptions 3.4 and 3.6(a) by Proposition 3.2.*

3.11.2 Restatement of all theory

***Theorem* 3.4** (Restatement of Theorem 3.1)**.** *Suppose Assumptions 3.4 and 3.5 hold. Suppose further that we stop on step (1) of the second iteration of Algorithm 3.2 for each* $k = 1, \ldots, K_{\max}$*, where* $K_{\max} \geq K$*. Then the estimates for* δ_*^2*,* $\boldsymbol{\tau}_*$ *and* $\boldsymbol{C}_\perp$ *from Algorithm 3.2 are such that*

$$
\begin{aligned}
|\hat{\delta}^2 - \delta_*^2|, \|\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}_*\|_2 = O_P\left(n^{-1}\right), \quad k \geq K \\
\|P_{\boldsymbol{C}_\perp} - P_{\hat{\boldsymbol{C}}_\perp}\|_F^2 = O_P\left\{n/\left(\lambda_K p\right) + (p\lambda_K)^{-1/2} + (n\lambda_K)^{-1}\right\}, \quad k = K \geq 1 \\
\|P_{\boldsymbol{C}_\perp} - P_{\hat{\boldsymbol{C}}_\perp} P_{\boldsymbol{C}_\perp}\|_F^2 = O_P\left\{n/\left(\lambda_K p\right) + (p\lambda_K)^{-1/2} + (n\lambda_K)^{-1}\right\}, \quad k \geq K \geq 1
\end{aligned}
$$

where $\hat{\delta}^2$ *and* $\hat{\boldsymbol{\tau}}$ *depend on* k*. Further,* $\|P_{\boldsymbol{C}_\perp} - P_{\hat{\boldsymbol{C}}_\perp}\|_F^2$*,* $\|P_{\boldsymbol{C}_\perp} - P_{\hat{\boldsymbol{C}}_\perp} P_{\boldsymbol{C}_\perp}\|_F^2$ *and the estimates* $\hat{\delta}^2, \hat{\boldsymbol{\tau}}$ *are invariant of the choice of* $\boldsymbol{Q}_X$*.*

***Theorem* 3.5** (Restatement of Theorem 3.2)**.** *Let* $\hat{K}$ *be as defined in* (3.10)*. Suppose Assumptions 3.4 and 3.5 hold,* $\lambda_K \geq \delta_*^2 + \epsilon$ *for some constant* $\epsilon \in (0,1)$ *and we sample* $\boldsymbol{Q}_X$ *uniformly over the space of all orthonormal bases for* $\boldsymbol{X}^\perp$*. Then* $\lim_{n,p\to\infty} \mathbb{P}\left(\hat{K} = K\right) = 1$*.*

***Remark* 3.12.** *We assume* $d = 0$ *in the statement of Theorem 3.5 for notational convenience to avoid carrying the subscript "2". When* $d > 0$*, we simply replace* $\boldsymbol{Y}$ *with* $\boldsymbol{Y}_2$*,* $\boldsymbol{C}$ *with* $\boldsymbol{C}_\perp$ *and* $\boldsymbol{V}_*$ *with* $\boldsymbol{W}_*$*.*

***Remark* 3.13.** *We prove that choosing* $\boldsymbol{Q}_X$ *uniformly at random from the set of all orthonormal bases for* $X^\perp$ *guarantees* ϕ*, the largest leverage score of* $\boldsymbol{W}_*^{-1/2}\boldsymbol{C}_\perp$*, is* $o_P(1)$ *as* $n \to \infty$ *in Section 3.11.3.*

The fact that we tend to recover all factors with eigenvalues λ_k, greater than δ_*^2 is consistent with Owen & Wang (2016), which showed that bi-cross validation cannot recover factors whose eigenvalues λ_k are smaller than δ_*^2 when samples are independent (Owen & Wang 2016, Section A.1). When there exist eigenvalues λ_k that are strictly smaller than δ_*^2 for all n and do not increase with the sample size, their corresponding effects $[\boldsymbol{L}]_{*k}$ must be such that $[\boldsymbol{L}]_{*k}^{\mathrm{T}}[\boldsymbol{L}]_{*k} = O(1/n)$. This means that only $O(1/n)$ genomic-units g will have $[\boldsymbol{L}]_{gk} \neq 0$ $(g = 1, \ldots, p)$. As the sample size increases, this implies this factor will become less and less relevant.

The limit of detection for parallel analysis is larger than that of CBCV. In fact, it has been shown that parallel analysis generally fails to recover factors with moderate to small eigenvalues in the presence of factors with larger eigenvalues (Dobriban 2017, Section 3.1). This is not true for CBCV. For example, when we simulated data according to Section 3.5.1 and (3.14) such that $\lambda_1 = \ldots = \lambda_5 = n$ but kept $\lambda_6, \ldots, \lambda_{10}$ as they were in Table 3.2, parallel analysis consistently estimate K to be 5, whereas CBCV consistently estimated K to be 10.

Theorem 3.6 (Restatement of Theorem 3.3)**.** *Suppose Assumptions 3.4, 3.5 and 3.6 hold, we estimate $\boldsymbol{C}$ according to* (3.12) *and $\boldsymbol{V}_g$ and $\boldsymbol{\beta}_g$ via restricted maximum likelihood (REML) and generalized least squares (GLS) using the design matrix $\begin{bmatrix}\boldsymbol{X}\,\hat{\boldsymbol{C}}\end{bmatrix}$. If the REML estimate $\hat{\boldsymbol{v}}_g = \left(\hat{v}_{g,1} \cdots \hat{v}_{g,b}\right)^{\mathrm{T}}$ is estimated on the parameter space Θ_*, the following asymptotic relations hold for the REMLS and GLS estimates $\hat{\boldsymbol{\beta}}_g$ and $\hat{\boldsymbol{V}}_g = \sum_{j=1}^{b} \hat{v}_{g,j}\boldsymbol{B}_j$:*

$$\|\hat{\boldsymbol{V}}_g - \boldsymbol{V}_g\|_2 = o_P(1) \tag{3.18}$$

$$\left(n^{-1}\hat{\boldsymbol{M}}_g\right)^{-1/2}\left(\hat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_g\right) \stackrel{\mathcal{D}}{=} \boldsymbol{Z} + o_P(1) \tag{3.19}$$

where

$$\begin{aligned}
&\boldsymbol{Z} \sim N\left(0, I_d\right)\\
&\hat{\boldsymbol{M}}_g = \left(n^{-1}\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{V}}_g^{-1}\boldsymbol{X}\right)^{-1} + \hat{\boldsymbol{\Omega}}_g\left(n^{-1}\hat{\boldsymbol{C}}_{\perp}^{\mathrm{T}}\hat{\boldsymbol{W}}_g^{-1}\hat{\boldsymbol{C}}_{\perp}\right)^{-1}\hat{\boldsymbol{\Omega}}_g^{\mathrm{T}}\\
&\hat{\boldsymbol{\Omega}}_g = \left(\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{V}}_g^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{V}}_g^{-1}\hat{\boldsymbol{C}}.
\end{aligned}$$

Further, $\|\hat{\boldsymbol{M}}_g\boldsymbol{M}_g^{-1} - I_d\|_2 = o_P(1)$, where $n^{-1}\boldsymbol{M}_g$ is the variance of the GLS estimate for $\boldsymbol{\beta}_g$ when $\boldsymbol{C}$ and $\boldsymbol{V}_g$ are known.

3.11.3 Sampling $\boldsymbol{Q}_X$ uniformly at random ensures leverage scores are well-behaved.

Here we state and prove a simple lemma describing sufficient conditions under which ϕ, the maximum leverage score of $\boldsymbol{W}_*^{-1/2}\boldsymbol{C}_\perp$, is $o_P(1)$ as $n \to \infty$. This is used in the proof of Theorem 3.5.

***Lemma* 3.1.** *Let $\boldsymbol{G} \in \mathbb{R}^{n\times r}$ be a random matrix and let $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ be a unitary matrix, independent of $\boldsymbol{G}$, that is sampled uniformly from the space of all unitary matrices in $\mathbb{R}^{n\times n}$. Define ϕ_n to be the maximum leverage score of $\boldsymbol{MG} \in \mathbb{R}^{n\times r}$. Then if $r = O(1)$, for all $\epsilon, \xi > 0$, there exists an $N > 0$ that does not depend on $\boldsymbol{G}$ such that $\mathbb{P}\left(\phi_n \geq \epsilon\right) < \xi$ for all $n \geq N$.*

Proof. It suffices to assume $\boldsymbol{G}$ is full rank with orthonormal columns. Since $r = O(1)$, it suffices to assume $r = 1$. Let $\tilde{g}_i = [\boldsymbol{M}]_{i*}^{\mathrm{T}}\boldsymbol{G}$. Then

$$\tilde{g}_i \sim \|\boldsymbol{Z}_i\|_2^{-1}\boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{G}, \quad \boldsymbol{Z}_i \sim N_n\left(\mathbf{0}, I_n\right) \quad (i = 1,\ldots,n)$$

and

$$\tilde{g}_i^2 \mid \boldsymbol{G} \sim n^{-1}w_i^{-1}z_i, \quad w_i \sim n^{-1}\chi_n^2, \quad z_i \sim \chi_1^2 \quad (i = 1,\ldots,n). \tag{3.20}$$

Then for any $\epsilon > 0$,

$$\mathbb{P}\left(\phi_n \geq \epsilon \mid \boldsymbol{G}\right) \leq \mathbb{P}\left(\sum_{i=1}^{n}\tilde{g}_i^4 \geq \epsilon^2 \mid \boldsymbol{G}\right) \leq c\epsilon^{-2}n^{-1}$$

where the constant c is the product of the second moments of w_i^{-1} and z_i defined in (3.20), which exist and are bounded from above for all n suitably large. □

***Remark* 3.14.** *The constant $c\epsilon^{-2}$ in the proof of Lemma 3.1 does not depend on the random* *matrix $\boldsymbol{G}$ or the number of columns r, since it is assumed that $r = O(1)$ as $n \to \infty$. Further*

$$\boldsymbol{W}_*^{-1/2}\boldsymbol{C}_\perp = \boldsymbol{M}^{\mathrm{T}}\left(\boldsymbol{Q}_*^{\mathrm{T}}\boldsymbol{V}_*\boldsymbol{Q}_*\right)^{-1/2}\boldsymbol{Q}_*^{\mathrm{T}}\boldsymbol{C}$$

where $\boldsymbol{M}$ and $\boldsymbol{Q}_*$ are defined in the statement of Theorem 3.5.

3.11.4 Proof strategy

We prove Theorems 3.4, 3.5 and 3.6 by first conditioning on $\boldsymbol{C}$ and proving analogous results when $\boldsymbol{C}$ is treated as a non-random matrix in Section 3.13. We then show that the assumptions necessary to prove these analogous theorems hold with probability tending to 1 as $n, p \to \infty$, which we show is sufficient to prove all theory in Section 3.14. We state Assumptions 3.7, 3.8 and 3.9 and Theorems 3.7, 3.8 and 3.9 under the assumption that $\boldsymbol{C}$ is a non-random matrix in Section 3.12, which are analogous versions of the assumptions and theory presented in Sections 3.11.1 and 3.11.2.

3.12 Assumptions and theory when $\boldsymbol{C}$ is a non-random matrix

3.12.1 Technical assumptions when $\boldsymbol{C}$ is a non-random matrix

***Assumption* 3.7.** *Let c_1 be as defined in Assumption 3.4 and define $\epsilon_1 = c^{-1}/\left(4 - c^{-1}\right)$. Assume:*

(a) Item (a) of Assumption 3.4 holds.

(b) Item (b) of Assumption 3.4 holds.

(c) Item (c) of Assumption 3.4 holds.

(d) $\boldsymbol{C} \in \mathbb{R}^{n \times K}$ is a non-random, unknown matrix.

(e) *Let* $\boldsymbol{W}_* = \boldsymbol{Q}_X^{\mathrm{T}} \boldsymbol{V}_* \boldsymbol{Q}_X$ *and let* $\lambda_1, \ldots, \lambda_K$ *be as defined in Assumption 3.4. The matrix*

$$\mathcal{I} = \boldsymbol{W}_*^{-1/2} \boldsymbol{C}_\perp \left(p^{-1} \boldsymbol{L}^{\mathrm{T}} \boldsymbol{L}\right) \boldsymbol{C}_\perp^{\mathrm{T}} \boldsymbol{W}_*^{-1/2} \tag{3.21}$$

has $K > 0$ *non-zero eigenvalues* $\gamma_1 > \cdots > \gamma_K > 0$ *with* $|\gamma_k/\lambda_k - 1| \leq \epsilon_1$ *for all* $k \in [K]$. *Further,* $\boldsymbol{\ell}_g^{\mathrm{T}} \left(n^{-1} \boldsymbol{C}_\perp^{\mathrm{T}} \boldsymbol{W}_*^{-1} \boldsymbol{C}_\perp\right) \boldsymbol{\ell}_g \leq c_1 + \epsilon_1$ *for all* $g \in [p]$.

(f) *Item (f) of Assumption 3.4 holds.*

(g) *Item (g) of Assumption 3.4 holds.*

***Remark* 3.15.** *Under (d),* $\mathbb{E}(\boldsymbol{Y}_2) = \boldsymbol{L}\boldsymbol{C}_\perp^{\mathrm{T}}$, $\mathcal{I}$ *and* $\boldsymbol{\ell}_g^{\mathrm{T}} \left(n^{-1} \boldsymbol{C}_\perp^{\mathrm{T}} \boldsymbol{W}_*^{-1} \boldsymbol{C}_\perp\right) \boldsymbol{\ell}_g$ *are identifiable. Under (e),* $\gamma_1/\gamma_K \leq c_1 \frac{1+\epsilon_1}{1-\epsilon_1}$, $\gamma_k \in \left[c_1^{-1} \frac{1-\epsilon_1}{1+\epsilon_1}, \frac{1+\epsilon_1}{1-\epsilon_1} c_1 n\right]$, $\left(\gamma_k - \gamma_{k+1}\right)/\gamma_k \geq (2c_1)^{-1}$ *for all* $k \in [K]$ *with* $\gamma_{k+1} = 0$.

***Remark* 3.16.** *The non-zero eigenvalues of* $\mathcal{I}$ *are also the eigenvalues of*

$$\left(np^{-1} \boldsymbol{L}^{\mathrm{T}} \boldsymbol{L}\right) \left(n^{-1} \boldsymbol{C}_\perp^{\mathrm{T}} \boldsymbol{W}_*^{-1} \boldsymbol{C}_\perp\right)$$

, where $\boldsymbol{C}_\perp^{\mathrm{T}} \boldsymbol{W}_*^{-1} \boldsymbol{C}_\perp$ *is invariant of the choice of* $\boldsymbol{Q}_X$.

***Assumption* 3.8.** *Assumption 3.5 holds.*

***Assumption* 3.9.** *Let* $c_3 > 0$ *and* $\lambda_1, \ldots, \lambda_K$ *be as defined in Assumption 3.4 and let* $M_3 > 0$ *be a large constant. Assume:*

(a) $p^{-1} \sum_{g=1}^{p} I\left(\boldsymbol{\beta}_g \neq 0\right) = o\left(n^{-3/2} \lambda_K\right)$ *as* $n, p \to \infty$.

(b) $\|\boldsymbol{\beta}_g\|_2 \leq c_3$ *for all* $g \in [p]$.

(c) *Let* $\boldsymbol{C} \in \mathbb{R}^{n \times K}$ *be any matrix such that* $\mathbb{E}(\boldsymbol{Y}_2) = \boldsymbol{L}\boldsymbol{C}_\perp^{\mathrm{T}}$ *for some* $\boldsymbol{L} \in \mathbb{R}^{p \times K}$. *Then for* $\boldsymbol{S} = \left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{V}_*^{-1} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{V}_*^{-1} \boldsymbol{C}$,

$$\|\boldsymbol{S} \left(n^{-1} \boldsymbol{C}_\perp^{\mathrm{T}} \boldsymbol{W}_*^{-1} \boldsymbol{C}_\perp\right)^{-1} \boldsymbol{S}^{\mathrm{T}}\|_2 \leq M_3.$$

3.12.2 The identifiability of $\boldsymbol{\beta}$, $\boldsymbol{C}$ and $\boldsymbol{L}$ in the model for $\boldsymbol{Y}$ conditional on $\boldsymbol{C}$

We first prove a simple proposition regarding the identifiability of $\operatorname{Im}(\boldsymbol{C})$ and $\boldsymbol{\beta}$ in the model for $\boldsymbol{Y} \mid \boldsymbol{C}$. This shows that one can develop a procedure to estimate both $\operatorname{Im}(\boldsymbol{C})$ and $\boldsymbol{\beta}$.

***Proposition* 3.3.** *Suppose Assumption 3.7 and Assumption 3.9(a) hold, and let $\boldsymbol{M}_g$ be the variance of the generalized least squares estimate for $\boldsymbol{\beta}_g$ when both $\boldsymbol{C}$ and $\boldsymbol{V}_g$ are known for $g \in [p]$. Then $\boldsymbol{\beta}$, $\boldsymbol{A} = \boldsymbol{S}\left(n^{-1}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{W}_*^{-1}\boldsymbol{C}\right)^{-1}\boldsymbol{S}^{\mathrm{T}}$ and $\boldsymbol{M}_g$ are identifiable for all $g \in [p]$ and invariant of the choice of $\boldsymbol{Q}_X$ for all $n \geq c_4$, where $c_4 > 0$ is a constant.*

Proof. The proof that $\boldsymbol{\beta}$ is identifiable is identical to that of Proposition 3.2 and is omitted. The identifiability of $\operatorname{Im}(\boldsymbol{C})$, $\boldsymbol{A}$ and $\boldsymbol{M}_g$ immediately follow. □

We next state and prove a simple corollary that shows one can identify the columns of $\boldsymbol{L}$ and $\boldsymbol{C}$ up to a sign flip by choosing a particular parametrization of $\boldsymbol{L}$ and $\boldsymbol{C}$. This parametrization does not change $\operatorname{Im}(\boldsymbol{C})$ or $\boldsymbol{\beta}$, and will be used in the statement of Lemma 3.2 in Section 3.12.3 below.

***Corollary* 3.1.** *Let $\mathcal{G} = \{\operatorname{diag}(a_1, \ldots, a_K) : a_1, \ldots, a_K \in \{-1, 1\}\}$, suppose Assumption 3.7 holds and define the parameter space*

$$\Lambda_{(0)} = \Big\{(\boldsymbol{L}, \boldsymbol{C}) \in \mathbb{R}^{p \times K} \times \mathbb{R}^{n \times K} : \mathbb{E}(\boldsymbol{Y}_2) = \boldsymbol{L}\boldsymbol{C}_{\perp}^{\mathrm{T}}, n^{-1}\boldsymbol{C}_{\perp}^{\mathrm{T}}\boldsymbol{W}_*^{-1}\boldsymbol{C}_{\perp} = I_K,$$
$$np^{-1}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L} = \operatorname{diag}(\gamma_1, \ldots, \gamma_K) \text{ for } \boldsymbol{C}_{\perp} = \boldsymbol{Q}_X^{\mathrm{T}}\boldsymbol{C}\Big\}.$$

Then $\Lambda_{(0)}$ is non-empty and if $\left\{\boldsymbol{L}^{(a)}, \boldsymbol{C}^{(a)}\right\}, \left\{\boldsymbol{L}^{(b)}, \boldsymbol{C}^{(b)}\right\} \in \Lambda_{(0)}$, then $\boldsymbol{L}^{(b)} = \boldsymbol{L}^{(a)}\Pi$ and $\boldsymbol{C}_{\perp}^{(b)} = \boldsymbol{C}_{\perp}^{(a)}\Pi$ for some $\Pi \in \mathcal{G}$. If Assumptions 3.7 and 3.9(a) hold, then

$$\Lambda_{(1)} = \Lambda_{(0)} \cap \left\{(\boldsymbol{L}, \boldsymbol{C}) \in \mathbb{R}^{p \times K} \times \mathbb{R}^{n \times K} : \mathbb{E}\left\{\boldsymbol{Y}\boldsymbol{V}_*^{-1}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_*^{-1}\boldsymbol{X}\right)^{-1}\right\} = \boldsymbol{\beta} + \boldsymbol{L}\boldsymbol{S}^{\mathrm{T}}\right.$$
$$\left.\text{for } \boldsymbol{S} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_*^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_*^{-1}\boldsymbol{C}\right\}$$

is non-empty for all $n \geq c_4$. Further, if $\left\{\boldsymbol{L}^{(a)}, \boldsymbol{C}^{(a)}\right\}, \left\{\boldsymbol{L}^{(b)}, \boldsymbol{C}^{(b)}\right\} \in \Lambda_{(1)}$, then $\boldsymbol{L}^{(b)} = \boldsymbol{L}^{(a)}\Pi$ and $\boldsymbol{C}^{(b)} = \boldsymbol{C}^{(a)}\Pi$ for some $\Pi \in \mathcal{G}$ for all $n \geq c_4$.

Proof. The proof of the first statement follows directly from the fact that the left and right singular vectors of a matrix are unique up to sign when the singular values are unique. The second statement is a direct consequence of Proposition 3.3. □

3.12.3 Statement of theory when $\boldsymbol{C}$ is a non-random matrix

***Theorem* 3.7** (Restatement of Theorem 3.4 when $\boldsymbol{C}$ is non-random)**.** *Suppose Assumptions 3.7 and 3.8 hold. Suppose further that we stop on step (1) of the second iteration of Algorithm 3.2 for each $k = 1, \ldots, K_{\max}$, where $K_{\max} \geq K$. Then the estimates for δ_*^2, $\boldsymbol{\tau}_*$ and $\boldsymbol{C}_\perp$ from Algorithm 3.2 are such that*

$$|\hat{\delta}^2 - \delta_*^2|, \|\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}_*\|_2 = O_P\left(n^{-1}\right), \quad k \geq K \tag{3.22a}$$

$$\|P_{\boldsymbol{C}_\perp} - P_{\hat{\boldsymbol{C}}_\perp}\|_F^2 = O_P\left\{n/\left(\gamma_K p\right) + \left(p\gamma_K\right)^{-1/2} + \left(n\gamma_K\right)^{-1}\right\}, \quad k = K \geq 1 \tag{3.22b}$$

$$\|P_{\boldsymbol{C}_\perp} - P_{\hat{\boldsymbol{C}}_\perp} P_{\boldsymbol{C}_\perp}\|_F^2 = O_P\left\{n/\left(\gamma_K p\right) + \left(p\gamma_K\right)^{-1/2} + \left(n\gamma_K\right)^{-1}\right\}, \quad k \geq K \geq 1 \tag{3.22c}$$

where $\hat{\delta}^2$ and $\hat{\boldsymbol{\tau}}$ depend on k. Further, the estimates $\hat{\delta}^2, \hat{\boldsymbol{\tau}}$ and the quantities $\|P_{\boldsymbol{C}_\perp} - P_{\hat{\boldsymbol{C}}_\perp}\|_F^2$, $\|P_{\boldsymbol{C}_\perp} - P_{\hat{\boldsymbol{C}}_\perp} P_{\boldsymbol{C}_\perp}\|_F^2$ are invariant of the choice of $\boldsymbol{Q}_X$.

***Theorem* 3.8** (Restatement of Theorem 3.5 when $\boldsymbol{C}$ is non-random)**.** *Let $\epsilon \in (0,1)$ and assume $\gamma_K \geq \delta_*^2 + \epsilon$. For any fold $f \in [F]$ and $i \in [n]$, let $\bar{\boldsymbol{y}}_{f,i}$, $\bar{\boldsymbol{C}}$, $\bar{\boldsymbol{c}}_i$, $\hat{\bar{\boldsymbol{C}}}$ and $\hat{\boldsymbol{L}}_{f,(-i)}$ be as defined in Algorighm 3.3. Suppose Assumptions 3.7 and 3.8 hold for $d = 0$, $K_{\max} \geq K$ and define $\hat{\bar{h}}_i$ to be the i^{th} leverage score of $\hat{\bar{\boldsymbol{C}}}$ (i.e. the i^{th} diagonal element of $P_{\hat{\bar{C}}}$). For each fold $f \in [F]$, suppose we modify the loss function in* (3.9) *to be*

$$\mathcal{L}_f(k) = \begin{cases} \sum\limits_{i=n}^{n} \|\bar{\boldsymbol{y}}_{f,i} - \hat{\boldsymbol{L}}_{f,(-i)}\hat{\bar{\boldsymbol{c}}}_i\|_2^2, & \max\limits_{i \in [n]} \eta \left(1 - \hat{\bar{h}}_i\right)^{-1} \leq m_n \\ \infty, & \textit{otherwise} \end{cases} \tag{3.23}$$

where $\eta > 0$ is a constant and

$$m_n = \min\left\{n^{1/2}, \frac{p/n}{\log(p/n)}, \frac{p^{1/2}/n^{1/4}}{\log\left(p^{1/2}/n^{1/4}\right)}\right\}.$$

If we define ϕ to be the maximum leverage score of $\boldsymbol{V}_*^{-1/2}\boldsymbol{C}$ and $\phi \leq \epsilon/(2K)$, then

$$\lim_{n,p\to\infty} \mathbb{P}\left(\hat{K} = K\right) = 1.$$

***Remark* 3.17.** *If $\boldsymbol{Q}_X$ and $\boldsymbol{Q}_*$ are matrices whose columns form an orthonormal basis for $\boldsymbol{X}^\perp$, then $\boldsymbol{Q}_X = \boldsymbol{Q}_*\boldsymbol{M}$ for some unitary matrix $\boldsymbol{M} \in \mathbb{R}^{(n-d)\times(n-d)}$. Therefore, if $\hat{\boldsymbol{C}}_{\perp,*}$ and $\hat{\boldsymbol{C}}_\perp$ are estimates from Algorithm 3.2 using $\boldsymbol{Q}_*$ and $\boldsymbol{Q}_X$, $\hat{\boldsymbol{C}}_\perp = \boldsymbol{M}^{\mathrm{T}}\hat{\boldsymbol{C}}_{\perp,*}$. Further, since $\hat{\delta}^2$ and $\hat{\boldsymbol{\tau}}$ do not depend on the choice of $\boldsymbol{M}$, then $\hat{\hat{\boldsymbol{C}}}_\perp = \boldsymbol{M}^{\mathrm{T}}\hat{\hat{\boldsymbol{C}}}_{\perp,*}$. Therefore, if $\boldsymbol{Q}_X$ is sampled uniformly over all orthonormal bases of $\boldsymbol{X}^\perp$, Lemma 3.1 and Remark 3.14 imply $\max_{i\in[n]}\left(1 - \hat{\hat{h}}_i\right)^{-1} = O_P(1)$ as $n \to \infty$. Therefore, there is no need to re-define the loss function in* (3.23) *if we assume $\boldsymbol{Q}_X$ is sampled uniformly over all orthonormal bases of $\boldsymbol{X}^\perp$. We choose to prove Theorem 3.8 using this modified loss function to illustrate the importance of the ratio p/n.*

We next state a lemma that was not stated in Section 3.4 but is useful in proving 3.9 below. It may also be useful in its own right, since it shows we can accurately estimate the part of $\boldsymbol{C}$ in the image of $\boldsymbol{X}$.

***Lemma* 3.2** (Accuracy of $\boldsymbol{\Omega}$)**.** *Suppose Assumptions 3.7, 3.8 and 3.9 hold, K is known and we estimate $\boldsymbol{C}_\perp$, $\boldsymbol{\tau}_*$ and δ_*^2 according to Algorithm 3.2. Define the estimate for $\boldsymbol{\Omega}$ to be*

$$\hat{\boldsymbol{\Omega}} = \boldsymbol{Y}_1^{\mathrm{T}}\hat{\boldsymbol{L}}\left\{\hat{\boldsymbol{L}}^{\mathrm{T}}\hat{\boldsymbol{L}} - p\hat{\delta}^2\left(\hat{\boldsymbol{C}}_\perp^{\mathrm{T}}\hat{\boldsymbol{W}}^{-1}\hat{\boldsymbol{C}}_\perp\right)^{-1}\right\}^{-1}. \tag{3.24}$$

Suppose that we use the IC3 identification criterion from Bai & Li (2012) to identify $\boldsymbol{L}$ and $\boldsymbol{C}$. That is, we assume *(i)* $(\boldsymbol{L}, \boldsymbol{C}) \in \Lambda_{(1)}$*, where* $\Lambda_{(1)}$ *is defined in the statement of Corollary 3.1.*

(ii) For any estimate $\hat{\boldsymbol{C}}_{\perp} \in \mathbb{R}^{(n-d)\times K}$*, the diagonal elements of* $\hat{\boldsymbol{C}}_{\perp}^{\mathrm{T}} \boldsymbol{C}_{\perp}$ *are non-negative.*

Then,

$$n^{1/2} \| \boldsymbol{\Omega} - \hat{\boldsymbol{\Omega}} \|_2 = o_P(1). \tag{3.25}$$

***Theorem* 3.9** (Restatement of Theorem 3.6 when $\boldsymbol{C}$ is non-random)**.** *Suppose Assumptions 3.7, 3.8 and 3.9 hold, we estimate* $\boldsymbol{C}$ *according to* (3.12)*, and* $\boldsymbol{V}_g$ *and* $\boldsymbol{\beta}_g$ *via restricted maximum likelihood (REML) and generalized least squares (GLS) using the design matrix* $\left[\boldsymbol{X}\, \hat{\boldsymbol{C}}\right]$*. If the REML estimate* $\hat{\boldsymbol{v}}_g = \left(\hat{v}_{g,1} \cdots \hat{v}_{g,b}\right)^{\mathrm{T}}$ *is estimated on the parameter space* Θ_**, the following asymptotic relations hold for the GLS estimate* $\hat{\boldsymbol{\beta}}_g$ *and* $\hat{\boldsymbol{V}}_g = \sum_{j=1}^{b} \hat{v}_{g,j} \boldsymbol{B}_j$*:*

$$\| \hat{\boldsymbol{V}}_g - \boldsymbol{V}_g \|_2 = o_P(1) \tag{3.26}$$

$$\left(n^{-1} \hat{\boldsymbol{M}}_g\right)^{-1/2} \left(\hat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_g\right) \overset{\mathcal{D}}{=} \boldsymbol{Z} + o_P(1) \tag{3.27}$$

where

$$\begin{aligned}
&\boldsymbol{Z} \sim N\left(0, I_d\right) \\
&\hat{\boldsymbol{M}}_g = \left(n^{-1} \boldsymbol{X}^{\mathrm{T}} \hat{\boldsymbol{V}}_g^{-1} \boldsymbol{X}\right)^{-1} + \hat{\boldsymbol{\Omega}}_g \left(n^{-1} \hat{\boldsymbol{C}}_{\perp}^{\mathrm{T}} \hat{\boldsymbol{W}}_g^{-1} \hat{\boldsymbol{C}}_{\perp}\right)^{-1} \hat{\boldsymbol{\Omega}}_g^{\mathrm{T}} \\
&\hat{\boldsymbol{\Omega}}_g = \left(\boldsymbol{X}^{\mathrm{T}} \hat{\boldsymbol{V}}_g^{-1} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\mathrm{T}} \hat{\boldsymbol{V}}_g^{-1} \hat{\boldsymbol{C}}.
\end{aligned}$$

Further, $\| \hat{\boldsymbol{M}}_g \boldsymbol{M}_g^{-1} - I_d \|_2 = o_P(1)$*, where* $n^{-1} \boldsymbol{M}_g$ *is the variance of the GLS estimate for* $\boldsymbol{\beta}_g$ *when* $\boldsymbol{C}$ *and* $\boldsymbol{V}_g$ *are known.*

***Remark* 3.18.** *The conditions under which Theorem 3.9 holds are invariant to the choice of parametrization of* $\boldsymbol{L}$ *and* $\boldsymbol{C}$*, and* $\hat{\boldsymbol{V}}_g$*,* $\hat{\boldsymbol{\beta}}_g$ *and* $\hat{\boldsymbol{M}}_g$ *depend only on* $\boldsymbol{X}$ *and* $\mathrm{Im}\left(\hat{\boldsymbol{C}}\right)$*.*

3.13 Proofs of all propositions, lemmas and theorems when $\boldsymbol{C}$ is a non-random matrix

Let $\|\boldsymbol{A}\|_2$ be the Euclidean norm if $\boldsymbol{A}$ is a vector or the largest singular value of $\boldsymbol{A}$ if $\boldsymbol{A}$ is a matrix. Unless otherwise stated, we use the notation that a matrix or vector $\boldsymbol{A} = O_P(a_n)$ if $\|\boldsymbol{A}\|_2 = O_P(a_n)$ for some sequence a_n. We also define

$$\tilde{\boldsymbol{B}}_j = \boldsymbol{Q}_X^{\mathrm{T}} \boldsymbol{B}_j \boldsymbol{Q}_X \quad (j = 1, \ldots, b).$$

We first prove a lemma that will be useful in the proof of Lemma 3.6 in Section 3.13.

***Lemma* 3.3.** *Suppose Assumption 3.7 holds and define*

$$f\left(\delta^2 \boldsymbol{\tau}\right) = -n^{-1} \log|\boldsymbol{Z}| + n^{-1} \operatorname{tr}\left(\boldsymbol{Z}\bar{\boldsymbol{V}}\right)$$

where $\boldsymbol{Z} = \boldsymbol{Z}\left(\delta^2\boldsymbol{\tau}\right) = \boldsymbol{V}\left(\delta^2\boldsymbol{\tau}\right)^{-1}$ *and* $\bar{\boldsymbol{V}} = p^{-1}\left(\boldsymbol{V}_1 + \cdots + \boldsymbol{V}_p\right)$. *Then for all* $\delta^2\boldsymbol{\tau} \in \Theta_*$,

$$f\left(\delta^2 \boldsymbol{\tau}\right) - f\left(\delta_*^2 \boldsymbol{\tau}_*\right) \geq c\|\delta^2\boldsymbol{\tau} - \delta_*^2\boldsymbol{\tau}_*\|_2^2$$

where $c > 0$ *is a constant not dependent on* n *or* p.

Proof. The function f is strictly convex in Θ_* when considered a function of the precision matrix $\boldsymbol{Z}$, where

$$\nabla_Z^2 f(\boldsymbol{Z}) = n^{-1}\boldsymbol{Z}^{-1} \otimes \boldsymbol{Z}^{-1} = n^{-1}\boldsymbol{V} \otimes \boldsymbol{V} \succeq n^{-1} 2\epsilon I_{n^2}$$

for some constant $\epsilon > 0$. Therefore,

$$f\left(\delta^2 \boldsymbol{\tau}\right) - f\left(\delta_*^2 \boldsymbol{\tau}_*\right) \geq n^{-1}\epsilon\|\boldsymbol{Z}\left(\delta^2\boldsymbol{\tau}\right) - \boldsymbol{Z}\left(\delta_*^2\boldsymbol{\tau}_*\right)\|_F^2.$$

by Taylor's Theorem. For $[\boldsymbol{M}]_{ij} = n^{-1} \operatorname{tr}\left(\boldsymbol{B}_i \boldsymbol{B}_j\right)$ $(i, j \in [b])$, we also have

$$\left(\delta^2 \boldsymbol{\tau} - \delta_*^2 \boldsymbol{\tau}_*\right)^{\mathrm{T}} \boldsymbol{M} \left(\delta^2 \boldsymbol{\tau} - \delta_*^2 \boldsymbol{\tau}_*\right) = n^{-1} \|\boldsymbol{V} - \boldsymbol{V}_*\|_F^2 = n^{-1} \|\boldsymbol{V}\left(\boldsymbol{Z} - \boldsymbol{Z}_*\right) \boldsymbol{V}_*\|_F^2$$
$$\leq c' n^{-1} \|\boldsymbol{Z} - \boldsymbol{Z}_*\|_F^2$$

where $c' > 0$ is some constant not dependent on n or p. The inequality follows from the fact that $\delta^2 \boldsymbol{\tau} \in \Theta_*$, meaning $\boldsymbol{V}\left(\delta^2 \boldsymbol{\tau}\right)$ has eigenvalues that are uniformly bounded above 0 and below ∞. That is, if $\boldsymbol{V} = \boldsymbol{U}_1 \boldsymbol{\Sigma}_1 \boldsymbol{U}_1^{\mathrm{T}}$ and $\boldsymbol{V}_* = \boldsymbol{U}_2 \boldsymbol{\Sigma}_2 \boldsymbol{U}_2^{\mathrm{T}}$,

$$\begin{aligned}
\|\boldsymbol{V}\left(\boldsymbol{Z} - \boldsymbol{Z}_*\right) \boldsymbol{V}_*\|_F^2 &= \operatorname{tr}\left\{\boldsymbol{\Sigma}_1^2 \boldsymbol{U}_1^{\mathrm{T}} \left(\boldsymbol{Z} - \boldsymbol{Z}_*\right) \boldsymbol{V}_*^2 \left(\boldsymbol{Z} - \boldsymbol{Z}_*\right) \boldsymbol{U}_1\right\} \leq c_1' \operatorname{tr}\left\{\left(\boldsymbol{Z} - \boldsymbol{Z}_*\right)^2 \boldsymbol{V}_*^2\right\} \\
&= c_1' \operatorname{tr}\left\{\boldsymbol{U}_2^{\mathrm{T}} \left(\boldsymbol{Z} - \boldsymbol{Z}_*\right)^2 \boldsymbol{U}_2 \boldsymbol{\Sigma}_2^2\right\} \leq c' \operatorname{tr}\left\{\boldsymbol{U}_2^{\mathrm{T}} \left(\boldsymbol{Z} - \boldsymbol{Z}_*\right)^2 \boldsymbol{U}_2\right\} \\
&= c' \|\boldsymbol{Z} - \boldsymbol{Z}_*\|_F^2.
\end{aligned}$$

Since we have assumed the eigenvalues of $\boldsymbol{M}$ are uniformly bounded above 0,

$$f\left(\delta^2 \boldsymbol{\tau}\right) - f\left(\delta_*^2 \boldsymbol{\tau}_*\right) \geq c \|\delta^2 \boldsymbol{\tau} - \delta_*^2 \boldsymbol{\tau}_*\|_2^2$$

for $c = a\epsilon / c'$ where $a > 0$ bounds the smallest eigenvalue of $\boldsymbol{M}$ from below. $\square$

We next prove a simple lemma that will be useful in the proofs of Theorems 3.7, 3.8 and 3.9, as well as Lemma 3.2.

Lemma 3.4. *Suppose Assumptions 3.7 and 3.8 hold. Then for any $\hat{\delta}^2, \hat{\boldsymbol{\tau}}$ such that $\hat{\delta}^2 \hat{\boldsymbol{\tau}} \in \Theta_*$, the following hold:*

(i) Let $\hat{\boldsymbol{W}} = \sum_{j=1}^{b} [\hat{\boldsymbol{\tau}}]_j \boldsymbol{Q}_X^{\mathrm{T}} \boldsymbol{B}_j \boldsymbol{Q}_X$ and let $\lambda_1 \geq \cdots \geq \lambda_K$ be the non-zero eigenvalues of

$$\hat{\boldsymbol{W}}^{-1/2} \boldsymbol{C}_{\perp} \left(p^{-1} \boldsymbol{L}^{\mathrm{T}} \boldsymbol{L}\right) \boldsymbol{C}_{\perp}^{\mathrm{T}} \hat{\boldsymbol{W}}^{-1/2}.$$

Then for some constant $c > 1$ that does not depend on n or p, $c^{-1}\gamma_K \leq \lambda_K \leq \lambda_1 \leq c\gamma_1$, where $\gamma_1, \ldots, \gamma_K$ *were defined in Assumption 3.7.*

(ii) $\boldsymbol{\ell}_g^{\mathrm{T}} \left(n^{-1} \boldsymbol{C}_\perp^{\mathrm{T}} \hat{\boldsymbol{W}}^{-1} \boldsymbol{C}_\perp \right)^{-1} \boldsymbol{\ell}_g \leq c'$ *(*$g = 1, \ldots, p$*) for some constant* $c' > 0$ *that does not depend on* n *or* p*.*

***Remark* 3.19.** *We abused notation in (i) to define the non-zero eigenvalues of*

$$\hat{\boldsymbol{W}}^{-1/2} \boldsymbol{C}_\perp \left(p^{-1} \boldsymbol{L}^{\mathrm{T}} \boldsymbol{L} \right) \boldsymbol{C}_\perp^{\mathrm{T}} \hat{\boldsymbol{W}}^{-1/2}.$$

These are not the same as the eigenvalues defined in Assumption 3.4 in Section 3.11.1.

Proof. By Assumptions 3.7 and 3.8, $\|\hat{\boldsymbol{W}}\|_2, \|\hat{\boldsymbol{W}}^{-1}\|_2$ are uniformly bounded above by a constant that does not depend on n or p. Let $c > 1$ be a constant. To prove Item (i), it suffices to show that for $\boldsymbol{D} \in \mathbb{R}^{(n-d) \times (n-d)}$ a diagonal matrix with entries bounded above c^{-1} and below c, and $\lambda_{\min}, \lambda_{\max}$ the minimum and maximum eigenvalues of

$$\boldsymbol{D}^{1/2} \boldsymbol{W}_*^{-1/2} \boldsymbol{C}_\perp \left(p^{-1} \boldsymbol{L}^{\mathrm{T}} \boldsymbol{L} \right) \boldsymbol{C}_\perp^{\mathrm{T}} \boldsymbol{W}_*^{-1/2} \boldsymbol{D}^{1/2},$$

respectively, that $c^{-1} \gamma_K \leq \lambda_{\min} \leq \lambda_{\max} \leq c \gamma_1$. Let $\boldsymbol{U} \boldsymbol{\Sigma} \boldsymbol{U}^{\mathrm{T}}$ be the eigen-decomposition of $\boldsymbol{W}_*^{-1/2} \boldsymbol{C}_\perp \left(p^{-1} \boldsymbol{L}^{\mathrm{T}} \boldsymbol{L} \right) \boldsymbol{C}_\perp^{\mathrm{T}} \boldsymbol{W}_*^{-1/2}$, where $\boldsymbol{U} \in \mathbb{R}^{(n-d) \times K}$ has orthonormal columns and

$$\boldsymbol{\Sigma} = \operatorname{diag}\left(\gamma_1, \ldots, \gamma_K \right).$$

Then the non-zero eigenvalues of $\boldsymbol{D}^{1/2} \boldsymbol{W}_*^{-1/2} \boldsymbol{C}_\perp \left(p^{-1} \boldsymbol{L}^{\mathrm{T}} \boldsymbol{L} \right) \boldsymbol{C}_\perp^{\mathrm{T}} \boldsymbol{W}_*^{-1/2} \boldsymbol{D}^{1/2}$ are also the eigenvalues of $\boldsymbol{\Sigma} \left(\boldsymbol{U}^{\mathrm{T}} \boldsymbol{D} \boldsymbol{U} \right)$, where $\|\boldsymbol{U}^{\mathrm{T}} \boldsymbol{D} \boldsymbol{U}\|_2, \|\left(\boldsymbol{U}^{\mathrm{T}} \boldsymbol{D} \boldsymbol{U} \right)^{-1}\|_2 \leq c$. Therefore,

$$\lambda_{\max} = \|\boldsymbol{\Sigma} \left(\boldsymbol{U}^{\mathrm{T}} \boldsymbol{D} \boldsymbol{U} \right)\|_2 \leq \gamma_1 c$$

and

$$\lambda_{\min}^{-1} = \|\left\{ \boldsymbol{\Sigma} \left(\boldsymbol{U}^{\mathrm{T}} \boldsymbol{D} \boldsymbol{U} \right) \right\}^{-1}\|_2 \leq \gamma_K^{-1} c,$$

which proves the result.

To prove (ii), let $\boldsymbol{R} = \left(n^{-1}\boldsymbol{C}_{\perp}^{\mathrm{T}}\boldsymbol{W}_{*}^{-1}\boldsymbol{C}_{\perp}\right)^{1/2}$, where $\|\boldsymbol{R}\boldsymbol{\ell}_g\|_2 \leq c_1$ for all $g = 1, \dots, p$ by Assumption 3.7. For $g \in [p]$, we then have,

$$\begin{aligned}
&\boldsymbol{\ell}_g^{\mathrm{T}}\left(n^{-1}\boldsymbol{C}_{\perp}^{\mathrm{T}}\hat{\boldsymbol{W}}^{-1}\boldsymbol{C}_{\perp}\right)^{-1}\boldsymbol{\ell}_g \\
&= \left(\boldsymbol{R}\boldsymbol{\ell}_g\right)^{\mathrm{T}}\left\{n^{-1}\left(\boldsymbol{W}_{*}^{-1/2}\boldsymbol{C}_{\perp}\boldsymbol{R}^{-1}\right)^{\mathrm{T}}\boldsymbol{W}_{*}^{1/2}\hat{\boldsymbol{W}}^{-1}\boldsymbol{W}_{*}^{1/2}\left(\boldsymbol{W}_{*}^{-1/2}\boldsymbol{C}_{\perp}\boldsymbol{R}^{-1}\right)\right\}^{-1}\left(\boldsymbol{R}\boldsymbol{\ell}_g\right),
\end{aligned}$$

where $n^{-1/2}\boldsymbol{W}_{*}^{-1/2}\boldsymbol{C}_{\perp}\boldsymbol{R}^{-1}$ has orthonormal columns. Since $\boldsymbol{W}_{*}^{1/2}\hat{\boldsymbol{W}}^{-1}\boldsymbol{W}_{*}^{1/2}$ has eigenvalues that are bounded away from 0 by a universal constant that does not depend on n or p by Assumptions 3.7 and 3.8, the result follows. □

***Remark* 3.20.** *We use the identification criterion*

$$\begin{aligned}
&n^{-1}\boldsymbol{C}_{\perp}^{\mathrm{T}}\hat{\boldsymbol{W}}^{-1}\boldsymbol{C}_{\perp} = I_K \\
&np^{-1}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L} = \mathrm{diag}\left(\lambda_1, \dots, \lambda_K\right)
\end{aligned}$$

where $\lambda_1 \geq \cdots \geq \lambda_K$, in the proofs of Theorems 3.7, 3.8 and 3.9 in Sections 3.13, 3.13 and 3.13. Note that $\lambda_1, \dots, \lambda_K$ are not the same as those defined in Assumption 3.4 in Section 3.11.1. Special care must be taken in the proofs of these theorems because the identification criterion is random, since $\hat{\boldsymbol{W}}$ is random. To cope with this, suppose $\left\{\boldsymbol{L}^{(0)}, \boldsymbol{C}^{(0)}\right\} \in \Lambda_{(1)}$, where $\Lambda_{(1)}$ was defined in the statement of Corollary 3.1. If $\boldsymbol{C}$ and $\boldsymbol{L}$ satisfy the aforementioned identification criterion, then

$$\begin{aligned}
\boldsymbol{C}_{\perp} &= \boldsymbol{C}_{\perp}^{(0)}\left[n^{-1}\left\{\boldsymbol{C}_{\perp}^{(0)}\right\}^{\mathrm{T}}\hat{\boldsymbol{W}}^{-1}\boldsymbol{C}_{\perp}^{(0)}\right]^{-1/2}\hat{\boldsymbol{U}} \\
\boldsymbol{L} &= \boldsymbol{L}^{(0)}\left[n^{-1}\left\{\boldsymbol{C}_{\perp}^{(0)}\right\}^{\mathrm{T}}\hat{\boldsymbol{W}}^{-1}\boldsymbol{C}_{\perp}^{(0)}\right]^{1/2}\hat{\boldsymbol{U}}
\end{aligned}$$

in Theorems 3.7 and 3.8, and

$$\begin{aligned}\boldsymbol{C} &= \boldsymbol{C}^{(0)}\left[n^{-1}\left\{\boldsymbol{C}_{\perp}^{(0)}\right\}^{\mathrm{T}}\hat{\boldsymbol{W}}^{-1}\boldsymbol{C}_{\perp}^{(0)}\right]^{-1/2}\hat{\boldsymbol{U}}\\ \boldsymbol{L} &= \boldsymbol{L}^{(0)}\left[n^{-1}\left\{\boldsymbol{C}_{\perp}^{(0)}\right\}^{\mathrm{T}}\hat{\boldsymbol{W}}^{-1}\boldsymbol{C}_{\perp}^{(0)}\right]^{1/2}\hat{\boldsymbol{U}}\end{aligned}$$

in Theorem 3.9, where $\hat{\boldsymbol{U}} \in \mathbb{R}^{K\times K}$ *is a unitary matrix that ensures*

$$np^{-1}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L} = \operatorname{diag}\left(\lambda_1,\ldots,\lambda_K\right).$$

By Assumptions 3.7 and 3.8, $n^{-1}\left\{\boldsymbol{C}_{\perp}^{(0)}\right\}^{\mathrm{T}}\hat{\boldsymbol{W}}^{-1}\boldsymbol{C}_{\perp}^{(0)}$ *has eigenvalues that are uniformly bounded above and below* c^{-1} *and* c *for some constant* $c > 1$*. Therefore,*

$$\begin{aligned}&\|n^{-1}\left\{\boldsymbol{C}_{\perp}^{(0)}\right\}^{\mathrm{T}}\hat{\boldsymbol{W}}^{-1}\boldsymbol{C}_{\perp}^{(0)} - I_K\|_2,\, |\lambda_k/\gamma_k - 1|,\, \|\hat{\boldsymbol{U}}\boldsymbol{\Pi} - I_K\|_2\\ =&O\left(\|\boldsymbol{W}_* - \hat{\boldsymbol{W}}\|_2\right) \quad (k = 1,\ldots,K)\end{aligned}$$

for some $\boldsymbol{\Pi} \in \mathcal{G}$*, where* $\mathcal{G}$ *was defined in the statement of Corollary 3.1. Note that* $\operatorname{Im}(\boldsymbol{C}) = \operatorname{Im}\left\{\boldsymbol{C}^{(0)}\right\}$.

Proof of Theorem 3.7

We next prove Theorem 3.7, which will be subsequently used in the proofs of Theorems 3.8 and 3.9 and Lemma 3.2. For ease of notation we will assume in this section that the data $\boldsymbol{Y}$ are generated according to model (3.2) with $d = 0$ and will use the incorrect, but simpler model in (3.7) with $d = 0$ as well in the proofs of Lemma 3.5, Theorem 3.10, Lemma 3.6 and Theorem 3.7 below. Since $\boldsymbol{L}$ and $\boldsymbol{C}$ are not unique and we are only interested in estimating P_C, δ_*^2 and $\boldsymbol{\tau}_*$, we may parametrize $\boldsymbol{L}$ and $\boldsymbol{C}$ to any propitious values so long as $\mathbb{E}(\boldsymbol{Y}) = \boldsymbol{L}\boldsymbol{C}^{\mathrm{T}}$. We assume $n^{-1}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{V}_*^{-1}\boldsymbol{C} = I_K$ and $np^{-1}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L} = \operatorname{diag}(\gamma_1,\ldots,\gamma_K)$ in this section. By Corollary 3.1, this uniquely identifies the columns of $\boldsymbol{C}$ and $\boldsymbol{L}$ up sign. Let

$\hat{\boldsymbol{C}} \in \mathbb{R}^{n \times r}$ be any estimate of $\boldsymbol{C} \in \mathbb{R}^{n \times K}$. Since we are only interested in the column spaces of $\boldsymbol{C}$ and $\hat{\boldsymbol{C}}$, we also assume without loss of generality that the first $\min(r, K)$ diagonal elements of $\hat{\boldsymbol{C}}^{\mathrm{T}}\boldsymbol{C} \in \mathbb{R}^{r \times K}$ are non-negative.

We first state a lemma about the extreme singular values of a Gaussian random matrix with independent rows.

***Lemma* 3.5.** *Let* $\boldsymbol{e}_g \sim N_n\left(0, \boldsymbol{V}_g\right)$ *be as defined in* (3.2a) *and suppose Assumption 3.7 holds. Then*

$$\|p^{-1}\sum_{g=1}^{p} \boldsymbol{e}_g\boldsymbol{e}_g^{\mathrm{T}} - p^{-1}\sum_{g=1}^{p} \boldsymbol{V}_g\|_2 = O_P\left(n^{1/2}p^{-1/2}\right).$$

Proof. This follows directly from Theorem 5.39 and Remark 5.40 in Eldar & Kutyniok (2012). □

***Theorem* 3.10.** *Suppose Assumptions 3.7 and 3.8 hold and define* $\epsilon = \|\hat{\boldsymbol{V}} - \boldsymbol{V}_*\|_2$, *where* $\hat{\boldsymbol{V}}$ *is an estimate of* $\boldsymbol{V}_*$ *with* $|\hat{\boldsymbol{V}}| = 1$, *and*

$$\bar{\boldsymbol{C}} = \hat{\boldsymbol{V}}^{-1/2}\boldsymbol{C}\left(\boldsymbol{C}^{\mathrm{T}}\hat{\boldsymbol{V}}^{-1}\boldsymbol{C}\right)^{-1/2}\hat{\boldsymbol{U}} \tag{3.28a}$$

$$\bar{\boldsymbol{Q}} = \boldsymbol{Q}_{\bar{\boldsymbol{C}}} \tag{3.28b}$$

$$\lambda_k = \boldsymbol{\lambda}_k\left(p^{-1}\hat{\boldsymbol{V}}^{-1/2}\boldsymbol{C}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{C}^{\mathrm{T}}\hat{\boldsymbol{V}}^{-1/2}\right). \tag{3.28c}$$

where $\hat{\boldsymbol{U}} \in \mathbb{R}^{K \times K}$ *contains the right singular vectors of* $\boldsymbol{L}\left(\boldsymbol{C}^{\mathrm{T}}\hat{\boldsymbol{V}}^{-1}\boldsymbol{C}\right)^{1/2}$ *and* $\boldsymbol{\lambda}_k(\boldsymbol{A})$ *is the* k^{th} *largest eigenvector of the matrix* $\boldsymbol{A}$. *Define the estimates* $\hat{\bar{\boldsymbol{C}}} \in \mathbb{R}^{n \times K}$ *and* $\hat{\lambda}_1, \ldots, \hat{\lambda}_K$ *to be the first* K *eigenvectors and eigenvalues of* $\hat{\boldsymbol{V}}^{-1/2}p^{-1}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}\hat{\boldsymbol{V}}^{-1/2}$, *respectively. Then if* $\epsilon/\gamma_K = o_P(1)$,

$$\|P_{\bar{\boldsymbol{C}}} - P_{\hat{\bar{\boldsymbol{C}}}}\|_F^2 = O_P\left[\left\{n^{1/2}\left(\gamma_K p\right)^{-1/2} + \epsilon\gamma_K^{-1}\right\}^2\right]. \tag{3.29}$$

Further, if $\epsilon = o_P(1)$,

$$\hat{\lambda}_k/\lambda_k = 1 + \delta_*^2/\lambda_k + O_P\left\{n\left(\gamma_K p\right)^{-1} + n^{1/2}\left(\gamma_K p\right)^{-1/2}\epsilon + \left(\gamma_K p\right)^{-1/2} + \epsilon\gamma_K^{-1}\right\} \tag{3.30}$$

$$\|\bar{\boldsymbol{C}}^{\mathrm{T}}\hat{\bar{\boldsymbol{C}}} - I_K\|_2 = O_P\left\{n\left(\gamma_K p\right)^{-1} + n^{1/2}\left(\gamma_K p\right)^{-1/2}\epsilon + \left(\gamma_K p\right)^{-1/2} + \epsilon\gamma_K^{-1}\right\}. \tag{3.31}$$

Proof. First, $(\lambda_k - \lambda_{k+1})/\lambda_{k+1} \geq c_1^{-1} + o_P(1)$ when $\epsilon = o_P(1)$ by item *(a)* of Assumption 3.7. We let $\gamma = \gamma_K$ and use a technique developed in Paul (2007) and define the rotated matrix $\bar{\boldsymbol{S}}$ to be

$$\begin{aligned}
\bar{\boldsymbol{S}} &= \begin{pmatrix} \bar{\boldsymbol{C}}^{\mathrm{T}} \\ \bar{\boldsymbol{Q}}^{\mathrm{T}} \end{pmatrix} \left(\gamma p\right)^{-1} \hat{\boldsymbol{V}}^{-1/2}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}\hat{\boldsymbol{V}}^{-1/2} \begin{pmatrix} \bar{\boldsymbol{C}} & \bar{\boldsymbol{Q}} \end{pmatrix} \\
&= \begin{Bmatrix} \left(\bar{\boldsymbol{L}} + \bar{\boldsymbol{E}}_1\right)^{\mathrm{T}}\left(\bar{\boldsymbol{L}} + \bar{\boldsymbol{E}}_1\right) & \left(\bar{\boldsymbol{L}} + \bar{\boldsymbol{E}}_1\right)^{\mathrm{T}}\bar{\boldsymbol{E}}_2 \\ \bar{\boldsymbol{E}}_2^{\mathrm{T}}\left(\bar{\boldsymbol{L}} + \bar{\boldsymbol{E}}_1\right) & \bar{\boldsymbol{E}}_2^{\mathrm{T}}\bar{\boldsymbol{E}}_2 \end{Bmatrix} && (3.32) \\
\bar{\boldsymbol{L}} &= n^{1/2}\left(\gamma p\right)^{-1/2}\boldsymbol{L}\left(n^{-1}\boldsymbol{C}^{\mathrm{T}}\hat{\boldsymbol{V}}^{-1}\boldsymbol{C}\right)^{1/2}\hat{\boldsymbol{U}} && (3.33) \\
\bar{\boldsymbol{E}}_1 &= \left(\gamma p\right)^{-1/2}\boldsymbol{E}\hat{\boldsymbol{V}}^{-1/2}\bar{\boldsymbol{C}} && (3.34) \\
\bar{\boldsymbol{E}}_2 &= \left(\gamma p\right)^{-1/2}\boldsymbol{E}\hat{\boldsymbol{V}}^{-1/2}\bar{\boldsymbol{Q}}. && (3.35)
\end{aligned}$$

We now get explicit error bounds for each term in $\bar{\boldsymbol{S}}$.

1.

$$\bar{\boldsymbol{L}}^{\mathrm{T}}\bar{\boldsymbol{E}}_1 = (\gamma p)^{-1/2}\,\bar{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{E}\hat{\boldsymbol{V}}^{-1}\left(n^{-1/2}\boldsymbol{C}\right)\underbrace{\left(n^{-1}\boldsymbol{C}^{\mathrm{T}}\hat{\boldsymbol{V}}^{-1}\boldsymbol{C}\right)^{-1/2}\hat{\boldsymbol{U}}}_{O_P(1)}$$

$$= \left(\underbrace{(\gamma p)^{-1/2}\,\bar{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{E}\boldsymbol{V}_*^{-1}\left(n^{-1/2}\boldsymbol{C}\right)}_{O_P\left\{(\gamma p)^{-1/2}\right\}} + \underbrace{(\gamma p)^{-1/2}\,\bar{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{E}\left(\hat{\boldsymbol{V}}^{-1} - \boldsymbol{V}_*^{-1}\right)\left(n^{-1/2}\boldsymbol{C}\right)}_{O_P\left\{n^{1/2}(\gamma p)^{-1/2}\epsilon\right\}}\right) O_P(1).$$

Therefore, $\|\bar{\boldsymbol{L}}^{\mathrm{T}}\bar{\boldsymbol{E}}_1\|_2 = O_P\left\{(\gamma p)^{-1/2} + n^{1/2}\,(\gamma p)^{-1/2}\,\epsilon\right\}$.

2.

$$\bar{\boldsymbol{L}}^{\mathrm{T}}\bar{\boldsymbol{E}}_2 = \underbrace{(\gamma p)^{-1/2}\,\bar{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{E}\hat{\boldsymbol{V}}^{-1/2}\hat{\boldsymbol{V}}^{1/2}\boldsymbol{Q}_{\boldsymbol{C}}}_{O_P\left\{n^{1/2}(\gamma p)^{-1/2}\right\}}\underbrace{\left(\boldsymbol{Q}_{\boldsymbol{C}}^{\mathrm{T}}\hat{\boldsymbol{V}}\boldsymbol{Q}_{\boldsymbol{C}}\right)^{-1/2}}_{O_P(1)}.$$

Therefore, $\|\bar{\boldsymbol{L}}^{\mathrm{T}}\bar{\boldsymbol{E}}_2\|_2 = O_P\left\{n^{1/2}\,(\gamma p)^{-1/2}\right\}$.

3.

$$\bar{\boldsymbol{E}}_1^{\mathrm{T}}\bar{\boldsymbol{E}}_1 = (\gamma p)^{-1}\,\bar{\boldsymbol{C}}^{\mathrm{T}}\boldsymbol{V}_*^{-1/2}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\boldsymbol{V}_*^{-1/2}\bar{\boldsymbol{C}} + O_P\left(\epsilon\gamma^{-1}\right).$$

Define $\tilde{\boldsymbol{C}} = \boldsymbol{V}_*^{-1/2}\boldsymbol{C}\left(\boldsymbol{C}^{\mathrm{T}}\boldsymbol{V}_*^{-1}\boldsymbol{C}\right)^{-1/2}$.

$$
\begin{aligned}
(\gamma p)^{-1}\,\bar{\boldsymbol{C}}^{\mathrm{T}}\hat{\boldsymbol{V}}^{-1/2}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\hat{\boldsymbol{V}}^{-1/2}\bar{\boldsymbol{C}} =& \gamma^{-1}\delta_*^2 I_K \\
&+ \gamma^{-1}\bar{\boldsymbol{C}}^{\mathrm{T}}\left(\hat{\boldsymbol{V}}^{-1/2}p^{-1}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\hat{\boldsymbol{V}}^{-1/2} - \delta_*^2 I_n\right)\bar{\boldsymbol{C}} \\
=& \gamma^{-1}\delta_*^2 I_K \\
&+ \underbrace{\gamma^{-1}\hat{\boldsymbol{U}}^{\mathrm{T}}\tilde{\boldsymbol{C}}^{\mathrm{T}}\left(\boldsymbol{V}_*^{-1/2}p^{-1}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\boldsymbol{V}_*^{-1/2} - \delta_*^2 I_n\right)\tilde{\boldsymbol{C}}\hat{\boldsymbol{U}}}_{O_P(\gamma^{-1}p^{-1/2})} \\
&+ O_P\left(\epsilon\gamma^{-1}\right).
\end{aligned}
$$

Therefore,

$$
\|\bar{\boldsymbol{E}}_1^{\mathrm{T}}\bar{\boldsymbol{E}}_1 - \gamma^{-1}\delta_*^2 I_K\|_2 = O_P\left(\epsilon\gamma^{-1} + \gamma^{-1}p^{-1/2}\right).
$$

4. Let $\tilde{\boldsymbol{Q}} = \boldsymbol{Q}_{\tilde{\boldsymbol{C}}}$. Then

$$
\begin{aligned}
\bar{\boldsymbol{E}}_1^{\mathrm{T}}\bar{\boldsymbol{E}}_2 =& (\gamma p)^{-1}\,\bar{\boldsymbol{C}}^{\mathrm{T}}\hat{\boldsymbol{V}}^{-1/2}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\hat{\boldsymbol{V}}^{-1/2}\bar{\boldsymbol{Q}} = (\gamma p)^{-1}\,\hat{\boldsymbol{U}}^{\mathrm{T}}\tilde{\boldsymbol{C}}^{\mathrm{T}}\boldsymbol{V}_*^{-1/2}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\boldsymbol{V}_*^{-1/2}\tilde{\boldsymbol{Q}} \\
&+ O_P\left(\epsilon\gamma^{-1}\right).
\end{aligned}
$$

Let $\tilde{\boldsymbol{E}} = p^{-1/2}\boldsymbol{E}\boldsymbol{V}_*^{-1/2}$, $\tilde{\boldsymbol{E}}_1 = \tilde{\boldsymbol{E}}\tilde{\boldsymbol{C}}$, $\tilde{\boldsymbol{E}}_2 = \tilde{\boldsymbol{E}}\tilde{\boldsymbol{Q}}$ and $\boldsymbol{A}_g = \boldsymbol{V}_*^{-1/2}\boldsymbol{V}_g\boldsymbol{V}_*^{-1/2}$. Then the k^{th} column of $\tilde{\boldsymbol{E}}_2^{\mathrm{T}}\tilde{\boldsymbol{E}}_1$ $(k = 1, \ldots, K)$ is

$$
\begin{aligned}
\boldsymbol{b}_k = p^{-1}\tilde{\boldsymbol{Q}}^{\mathrm{T}}\tilde{\boldsymbol{E}}^{\mathrm{T}}\tilde{\boldsymbol{E}}\tilde{\boldsymbol{c}}_k = \tilde{\boldsymbol{Q}}^{\mathrm{T}}&\left(p^{-1}\tilde{\boldsymbol{E}}^{\mathrm{T}}\tilde{\boldsymbol{E}} - p^{-1}\sum_{g=1}^{p}\boldsymbol{A}_g\right)\tilde{\boldsymbol{c}}_k + \tilde{\boldsymbol{Q}}^{\mathrm{T}}\underbrace{p^{-1}\sum_{g=1}^{p}\boldsymbol{A}_g}_{=\delta_*^2 I_n}\tilde{\boldsymbol{c}}_k \\
= p^{-1}\sum_{g=1}^{p}\tilde{\boldsymbol{Q}}^{\mathrm{T}}\boldsymbol{A}_g^{1/2}&\left(\boldsymbol{r}_g\boldsymbol{r}_g^{\mathrm{T}} - \delta_*^2 I_n\right)\boldsymbol{A}_g^{1/2}\tilde{\boldsymbol{c}}_k = p^{-1}\sum_{g=1}^{p}\boldsymbol{b}_{g,k}
\end{aligned}
$$

where $\boldsymbol{r}_g \sim N\left(\boldsymbol{0}, I_n\right)$, $\boldsymbol{r}_g$ and $\boldsymbol{r}_{g'}$ are independent for $g \neq g'$ $(g = 1, \ldots, p; g' = 1, \ldots, p)$

and $\mathbb{E}\left(\boldsymbol{b}_k\right) = \mathbf{0}$. Therefore,

$$\mathbb{E}\left(\|\boldsymbol{b}_k\|_2^2\right) = \sum_{i=1}^{n-K} \mathbb{E}\left(\boldsymbol{b}_{ki}^2\right) = \sum_{i=1}^{n-K} \mathbb{V}\left(\boldsymbol{b}_{ki}\right) = O\left\{n\,\mathbb{V}\left(\boldsymbol{b}_{k1}\right)\right\} = O\left(np^{-1}\right)$$

where the second to last and last equalities follow because $\|\boldsymbol{A}_g\|_2$ is uniformly bounded by item *(a)* of Assumption 3.7. Therefore, $\|\bar{\boldsymbol{E}}_2^{\mathrm{T}}\bar{\boldsymbol{E}}_1\|_2 = O_P\left(\gamma^{-1}n^{1/2}p^{-1/2} + \epsilon\gamma^{-1}\right)$.

Let $\mu_k = \lambda_k/\gamma$ and define $\begin{pmatrix}\hat{\boldsymbol{v}}_k^{\mathrm{T}} & \hat{\boldsymbol{z}}_k^{\mathrm{T}}\end{pmatrix}^{\mathrm{T}} \in \mathbb{R}^n$ to be the k^{th} normalized eigenvector of $\bar{\boldsymbol{S}}$, where $\hat{\boldsymbol{v}}_k \in \mathbb{R}^K$ and $\hat{\boldsymbol{z}}_k \in \mathbb{R}^{n-K}$. All of this proves that $\hat{\mu}_k = \mu_k + \delta_*^2/\gamma + o_P(1)$ by Weyl's theorem. To prove sharper bounds, we set set up the eigenvalue equations

$$\begin{aligned}
\hat{\mu}_k\hat{\boldsymbol{v}}_k &= \left(\bar{\boldsymbol{L}} + \bar{\boldsymbol{E}}_1\right)^{\mathrm{T}}\left(\bar{\boldsymbol{L}} + \bar{\boldsymbol{E}}_1\right)\hat{\boldsymbol{v}}_k + \left(\bar{\boldsymbol{L}} + \bar{\boldsymbol{E}}_1\right)^{\mathrm{T}}\bar{\boldsymbol{E}}_2\hat{\boldsymbol{z}}_k \\
\hat{\mu}_k\hat{\boldsymbol{z}}_k &= \bar{\boldsymbol{E}}_2^{\mathrm{T}}\left(\bar{\boldsymbol{L}} + \bar{\boldsymbol{E}}_1\right)\hat{\boldsymbol{v}}_k + \bar{\boldsymbol{E}}_2^{\mathrm{T}}\bar{\boldsymbol{E}}_2\hat{\boldsymbol{z}}_k.
\end{aligned}$$

A little algebra shows that on the event $\hat{\mu}_k I_n - \bar{\boldsymbol{E}}_2^{\mathrm{T}}\bar{\boldsymbol{E}}_2$ is invertible,

$$\hat{\mu}_k\hat{\boldsymbol{v}}_k = \left\{\left(\bar{\boldsymbol{L}} + \bar{\boldsymbol{E}}_1\right)^{\mathrm{T}}\left(\bar{\boldsymbol{L}} + \bar{\boldsymbol{E}}_1\right) + \left(\bar{\boldsymbol{L}} + \bar{\boldsymbol{E}}_1\right)^{\mathrm{T}}\bar{\boldsymbol{E}}_2\left(\hat{\mu}_k I_n - \bar{\boldsymbol{E}}_2^{\mathrm{T}}\bar{\boldsymbol{E}}_2\right)^{-1}\bar{\boldsymbol{E}}_2^{\mathrm{T}}\left(\bar{\boldsymbol{L}} + \bar{\boldsymbol{E}}_1\right)\right\}\hat{\boldsymbol{v}}_k \tag{3.36}$$

$$\hat{\boldsymbol{z}}_k = \left(\hat{\mu}_k I_n - \bar{\boldsymbol{E}}_2^{\mathrm{T}}\bar{\boldsymbol{E}}_2\right)^{-1}\bar{\boldsymbol{E}}_2^{\mathrm{T}}\left(\bar{\boldsymbol{L}} + \bar{\boldsymbol{E}}_1\right)\hat{\boldsymbol{v}}_k \tag{3.37}$$

where $\hat{\mu}_k I_n - \bar{\boldsymbol{E}}_2^{\mathrm{T}}\bar{\boldsymbol{E}}_2$ is invertible with probability tending to 1 because $\hat{\mu}_k = \mu_k + \delta_*^2/\gamma + o_P(1)$

and by Lemma 3.5. By what we showed above, we then have that

$$\|\hat{\boldsymbol{z}}_k\|_2 = O_P\left\{n^{1/2}\left(\gamma p\right)^{-1/2} + \epsilon\gamma^{-1}\right\} \tag{3.38a}$$

$$\begin{aligned}\hat{\mu}_k &= \boldsymbol{\lambda}_k\left\{\left(\bar{\boldsymbol{L}} + \bar{\boldsymbol{E}}_1\right)^{\mathrm{T}}\left(\bar{\boldsymbol{L}} + \bar{\boldsymbol{E}}_1\right)\right\} + O_P\left[\left\{n^{1/2}\left(\gamma p\right)^{-1/2} + \epsilon\gamma^{-1}\right\}^2\right]\\ &= \mu_k + \delta_*^2\gamma^{-1} + O_P\left\{n\left(\gamma p\right)^{-1} + n^{1/2}\left(\gamma p\right)^{-1/2}\epsilon + \left(\gamma p\right)^{-1/2} + \epsilon\gamma^{-1}\right\}\end{aligned} \tag{3.38b}$$

$$\|\hat{\boldsymbol{v}}_k - \boldsymbol{v}_k\|_2 = O_P\left[\left\{n^{1/2}\left(\gamma p\right)^{-1/2} + \epsilon\gamma^{-1}\right\}^2\right] \tag{3.38c}$$

$$\|\hat{\boldsymbol{v}}_k - \boldsymbol{a}_k\|_2 = O_P\left\{n\left(\gamma p\right)^{-1} + n^{1/2}\left(\gamma p\right)^{-1/2}\epsilon + \left(\gamma p\right)^{-1/2} + \epsilon\gamma^{-1}\right\} \tag{3.38d}$$

where $\boldsymbol{v}_k$ is the k^{th} eigenvector of $\left(\bar{\boldsymbol{L}} + \bar{\boldsymbol{E}}_1\right)^{\mathrm{T}}\left(\bar{\boldsymbol{L}} + \bar{\boldsymbol{E}}_1\right)$ and $\boldsymbol{a}_k \in \mathbb{R}^K$ is the k^{th} standard basis vector. Equation (3.38a) follow from Weyl's Theorem (i.e. $\hat{\mu}_k = \mu_k + \delta_*^2/\lambda_k + o_P(1)$) and (3.38b) follows from Theorem 3.5 of Auffinger & Tang (2015) and the assumption that $(\lambda_k - \lambda_{k+1})/\lambda_{k+1} \geq c_1^{-1} + o_P(1)$. Since the left and right singular vectors of any matrix are unique up to sign parity, a second application of Theorem 3.5 of Auffinger & Tang (2015), along with the assumption that $(\lambda_k - \lambda_{k+1})/\lambda_{k+1} \geq c_1^{-1} + o_P(1)$, proves (3.38c) and (3.38d). This proves (3.30).

Define $\hat{\boldsymbol{v}} = \begin{pmatrix}\hat{\boldsymbol{v}}_1 & \cdots & \hat{\boldsymbol{v}}_K\end{pmatrix}$ and $\hat{\boldsymbol{z}} = \begin{pmatrix}\hat{\boldsymbol{z}}_1 & \cdots & \hat{\boldsymbol{z}}_K\end{pmatrix}$. Since $\hat{\boldsymbol{v}}^{\mathrm{T}}\hat{\boldsymbol{v}} + \hat{\boldsymbol{z}}^{\mathrm{T}}\hat{\boldsymbol{z}} = I_K$ and by (3.38a),

$$\|\hat{\boldsymbol{v}}\hat{\boldsymbol{v}}^{\mathrm{T}} - I_K\|_2 = O_P\left[\left\{n^{1/2}\left(\gamma p\right)^{-1/2} + \epsilon\gamma^{-1}\right\}^2\right]. \tag{3.39}$$

Then

$$\hat{\bar{\boldsymbol{C}}} = \bar{\boldsymbol{C}}\hat{\boldsymbol{v}} + \bar{\boldsymbol{Q}}\hat{\boldsymbol{z}} \tag{3.40}$$

with

$$
\begin{aligned}
\|P_{\hat{\bar{\boldsymbol{C}}}} - P_{\bar{\boldsymbol{C}}}\|_F^2 &= 2K - 2\,\mathrm{tr}\left[\bar{\boldsymbol{C}}^{\mathrm{T}}\hat{\bar{\boldsymbol{C}}}\hat{\bar{\boldsymbol{C}}}^{\mathrm{T}}\bar{\boldsymbol{C}}\right] = 2K - 2\,\mathrm{tr}\left[\hat{\boldsymbol{v}}\hat{\boldsymbol{v}}^{\mathrm{T}}\right] \\
&= O_P\left[\left\{n^{1/2}\left(\gamma p\right)^{-1/2} + \epsilon\gamma^{-1}\right\}^2\right],
\end{aligned}
$$

which proves (3.29). Lastly, (3.31) follows from (3.38c) and (3.38d). □

***Corollary* 3.2** (Used in the proof of Theorems 3.7 and 3.9)**.** *Suppose the assumptions of Theorem 3.10 hold (including the assumption that $\epsilon = o_P(1)$) and let $\boldsymbol{M}_g \in \mathbb{R}^{n\times n}$ be a random symmetric positive definite matrix that is only a function of $\boldsymbol{e}_g$, the g^{th} row of $\boldsymbol{E}$, with $\|\boldsymbol{M}_g\|_2 = 1$ and smallest eigenvalue uniformly bounded away from 0. Define $\hat{\boldsymbol{C}} = n^{1/2}\hat{\boldsymbol{V}}^{1/2}\hat{\bar{\boldsymbol{C}}}$ and let $\hat{\boldsymbol{A}} = \left(n^{-1}\boldsymbol{C}^{\mathrm{T}}\hat{\boldsymbol{V}}^{-1}\boldsymbol{C}\right)^{-1/2}\hat{\boldsymbol{U}}$, where $\boldsymbol{C}$ and $\hat{\boldsymbol{U}}$ are defined in the statement of Theorem 3.10. Then using the same notation as the statement of Theorem 3.10,*

$$\|n^{-1}\hat{\boldsymbol{C}}^{\mathrm{T}}\boldsymbol{M}_g\hat{\boldsymbol{C}} - n^{-1}\hat{\boldsymbol{A}}^{\mathrm{T}}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{M}_g\boldsymbol{C}\hat{\boldsymbol{A}}\|_2 = O_P\left(\eta\right) \tag{3.41}$$

$$\|n^{-1}\hat{\boldsymbol{C}}^{\mathrm{T}}\boldsymbol{M}_g\boldsymbol{C}\hat{\boldsymbol{A}} - n^{-1}\hat{\boldsymbol{A}}^{\mathrm{T}}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{M}_g\boldsymbol{C}\hat{\boldsymbol{A}}\|_2 = O_P\left(\eta\right) \tag{3.42}$$

$$\|P_{\boldsymbol{C}} - P_{\hat{\boldsymbol{C}}}\|_F^2 = O_P\left(\eta\right). \tag{3.43}$$

where $\eta = n\left(\gamma_K p\right)^{-1} + \left(\gamma_K p\right)^{-1/2} + n^{1/2}\left(\gamma_K p\right)^{-1/2}\epsilon + \epsilon\gamma_K^{-1}$.

Proof. The proof utilizes objects defined in Theorem 3.10. We see that

$$
\begin{aligned}
n^{-1}\hat{\boldsymbol{C}}^{\mathrm{T}}\boldsymbol{M}_g\hat{\boldsymbol{C}} =& \hat{\bar{\boldsymbol{C}}}^{\mathrm{T}}\hat{\boldsymbol{V}}^{1/2}\boldsymbol{M}_g\hat{\boldsymbol{V}}^{1/2}\hat{\bar{\boldsymbol{C}}} = \hat{\boldsymbol{v}}^{\mathrm{T}}\bar{\boldsymbol{C}}^{\mathrm{T}}\hat{\boldsymbol{V}}^{1/2}\boldsymbol{M}_g\hat{\boldsymbol{V}}^{1/2}\bar{\boldsymbol{C}}\hat{\boldsymbol{v}} + \hat{\boldsymbol{v}}^{\mathrm{T}}\bar{\boldsymbol{C}}^{\mathrm{T}}\hat{\boldsymbol{V}}^{1/2}\boldsymbol{M}_g\hat{\boldsymbol{V}}^{1/2}\bar{\boldsymbol{Q}}\hat{\boldsymbol{z}} \\
&+ \left(\hat{\boldsymbol{v}}^{\mathrm{T}}\bar{\boldsymbol{C}}^{\mathrm{T}}\hat{\boldsymbol{V}}^{1/2}\boldsymbol{M}_g\hat{\boldsymbol{V}}^{1/2}\bar{\boldsymbol{Q}}\hat{\boldsymbol{z}}\right)^{\mathrm{T}} + \underbrace{\hat{\boldsymbol{z}}^{\mathrm{T}}\bar{\boldsymbol{Q}}^{\mathrm{T}}\hat{\boldsymbol{V}}^{1/2}\boldsymbol{M}_g\hat{\boldsymbol{V}}^{1/2}\bar{\boldsymbol{Q}}\hat{\boldsymbol{z}}}_{O_P\left[\left\{n(\gamma_K p)^{-1} + \epsilon\gamma_K^{-1}\right\}^2\right] = O_P(\eta)}.
\end{aligned}
$$

First,

$$
\begin{aligned}
&\|\hat{\boldsymbol{v}}^{\mathrm{T}}\bar{\boldsymbol{C}}^{\mathrm{T}}\hat{\boldsymbol{V}}^{1/2}\boldsymbol{M}_g\hat{\boldsymbol{V}}^{1/2}\bar{\boldsymbol{C}}\hat{\boldsymbol{v}} - n^{-1}\hat{\boldsymbol{A}}^{\mathrm{T}}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{M}_g\boldsymbol{C}\hat{\boldsymbol{A}}\|_2 \\
=&\|\hat{\boldsymbol{v}}^{\mathrm{T}}\bar{\boldsymbol{C}}^{\mathrm{T}}\hat{\boldsymbol{V}}^{1/2}\boldsymbol{M}_g\hat{\boldsymbol{V}}^{1/2}\bar{\boldsymbol{C}}\hat{\boldsymbol{v}} - \bar{\boldsymbol{C}}^{\mathrm{T}}\hat{\boldsymbol{V}}^{1/2}\boldsymbol{M}_g\hat{\boldsymbol{V}}^{1/2}\bar{\boldsymbol{C}}\|_2 = O_P\left(\eta\right).
\end{aligned}
$$

Next, by the proof of Theorem 3.10,

$$
\|\hat{\boldsymbol{z}} - \bar{\boldsymbol{E}}_2^{\mathrm{T}}\left(\bar{\boldsymbol{L}} + \bar{\boldsymbol{E}}_1\right)\|_2 = O_P\left(\eta\right).
$$

Therefore, we need only understand how

$$
\begin{aligned}
&\bar{\boldsymbol{C}}^{\mathrm{T}}\hat{\boldsymbol{V}}^{1/2}\boldsymbol{M}_g\hat{\boldsymbol{V}}^{1/2}\bar{\boldsymbol{Q}}\bar{\boldsymbol{E}}_2^{\mathrm{T}}\left(\bar{\boldsymbol{L}} + \bar{\boldsymbol{E}}_1\right) = \left(n\gamma_K p\right)^{-1/2}\hat{\boldsymbol{A}}^{\mathrm{T}}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{M}_g\hat{\boldsymbol{V}}\boldsymbol{Q}_{\boldsymbol{C}}\left(\boldsymbol{Q}_{\boldsymbol{C}}^{\mathrm{T}}\hat{\boldsymbol{V}}\boldsymbol{Q}_{\boldsymbol{C}}\right)^{-1} \\
&\times\boldsymbol{Q}_{\boldsymbol{C}}^{\mathrm{T}}\boldsymbol{E}^{\mathrm{T}}\bar{\boldsymbol{L}} + \gamma_K^{-1}n^{-1}\hat{\boldsymbol{A}}^{\mathrm{T}}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{M}_g\hat{\boldsymbol{V}}\boldsymbol{Q}_{\boldsymbol{C}}\left(\boldsymbol{Q}_{\boldsymbol{C}}^{\mathrm{T}}\hat{\boldsymbol{V}}\boldsymbol{Q}_{\boldsymbol{C}}\right)^{-1}\boldsymbol{Q}_{\boldsymbol{C}}^{\mathrm{T}}\left(p^{-1}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\right)\hat{\boldsymbol{V}}^{-1}\boldsymbol{C}\hat{\boldsymbol{A}} \\
=&\left(\gamma_K pn\right)^{-1/2}\hat{\boldsymbol{A}}^{\mathrm{T}}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{M}_g\boldsymbol{V}_*\boldsymbol{Q}_{\boldsymbol{C}}\left(\boldsymbol{Q}_{\boldsymbol{C}}^{\mathrm{T}}\boldsymbol{V}_*\boldsymbol{Q}_{\boldsymbol{C}}\right)^{-1}\boldsymbol{Q}_{\boldsymbol{C}}^{\mathrm{T}}\boldsymbol{E}^{\mathrm{T}}\bar{\boldsymbol{L}} \\
&+\gamma_K^{-1}n^{-1}\hat{\boldsymbol{A}}^{\mathrm{T}}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{M}_g\boldsymbol{V}_*\boldsymbol{Q}_{\boldsymbol{C}}\left(\boldsymbol{Q}_{\boldsymbol{C}}^{\mathrm{T}}\boldsymbol{V}_*\boldsymbol{Q}_{\boldsymbol{C}}\right)^{-1}\boldsymbol{Q}_{\boldsymbol{C}}^{\mathrm{T}}\left(p^{-1}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\right)\boldsymbol{V}_*^{-1}\boldsymbol{C}\hat{\boldsymbol{A}} \\
&+O_P\left\{\epsilon\gamma_K^{-1} + n^{1/2}\left(\gamma_K p\right)^{-1/2}\epsilon\right\}
\end{aligned}
$$

behaves. We can ignore $\hat{\boldsymbol{A}}$ because $\|\hat{\boldsymbol{A}}\|_2 = O(1)$ by Assumptions 3.7 and 3.8. We can write the first term as

$$
\begin{aligned}
&\underbrace{\left(\gamma_K p\right)^{-1}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{M}_g\boldsymbol{V}_*\boldsymbol{Q}_{\boldsymbol{C}}\left(\boldsymbol{Q}_{\boldsymbol{C}}^{\mathrm{T}}\boldsymbol{V}_*\boldsymbol{Q}_{\boldsymbol{C}}\right)^{-1}\boldsymbol{Q}_{\boldsymbol{C}}^{\mathrm{T}}\boldsymbol{e}_g\boldsymbol{\ell}_g^{\mathrm{T}}}_{O_P\left\{n\left(\gamma_K p\right)^{-1}\right\}} + \\
&\underbrace{\left(\gamma_K np\right)^{-1/2}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{M}_g\boldsymbol{V}_*\boldsymbol{Q}_{\boldsymbol{C}}\left(\boldsymbol{Q}_{\boldsymbol{C}}^{\mathrm{T}}\boldsymbol{V}_*\boldsymbol{Q}_{\boldsymbol{C}}\right)^{-1}\boldsymbol{Q}_{\boldsymbol{C}}^{\mathrm{T}}\boldsymbol{E}_{-g}^{\mathrm{T}}\bar{\boldsymbol{L}}_{-g}}_{O_P\left\{\left(\gamma_K p\right)^{-1/2}\right\}} \\
=&O_P\left(\eta\right)
\end{aligned}
$$

where the subscript $-g$ means we remove the g^{th} row from the matrix. The second term can also be decomposed in the same way:

$$
\begin{aligned}
&\underbrace{\gamma_K^{-1} n^{-1} \boldsymbol{C}^{\mathrm{T}} \boldsymbol{M}_g \boldsymbol{V}_* \boldsymbol{Q}_{\boldsymbol{C}} \left(\boldsymbol{Q}_{\boldsymbol{C}}^{\mathrm{T}} \boldsymbol{V}_* \boldsymbol{Q}_{\boldsymbol{C}}\right)^{-1} \boldsymbol{Q}_{\boldsymbol{C}}^{\mathrm{T}} \left(p^{-1} \boldsymbol{e}_g \boldsymbol{e}_g^{\mathrm{T}}\right) \boldsymbol{V}_*^{-1} \boldsymbol{C}}_{O_P\left\{n^{1/2}(\gamma_K p)^{-1}\right\}} + \\
&\underbrace{\gamma_K^{-1} n^{-1} \boldsymbol{C}^{\mathrm{T}} \boldsymbol{M}_g \boldsymbol{V}_* \boldsymbol{Q}_{\boldsymbol{C}} \left(\boldsymbol{Q}_{\boldsymbol{C}}^{\mathrm{T}} \boldsymbol{V}_* \boldsymbol{Q}_{\boldsymbol{C}}\right)^{-1} \boldsymbol{Q}_{\boldsymbol{C}}^{\mathrm{T}} \left(p^{-1} \boldsymbol{E}_{-g}^{\mathrm{T}} \boldsymbol{E}_{-g}\right) \boldsymbol{V}_*^{-1} \boldsymbol{C}}_{O_P\left\{(\gamma_K p)^{-1/2}\right\}} \\
&= O_P(\eta).
\end{aligned}
$$

This completes the proof for (3.41). The proof of (3.42) is identical to the analysis above and is not shown.

To prove (3.43), we see that

$$
\begin{aligned}
\|P_{\boldsymbol{C}} - P_{\hat{\boldsymbol{C}}}\|_F^2 &= 2I_K - 2\operatorname{tr}\left(P_{\boldsymbol{C}} P_{\hat{\boldsymbol{C}}}\right) = 2I_K - 2\operatorname{tr}\left\{\boldsymbol{C}^{\mathrm{T}} \hat{\boldsymbol{C}} \left(\hat{\boldsymbol{C}}^{\mathrm{T}} \hat{\boldsymbol{C}}\right)^{-1} \hat{\boldsymbol{C}}^{\mathrm{T}} \boldsymbol{C} \left(\boldsymbol{C}^{\mathrm{T}} \boldsymbol{C}\right)^{-1}\right\} \\
&= O_P(\eta).
\end{aligned}
$$

□

***Corollary* 3.3.** *Suppose the conditions of Corollary 3.2 hold. Then using the same notation as Theorem 3.10,*

$$
\gamma_K \left\{ \bar{\boldsymbol{S}} - \sum_{k=1}^{K} \hat{\mu}_k \begin{pmatrix} \hat{\boldsymbol{v}}_k \\ \hat{\boldsymbol{z}}_k \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{v}}_k^{\mathrm{T}} & \hat{\boldsymbol{z}}_k \end{pmatrix} \right\} = \begin{pmatrix} \underbrace{o_p(1)}_{K \times K} & \underbrace{o_P(1)}_{K \times (n-K)} \\ \underbrace{o_P(1)}_{(n-K) \times K} & \underbrace{\bar{\boldsymbol{Q}}^{\mathrm{T}} \hat{\boldsymbol{V}}^{-1/2} p^{-1} \boldsymbol{E}^{\mathrm{T}} \boldsymbol{E} \hat{\boldsymbol{V}}^{-1/2} \bar{\boldsymbol{Q}} + o_P(1)}_{(n-K) \times (n-K)} \end{pmatrix}
\tag{3.44}
$$

where $o_P(1)$ here is short for a rank K (i.e. K non-zero singular values) matrix whose 2-norm converges to 0 in probability.

Proof. We see that

$$\hat{\mu}_k \begin{pmatrix} \hat{\boldsymbol{v}}_k \\ \hat{\boldsymbol{z}}_k \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{v}}_k^{\mathrm{T}} & \hat{\boldsymbol{z}}_k \end{pmatrix} = \hat{\mu}_k \begin{pmatrix} \hat{\boldsymbol{v}}_k \hat{\boldsymbol{v}}_k^{\mathrm{T}} & \hat{\boldsymbol{v}}_k \hat{\boldsymbol{z}}_k^{\mathrm{T}} \\ \hat{\boldsymbol{z}}_k \hat{\boldsymbol{v}}_k^{\mathrm{T}} & \hat{\boldsymbol{z}}_k \hat{\boldsymbol{z}}_k^{\mathrm{T}} \end{pmatrix} \quad (k = 1, \dots, K).$$

First, let $\bar{\boldsymbol{N}} = \bar{\boldsymbol{L}} + \bar{\boldsymbol{E}}_1$. In (3.38) we defined $\boldsymbol{v}_k$ to be the k^{th} eigenvector of $\bar{\boldsymbol{N}}^{\mathrm{T}} \bar{\boldsymbol{N}}$. We then have that

$$\|\hat{\mu}_k \hat{\boldsymbol{v}}_k \hat{\boldsymbol{v}}_k^{\mathrm{T}} - \boldsymbol{\lambda}_k \left(\bar{\boldsymbol{N}}^{\mathrm{T}} \bar{\boldsymbol{N}} \right) \boldsymbol{v}_k \boldsymbol{v}_k^{\mathrm{T}} \|_2 = o_P \left(\gamma_K^{-1} \right),$$

meaning

$$\|\bar{\boldsymbol{N}}^{\mathrm{T}} \bar{\boldsymbol{N}} - \sum_{k=1}^{K} \hat{\mu}_k \hat{\boldsymbol{v}}_k \hat{\boldsymbol{v}}_k^{\mathrm{T}} \|_2 = o_P \left(\gamma_K^{-1} \right).$$

Next, we have

$$\begin{aligned} &\hat{\mu}_k \hat{\boldsymbol{z}}_k - \hat{\mu}_k \left(\hat{\mu}_k I_{n-K} - \bar{\boldsymbol{E}}_2^{\mathrm{T}} \bar{\boldsymbol{E}} \right)^{-1} \bar{\boldsymbol{E}}_2^{\mathrm{T}} \bar{\boldsymbol{N}} \hat{\boldsymbol{v}}_k = \frac{\hat{\mu}_k}{\hat{\mu}_k - \delta_*^2 / \gamma_K} \bar{\boldsymbol{E}}_2^{\mathrm{T}} \bar{\boldsymbol{N}} \hat{\boldsymbol{v}}_k \\ &+ O_P \left\{ \gamma_K^{-1} \left(n^{1/2} p^{-1/2} + \epsilon \right) \left(n^{1/2} p^{-1/2} \gamma_K^{-1/2} + \epsilon \gamma_K^{-1} \right) \right\} \\ =&\frac{\mu_k + \delta_*^2 / \gamma_K}{\mu_k} \bar{\boldsymbol{E}}_2^{\mathrm{T}} \bar{\boldsymbol{N}} \hat{\boldsymbol{v}}_k + o_P \left(\gamma_K^{-1} \right) = \bar{\boldsymbol{E}}_2^{\mathrm{T}} \bar{\boldsymbol{N}} \hat{\boldsymbol{v}}_k + o_P \left(\gamma_K^{-1} \right) \end{aligned}$$

where the second equality follows because $\bar{\boldsymbol{E}}_2^{\mathrm{T}} \bar{\boldsymbol{N}} = O_P \left\{ n^{1/2} \left(\gamma_K p \right)^{-1/2} + \epsilon / \gamma_K \right\}$ and

$$\hat{\mu}_k - \left(\mu_k + \delta_*^2 / \gamma_K \right) = O_P \left\{ n \left(p \gamma_K \right)^{-1} + n^{1/2} \left(\gamma_K p \right)^{-1/2} \epsilon + \left(\gamma_K p \right)^{-1/2} + \epsilon / \gamma_K \right\}.$$

Therefore,

$$\|\bar{\boldsymbol{E}}_2^{\mathrm{T}} \bar{\boldsymbol{N}} - \sum_{k=1}^{K} \hat{\mu}_k \hat{\boldsymbol{z}}_k \hat{\boldsymbol{v}}_k^{\mathrm{T}} \|_2 = \|\bar{\boldsymbol{E}}_2^{\mathrm{T}} \bar{\boldsymbol{N}} \left(I_K - \hat{\boldsymbol{v}} \hat{\boldsymbol{v}}^{\mathrm{T}} \right) \|_2 + o_P \left(\gamma_K^{-1} \right) = o_P \left(\gamma_K^{-1} \right).$$

Lastly, $\|\hat{\boldsymbol{z}}_k\|_2^2 = o_P\left(1/\gamma_K\right)$, which completes the proof. $\square$

We now prove a crucial lemma that states that we can accurately estimate $\boldsymbol{V}_*$ when the starting point is sufficiently close to $\boldsymbol{V}_*$.

***Lemma* 3.6.** *Suppose Assumptions 3.7 and 3.8 hold and let $k \geq K$. Define $\hat{\boldsymbol{V}}$ to be an initial estimate of $\boldsymbol{V}_*$ such that $\|\hat{\boldsymbol{V}} - \boldsymbol{V}_*\|_2 = \epsilon = o_P(1)$ and $\hat{\delta}^{(f)2}, \hat{\boldsymbol{V}}^{(f)}$ to be the estimates after we run step (1) and then (2) of Algorithm 3.2 using $\hat{\boldsymbol{V}}$ as a starting point. Then the maximum of step (2) is such that*

$$
\begin{aligned}
\|\hat{\boldsymbol{V}}^{(f)} - \boldsymbol{V}_*\|_2 &= O_P\left(n^{-1}\right)\\
\hat{\delta}^{(f)2} - \delta_*^2 &= O_P\left(n^{-1}\right).
\end{aligned}
$$

Proof. It suffices to prove this lemma by proving that $\|\delta^{(f)2}\hat{\boldsymbol{V}}^{(f)} - \delta_*^2\boldsymbol{V}_*\| = \mathcal{O}_P\left(n^{-1}\right)$, since $\delta^2 = \exp\left\{n^{-1}\log|\delta^2\boldsymbol{V}\left(\boldsymbol{\tau}\right)|\right\}$ is Lipschitz continuous in $\delta^2\boldsymbol{\tau}$ and bounded away from 0 in Θ_*. Therefore, we ignore the requirement that $\log|\boldsymbol{V}| = 0$ and re-define $\boldsymbol{\tau}$ to be $\boldsymbol{\tau} \leftarrow \delta^2\boldsymbol{\tau}$ in the proof of this lemma. We continue to use the objects $\bar{\boldsymbol{C}}$, $\bar{\boldsymbol{Q}}$ and $\hat{\bar{\boldsymbol{C}}}$ defined in (3.28a), (3.28b) and (3.40), and also define $\hat{\boldsymbol{Q}} = \boldsymbol{Q}_{\hat{\boldsymbol{C}}}$ and $\hat{\bar{\boldsymbol{Q}}} = \boldsymbol{Q}_{\hat{\bar{\boldsymbol{C}}}}$.

We first assume that $k = K$. Recall step *(1)* in Algorithm 3.2 is to estimate $\bar{\boldsymbol{C}}$, and step *(2)* computes $\hat{\boldsymbol{V}}^{(f)}$ as

$$
\begin{aligned}
\hat{\boldsymbol{V}}^{(f)} = \underset{\substack{\boldsymbol{V}=\boldsymbol{V}(\boldsymbol{\tau})\\ \boldsymbol{\tau}\in\Theta_*}}{\arg\max}\Big[&-\left(n-K\right)^{-1}\log|\hat{\boldsymbol{Q}}^{\mathrm{T}}\boldsymbol{V}\hat{\boldsymbol{Q}}|\\
&-\left(n-K\right)^{-1}\operatorname{tr}\left\{\hat{\boldsymbol{Q}}^{\mathrm{T}}p^{-1}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}\hat{\boldsymbol{Q}}\left(\hat{\boldsymbol{Q}}^{\mathrm{T}}\boldsymbol{V}\hat{\boldsymbol{Q}}\right)^{-1}\right\}\Big].
\end{aligned}
\tag{3.45}
$$

where Θ_* is defined in Assumption 3.8. We therefore need to understand how $\hat{\boldsymbol{Q}}^{\mathrm{T}}p^{-1}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}\hat{\boldsymbol{Q}}$

behaves. First, note that we can express $\hat{\hat{\boldsymbol{Q}}}$ in terms of $\hat{\boldsymbol{Q}}$:

$$\hat{\hat{\boldsymbol{Q}}} = \hat{\boldsymbol{V}}^{1/2}\hat{\boldsymbol{Q}}\left(\hat{\boldsymbol{Q}}^{\mathrm{T}}\hat{\boldsymbol{V}}\hat{\boldsymbol{Q}}\right)^{-1/2}.$$

Using the results of Corollary 3.3, we get that

$$\begin{aligned}
\hat{\boldsymbol{Q}}^{\mathrm{T}} p^{-1}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}\hat{\boldsymbol{Q}} &= \left(\hat{\boldsymbol{Q}}^{\mathrm{T}}\hat{\boldsymbol{V}}\hat{\boldsymbol{Q}}\right)^{1/2}\hat{\hat{\boldsymbol{Q}}}^{\mathrm{T}}\hat{\boldsymbol{V}}^{-1/2}p^{-1}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}\hat{\boldsymbol{V}}^{-1/2}\hat{\hat{\boldsymbol{Q}}}\left(\hat{\boldsymbol{Q}}^{\mathrm{T}}\hat{\boldsymbol{V}}\hat{\boldsymbol{Q}}\right)^{1/2}\\
&= \left(\hat{\boldsymbol{Q}}^{\mathrm{T}}\hat{\boldsymbol{V}}\hat{\boldsymbol{Q}}\right)^{1/2}\hat{\hat{\boldsymbol{Q}}}^{\mathrm{T}}\begin{pmatrix}\bar{\boldsymbol{C}} & \bar{\boldsymbol{Q}}\end{pmatrix}\left(\gamma_K\bar{\boldsymbol{S}}\right)\begin{pmatrix}\bar{\boldsymbol{C}}^{\mathrm{T}}\\ \bar{\boldsymbol{Q}}^{\mathrm{T}}\end{pmatrix}\hat{\hat{\boldsymbol{Q}}}\left(\hat{\boldsymbol{Q}}^{\mathrm{T}}\hat{\boldsymbol{V}}\hat{\boldsymbol{Q}}\right)^{1/2}\\
&= \left(\hat{\boldsymbol{Q}}^{\mathrm{T}}\hat{\boldsymbol{V}}\hat{\boldsymbol{Q}}\right)^{1/2}\hat{\hat{\boldsymbol{Q}}}^{\mathrm{T}}\begin{pmatrix}\bar{\boldsymbol{C}} & \bar{\boldsymbol{Q}}\end{pmatrix}\left\{\gamma_K\bar{\boldsymbol{S}} - \sum_{j=1}^{K}\gamma_K\hat{\mu}_j\begin{pmatrix}\hat{\boldsymbol{v}}_j\\ \hat{\boldsymbol{z}}_j\end{pmatrix}\begin{pmatrix}\hat{\boldsymbol{v}}_j^{\mathrm{T}} & \hat{\boldsymbol{z}}_j^{\mathrm{T}}\end{pmatrix}\right\}\\
&\quad\times\begin{pmatrix}\bar{\boldsymbol{C}}^{\mathrm{T}}\\ \bar{\boldsymbol{Q}}^{\mathrm{T}}\end{pmatrix}\hat{\hat{\boldsymbol{Q}}}\left(\hat{\boldsymbol{Q}}^{\mathrm{T}}\hat{\boldsymbol{V}}\hat{\boldsymbol{Q}}\right)^{1/2}\\
&= \left(\hat{\boldsymbol{Q}}^{\mathrm{T}}\hat{\boldsymbol{V}}\hat{\boldsymbol{Q}}\right)^{1/2}\hat{\hat{\boldsymbol{Q}}}^{\mathrm{T}}\bar{\boldsymbol{Q}}\left(\bar{\boldsymbol{Q}}^{\mathrm{T}}\hat{\boldsymbol{V}}^{-1/2}p^{-1}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\hat{\boldsymbol{V}}^{-1/2}\bar{\boldsymbol{Q}}\right)\bar{\boldsymbol{Q}}^{\mathrm{T}}\hat{\hat{\boldsymbol{Q}}}\left(\hat{\boldsymbol{Q}}^{\mathrm{T}}\hat{\boldsymbol{V}}\hat{\boldsymbol{Q}}\right)^{1/2}\\
&\quad+ \underbrace{o_P\left(1\right)}_{\text{rank } K \text{ matrix with this 2-norm}}. \qquad (3.46)
\end{aligned}$$

where the third equality is because

$$0 = \hat{\hat{\boldsymbol{Q}}}^{\mathrm{T}}\hat{\hat{\boldsymbol{C}}} = \left\{\hat{\hat{\boldsymbol{Q}}}^{\mathrm{T}}\left(\bar{\boldsymbol{C}}\,\bar{\boldsymbol{Q}}\right)\right\}\begin{pmatrix}\hat{\boldsymbol{v}}\\ \hat{\boldsymbol{z}}\end{pmatrix}.$$

Since the likelihood function in (3.45) is depends on $\hat{\boldsymbol{Q}}^{\mathrm{T}}p^{-1}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}\hat{\boldsymbol{Q}}$ through $1/(n-K)\,\mathrm{tr}\left(\cdot\right)$, the rank K matrix with Frobenius norm $o_P\left(1\right)$ will contribute $o_P\left(1/n\right)$ to the likelihood, score function and Hessian, and can therefore be ignored. Let σ_i be the i^{th} singular value of $\bar{\boldsymbol{Q}}^{\mathrm{T}}\hat{\hat{\boldsymbol{Q}}}$. Since $\bar{\boldsymbol{Q}}$ and $\hat{\hat{\boldsymbol{Q}}}$ have orthonormal columns, $\sigma_i \leq 1$. Further, the proof of Theorem

3.10 shows that

$$0 \leq \sum_{i=1}^{n-K}\left(1-\sigma_i\right) \leq \sum_{i=1}^{n-K}\left(1-\sigma_i^2\right) = \operatorname{tr}\left(I_{n-K}\right) - \operatorname{tr}\left(\bar{\boldsymbol{Q}}^{\mathrm{T}}\hat{\bar{\boldsymbol{Q}}}\hat{\bar{\boldsymbol{Q}}}^{\mathrm{T}}\bar{\boldsymbol{Q}}\right) = 2^{-1}\|P_{\bar{\boldsymbol{Q}}} - P_{\hat{\bar{\boldsymbol{Q}}}}\|_F^2$$
$$=2^{-1}\|P_{\bar{\boldsymbol{C}}} - P_{\hat{\bar{\boldsymbol{C}}}}\|_F^2 = o_P(1).$$

Next let $\boldsymbol{U}_1, \boldsymbol{U}_2 \in \mathbb{R}^{(n-K)\times(n-K)}$ be any unitary matrices. Then

$$\begin{aligned}
&\operatorname{tr}\left[\left(\hat{\boldsymbol{Q}}^{\mathrm{T}}\hat{\boldsymbol{V}}\hat{\boldsymbol{Q}}\right)^{1/2}\hat{\bar{\boldsymbol{Q}}}^{\mathrm{T}}\bar{\boldsymbol{Q}}\left(\bar{\boldsymbol{Q}}^{\mathrm{T}}\hat{\boldsymbol{V}}^{-1/2}p^{-1}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\hat{\boldsymbol{V}}^{-1/2}\bar{\boldsymbol{Q}}\right)\bar{\boldsymbol{Q}}^{\mathrm{T}}\hat{\bar{\boldsymbol{Q}}}\left(\hat{\boldsymbol{Q}}^{\mathrm{T}}\hat{\boldsymbol{V}}\hat{\boldsymbol{Q}}\right)^{1/2}\right.\\
&\left.\times\left\{\hat{\boldsymbol{Q}}^{\mathrm{T}}\boldsymbol{V}\left(\boldsymbol{\tau}\right)\hat{\boldsymbol{Q}}\right\}^{-1}\right] = \operatorname{tr}\left[\left(\boldsymbol{U}_1^{\mathrm{T}}\hat{\boldsymbol{Q}}^{\mathrm{T}}\hat{\boldsymbol{V}}\hat{\boldsymbol{Q}}\boldsymbol{U}_1\right)^{1/2}\boldsymbol{U}_1^{\mathrm{T}}\hat{\bar{\boldsymbol{Q}}}^{\mathrm{T}}\bar{\boldsymbol{Q}}\boldsymbol{U}_2\right.\\
&\times\left(\boldsymbol{U}_2^{\mathrm{T}}\bar{\boldsymbol{Q}}^{\mathrm{T}}\hat{\boldsymbol{V}}^{-1/2}p^{-1}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\hat{\boldsymbol{V}}^{-1/2}\bar{\boldsymbol{Q}}\boldsymbol{U}_2\right)\boldsymbol{U}_2^{\mathrm{T}}\bar{\boldsymbol{Q}}^{\mathrm{T}}\hat{\bar{\boldsymbol{Q}}}\boldsymbol{U}_1\left(\boldsymbol{U}_1^{\mathrm{T}}\hat{\boldsymbol{Q}}^{\mathrm{T}}\hat{\boldsymbol{V}}\hat{\boldsymbol{Q}}\boldsymbol{U}_1\right)^{1/2}\\
&\left.\times\left\{\boldsymbol{U}_1^{\mathrm{T}}\hat{\boldsymbol{Q}}^{\mathrm{T}}\boldsymbol{V}\left(\boldsymbol{\tau}\right)\hat{\boldsymbol{Q}}\boldsymbol{U}_1\right\}^{-1}\right].
\end{aligned}$$

In particular, we may choose $\boldsymbol{U}_1, \boldsymbol{U}_2$ to be the left and right singular vectors of $\hat{\bar{\boldsymbol{Q}}}^{\mathrm{T}}\bar{\boldsymbol{Q}}$, meaning it suffices to assume $\hat{\bar{\boldsymbol{Q}}}^{\mathrm{T}}\bar{\boldsymbol{Q}}$ is diagonal, since both $\bar{\boldsymbol{Q}}$ and $\hat{\bar{\boldsymbol{Q}}}$ are unique up to a rotation of their rows. And since the singular values of

$$\left(\bar{\boldsymbol{Q}}^{\mathrm{T}}\hat{\boldsymbol{V}}^{-1/2}p^{-1}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\hat{\boldsymbol{V}}^{-1/2}\bar{\boldsymbol{Q}}\right)\bar{\boldsymbol{Q}}^{\mathrm{T}}\hat{\bar{\boldsymbol{Q}}}\left(\hat{\boldsymbol{Q}}^{\mathrm{T}}\hat{\boldsymbol{V}}\hat{\boldsymbol{Q}}\right)^{1/2}\left\{\hat{\boldsymbol{Q}}^{\mathrm{T}}\boldsymbol{V}\left(\boldsymbol{\tau}\right)\hat{\boldsymbol{Q}}\right\}^{-1}\left(\hat{\boldsymbol{Q}}^{\mathrm{T}}\hat{\boldsymbol{V}}\hat{\boldsymbol{Q}}\right)^{1/2}$$

are uniformly bounded from above in probability for all $\boldsymbol{\tau} \in \boldsymbol{\Theta}_*$,

$$\begin{aligned}
&(n-K)^{-1}\operatorname{tr}\left[\left(\hat{\boldsymbol{Q}}^{\mathrm{T}}\hat{\boldsymbol{V}}\hat{\boldsymbol{Q}}\right)^{1/2}\hat{\bar{\boldsymbol{Q}}}^{\mathrm{T}}\bar{\boldsymbol{Q}}\left(\bar{\boldsymbol{Q}}^{\mathrm{T}}\hat{\boldsymbol{V}}^{-1/2}p^{-1}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\hat{\boldsymbol{V}}^{-1/2}\bar{\boldsymbol{Q}}\right)\bar{\boldsymbol{Q}}^{\mathrm{T}}\hat{\bar{\boldsymbol{Q}}}\left(\hat{\boldsymbol{Q}}^{\mathrm{T}}\hat{\boldsymbol{V}}\hat{\boldsymbol{Q}}\right)^{1/2}\right.\\
&\left.\times\left\{\hat{\boldsymbol{Q}}^{\mathrm{T}}\boldsymbol{V}\left(\boldsymbol{\tau}\right)\hat{\boldsymbol{Q}}\right\}^{-1}\right]\\
=&(n-K)^{-1}\operatorname{tr}\left[\left(\hat{\boldsymbol{Q}}^{\mathrm{T}}\hat{\boldsymbol{V}}\hat{\boldsymbol{Q}}\right)^{1/2}\left(\bar{\boldsymbol{Q}}^{\mathrm{T}}\hat{\boldsymbol{V}}^{-1/2}p^{-1}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\hat{\boldsymbol{V}}^{-1/2}\bar{\boldsymbol{Q}}\right)\left(\hat{\boldsymbol{Q}}^{\mathrm{T}}\hat{\boldsymbol{V}}\hat{\boldsymbol{Q}}\right)^{1/2}\right.\\
&\left.\times\left\{\hat{\boldsymbol{Q}}^{\mathrm{T}}\boldsymbol{V}\left(\boldsymbol{\tau}\right)\hat{\boldsymbol{Q}}\right\}^{-1}\right] + o_P\left(n^{-1}\right).
\end{aligned}$$

Next, we can re-write $\bar{\boldsymbol{Q}}$ as

$$\bar{\boldsymbol{Q}} = \left(\hat{\bar{\boldsymbol{Q}}}\hat{\bar{\boldsymbol{Q}}}^{\mathrm{T}} + \hat{\bar{\boldsymbol{C}}}\hat{\bar{\boldsymbol{C}}}^{\mathrm{T}}\right)\bar{\boldsymbol{Q}} = \hat{\bar{\boldsymbol{Q}}}\left(\hat{\bar{\boldsymbol{Q}}}^{\mathrm{T}}\bar{\boldsymbol{Q}}\right) + \underbrace{o_P\left(1\right)}_{\text{rank } K \text{ matrix with this 2-norm}}.$$

Therefore,

$$\begin{aligned}
&(n-K)^{-1}\operatorname{tr}\left[\left(\hat{\boldsymbol{Q}}^{\mathrm{T}}\hat{\boldsymbol{V}}\hat{\boldsymbol{Q}}\right)^{1/2}\left(\bar{\boldsymbol{Q}}^{\mathrm{T}}\hat{\boldsymbol{V}}^{-1/2}p^{-1}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\hat{\boldsymbol{V}}^{-1/2}\bar{\boldsymbol{Q}}\right)\left(\hat{\boldsymbol{Q}}^{\mathrm{T}}\hat{\boldsymbol{V}}\hat{\boldsymbol{Q}}\right)^{1/2}\right.\\
&\left.\times\left\{\hat{\boldsymbol{Q}}^{\mathrm{T}}\boldsymbol{V}\left(\boldsymbol{\tau}\right)\hat{\boldsymbol{Q}}\right\}^{-1}\right] = (n-K)^{-1}\operatorname{tr}\left[\left(\hat{\boldsymbol{Q}}^{\mathrm{T}}\hat{\boldsymbol{V}}\hat{\boldsymbol{Q}}\right)^{1/2}\hat{\bar{\boldsymbol{Q}}}^{\mathrm{T}}\hat{\boldsymbol{V}}^{-1/2}\left(p^{-1}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\right)\hat{\boldsymbol{V}}^{-1/2}\hat{\bar{\boldsymbol{Q}}}\right.\\
&\left.\times\left(\hat{\boldsymbol{Q}}^{\mathrm{T}}\hat{\boldsymbol{V}}\hat{\boldsymbol{Q}}\right)^{1/2}\left\{\hat{\boldsymbol{Q}}^{\mathrm{T}}\boldsymbol{V}\left(\boldsymbol{\tau}\right)\hat{\boldsymbol{Q}}\right\}^{-1}\right] + o_P\left(n^{-1}\right)\\
&= (n-K)^{-1}\operatorname{tr}\left[\hat{\boldsymbol{Q}}^{\mathrm{T}}\left(p^{-1}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\right)\hat{\boldsymbol{Q}}\left\{\hat{\boldsymbol{Q}}^{\mathrm{T}}\boldsymbol{V}\left(\boldsymbol{\tau}\right)\hat{\boldsymbol{Q}}\right\}^{-1}\right] + o_P\left(n^{-1}\right).
\end{aligned}$$

where the last equality follows because $\hat{\bar{\boldsymbol{Q}}}$ and $\hat{\boldsymbol{Q}}$ satisfy $\hat{\bar{\boldsymbol{Q}}} = \hat{\boldsymbol{V}}^{1/2}\hat{\boldsymbol{Q}}\left(\hat{\boldsymbol{Q}}^{\mathrm{T}}\hat{\boldsymbol{V}}\hat{\boldsymbol{Q}}\right)^{-1/2}$. The likelihood function in (3.45) (which is $O_P(1)$) can then be re-written up to factors that are $o_P(1/n)$ as

$$\hat{l}\left(\boldsymbol{\tau}\right) = -\left(n-K\right)^{-1}\log|\hat{\boldsymbol{Q}}^{\mathrm{T}}\boldsymbol{V}\left(\boldsymbol{\tau}\right)\hat{\boldsymbol{Q}}| - (n-K)^{-1}\operatorname{tr}\left[\hat{\boldsymbol{Q}}^{\mathrm{T}}\left(p^{-1}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\right)\hat{\boldsymbol{Q}}\left\{\hat{\boldsymbol{Q}}^{\mathrm{T}}\boldsymbol{V}\left(\boldsymbol{\tau}\right)\hat{\boldsymbol{Q}}\right\}^{-1}\right]$$

We will now show that $\hat{l}\left(\boldsymbol{\tau}\right) - l\left(\boldsymbol{\tau}\right) = O_P(1/n)$, where

$$l\left(\boldsymbol{\tau}\right) = -n^{-1}\log|\boldsymbol{V}\left(\boldsymbol{\tau}\right)| - n^{-1}\operatorname{tr}\left\{\left(p^{-1}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\right)\boldsymbol{V}\left(\boldsymbol{\tau}\right)^{-1}\right\}. \tag{3.47}$$

First,

$$\begin{aligned}
\log|\boldsymbol{V}| &= \log|\hat{\boldsymbol{Q}}^{\mathrm{T}}\boldsymbol{V}\hat{\boldsymbol{Q}}| + \log|\hat{\boldsymbol{C}}^{\mathrm{T}}\boldsymbol{V}\hat{\boldsymbol{C}} - \hat{\boldsymbol{C}}^{\mathrm{T}}\boldsymbol{V}\hat{\boldsymbol{Q}}\left(\hat{\boldsymbol{Q}}^{\mathrm{T}}\boldsymbol{V}\hat{\boldsymbol{Q}}\right)^{-1}\hat{\boldsymbol{Q}}^{\mathrm{T}}\boldsymbol{V}\hat{\boldsymbol{C}}|\\
&= \log|\hat{\boldsymbol{Q}}^{\mathrm{T}}\boldsymbol{V}\hat{\boldsymbol{Q}}| + \log|\left(\hat{\boldsymbol{C}}^{\mathrm{T}}\boldsymbol{V}^{-1}\hat{\boldsymbol{C}}\right)^{-1}| = \log|\hat{\boldsymbol{Q}}^{\mathrm{T}}\boldsymbol{V}\hat{\boldsymbol{Q}}| + O(1).
\end{aligned}$$

where we have abused notation here and defined $\hat{\boldsymbol{C}}$ to be a $n\times K$ matrix with orthonormal columns that is orthogonal to $\hat{\boldsymbol{Q}}$. The second equality follows from the fact that

$$I_n - \boldsymbol{V}^{1/2}\hat{\boldsymbol{Q}}\left(\hat{\boldsymbol{Q}}^{\mathrm{T}}\boldsymbol{V}\hat{\boldsymbol{Q}}\right)^{-1}\hat{\boldsymbol{Q}}^{\mathrm{T}}\boldsymbol{V}^{1/2} = \boldsymbol{V}^{-1/2}\hat{\boldsymbol{C}}^{\mathrm{T}}\left(\hat{\boldsymbol{C}}^{\mathrm{T}}\boldsymbol{V}^{-1}\hat{\boldsymbol{C}}\right)^{-1}\hat{\boldsymbol{C}}\boldsymbol{V}^{-1/2}.$$

Let $\boldsymbol{S} = p^{-1}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}$. Then

$$\boldsymbol{S} = \hat{\boldsymbol{Q}}\hat{\boldsymbol{Q}}^{\mathrm{T}}\boldsymbol{S}\hat{\boldsymbol{Q}}\hat{\boldsymbol{Q}}^{\mathrm{T}} + \text{rank } K \text{ matrix with eigenvalues that are } O_P(1).$$

Therefore,

$$l\left(\boldsymbol{\tau}\right) = -\left(n-K\right)^{-1}\log|\hat{\boldsymbol{Q}}^{\mathrm{T}}\boldsymbol{V}\hat{\boldsymbol{Q}}| - \left(n-K\right)^{-1}\operatorname{tr}\left[\hat{\boldsymbol{Q}}^{\mathrm{T}}\boldsymbol{S}\hat{\boldsymbol{Q}}\hat{\boldsymbol{Q}}^{\mathrm{T}}\boldsymbol{V}^{-1}\hat{\boldsymbol{Q}}\right] + O_P\left(n^{-1}\right).$$

Further, we can re-write $\boldsymbol{V}^{-1}$ as

$$\begin{aligned}\boldsymbol{V}^{-1} &= \begin{bmatrix}\hat{\boldsymbol{Q}} & \hat{\boldsymbol{C}}\end{bmatrix}\begin{bmatrix}\hat{\boldsymbol{Q}}^{\mathrm{T}}\boldsymbol{V}\hat{\boldsymbol{Q}} & \hat{\boldsymbol{Q}}^{\mathrm{T}}\boldsymbol{V}\hat{\boldsymbol{C}}\\ \hat{\boldsymbol{C}}^{\mathrm{T}}\boldsymbol{V}\hat{\boldsymbol{Q}} & \hat{\boldsymbol{C}}^{\mathrm{T}}\boldsymbol{V}\hat{\boldsymbol{C}}\end{bmatrix}^{-1}\begin{bmatrix}\hat{\boldsymbol{Q}}^{\mathrm{T}}\\ \hat{\boldsymbol{C}}^{\mathrm{T}}\end{bmatrix}\\ &= \hat{\boldsymbol{Q}}\left(\hat{\boldsymbol{Q}}^{\mathrm{T}}\boldsymbol{V}\hat{\boldsymbol{Q}}\right)^{-1}\hat{\boldsymbol{Q}}^{\mathrm{T}} + \text{Matrix of rank } K \text{ with bounded singular values.}\end{aligned} \tag{3.48}$$

We then get that

$$l\left(\boldsymbol{\tau}\right) = -\left(n-K\right)^{-1}\log|\hat{\boldsymbol{Q}}^{\mathrm{T}}\boldsymbol{V}\hat{\boldsymbol{Q}}| - \left(n-K\right)^{-1}\operatorname{tr}\left\{\hat{\boldsymbol{Q}}^{\mathrm{T}}\boldsymbol{S}\hat{\boldsymbol{Q}}\left(\hat{\boldsymbol{Q}}^{\mathrm{T}}\boldsymbol{V}\hat{\boldsymbol{Q}}\right)^{-1}\right\} + O_P\left(n^{-1}\right),$$

as desired.

Define the function

$$h_n\left(\boldsymbol{\tau}\right) = -n^{-1}\log|\boldsymbol{V}\left(\boldsymbol{\tau}\right)| - n^{-1}\operatorname{tr}\left[\boldsymbol{V}_*\left\{\boldsymbol{V}\left(\boldsymbol{\tau}\right)\right\}^{-1}\right].$$

Since the eigenvalues of $\boldsymbol{V}(\boldsymbol{\tau})$ are uniformly bounded from above and below on Θ_*, we have $\sup_{\boldsymbol{x}\in\boldsymbol{\Theta}_*}|\hat{l}^{(n)}(\boldsymbol{x})-l^{(n)}(\boldsymbol{x})| = O_P\left(n^{-1}\right)$ (the superscript (n) is to indicate this is with sample size

n). And since $\sup_{\boldsymbol{x}\in\Theta_*} |h_n(\boldsymbol{x}) - l^{(n)}(\boldsymbol{x})| = O_P\left(n^{1/2}p^{-1/2}\right) = o_P(1)$, $\sup_{\boldsymbol{x}\in\Theta_*} |h_n(\boldsymbol{x}) - \hat{l}^{(n)}(\boldsymbol{x})| = o_P(1)$. We can then apply Lemma 3.3 to show that $\|\hat{\boldsymbol{\tau}}^f - \boldsymbol{\tau}_*\|_2 = o_P(1)$.

Let $i, j \in [b]$. To find the rate, we simply use an identical analysis to write the score and Hessian of the objective function in (3.45) up to terms that are $O_P(1/n)$, which gives us

$$
\begin{aligned}
\boldsymbol{s}^{(n)}(\boldsymbol{\tau})_j &= \frac{d}{d\boldsymbol{\tau}_j} l^{(n)}(\boldsymbol{\tau}) = -n^{-1}\operatorname{tr}\left\{\boldsymbol{V}(\boldsymbol{\tau})^{-1}\boldsymbol{B}_j\right\} + n^{-1}\operatorname{tr}\left\{\boldsymbol{V}(\boldsymbol{\tau})^{-1}\boldsymbol{B}_j\boldsymbol{V}(\boldsymbol{\tau})^{-1}\boldsymbol{S}\right\}\\
\hat{\boldsymbol{s}}^{(n)}(\boldsymbol{\tau})_j &= \frac{d}{d\boldsymbol{\tau}_j} \hat{l}^{(n)}(\boldsymbol{\tau}) = -n^{-1}\operatorname{tr}\left\{\hat{\boldsymbol{A}}(\boldsymbol{\tau})^{-1}\hat{\boldsymbol{B}}_j\right\} + n^{-1}\operatorname{tr}\left\{\hat{\boldsymbol{A}}(\boldsymbol{\tau})^{-1}\hat{\boldsymbol{B}}_j\hat{\boldsymbol{A}}(\boldsymbol{\tau})^{-1}\hat{\boldsymbol{Q}}^{\mathrm{T}}\boldsymbol{S}\hat{\boldsymbol{Q}}\right\}\\
&\quad + O_P\left(n^{-1}\right)\\
H^{(n)}(\boldsymbol{\tau})_{ij} &= \frac{d^2}{d\boldsymbol{\tau}_j d\boldsymbol{\tau}_i} l^{(n)}(\boldsymbol{\tau}) = n^{-1}\operatorname{tr}\left\{\boldsymbol{V}(\boldsymbol{\tau})^{-1}\boldsymbol{B}_i\boldsymbol{V}(\boldsymbol{\tau})^{-1}\boldsymbol{B}_j\right\}\\
&\quad - 2n^{-1}\operatorname{tr}\left\{\boldsymbol{V}(\boldsymbol{\tau})^{-1}\boldsymbol{B}_i\boldsymbol{V}(\boldsymbol{\tau})^{-1}\boldsymbol{B}_j\boldsymbol{V}(\boldsymbol{\tau})^{-1}\boldsymbol{S}\right\}\\
\hat{H}^{(n)}(\boldsymbol{\tau})_{ij} &= \frac{d^2}{d\boldsymbol{\tau}_j d\boldsymbol{\tau}_i} \hat{l}^{(n)}(\boldsymbol{\tau}) = n^{-1}\operatorname{tr}\left\{\hat{\boldsymbol{A}}(\boldsymbol{\tau})^{-1}\hat{\boldsymbol{B}}_i\hat{\boldsymbol{A}}(\boldsymbol{\tau})^{-1}\hat{\boldsymbol{B}}_j\right\}\\
&\quad - 2n^{-1}\operatorname{tr}\left\{\hat{\boldsymbol{A}}(\boldsymbol{\tau})^{-1}\hat{\boldsymbol{B}}_i\hat{\boldsymbol{A}}(\boldsymbol{\tau})^{-1}\hat{\boldsymbol{B}}_j\hat{\boldsymbol{A}}(\boldsymbol{\tau})^{-1}\hat{\boldsymbol{Q}}^{\mathrm{T}}\boldsymbol{S}\hat{\boldsymbol{Q}}\right\} + O_P\left(n^{-1}\right)
\end{aligned}
$$

where $\hat{\boldsymbol{B}}_j = \hat{\boldsymbol{Q}}^{\mathrm{T}}\boldsymbol{B}_j\hat{\boldsymbol{Q}}$ and $\hat{\boldsymbol{A}} = \hat{\boldsymbol{Q}}^{\mathrm{T}}\boldsymbol{V}\hat{\boldsymbol{Q}}$. An identical analysis as the one used to show the likelihoods differ by $O_P\left(n^{-1}\right)$ shows that $\|\boldsymbol{s}^{(n)} - \hat{\boldsymbol{s}}^{(n)}\|_2, \|\boldsymbol{H}^{(n)} - \hat{\boldsymbol{H}}^{(n)}\|_2 = O_P\left(n^{-1}\right)$, where for $i, j \in [b]$,

$$
\left[\boldsymbol{H}^{(n)}(\boldsymbol{\tau}_*)\right]_{ij} = -n^{-1}\operatorname{tr}\left\{\boldsymbol{V}_*^{-1}\boldsymbol{B}_i\boldsymbol{V}_*^{-1}\boldsymbol{B}_j\right\} + o_P(1)
$$

and for $\boldsymbol{M}_{ij} = n^{-1}\operatorname{tr}\left\{\boldsymbol{V}_*^{-1}\boldsymbol{B}_i\boldsymbol{V}_*^{-1}\boldsymbol{B}_j\right\}$, $\boldsymbol{M} \succeq \epsilon I_b$ for some $\epsilon > 0$ by Assumptions 3.7 and 3.8. Let $\hat{\boldsymbol{\tau}}$ be the solution to (3.45) that ignores the inequality and equality constraints defined by $\boldsymbol{A}_{\mathcal{I}}$ and $\boldsymbol{A}_{\mathcal{E}}$. Then for $\bar{\boldsymbol{\tau}} = \lambda\boldsymbol{\tau}_* + (1-\lambda)\hat{\boldsymbol{\tau}}$ for some $\lambda \in [0,1]$,

$$\begin{aligned}
\mathbf{0} = \hat{\boldsymbol{s}}^{(n)}\left(\hat{\boldsymbol{\tau}}\right) &= \hat{\boldsymbol{s}}^{(n)}\left(\boldsymbol{\tau}_*\right) + \hat{\boldsymbol{H}}^{(n)}\left(\bar{\boldsymbol{\tau}}\right)\left(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}_*\right) \\
&= \boldsymbol{s}^{(n)}\left(\boldsymbol{\tau}_*\right) + \boldsymbol{H}^{(n)}\left(\bar{\boldsymbol{\tau}}\right)\left(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}_*\right) + O_P\left(n^{-1}\right) \\
&= O_P\left\{(np)^{-1/2} + n^{-1}\right\} + \boldsymbol{H}^{(n)}\left(\bar{\boldsymbol{\tau}}\right)\left(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}_*\right),
\end{aligned}$$

meaning $\|\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}_*\|_2 = O_P\left(1/n\right)$. Since $\boldsymbol{\tau}_* \in \Theta_*$, we then have

$$\begin{aligned}
\hat{l}^{(n)}\left(\hat{\boldsymbol{\tau}}\right) \geq \hat{l}^{(n)}\left(\hat{\boldsymbol{\tau}}^f\right) \geq \hat{l}^{(n)}\left(\boldsymbol{\tau}_*\right) &= \hat{l}^{(n)}\left(\hat{\boldsymbol{\tau}}\right) + 2^{-1}\left(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}_*\right)^{\mathrm{T}} \hat{\boldsymbol{H}}^{(n)}\left(\bar{\boldsymbol{\tau}}\right)\left(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}_*\right) \\
&= \hat{l}^{(n)}\left(\hat{\boldsymbol{\tau}}\right) + O_P\left(n^{-2}\right).
\end{aligned}$$

Since the b eigenvalues of $\hat{\boldsymbol{H}}^{(n)}\left(\bar{\boldsymbol{\tau}}\right)$ are uniformly bounded away from 0 with probability tending to 1 for any $\bar{\boldsymbol{\tau}}$ such that $\|\bar{\boldsymbol{\tau}} - \boldsymbol{\tau}_*\|_2 = o_P(1)$, $\|\hat{\boldsymbol{\tau}} - \hat{\boldsymbol{\tau}}^f\|_2 = O_P\left(1/n\right)$. Therefore, $\|\hat{\boldsymbol{\tau}}^f - \boldsymbol{\tau}_*\|_2 = O_P(1/n)$ when $K = k$.

When $k > K$, define $\hat{\boldsymbol{Q}}_k \in \mathbb{R}^{n\times(n-k)}$ to be the analogue of $\hat{\boldsymbol{Q}}$ above. Since $\hat{\boldsymbol{Q}}_k$ is orthogonal to $\hat{\boldsymbol{C}} \in \mathbb{R}^{n\times K}$, its image must be a proper subspace of the image of $\hat{\boldsymbol{Q}}_K$. Therefore, $\hat{\boldsymbol{Q}}_k = \hat{\boldsymbol{Q}}_K\boldsymbol{U}_k$ for some matrix $\boldsymbol{U}_k \in \mathbb{R}^{(n-K)\times(n-k)}$ with orthonormal columns. By what we have shown above and because $k - K$ is at most finite,

$$\begin{aligned}
\hat{l}\left(\boldsymbol{\tau}\right) =& -(n-k)^{-1}\log|\boldsymbol{U}_k^{\mathrm{T}}\hat{\boldsymbol{Q}}_K^{\mathrm{T}}\boldsymbol{V}\left(\boldsymbol{\tau}\right)\hat{\boldsymbol{Q}}_K\boldsymbol{U}_k| \\
&- (n-k)^{-1}\operatorname{tr}\left[\boldsymbol{U}_k^{\mathrm{T}}\hat{\boldsymbol{Q}}_K^{\mathrm{T}}\left(p^{-1}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}\right)\hat{\boldsymbol{Q}}_K\boldsymbol{U}_k\left\{\boldsymbol{U}_k^{\mathrm{T}}\hat{\boldsymbol{Q}}_K^{\mathrm{T}}\boldsymbol{V}\left(\boldsymbol{\tau}\right)\hat{\boldsymbol{Q}}_K\boldsymbol{U}_k\right\}^{-1}\right] \\
=& -(n-k)^{-1}\log|\boldsymbol{U}_k^{\mathrm{T}}\hat{\boldsymbol{Q}}_K^{\mathrm{T}}\boldsymbol{V}\left(\boldsymbol{\tau}\right)\hat{\boldsymbol{Q}}_K\boldsymbol{U}_k| \\
&- (n-k)^{-1}\operatorname{tr}\left[\boldsymbol{U}_k^{\mathrm{T}}\hat{\boldsymbol{Q}}_K^{\mathrm{T}}\left(p^{-1}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\right)\hat{\boldsymbol{Q}}_K\boldsymbol{U}_k\left\{\boldsymbol{U}_k^{\mathrm{T}}\hat{\boldsymbol{Q}}_K^{\mathrm{T}}\boldsymbol{V}\left(\boldsymbol{\tau}\right)\hat{\boldsymbol{Q}}_K\boldsymbol{U}_k\right\}^{-1}\right] \\
&+ O_P\left(n^{-1}\right) = (n-K)^{-1}\log|\hat{\boldsymbol{Q}}_K^{\mathrm{T}}\boldsymbol{V}\left(\boldsymbol{\tau}\right)\hat{\boldsymbol{Q}}_K| \\
&- (n-K)^{-1}\operatorname{tr}\left[\hat{\boldsymbol{Q}}_K^{\mathrm{T}}\left(p^{-1}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\right)\hat{\boldsymbol{Q}}_K\left\{\hat{\boldsymbol{Q}}_K^{\mathrm{T}}\boldsymbol{V}\left(\boldsymbol{\tau}\right)\hat{\boldsymbol{Q}}_K\right\}^{-1}\right] + O_P\left(n^{-1}\right).
\end{aligned}$$

An identical analysis can be used to show that the score function and Hessian matrix when $k > K$ differ from the score function and Hessian matrix when $k = K$ by something at most $O_P\left(n^{-1}\right)$, which proves the claim when $k > K$. $\square$

We can now prove Theorem 3.7.

Proof of Theorem 3.7. Again, for values δ^2 and $\boldsymbol{\tau}$ such that $\delta^2\boldsymbol{\tau} \in \Theta_*$, we redefine $\boldsymbol{\tau} \leftarrow \delta^2\boldsymbol{\tau}$ for notational convenience to prove $\|\boldsymbol{\tau}_* - \hat{\boldsymbol{\tau}}\|_2 = O_P(1/n)$, which will prove (3.22a) by the analysis in the first paragraph of the proof of Lemma 3.6. Results (3.22b) and (3.22c) will follow by Corollary 3.2 because the column space of $\boldsymbol{C}_\perp$ and $\hat{\boldsymbol{C}}_\perp$ is invariant to scalar multiplication.

For any value of $\hat{\boldsymbol{V}}$, define $\epsilon = \|\hat{\boldsymbol{V}} - \boldsymbol{V}_*\|_2$. We use a similar technique as we did in the proof of Lemma 3.6 and work with the $\operatorname{tr}(\cdot)$ component in the likelihood given by (3.45), where

$$p^{-1}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y} = p^{-1}\boldsymbol{C}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{C} + p^{-1}\boldsymbol{C}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{E} + \left(p^{-1}\boldsymbol{C}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{E}\right)^{\mathrm{T}} + p^{-1}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}. \tag{3.49}$$

Suppose first that $\gamma_1 = o(n)$. Then for any matrix $\hat{\boldsymbol{Q}} \in \mathbb{R}^{n\times k}$ with orthonormal columns (k bounded from above),

$$\begin{aligned} n^{-1}\operatorname{tr}\left\{\hat{\boldsymbol{Q}}^{\mathrm{T}}\left(p^{-1}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}\right)\hat{\boldsymbol{Q}}\left(\hat{\boldsymbol{Q}}^{\mathrm{T}}\boldsymbol{V}\hat{\boldsymbol{Q}}\right)^{-1}\right\} =& n^{-1}\operatorname{tr}\left\{\hat{\boldsymbol{Q}}^{\mathrm{T}}\left(p^{-1}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\right)\hat{\boldsymbol{Q}}\left(\hat{\boldsymbol{Q}}^{\mathrm{T}}\boldsymbol{V}\hat{\boldsymbol{Q}}\right)^{-1}\right\} \\ &+ o_P(1) \end{aligned}$$

since the rank k matrix $\boldsymbol{C}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{E}$ has maximum singular value

$$\|p^{-1}\boldsymbol{C}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{E}\|_2 = O_P\left\{(\gamma_1 n)^{1/2}\, p^{-1/2}\right\}$$

and $\|p^{-1}\boldsymbol{C}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{C}\|_2 = O(\gamma_1)$. Therefore, $\epsilon = o_P(1)$ for all estimates of $\boldsymbol{V}_*$ when $\gamma_1 = o(n)$, meaning all conditions of Theorem 3.10 and Lemma 3.6 will be satisfied on the first iteration of Algorithm 3.2, meaning $\epsilon = O_P\left(n^{-1}\right)$ in step *(2)* of the first iteration when $k = K$. Therefore, the estimate for $\boldsymbol{C}$ in the second iteration will satisfy (3.29), (3.30) and (3.31) from Theorem 3.10 and (3.41), (3.42) for Corollary 3.2 with $\epsilon = O_P\left(n^{-1}\right)$ when $k = K$. And since we use $\hat{\boldsymbol{V}}$ as a starting point for $k+1$, $\epsilon = O_P\left(n^{-1}\right)$ for all subsequent $k's$ $(k \geq K)$.

Next, suppose $\gamma_K \asymp n$ and $k = K$. Then after step *(1)* in the first iteration, we have for $\bar{\boldsymbol{C}}$ defined in (3.28a) and $\hat{\boldsymbol{v}} = \begin{bmatrix}\hat{\boldsymbol{v}}_1 & \cdots & \hat{\boldsymbol{v}}_K\end{bmatrix}$, $\hat{\boldsymbol{z}} = \begin{bmatrix}\hat{\boldsymbol{z}}_1 & \cdots & \hat{\boldsymbol{z}}_K\end{bmatrix}$ defined in (3.38),

$$\begin{aligned}
\|n^{-1/2}\hat{\boldsymbol{Q}}^{\mathrm{T}}\boldsymbol{C}\|_2 &= \|n^{-1/2}\hat{\boldsymbol{Q}}^{\mathrm{T}}\hat{\boldsymbol{V}}^{1/2}\hat{\boldsymbol{V}}^{-1/2}\boldsymbol{C}\|_2 = O\left(\|\hat{\bar{\boldsymbol{Q}}}^{\mathrm{T}}\bar{\boldsymbol{C}}\|_2\right) = O\left(\|\hat{\bar{\boldsymbol{Q}}}^{\mathrm{T}}\left(\bar{\boldsymbol{C}} - \hat{\bar{\boldsymbol{C}}}\hat{\boldsymbol{v}}^{\mathrm{T}}\right)\|_2\right)\\
&= O\left(\|\hat{\bar{\boldsymbol{Q}}}^{\mathrm{T}}\left\{\bar{\boldsymbol{C}}\left(I_K - \hat{\boldsymbol{v}}\hat{\boldsymbol{v}}^{\mathrm{T}}\right) + \bar{\boldsymbol{Q}}\hat{\boldsymbol{z}}\hat{\boldsymbol{v}}^{\mathrm{T}}\right\}\|_2\right)\\
&= O_P\left\{n^{1/2}\left(\gamma_K p\right)^{-1/2} + \epsilon\gamma_K^{-1}\right\}.
\end{aligned}$$

When completing step *(2)* of the first iteration to re-estimate $\boldsymbol{V}_*$, we have

$$\begin{aligned}
n^{-1}\operatorname{tr}\left\{\hat{\boldsymbol{Q}}^{\mathrm{T}}\left(p^{-1}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}\right)\hat{\boldsymbol{Q}}\left(\hat{\boldsymbol{Q}}^{\mathrm{T}}\boldsymbol{V}\hat{\boldsymbol{Q}}\right)^{-1}\right\} =& n^{-1}\operatorname{tr}\left\{\hat{\boldsymbol{Q}}^{\mathrm{T}}\left(p^{-1}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\right)\hat{\boldsymbol{Q}}\left(\hat{\boldsymbol{Q}}^{\mathrm{T}}\boldsymbol{V}\hat{\boldsymbol{Q}}\right)^{-1}\right\}\\
&+ O_P\left(n^{-1}\right),
\end{aligned}$$

meaning $\epsilon = O_P\left(n^{-1}\right)$ after this step. Therefore, step *(1)* of the second iteration will give us an estimate of $\boldsymbol{C}$ that satisfies (3.29), (3.30) and (3.31) from Theorem 3.10 and (3.41), (3.42) for Corollary 3.2 with $\epsilon = O_P\left(n^{-1}\right)$. This is also true when $k > K$ by Lemma 3.6, since we use the previous estimate of $\boldsymbol{V}_*$ as a starting point for step *(1)*. Equation (3.22c) then follows because

$$\begin{aligned}
\|P_C - P_{\hat{C}_k}P_C\|_F^2 &= \operatorname{tr}\left(P_C\right) + \operatorname{tr}\left(P_{\hat{C}_k}P_C\right) - 2\operatorname{tr}\left(P_{\hat{C}_k}P_C\right) = K - \operatorname{tr}\left(P_{\hat{C}_k}P_C\right)\\
&\leq K - \operatorname{tr}\left(P_{\hat{C}_K}P_C\right) = K - \operatorname{tr}\left\{\boldsymbol{C}^{\mathrm{T}}\hat{\boldsymbol{C}}_K\left(\hat{\boldsymbol{C}}_K^{\mathrm{T}}\hat{\boldsymbol{C}}_K\right)^{-1}\hat{\boldsymbol{C}}_K^{\mathrm{T}}\boldsymbol{C}\left(\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C}\right)^{-1}\right\}\\
&= O_P\left\{n\left(\gamma_K p\right)^{-1} + \left(\gamma_K p\right)^{-1/2} + n^{-1}\gamma_K^{-1}\right\}
\end{aligned}$$

where the column space of $\hat{\boldsymbol{C}}_K \in \mathbb{R}^{n\times K}$ is a subspace of the column space of $\hat{\boldsymbol{C}}_k \in \mathbb{R}^{n\times k}$, since $k \geq K$. The final equality follows from Corollary 3.2.

Lastly, let $\boldsymbol{Q}'_X \in \mathbb{R}^{n\times(n-d)}$ be another matrix whose columns form an orthonormal basis for the null space of $\boldsymbol{X}$. Then $\boldsymbol{Q}'_X = \boldsymbol{Q}_X \boldsymbol{U}$ for some unitary matrix $\boldsymbol{U} \in \mathbb{R}^{(n-d)\times(n-d)}$. Define $\boldsymbol{C}'_\perp = \left(\boldsymbol{Q}'_X\right)^{\mathrm{T}} \boldsymbol{C} = \boldsymbol{U}^{\mathrm{T}} \boldsymbol{C}_\perp$ and $\hat{\boldsymbol{C}}'_\perp$ to be the estimate obtained by using $\boldsymbol{Q}'_X$ at any point in the procedure. It is easy to show that $\hat{\boldsymbol{C}}'_\perp = \boldsymbol{U}^{\mathrm{T}} \hat{\boldsymbol{C}}_\perp$, where $\hat{\boldsymbol{C}}_\perp$ is the estimate obtained using $\boldsymbol{Q}_X$. This proves $\|P_{\boldsymbol{C}_\perp} - P_{\hat{\boldsymbol{C}}_\perp}\|_F^2$ and $\|P_{\boldsymbol{C}_\perp} - P_{\hat{\boldsymbol{C}}_\perp} P_{\boldsymbol{C}_\perp}\|_F^2$ are invariant of the choice of $\boldsymbol{Q}_X$. Lastly, using restricted maximum likelihood to estimate $\delta_*^2, \boldsymbol{\tau}$ implies the estimates only depend on the column space of $\left[\boldsymbol{X}\, \boldsymbol{Q}_X \hat{\boldsymbol{C}}_\perp\right]$. And since $\boldsymbol{Q}'_X \hat{\boldsymbol{C}}'_\perp = \boldsymbol{Q}_X \hat{\boldsymbol{C}}_\perp$, this completes the proof. □

Proof of Theorem 3.8

We use Theorem 3.7 to prove Theorem 3.8. Just as we did in Section 3.13, we assume $\boldsymbol{Y}$ is generated according to model (3.2) with $d = 0$ and will use the incorrect, but simpler model in (3.7) with $d = 0$.

We first show that for any fold f and $k \leq K_{\max}$, the loss function in (3.23) only depends on the column space of $\hat{\hat{\boldsymbol{C}}} = \hat{\boldsymbol{V}}_{(-f)}^{-1/2} \hat{\boldsymbol{C}}$.

Lemma 3.7. *For any fold f and $k \leq K_{\max}$, the loss function in* (3.23) *only depends on $\hat{\boldsymbol{V}}_{(-f)}$ and the column space of $\hat{\hat{\boldsymbol{C}}} = \hat{\boldsymbol{V}}_{(-f)}^{-1/2} \hat{\boldsymbol{C}}$.*

Proof. Suppose $\hat{\hat{\boldsymbol{C}}}^{(1)}$ and $\hat{\hat{\boldsymbol{C}}}^{(2)}$ have the same column space. Then $\hat{\hat{\boldsymbol{C}}}^{(1)} \boldsymbol{R} = \hat{\hat{\boldsymbol{C}}}^{(2)}$ for some invertible matrix $\boldsymbol{R}$. Therefore, $\hat{\hat{\boldsymbol{C}}}^{(1)}$ and $\hat{\hat{\boldsymbol{C}}}^{(2)}$ have the same leverage scores. Let $\hat{\hat{\boldsymbol{c}}}_i^{(r)}$ be the ith row of $\hat{\hat{\boldsymbol{C}}}^{(r)}$ for $r = 1, 2$. If we use $\hat{\hat{\boldsymbol{C}}}^{(2)}$ to determine the loss,

$$
\begin{aligned}
\hat{\boldsymbol{L}}_{f,(-i)}^{(2)} \hat{\hat{\boldsymbol{c}}}_i^{(2)} &= \bar{\boldsymbol{Y}}_{f,(-i)} \hat{\hat{\boldsymbol{C}}}_{(-i)}^{(2)} \left[\left\{ \hat{\hat{\boldsymbol{C}}}_{(-i)}^{(2)} \right\}^{\mathrm{T}} \hat{\hat{\boldsymbol{C}}}_{(-i)}^{(2)} \right]^{-1} \hat{\hat{\boldsymbol{c}}}_i^{(2)} \\
&= \bar{\boldsymbol{Y}}_{f,(-i)} \hat{\hat{\boldsymbol{C}}}_{(-i)}^{(1)} \boldsymbol{R} \boldsymbol{R}^{-1} \left[\left\{ \hat{\hat{\boldsymbol{C}}}_{(-i)}^{(1)} \right\}^{\mathrm{T}} \hat{\hat{\boldsymbol{C}}}_{(-i)}^{(1)} \right]^{-1} \boldsymbol{R}^{-T} \boldsymbol{R}^{\mathrm{T}} \hat{\hat{\boldsymbol{c}}}_i^{(1)} = \hat{\boldsymbol{L}}_{f,(-i)}^{(1)} \hat{\hat{\boldsymbol{c}}}_i^{(1)}
\end{aligned}
$$

which completes the proof. $\square$

For any fold f and $k \leq K_{\max}$, $\hat{\boldsymbol{V}}_{(-f)}$ and the column space of $\hat{\hat{\boldsymbol{C}}} \in \mathbb{R}^{n \times k}$, as estimated in Algorithm 3.2, are invariant the choice of parametrization of $\hat{\boldsymbol{C}}$ (and therefore $\hat{\hat{\boldsymbol{C}}}$). Therefore, it suffices to use the output of Algorithm 3.2 to for $k = 1, \ldots, K_{\max}$ to estimate K, i.e. $n^{-1/2}\hat{\hat{\boldsymbol{C}}}$ are the first k right singular values of $\bar{\boldsymbol{Y}}_{(-f)} = \boldsymbol{Y}_{(-f)}\hat{\boldsymbol{V}}_{(-f)}^{-1/2}$.

Proof of Theorem 3.8. Fix some fold f and define δ_{f*}^2 and $\boldsymbol{V}_{f*}$ to be the analogues of δ_*^2 and $\boldsymbol{V}_*$ for fold f. Let $\boldsymbol{\pi} : [p] \to [p]$ be a permutation sampled uniformly from the set of all permutations on $[p]$. All conditional expectations and variances calculated below are with reference to the sigma algebra $\sigma\left(\boldsymbol{Y}_{(-f)}, \boldsymbol{\pi}\right)$, where $\boldsymbol{Y}_f = \boldsymbol{L}_f\boldsymbol{C}^{\mathrm{T}} + \boldsymbol{E}_f \in \mathbb{R}^{p_f \times n}$ is the test data used to estimate $\boldsymbol{L}_f$ and evaluate the loss and $\boldsymbol{Y}_{(-f)} = \boldsymbol{L}_{(-f)}\boldsymbol{C}^{\mathrm{T}} + \boldsymbol{E}_{(-f)} \in \mathbb{R}^{(p-p_f)\times n}$ is the training data used to estimate $\boldsymbol{C}$ and $\boldsymbol{V}_*$.

We first prove that the kth non-zero eigenvalue of $\boldsymbol{V}_{f*}^{-1/2}\boldsymbol{C}\left(p^{-1}\boldsymbol{L}_f^{\mathrm{T}}\boldsymbol{L}_f\right)\boldsymbol{C}^{\mathrm{T}}\boldsymbol{V}_{f*}^{-1/2}$ is $\gamma_k\left\{1 + o_P(1)\right\}$ as $n \to \infty$, where the randomness is in the choice of permutation π to partition the rows of $\boldsymbol{Y}$ into F folds. To do so, it suffices to assume that $n^{-1}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{V}_*^{-1}\boldsymbol{C} = I_K$ and $np^{-1}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L} = \mathrm{diag}\left(\gamma_1, \ldots, \gamma_K\right)$. Note that the non-zero eigenvalues of

$$\gamma_K^{-1}\boldsymbol{V}_{f*}^{-1/2}\boldsymbol{C}\left(p^{-1}\boldsymbol{L}_f^{\mathrm{T}}\boldsymbol{L}_f\right)\boldsymbol{C}^{\mathrm{T}}\boldsymbol{V}_{f*}^{-1/2}$$

are also the eigenvalues of $\left\{n\left(\gamma_K p\right)^{-1}\boldsymbol{L}_f^{\mathrm{T}}\boldsymbol{L}_f\right\}\left(n^{-1}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{V}_{f*}^{-1}\boldsymbol{C}\right)$, where because $\|\boldsymbol{V}_{f*} - \boldsymbol{V}_*\|_2 = O_P\left(p^{-1/2}\right)$,

$$\|n^{-1}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{V}_{f*}^{-1}\boldsymbol{C} - I_K\|_2 = O_P\left(p^{-1/2}\right).$$

We also have

$$\mathbb{E}_{\boldsymbol{\pi}}\left(\frac{n}{\gamma_K p_f}\boldsymbol{L}_f^{\mathrm{T}}\boldsymbol{L}_f\right) = \frac{n}{\gamma_K p}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L},$$

and for $r, s \in [K]$ and some constant $c > 0$,

$$\begin{aligned}
\mathbb{V}\left\{\frac{n}{\gamma_K p_f}\sum_{g=1}^{p_f}\left[\boldsymbol{\ell}_{\boldsymbol{\pi}(g)}\right]_r\left[\boldsymbol{\ell}_{\boldsymbol{\pi}(g)}\right]_s\right\} &\leq \sum_{g=1}^{p_f}\mathbb{V}\left\{\frac{n}{\gamma_K p_f}\left[\boldsymbol{\ell}_{\boldsymbol{\pi}(g)}\right]_r\left[\boldsymbol{\ell}_{\boldsymbol{\pi}(g)}\right]_s\right\} \\
&\leq \left(\frac{n}{\gamma_K p_f}\right)^2\sum_{g=1}^{p_f}\mathbb{E}\left\{\left[\boldsymbol{\ell}_{\boldsymbol{\pi}(g)}\right]_r^2\left[\boldsymbol{\ell}_{\boldsymbol{\pi}(g)}\right]_s^2\right\} \\
&\leq \left(\frac{n}{\gamma_K p_f}\right)^2 c^2\sum_{g=1}^{p_f}\mathbb{E}\left\{\left[\boldsymbol{\ell}_{\boldsymbol{\pi}(g)}\right]_r^2\right\} \\
&= \frac{nc^2}{\gamma_K p^2}\sum_{g=1}^{p_f}\mathbb{E}\left\{\frac{n}{\gamma_K}\left[\boldsymbol{\ell}_{\boldsymbol{\pi}(g)}\right]_r^2\right\} \\
&= \frac{nc^2\left\{n\left(\gamma_K p\right)^{-1}\left[\boldsymbol{L}\right]_{*r}^{\mathrm{T}}\left[\boldsymbol{L}\right]_{*r}\right\}p_f}{\gamma_K p^2} \\
&\to 0 \text{ as } n \to \infty,
\end{aligned}$$

where the first inequality follows from the fact that $\left[\boldsymbol{\ell}_g\right]_r\left[\boldsymbol{\ell}_g\right]_s$ is being sampled without replacement from a finite population, meaning successive draws are negatively correlated, and the third inequality follows because the magnitude entries of $\boldsymbol{L}$ are uniformly bounded by a constant $c > 0$ by Assumption 3.7. Therefore, the eigenvalues of $n/\left(\gamma_K p_f\right)\boldsymbol{L}_f^{\mathrm{T}}\boldsymbol{L}_f$ are $\gamma_k\gamma_K^{-1} + o_P(1)$ as $n \to \infty$ ($k = 1, \ldots, K$), meaning the results of Theorems 3.7 and 3.10, as well as Lemma 3.6, hold for any subset of p_f rows of $\boldsymbol{Y}$ that are chosen uniformly at random (where $p_f/p \asymp 1$) with probability tending to 1 as $n \to \infty$.

Let ϕ be the maximum leverage score of $\boldsymbol{V}_*^{-1/2}\boldsymbol{C}$ and define $\alpha = (1-\phi)^{-1}$. We note that by the proofs of Theorem 3.10 and Lemma 3.6, the leverage scores $\hat{\hat{h}}_i$ are such that $\max_{i\in[n]}\hat{\hat{h}}_i = \phi + o_P(1)$ and $\max_{i\in[n]}\left(1-\hat{\hat{h}}_i\right)^{-1} = \alpha + o_P(1)$ when $k = K$. Therefore, to prove the theorem it suffices to assume that $\max_{i\in[n]}\hat{\hat{h}}_i < 1 - \eta/m_n$ for all $k \leq K_{\max}$.

Let $\bar{\boldsymbol{C}} = \hat{\boldsymbol{V}}_{(-f)}^{-1/2}\boldsymbol{C} \in \mathbb{R}^{n\times K}$ and define $\bar{\boldsymbol{c}}_i$ to be the ith row of $\bar{\boldsymbol{C}}$. We define $g(k)$ to be

$$g(k) = \frac{1}{\delta_{f*}^2 n p_f}\mathcal{L}(k) = \frac{1}{\delta_{f*}^2 n p_f}\sum_{i=1}^{n}\|\bar{\boldsymbol{y}}_{f,i} - \hat{\boldsymbol{L}}_{f,(-i)}\hat{\bar{\boldsymbol{c}}}_i\|_2^2 = \underbrace{\frac{1}{\delta_{f*}^2 n p_f}\sum_{i=1}^{n}\|\boldsymbol{L}_f\bar{\boldsymbol{c}}_i - \hat{\boldsymbol{L}}_{f,(-i)}\hat{\bar{\boldsymbol{c}}}_i\|_2^2}_{(1)}$$
$$+ \underbrace{\frac{1}{\delta_{f*}^2 n}\operatorname{tr}\left\{\hat{\boldsymbol{V}}_{(-f)}^{-1}\left(p_f^{-1}\boldsymbol{E}_f^{\mathrm{T}}\boldsymbol{E}_f\right)\right\}}_{(2)} - \underbrace{\frac{2}{\delta_{f*}^2 n}\sum_{i=1}^{n} p_f^{-1}\left(\boldsymbol{L}_f\bar{\boldsymbol{c}}_i - \hat{\boldsymbol{L}}_{f,(-i)}\hat{\bar{\boldsymbol{c}}}_i\right)^{\mathrm{T}}\boldsymbol{E}_f\hat{\boldsymbol{V}}_{(-f)}^{-1/2}\boldsymbol{a}_i}_{(3)}\,. \tag{3.50}$$

where $\boldsymbol{a}_i \in \mathbb{R}^n$ is the standard basis vector with 1 in the ith position and zeros everywhere else. Two useful equalities to be used throughout the proof are

$$\left\{\hat{\bar{\boldsymbol{C}}}_{(-i)}^{\mathrm{T}}\hat{\bar{\boldsymbol{C}}}_{(-i)}\right\}^{-1}\hat{\bar{\boldsymbol{c}}}_i = \frac{1}{1-\hat{\bar{h}}_i}\left(\hat{\bar{\boldsymbol{C}}}^{\mathrm{T}}\hat{\bar{\boldsymbol{C}}}\right)^{-1}\hat{\bar{\boldsymbol{c}}}_i$$
$$\hat{\bar{\boldsymbol{C}}}\left\{\hat{\bar{\boldsymbol{C}}}_{(-i)}^{\mathrm{T}}\hat{\bar{\boldsymbol{C}}}_{(-i)}\right\}^{-1}\hat{\bar{\boldsymbol{c}}}_i = \frac{1}{1-\hat{\bar{h}}_i}\hat{\bar{\boldsymbol{H}}}_i$$

where $\hat{\bar{\boldsymbol{H}}}_i$ is the i^{th} column of $P_{\hat{\bar{\boldsymbol{C}}}} = \hat{\bar{\boldsymbol{H}}}$. We also define $\boldsymbol{A}_{(-i)} = I_n - \boldsymbol{a}_i\boldsymbol{a}_i^{\mathrm{T}} \in \mathbb{R}^{n\times n}$, and let $\hat{\boldsymbol{v}}_{(-f)}$, $\hat{\boldsymbol{z}}_{(-f)}$ be the analogues of $\hat{\boldsymbol{v}}$, $\hat{\boldsymbol{z}}$ defined in (3.36), (3.37) for $\boldsymbol{Y}_{(-f)}$. The asymptotic properties of $\hat{\boldsymbol{v}}_{(-f)}$ and $\hat{\boldsymbol{z}}_{(-f)}$ are given in (3.38) and (3.39). For each $k = 1,\ldots,K_{\max}$, we assume that $n^{-1}\bar{\boldsymbol{C}}^{\mathrm{T}}\bar{\boldsymbol{C}} = I_K$ and that $\boldsymbol{L}_{(-f)}^{\mathrm{T}}\boldsymbol{L}_{(-f)}$ is diagonal with non-increasing elements. This is without loss of generality by Assumption 3.8 and by the identifiability of $\boldsymbol{L}\boldsymbol{C}^{\mathrm{T}}$. Note that by the above analysis, this implies $\gamma_K^{-1}\|\frac{n}{p_{(-f)}}\boldsymbol{L}_{(-f)}^{\mathrm{T}}\boldsymbol{L}_{(-f)} - \frac{n}{p_f}\boldsymbol{L}_f^{\mathrm{T}}\boldsymbol{L}_f\|_2 = o_P(1)$ as $n\to\infty$. We now go through (1), (2) and (3) of (3.50) to expand each expression into terms with calculable conditional expectations and variances to prove the claim.

(1)

$$\frac{1}{\delta_{f*}^2 n p_f}\sum_{i=1}^n \|\boldsymbol{L}_f\bar{\boldsymbol{c}}_i - \hat{\boldsymbol{L}}_{f,(-i)}\hat{\bar{\boldsymbol{c}}}_i\|_2^2 = \frac{1}{\delta_{f*}^2 n p_f}\sum_{i=1}^n \|\frac{\boldsymbol{L}_f\bar{\boldsymbol{c}}_i - \boldsymbol{L}_f\bar{\boldsymbol{C}}^{\mathrm{T}}\hat{\bar{\boldsymbol{C}}}\left(\hat{\bar{\boldsymbol{C}}}^{\mathrm{T}}\hat{\bar{\boldsymbol{C}}}\right)^{-1}\hat{\bar{\boldsymbol{c}}}_i}{\left(1-\hat{\bar{h}}_i\right)}\|_2^2$$
$$+ n^{-1}\sum_{i=1}^n \left(\frac{1}{1-\hat{\bar{h}}_i}\right)^2 \hat{\bar{\boldsymbol{H}}}_i^{\mathrm{T}}\boldsymbol{A}_{(-i)}\hat{\boldsymbol{V}}_{(-f)}^{-1/2}\left(\delta_{f*}^{-2}p_f^{-1}\boldsymbol{E}_f^{\mathrm{T}}\boldsymbol{E}_f\right)\hat{\boldsymbol{V}}_{(-f)}^{-1/2}\boldsymbol{A}_{(-i)}\hat{\bar{\boldsymbol{H}}}_i$$
$$- \frac{2}{\delta_{f*}^2 n^{3/2}p_f^{1/2}}\sum_{i=1}^n \left(\frac{1}{1-\hat{\bar{h}}_i}\right)^2 \hat{\bar{\boldsymbol{H}}}_i^{\mathrm{T}}\boldsymbol{A}_{(-i)}\hat{\boldsymbol{V}}_{(-f)}^{-1/2}\boldsymbol{E}_f^{\mathrm{T}}\boldsymbol{\Delta}_i \tag{3.51}$$

where

$$\boldsymbol{\Delta}_i = \left(n^{1/2}p^{-1/2}\boldsymbol{L}_f\right)\bar{\boldsymbol{c}}_i - \left(n^{1/2}p^{-1/2}\boldsymbol{L}_f\right)\bar{\boldsymbol{C}}^{\mathrm{T}}\hat{\bar{\boldsymbol{C}}}\left(\hat{\bar{\boldsymbol{C}}}^{\mathrm{T}}\hat{\bar{\boldsymbol{C}}}\right)^{-1}\hat{\bar{\boldsymbol{c}}}_i$$
$$= \tilde{\boldsymbol{L}}_f\bar{\boldsymbol{c}}_i - \tilde{\boldsymbol{L}}_f\bar{\boldsymbol{C}}^{\mathrm{T}}\hat{\bar{\boldsymbol{C}}}\left(\hat{\bar{\boldsymbol{C}}}^{\mathrm{T}}\hat{\bar{\boldsymbol{C}}}\right)^{-1}\hat{\bar{\boldsymbol{c}}}_i. \tag{3.52}$$

We now derive the asymptotic properties of (3.51) when $\underline{k=K}$, $\underline{k<K}$ and $\underline{k>K}$ in (a), (b) and (c) below.

(a) $\underline{k=K}$,

$$\frac{1}{\delta_{f*}^2 n p_f}\sum_{i=1}^n \|\frac{\boldsymbol{L}_f\bar{\boldsymbol{c}}_i - \boldsymbol{L}_f\bar{\boldsymbol{C}}^{\mathrm{T}}\hat{\bar{\boldsymbol{C}}}\left(\hat{\bar{\boldsymbol{C}}}^{\mathrm{T}}\hat{\bar{\boldsymbol{C}}}\right)^{-1}\hat{\bar{\boldsymbol{c}}}_i}{\left(1-\hat{\bar{h}}_i\right)}\|_2^2$$
$$= \delta_{f*}^{-2}\operatorname{tr}\left\{p_f^{-1}\boldsymbol{L}_f^{\mathrm{T}}\boldsymbol{L}_f\left(n^{-1}\bar{\boldsymbol{C}}^{\mathrm{T}}P_{\hat{\bar{\boldsymbol{C}}}}^{\perp}\bar{\boldsymbol{C}}\right)\right\}\{\alpha+o_P(1)\}$$
$$\underbrace{=}_{\text{Theorem 3.10}} \delta_{f*}^{-2}\operatorname{tr}\left[p_f^{-1}\boldsymbol{L}_f^{\mathrm{T}}\boldsymbol{L}_f\left\{I_K - \hat{\boldsymbol{v}}_{(-f)}\hat{\boldsymbol{v}}_{(-f)}^{\mathrm{T}}\right\}\right]\{\alpha+o_P(1)\}$$
$$\underbrace{=}_{\substack{\text{Theorem 3.10 and}\\ \text{Lemma 3.6}}} O_P\left(p^{-1}+n^{-2}\right) = o_P\left(n^{-1}\right).$$

To understand the second term,

$$
\begin{aligned}
&\mathbb{E}\left\{n^{-1}\sum_{i=1}^{n}\left(\frac{1}{1-\hat{\hat{h}}_i}\right)^2 \hat{\hat{\boldsymbol{H}}}_i^{\mathrm{T}}\boldsymbol{A}_{(-i)}\hat{\boldsymbol{V}}_{(-f)}^{-1/2}\left(\delta_{f*}^{-2}p_f^{-1}\boldsymbol{E}_f^{\mathrm{T}}\boldsymbol{E}_f\right)\hat{\boldsymbol{V}}_{(-f)}^{-1/2}\boldsymbol{A}_{(-i)}\hat{\hat{\boldsymbol{H}}}_i\right.\\
&\left. \mid \boldsymbol{Y}_{(-f)},\boldsymbol{\pi}\right\} = n^{-1}\sum_{i=1}^{n}\left(\frac{1}{1-\hat{\hat{h}}_i}\right)^2 \hat{\hat{\boldsymbol{H}}}_i^{\mathrm{T}}\boldsymbol{A}_{(-i)}\hat{\boldsymbol{V}}_{(-f)}^{-1/2}\boldsymbol{V}_{f*}\hat{\boldsymbol{V}}_{(-f)}^{-1/2}\boldsymbol{A}_{(-i)}\hat{\hat{\boldsymbol{H}}}_i\\
=&\frac{\alpha+o_P(1)}{n}\sum_{i=1}^{n}\left(\frac{1}{1-\hat{\hat{h}}_i}\right)\hat{\hat{\boldsymbol{H}}}_i^{\mathrm{T}}\boldsymbol{A}_{(-i)}\hat{\boldsymbol{V}}_{(-f)}^{-1/2}\boldsymbol{V}_{f*}\hat{\boldsymbol{V}}_{(-f)}^{-1/2}\boldsymbol{A}_{(-i)}\hat{\hat{\boldsymbol{H}}}_i\\
=&\frac{\alpha+o_P(1)}{n}\operatorname{tr}\left\{\hat{\boldsymbol{V}}_{(-f)}^{-1/2}\boldsymbol{V}_{f*}\hat{\boldsymbol{V}}_{(-f)}^{-1/2}\sum_{i=1}^{n}\frac{1}{1-\hat{\hat{h}}_i}\boldsymbol{A}_{(-i)}\hat{\hat{\boldsymbol{H}}}_i\hat{\hat{\boldsymbol{H}}}_i^{\mathrm{T}}\boldsymbol{A}_{(-i)}\right\}
\end{aligned}
$$

where $\sum_{i=1}^{n}\left(1-\hat{\hat{h}}_i\right)^{-1}\boldsymbol{A}_{(-i)}\hat{\hat{\boldsymbol{H}}}_i\hat{\hat{\boldsymbol{H}}}_i^{\mathrm{T}}\boldsymbol{A}_{(-i)}$ is positive semi-definite with

$$
\begin{aligned}
&\operatorname{tr}\left\{\sum_{i=1}^{n}\left(1-\hat{\hat{h}}_i\right)^{-1}\boldsymbol{A}_{(-i)}\hat{\hat{\boldsymbol{H}}}_i\hat{\hat{\boldsymbol{H}}}_i^{\mathrm{T}}\boldsymbol{A}_{(-i)}\right\}\\
=&\operatorname{tr}\left\{\sum_{i=1}^{n}\left(1-\hat{\hat{h}}_i\right)^{-1}\left(\hat{\hat{\boldsymbol{H}}}_i-\hat{\hat{h}}_i\boldsymbol{a}_i\right)\left(\hat{\hat{\boldsymbol{H}}}_i-\hat{\hat{h}}_i\boldsymbol{a}_i\right)^{\mathrm{T}}\right\}=\sum_{i=1}^{n}\hat{\hat{h}}_i=K.
\end{aligned}
$$

Therefore,

$$
n^{-1}\sum_{i=1}^{n}\left(\frac{1}{1-\hat{\hat{h}}_i}\right)^2 \hat{\hat{\boldsymbol{H}}}_i^{\mathrm{T}}\boldsymbol{A}_{(-i)}\hat{\boldsymbol{V}}_{(-f)}^{-1/2}\boldsymbol{V}_{f*}\hat{\boldsymbol{V}}_{(-f)}^{-1/2}\boldsymbol{A}_{(-i)}\hat{\hat{\boldsymbol{H}}}_i = K/n\left\{\alpha+o_P(1)\right\}
$$

when $\underline{k=K}$. To then calculate the conditional variance of the second term in (3.51), we see that the second term can be re-written as

$$
\begin{aligned}
&\left(np_f\delta_{f*}^2\right)^{-1}\operatorname{tr}\left[\boldsymbol{E}_f\hat{\boldsymbol{V}}_{(-f)}^{-1/2}\left\{\sum_{i=1}^{n}\left(1-\hat{\hat{h}}_i\right)^{-2}\boldsymbol{A}_{(-i)}\hat{\hat{\boldsymbol{H}}}_i\hat{\hat{\boldsymbol{H}}}_i^{\mathrm{T}}\boldsymbol{A}_{(-i)}\right\}\hat{\boldsymbol{V}}_{(-f)}^{-1/2}\boldsymbol{E}_f^{\mathrm{T}}\right]\\
=&\left(np_f\delta_{f*}^2\right)^{-1}\sum_{g=1}^{p_f}\bar{\boldsymbol{e}}_{f_g}^{\mathrm{T}}\hat{\boldsymbol{M}}_{g,(-f)}\bar{\boldsymbol{e}}_{f_g}\\
&\bar{\boldsymbol{e}}_{f_g}\sim N_n\left(\mathbf{0},I_n\right)
\end{aligned}
$$

where

$$\hat{\boldsymbol{M}}_{g,(-f)} = \boldsymbol{V}_g^{1/2}\hat{\boldsymbol{V}}_{(-f)}^{-1/2}\left\{\sum_{i=1}^{n}\left(1-\hat{\bar{h}}_i\right)^{-2}\boldsymbol{A}_{(-i)}\hat{\bar{\boldsymbol{H}}}_i\hat{\bar{\boldsymbol{H}}}_i^{\mathrm{T}}\boldsymbol{A}_{(-i)}\right\}\hat{\boldsymbol{V}}_{(-f)}^{-1/2}\boldsymbol{V}_g^{1/2} \succeq \boldsymbol{0}$$

and $\operatorname{tr}\left\{\hat{\boldsymbol{M}}_{g,(-f)}\right\} \leq cK\max_{i\in[n]}\left(1-\hat{\bar{h}}_i\right)^{-1}$ where $c > 0$ is a constant and $\max_{i\in[n]}\left(1-\hat{\bar{h}}_i\right)^{-1} = \alpha + o_P(1)$. For any $g \in [p]$,

$$\mathbb{V}\left\{\bar{\boldsymbol{e}}_{f_g}^{\mathrm{T}}\hat{\boldsymbol{M}}_{g,(-f)}\bar{\boldsymbol{e}}_{f_g} \mid \boldsymbol{Y}_{(-f)}, \boldsymbol{\pi}\right\} = 2\operatorname{tr}\left\{\hat{\boldsymbol{M}}_{g,(-f)}^2\right\} \leq 2\operatorname{tr}\left\{\hat{\boldsymbol{M}}_{g,(-f)}\right\}^2.$$

Therefore,

$$\operatorname{SD}\left\{n^{-1}\sum_{i=1}^{n}\left(\frac{1}{1-\hat{\bar{h}}_i}\right)^2\hat{\bar{\boldsymbol{H}}}_i^{\mathrm{T}}\boldsymbol{A}_{(-i)}\hat{\boldsymbol{V}}_{(-f)}^{-1/2}\left(\delta_{f*}^{-2}p_f^{-1}\boldsymbol{E}_f^{\mathrm{T}}\boldsymbol{E}_f\right)\hat{\boldsymbol{V}}_{(-f)}^{-1/2}\boldsymbol{A}_{(-i)}\hat{\bar{\boldsymbol{H}}}_i \right.$$
$$\left. \mid \boldsymbol{Y}_{(-f)}, \boldsymbol{\pi}\right\} = O_P\left(n^{-1}p^{-1/2}\right) = o_P\left(n^{-1}\right)$$

when $\underline{k = K}$. We lastly need to understand the third term in (3.51) when $\underline{k = K}$. This term clearly has conditional expectation equal to 0. To calculate the conditional variance, we can use the proof of Theorem 3.10 to show that

$$\boldsymbol{\Delta}_i = \tilde{\boldsymbol{L}}_f\left\{I_K - \hat{\boldsymbol{v}}_{(-f)}\hat{\boldsymbol{v}}_{(-f)}^{\mathrm{T}}\right\}\bar{\boldsymbol{c}}_i - n^{1/2}\tilde{\boldsymbol{L}}_f\hat{\boldsymbol{v}}_{(-f)}\hat{\boldsymbol{z}}_{(-f)}^{\mathrm{T}}\bar{\boldsymbol{q}}_i$$

where $\bar{\boldsymbol{q}}_i$ is the i^{th} row of $\boldsymbol{Q}_{\bar{\boldsymbol{C}}}$. By (3.38) and (3.39),

$$\|\tilde{\boldsymbol{L}}_f\left\{I_K - \hat{\boldsymbol{v}}_{(-f)}\hat{\boldsymbol{v}}_{(-f)}^{\mathrm{T}}\right\}\|_2 = O_P\left\{np^{-1}\gamma_K^{-1/2} + \gamma_K^{-3/2}n^{-2}\right\} \tag{3.53a}$$
$$\|\tilde{\boldsymbol{L}}_f\hat{\boldsymbol{v}}_{(-f)}\hat{\boldsymbol{z}}_{(-f)}^{\mathrm{T}}\|_2 = O_P\left(n^{1/2}p^{-1/2} + n^{-1}\gamma_K^{-1/2}\right). \tag{3.53b}$$

We can then re-write the third term in (3.51) as

$$
\begin{aligned}
t_3 = & -2\delta_{f*}^{-2} n^{-3/2} p^{-1/2} \sum_{g=1}^{p} \left[\tilde{\boldsymbol{\ell}}_g^{\mathrm{T}} \left\{ I_K - \hat{\boldsymbol{v}}_{(-f)} \hat{\boldsymbol{v}}_{(-f)}^{\mathrm{T}} \right\} \hat{\boldsymbol{N}}_{(-f)} \boldsymbol{e}_{f_g} \right. \\
& \left. + \tilde{\boldsymbol{\ell}}_g^{\mathrm{T}} \hat{\boldsymbol{v}}_{(-f)} \hat{\boldsymbol{z}}_{(-f)}^{\mathrm{T}} \hat{\boldsymbol{S}}_{(-f)} \boldsymbol{e}_{f_g} \right]
\end{aligned}
$$

where

$$
\begin{aligned}
\hat{\boldsymbol{N}}_{(-f)} = & \sum_{i=1}^{n} \left(1 - \hat{\bar{h}}_i\right)^{-2} \bar{\boldsymbol{c}}_i \hat{\bar{\boldsymbol{H}}}_i^{\mathrm{T}} \boldsymbol{A}_{(-i)} \hat{\boldsymbol{V}}_{(-f)}^{-1/2} = \bar{\boldsymbol{C}}^{\mathrm{T}} \operatorname{diag} \left\{ \left(1 - \hat{\bar{h}}_i\right)^{-2} \right\} \hat{\bar{\boldsymbol{H}}} \hat{\boldsymbol{V}}_{(-f)}^{-1/2} \\
& - \bar{\boldsymbol{C}}^{\mathrm{T}} \operatorname{diag} \left\{ \hat{\bar{h}}_i \left(1 - \hat{\bar{h}}_i\right)^{-2} \right\} \hat{\boldsymbol{V}}_{(-f)}^{-1/2} \\
\hat{\boldsymbol{S}}_{(-f)} = & n^{1/2} \sum_{i=1}^{n} \left(1 - \hat{\bar{h}}_i\right)^{-2} \bar{\boldsymbol{q}}_i \hat{\bar{\boldsymbol{H}}}_i^{\mathrm{T}} \boldsymbol{A}_{(-i)} \hat{\boldsymbol{V}}_{(-f)}^{-1/2} \\
= & n^{1/2} \boldsymbol{Q}_{\bar{\boldsymbol{C}}}^{\mathrm{T}} \operatorname{diag} \left\{ \left(1 - \hat{\bar{h}}_i\right)^{-2} \right\} \hat{\bar{\boldsymbol{H}}} \hat{\boldsymbol{V}}_{(-f)}^{-1/2} \\
& - n^{1/2} \boldsymbol{Q}_{\bar{\boldsymbol{C}}}^{\mathrm{T}} \operatorname{diag} \left\{ \hat{\bar{h}}_i \left(1 - \hat{\bar{h}}_i\right)^{-2} \right\} \hat{\boldsymbol{V}}_{(-f)}^{-1/2}
\end{aligned}
$$

First, $\|\hat{\boldsymbol{N}}_{(-f)}\|_2, \|\hat{\boldsymbol{S}}_{(-f)}\|_2 \leq c n^{1/2} \max_{i \in [n]} \left\{ \left(1 - \hat{\bar{h}}_i\right)^{-2} \right\}$, where $c > 0$ is some constant not dependent on n or p and $\hat{\bar{M}} = \max_{i \in [n]} \left\{ \left(1 - \hat{\bar{h}}_i\right)^{-2} \right\} = \alpha^2 + o_P(1)$. Next,

$$
\begin{aligned}
\mathbb{V} \left\{ \tilde{\boldsymbol{\ell}}_g^{\mathrm{T}} \hat{\boldsymbol{v}}_{(-f)} \hat{\boldsymbol{z}}_{(-f)}^{\mathrm{T}} \hat{\boldsymbol{S}}_{(-f)} \boldsymbol{e}_{f_g} \mid \boldsymbol{Y}_{(-f)}, \boldsymbol{\pi} \right\} & \leq c^2 n \|\tilde{\boldsymbol{\ell}}_g^{\mathrm{T}} \hat{\boldsymbol{v}}_{(-f)} \hat{\boldsymbol{z}}_{(-f)}^{\mathrm{T}}\|_2^2 \hat{\bar{M}}^2 \|\boldsymbol{V}_g\|_2^2 \\
& \leq c' n \|\tilde{\boldsymbol{\ell}}_g^{\mathrm{T}} \hat{\boldsymbol{v}}_{(-f)} \hat{\boldsymbol{z}}_{(-f)}^{\mathrm{T}}\|_2^2 \hat{\bar{M}}^2
\end{aligned}
$$

and similarly

$$
\begin{aligned}
& \mathbb{V} \left[\tilde{\boldsymbol{\ell}}_g^{\mathrm{T}} \left\{ I_K - \hat{\boldsymbol{v}}_{(-f)} \hat{\boldsymbol{v}}_{(-f)}^{\mathrm{T}} \right\} \hat{\boldsymbol{N}}_{(-f)} \boldsymbol{e}_{f_g} \mid \boldsymbol{Y}_{(-f)}, \boldsymbol{\pi} \right] \\
& \leq c' \|\tilde{\boldsymbol{\ell}}_g^{\mathrm{T}} \left\{ I_K - \hat{\boldsymbol{v}}_{(-f)} \hat{\boldsymbol{v}}_{(-f)}^{\mathrm{T}} \right\}\|_2^2 n \hat{\bar{M}}^2.
\end{aligned}
$$

for some constant $c' > 0$ not dependent on n or p. Therefore,

$$\mathbb{V}\left(t_3 \mid \boldsymbol{Y}_{(-f)}, \boldsymbol{\pi}\right) \leq O\left(n^{-2}p^{-1}\right)\hat{\hat{M}}^2\left[\|\tilde{\boldsymbol{L}}_f\left\{I_K - \hat{\boldsymbol{v}}_{(-f)}\hat{\boldsymbol{v}}_{(-f)}^{\mathrm{T}}\right\}\|_F^2 \right.$$
$$\left. + \|\tilde{\boldsymbol{L}}_f\hat{\boldsymbol{v}}_{(-f)}\hat{\boldsymbol{z}}_{(-f)}^{\mathrm{T}}\|_F^2\right].$$

By (3.53),

$$\begin{aligned}\|\tilde{\boldsymbol{L}}_f\left\{I_K - \hat{\boldsymbol{v}}_{(-f)}\hat{\boldsymbol{v}}_{(-f)}^{\mathrm{T}}\right\}\|_F^2 + \|\tilde{\boldsymbol{L}}_f\hat{\boldsymbol{v}}_{(-f)}\hat{\boldsymbol{z}}_{(-f)}^{\mathrm{T}}\|_F^2 \leq & K\|\tilde{\boldsymbol{L}}_f\left\{I_K - \hat{\boldsymbol{v}}_{(-f)}\hat{\boldsymbol{v}}_{(-f)}^{\mathrm{T}}\right\}\|_2^2 \\ & + K\|\tilde{\boldsymbol{L}}_f\hat{\boldsymbol{v}}_{(-f)}\hat{\boldsymbol{z}}_{(-f)}^{\mathrm{T}}\|_2^2 \\ = & o_P(1),\end{aligned}$$

meaning $\mathbb{V}\left(t_3 \mid \boldsymbol{Y}_{(-f)}, \boldsymbol{\pi}\right) = o_P\left(n^{-2}\right)$. Therefore, the third term in (3.51) is $o_P\left(n^{-1}\right)$ when $\underline{k = K}$. Putting this all together, we get

$$\frac{1}{\delta_{f*}^2 n p_f}\sum_{i=1}^n \|\boldsymbol{L}_f\bar{\boldsymbol{c}}_i - \hat{\boldsymbol{L}}_{f,(-i)}\hat{\hat{\boldsymbol{c}}}_i\|_2^2 = K/n\left\{\alpha + o_P(1)\right\}$$

when $\underline{k = K}$.

(b) $\underline{k < K}$. First,

$$\begin{aligned}&\frac{1}{\delta_{f*}^2 n p_f}\sum_{i=1}^n \|\frac{\boldsymbol{L}_f\bar{\boldsymbol{c}}_i - \boldsymbol{L}_f\bar{\boldsymbol{C}}^{\mathrm{T}}\hat{\hat{\boldsymbol{C}}}\left(\hat{\hat{\boldsymbol{C}}}^{\mathrm{T}}\hat{\hat{\boldsymbol{C}}}\right)^{-1}\hat{\hat{\boldsymbol{c}}}_i}{\left(1 - \hat{\hat{h}}_i\right)}\|_2^2 \\ &\geq \frac{1}{\delta_{f*}^2}\operatorname{tr}\left\{p_f^{-1}\boldsymbol{L}_f^{\mathrm{T}}\boldsymbol{L}_f\left(n^{-1}\bar{\boldsymbol{C}}^{\mathrm{T}}P_{\hat{\hat{\boldsymbol{C}}}}^{\perp}\bar{\boldsymbol{C}}\right)\right\}.\end{aligned}$$

When $\gamma_K \to \infty$, this term is $\gtrsim \gamma_K/n\left\{1 + o_P(1)\right\}$. When $\gamma_K = O(1)$ and $\delta_*^{-2}\gamma_K \geq 1 + \epsilon$ for some $\epsilon > 0$, then this term is

$$\geq n^{-1}\sum_{r=k+1}^K \delta_*^{-2}\gamma_r\left\{1 + o_P(1)\right\} \geq \left(K - k + \epsilon\right)/n\left\{1 + o_P(1)\right\}.$$

Next,

$$n^{-1} \sum_{i=1}^{n} \left(\frac{1}{1-\hat{\hat{h}}_i} \right)^2 \hat{\hat{\boldsymbol{H}}}_i^{\mathrm{T}} \boldsymbol{A}_{(-i)} \hat{\boldsymbol{V}}_{(-f)}^{-1/2} \left(\delta_{f*}^{-2} p_f^{-1} \boldsymbol{E}_f^{\mathrm{T}} \boldsymbol{E}_f \right) \hat{\boldsymbol{V}}_{(-f)}^{-1/2} \boldsymbol{A}_{(-i)} \hat{\hat{\boldsymbol{H}}}_i \geq 0.$$

However, when $\gamma_K = O(1)$, $\|\boldsymbol{V}_{f*} - \hat{\boldsymbol{V}}_f\|_2 = O_P\left(n^{-1}\right)$ for any $k < K$ by the proof of Theorem 3.7. Therefore, the analysis when k was equal to K shows that

$$\begin{aligned} & n^{-1} \sum_{i=1}^{n} \left(\frac{1}{1-\hat{\hat{h}}_i} \right)^2 \hat{\hat{\boldsymbol{H}}}_i^{\mathrm{T}} \boldsymbol{A}_{(-i)} \hat{\boldsymbol{V}}_{(-f)}^{-1/2} \left(\delta_{f*}^{-2} p_f^{-1} \boldsymbol{E}_f^{\mathrm{T}} \boldsymbol{E}_f \right) \hat{\boldsymbol{V}}_{(-f)}^{-1/2} \boldsymbol{A}_{(-i)} \hat{\hat{\boldsymbol{H}}}_i \\ & \geq k/n \left\{1 + o_P(1)\right\} \end{aligned}$$

when $\gamma_K = O(1)$ and $\underline{k < K}$. It therefore remains to bound the third term in (3.51) when $\underline{k < K}$. To do so, we perform a similar analysis as when k was assumed to be K. We can write the third term in (3.51) as

$$t_3 = \delta_{f*}^{-2} n^{-3/2} p^{-1/2} \sum_{g=1}^{p_f} \left\{ \tilde{\boldsymbol{\ell}}_g^{\mathrm{T}} \hat{\boldsymbol{N}}_{(-f)} \boldsymbol{e}_{f_g} - \tilde{\boldsymbol{\ell}}_g^{\mathrm{T}} \hat{\boldsymbol{U}}_{(-f)} \boldsymbol{e}_{f_g} \right\}$$

where

$$\begin{aligned} \hat{\boldsymbol{N}}_{(-f)} =& \sum_{i=1}^{n} \left(1 - \hat{\hat{h}}_i\right)^{-2} \bar{\boldsymbol{c}}_i \hat{\hat{\boldsymbol{H}}}_i^{\mathrm{T}} \boldsymbol{A}_{(-i)} \hat{\boldsymbol{V}}_{(-f)}^{-1/2} = \bar{\boldsymbol{C}}^{\mathrm{T}} \operatorname{diag}\left\{ \left(1 - \hat{\hat{h}}_i\right)^{-2} \right\} \hat{\hat{\boldsymbol{H}}} \hat{\boldsymbol{V}}_{(-f)}^{-1/2} \\ & - \bar{\boldsymbol{C}}^{\mathrm{T}} \operatorname{diag}\left\{ \hat{\hat{h}}_i \left(1 - \hat{\hat{h}}_i\right)^{-2} \right\} \hat{\boldsymbol{V}}_{(-f)}^{-1/2} \\ \hat{\boldsymbol{U}}_{(-f)} =& \bar{\boldsymbol{C}}^{\mathrm{T}} \hat{\hat{\boldsymbol{H}}} \operatorname{diag}\left\{ \left(1 - \hat{\hat{h}}_i\right)^{-2} \right\} \hat{\hat{\boldsymbol{H}}} \hat{\boldsymbol{V}}_{(-f)}^{-1/2} - \bar{\boldsymbol{C}}^{\mathrm{T}} \hat{\hat{\boldsymbol{H}}} \operatorname{diag}\left\{ \hat{\hat{h}}_i \left(1 - \hat{\hat{h}}_i\right)^{-2} \right\} \\ & \times \hat{\boldsymbol{V}}_{(-f)}^{-1/2}. \end{aligned}$$

Again, $\|\hat{\boldsymbol{N}}_{(-f)}\|_2, \|\hat{\boldsymbol{U}}_{(-f)}\|_2 \leq c n^{1/2} \hat{\hat{M}}$, where $\hat{\hat{M}} = \max_{i \in [n]} \left\{ \left(1 - \hat{\hat{h}}_i\right)^{-2} \right\}$ and

$c > 0$ is a constant not dependent on n or p. Therefore,

$$\mathbb{V}\left(t_3 \mid \boldsymbol{Y}_{(-f)}, \boldsymbol{\pi}\right) \leq O\left(n^{-2}p^{-1}\right) \hat{\hat{M}}^2 \|\tilde{\boldsymbol{L}}\|_F^2 = O\left(n^{-2}p^{-1}\gamma_K\right) \hat{\hat{M}}^2 \tag{3.54}$$

When $\gamma_K = o(n)$, $\hat{\hat{M}} = O_P(1)$, meaning

$$\gamma_K^{1/2} n^{-1} p^{-1/2} \hat{\hat{M}} = o_P\left(\gamma_K/n\right).$$

When $\gamma_K \asymp n$,

$$\begin{aligned}
&\gamma_K^{1/2} n^{-1} p^{-1/2} \hat{\hat{M}} = o_P\left(\gamma_K/n\right) \\
\Leftrightarrow & n^{-1/2} p^{-1/2} \hat{\hat{M}} = o_P(1) \\
\Leftarrow & n^{-1/2} \max_{i \in [n]} \left\{\left(1 - \hat{\hat{h}}_i\right)^{-1}\right\} \leq c
\end{aligned}$$

for some constant $c > 0$ not dependent on n or p, which is true by assumption. Since $\max_{i \in [n]} \left(1 - \hat{\hat{h}}_i\right)^{-1} = \alpha + o_P(1)$ when $\gamma_K = o(n)$, this proves that

$$\begin{aligned}
&\frac{1}{\delta_{f*}^2 n p_f} \sum_{i=1}^n \|\boldsymbol{L}_f \bar{\boldsymbol{c}}_i - \hat{\boldsymbol{L}}_{f,(-i)} \hat{\bar{\boldsymbol{c}}}_i\|_2^2 \\
&\begin{cases} \gtrsim \gamma_K/n \left\{1 + o_P(1)\right\} & \text{if } \gamma_K \to \infty \\ \geq (K+\epsilon)/n \left\{1 + o_P(1)\right\} & \text{if } \gamma_K = O(1), \delta_*^{-2}\gamma_K \geq 1 + \epsilon \end{cases}
\end{aligned}$$

for some constant $\epsilon > 0$ when $\underline{k < K}$.

(c) $\underline{k > K}$. First,

$$\frac{1}{\delta_{f*}^2 n p_f} \sum_{i=1}^n \left\| \frac{\boldsymbol{L}_f \bar{\boldsymbol{c}}_i - \boldsymbol{L}_f \bar{\boldsymbol{C}}^{\mathrm{T}} \hat{\bar{\boldsymbol{C}}} \left(\hat{\bar{\boldsymbol{C}}}^{\mathrm{T}} \hat{\bar{\boldsymbol{C}}}\right)^{-1} \hat{\bar{\boldsymbol{c}}}_i}{\left(1 - \hat{\hat{h}}_i\right)} \right\|_2^2 \geq 0.$$

Next, the second term of (3.51) is such that

$$
\begin{aligned}
&n^{-1}\sum_{i=1}^{n}\left(\frac{1}{1-\hat{\bar{h}}_i}\right)^2 \hat{\bar{\boldsymbol{H}}}_i^{\mathrm{T}}\boldsymbol{A}_{(-i)}\hat{\boldsymbol{V}}_{(-f)}^{-1/2}\left(\delta_{f*}^{-2}p_f^{-1}\boldsymbol{E}_f^{\mathrm{T}}\boldsymbol{E}_f\right)\hat{\boldsymbol{V}}_{(-f)}^{-1/2}\boldsymbol{A}_{(-i)}\hat{\bar{\boldsymbol{H}}}_i\\
\geq &n^{-1}\sum_{i=1}^{n}\left(\frac{1}{1-\hat{\bar{h}}_i}\right) \hat{\bar{\boldsymbol{H}}}_i^{\mathrm{T}}\boldsymbol{A}_{(-i)}\hat{\boldsymbol{V}}_{(-f)}^{-1/2}\left(\delta_{f*}^{-2}p_f^{-1}\boldsymbol{E}_f^{\mathrm{T}}\boldsymbol{E}_f\right)\hat{\boldsymbol{V}}_{(-f)}^{-1/2}\boldsymbol{A}_{(-i)}\hat{\bar{\boldsymbol{H}}}_i.
\end{aligned}
$$

By Theorem 3.7, $\|\hat{\boldsymbol{V}}_{(-f)}-\boldsymbol{V}_{(-f)*}\|_2 = O_P\left(n^{-1}\right)$ whenever $k \geq K$, meaning we can apply the same techniques we used to derive the conditional mean and variance of this term when k was assumed equal to K to show that

$$
\begin{aligned}
&\mathbb{E}\left\{n^{-1}\sum_{i=1}^{n}\left(\frac{1}{1-\hat{\bar{h}}_i}\right) \hat{\bar{\boldsymbol{H}}}_i^{\mathrm{T}}\boldsymbol{A}_{(-i)}\hat{\boldsymbol{V}}_{(-f)}^{-1/2}\left(\delta_{f*}^{-2}p_f^{-1}\boldsymbol{E}_f^{\mathrm{T}}\boldsymbol{E}_f\right)\hat{\boldsymbol{V}}_{(-f)}^{-1/2}\boldsymbol{A}_{(-i)}\hat{\bar{\boldsymbol{H}}}_i\right.\\
&\left.\mid \boldsymbol{Y}_{(-f)},\boldsymbol{\pi}\right\} = k/n\left\{1+o_P(1)\right\},
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathrm{SD}\left\{n^{-1}\sum_{i=1}^{n}\left(\frac{1}{1-\hat{\bar{h}}_i}\right) \hat{\bar{\boldsymbol{H}}}_i^{\mathrm{T}}\boldsymbol{A}_{(-i)}\hat{\boldsymbol{V}}_{(-f)}^{-1/2}\left(\delta_{f*}^{-2}p_f^{-1}\boldsymbol{E}_f^{\mathrm{T}}\boldsymbol{E}_f\right)\hat{\boldsymbol{V}}_{(-f)}^{-1/2}\boldsymbol{A}_{(-i)}\hat{\bar{\boldsymbol{H}}}_i\right.\\
&\left.\mid \boldsymbol{Y}_{(-f)},\boldsymbol{\pi}\right\} = O_P\left(n^{-1}p^{-1/2}\right) = o_P\left(n^{-1}\right).
\end{aligned}
$$

It remains to understand the third term in (3.51) when $\underline{k > K}$. To do so, we need to find an expression for $\boldsymbol{\Delta}_i$. When $k > K$,

$$
\hat{\bar{\boldsymbol{C}}} = \left[\bar{\boldsymbol{C}}\hat{\boldsymbol{v}}_{(-f)} + n^{1/2}\boldsymbol{Q}_{\bar{C}}\hat{\boldsymbol{z}}_{(-f)}\ \ \hat{\boldsymbol{R}}_{(-f)}\right] \in \mathbb{R}^{n\times k}. \tag{3.55}
$$

where $\hat{\boldsymbol{R}}_{(-f)} \in \mathbb{R}^{n\times(k-K)}$. The asymptotic properties of $\hat{\boldsymbol{v}}_{(-f)}$ and $\hat{\boldsymbol{z}}_{(-f)}$ when $k > K$ are the same as those when $k = K$ because $\|\hat{\boldsymbol{V}}_{(-f)}-\boldsymbol{V}_{(-f)*}\|_2 = O_P\left(n^{-1}\right)$

whenever $k \geq K$ by Theorem 3.7. Since the columns of $\hat{\bar{\boldsymbol{C}}}$ are orthogonal,

$$\mathbf{0} = \hat{\boldsymbol{v}}_{(-f)}^{\mathrm{T}} \bar{\boldsymbol{C}}^{\mathrm{T}} \hat{\boldsymbol{R}}_{(-f)} + n^{1/2} \hat{\boldsymbol{z}}_{(-f)}^{\mathrm{T}} \boldsymbol{Q}_{\bar{C}}^{\mathrm{T}} \hat{\boldsymbol{R}}_{(-f)},$$

meaning

$$n^{-1} \bar{\boldsymbol{C}}^{\mathrm{T}} \hat{\boldsymbol{R}}_{(-f)} = O_P \left\{ n^{1/2} \left(\gamma_K p \right)^{-1/2} + \left(n \gamma_K \right)^{-1} \right\}.$$

Let $\hat{\boldsymbol{r}}_i \in \mathbb{R}^{k-K}$ be the i^{th} row of $\hat{\boldsymbol{R}}_{(-f)}$. Then

$$\boldsymbol{\Delta}_i = \tilde{\boldsymbol{L}} \left\{ I_K - \hat{\boldsymbol{v}}_{(-f)} \hat{\boldsymbol{v}}_{(-f)}^{\mathrm{T}} \right\} \bar{\boldsymbol{c}}_i - n^{1/2} \tilde{\boldsymbol{L}} \hat{\boldsymbol{v}}_{(-f)} \hat{\boldsymbol{z}}_{(-f)}^{\mathrm{T}} \bar{\boldsymbol{q}}_i - \tilde{\boldsymbol{L}} \left\{ n^{-1} \bar{\boldsymbol{C}}^{\mathrm{T}} \hat{\boldsymbol{R}}_{(-f)} \right\} \hat{\boldsymbol{r}}_i \tag{3.56}$$

Using (3.53) and the fact that

$$\| \tilde{\boldsymbol{L}} \left\{ n^{-1} \bar{\boldsymbol{C}}^{\mathrm{T}} \hat{\boldsymbol{R}}_{(-f)} \right\} \|_2 = O_P \left(n^{1/2} p^{-1/2} + \gamma_K^{-1/2} n^{-1} \right), \tag{3.57}$$

we can use the analysis when $k = K$ to show that

$$\begin{aligned} sd_k^2 =& \mathbb{V} \left\{ \frac{2}{\delta_{f*}^2 n^{3/2} p_f^{1/2}} \sum_{i=1}^{n} \left(\frac{1}{1 - \hat{\bar{h}}_i} \right)^2 \hat{\bar{\boldsymbol{H}}}_i^{\mathrm{T}} \boldsymbol{A}_{(-i)} \hat{\boldsymbol{V}}_{(-f)}^{-1/2} \boldsymbol{E}_f^{\mathrm{T}} \boldsymbol{\Delta}_i \mid \boldsymbol{Y}_{(-f)}, \boldsymbol{\pi} \right\} \\ \leq& O\left(n^{-2} p^{-1} \right) \hat{\bar{M}}^2 \left[\| \tilde{\boldsymbol{L}}_f \left\{ I_K - \hat{\boldsymbol{v}}_{(-f)} \hat{\boldsymbol{v}}_{(-f)}^{\mathrm{T}} \right\} \|_F^2 + \| \tilde{\boldsymbol{L}}_f \hat{\boldsymbol{v}}_{(-f)} \hat{\boldsymbol{z}}_{(-f)}^{\mathrm{T}} \|_F^2 \right. \\ & \left. + \| \tilde{\boldsymbol{L}} \left\{ n^{-1} \bar{\boldsymbol{C}}^{\mathrm{T}} \hat{\boldsymbol{R}}_{(-f)} \right\} \|_F^2 \right] \\ =& O\left(n^{-2} p^{-1} \right) \hat{\bar{M}}^2 O_P \left\{ \left(n^{1/2} p^{-1/2} + \gamma_K^{-1/2} n^{-1} \right)^2 \right\}. \end{aligned} \tag{3.58}$$

Therefore,

$$sd_k = O_P \left(n^{-1/2} p^{-1} + n^{-2} p^{-1/2} \gamma_K^{-1/2} \right) \hat{\bar{M}}$$

where $\hat{\hat{M}} = \max_{i\in[n]} \left\{ \left(1 - \hat{\hat{h}}_i\right)^{-2} \right\}$. First,

$$\begin{aligned} & n^{-2} p^{-1/2} \gamma_K^{-1/2} \hat{\hat{M}} = o_P\left(n^{-1}\right) \\ \Leftrightarrow & n^{-1} p^{-1/2} \gamma_K^{-1/2} \hat{\hat{M}} = o_P(1) \\ \Leftrightarrow & n^{-1/2} p^{-1/4} \gamma_K^{-1/4} \hat{\hat{M}}^{1/2} = o_P(1). \end{aligned}$$

The last equality follows by assumptions on $\hat{\hat{M}}^{1/2}$. Next,

$$\begin{aligned} & n^{-1/2} p^{-1} \hat{\hat{M}} = o_P\left(n^{-1}\right) \\ \Leftrightarrow & n^{1/4} p^{-1/2} \hat{\hat{M}}^{1/2} = o_P\left(n^{-1}\right) \end{aligned}$$

where the last equality follows by the assumptions on $\hat{\hat{M}}^{1/2}$. Therefore, $sd_k = o_P\left(n^{-1}\right)$.

Putting (a), (b) and (c) together, we have shown that when $\gamma_K \to \infty$,

$$\frac{1}{\delta_{f*}^2 n p_f} \sum_{i=1}^{n} \|\boldsymbol{L}_f \bar{\boldsymbol{c}}_i - \hat{\boldsymbol{L}}_{f,(-i)} \hat{\bar{\boldsymbol{c}}}_i\|_2^2 \begin{cases} \gtrsim \gamma_K/n \left\{1 + o_P(1)\right\} & \text{if } k < K \\ = K/n \left\{\alpha + o_P(1)\right\} & \text{if } k = K \\ \geq k/n \left\{1 + o_P(1)\right\} & \text{if } k > K \end{cases}$$

and when $\gamma_K = O(1)$ and $\gamma_K \geq 1 + \epsilon$,

$$\frac{1}{\delta_{f*}^2 n p_f} \sum_{i=1}^{n} \|\boldsymbol{L}_f \bar{\boldsymbol{c}}_i - \hat{\boldsymbol{L}}_{f,(-i)} \hat{\bar{\boldsymbol{c}}}_i\|_2^2 \begin{cases} \geq (K+\epsilon)/n \left\{1 + o_P(1)\right\} & \text{if } k < K \\ = K/n \left\{\alpha + o_P(1)\right\} & \text{if } k = K \\ \geq k/n \left\{1 + o_P(1)\right\} & \text{if } k > K \end{cases}$$

for some constant $\epsilon > 0$.

(2) The second term of (3.50) is

$$\delta_{f*}^{-2} n^{-1} \operatorname{tr}\left\{\hat{\boldsymbol{V}}_{(-f)}^{-1}\left(p_f^{-1}\boldsymbol{E}_f^{\mathrm{T}}\boldsymbol{E}_f\right)\right\} = \delta_{f*}^{-2} n^{-1} p_f^{-1} \sum_{g=1}^{p_f} \boldsymbol{e}_g^{\mathrm{T}} \hat{\boldsymbol{V}}_{(-f)}^{-1} \boldsymbol{e}_g,$$

where

$$\mathbb{V}\left\{\boldsymbol{e}_g^{\mathrm{T}} \hat{\boldsymbol{V}}_{(-f)}^{-1} \boldsymbol{e}_g \mid \boldsymbol{Y}_{(-f)}, \boldsymbol{\pi}\right\} \leq nc$$

for some constant $c > 0$ not dependent on n or p. Therefore, for $k \in \{0, 1, \ldots, K_{\max}\}$,

$$\mathrm{SD}\left[\delta_{f*}^{-2} n^{-1} \operatorname{tr}\left\{\hat{\boldsymbol{V}}_{(-f)}^{-1}\left(p_f^{-1}\boldsymbol{E}_f^{\mathrm{T}}\boldsymbol{E}_f\right)\right\} \mid \boldsymbol{Y}_{(-f)}, \boldsymbol{\pi}\right] = O\left(n^{-1/2} p^{-1/2}\right) = o\left(n^{-1}\right).$$

We now need only calculate the conditional expectation for different values of k. First,

$$\mathbb{E}\left[\delta_{f*}^{-2} n^{-1} \operatorname{tr}\left\{\hat{\boldsymbol{V}}_{(-f)}^{-1}\left(p_f^{-1}\boldsymbol{E}_f^{\mathrm{T}}\boldsymbol{E}_f\right)\right\} \mid \boldsymbol{Y}_{(-f)}, \boldsymbol{\pi}\right] = n^{-1} \operatorname{tr}\left\{\hat{\boldsymbol{V}}_{(-f)}^{-1} \boldsymbol{V}_{f*}\right\}.$$

Since $\log|\hat{\boldsymbol{V}}_{(-f)}^{-1} \boldsymbol{V}_{f*}| = \log|\hat{\boldsymbol{V}}_{(-f)}^{-1}| + \log|\boldsymbol{V}_{f*}| = 0$, $n^{-1} \operatorname{tr}\left\{\hat{\boldsymbol{V}}_{(-f)}^{-1} \boldsymbol{V}_{f*}\right\} \geq 1$ for all $k \in \{0, 1, \ldots, K_{\max}\}$, with equality if and only if $\hat{\boldsymbol{V}}_{(-f)} = \boldsymbol{V}_{f*}$ by Jensen's Inequality. Therefore, we only need to show that

$$\mathbb{E}\left[\delta_{f*}^{-2} n^{-1} \operatorname{tr}\left\{\hat{\boldsymbol{V}}_{(-f)}^{-1}\left(p_f^{-1}\boldsymbol{E}_f^{\mathrm{T}}\boldsymbol{E}_f\right)\right\} \mid \boldsymbol{Y}_{(-f)}, \boldsymbol{\pi}\right] = 1 + o_P\left(n^{-1}\right)$$

when $\underline{k = K}$. By Taylor's Theorem we have

$$\begin{aligned} n^{-1} \operatorname{tr}\left\{\hat{\boldsymbol{V}}_{(-f)}^{-1} \boldsymbol{V}_{f*}\right\} =& 1 - n^{-1} \operatorname{tr}\left\{\boldsymbol{V}_{f*}^{-1}\left(\hat{\boldsymbol{V}}_{(-f)} - \boldsymbol{V}_{f*}\right) \boldsymbol{V}_{f*}^{-1} \boldsymbol{V}_{f*}\right\} \\ &+ O_P\left\{\left(n^{-1} + p^{-1/2}\right)^2\right\} = 1 - n^{-1} \operatorname{tr}\left\{\boldsymbol{V}_{f*}^{-1}\left(\hat{\boldsymbol{V}}_{(-f)} - \boldsymbol{V}_{f*}\right)\right\} \\ &+ o_P\left(n^{-1}\right). \end{aligned}$$

Therefore,

$$1 \leq n^{-1} \operatorname{tr}\left\{\hat{\boldsymbol{V}}_{(-f)}^{-1} \boldsymbol{V}_{f*}\right\} = 1 - n^{-1} \operatorname{tr}\left\{\boldsymbol{V}_{f*}^{-1}\left(\hat{\boldsymbol{V}}_{(-f)} - \boldsymbol{V}_{f*}\right)\right\} + o_P\left(n^{-1}\right)$$
$$= 2 - n^{-1} \operatorname{tr}\left\{\boldsymbol{V}_{f*}^{-1} \hat{\boldsymbol{V}}_{(-f)}\right\} + o_P\left(n^{-1}\right) \leq 1 + o_P\left(n^{-1}\right),$$

which implies

$$n^{-1} \operatorname{tr}\left\{\hat{\boldsymbol{V}}_{(-f)}^{-1} \boldsymbol{V}_{f*}\right\} = 1 + o_P\left(n^{-1}\right).$$

Therefore,

$$\delta_{f*}^{-2} n^{-1} \operatorname{tr}\left\{\hat{\boldsymbol{V}}_{(-f)}^{-1}\left(p_f^{-1} \boldsymbol{E}_f^{\mathrm{T}} \boldsymbol{E}_f\right)\right\} \begin{cases} \geq 1 + o_P\left(n^{-1}\right) & \text{if } k \neq K \\ = 1 + o_P\left(n^{-1}\right) & \text{if } k = K \end{cases}.$$

(3) The third term of equation (3.50), up to a scalar constant, is

$$\begin{aligned} &\delta_{f*}^{-2} n^{-1} p_f^{-1} \sum_{i=1}^{n}\left(\boldsymbol{L}_f \bar{\boldsymbol{c}}_i - \hat{\boldsymbol{L}}_{f,(-i)} \hat{\bar{\boldsymbol{c}}}_i\right)^{\mathrm{T}} \boldsymbol{E}_f \hat{\boldsymbol{V}}_{(-f)}^{-1/2} \boldsymbol{a}_i \\ =&\delta_{f*}^{-2} \underbrace{n^{-3/2} p^{-1/2} \sum_{i=1}^{n}\left(1 - \hat{\bar{h}}_i\right)^{-1} \boldsymbol{\Delta}_i^{\mathrm{T}} \boldsymbol{E}_f \hat{\boldsymbol{V}}_{(-f)}^{-1/2} \boldsymbol{a}_i}_{\text{(a)}} \\ &+ \underbrace{\delta_{f*}^{-2} n^{-1} p_f^{-1} \sum_{i=1}^{n}\left(1 - \hat{\bar{h}}_i\right)^{-1} \hat{\boldsymbol{H}}_i^{\mathrm{T}} \boldsymbol{A}_{(-i)} \hat{\boldsymbol{V}}_{(-f)}^{-1/2} \boldsymbol{E}_f^{\mathrm{T}} \boldsymbol{E}_f \hat{\boldsymbol{V}}_{(-f)}^{-1/2} \boldsymbol{a}_i}_{\text{(b)}} \end{aligned} \tag{3.59}$$

where $\boldsymbol{\Delta}_i$ was defined in (3.52). We derive the asymptotic properties of (a) and (b) for each $k \in \{0, 1, \ldots, K_{\max}\}$ below.

(a) The conditional expectation of this term is 0. We derive its its conditional variance by partitioning $k \in \{0, 1, \ldots, K_{\max}\}$ into $k \leq K$ and $k > K$.

(i) $\underline{k \leq K}$. We can re-write (a) from (3.59) as

$$n^{-3/2}p^{-1/2}\sum_{g=1}^{p_f}\left\{\tilde{\boldsymbol{\ell}}_g^{\mathrm{T}}\boldsymbol{F}_{(-f)}\boldsymbol{e}_{f_g} - \tilde{\boldsymbol{\ell}}_g^{\mathrm{T}}\boldsymbol{G}_{(-f)}\boldsymbol{e}_{f_g}\right\}$$
$$\boldsymbol{F}_{(-f)} = \bar{\boldsymbol{C}}^{\mathrm{T}}\,\mathrm{diag}\left\{\left(1-\hat{\bar{h}}_i\right)^{-1}\right\}\hat{\boldsymbol{V}}_{(-f)}^{-1/2}$$
$$\boldsymbol{G}_{(-f)} = \bar{\boldsymbol{C}}^{\mathrm{T}}\hat{\boldsymbol{H}}\,\mathrm{diag}\left\{\left(1-\hat{\bar{h}}_i\right)^{-1}\right\}\hat{\boldsymbol{V}}_{(-f)}^{-1/2}.$$

Using an identical method used to show (3.54), we can show that

$$\mathrm{SD}\left[n^{-3/2}p^{-1/2}\sum_{g=1}^{p_f}\left\{\tilde{\boldsymbol{\ell}}_g^{\mathrm{T}}\boldsymbol{F}_{(-f)}\boldsymbol{e}_{f_g} - \tilde{\boldsymbol{\ell}}_g^{\mathrm{T}}\boldsymbol{G}_{(-f)}\boldsymbol{e}_{f_g}\right\} \mid \boldsymbol{Y}_{(-f)}, \boldsymbol{\pi}\right]$$
$$= O_P\left(n^{-1}p^{-1/2}\gamma_K^{1/2}\right)M_k$$

where $M_k = \max_{i\in[n]}\left(1-\hat{\bar{h}}_i\right)^{-1}$. When $k=K$, we have that $M_k = O_P(1)$ and $n^{-1}p^{-1/2}\gamma_K^{1/2} = o\left(n^{-1}\right)$. When $k<K$,

$$n^{-1}p^{-1/2}\gamma_K^{1/2}M_k = o_P\left(\gamma_K/n\right)$$
$$\Leftarrow p^{-1/2}M_k = o_P(1).$$

The last equality follows by assumption on M_k, and proves that (a) from (3.59) is $o_P\left(n^{-1}\right)$ when $\underline{k \leq K}$.

(ii) $\underline{k > K}$. We perform the same analysis as we did in part (i) above, except with $\boldsymbol{\Delta}_i$ defined in (3.56), where the asymptotic properties of each term in (3.56) are given in (3.53) and (3.57). We can re-write (a) from (3.59) as

$$n^{-3/2}p^{-1/2}\sum_{g=1}^{p_f}\left[\tilde{\boldsymbol{\ell}}_g^{\mathrm{T}}\left\{I_K - \hat{\boldsymbol{v}}_{(-f)}\hat{\boldsymbol{v}}_{(-f)}^{\mathrm{T}}\right\}\boldsymbol{F}_{(-f)} - \tilde{\boldsymbol{\ell}}_g^{\mathrm{T}}\hat{\boldsymbol{v}}_{(-f)}\hat{\boldsymbol{z}}_{(-f)}\boldsymbol{J}_{1,(-f)}\right.$$
$$\left.-\tilde{\boldsymbol{\ell}}_g^{\mathrm{T}}\left\{n^{-1}\bar{\boldsymbol{C}}^{\mathrm{T}}\hat{\boldsymbol{R}}_{(-f)}\right\}\boldsymbol{J}_{2,(-f)}\right]\boldsymbol{e}_{f_g}$$

where

$$
\begin{aligned}
\boldsymbol{F}_{(-f)} &= \bar{\boldsymbol{C}}^{\mathrm{T}} \operatorname{diag}\left\{\left(1-\hat{\bar{h}}_i\right)^{-1}\right\} \hat{\boldsymbol{V}}_{(-f)}^{-1/2} \\
\boldsymbol{J}_{1,(-f)} &= n^{1/2} \boldsymbol{Q}_{\bar{C}}^{\mathrm{T}} \operatorname{diag}\left\{\left(1-\hat{\bar{h}}_i\right)^{-1}\right\} \hat{\boldsymbol{V}}_{(-f)}^{-1/2} \\
\boldsymbol{J}_{2,(-f)} &= \hat{\boldsymbol{R}}_{(-f)}^{\mathrm{T}} \operatorname{diag}\left\{\left(1-\hat{\bar{h}}_i\right)^{-1}\right\} \hat{\boldsymbol{V}}_{(-f)}^{-1/2}
\end{aligned}
$$

and where $\hat{\boldsymbol{R}}_{(-f)}$ was defined in (3.55). We can then use the same techniques that we used to show (3.58) to show that

$$
\begin{aligned}
&\mathrm{SD}\left\{n^{-3/2} p^{-1/2} \sum_{i=1}^{n}\left(1-\hat{\bar{h}}_i\right)^{-1} \boldsymbol{\Delta}_i^{\mathrm{T}} \boldsymbol{E}_f \hat{\boldsymbol{V}}_{(-f)}^{-1/2} \boldsymbol{a}_i \mid \boldsymbol{Y}_{(-f)}, \boldsymbol{\pi}\right\} \\
=&O_P\left(n^{-1/2} p^{-1}+n^{-2} p^{-1/2} \gamma_K^{-1/2}\right) M_k=o_P\left(n^{-1}\right)
\end{aligned}
$$

where the last equality follows by the assumptions on M_k.

Items (i) and (ii) show that

$$
n^{-3/2} p^{-1/2} \sum_{i=1}^{n}\left(1-\hat{\bar{h}}_i\right)^{-1} \boldsymbol{\Delta}_i^{\mathrm{T}} \boldsymbol{E}_f \hat{\boldsymbol{V}}_{(-f)}^{-1/2} \boldsymbol{a}_i=\begin{cases} o_P\left(\gamma_K / n\right) & \text{if } k<K \\ o_P\left(n^{-1}\right) & \text{if } k \geq K . \end{cases}
$$

(b) We first derive the conditional expectation of this term.

$$
\begin{aligned}
&\mathbb{E}\left\{\delta_{f*}^{-2} n^{-1} p_f^{-1} \sum_{i=1}^{n}\left(1-\hat{\bar{h}}_i\right)^{-1} \hat{\boldsymbol{H}}_i^{\mathrm{T}} \boldsymbol{A}_{(-i)} \hat{\boldsymbol{V}}_{(-f)}^{-1/2} \boldsymbol{E}_f^{\mathrm{T}} \boldsymbol{E}_f \hat{\boldsymbol{V}}_{(-f)}^{-1/2} \boldsymbol{a}_i \mid \boldsymbol{Y}_{(-f)}, \boldsymbol{\pi}\right\} \\
=&n^{-1} \sum_{i=1}^{n} \frac{1}{1-\hat{\bar{h}}_i} \hat{\bar{\boldsymbol{H}}}_i^{\mathrm{T}} \boldsymbol{A}_{(-i)} \hat{\boldsymbol{V}}_{(-f)}^{-1/2} \boldsymbol{V}_{f*} \hat{\boldsymbol{V}}_{(-f)}^{-1/2} \boldsymbol{a}_i \\
=&n^{-1} \operatorname{tr}\left\{\hat{\boldsymbol{V}}_{(-f)}^{-1/2} \boldsymbol{V}_{f*} \hat{\boldsymbol{V}}_{(-f)}^{-1/2} \sum_{i=1}^{n} \frac{1}{1-\hat{\bar{h}}_i} \boldsymbol{A}_{(-i)} \hat{\bar{\boldsymbol{H}}}_i \boldsymbol{a}_i^{\mathrm{T}}\right\}
\end{aligned}
$$

where

$$\operatorname{tr}\left\{\sum_{i=1}^{n}\frac{1}{1-\hat{\bar{h}}_i}\boldsymbol{A}_{(-i)}\hat{\bar{\boldsymbol{H}}}_i\boldsymbol{a}_i^{\mathrm{T}}\right\}=0$$
$$\|\sum_{i=1}^{n}\frac{1}{1-\hat{\bar{h}}_i}\boldsymbol{A}_{(-i)}\hat{\bar{\boldsymbol{H}}}_i\boldsymbol{a}_i^{\mathrm{T}}\|_F^2=\sum_{i=1}^{n}\frac{\hat{\bar{h}}_i}{1-\hat{\bar{h}}_i}.$$

Using the Cauchy-Schwartz inequality and the proof of Theorem 3.7 (see equation (3.49)), we get that

$$\begin{aligned}
&|n^{-1}\operatorname{tr}\left\{\hat{\boldsymbol{V}}_{(-f)}^{-1/2}\boldsymbol{V}_{f*}\hat{\boldsymbol{V}}_{(-f)}^{-1/2}\sum_{i=1}^{n}\frac{1}{1-\hat{\bar{h}}_i}\boldsymbol{A}_{(-i)}\hat{\bar{\boldsymbol{H}}}_i\boldsymbol{a}_i^{\mathrm{T}}\right\}|\\
&\leq\left(n^{-1}\operatorname{tr}\left[\left\{I-\hat{\boldsymbol{V}}_{(-f)}^{-1/2}\boldsymbol{V}_{f*}\hat{\boldsymbol{V}}_{(-f)}^{-1/2}\right\}^2\right]\right)^{1/2}\left(n^{-1}\|\sum_{i=1}^{n}\frac{1}{1-\hat{\bar{h}}_i}\boldsymbol{A}_{(-i)}\hat{\bar{\boldsymbol{H}}}_i\boldsymbol{a}_i^{\mathrm{T}}\|_F^2\right)^{1/2}\\
&=\begin{cases}O_P\left\{\gamma_K n^{-3/2}+(np)^{-1/2}\right\}M_k^{1/2} & \text{if } k<K\\ O_P\left\{n^{-3/2}+(np)^{-1/2}\right\}M_k^{1/2} & \text{if } k\geq K\end{cases}
\end{aligned}$$

where $M_k=\max_{i\in[n]}\left(1-\hat{\bar{h}}_i\right)^{-1}$. The subscript k is to indicate that M_k is a function of $k\in\{0,1,\ldots,K_{\max}\}$. By assumption,

$$M_k=\begin{cases}O_P(1) & \text{if } k=K \text{ or } k<K \text{ and } \gamma_K=o(n)\\ O\left[\min\left\{n^{1/2},\frac{p/n}{\log(p/n)}\right\}\right] & \text{otherwise.}\end{cases}$$

Therefore, when $k>K$,

$$\begin{aligned}
&M_k^{1/2}\left(n^{-3/2}+n^{-1/2}p^{-1/2}\right)=o_P\left(n^{-1}\right)\\
\Leftrightarrow &M_k^{1/2}\left(n^{-1/2}+n^{1/2}p^{-1/2}\right)=o_P\left(1\right)\\
\Leftrightarrow &M_k\left(n^{-1}+np^{-1}\right)=o_P(1),
\end{aligned}$$

where the last equality holds by our assumption on M_k. When $k < K$,

$$\begin{aligned} &M_k^{1/2}\left(\gamma_K n^{-3/2} + n^{-1/2}p^{-1/2}\right) = o_P\left(\gamma_K/n\right) \\ \Leftrightarrow &M_k^{1/2}\left(n^{-1/2} + \gamma_K^{-1} n^{1/2} p^{-1/2}\right) = o_P(1) \end{aligned}$$

where the last equality holds by our assumption on M_k. This proves that

$$\mathbb{E}\left\{\delta_{f*}^{-2} n^{-1} p_f^{-1} \sum_{i=1}^{n} \left(1 - \hat{\bar{h}}_i\right)^{-1} \hat{\boldsymbol{H}}_i^{\mathrm{T}} \boldsymbol{A}_{(-i)} \hat{\boldsymbol{V}}_{(-f)}^{-1/2} \boldsymbol{E}_f^{\mathrm{T}} \boldsymbol{E}_f \hat{\boldsymbol{V}}_{(-f)}^{-1/2} \boldsymbol{a}_i \mid \boldsymbol{Y}_{(-f)}, \boldsymbol{\pi}\right\}$$
$$= \begin{cases} o_P\left(\gamma_K/n\right) & \text{if } k < K \\ o_P\left(n^{-1}\right) & \text{if } k \geq K. \end{cases}$$

To then calculate the conditional variance, we see that we can-rewrite (b) from (3.59), up to a scalar constant, as

$$n^{-1} p_f^{-1} \sum_{g=1}^{p_f} \boldsymbol{e}_{f_g}^{\mathrm{T}} \hat{\boldsymbol{V}}_{(-f)}^{-1/2} \boldsymbol{D}_{(-f)} \hat{\boldsymbol{V}}_{(-f)}^{-1/2} \boldsymbol{e}_{f_g}$$

where

$$\boldsymbol{D}_{(-f)} = 2^{-1} \sum_{i=1}^{n} \left(1 - \hat{\bar{h}}_i\right)^{-1} \boldsymbol{a}_i \hat{\boldsymbol{H}}_i^{\mathrm{T}} \boldsymbol{A}_{(-i)} + 2^{-1} \sum_{i=1}^{n} \left(1 - \hat{\bar{h}}_i\right)^{-1} \boldsymbol{A}_{(-i)} \hat{\boldsymbol{H}}_i \boldsymbol{a}_i^{\mathrm{T}}.$$

From the analysis above, $\operatorname{tr}\left\{\boldsymbol{D}_{(-f)}\right\} = 0$ and

$$\begin{aligned} \operatorname{tr}\left\{\boldsymbol{D}_{(-f)}^2\right\} &= \|\boldsymbol{D}_{(-f)}\|_F^2 \leq \|\sum_{i=1}^{n} \left(1 - \hat{\bar{h}}_i\right)^{-1} \boldsymbol{a}_i \hat{\boldsymbol{H}}_i^{\mathrm{T}} \boldsymbol{A}_{(-i)}\|_F^2 = \sum_{i=1}^{n} \frac{\hat{\bar{h}}_i}{1 - \hat{\bar{h}}_i} \\ &\leq k M_k. \end{aligned}$$

Therefore, for any $g \in [p]$,

$$\mathbb{V}\left\{\boldsymbol{e}_{f_g}^{\mathrm{T}}\hat{\boldsymbol{V}}_{(-f)}^{-1/2}\boldsymbol{D}_{(-f)}\hat{\boldsymbol{V}}_{(-f)}^{-1/2}\boldsymbol{e}_{f_g} \mid \boldsymbol{Y}_{(-f)}, \boldsymbol{\pi}\right\} = 2\,\mathrm{tr}\left\{\boldsymbol{D}_{(-f)}\boldsymbol{A}_{(-f),g}^2\boldsymbol{D}_{(-f)}\boldsymbol{A}_{(-f),g}^2\right\}$$
$$\boldsymbol{A}_{(-f),g}^2 = \hat{\boldsymbol{V}}_{(-f)}^{-1/2}\boldsymbol{V}_g\hat{\boldsymbol{V}}_{(-f)}^{-1/2}.$$

Let $\boldsymbol{U}_g\boldsymbol{\Sigma}_g\boldsymbol{U}_g^{\mathrm{T}}$ be the eigen decomposition of $\boldsymbol{A}_{(-f),g}$, where $\boldsymbol{U}_g$ is a unitary matrix and $\boldsymbol{\Sigma}_g$ is a diagonal matrix with positive elements that are bounded above by a constant (by assumptions on $\boldsymbol{V}_g$ and $\hat{\boldsymbol{V}}_{(-f)}$). Let $\tilde{\boldsymbol{D}}_{(-f)} = \boldsymbol{U}_g^{\mathrm{T}}\boldsymbol{D}_{(-f)}\boldsymbol{U}_g$. Then for some constant $c > 0$,

$$\begin{aligned}\mathrm{tr}\left\{\boldsymbol{D}_{(-f)}\boldsymbol{A}_{(-f),g}^2\boldsymbol{D}_{(-f)}\boldsymbol{A}_{(-f),g}^2\right\} &= \mathrm{tr}\left\{\boldsymbol{\Sigma}_g\tilde{\boldsymbol{D}}_{(-f)}\boldsymbol{\Sigma}_g\boldsymbol{\Sigma}_g\tilde{\boldsymbol{D}}_{(-f)}\boldsymbol{\Sigma}_g\right\}\\ &\leq c\,\mathrm{tr}\left\{\tilde{\boldsymbol{D}}_{(-f)}^2\right\} = c\,\mathrm{tr}\left\{\boldsymbol{D}_{(-f)}^2\right\} \leq ckM_k.\end{aligned}$$

Therefore, for some constant $c' > 0$ not dependent on n or p,

$$\begin{aligned}\mathrm{SD}\left\{n^{-1}p_f^{-1}\sum_{g=1}^{p_f}\boldsymbol{e}_{f_g}^{\mathrm{T}}\hat{\boldsymbol{V}}_{(-f)}^{-1/2}\boldsymbol{D}_{(-f)}\hat{\boldsymbol{V}}_{(-f)}^{-1/2}\boldsymbol{e}_{f_g} \mid \boldsymbol{Y}_{(-f)}, \boldsymbol{\pi}\right\} &\leq c'n^{-1}p^{-1/2}M_k^{1/2}\\ &= o_P\left(n^{-1}\right).\end{aligned}$$

The last equality follows because $M_k = O_P(1)$ when $k = K$ and is $o\,(p)$ otherwise by assumption. This shows that

$$\begin{aligned}&\delta_{f*}^{-2}n^{-1}p_f^{-1}\sum_{i=1}^{n}\left(1-\hat{\hat{h}}_i\right)^{-1}\hat{\boldsymbol{H}}_i^{\mathrm{T}}\boldsymbol{A}_{(-i)}\hat{\boldsymbol{V}}_{(-f)}^{-1/2}\boldsymbol{E}_f^{\mathrm{T}}\boldsymbol{E}_f\hat{\boldsymbol{V}}_{(-f)}^{-1/2}\boldsymbol{a}_i\\ &= \begin{cases} o_P\left(\gamma_K/n\right) & \text{if } k < K \\ o_P\left(n^{-1}\right) & \text{if } k \geq K \end{cases}.\end{aligned}$$

Putting (1), (2) and (3) together, we get that when $\gamma_K \to \infty$,

$$n\{g(k)-1\} \begin{cases} \gtrsim \gamma_K \{1+o_P(1)\} & \text{if } k<K \\ = K\{\alpha + o_P(1)\} & \text{if } k=K \\ \geq k\{1+o_P(1)\} & \text{if } k>K \end{cases} \cdot$$

When $\gamma_K = O(1)$ and $\delta_*^{-2}\gamma_K \geq 1+\epsilon$ for some constant $\epsilon > 0$,

$$n\{g(k)-1\} \begin{cases} \geq (K+\epsilon)\{1+o_P(1)\} & \text{if } k<K \\ = K\{\alpha + o_P(1)\} & \text{if } k=K \\ \geq k\{1+o_P(1)\} & \text{if } k>K \end{cases} \cdot$$

Lastly,

$$K\alpha < K+\epsilon \Leftrightarrow \alpha < 1+\epsilon/K \Leftrightarrow \phi < \frac{\epsilon/K}{1+\epsilon/K} \Leftarrow \phi < \epsilon/(2K)$$

which completes the proof. □

Lemma 3.2 and Theorem 3.9

In this section we prove Lemma 3.2 and Theorem 3.9 which justify performing inference on $\boldsymbol{\beta}$ using the estimated design matrix $\hat{\boldsymbol{C}}$ when K is known. We first prove a technical lemma that will be useful in both proofs.

***Lemma* 3.8.** *Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ be a positive definite matrix with smallest eigenvalue 1 and largest eigenvalue λ, and suppose Assumptions 3.7, 3.8 and 3.9 hold. For any $\boldsymbol{C}$, let $\boldsymbol{R}^2 = n^{-1}\boldsymbol{C}_\perp^{\mathrm{T}}\boldsymbol{W}_*^{-1}\boldsymbol{C}_\perp$. Then for*

$$\boldsymbol{M} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{C}\boldsymbol{R}^{-1},$$

$\|\boldsymbol{M}\|_2$ *is identifiable with* $\|\boldsymbol{M}\|_2 \leq c + c'\lambda$*, where* $c, c' > 0$ *are not dependent on* n *or* p*.*

Proof. The identifiability of $\|\boldsymbol{M}\|_2$ follows by the proof of Proposition 3.3. Note that we can write

$$\boldsymbol{C} = \boldsymbol{X}\boldsymbol{S} + \boldsymbol{V}_{*}\boldsymbol{Q}_X\boldsymbol{W}_{*}^{-1}\boldsymbol{C}_{\perp}$$

where $\boldsymbol{S}$ is defined in Assumption 3.9. Therefore,

$$\boldsymbol{M} = \boldsymbol{S}\boldsymbol{R}^{-1} + \left(n^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{X}\right)^{-1} n^{-1/2}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{V}_{*}\boldsymbol{Q}_X\boldsymbol{W}_{*}^{-1/2}\left(n^{-1/2}\boldsymbol{W}_{*}^{-1/2}\boldsymbol{C}_{\perp}\boldsymbol{R}^{-1}\right)$$

where $\|\boldsymbol{S}\boldsymbol{R}^{-1}\|_2$ is bounded from above by Assumption 3.9. This proves the result. □

First, $\boldsymbol{\beta}$ is identifiable regardless of the parametrization of $\boldsymbol{L}$ and $\boldsymbol{C}$ by Proposition 3.3. Second, restricted maximum likelihood and generalized least squares estimates for $\boldsymbol{V}_g$ ($g = 1, \ldots, p$) and $\boldsymbol{\beta}$ only depend on $\boldsymbol{X}$ and the column space of $\hat{\boldsymbol{C}}$. Therefore, we can assume without loss of generality that $n^{-1}\boldsymbol{C}_{\perp}^{\mathrm{T}}\hat{\boldsymbol{W}}^{-1}\boldsymbol{C}_{\perp} = I_K$ and $np^{-1}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}$ is diagonal with decreasing elements in the proof of Theorem 3.9. In the statement of Lemma 3.2, we assume $n^{-1}\boldsymbol{C}_{\perp}^{\mathrm{T}}\boldsymbol{W}_{*}^{-1}\boldsymbol{C}_{\perp} = I_K$. However, by Remark 3.20, the results of Theorem 3.7 and Lemma 3.8,

$$\|\boldsymbol{\Omega} - \boldsymbol{\Omega}\left(n^{-1}\boldsymbol{C}_{\perp}^{\mathrm{T}}\hat{\boldsymbol{W}}^{-1}\boldsymbol{C}_{\perp}\right)^{-1/2}\hat{\boldsymbol{U}}\|_2 = O_P\left(n^{-1}\right)$$

where here $\boldsymbol{U} = I_K + O_P\left(n^{-1}\right)$. Therefore, we assume without loss of generality that $n^{-1}\boldsymbol{C}_{\perp}^{\mathrm{T}}\hat{\boldsymbol{W}}^{-1}\boldsymbol{C}_{\perp} = I_K$ and $np^{-1}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}$ is diagonal with decreasing elements in the proofs of Lemma 3.2 and Theorem 3.9.

Proof of Lemma 3.2. Let $m = n - d$. Once we have estimated $\boldsymbol{C}_{\perp}$, δ_{*}^{2} and $\boldsymbol{V}_{*}$ from Algorithm

3.2, we define

$$\tilde{\boldsymbol{X}} = \hat{\boldsymbol{V}}^{-1/2}\boldsymbol{X} \in \mathbb{R}^{n\times d}$$
$$\tilde{\boldsymbol{C}} = \hat{\boldsymbol{V}}^{-1/2}\boldsymbol{C} \in \mathbb{R}^{n\times K}$$
$$\tilde{\boldsymbol{C}}_{\perp} = \boldsymbol{Q}_{\tilde{X}}^{\mathrm{T}}\tilde{\boldsymbol{C}} \in \mathbb{R}^{m\times K}.$$

From our discussion above, it suffices to assume that $m^{-1}\tilde{\boldsymbol{C}}_{\perp}^{\mathrm{T}}\tilde{\boldsymbol{C}}_{\perp} = I_K$ and $mp^{-1}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L} = \operatorname{diag}(\lambda_1, \ldots, \lambda_K)$ where $\lambda_1, \lambda_K \asymp \gamma_K$ and $(\lambda_k - \lambda_{k+1})/\lambda_{k+1} \geq c_1^{-1} + o_P(1)$ where c_1 is defined in Assumption 3.7. Therefore, all conditions of Theorem 3.10 and Lemma 3.6 are satisfied and (3.38) holds with $\epsilon = 1/n$. This means that for $\hat{\boldsymbol{L}}$ defined as

$$\hat{\boldsymbol{L}} = \boldsymbol{Y}_2\hat{\boldsymbol{W}}^{-1}\hat{\boldsymbol{C}}_{\perp}\left(\hat{\boldsymbol{C}}_{\perp}^{\mathrm{T}}\hat{\boldsymbol{W}}^{-1}\hat{\boldsymbol{C}}_{\perp}\right)^{-1}$$

and $\hat{\mu}_k$ defined in the proof of Theorem 3.10 (see (3.36) and (3.37)),

$$\begin{aligned}
&m(\gamma_K p)^{-1}\hat{\boldsymbol{L}}^{\mathrm{T}}\hat{\boldsymbol{L}} - m\gamma_K^{-1}\hat{\delta}^2\left(\hat{\boldsymbol{C}}_{\perp}^{\mathrm{T}}\hat{\boldsymbol{W}}^{-1}\hat{\boldsymbol{C}}_{\perp}\right)^{-1} = \operatorname{diag}(\hat{\mu}_1, \ldots, \hat{\mu}_K) - \gamma_K^{-1}\hat{\delta}^2 I_K \\
=&m(\gamma_K p)^{-1}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L} + O_P\left\{n(p\gamma_K)^{-1} + (p\gamma_K)^{-1/2} + (n\gamma_K)^{-1}\right\} \\
=&m(\gamma_K p)^{-1}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L} + o_P\left(n^{-1/2}\right)
\end{aligned}$$

Therefore, to prove Lemma 3.2, we only need to show that

$$\|m(\gamma_K p)^{-1}\hat{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{Y}_1 - \left\{m(\gamma_K p)^{-1}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{\Omega}^{\mathrm{T}}\right\}\|_2 = o_P\left(n^{-1/2}\right). \tag{3.60}$$

We first note that

$$\tilde{\boldsymbol{C}}_{\perp} = \boldsymbol{Q}_{\tilde{X}}^{\mathrm{T}}\tilde{\boldsymbol{C}} = \left(\boldsymbol{Q}_X^{\mathrm{T}}\hat{\boldsymbol{V}}\boldsymbol{Q}_X\right)^{-1/2}\left(\hat{\boldsymbol{V}}^{1/2}\boldsymbol{Q}_X\right)^{\mathrm{T}}\hat{\boldsymbol{V}}^{-1/2}\boldsymbol{C} = \hat{\boldsymbol{W}}^{-1/2}\boldsymbol{C}_{\perp}.$$

Algorithm 3.2 will return

$$\hat{\tilde{\boldsymbol{C}}}_\perp = \tilde{\boldsymbol{C}}_\perp \hat{\boldsymbol{v}} + \sqrt{m}\boldsymbol{Q}_{\tilde{C}_\perp}\hat{\boldsymbol{z}},$$

an estimate of $\tilde{\boldsymbol{C}}_\perp$, where $\hat{\boldsymbol{v}}$ and $\hat{\boldsymbol{z}}$ are defined in (3.36) and (3.37) with asymptotic properties given in (3.38). Next, since $\boldsymbol{Q}_{\tilde{X}} = \hat{\boldsymbol{V}}^{1/2}\boldsymbol{Q}_X\hat{\boldsymbol{W}}^{-1/2}$,

$$\begin{aligned} \hat{\boldsymbol{L}} &= \boldsymbol{Y}\boldsymbol{Q}_X\hat{\boldsymbol{W}}^{-1}\hat{\boldsymbol{C}}_\perp\left(\hat{\boldsymbol{C}}_\perp^{\mathrm{T}}\hat{\boldsymbol{W}}^{-1}\hat{\boldsymbol{C}}_\perp\right)^{-1} = \boldsymbol{Y}\boldsymbol{Q}_X\hat{\boldsymbol{W}}^{-1/2}\hat{\tilde{\boldsymbol{C}}}_\perp\left(\hat{\tilde{\boldsymbol{C}}}_\perp^{\mathrm{T}}\hat{\tilde{\boldsymbol{C}}}_\perp\right)^{-1} \\ &= m^{-1}\boldsymbol{L}\tilde{\boldsymbol{C}}_\perp^{\mathrm{T}}\hat{\tilde{\boldsymbol{C}}}_\perp + m^{-1}\boldsymbol{E}\hat{\boldsymbol{V}}^{-1/2}\boldsymbol{Q}_{\tilde{X}}\hat{\tilde{\boldsymbol{C}}}_\perp = \boldsymbol{L}\hat{\boldsymbol{v}} + m^{-1}\boldsymbol{E}\hat{\boldsymbol{V}}^{-1/2}\boldsymbol{Q}_{\tilde{X}}\hat{\tilde{\boldsymbol{C}}}_\perp \end{aligned} \tag{3.61}$$

and

$$\begin{aligned} m\left(\gamma_K p\right)^{-1}\hat{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{Y}_1 =& m\left(\gamma_K p\right)^{-1}\hat{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{Y}\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}\left(\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}}\right)^{-1} \\ &= \underbrace{m\left(\gamma_K p\right)^{-1}\hat{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{\beta}}_{(1)} + \underbrace{m\left(\gamma_K p\right)^{-1}\hat{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{\Omega}^{\mathrm{T}}}_{(2)} \\ &\quad + \underbrace{m\left(\gamma_K p\right)^{-1}\hat{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{E}\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}\left(\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}}\right)^{-1}}_{(3)}. \end{aligned}$$

We will go through each of these terms to prove (3.60).

(1)

$$m\left(\gamma_K p\right)^{-1}\hat{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{\beta} = m\left(\gamma_K p\right)^{-1}\hat{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{\beta} + \left(\gamma_K p\right)^{-1}\hat{\tilde{\boldsymbol{C}}}_\perp^{\mathrm{T}}\boldsymbol{Q}_{\tilde{X}}^{\mathrm{T}}\hat{\boldsymbol{V}}^{-1/2}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{\beta}.$$

By the assumptions on $\boldsymbol{\beta}$ and $\boldsymbol{L}$, the first term is $o_P\left(n^{-1/2}\right)$. For the second term, it suffices to assume $d = 1$. Then

$$\mathrm{Var}\left(p^{-1/2}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{\beta}\right) = p^{-1}\sum_{g=1}^{p}\beta_g^2\boldsymbol{V}_g.$$

Therefore,

$$\mathbb{E}\|p^{-1/2}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{\beta}\|_2^2 \asymp ns_1$$

where $s_1 = p^{-1}\sum_{g=1}^{p} I\left(\beta_g \neq 0\right)$. By Assumption 3.9,

$$n^{1/2}\left\{\left(\gamma_K^{-1}n^{1/2}p^{-1/2}\right)(ns_1)^{1/2}\right\} \lesssim n^{3/4}(\gamma_K p)^{-1/2} = \left\{n^{3/2}(\gamma_K p)^{-1}\right\}^{1/2} \to 0.$$

Therefore, the second term is $o_P\left(n^{-1/2}\right)$.

(2)

$$m(\gamma_K p)^{-1}\hat{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{L} = \hat{\boldsymbol{v}}^{\mathrm{T}}m(\gamma_K p)^{-1}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L} + (\gamma_K mp)^{-1/2}\hat{\tilde{\boldsymbol{C}}}_{\perp}^{\mathrm{T}}\boldsymbol{Q}_{\tilde{X}}^{\mathrm{T}}\hat{\boldsymbol{V}}^{-1/2}\boldsymbol{E}^{\mathrm{T}}\bar{\boldsymbol{L}}.$$

By (3.38d), the first term is such that

$$\|\hat{\boldsymbol{v}}^{\mathrm{T}}m(\gamma_K p)^{-1}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L} - m(\gamma_K p)^{-1}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\|_2 = o_P\left(n^{-1/2}\right).$$

For the second term,

$$\begin{aligned}
&(\gamma_K mp)^{-1/2}\hat{\tilde{\boldsymbol{C}}}_{\perp}^{\mathrm{T}}\boldsymbol{Q}_{\tilde{X}}^{\mathrm{T}}\hat{\boldsymbol{V}}^{-1/2}\boldsymbol{E}^{\mathrm{T}}\bar{\boldsymbol{L}} = (\gamma_K mp)^{-1/2}\hat{\boldsymbol{v}}^{\mathrm{T}}\tilde{\boldsymbol{C}}_{\perp}^{\mathrm{T}}\boldsymbol{Q}_{\tilde{X}}^{\mathrm{T}}\hat{\boldsymbol{V}}^{-1/2}\boldsymbol{E}^{\mathrm{T}}\bar{\boldsymbol{L}}\\
&+ \underbrace{\hat{\boldsymbol{z}}^{\mathrm{T}}\boldsymbol{Q}_{\tilde{C}}^{\mathrm{T}}\boldsymbol{Q}_{\tilde{X}}^{\mathrm{T}}\hat{\boldsymbol{V}}^{-1/2}}_{O_P\left\{n^{1/2}(\gamma_K p)^{-1/2}+(n\gamma_K)^{-1}\right\}}\underbrace{(\gamma_K p)^{-1/2}\boldsymbol{E}^{\mathrm{T}}\bar{\boldsymbol{L}}}_{O_P\left\{n^{1/2}(\gamma_K p)^{-1/2}\right\}}\\
=&(\gamma_K mp)^{-1/2}\hat{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{C}_{\perp}^{\mathrm{T}}\left(\boldsymbol{Q}_X^{\mathrm{T}}\hat{\boldsymbol{V}}\boldsymbol{Q}_X\right)^{-1}\boldsymbol{Q}_X^{\mathrm{T}}\boldsymbol{E}^{\mathrm{T}}\bar{\boldsymbol{L}} + o_P\left(n^{-1/2}\right)\\
=&(\gamma_K p)^{-1/2}\hat{\boldsymbol{v}}^{\mathrm{T}}\left(m^{-1/2}\boldsymbol{C}_{\perp}\right)^{\mathrm{T}}\left(\boldsymbol{Q}_X^{\mathrm{T}}\boldsymbol{V}_*\boldsymbol{Q}_X\right)^{-1}\boldsymbol{Q}_X^{\mathrm{T}}\boldsymbol{E}^{\mathrm{T}}\bar{\boldsymbol{L}} + o_P\left(n^{-1/2}\right)\\
=&o_P\left(n^{-1/2}\right).
\end{aligned}$$

(3) Lastly,

$$
\begin{aligned}
& m\left(\gamma_K p\right)^{-1} \hat{\boldsymbol{L}}^{\mathrm{T}} \boldsymbol{E} \hat{\boldsymbol{V}}^{-1} \boldsymbol{X}\left(\tilde{\boldsymbol{X}}^{\mathrm{T}} \tilde{\boldsymbol{X}}\right)^{-1}=m\left(\gamma_K p\right)^{-1} \hat{\boldsymbol{v}} \boldsymbol{L}^{\mathrm{T}} \boldsymbol{E} \hat{\boldsymbol{V}}^{-1} \boldsymbol{X}\left(\tilde{\boldsymbol{X}}^{\mathrm{T}} \tilde{\boldsymbol{X}}\right)^{-1} \\
& \quad+\left(\gamma_K p\right)^{-1} \hat{\tilde{\boldsymbol{C}}}_{\perp}^{\mathrm{T}} \boldsymbol{Q}_{\tilde{X}}^{\mathrm{T}} \hat{\boldsymbol{V}}^{-1/2} \boldsymbol{E}^{\mathrm{T}} \boldsymbol{E} \hat{\boldsymbol{V}}^{-1} \boldsymbol{X}\left(\tilde{\boldsymbol{X}}^{\mathrm{T}} \tilde{\boldsymbol{X}}\right)^{-1} \\
& =\left(\gamma_K p\right)^{-1/2} \bar{\boldsymbol{L}}^{\mathrm{T}} \boldsymbol{E} \hat{\boldsymbol{V}}^{-1}\left\{m^{1/2} \boldsymbol{X}\left(\tilde{\boldsymbol{X}}^{\mathrm{T}} \tilde{\boldsymbol{X}}\right)^{-1}\right\} \\
& \quad+\gamma_K^{-1} \hat{\boldsymbol{v}}^{\mathrm{T}} \boldsymbol{C}_{\perp}^{\mathrm{T}}\left(\boldsymbol{Q}_X^{\mathrm{T}} \hat{\boldsymbol{V}} \boldsymbol{Q}_X\right)^{-1} \boldsymbol{Q}_X^{\mathrm{T}} p^{-1} \boldsymbol{E}^{\mathrm{T}} \boldsymbol{E} \hat{\boldsymbol{V}}^{-1} \boldsymbol{X}\left(\tilde{\boldsymbol{X}}^{\mathrm{T}} \tilde{\boldsymbol{X}}\right)^{-1} \\
& \quad+\gamma_K^{-1} \hat{\boldsymbol{z}}^{\mathrm{T}} \boldsymbol{Q}_{\tilde{C}_{\perp}}^{\mathrm{T}} \boldsymbol{Q}_{\tilde{X}}^{\mathrm{T}} \hat{\boldsymbol{V}}^{-1/2} p^{-1} \boldsymbol{E}^{\mathrm{T}} \boldsymbol{E} \hat{\boldsymbol{V}}^{-1/2}\left\{m^{1/2} \tilde{\boldsymbol{X}}\left(\tilde{\boldsymbol{X}}^{\mathrm{T}} \tilde{\boldsymbol{X}}\right)^{-1}\right\} .
\end{aligned}
$$

First,

$$
\begin{aligned}
& \left(\gamma_K p\right)^{-1/2} \bar{\boldsymbol{L}}^{\mathrm{T}} \boldsymbol{E} \hat{\boldsymbol{V}}^{-1}\left\{m^{1/2} \boldsymbol{X}\left(\tilde{\boldsymbol{X}}^{\mathrm{T}} \tilde{\boldsymbol{X}}\right)^{-1}\right\} \\
& =\left(\gamma_K p\right)^{-1/2} \bar{\boldsymbol{L}}^{\mathrm{T}} \boldsymbol{E} \boldsymbol{V}_*^{-1}\left\{m^{1/2} \boldsymbol{X}\left(\tilde{\boldsymbol{X}}^{\mathrm{T}} \tilde{\boldsymbol{X}}\right)^{-1}\right\}+o_P\left(n^{-1/2}\right)=o_P\left(n^{-1/2}\right) .
\end{aligned}
$$

Next,

$$
\begin{aligned}
& \gamma_K^{-1} \hat{\boldsymbol{v}}^{\mathrm{T}} \boldsymbol{C}_{\perp}^{\mathrm{T}}\left(\boldsymbol{Q}_X^{\mathrm{T}} \hat{\boldsymbol{V}} \boldsymbol{Q}_X\right)^{-1} \boldsymbol{Q}_X^{\mathrm{T}} \frac{1}{p} \boldsymbol{E}^{\mathrm{T}} \boldsymbol{E} \hat{\boldsymbol{V}}^{-1} \boldsymbol{X}\left(\tilde{\boldsymbol{X}}^{\mathrm{T}} \tilde{\boldsymbol{X}}\right)^{-1} \\
& =\gamma_K^{-1} \hat{\boldsymbol{v}}^{\mathrm{T}} \boldsymbol{C}_{\perp}^{\mathrm{T}} \boldsymbol{W}_*^{-1} \boldsymbol{Q}_X^{\mathrm{T}} p^{-1} \boldsymbol{E}^{\mathrm{T}} \boldsymbol{E} \boldsymbol{V}_*^{-1} \boldsymbol{X}\left(\tilde{\boldsymbol{X}}^{\mathrm{T}} \tilde{\boldsymbol{X}}\right)^{-1}+O_P\left\{\left(\gamma_K n\right)^{-1}\right\} .
\end{aligned}
$$

Let $\boldsymbol{A} \in \mathbb{R}^{m \times K}$ be a matrix whose columns form an orthonormal basis for the column space of $\boldsymbol{C}_{\perp}$, which is not dependent on $\hat{\boldsymbol{V}}$, and define the non-random matrices

$$
\begin{aligned}
& \boldsymbol{f}=\boldsymbol{Q}_X \boldsymbol{W}_*^{-1} \boldsymbol{A} \in \mathbb{R}^{n \times K} \\
& \boldsymbol{u}=\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{X}\right)^{-1/2} \in \mathbb{R}^{n \times d},
\end{aligned}
$$

both of which have bounded 2-norm and $\boldsymbol{f}^{\mathrm{T}} \boldsymbol{u}=\mathbf{0}$. Since K, d are fixed, it suffices to assume that $K = d = 1$. We first see that

$$\mathbb{E}\left(\boldsymbol{f}^{\mathrm{T}} p^{-1} \boldsymbol{E}^{\mathrm{T}} \boldsymbol{E} \boldsymbol{V}_*^{-1} \boldsymbol{u}\right) = \boldsymbol{f}^{\mathrm{T}} \left(p^{-1} \sum_{g=1}^{p} \boldsymbol{V}_g\right) \boldsymbol{V}_*^{-1} \boldsymbol{u} = \boldsymbol{f}^{\mathrm{T}} \boldsymbol{V}_* \boldsymbol{V}_*^{-1} \boldsymbol{u} = 0$$

and

$$\mathbb{V}\left(\boldsymbol{f}^{\mathrm{T}} p^{-1} \boldsymbol{E}^{\mathrm{T}} \boldsymbol{E} \boldsymbol{V}_*^{-1} \boldsymbol{u}\right) = p^{-2} \sum_{g=1}^{p} \operatorname{Var}\left(\tilde{e}_{g,1} \tilde{e}_{g,2}\right) \asymp p^{-1}$$

where $\tilde{e}_{g,1} = \boldsymbol{f}^{\mathrm{T}} \boldsymbol{e}_g \sim N\left(0, \boldsymbol{f}^{\mathrm{T}} \boldsymbol{V}_g \boldsymbol{f}\right)$ and $\tilde{e}_{g,2} = \boldsymbol{u}^{\mathrm{T}} \boldsymbol{V}_*^{-1} \boldsymbol{e}_g \sim N\left(0, \boldsymbol{u}^{\mathrm{T}} \boldsymbol{V}_*^{-1} \boldsymbol{V}_g \boldsymbol{V}_*^{-1} \boldsymbol{u}\right)$. Therefore,

$$\begin{aligned}
&\|\gamma_K^{-1} \hat{\boldsymbol{v}}^{\mathrm{T}} \boldsymbol{C}_{\perp}^{\mathrm{T}} \left(\boldsymbol{Q}_X^{\mathrm{T}} \hat{\boldsymbol{V}} \boldsymbol{Q}_X\right)^{-1} \boldsymbol{Q}_X^{\mathrm{T}} p^{-1} \boldsymbol{E}^{\mathrm{T}} \boldsymbol{E} \hat{\boldsymbol{V}}^{-1} \boldsymbol{X} \left(\tilde{\boldsymbol{X}}^{\mathrm{T}} \tilde{\boldsymbol{X}}\right)^{-1} \|_2 \\
=& O_P\left\{\gamma_K^{-1} p^{-1/2} + (\gamma_K n)^{-1}\right\} = o_P\left(n^{-1/2}\right).
\end{aligned}$$

Finally,

$$\begin{aligned}
&\gamma_K^{-1} \hat{\boldsymbol{z}}^{\mathrm{T}} \boldsymbol{Q}_{\tilde{C}_\perp}^{\mathrm{T}} \boldsymbol{Q}_{\tilde{X}}^{\mathrm{T}} \hat{\boldsymbol{V}}^{-1/2} p^{-1} \boldsymbol{E}^{\mathrm{T}} \boldsymbol{E} \hat{\boldsymbol{V}}^{-1/2} \left\{m^{1/2} \tilde{\boldsymbol{X}} \left(\tilde{\boldsymbol{X}}^{\mathrm{T}} \tilde{\boldsymbol{X}}\right)^{-1}\right\} \\
=& \gamma_K^{-1} \hat{\boldsymbol{z}}^{\mathrm{T}} \boldsymbol{Q}_{\tilde{C}_\perp}^{\mathrm{T}} \boldsymbol{Q}_{\tilde{X}}^{\mathrm{T}} \boldsymbol{V}_*^{-1/2} p^{-1} \boldsymbol{E}^{\mathrm{T}} \boldsymbol{E} \boldsymbol{V}_*^{-1/2} \left\{m^{1/2} \tilde{\boldsymbol{X}} \left(\tilde{\boldsymbol{X}}^{\mathrm{T}} \tilde{\boldsymbol{X}}\right)^{-1}\right\} + o_P\left\{(n\gamma_K)^{-1}\right\} \\
=& \gamma_K^{-1} \hat{\boldsymbol{z}}^{\mathrm{T}} \boldsymbol{Q}_{\tilde{C}_\perp}^{\mathrm{T}} \boldsymbol{Q}_{\tilde{X}}^{\mathrm{T}} \boldsymbol{V}_*^{-1/2} \left(p^{-1} \boldsymbol{E}^{\mathrm{T}} \boldsymbol{E} - \boldsymbol{V}_*\right) \boldsymbol{V}_*^{-1/2} \left\{m^{1/2} \tilde{\boldsymbol{X}} \left(\tilde{\boldsymbol{X}}^{\mathrm{T}} \tilde{\boldsymbol{X}}\right)^{-1}\right\} \\
&+ o_P\left\{(n\gamma_K)^{-1}\right\}
\end{aligned}$$

where

$$\begin{aligned}
&\|\gamma_K^{-1} \hat{\boldsymbol{z}}^{\mathrm{T}} \boldsymbol{Q}_{\tilde{C}_\perp}^{\mathrm{T}} \boldsymbol{Q}_{\tilde{X}}^{\mathrm{T}} \boldsymbol{V}_*^{-1/2} \left(p^{-1} \boldsymbol{E}^{\mathrm{T}} \boldsymbol{E} - \boldsymbol{V}_*\right) \boldsymbol{V}_*^{-1/2} \left\{m^{1/2} \tilde{\boldsymbol{X}} \left(\tilde{\boldsymbol{X}}^{\mathrm{T}} \tilde{\boldsymbol{X}}\right)^{-1}\right\} \|_2 \\
=& O_P\left(n \gamma_K^{-3/2} p^{-1}\right) = o_P\left(n^{-1/2}\right).
\end{aligned}$$

This proves (3.60) and completes the proof.

$\square$

Proof of Theorem 3.9. Fix a $g \in [p]$ and define $\boldsymbol{V}(\boldsymbol{\theta}) = \sum_{j=1}^{b} [\boldsymbol{\theta}]_j \boldsymbol{B}_j$ for $\boldsymbol{\theta} \in \mathbb{R}^b$ and $\boldsymbol{W}_g = \boldsymbol{Q}_X^{\mathrm{T}} \boldsymbol{V}_g \boldsymbol{Q}_X$. We will first show (3.26). Define

$$\begin{aligned} l_{n,g}(\boldsymbol{\theta}) &= n^{-1} \log|\boldsymbol{V}(\boldsymbol{\theta})| - n^{-1} \boldsymbol{e}_g^{\mathrm{T}} \boldsymbol{V}(\boldsymbol{\theta})^{-1} \boldsymbol{e}_g \\ f_{n,g}(\boldsymbol{\theta}) &= n^{-1} \log|\boldsymbol{V}(\boldsymbol{\theta})| - n^{-1} \operatorname{tr}\left\{\boldsymbol{V}(\boldsymbol{\theta})^{-1} \boldsymbol{V}(\boldsymbol{v}_g)\right\} \end{aligned}$$

Then the function $G_n(\boldsymbol{\theta}) = l_{n,g}(\boldsymbol{\theta}) - f_{n,g}(\boldsymbol{\theta})$ is stochastically equicontinuous on the compact space Θ_*. Since $|G_n(\boldsymbol{\theta})| = o_P(1)$ for all $\boldsymbol{\theta} \in \Theta_*$,

$$\sup_{\boldsymbol{\theta} \in \Theta_*} |l_{n,g}(\boldsymbol{\theta}) - f_{n,g}(\boldsymbol{\theta})| = o_P(1).$$

Next, $-\left[\nabla^2_{\boldsymbol{\theta}} f_{n,g}(\boldsymbol{v}_g)\right]_{ij} = n^{-1} \operatorname{tr}\left\{\boldsymbol{B}_i \boldsymbol{V}(\boldsymbol{v}_g)^{-1} \boldsymbol{B}_j \boldsymbol{V}(\boldsymbol{v}_g)^{-1}\right\}$ for all $i, j \in [b]$, meaning $-\nabla^2_{\boldsymbol{\theta}} f_{n,g}(\boldsymbol{v}_g)$ has eigenvalues that are uniformly bounded above ϵ by Assumption 3.7, where $\epsilon > 0$ is a constant not dependent on n or p. And by Lemma 3.3, $f_{n,g}(\boldsymbol{\theta})$ has a unique maximum at $\boldsymbol{\theta} = \boldsymbol{v}_g$. The restricted maximum likelihood problem we are interested in solving is

$$\begin{aligned} &\arg\max_{\boldsymbol{\theta} \in \Theta_*} \hat{l}_{n,g}(\boldsymbol{\theta}) \\ &\hat{l}_{n,g}(\boldsymbol{\theta}) = -n^{-1} \log|\boldsymbol{Q}_{\hat{C}_\perp}^{\mathrm{T}} \boldsymbol{W}(\boldsymbol{\theta}) \boldsymbol{Q}_{\hat{C}_\perp}| - n^{-1} \boldsymbol{y}_{g2}^{\mathrm{T}} \boldsymbol{Q}_{\hat{C}_\perp} \left\{\boldsymbol{Q}_{\hat{C}_\perp}^{\mathrm{T}} \boldsymbol{W}(\boldsymbol{\theta}) \boldsymbol{Q}_{\hat{C}_\perp}\right\}^{-1} \boldsymbol{Q}_{\hat{C}_\perp}^{\mathrm{T}} \boldsymbol{y}_{g2}. \end{aligned} \tag{3.62}$$

We use the same technique used in the proof of Lemma 3.6 to show that

$$\sup_{\boldsymbol{\theta} \in \Theta_*} |l_{n,g}(\boldsymbol{\theta}) - \hat{l}_{n,g}(\boldsymbol{\theta})| = O_P\left\{n^{1/2} (\gamma_K p)^{-1/2} + n^{-1/2}\right\}, \tag{3.63}$$

which can also be used to show that

$$|\nabla l_{n,g} - \nabla \hat{l}_{n,g}|, |\nabla^2 l_{n,g} - \nabla^2 \hat{l}_{n,g}| = O_P\left\{n^{1/2}\left(\gamma_K p\right)^{-1/2} + n^{-1/2}\right\}.$$

First, from (3.38) we get that

$$\begin{aligned} n^{-1/2}\|\boldsymbol{Q}_{\hat{C}_\perp}^{\mathrm{T}}\boldsymbol{C}_\perp\|_2 &= O\left\{\|\boldsymbol{Q}_{\hat{\bar{C}}_\perp}^{\mathrm{T}}\left(\bar{\boldsymbol{C}}_\perp - \hat{\bar{\boldsymbol{C}}}_\perp\right)\|_2\right\} = O\left(\|\bar{\boldsymbol{C}}_\perp - \hat{\bar{\boldsymbol{C}}}_\perp\|_2\right) \\ &= O_P\left\{n^{1/2}\left(\gamma_K p\right)^{-1/2} + n^{-1}\right\} \end{aligned}$$

where $\bar{\boldsymbol{C}}_\perp = n^{-1/2}\hat{\boldsymbol{W}}^{-1/2}\boldsymbol{C}_\perp$ and $\hat{\bar{\boldsymbol{C}}}_\perp = n^{-1/2}\hat{\boldsymbol{W}}^{-1/2}\hat{\boldsymbol{C}}_\perp$. Define

$$\eta = n^{1/2}\left(\gamma_K p\right)^{-1/2} + n^{-1}.$$

Then

$$\begin{aligned} & n^{-1}\boldsymbol{y}_{g_2}^{\mathrm{T}}\boldsymbol{Q}_{\hat{C}_\perp}\left\{\boldsymbol{Q}_{\hat{C}_\perp}^{\mathrm{T}}\boldsymbol{W}\left(\boldsymbol{\theta}\right)\boldsymbol{Q}_{\hat{C}_\perp}\right\}^{-1}\boldsymbol{Q}_{\hat{C}_\perp}^{\mathrm{T}}\boldsymbol{y}_{g_2} \\ =& \underbrace{n^{-1}\boldsymbol{\ell}_g^{\mathrm{T}}\boldsymbol{C}_\perp^{\mathrm{T}}\boldsymbol{Q}_{\hat{C}_\perp}\left\{\boldsymbol{Q}_{\hat{C}_\perp}^{\mathrm{T}}\boldsymbol{W}\left(\boldsymbol{\theta}\right)\boldsymbol{Q}_{\hat{C}_\perp}\right\}^{-1}\boldsymbol{Q}_{\hat{C}_\perp}^{\mathrm{T}}\boldsymbol{C}_\perp\boldsymbol{\ell}_g}_{o_P(\eta)} \\ &+ \underbrace{n^{-1}\boldsymbol{\ell}_g^{\mathrm{T}}\boldsymbol{C}_\perp^{\mathrm{T}}\boldsymbol{Q}_{\hat{C}_\perp}\left\{\boldsymbol{Q}_{\hat{C}_\perp}^{\mathrm{T}}\boldsymbol{W}\left(\boldsymbol{\theta}\right)\boldsymbol{Q}_{\hat{C}_\perp}\right\}^{-1}\boldsymbol{Q}_{\hat{C}_\perp}^{\mathrm{T}}\boldsymbol{e}_{g_2}}_{O_P(\eta)} \\ &+ \left[n^{-1}\boldsymbol{\ell}_g^{\mathrm{T}}\boldsymbol{C}_\perp^{\mathrm{T}}\boldsymbol{Q}_{\hat{C}_\perp}\left\{\boldsymbol{Q}_{\hat{C}_\perp}^{\mathrm{T}}\boldsymbol{W}\left(\boldsymbol{\theta}\right)\boldsymbol{Q}_{\hat{C}_\perp}\right\}^{-1}\boldsymbol{Q}_{\hat{C}_\perp}^{\mathrm{T}}\boldsymbol{e}_{g_2}\right]^{\mathrm{T}} \\ &+ \underbrace{n^{-1}\boldsymbol{e}_{g_2}^{\mathrm{T}}\boldsymbol{Q}_{\hat{C}_\perp}\left\{\boldsymbol{Q}_{\hat{C}_\perp}^{\mathrm{T}}\boldsymbol{W}\left(\boldsymbol{\theta}\right)\boldsymbol{Q}_{\hat{C}_\perp}\right\}^{-1}\boldsymbol{Q}_{\hat{C}_\perp}^{\mathrm{T}}\boldsymbol{e}_{g_2}}_{n^{-1}e_g^{\mathrm{T}}\boldsymbol{V}(\boldsymbol{\theta})^{-1}e_g + O_P(n^{-1})} = n^{-1}e_g^{\mathrm{T}}\boldsymbol{V}\left(\boldsymbol{\theta}\right)^{-1}e_g + O_P\left(\eta\right). \end{aligned}$$

And since

$$\sup_{\boldsymbol{\theta}\in\Theta_*}|n^{-1}\log|\boldsymbol{Q}_{\hat{C}_\perp}^{\mathrm{T}}\boldsymbol{W}\left(\boldsymbol{\theta}\right)\boldsymbol{Q}_{\hat{C}_\perp}| - n^{-1}\log|\boldsymbol{V}\left(\boldsymbol{\theta}\right)|| = O_P\left(n^{-1}\right),$$

(3.63) follows. Therefore,

$$\sup_{\boldsymbol{\theta}\in\Theta_*} |\hat{l}_{n,g}(\boldsymbol{\theta}) - f_{n,g}(\boldsymbol{\theta})| = o_P(1),$$

meaning $\hat{\boldsymbol{v}}_g$, the solution to (3.62), satisfies $\|\hat{\boldsymbol{v}}_g - \boldsymbol{v}_g\|_2 = o_P(1)$. Define $\hat{\boldsymbol{\Gamma}}_g \in \mathbb{R}^{b\times r_g}$, $r_g \leq b$, to be a matrix whose columns form an orthonormal basis for the null space of $\boldsymbol{A}_g$, where $\boldsymbol{A}_g$ has b columns and forms the set of equality constraints at the optimum $\hat{\boldsymbol{v}}_g$. If there are no equality constraints satisfied at the optimum, $\hat{\boldsymbol{\Gamma}}_g = I_b$. Note that if $\boldsymbol{x}$ is a row of $\boldsymbol{A}_{\mathcal{I}}$ defined in Section 3.2 and $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{v}_g > 0$, then, $\boldsymbol{x}^{\mathrm{T}}\hat{\boldsymbol{v}}_g > 0$ with probability tending to 1. Therefore, $\boldsymbol{v}_g$ will lie in the column space of $\hat{\boldsymbol{\Gamma}}_g$ with probability tending to 1. For $\bar{\boldsymbol{v}}_g = \lambda\boldsymbol{v}_g + (1-\lambda)\hat{\boldsymbol{v}}_g$ for some $\lambda \in [0,1]$, we then have

$$\begin{aligned}\boldsymbol{0} = \hat{\boldsymbol{\Gamma}}_g^{\mathrm{T}}\nabla\hat{l}_{n,g}\left(\hat{\boldsymbol{v}}_g\right) &= \hat{\boldsymbol{\Gamma}}_g^{\mathrm{T}}\nabla\hat{l}_{n,g}\left(\boldsymbol{v}_g\right) + \hat{\boldsymbol{\Gamma}}_g^{\mathrm{T}}\nabla^2\hat{l}_{n,g}\left(\bar{\boldsymbol{v}}_g\right)\left(\hat{\boldsymbol{v}}_g - \boldsymbol{v}_g\right)\\ &= \hat{\boldsymbol{\Gamma}}_g^{\mathrm{T}}\nabla l_{n,g}\left(\boldsymbol{v}_g\right) + \hat{\boldsymbol{\Gamma}}_g^{\mathrm{T}}\nabla^2 l_{n,g}\left(\bar{\boldsymbol{v}}_g\right)\hat{\boldsymbol{\Gamma}}_g\left(\hat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_g\right) + O_P(\eta)\end{aligned}$$

This proves that

$$\|\boldsymbol{V}_g - \hat{\boldsymbol{V}}_g\|_2 = O_P\left\{n^{1/2}\left(\gamma_K p\right)^{-1/2} + n^{-1/2}\right\},$$

and therefore (3.26), since $\|\nabla l_{n,g}\left(\boldsymbol{v}_g\right)\|_2 = O_P\left(n^{-1/2}\right)$ and the eigenvalues of $\nabla^2 l_{n,g}\left(\bar{\boldsymbol{v}}_g\right)$ are bounded above 0 (and below ∞) with probability tending to 1. We can then express $\hat{\boldsymbol{V}}_g^{-1}$ and $\hat{\boldsymbol{W}}_g$ as

$$\|\hat{\boldsymbol{V}}_g^{-1} - \left(\boldsymbol{V}_g^{-1} + \sum_{j=1}^{b}\epsilon_{g,j}\boldsymbol{V}_g^{-1}\boldsymbol{B}_j\boldsymbol{V}_g^{-1}\right)\|_2 = o_P\left(n^{-1/2}\right) \tag{3.64a}$$

$$\|\hat{\boldsymbol{W}}_g^{-1} - \left(\boldsymbol{W}_g^{-1} + \sum_{j=1}^{b}\epsilon_{g,j}\boldsymbol{W}_g^{-1}\tilde{\boldsymbol{B}}_j\boldsymbol{W}_g^{-1}\right)\|_2 = o_P\left(n^{-1/2}\right) \tag{3.64b}$$

$$\epsilon_{g,j} = v_{g,j} - \hat{v}_{g,j}, \tag{3.64c}$$

which we will use to prove (3.27).

Define $\hat{\boldsymbol{D}} = \begin{bmatrix} \boldsymbol{X} & \hat{\boldsymbol{C}} \end{bmatrix}$. The generalized least squares estimate for $\boldsymbol{\beta}_g$ is

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_g &= \left[\left(\hat{\boldsymbol{D}}^{\mathrm{T}}\hat{\boldsymbol{V}}_g^{-1}\hat{\boldsymbol{D}}\right)^{-1}\hat{\boldsymbol{D}}^{\mathrm{T}}\hat{\boldsymbol{V}}_g^{-1}\boldsymbol{y}_g\right]_{1:d} = \left(\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{V}}_g^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{V}}_g^{-1}\boldsymbol{y}_g - \hat{\boldsymbol{\Omega}}_g\hat{\boldsymbol{\ell}}_g \\
&= \boldsymbol{\beta}_g + \boldsymbol{\Omega}_g\boldsymbol{\ell}_g - \hat{\boldsymbol{\Omega}}_g\hat{\boldsymbol{\ell}}_g + \left(\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{V}}_g^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{V}}_g^{-1}\boldsymbol{e}_g \qquad (3.65)
\end{aligned}$$

where here

$$\begin{aligned}
\hat{\boldsymbol{\ell}}_g &= \left(\hat{\boldsymbol{C}}_\perp^{\mathrm{T}}\hat{\boldsymbol{W}}_g^{-1}\hat{\boldsymbol{C}}_\perp\right)^{-1}\hat{\boldsymbol{C}}_\perp^{\mathrm{T}}\hat{\boldsymbol{W}}_g^{-1}\boldsymbol{Q}_X^{\mathrm{T}}\boldsymbol{y}_g \\
\hat{\boldsymbol{\Omega}}_g &= \left(\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{V}}_g^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{V}}_g^{-1}\hat{\boldsymbol{C}} \\
\boldsymbol{\Omega}_g &= \left(\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{V}}_g^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{V}}_g^{-1}\boldsymbol{C}.
\end{aligned}$$

Since $\hat{\boldsymbol{\beta}}_g$ only depends on the column space of $\hat{\boldsymbol{C}}$ and $\boldsymbol{\beta}_g$ is identifiable regardless of the parametrization of $\boldsymbol{C}$, it suffices to assume

$$\begin{aligned}
&(n-d)^{-1}\hat{\boldsymbol{C}}_\perp^{\mathrm{T}}\hat{\boldsymbol{W}}^{-1}\hat{\boldsymbol{C}}_\perp = I_K, \quad np^{-1}\hat{\boldsymbol{L}}^{\mathrm{T}}\hat{\boldsymbol{L}} \text{ is diagonal with non-increasing elements} \\
&(n-d)^{-1}\boldsymbol{C}_\perp^{\mathrm{T}}\hat{\boldsymbol{W}}^{-1}\boldsymbol{C}_\perp = I_K, \quad (n-d)p^{-1}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L} = \mathrm{diag}\left(\lambda_1, \ldots, \lambda_K\right)
\end{aligned}$$

and that the diagonal elements of $\boldsymbol{C}_\perp^{\mathrm{T}}\hat{\boldsymbol{C}}_\perp$ are positive, where $\lambda_1, \ldots, \lambda_K$ are defined in Lemma 3.4. Corollary 3.1, Lemma 3.4 and Remark 3.20 shows that this uniquely identifies $\boldsymbol{L}$ and $\boldsymbol{C}$. We discussed the behavior of this random identification criterion in Remark 3.20.

To prove the theorem, we will prove two relations in lemmas 3.9 and 3.10:

$$n^{1/2}\left\{\boldsymbol{\Omega}_g\left(\boldsymbol{\ell}_g - \hat{\boldsymbol{\ell}}_g\right) + \left(\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{V}}_g^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{V}}_g^{-1}\boldsymbol{e}_g\right\} = \boldsymbol{F} + o_P(1) \qquad (3.66)$$

$$n^{1/2}\|\hat{\boldsymbol{\Omega}}_g - \boldsymbol{\Omega}_g\|_2 = o_P(1) \qquad (3.67)$$

where

$$\boldsymbol{F} \sim N\left\{\mathbf{0}, \left(n^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_g^{-1}\boldsymbol{X}\right)^{-1} + \tilde{\boldsymbol{\Omega}}_g\left(n^{-1}\boldsymbol{C}_{\perp}^{\mathrm{T}}\boldsymbol{W}_g^{-1}\boldsymbol{C}_{\perp}\right)^{-1}\tilde{\boldsymbol{\Omega}}_g^{\mathrm{T}}\right\} \tag{3.68}$$
$$\tilde{\boldsymbol{\Omega}}_g = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_g^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_g^{-1}\boldsymbol{C}.$$

***Lemma* 3.9.** *Under the assumptions of Theorem 3.9, relation* (3.66) *holds.*

Proof. First, note that because $\|\boldsymbol{V}_g - \hat{\boldsymbol{V}}_g\|_2 = o_P(1)$, $\|\tilde{\boldsymbol{\Omega}}_g - \boldsymbol{\Omega}_g\|_2 = o_P(1)$, meaning it suffices to prove the lemma by replacing $\tilde{\boldsymbol{\Omega}}_g$ with $\boldsymbol{\Omega}_g$ in (3.68). To prove the Lemma we need to understand how $\hat{\boldsymbol{\ell}}_g$ behaves. First,

$$\begin{aligned}
\hat{\boldsymbol{\ell}}_g &= \left(\hat{\boldsymbol{C}}_{\perp}^{\mathrm{T}}\hat{\boldsymbol{W}}_g^{-1}\hat{\boldsymbol{C}}_{\perp}\right)^{-1}\hat{\boldsymbol{C}}_{\perp}^{\mathrm{T}}\hat{\boldsymbol{W}}_g^{-1}\boldsymbol{Q}_X^{\mathrm{T}}\boldsymbol{y}_g\\
&= \left(\hat{\boldsymbol{C}}_{\perp}^{\mathrm{T}}\hat{\boldsymbol{W}}_g^{-1}\hat{\boldsymbol{C}}_{\perp}\right)^{-1}\hat{\boldsymbol{C}}_{\perp}^{\mathrm{T}}\hat{\boldsymbol{W}}_g^{-1}\boldsymbol{C}_{\perp}\boldsymbol{\ell}_g + \left(\hat{\boldsymbol{C}}_{\perp}^{\mathrm{T}}\hat{\boldsymbol{W}}_g^{-1}\hat{\boldsymbol{C}}_{\perp}\right)^{-1}\hat{\boldsymbol{C}}_{\perp}^{\mathrm{T}}\hat{\boldsymbol{W}}_g^{-1}\boldsymbol{Q}_X^{\mathrm{T}}\boldsymbol{e}_g\\
&= \boldsymbol{\ell}_g + \left(\boldsymbol{C}_{\perp}^{\mathrm{T}}\hat{\boldsymbol{W}}_g^{-1}\boldsymbol{C}_{\perp}\right)^{-1}\hat{\boldsymbol{C}}_{\perp}^{\mathrm{T}}\hat{\boldsymbol{W}}_g^{-1}\boldsymbol{Q}_X^{\mathrm{T}}\boldsymbol{e}_g + o_P\left(n^{-1/2}\right).
\end{aligned}$$

The last equality follows from Corollary 3.2 and because

$$\begin{aligned}
&\left(n^{-1}\hat{\boldsymbol{C}}_{\perp}^{\mathrm{T}}\hat{\boldsymbol{W}}_g^{-1}\hat{\boldsymbol{C}}_{\perp}\right)^{-1} = \left(n^{-1}\hat{\boldsymbol{C}}_{\perp}^{\mathrm{T}}\boldsymbol{W}_g^{-1}\hat{\boldsymbol{C}}_{\perp} + \sum_{j=1}^{b}\epsilon_{g,j}n^{-1}\hat{\boldsymbol{C}}_{\perp}^{\mathrm{T}}\boldsymbol{W}_g^{-1}\tilde{\boldsymbol{B}}_j\boldsymbol{W}_g^{-1}\hat{\boldsymbol{C}}_{\perp}\right)^{-1}\\
&+ o_P\left(n^{-1/2}\right) = \left(n^{-1}\hat{\boldsymbol{C}}_{\perp}^{\mathrm{T}}\boldsymbol{W}_g^{-1}\hat{\boldsymbol{C}}_{\perp}\right)^{-1} + \sum_{j=1}^{b}\epsilon_{g,j}\left\{\left(n^{-1}\hat{\boldsymbol{C}}_{\perp}^{\mathrm{T}}\boldsymbol{W}_g^{-1}\hat{\boldsymbol{C}}_{\perp}\right)^{-1}\right.\\
&\times\left.\left(n^{-1}\hat{\boldsymbol{C}}_{\perp}^{\mathrm{T}}\boldsymbol{W}_g^{-1}\tilde{\boldsymbol{B}}_j\boldsymbol{W}_g^{-1}\hat{\boldsymbol{C}}_{\perp}\right)\left(n^{-1}\hat{\boldsymbol{C}}_{\perp}^{\mathrm{T}}\boldsymbol{W}_g^{-1}\hat{\boldsymbol{C}}_{\perp}\right)^{-1}\right\} + o_P\left(n^{-1/2}\right)\\
=&\left(n^{-1}\boldsymbol{C}_{\perp}^{\mathrm{T}}\boldsymbol{W}_g^{-1}\boldsymbol{C}_{\perp}\right)^{-1} + \sum_{j=1}^{b}\epsilon_{g,j}\left\{\left(n^{-1}\boldsymbol{C}_{\perp}^{\mathrm{T}}\boldsymbol{W}_g^{-1}\boldsymbol{C}_{\perp}\right)^{-1}\right.\\
&\times\left.\left(n^{-1}\boldsymbol{C}_{\perp}^{\mathrm{T}}\boldsymbol{W}_g^{-1}\tilde{\boldsymbol{B}}_j\boldsymbol{W}_g^{-1}\boldsymbol{C}_{\perp}\right)\left(n^{-1}\boldsymbol{C}_{\perp}^{\mathrm{T}}\boldsymbol{W}_g^{-1}\boldsymbol{C}_{\perp}\right)^{-1}\right\} + o_P\left(n^{-1/2}\right)\\
=&\left(n^{-1}\boldsymbol{C}_{\perp}^{\mathrm{T}}\hat{\boldsymbol{W}}_g^{-1}\boldsymbol{C}_{\perp}\right)^{-1} + o_P\left(n^{-1/2}\right),
\end{aligned}$$

where we can substitute $n^{-1}\boldsymbol{C}_{\perp}^{\mathrm{T}}\hat{\boldsymbol{W}}_g^{-1}\boldsymbol{C}_{\perp}$ for $\hat{\boldsymbol{C}}_{\perp}^{\mathrm{T}}\hat{\boldsymbol{W}}_g^{-1}\boldsymbol{C}_{\perp}$ in the above analysis as well by

Corollary 3.2. Next, $\boldsymbol{Q}_X^{\mathrm{T}}\boldsymbol{e}_g$ is independent of $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_g^{-1}\boldsymbol{e}_g$ by Craig's Theorem and

$$\begin{aligned}n^{-1}\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{V}}_g^{-1}\boldsymbol{e}_g &= n^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_g^{-1}\boldsymbol{e}_g + \sum_{j=0}^{b}\epsilon_{g,j}n^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_g^{-1}\boldsymbol{B}_j\boldsymbol{V}_g^{-1}\boldsymbol{e}_g + o_P\left(n^{-1/2}\right)\\ &= n^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_g^{-1}\boldsymbol{e}_g + o_P\left(n^{-1/2}\right).\end{aligned}$$

Therefore, the proof will be complete if we can show

$$\|n^{-1}\hat{\boldsymbol{C}}_\perp^{\mathrm{T}}\hat{\boldsymbol{W}}_g^{-1}\boldsymbol{Q}_X^{\mathrm{T}}\boldsymbol{e}_g - n^{-1}\boldsymbol{C}_\perp^{\mathrm{T}}\boldsymbol{W}_g^{-1}\boldsymbol{Q}_X^{\mathrm{T}}\boldsymbol{e}_g\|_2 = o_P\left(n^{-1/2}\right),$$

since a simple application of Slutsky's Theorem would give us the result. To show this, we first note that

$$\begin{aligned}&\|n^{-1}\hat{\boldsymbol{C}}_\perp^{\mathrm{T}}\hat{\boldsymbol{W}}_g^{-1}\boldsymbol{Q}_X^{\mathrm{T}}\boldsymbol{e}_g - n^{-1}\boldsymbol{C}_\perp^{\mathrm{T}}\boldsymbol{W}_g^{-1}\boldsymbol{Q}_X^{\mathrm{T}}\boldsymbol{e}_g\|_2\\ \leq&\|n^{-1}\hat{\boldsymbol{C}}_\perp^{\mathrm{T}}\boldsymbol{W}_g^{-1}\boldsymbol{Q}_X^{\mathrm{T}}\boldsymbol{e}_g - n^{-1}\boldsymbol{C}_\perp^{\mathrm{T}}\boldsymbol{W}_g^{-1}\boldsymbol{Q}_X^{\mathrm{T}}\boldsymbol{e}_g\|_2\\ &+\sum_{j=0}^{b}\epsilon_{g,j}\|n^{-1}\hat{\boldsymbol{C}}_\perp^{\mathrm{T}}\boldsymbol{W}_g^{-1}\tilde{\boldsymbol{B}}_j\boldsymbol{W}_g^{-1}\boldsymbol{Q}_X^{\mathrm{T}}\boldsymbol{e}_g - n^{-1}\boldsymbol{C}_\perp^{\mathrm{T}}\boldsymbol{W}_g^{-1}\tilde{\boldsymbol{B}}_j\boldsymbol{W}_g^{-1}\boldsymbol{Q}_X^{\mathrm{T}}\boldsymbol{e}_g\|_2 + o_P\left(n^{-1/2}\right).\end{aligned}$$

We will first prove

$$\|n^{-1}\hat{\boldsymbol{C}}_\perp^{\mathrm{T}}\boldsymbol{W}_g^{-1}\boldsymbol{Q}_X^{\mathrm{T}}\boldsymbol{e}_g - n^{-1}\boldsymbol{C}_\perp^{\mathrm{T}}\boldsymbol{W}_g^{-1}\boldsymbol{Q}_X^{\mathrm{T}}\boldsymbol{e}_g\|_2 = o_P\left(n^{-1/2}\right) \tag{3.69}$$

and then an identical analysis can be used to show

$$\|n^{-1}\hat{\boldsymbol{C}}_\perp^{\mathrm{T}}\boldsymbol{W}_g^{-1}\tilde{\boldsymbol{B}}_j\boldsymbol{W}_g^{-1}\boldsymbol{Q}_X^{\mathrm{T}}\boldsymbol{e}_g - n^{-1}\boldsymbol{C}_\perp^{\mathrm{T}}\boldsymbol{W}_g^{-1}\tilde{\boldsymbol{B}}_j\boldsymbol{W}_g^{-1}\boldsymbol{Q}_X^{\mathrm{T}}\boldsymbol{e}_g\|_2 = o_P\left(n^{-1/2}\right).$$

To prove (3.69), we first define

$$\bar{\boldsymbol{C}}_\perp = (n-d)^{-1/2}\hat{\boldsymbol{W}}^{-1/2}\boldsymbol{C}_\perp$$
$$\bar{\boldsymbol{Q}}_\perp = \boldsymbol{Q}_{\bar{C}_\perp}.$$

Then using (3.36), (3.37) and (3.38) in the proof of Theorem 3.10, we have that

$$\hat{\boldsymbol{C}}_\perp = (n-d)^{1/2}\hat{\boldsymbol{W}}^{1/2}\hat{\bar{\boldsymbol{C}}}_\perp = (n-d)^{1/2}\hat{\boldsymbol{W}}^{1/2}\bar{\boldsymbol{C}}_\perp\hat{\boldsymbol{v}} + (n-d)^{1/2}\hat{\boldsymbol{W}}^{1/2}\bar{\boldsymbol{Q}}_\perp\hat{\boldsymbol{z}}. \tag{3.70}$$

We first see that

$$\begin{aligned} n^{-1}(n-d)^{1/2}\hat{\boldsymbol{v}}^{\mathrm{T}}\bar{\boldsymbol{C}}_\perp^{\mathrm{T}}\hat{\boldsymbol{W}}^{1/2}\boldsymbol{W}_g^{-1}\boldsymbol{Q}_X^{\mathrm{T}}\boldsymbol{e}_g &= n^{-1}(n-d)^{1/2}\bar{\boldsymbol{C}}_\perp^{\mathrm{T}}\hat{\boldsymbol{W}}^{1/2}\boldsymbol{W}_g^{-1}\boldsymbol{Q}_X^{\mathrm{T}}\boldsymbol{e}_g + o_P\left(n^{-1/2}\right) \\ &= n^{-1}\boldsymbol{C}_\perp^{\mathrm{T}}\boldsymbol{W}_g^{-1}\boldsymbol{Q}_X^{\mathrm{T}}\boldsymbol{e}_g + o_P\left(n^{-1/2}\right). \end{aligned}$$

To then show (3.69) and complete the proof, we need only show that

$$\|\hat{\boldsymbol{z}}^{\mathrm{T}}\bar{\boldsymbol{Q}}_\perp^{\mathrm{T}}\hat{\boldsymbol{W}}^{1/2}\boldsymbol{W}_g^{-1}\boldsymbol{Q}_X^{\mathrm{T}}\boldsymbol{e}_g\|_2 = o_P\left(1\right).$$

$\hat{\boldsymbol{z}}$ is such that

$$\|\hat{\boldsymbol{z}} - (p\gamma_K)^{-1/2}\,\bar{\boldsymbol{Q}}_\perp^{\mathrm{T}}\hat{\boldsymbol{W}}^{-1/2}\tilde{\boldsymbol{E}}^{\mathrm{T}}\left\{(p\gamma_K)^{-1/2}\,\tilde{\boldsymbol{E}}\hat{\boldsymbol{W}}^{-1/2}\bar{\boldsymbol{C}}_\perp + \bar{\boldsymbol{L}}\right\}\|_2 = o_P\left(n^{-1/2}\right)$$

where $\tilde{\boldsymbol{E}} = \boldsymbol{E}\boldsymbol{Q}_X$. Therefore,

$$\begin{aligned} \hat{\boldsymbol{z}}^{\mathrm{T}}\bar{\boldsymbol{Q}}_\perp^{\mathrm{T}}\hat{\boldsymbol{W}}^{1/2}\boldsymbol{W}_g^{-1}\boldsymbol{Q}_X^{\mathrm{T}}\boldsymbol{e}_g &= \underbrace{(p\gamma_K)^{-1/2}\,\bar{\boldsymbol{L}}^{\mathrm{T}}\tilde{\boldsymbol{E}}\boldsymbol{Q}_{C_\perp}\left(\boldsymbol{Q}_{C_\perp}^{\mathrm{T}}\hat{\boldsymbol{W}}\boldsymbol{Q}_{C_\perp}\right)^{-1}\boldsymbol{Q}_{C_\perp}^{\mathrm{T}}\hat{\boldsymbol{W}}\boldsymbol{W}_g^{-1}\tilde{\boldsymbol{e}}_g}_{(1)} \\ &+ \underbrace{\gamma_K^{-1}\,(n-d)^{-1/2}\,\boldsymbol{C}_\perp^{\mathrm{T}}\hat{\boldsymbol{W}}^{-1}\left(p^{-1}\tilde{\boldsymbol{E}}^{\mathrm{T}}\tilde{\boldsymbol{E}}\right)\boldsymbol{Q}_{C_\perp}\left(\boldsymbol{Q}_{C_\perp}^{\mathrm{T}}\hat{\boldsymbol{W}}\boldsymbol{Q}_{C_\perp}\right)^{-1}\boldsymbol{Q}_{C_\perp}^{\mathrm{T}}\hat{\boldsymbol{W}}\boldsymbol{W}_g^{-1}\tilde{\boldsymbol{e}}_g}_{(2)} + o_P(1). \end{aligned}$$

We now go through each one of these terms.

(1) Since this is a continuously differentiable function of $\hat{\boldsymbol{W}}$ and $\|\hat{\boldsymbol{W}}-\boldsymbol{W}^*\|_2 = O_P\left(n^{-1}\right)$,

$$\left\| (p\gamma_K)^{-1/2}\, \bar{\boldsymbol{L}}^{\mathrm{T}} \tilde{\boldsymbol{E}} \boldsymbol{Q}_{C_\perp} \left(\boldsymbol{Q}_{C_\perp}^{\mathrm{T}} \hat{\boldsymbol{W}} \boldsymbol{Q}_{C_\perp}\right)^{-1} \boldsymbol{Q}_{C_\perp}^{\mathrm{T}} \hat{\boldsymbol{W}} \boldsymbol{W}_g^{-1} \tilde{\boldsymbol{e}}_g \right.$$
$$\left. - (p\gamma_K)^{-1/2}\, \bar{\boldsymbol{L}}^{\mathrm{T}} \tilde{\boldsymbol{E}} \boldsymbol{Q}_{C_\perp} \left(\boldsymbol{Q}_{C_\perp}^{\mathrm{T}} \boldsymbol{W}^* \boldsymbol{Q}_{C_\perp}\right)^{-1} \boldsymbol{Q}_{C_\perp}^{\mathrm{T}} \boldsymbol{W}^* \boldsymbol{W}_g^{-1} \tilde{\boldsymbol{e}}_g \right\|_2$$
$$= O_P\left\{ n^{1/2} (p\gamma_K)^{-1/2} \right\} = o_P(1).$$

Define the non-random matrix

$$\boldsymbol{M}_g = \boldsymbol{Q}_{C_\perp} \left(\boldsymbol{Q}_{C_\perp}^{\mathrm{T}} \boldsymbol{W}^* \boldsymbol{Q}_{C_\perp}\right)^{-1} \boldsymbol{Q}_{C_\perp}^{\mathrm{T}} \boldsymbol{W}^* \boldsymbol{W}_g^{-1} \in \mathbb{R}^{(n-d)\times(n-d)}.$$

and the random vector $\boldsymbol{w}_g = \boldsymbol{M}_g \tilde{\boldsymbol{e}}_g \sim N\left(0, \boldsymbol{M}_g \boldsymbol{W}_g \boldsymbol{M}_g^{\mathrm{T}}\right)$. Then

$$(p\gamma_K)^{-1/2}\, \bar{\boldsymbol{L}}^{\mathrm{T}} \tilde{\boldsymbol{E}} \boldsymbol{w}_g = (p\gamma_K)^{-1/2} \sum_{h\neq g} \bar{\boldsymbol{\ell}}_h \tilde{\boldsymbol{e}}_h^{\mathrm{T}} \boldsymbol{w}_g + \underbrace{(p\gamma_K)^{-1/2}\, \bar{\boldsymbol{\ell}}_g \tilde{\boldsymbol{e}}_g^{\mathrm{T}} \boldsymbol{w}_g}_{O_P\left\{n^{3/2}(\gamma_K p)^{-1}\right\} = o_P(1)}$$

with

$$\mathbb{E}\left\{ (p\gamma_K)^{-1/2} \sum_{h\neq g} \bar{\boldsymbol{\ell}}_h \tilde{\boldsymbol{e}}_h^{\mathrm{T}} \boldsymbol{w}_g \right\} = \mathbb{E}\left\{ (p\gamma_K)^{-1/2} \sum_{h\neq g} \bar{\boldsymbol{\ell}}_h \tilde{\boldsymbol{e}}_h^{\mathrm{T}} \boldsymbol{w}_g \mid \boldsymbol{w}_g \right\} = \boldsymbol{0}$$

$$\begin{aligned}
\mathbb{V}\left\{ (p\gamma_K)^{-1/2} \sum_{h\neq g} \bar{\boldsymbol{\ell}}_h \tilde{\boldsymbol{e}}_h^{\mathrm{T}} \boldsymbol{w}_g \mid \boldsymbol{w}_g \right\} &= (p\gamma_K)^{-1} \sum_{h\neq g} \mathbb{V}\left(\bar{\boldsymbol{\ell}}_h \tilde{\boldsymbol{e}}_h^{\mathrm{T}} \boldsymbol{w}_g \mid \boldsymbol{w}_g\right) \\
&= \frac{1}{\lambda p} \sum_{h\neq g} \left(\boldsymbol{w}_g^{\mathrm{T}} \boldsymbol{W}_h \boldsymbol{w}_g\right) \bar{\boldsymbol{\ell}}_h \bar{\boldsymbol{\ell}}_h^{\mathrm{T}} \\
&\preceq c\,(p\gamma_K)^{-1} \|\boldsymbol{w}_g\|_2^2 \bar{\boldsymbol{L}}^{\mathrm{T}} \bar{\boldsymbol{L}}
\end{aligned}$$

where c bounds the eigenvalues of $\boldsymbol{W}_h$ $(h = 1, \ldots, p)$ from above. Therefore,

$$\mathbb{V}\left\{(p\gamma_K)^{-1/2}\sum_{h\neq g}\bar{\boldsymbol{\ell}}_h\tilde{\boldsymbol{e}}_h^{\mathrm{T}}\boldsymbol{w}_g\right\} = O\left\{n\left(p\gamma_K\right)^{-1}\right\} = o(1).$$

This shows that $\|1.)\|_2 = o_P(1)$.

(2) For this, we use the same technique to replace $\hat{\boldsymbol{W}}$ with $\boldsymbol{W}^*$, which only differs from 2.) by $O_P\left(n^{-1/2}\right) = o_P(1)$ in 2-norm. Next, we define non-random the matrix $\boldsymbol{A} \in \mathbb{R}^{(n-d)\times K}$ to be a matrix whose columns form an orthonormal basis for the column space of $\boldsymbol{C}_{\perp}$. For $\boldsymbol{w}_g$ defined above, note that $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{w}_g = \boldsymbol{0}$. Therefore,

$$\begin{aligned}2.) = \gamma_K^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{W}^{*-1}\left(p^{-1}\tilde{\boldsymbol{E}}^{\mathrm{T}}\tilde{\boldsymbol{E}}\right)\boldsymbol{w}_g =& \gamma_K^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{W}^{*-1}\left(p^{-1}\tilde{\boldsymbol{E}}_{-g}^{\mathrm{T}}\tilde{\boldsymbol{E}}_{-g}\right)\boldsymbol{w}_g \\ &+ \underbrace{\left(\gamma_K p\right)^{-1}\boldsymbol{A}^{\mathrm{T}}\tilde{\boldsymbol{e}}_g\tilde{\boldsymbol{e}}_g^{\mathrm{T}}\boldsymbol{w}_g}_{O_P\left\{n(\gamma_K p)^{-1}\right\} = o_P(1)}\end{aligned}$$

where $\tilde{\boldsymbol{E}}_{-g}$ is the $\tilde{\boldsymbol{E}}$ with the g^{th} row removed. Note that

$$\begin{aligned}&\mathbb{E}\left\{\boldsymbol{A}^{\mathrm{T}}\boldsymbol{W}^{*-1}\left(p^{-1}\tilde{\boldsymbol{E}}_{-g}^{\mathrm{T}}\tilde{\boldsymbol{E}}_{-g}\right)\boldsymbol{w}_g \mid \boldsymbol{w}_g\right\} = p^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{W}^{*-1}\boldsymbol{W}_g\boldsymbol{w}_g \\ \sim & p^{-1}N_K\left(\boldsymbol{0}, \boldsymbol{A}^{\mathrm{T}}\boldsymbol{W}^{*-1}\boldsymbol{W}_g\boldsymbol{M}_g\boldsymbol{W}_g\boldsymbol{M}_g\boldsymbol{W}_g\boldsymbol{W}^{*-1}\boldsymbol{A}\right)\end{aligned}$$

Since K is fixed and finite, it suffices to assume $K = 1$. If we let $\boldsymbol{u} = \boldsymbol{W}^{*-1}\boldsymbol{A} \in \mathbb{R}^{(n-d)\times K}$ (which has bounded 2-norm), then

$$\begin{aligned}\mathbb{V}\left\{\boldsymbol{A}^{\mathrm{T}}\boldsymbol{W}^{*-1}\left(p^{-1}\tilde{\boldsymbol{E}}_{-g}^{\mathrm{T}}\tilde{\boldsymbol{E}}_{-g}\right)\boldsymbol{w}_g \mid \boldsymbol{w}_g\right\} &= p^{-2}\sum_{h\neq g}\mathbb{V}\left\{\boldsymbol{u}^{\mathrm{T}}\tilde{\boldsymbol{e}}_h\tilde{\boldsymbol{e}}_h^{\mathrm{T}}\boldsymbol{w}_g \mid \boldsymbol{w}_g\right\} \\ &= p^{-2}O\left(p\|\boldsymbol{w}_g\|_2^2\right) = O\left(p^{-1}\|\boldsymbol{w}_g\|_2^2\right).\end{aligned}$$

This shows that $\|2.)\|_2 = o_P(1)$, and completes the proof.

□

***Lemma* 3.10.** *Under the assumptions Theorem 3.9, relation* (3.67) *holds.*

Proof. From the expression for $\boldsymbol{C}$ and $\hat{\boldsymbol{C}}$ in item *(c)* of Algorithm 3.1 and equation (3.12), we can write $\boldsymbol{\Omega}_g$ and $\hat{\boldsymbol{\Omega}}_g$ as

$$\boldsymbol{\Omega}_g = \left(\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{V}}_g^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{V}}_g^{-1}\boldsymbol{C} = \boldsymbol{\Omega} + \left(\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{V}}_g^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{V}}_g^{-1}\hat{\boldsymbol{V}}\boldsymbol{Q}_{\boldsymbol{X}}\hat{\boldsymbol{W}}^{-1}\boldsymbol{C}_{\perp}$$
$$\hat{\boldsymbol{\Omega}}_g = \left(\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{V}}_g^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{V}}_g^{-1}\hat{\boldsymbol{C}} = \hat{\boldsymbol{\Omega}} + \left(\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{V}}_g^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{V}}_g^{-1}\hat{\boldsymbol{V}}\boldsymbol{Q}_{\boldsymbol{X}}\hat{\boldsymbol{W}}^{-1}\hat{\boldsymbol{C}}_{\perp}.$$

By Lemma 3.2, $\|\boldsymbol{\Omega} - \hat{\boldsymbol{\Omega}}\|_2 = o_P\left(n^{-1/2}\right)$. Therefore, to prove the current lemma, we need only show that

$$\|n^{-1}\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{V}}_g^{-1}\hat{\boldsymbol{V}}\boldsymbol{Q}_{\boldsymbol{X}}\hat{\boldsymbol{W}}^{-1}\boldsymbol{C}_{\perp} - n^{-1}\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{V}}_g^{-1}\hat{\boldsymbol{V}}\boldsymbol{Q}_{\boldsymbol{X}}\hat{\boldsymbol{W}}^{-1}\hat{\boldsymbol{C}}_{\perp}\|_2 = o_P\left(n^{-1/2}\right).$$

Just as we did in the proof of Lemma 3.9, we use (3.64) to expand the above equation:

$$\begin{aligned}
&\|n^{-1}\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{V}}_g^{-1}\hat{\boldsymbol{V}}\boldsymbol{Q}_{\boldsymbol{X}}\hat{\boldsymbol{W}}^{-1}\boldsymbol{C}_{\perp} - n^{-1}\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{V}}_g^{-1}\hat{\boldsymbol{V}}\boldsymbol{Q}_{\boldsymbol{X}}\hat{\boldsymbol{W}}^{-1}\hat{\boldsymbol{C}}_{\perp}\|_2 \leq \\
&\|n^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_g^{-1}\hat{\boldsymbol{V}}\boldsymbol{Q}_{\boldsymbol{X}}\hat{\boldsymbol{W}}^{-1}\boldsymbol{C}_{\perp} - n^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_g^{-1}\hat{\boldsymbol{V}}\boldsymbol{Q}_{\boldsymbol{X}}\hat{\boldsymbol{W}}^{-1}\hat{\boldsymbol{C}}_{\perp}\|_2 \\
+&\sum_{j=1}^{b}\epsilon_{g,j}\|n^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_g^{-1}\boldsymbol{B}_j\boldsymbol{V}_g^{-1}\hat{\boldsymbol{V}}\boldsymbol{Q}_{\boldsymbol{X}}\hat{\boldsymbol{W}}^{-1}\boldsymbol{C}_{\perp} - n^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_g^{-1}\boldsymbol{B}_j\boldsymbol{V}_g^{-1}\hat{\boldsymbol{V}}\boldsymbol{Q}_{\boldsymbol{X}}\hat{\boldsymbol{W}}^{-1}\hat{\boldsymbol{C}}_{\perp}\|_2 \\
+&o_P\left(n^{-1/2}\right)
\end{aligned}$$

where again we will first show

$$\|n^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_g^{-1}\hat{\boldsymbol{V}}\boldsymbol{Q}_{\boldsymbol{X}}\hat{\boldsymbol{W}}^{-1}\boldsymbol{C}_{\perp} - n^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_g^{-1}\hat{\boldsymbol{V}}\boldsymbol{Q}_{\boldsymbol{X}}\hat{\boldsymbol{W}}^{-1}\hat{\boldsymbol{C}}_{\perp}\|_2 = o_P\left(n^{-1/2}\right). \tag{3.71}$$

An identical argument can then be applied to show that

$$\begin{aligned}
&\|n^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_g^{-1}\boldsymbol{B}_j\boldsymbol{V}_g^{-1}\hat{\boldsymbol{V}}\boldsymbol{Q}_{\boldsymbol{X}}\hat{\boldsymbol{W}}^{-1}\boldsymbol{C}_{\perp} - n^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_g^{-1}\boldsymbol{B}_j\boldsymbol{V}_g^{-1}\hat{\boldsymbol{V}}\boldsymbol{Q}_{\boldsymbol{X}}\hat{\boldsymbol{W}}^{-1}\hat{\boldsymbol{C}}_{\perp}\|_2 \\
=&o_P\left(n^{-1/2}\right).
\end{aligned}$$

First, by the expression for $\hat{\boldsymbol{C}}_\perp$ in (3.70),

$$n^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_g^{-1}\hat{\boldsymbol{V}}\boldsymbol{Q}_{\boldsymbol{X}}\hat{\boldsymbol{W}}^{-1}\hat{\boldsymbol{C}}_\perp = n^{-1}(n-d)^{1/2}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_g^{-1}\hat{\boldsymbol{V}}\boldsymbol{Q}_{\boldsymbol{X}}\hat{\boldsymbol{W}}^{-1/2}\bar{\boldsymbol{C}}_\perp\hat{\boldsymbol{v}} + n^{-1}(n-d)^{1/2}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_g^{-1}\hat{\boldsymbol{V}}\boldsymbol{Q}_{\boldsymbol{X}}\hat{\boldsymbol{W}}^{-1/2}\bar{\boldsymbol{Q}}_\perp\hat{\boldsymbol{z}}$$

To show (3.71),

$$\|n^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_g^{-1}\hat{\boldsymbol{V}}\boldsymbol{Q}_{\boldsymbol{X}}\hat{\boldsymbol{W}}^{-1}\boldsymbol{C}_\perp - n^{-1}(n-d)^{1/2}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_g^{-1}\hat{\boldsymbol{V}}\boldsymbol{Q}_{\boldsymbol{X}}\hat{\boldsymbol{W}}^{-1/2}\bar{\boldsymbol{C}}_\perp\hat{\boldsymbol{v}}\|_2 = o_P\left(n^{-1/2}\right)$$

by (3.70) and because $\|\hat{\boldsymbol{v}} - I_K\|_2 = o_P\left(n^{-1/2}\right)$. Therefore, we need only show that

$$\|\boldsymbol{u}^{\mathrm{T}}\hat{\boldsymbol{V}}\boldsymbol{Q}_{\boldsymbol{X}}\hat{\boldsymbol{W}}^{-1/2}\bar{\boldsymbol{Q}}_\perp\hat{\boldsymbol{z}}\|_2 = o_P\left(n^{-1/2}\right)$$

where $\boldsymbol{u} = n^{-1/2}\boldsymbol{V}_g^{-1}\boldsymbol{X} \in \mathbb{R}^{n\times d}$ has bounded 2-norm. By the expression for $\hat{\boldsymbol{z}}$ in (3.37),

$$\begin{aligned}\boldsymbol{u}^{\mathrm{T}}\hat{\boldsymbol{V}}\boldsymbol{Q}_{\boldsymbol{X}}\hat{\boldsymbol{W}}^{-1/2}\bar{\boldsymbol{Q}}_\perp\hat{\boldsymbol{z}} &= (\gamma_K p)^{-1/2}\boldsymbol{u}^{\mathrm{T}}\hat{\boldsymbol{V}}\boldsymbol{Q}_{\boldsymbol{X}}\boldsymbol{Q}_{\boldsymbol{C}_\perp}\left(\boldsymbol{Q}_{\boldsymbol{C}_\perp}^{\mathrm{T}}\hat{\boldsymbol{W}}\boldsymbol{Q}_{\boldsymbol{C}_\perp}\right)^{-1}\boldsymbol{Q}_{\boldsymbol{C}_\perp}^{\mathrm{T}}\tilde{\boldsymbol{E}}^{\mathrm{T}}\bar{\boldsymbol{L}}\\
&+ \gamma_K^{-1}(n-d)^{-1/2}\boldsymbol{u}^{\mathrm{T}}\hat{\boldsymbol{V}}\boldsymbol{Q}_{\boldsymbol{X}}\boldsymbol{Q}_{\boldsymbol{C}_\perp}\left(\boldsymbol{Q}_{\boldsymbol{C}_\perp}^{\mathrm{T}}\hat{\boldsymbol{W}}\boldsymbol{Q}_{\boldsymbol{C}_\perp}\right)^{-1}\boldsymbol{Q}_{\boldsymbol{C}_\perp}^{\mathrm{T}}\left(p^{-1}\tilde{\boldsymbol{E}}^{\mathrm{T}}\tilde{\boldsymbol{E}}\right)\hat{\boldsymbol{W}}^{-1}\boldsymbol{C}_\perp\\
&+ o_P\left(n^{-1/2}\right)\end{aligned}$$

where $\tilde{\boldsymbol{E}} = \boldsymbol{E}\boldsymbol{Q}_{\boldsymbol{X}}$. Since $\|\hat{\boldsymbol{V}} - \boldsymbol{V}^*\|_2, \|\hat{\boldsymbol{W}} - \boldsymbol{W}^*\|_2 = O_P\left(n^{-1}\right)$, we can use identical techniques used in the proof of Lemma 3.9 to show that

$$\|\boldsymbol{u}^{\mathrm{T}}\hat{\boldsymbol{V}}\boldsymbol{Q}_{\boldsymbol{X}}\hat{\boldsymbol{W}}^{-1/2}\bar{\boldsymbol{Q}}_\perp\hat{\boldsymbol{z}}\|_2 = O_P\left\{n^{-1} + (\gamma_K p)^{-1/2}\right\} = o_P\left(n^{-1/2}\right)$$

which completes the proof. □

Going back to the expression for $\hat{\boldsymbol{\beta}}_g$ in (3.65), Lemmas 3.9 and 3.10 show that

$$\begin{aligned} n^{1/2}\left(\hat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_g\right) =& n^{1/2}\boldsymbol{\Omega}_g\left(\boldsymbol{\ell}_g - \hat{\boldsymbol{\ell}}_g\right) + n^{1/2}\left(\boldsymbol{\Omega}_g - \hat{\boldsymbol{\Omega}}_g\right)\hat{\boldsymbol{\ell}}_g \\ &+ n^{1/2}\left(\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{V}}_g^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{V}}_g^{-1}\boldsymbol{e}_g = \boldsymbol{F} + o_P(1) \end{aligned}$$

where $\boldsymbol{F}$ was defined in (3.68). To complete the proof, Corollary 3.2 shows that

$$\|n^{-1}\boldsymbol{C}_\perp^{\mathrm{T}}\boldsymbol{W}_g^{-1}\boldsymbol{C}_\perp - \hat{\boldsymbol{C}}_\perp^{\mathrm{T}}\boldsymbol{W}_g^{-1}\hat{\boldsymbol{C}}_\perp\|_2 = o_P(1).$$

Next

$$\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_g^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_g^{-1}\boldsymbol{C} = \boldsymbol{\Omega} + \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_g^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_g^{-1}\hat{\boldsymbol{V}}\boldsymbol{Q}_X\hat{\boldsymbol{W}}^{-1}\boldsymbol{C}_\perp,$$

where Theorem 3.10 and Lemma 3.2 show that

$$\|\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_g^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_g^{-1}\boldsymbol{C} - \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_g^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_g^{-1}\hat{\boldsymbol{C}}\|_2 = o_P(1).$$

The result then follows from the fact that $\|\boldsymbol{V}_g - \hat{\boldsymbol{V}}_g\|_2 = o_P(1)$.

□

3.14 Completing the proofs of Theorems 3.4, 3.5 and 3.6

Proof of Theorem 3.4. All rates given in the proof of Theorem 3.7 only depend on $\boldsymbol{C}$ through the constants c_1, ϵ_1, c_2 and $\gamma_1, \ldots, \gamma_K$ defined in Assumptions 3.7 and 3.8. That is, (3.22) of Theorem 3.7 can be expressed as follows: Let $T = T\left(\boldsymbol{C}, \boldsymbol{e}_1, \ldots, \boldsymbol{e}_p\right)$ be a statistic and define the event

$$E = \left\{\boldsymbol{C} \text{ is such that the assumptions needed to prove Theorem 3.7 hold}\right\}.$$

Then for all $\eta > 0$, there exists a constant M_η large enough such that for all $\boldsymbol{C} \in E$,

$$\mathbb{P}\left(|T| \geq M_\eta \mid \boldsymbol{C}\right) \leq \eta.$$

To prove the theorem, note that by Assumptions 3.4 and 3.5, $\mathbb{P}(E) \to 1$ as $n \to \infty$. Therefore, as $n \to \infty$,

$$\mathbb{P}\left(|T| \geq M_\eta\right) \leq \mathbb{E}\left\{I(E)\,\mathbb{P}\left(|T| \geq M_\eta \mid \boldsymbol{C}\right)\right\} + \mathbb{P}(E^c) \leq \eta + o(1)$$

where E^c is the complement of E. $\square$

Proof of Theorem 3.5. Define the statistic $T_k = n\{g(k) - 1\}$ for $k = 0, 1, \ldots, K_{\max}$, where $g(k)$ is defined in (3.50) of the proof of Theorem 3.8. Just as we did in the proof of Theorem 3.4, define the event

$$E = \{\boldsymbol{C} \text{ is such that the assumptions needed to prove Theorem 3.8 hold}\}.$$

The proof of Theorem 3.8 shows that for all $\xi > 0$, there exists an N large enough so that

$$\mathbb{P}\left(\operatorname*{arg\,min}_{k \in \{0,1,\ldots,K_{\max}\}} T_k \neq K \mid \boldsymbol{C}\right) < \xi$$

for all $\boldsymbol{C} \in E$ and $n \geq N$. By Lemma 3.1, Remark 3.17 and since $P(E) \to 1$ as $n \to \infty$, the result follows. $\square$

Proof of Theorem 3.6. Let

$$E = \{\boldsymbol{C} \text{ is such that the assumptions needed to prove Theorem 3.9 hold}\}.$$

Fix a $g \in [p]$, let $\hat{\boldsymbol{Z}}_g = \hat{\boldsymbol{M}}_n^{-1/2}\left(\hat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_g\right)$ and assume $\boldsymbol{C} \in E$. The proof of Theorem 3.9 shows that there exists a measurable function $\tilde{\boldsymbol{Z}}_{n,g} = \tilde{\boldsymbol{Z}}_{n,g}\left(\boldsymbol{C}, \boldsymbol{e}_1, \ldots, \boldsymbol{e}_p\right) \in \mathbb{R}^d$ with

$\tilde{\boldsymbol{Z}}_{n,g} \mid \boldsymbol{C} \sim N(\boldsymbol{0}, I_d)$ such that for all $\xi, \epsilon > 0$, there exists an $M > 0$ such that

$$\mathbb{P}\left(\|\hat{\boldsymbol{Z}}_g - \tilde{\boldsymbol{Z}}_{n,g}\|_2 \geq \epsilon \mid \boldsymbol{C}\right), \mathbb{P}\left(\|\hat{\boldsymbol{M}}_g \boldsymbol{M}_g^{-1} - I_d\|_2 \geq \epsilon \mid \boldsymbol{C}\right) < \xi$$

for all $n \geq M$, where $\hat{\boldsymbol{M}}_g, \boldsymbol{M}_g$ are defined in the statement of Theorem 3.6. Since $P(E) \to 1$ as $n \to \infty$, we simply define $\boldsymbol{Z}_{n,g} = I(\boldsymbol{C} \in E)\, \tilde{\boldsymbol{Z}}_{n,g} + I(\boldsymbol{C} \in E^c)\, \tilde{\boldsymbol{S}}_{n,g}$ for an arbitrary random variable $\tilde{\boldsymbol{S}}_{n,g} \sim N(\boldsymbol{0}, I_d)$ that is independent on $\boldsymbol{C}$. We then see that

$$\mathbb{P}\left(\|\hat{\boldsymbol{Z}}_g - \boldsymbol{Z}_{n,g}\|_2 \geq \epsilon\right) = \mathbb{E}\left\{I(\boldsymbol{C} \in E)\, \mathbb{P}\left(\|\hat{\boldsymbol{Z}}_g - \tilde{\boldsymbol{Z}}_{n,g}\|_2 \geq \epsilon \mid \boldsymbol{C}\right)\right\} + o(1) \leq \xi$$

for all n large enough. The result then follows. $\square$

CHAPTER 4

LONGITUDINAL STUDIES REVEAL STRONG GENETIC AND WEAK NON-GENETIC COMPONENTS OF ETHNICITY-DEPENDENT BLOOD DNA METHYLATION LEVELS

4.1 Introduction

Epigenetic patterning in human genomes reflects the contributions of genetic variation (Bell et al. 2011, Smith et al. 2014) exposure histories (Chatterton et al. 2017, Goodrich et al. 2016, Joubert et al. 2016, Kippler et al. 2013, C. et al. 2013, Rzehak et al. 2016), and biological factors, such as age (Bocklandt et al. 2011, Horvath 2013, Horvath et al. 2014, Johnson et al. 2012, Knight et al. 2016, Levine & Crimmins 2014, Marioni et al. 2015, Parets et al. 2013, Schroeder et al. 2011, Davey Smith et al. 2015), ethnicity (Adkins et al. 2011, Galanter et al. 2017, Heyn et al. 2013, Moen et al. 2013, Mozhui et al. 2015, Rahmani et al. 2017) and disease status (Chan et al. 2017, Ladd-Acosta et al. 2013, Nicodemus-Johnson et al. 2016, Rutledge et al. 2017), among others. However, little work has been done to elucidate the relative contributions or longitudinal dynamics of each on epigenetic patterning.

To directly examine the relationship between age, ethnicity, genetic variation, early life exposures and allergic phenotypes and an epigenetic mark, we studied global DNA methylation patterns at over 750,000 CpG sites on the EPIC array in cord blood mononuclear cells (CBMCs) collected at birth and in peripheral blood mononuclear cells (PBMCs) collected at 7 years of age from 196 children participating in the Urban Environment and Childhood Asthma (URECA) birth cohort study (Gern et al. 2009, O'Connor et al. 2018). This cohort is part of the NIAID-funded Inner City Asthma Consortium and is comprised of children primarily of Black and Hispanic self-reported ethnicity, with a mother and/or father with a history of at least one allergic disease living in poor urban areas (see O'Connor et al. (2018)

for details of enrollment criteria). Mothers of children in the URECA study were enrolled during pregnancy and children were followed from birth through at least 7 years of age.

The longitudinal design of the URECA study provided us with the resolution to partition genetic from non-genetic effects on ancestry-associated DNA methylation patterns, and yielded new insight into the factors affecting DNA methylation patterns at CpG sites in mononuclear (immune) cells during early life in ethnically admixed children. Using a novel statistical method that provides a general framework for analyzing longitudinal genetic and epigenetic data, we show that race/ethnicity-dependent methylation patterns are conserved over the first 7 years of life and that these patterns are strongly influenced, and often mediated, by local genotype. Further, chronological age, but not measured exposures during pre- or post-natal periods or disease status by age 7, was associated with methylation patterns in these children. Considering the results of our study and those of a recently published comprehensive review on environmental epigenetics research (Breton et al. 2017), we suggest that methylation levels in blood are not as responsive to environmental exposures as previously suggested (Galanter et al. 2017), at least during the first 7 years of life.

4.2 Statistical methods

4.2.1 Univariate statistical models

We used a standard linear regression model to determine the variable $\boldsymbol{X}$'s effect on methylation at age $a = 0$ or 7:

$$\boldsymbol{y}_g^{(a)} = \boldsymbol{X}\beta_g^{(a)} + \boldsymbol{Z}\boldsymbol{\gamma}_g^{(a)} + \boldsymbol{C}\boldsymbol{\ell}_g^{(a)} + \boldsymbol{e}_g^{(a)}, \quad \boldsymbol{e}_g^{(a)} \sim N_n\left(0, \sigma_{0,g}^2 I_n\right), \tag{4.1}$$

where we use $\sigma_{0,g}^2$ instead of σ_g^2 to distinguish it from the residual variance σ_g^2 in (4.3). Here, $\boldsymbol{y}_g^{(a)}$ is an n-vector containing the M-values at CpG site g at age a, where $g = 1, \ldots, p = 784, 484$. Unless otherwise stated, the covariates $\boldsymbol{Z}$ included gender, sample collection site and plate number (the methylation data were collected on a total of five plates). We estimated the unobserved confounders $\boldsymbol{C}$ using the method described in Chapter 2. Our estimate for $\boldsymbol{C}$ was highly correlated with estimated PBMC cell proportions, which were estimated prior to collecting methylation data in 190 of the 196 PBMC samples, and our results did not change when we included estimated cell proportion in $\boldsymbol{Z}$. Therefore, we chose to not include estimated cell proportions in $\boldsymbol{Z}$ in PBMCs and were forced to leave them out when analyzing the cord blood data, whose cell proportions were not estimated. We then computed a P value to test the hypothesis $\beta_g^{(a)} = 0$ for each of the $p = 784,484$ CpG sites and used the software qvalue (Storey et al. 2015) to compute a q-value for each site. We used the q-values to control the false discovery rate at a nominal level.

To find CpGs whose methylation changed from birth to age seven, we modeled the difference in methylation as

$$\boldsymbol{y}_g^{(7)} - \boldsymbol{y}_g^{(0)} = \mathbf{1}_n \beta_g^{(0\to 7)} + \boldsymbol{Z}\boldsymbol{\gamma}_g + \boldsymbol{C}\boldsymbol{\ell}_g + \boldsymbol{e}_g, \quad \boldsymbol{e}_g \sim N_n\left(0, \sigma_g^2 I_n\right). \tag{4.2}$$

where $\mathbf{1}_n \in \mathbb{R}^n$ is the vector of all 1's and $\beta_g^{(0\to 7)}$ is the average change in methylation between age seven and age 0, holding all else constant. $\boldsymbol{Z}$ included gestational age, PC1 of Figure 4.1 and sample collection site. We estimated $\boldsymbol{C}$ using the method described in Chapter 2.

4.2.2 Joint modeling of methylation at birth and seven

Let $\boldsymbol{X} \in \mathbb{R}^n$ be either self-reported race or inferred genetic ancestry. We used ideas from Flutre et al. (2013) and modeled the methylation levels at each CpG $g = 1, \ldots, p = 784,484$

as

$$\left\{\boldsymbol{y}_g^{(0)}\,\boldsymbol{y}_g^{(7)}\right\}^{\mathrm{T}} \sim N_{2n}\left[\boldsymbol{X}\oplus\boldsymbol{X}\left\{\beta_g^{(0)}\,\beta_g^{(7)}\right\}^{\mathrm{T}} + \boldsymbol{Z}\boldsymbol{\gamma}_g + \boldsymbol{C}\boldsymbol{\ell}_g, \sigma_g^2 I_{2n} + \delta_g^2\boldsymbol{B}\right] \tag{4.3a}$$

$$\begin{aligned}\left(\sigma_g^2+\delta_g^2\right)^{-1/2}\left\{\beta_g^{(0)}\,\beta_g^{(7)}\right\}^{\mathrm{T}} \sim & \pi_{(0,0)}\delta_{(0,0)} + \sum_{k=1}^{K}\pi_{(1,0)}^{(k)}\begin{Bmatrix} N_1\left(0,\tau_k^2\right) \\ \delta_0 \end{Bmatrix} \\ & + \sum_{k=1}^{K}\pi_{(0,1)}^{(k)}\begin{Bmatrix} \delta_0 \\ N_1\left(0,\tau_k^2\right) \end{Bmatrix} \\ & + \sum_{s=1}^{S}\sum_{k=1}^{K}\pi_{(1,1)}^{(k,s)} N_2\left\{0,\tau_k^2\begin{pmatrix} 1 & \rho_s \\ \rho_s & 1 \end{pmatrix}\right\}\end{aligned} \tag{4.3b}$$

where $\boldsymbol{Z}\in\mathbb{R}^{2m\times r}$ are observed nuisance covariates, $\boldsymbol{B}$ is a partition matrix that partitions samples by individuals and $\beta_g^{(0)}$, $\beta_g^{(7)}$ are the effects of $\boldsymbol{X}$ on $\boldsymbol{y}_g^{(0)}$ and $\boldsymbol{y}_g^{(7)}$, the methylation at birth and age seven at CpG g. $\delta_{(0,0)}$ is the point mass at the origin in $\mathbb{R}^2$ and the probability weights are such that

$$\pi_{(0,0)} + \sum_{k=1}^{K}\pi_{(0,1)}^{(k)} + \sum_{k=1}^{K}\pi_{(1,0)}^{(k)} + \sum_{s=1}^{S}\sum_{k=1}^{K}\pi_{(1,1)}^{(k,s)} = 1.$$

We estimated $\boldsymbol{C}\in\mathbb{R}^{2n\times J}$ using CBCV-CorrConf, described in Chapter 3, and subsequently rotated the observed and estimated nuisance covariates $\left[\hat{\boldsymbol{C}}\,\boldsymbol{Z}\right]$ out of (4.3) by defining $\tilde{\boldsymbol{y}}_g = \boldsymbol{Q}_{[\hat{C}\,Z]}^{\mathrm{T}}\boldsymbol{y}_g$, where $\boldsymbol{Q}_{[\hat{C}\,Z]}$ is a matrix whose columns form an orthonormal basis for the null space of $\left[\hat{\boldsymbol{C}}\,\boldsymbol{Z}\right]$. We then estimated σ_g^2 and δ_g^2 with restricted maximum likelihood using the eigenspace rotation paradigm described in Zhou & Stephens (2014), and set $\rho_s\in\{0,1/3,2/3,1\}$. We let $\tau_k^2\in\left\{0.05^2,0.1^2,0.15^2,0.2^2,0.25^2\right\}$ when $\boldsymbol{X}$ was inferred genetic ancestry and $\tau_k^2\in\left\{0.1^2,0.15^2,0.225^2;0.3^2,0.375^2\right\}$ when $\boldsymbol{X}$ was self-reported race. We set τ_4 by first performing a univariate analysis and then estimating the variance of the effect sizes for CpGs with q-values ≤ 0.05, and τ_1 was such that if $\beta_g^{(a)} \sim N_1\left\{0, \left(\sigma_g^2 + \delta_g^2\right)\tau_1^2\right\}$, the expected number of CpGs significant at the Bonferroni threshold $0.05/p$ in a univariate analysis would be smaller than 1 for $a = 0, 7$. Finally, we followed Stephens (2016) and estimated the prior weights with empirical Bayes by maximizing the penalized log-likelihood function

$$\hat{\boldsymbol{\pi}} = \underset{\substack{\boldsymbol{\pi} \geq \mathbf{0} \\ \mathbf{1}_{1+(2+S)K}^{\mathrm{T}}\boldsymbol{\pi}=1}}{\arg\max} \left\{\sum_{g=1}^{p} l\left(\boldsymbol{\pi} \mid \tilde{\boldsymbol{y}}_g \hat{\sigma}_g^2, \hat{\delta}_g^2\right) + \sum_{j=1}^{1+(2+S)K} \left(\lambda_j - 1\right)\log\left(\boldsymbol{\pi}_j\right)\right\}$$

$$\boldsymbol{\pi} = \left\{\pi_{(0,0)}, \pi_{(1,0)}^{(1)}, \ldots, \pi_{(1,0)}^{(K)}, \pi_{(0,1)}^{(1)}, \ldots, \pi_{(0,1)}^{(K)}, \pi_{(1,1)}^{(1,1)}, \ldots, \pi_{(1,1)}^{(S,K)}\right\}^{\mathrm{T}}$$

where $\boldsymbol{\pi}_j$ is the jth element of $\boldsymbol{\pi}$ and $l\left(\boldsymbol{\pi} \mid \tilde{\boldsymbol{y}}_g \hat{\sigma}_g^2, \hat{\delta}_g^2\right)$ is the log-likelihood that simply replaces σ_g^2, δ_g^2 with $\hat{\sigma}_g^2, \hat{\delta}_g^2$. We set $\lambda_1 = 100$ and $\lambda_j = 1$ for all $j > 1$ to encourage conservative inference. The regularization term λ_1 had a relatively little effect on the estimate of $\boldsymbol{\pi}$ because p was so large. We then estimated the posterior distribution as

$$\hat{\mathrm{pr}}\left\{\beta_g^{(0)}, \beta_g^{(7)} \mid \mathrm{Data}\right\} = \mathrm{pr}\left\{\beta_g^{(0)}, \beta_g^{(7)} \mid \tilde{\boldsymbol{y}}_g, \boldsymbol{\pi} = \hat{\boldsymbol{\pi}}, \sigma_g^2 = \hat{\sigma}_g^2, \delta_g^2 = \hat{\delta}_g^2\right\} \quad (g = 1, \ldots, p).$$

4.3 Results

Our study included $n = 196$ children participants in the URECA cohort who had stored cord blood mononuclear cells (CBMCs) and peripheral blood mononuclear cells (PBMCs) collected at birth and age 7, respectively (Gern et al. 2009), and passed quality control checks as described in Methods. The URECA children were classified by parent- or guardian-reported race into one of the following categories: Black, $n_{\mathrm{Black}} = 147$; Hispanic, $n_{\mathrm{Hispanic}} = 39$; White, $n_{\mathrm{White}} = 1$; Mixed race $n_{\mathrm{Mixed}} = 7$, and Other, $n_{\mathrm{Other}} = 2$. A description of the study population is shown in Table 4.1. Genetic ancestry, assessed using principle component analysis (PCA), revealed varying proportions of African and European ancestry along PC1. Because there was little separation along PC2 (Figure 4.1) and no genome-wide significant correlation between PC2 through PC10 and methylation levels at either age, we defined PC1 as inferred genetic ancestry (IGA). The reported races (RR) of the children are also shown in Figure 4.1. We included the self-reported Black and Hispanic children in the analyses of reported race below.

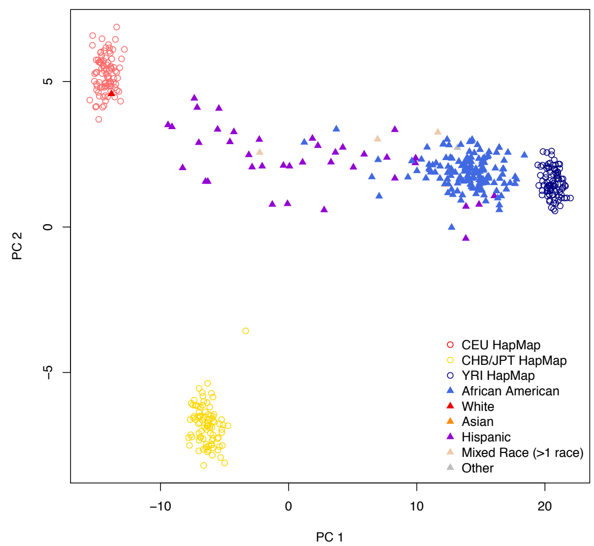


Figure 4.1: Estimated ancestry principal components (PCs) 1 and 2. Nearly all the variation in ancestry separates along PC1 in the URECA sample. Filled triangles represent the 196 URECA children in this study, with their self-reported race shown in different colors. Open circles are reference control samples from HapMap; red = Utah residents from northern and western Europe (CEU), yellow = east Asian (Chinese and Japanese); dark blue = Africans from Nigeria (Yoruban).

Table 4.1: *Covariates for the $n = 196$ URECA children in our study, stratified by self-reported race.*

	Black	**Hispanic**	**White**	**Mixed**	**Other**
Sample Size	147	39	1	7	2
Males (%)	71 (48%)	25 (64%)	0 (0%)	4 (57%)	0 (0%)
Asthma diagnosis at age 7 (%)	38 (26%)	12 (31%)	0 (0%)	2 (29%)	0 (0%)
Gestational age at birth, in weeks (mean [range])	39.0 [34,42]	38.9 [35,41]	36.0	39.1 [37,40]	39.0 [38,40]
Sample Collection Site					
Baltimore (%)	64 (44%)	1 (3%)	1 (100%)	3 (43%)	2 (100%)
Boston (%)	17 (12%)	5 (13%)	0 (0%)	2 (29%)	0 (0%)
New York (%)	23 (16%)	32 (82%)	0 (0%)	1 (14%)	0 (0%)
St. Louis (%)	43 (29%)	1 (3%)	0 (0%)	1 (14%)	0 (0%)

4.3.1 Reported race effects on DNA methylation patterns are conserved in magnitude and direction between birth and age 7

Previous cross-sectional studies have revealed associations between measures of ancestry (genetic ancestry or self-reported race) and DNA methylation at birth (Adkins et al. 2011, Mozhui et al. 2015) and later in life (Galanter et al. 2017, Heyn et al. 2013, Moen et al. 2013, Rahmani et al. 2017, Chan et al. 2017). These correlations were generally attributed to the combined effects of genetic variation and environmental exposures (Galanter et al. 2017, Heyn et al. 2013, Moen et al. 2013, Mozhui et al. 2015). However, because of the cross-sectional nature of those studies, it is not known if the associations between ancestry and methylation patterns present at birth persist (or change) in childhood. Moreover, because ancestry is typically confounded with environmental exposures (Nguyen et al. 2014), it has been proposed that ancestry effects on methylation levels may reflect the effects of exposure histories that also vary by race or ethnicity (Galanter et al. 2017). Alternatively, ancestry effects on DNA methylation patterns could also be due to genetic differences. In such cases, we would expect ancestry-associated methylation patterns to be conserved from birth to later childhood. Using the longitudinal data from the URECA cohort, we tested this hypothesis by addressing three questions. What is the correlation between ancestry and methylation levels at individual CpG sites at birth and age 7? Is the direction and magnitude of the correlation between ancestry and methylation levels conserved between birth and age 7? Are there any CpGs for which the correlation between methylation and ancestry at birth is significantly different from the correlation between methylation and ancestry at age 7?

Standard hypothesis testing can be used to answer the first question but is not appropriate for answering the second or third because failure to reject the null hypothesis that the effects are equal at birth and age 7 does not imply the null hypothesis is true. Additionally, because our studies were conducted in cord blood cells at birth and peripheral blood cells at age 7, ancestry effects at birth and age 7 may differ slightly due to differences in cell composition (Fu et al. 2012). To circumvent these issues, we used Model (4.3) and let the data determine both the strength of the correlation between inferred genetic ancestry (based on genotypes) or reported race (based on questionnaires) and methylation, and how similar the correlations are at birth and age 7. We then answered the first, second and third questions by defining and estimating the conserved (*con*) and discordant (*dis*) sign rates for each CpG $g = 1, \dots, p$:

$$
\begin{aligned}
con_g &= \hat{\mathbb{P}}\left\{\beta_g^{(0)} > 0, \beta_g^{(7)} > 0 \mid \text{Data}\right\} \vee \hat{\mathbb{P}}\left\{\beta_g^{(0)} < 0, \beta_g^{(7)} < 0 \mid \text{Data}\right\} \\
dis_g &= \hat{\mathbb{P}}\left[\left\{\beta_g^{(0)} > 0, \beta_g^{(7)} \leq 0\right\} \cup \left\{\beta_g^{(0)} < 0, \beta_g^{(7)} \geq 0\right\} \cup \left\{\beta_g^{(0)} \geq 0, \beta_g^{(7)} < 0\right\} \right. \\
&\qquad \left. \cup \left\{\beta_g^{(0)} \leq 0, \beta_g^{(7)} > 0\right\} \mid \text{Data}\right].
\end{aligned}
$$

For a given posterior probability threshold, these quantities partition the ancestry-associated CpGs into two groups: those whose ancestry effects were non-zero and conserved from birth to age 7, and those whose ancestry effects were different at birth and age 7. Figures 4.2 and 4.3 provides insight into how the conserved sign rate compares with standard univariate P

values.

Using our estimates for $\boldsymbol{\pi}$ defined in Section 4.2.2, we could then estimate the proportion of CpGs whose self-reported race effects at birth and age 7 were completely unrelated ($\rho_s = 0, \tau_k > 0.1$), moderately similar ($\rho_s = 1/3, \tau_k > 0.1$), very similar ($\rho_s = 2/3, \tau_k > 0.1$) or identical ($\rho_s = 1, \tau_k > 0.1$). Note that if a non-trivial fraction of CpG sites had ancestry effects that were in opposite directions at birth and age 7, they would be assigned to the first bin ($\rho_s = 0, \tau_k > 0.1$). We excluded $\tau_k = 0.1$ because reported race effects with this standard deviation were too small, on average, to differentiate from 0 in this sample size. As seen in Figure 4.4, we estimated that 0.2% of the CpGs with non-zero reported effects at both ages had unrelated or moderately similar reported race effects, whereas 30.7% fell in the very similar bin and 69.1% had identical reported race effects at birth and age 7. These data indicate that when reported race effects on methylation are present (i.e., non-zero) at both birth and age 7, they tend to be very similar or exactly the same at both ages with respect to both direction and magnitude.

We then estimated the conserved and discordant sign rates for all 784,484 probes and classified a CpG as a reported race associated-CpG (RR-CpG) if its conserved or discordant sign rate was above 0.80 (i.e. $con_g \geq 0.8$ or $dis_g \geq 0.8$). At this threshold, we identified 2,162 RR-CpGs, 2,157 (99.8%) of which were conserved in sign ($con_g \geq 0.8$). Black individuals tended to have higher methylation levels at 1,288 of the 2,157 conserved RR-CpGs (60%) ($P = 8.6 \times 10{-}38$). We observed the same trend when we substituted inferred genetic ancestry for reported race, indicating that individuals with more African ancestry tended to have more methylation. This is in accordance with the study of Moen et al. (2013), which used the Illumina 450K array to quantify the differences in methylation between European and African populations. The fact that only 5 of the 2,162 RR-CpGs had discordant reported race effects at birth and age 7 ($dis_g \geq 0.8$) corroborates the observations made in the previous paragraph and answers the second question in the affirmative: if methylation is correlated with reported race at birth, the magnitude and direction of the correlation is almost certainly

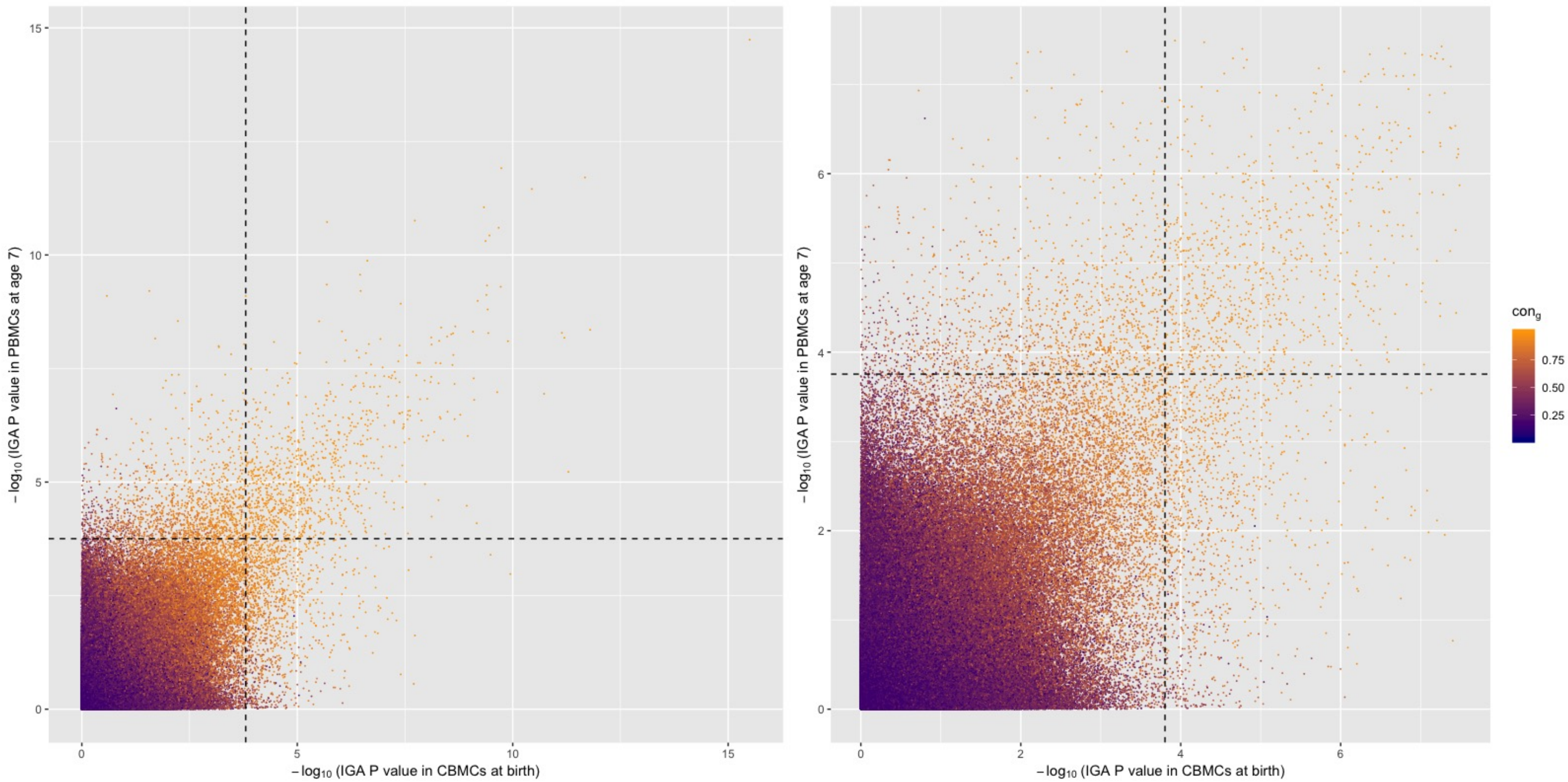


Figure 4.2: Relationship between inferred genetic ancestry P values at birth and age 7 and estimated conserved sign rates (con_g). The P values were estimated using the standard regression model defined in Section 4.2.1 and the dashed lines indicate the 5% FDR threshold, calculated with qvalue (Storey et al. 2015). The plot on the right is a zoomed-in version of the plot on the left.

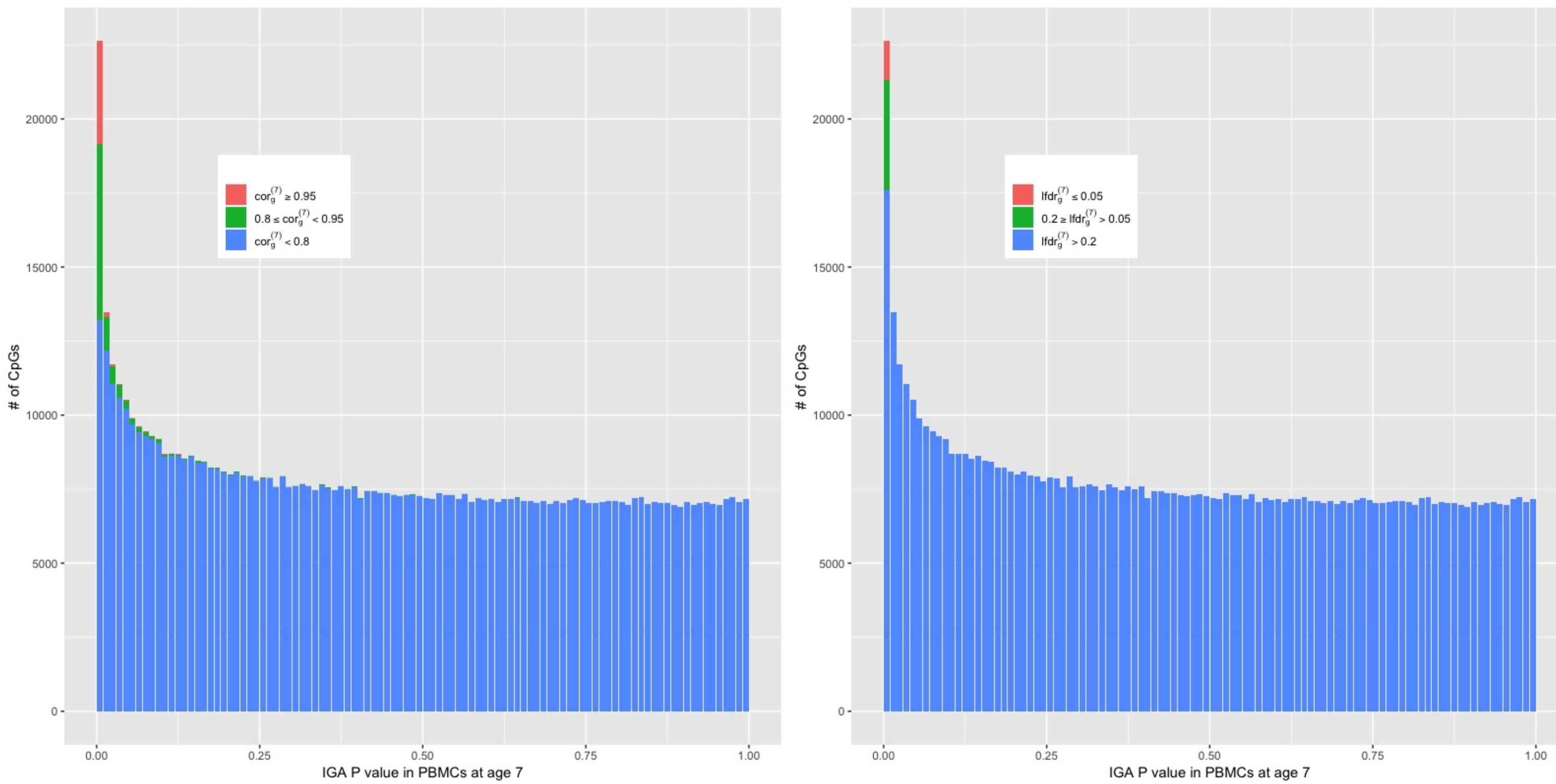


Figure 4.3: Relationship between $cor_g^{(7)} = \max\left[\hat{\mathbb{P}}\left\{\beta_g^{(7)} > 0 \mid \text{Data}\right\}, \hat{\mathbb{P}}\left\{\beta_g^{(7)} < 0 \mid \text{Data}\right\}\right]$ and P values for the effect of inferred genetic ancestry on methylation in PBMCs at age 7. The P values were estimated using the regression model defined in Section 4.2.1 and local false discovery rates, $lfdr_g^{(7)} = \hat{\mathbb{P}}\left\{\beta_g^{(7)} = 0 \mid \text{Data at age 7}\right\}$, were calculated using q-value (Storey et al. 2015).

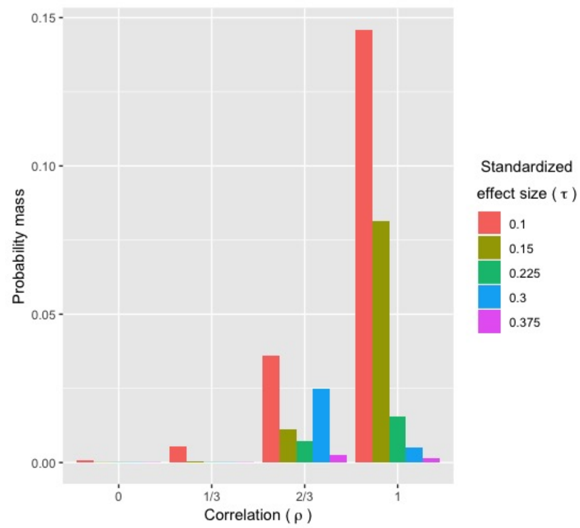


Figure 4.4: Probability mass of $\hat{\pi}^{k,s}_{(1,1)}$, where the correlation coefficient ρ and effect magnitude τ are the same as those defined in (4.3), but without the subscripts s and k. This trend was echoed in the inferred genetic ancestry analysis.

conserved at age 7 (and vice-versa).

4.3.2 Inferred genetic ancestry is more correlated with methylation than is self-reported race

The observed correlations between ancestry and methylation levels may reflect differences in environmental exposures (Galanter et al. 2017, Moen et al. 2013), due to associations between race or ethnicity with socio-cultural, nutritional, and geographic exposures, among others (Nguyen et al. 2014). In fact, Galanter et al. (2017) showed in a cross-sectional study that self-reported ethnicity explained a substantial portion of the variability in whole blood DNA methylation patterns from Latino children of diverse ethnicities, even more so than genetic ancestry. They concluded that ethnicity captures genetic, as well as the socio-cultural and environmental differences that influence methylation levels. If this were the case in the URECA children, the effect of inferred genetic ancestry on methylation levels should be no larger than that of reported race. To assess this possibility, we substituted inferred genetic ancestry for reported race in the analyses described above. This analysis revealed

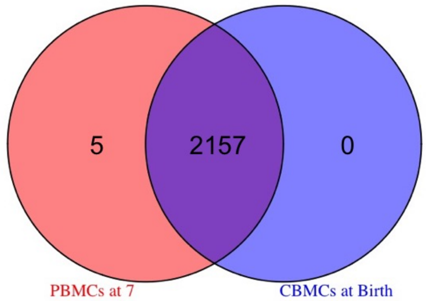
8,597 inferred genetic ancestry-associated CpGs (IGA-CpGs), 8,579 (99.8%) of which were conserved in sign ($con_g \geq 0.80$). This was significantly more than the 2,162 RR-CpGs identified in the reported race analysis above (Figures 4.5a and 4.5b).

To explore this further, we examined the overlap between RR-CpGs and IGA-CpGs (Figure 4.5c). Because reported race is an estimate of inferred genetic ancestry, there is a substantial overlap between IGA-CpGs and RR-CpGs. However, contrary to the results from the Galanter et al. study, the RR-CpGs that we identify are nearly a subset of the IGA-CpGs. This indicates that while IGA-CpGs include most RR-CpGs, reported race does not capture most of the variation in methylation attributable to genetic ancestry in these children.

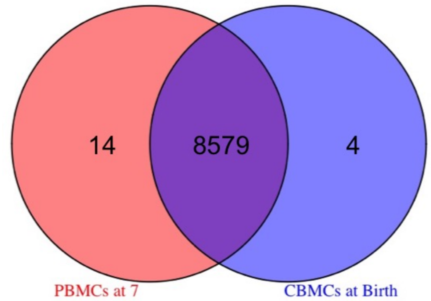
4.3.3 The observed correlations between DNA methylation and reported race are primarily genetic

To further address the question of whether reported race effects on methylation at either birth or age 7 were primarily due to genetic variation or to environmental exposures, we used local genetic variation (within 5kb of a CpG site) and DNA methylation data at birth and age 7 in the 147 self-reported Black children in our study to map methylation quantititave trait loci (meQTLs). Of the 519,622 CpGs within 5kb of a SNP, 65,068 and 70,898 had at least one meQTL in CBMCs at birth and in PBMCs at age 7, respectively, at an FDR of 5%. In addition, 51% of all RR-CpGs with at least one SNP in the 5kb window had at least one meQTL at birth or age 7 at an FDR of 5%, which was a significant enrichment when compared to the 12% observed for non-RR-CpGs (Figures 4.6a and 4.6b).

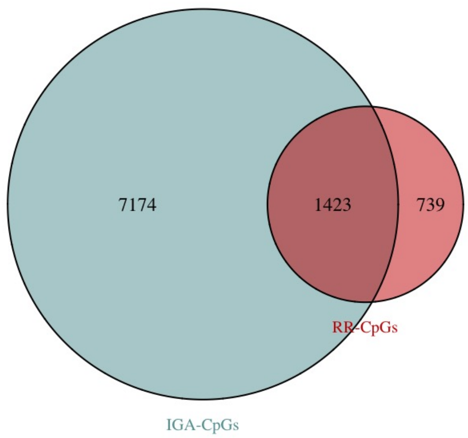
To provide additional evidence that local genotype mediates the effect of reported race on methylation, we used logistic regression to regress the genotype of each of the 269,622 SNPs in our study set onto reported race. The goal was to determine the fraction of RR-CpGs that were mediated through local genotype, i.e. RR-CpGs with both edges a and c in Figure 3a. Since the genotype at most SNPs will be related to self-resported race, a reasonable



(a)



(b)



(c)

Figure 4.5: Overlapping ancestry CpGs at birth and at age 7. (a): RR-CpGs with $con_g \geq 0.8$ (violet) or $dis_g \geq 0.8$ (red or blue). A discordant RR-CpG was classified as significant at birth but not at age 7 (blue) if the marginal posterior probability that the effect was non-zero at birth was greater than that at age 7. Discordant RR-CpGs that were significant at age 7 but not at birth (red) were defined analogously. (b): The same as (a), but for IGA-CpGs. (c): The overlap between RR-CpGs ($con_g \geq 0.8$ or $dis_g \geq 0.8$) and IGA-CpGs ($con_g \geq 0.8$ or $dis_g \geq 0.8$).

upper bound for this quantity is 51%, the fraction of RR-CpGs with at least one meQTL in their 5kb window. To determine a lower bound, we used the results of the abovementioned logisted regression to estimate that over 26% of all RR-CpGs with at least one SNP in their ±5kb windows had both edges a and c (see Section 4.5.3 for calculation details).

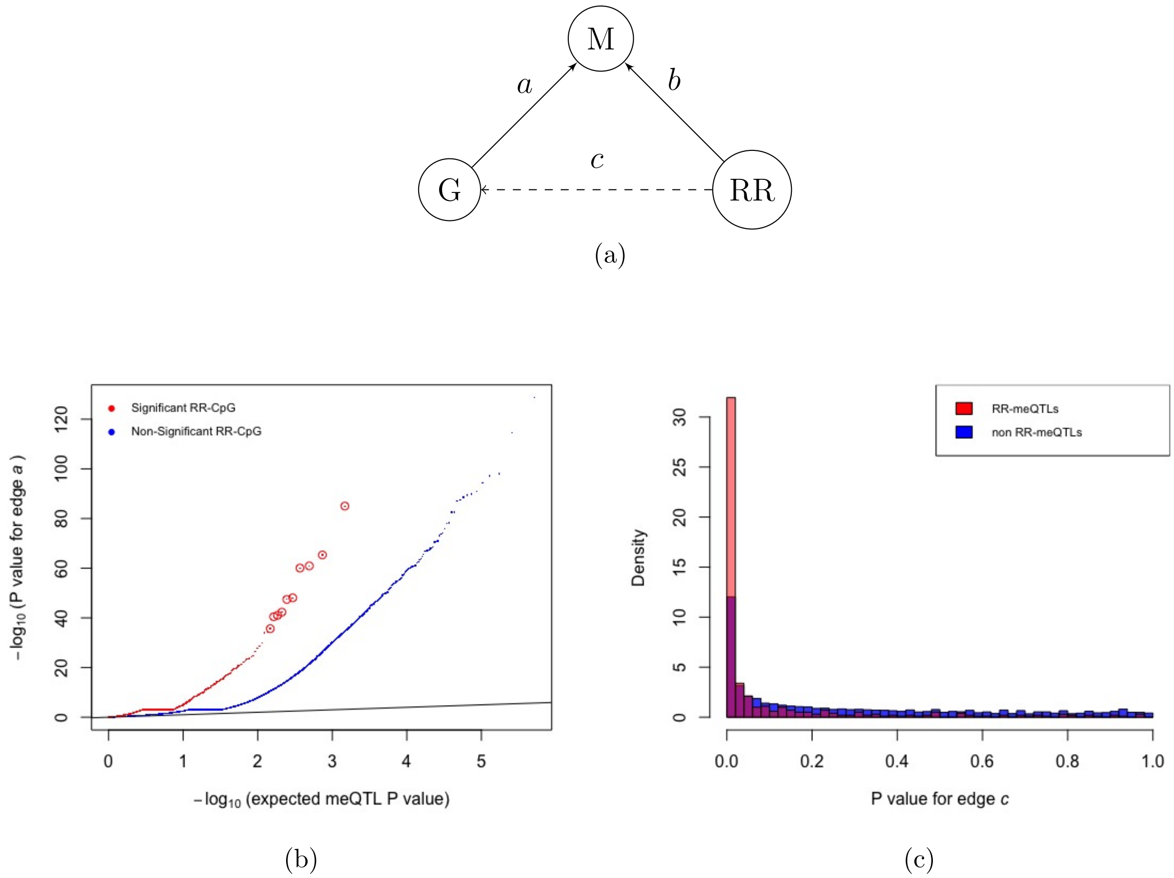


Figure 4.6: RR-CpGs are enriched for CpGs with meQTLs. (a) Illustration of the causal relationship between the methylation (M) at a CpG site, the genotype (G) at the SNP within ±5kb of the CpG that had the smallest meQTL P value and self-reported race (RR). Each graph corresponds to a unique CpG. (b) Plots of the meQTL P value for edge a in CBMCs at birth, where CpGs were stratified by whether or not it was an RR-CpG ($con_g \geq 0.8$ or $dis_g \geq 0.8$). The ten enlarged red circles are just for visual aid. (c) Plots of the logistic regression P value for edge c (Genotype $\sim$ RR) stratified by whether or not the SNP was a RR-meQTL, which was defined as an meQTL whose target CpG was a RR-CpG.

4.3.4 Environmental exposures have nearly undetectable effects on methylation levels in this cohort

We next sought to determine the extent to which environmental exposures affected CMBC and PBMC methylation levels at birth and age 7 in the URECA cohort. None of the direct or indirect measures of exposures that were available in this cohort were associated with methylation levels at either age. These included maternal asthma, maternal infections during pregnancy, pet ownership, bedroom allergens, mothers stress, anxiety and depression metrics, maternal cotinine levels during pregnancy, number of smokers in the household, number of siblings, number of previous live births, daycare attendance, number of colds at age 2 or 3 years, and allergic sensitization or asthma in the child (see Supplement for details). We did, however, identify 16,172 age-related CpGs (i.e., CpGs whose methylation changed from birth to age 7 at a 5% FDR). These 16,172 CpG were strongly enriched for CpGs used to predict gestational age in Knight et al. (2016) and to predict chronological age in Horvath (2013) (see Figure 4.7). Moreover, the estimates of the age effects among these age-related CpGs showed the same direction of change as their corresponding estimated gestational age effects at birth in 97% of the 16,172 age-related CpGs, which included 14,186 gestational age-associated effects that were not significant at a 5% FDR threshold. This concordance in direction of effect is unlikely to occur by chance ($P \leq 10^{-119}$, see Section 4.5.2 for calculation), and indicates that the majority of the changes in mean methylation levels from birth to age 7 is due to aging-related mechanisms rather than age-dependent environmental exposures.

Although most of the reported race-associated methylation patterns can be attributed to genetic variation in the URECA children, we hypothesized that any non-genetic component to these patterns would be greater at age 7 than at birth, due to accumulated exposures over the first 7 years of life. To directly test this hypothesis, we used our Bayesian model to estimate the proportion of CpGs in our study whose methylation was associated with reported race at age 7 but not at birth, as well as the proportion of CpGs that were associated with reported race at birth but not at age 7. Although we estimated the former to be 14% and the latter to be less than 1.5%, we were only able to identify 5 discordant reported race CpGs using a liberal threshold of $dis_g \geq 0.8$. This is because the effects due to ancestry at age 7 among this set of CpGs associated with reported race at age 7 but not at birth were far to small to confidently assign their directions (Figure 4.8).

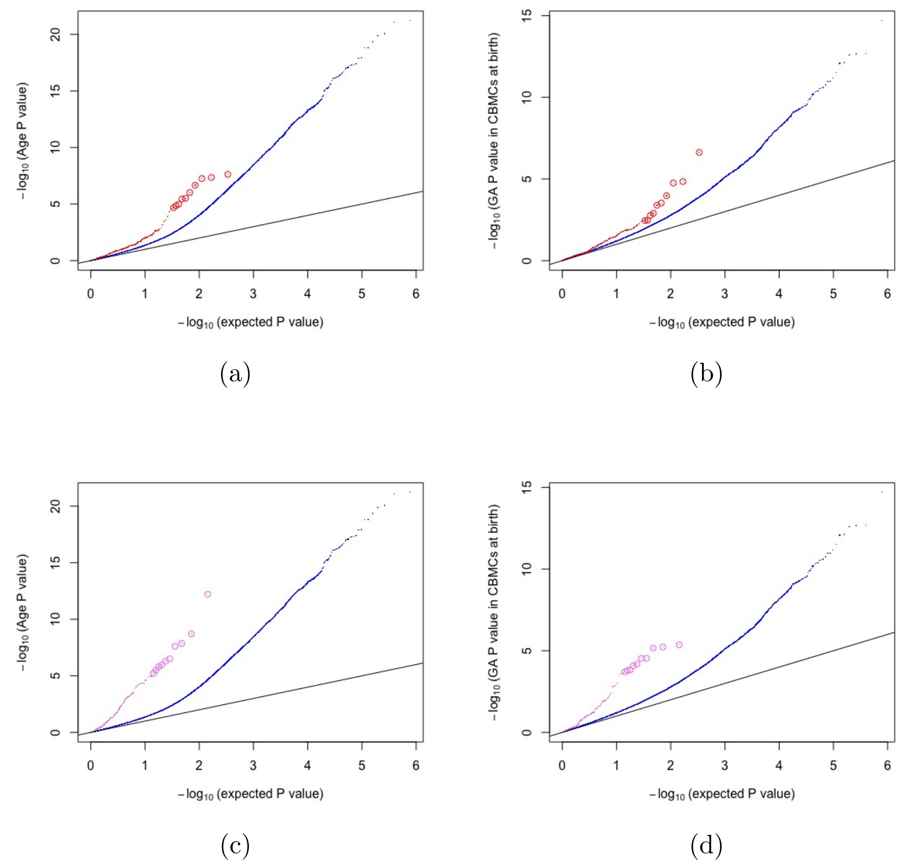


Figure 4.7: Distribution of P values for age (birth to age 7) (panels (a) and (c)) and gestational age (GA) (panels (b) and (d)). Upper panels, stratified by whether or not the CpG was one of the 353 CpGs used to build the linear predictor of age in Horvath (2013) (red dots). The blue dots are all other CpGs; the 10 enlarged red circles are for visual aid. Lower panels, stratified by whether or not the CpG was one of the 148 CpGs used to build the linear predictor of age in Knight et al. (2016). The blue dots are all of the other CpGs; the 10 enlarged violet circles are just for visual aid.

4.4 Discussion

The relationships between DNA methylation, chronological age, and ancestry have the potential to shed light on disease etiology and may help determine the relative genetic and environmental contributions to the observed inter-individual variability of the epigenome (Bocklandt et al. 2011, Horvath 2013, Horvath et al. 2014, Johnson et al. 2012, Knight et al. 2016, Levine & Crimmins 2014, Marioni et al. 2015, Adkins et al. 2011, Galanter et al. 2017, Heyn et al. 2013, Moen et al. 2013, Mozhui et al. 2015, Rahmani et al. 2017). While it

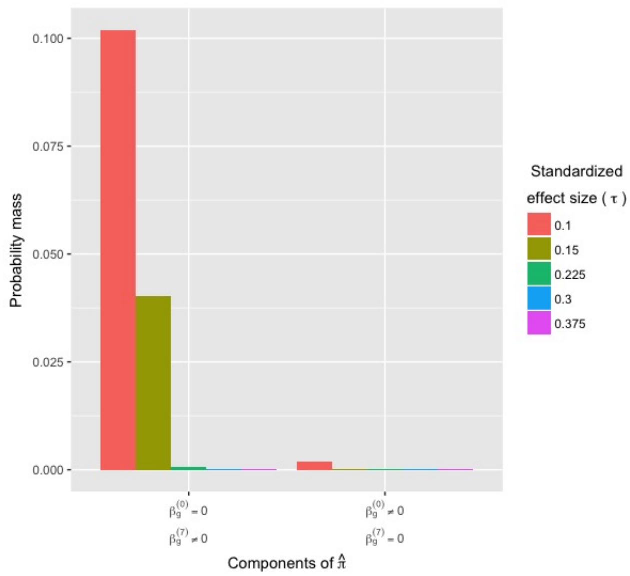


Figure 4.8: Probability mass of components of $\hat{\pi}^{(k)}_{(1,0)}$ ($\beta_g^{(0)} \neq 0, \beta_g^{(7)} = 0$) and $\hat{\pi}^{(k)}_{(0,1)}$ ($\beta_g^{(0)} = 0, \beta_g^{(7)} \neq 0$) in the reported race analyses. The corresponding plot for the inferred genetic ancestry analysis was nearly identical.

has previously been shown that ancestry is related to DNA methylation in cross-sectional studies (Adkins et al. 2011, Galanter et al. 2017, Heyn et al. 2013, Moen et al. 2013, Mozhui et al. 2015, Rahmani et al. 2017) and that statistically significant meQTLs are conserved as individuals age (Gaunt et al. 2016), it has yet to be shown whether or not race/ethnicity-dependent methylation marks are conserved as children age.

Even though there was substantial change in blood methylation levels over time among children in this cohort, inferred genetic ancestry and self-reported race effects on methylation were overwhelmingly conserved in both direction and magnitude from birth to age 7. This result, as well as our novel Bayesian inference paradigm used to obtain it, is important in and of itself because it provides an example of, and a general method for identifying, DNA methylation patterns that are conserved over time, and differentiating between environmentally responsive and temporally stable DNA methylation marks has been highlighted as both a gap in current knowledge and a critical area of future epigenetic research (Martin & Fry 2018). The consistency of our estimates for inferred genetic ancestry and reported race effects on methylation levels also demonstrates the fidelity of our processing pipeline that accounts for unobserved factors, including cell composition, because failure to account for latent covariates can lead to biased and irreproducible estimates (Peixoto et al. 2015, Yao et al. 2012).

While the observation that reported race effects are conserved from birth to age 7 gives credence to the hypothesis that the effects are genetic in nature, it does not rule out the possibility of environmental components or gene-environment interactions that could result in race/ethnicity-associated methylation effects prior to birth and persist as the child ages. To further explore this, we showed that the reported race associated CpGs (RR-CpGs) were enriched among CpGs with meQTLs, indicating that methylation levels at many of the RR-CpGs are mediated by local genotype and that much of the reported race-methylation correlation could be attributed to genetic variation. Moreover, the RR-CpGs were only a small subset of inferred genetic ancestry associated CpGs (IGA-CpGs) in our study. This is opposite to the findings of Galanter et al. (2017), who argued that ethnicity-dependent methylation patterns in admixed populations capture both genetic variation and differences in accumulated exposures. Our results provide evidence for genetics accounting for an overwhelming majority of the correlation between methylation and reported race, which suggests the non-genetic contribution to variability in blood methylation may be smaller than previously thought.

Our observations in support of strong genetically and weak environmentally determined reported race-associated methylation patterns in blood may seem paradoxical to the plethora of studies showing that DNA methylation levels in blood cells are associated with environmental exposures, such as cadmium, arsenic and smoking, to name a few (Joubert et al. 2016, Kippler et al. 2013, C. et al. 2013, Rzehak et al. 2016, Galanter et al. 2017, Joubert et al. 2012, Lee et al. 2015, Markunas et al. 2014, Richmond et al. 2015, Wu et al. 2018). Whereas the estimated genetic effect sizes in our study are substantially larger than many of the environmentally-associated effects on methylation patterns previously reported, the effects of many environmental exposures on methylation in blood are probably too small to estimate with even moderate to large sample sizes (Breton et al. 2017). For example, it was only by performing a meta-analysis in 6,685 individuals that Joubert et al. (2016) were able to identify 6,000 CpGs whose DNA methylation levels in blood from infants and adolescents were associated maternal smoking exposure. In one sense, we were able to corroborate previous observations of small non-genetic effects on methylation in blood by showing that while methylation patterns at an estimated 14% of all CpGs in our study were not correlated with reported race at birth but correlated with reported race at age 7, the correlation at individual CpGs at age 7 was too small to be identified as statistically significant. We were also not able to detect any statistically significant correlations between methylation at birth or at age 7 and any of the exposure variables measured in this cohort. In particular, we note that cord blood cotinine levels, a measure of *in utero* tobacco smoke exposure, were above the level of detection in only 34 of the 196 mothers in our study.

An unsurprising feature of these longitudinal data is that average methylation levels of 16,172 CpGs changed significantly from birth to age 7. However, what was quite remarkable was that the direction of the change in 97% of those CpGs matched the direction of the corresponding correlation between methylation levels and gestational age at birth. Not only does this further suggest that methylation levels of the vast majority of the 16,172 age-related CpGs were in fact changing due to age-related mechanisms and not due to differences in environmental exposures at birth and age 7, but also that the "epigenetic clock" present at birth may be the same as that present later in life. While we do not have the data to explore this further, this remains an important avenue of future research.

The results of our study suggest that DNA methylation levels in blood cells are fairly robust to environmental exposures, including those that are correlated with self reported race. A better understanding of tissue-specific methylation responses to environmental exposures could inform the design of future studies and provide insights into the mechanisms through which exposures and gene-environment interactions influence health and disease.

4.5 Additional statistical methods

4.5.1 meQTL analysis

We used self-reported African American individuals in the following linear regression model to detect meQTLs:

$$\boldsymbol{y}_g = \boldsymbol{x}_s \beta_{s,g} + \boldsymbol{Z}\gamma_{s,g} + \boldsymbol{C}_{\text{meQTL}} \ell_{s,g} + \boldsymbol{e}_{s,g}, \quad \boldsymbol{e}_{s,g} \sim N_n\left(0, \sigma_{s,g}^2 I_n\right) \tag{4.5}$$

where $\boldsymbol{y}_g$ contained the methylation M-value measurements at CpG g and $\boldsymbol{x}_s$ the genotype at SNP s. $\boldsymbol{Z}$ contained the first two components of inferred genetic ancestry, gestational age at birth, gender, sample collection site and methylation plate number. We approximated $\boldsymbol{C}_{\text{meQTL}}$ by first regressing out $\boldsymbol{Z}$ from the methylation data matrix to get the residual data matrix, $\boldsymbol{R} \in \mathbb{R}^{784,484 \times n}$, used CATE v1.0.4 (Owen & Wang 2016) to estimate there were nine additional components explaining a substantial portion of the variability in $\boldsymbol{R}$, and let $\hat{\boldsymbol{C}}_{\text{meQTL}}$ be the first nine right singular values of $\boldsymbol{R}$. We then selected pairs of CpGs and SNPs by only considering SNPs within a 5kB window of a CpG and used the test statistic

$$t_g = \min_{\substack{\text{SNP } s \pm 5\text{kB} \\ \text{from CpG } g}} \left\{P \text{ value testing if } \beta_{s,g} = 0 \text{ in (4.5)}\right\} \tag{4.6}$$

to test the hypothesis that there were no meQTLs in the 5kB window of CpG g and defined the representative meQTL for CpG g as the SNP that gives the smallest P value. When there was only a single SNP in the window, the P value for the test was t_g. Otherwise, the corresponding P value was computed by randomly permuting the entries of $\tilde{\boldsymbol{y}}_g$, the residuals of $\boldsymbol{y}_g$ after regressing out $\boldsymbol{Z}$ and $\hat{\boldsymbol{C}}_{\text{meQTL}}$, and estimating $t_g^{(b)}$ (the bootstrapped null version of (4.6)) using $\tilde{\boldsymbol{x}}_s$ (the residuals of $\boldsymbol{x}_s$ after regressing out $\boldsymbol{Z}$ and $\hat{\boldsymbol{C}}_{\text{meQTL}}$) for $b = 1, \ldots, B$. The P value was then $\frac{\#\left\{t_g^{(b)} \leq t_g\right\}+1}{B+1}$. B was set to 1,000 in both the cord blood and PBMC methylation data sets.

4.5.2 Upper bound on P value from Section 4.3.4

We estimated β_g^{GA}, the effect of gestational age on methylation at birth, using Model (4.1) and $\beta_g^{(0\to 7)}$, the effect of chronological age, using Model (4.2). Define the estimated gestational and chronological age effects to be $\hat{\beta}_g^{\mathrm{GA}}$ and $\hat{\beta}_g^{(0\to 7)}$, respectively. We approximate their joint distribution to get an approximate upper bound for the expected number of pairs $\left(\beta_g^{\mathrm{GA}}, \beta_g^{(0\to 7)}\right)$ out of all 16,172 age-related CpG sites that had the same sign, under the null hypothesis that β_g^{GA} and $\beta_g^{(0\to 7)}$ were generated independently for each $g \in [p]$.

Assume the variance model for the data at birth and age seven is given by (4.3a) and let $r_g = \frac{\delta_g^2}{\delta_g^2+\sigma_g^2}$. Then using the estimates for $\boldsymbol{C}$ and the observed nuisance covariates $\boldsymbol{Z}$ in Models (4.1) and (4.2), we estimated the correlation between $\hat{\beta}_g^{\mathrm{GA}}$ and $\hat{\beta}_g^{(0\to 7)}$ to be $0.66\left(1-r_g\right)$. We then have that conditional on the true effects $\left(\beta_g^{\mathrm{GA}},\, \beta_g^{(0\to 7)}\right)^{\mathrm{T}}$,

$$\left(\hat{\beta}_g^{\mathrm{GA}},\, \hat{\beta}_g^{(0\to 7)}\right)^{\mathrm{T}} \approx N_2\left\{\left(\beta_g^{\mathrm{GA}},\, \beta_g^{(0\to 7)}\right)^{\mathrm{T}}, \operatorname{diag}\left(c_g, d_g\right)\begin{pmatrix} 1 & 0.66\left(1-\hat{r}_g\right) \\ 0.66\left(1-\hat{r}_g\right) & 1 \end{pmatrix} \times \right.$$
$$\left. \times \operatorname{diag}\left(c_g, d_g\right)\right\} \tag{4.7}$$

for each $g \in [p]$, where c_g, d_g are positive constants. Let A_g be the event that CpG g is an age-CpG at a 5% FDR. We assume that $A_g = \left\{|z|_g \geq t\right\}$, where $z_g = \hat{\beta}_g^{0\to 7}/d_g$ is the z-score corresponding to $\hat{\beta}_g^{0\to 7}$ and t can be estimated as the smallest z-score with a q-value less than 0.05. The empirical distributions of $\left\{\hat{\beta}_g^{(0\to 7)}\right\}_{g\in\{5\% \text{ FDR age CpGs}\}}$ and $\left\{\hat{\beta}_g^{GA}\right\}_{g\in\{5\% \text{ FDR gestational age CpGs}\}}$ were approximately symmetric around 0, which we took to imply $\left\{\beta_g^{(0\to 7)}\right\}_{g\in[p]}$ and $\left\{\beta_g^{GA}\right\}_{g\in[p]}$ were symmetric around 0. For simplicity, we assume for density functions

$$h_{GA}(\cdot) = \sum_{r=1}^{R} \pi_r^{(GA)} N_1\left(\cdot; 0, \phi_r^{(GA)}\right) \qquad h_{(0\to 7)}(\cdot) = \sum_{j=1}^{J} \pi_j^{(0\to 7)} N_1\left(\cdot; 0, \phi_j^{(0\to 7)}\right),$$

$\beta_g^{\mathrm{GA}} \overset{i.i.d}{\sim} h_{GA}(\cdot)$ and $\beta_g^{(0\to 7)} \overset{i.i.d}{\sim} h_{(0\to 7)}(\cdot)$. Such mixture normal densities can approximate a large class of parametric and non-parametric distributions (Norets 2010). Define $X_g, Y_g \in \mathbb{R}$ to be such that

$$\left(X_g, Y_g\right)^{\mathrm{T}} \sim N_2\left\{0, \begin{pmatrix} 1 & 0.66\left(1-\hat{r}_g\right) \\ 0.66\left(1-\hat{r}_g\right) & 1 \end{pmatrix}\right\}.$$

Then under the null hypothesis that $\beta_g^{(0\to 7)}$ and β_g^{GA} are independent and assuming (4.7) is correct,

$$P\left\{\hat{\beta}_g^{\mathrm{GA}}\hat{\beta}_g^{(0\to 7)} > 0 \mid A_g\right\} \leq \frac{\mathbb{P}\left(\left\{X_g Y_g > 0\right\} \cap \left\{\left|Y_g\right| \geq t\right\}\right)}{\mathbb{P}\left(\left|Y_g\right| \geq t\right)}.$$

We can easily estimate the above upper bound. Therefore, conditional on knowing whether or not each CpG is an age-associated CpG,

$$\begin{aligned}\mu &= \mathbb{E}\left\{\sum_{g\in\{5\%\text{ FDR age CpGs}\}} 1\left(\hat{\beta}_g^{\mathrm{GA}}\hat{\beta}_g^{(0\to 7)} > 0\right)\right\}\\ &= \sum_{g\in\{5\%\text{ FDR age CpGs}\}} P\left\{\hat{\beta}_g^{\mathrm{GA}}\hat{\beta}_g^{(0\to 7)} > 0 \mid A_g\right\} \leq 14,236\end{aligned}$$

under the null hypothesis. Since the maximum variance for a Bernoulli random variable is 1/4, an approximate lower bound for the test-statistic is

$$\frac{0.97 \times 16,172 - 14,236}{\sqrt{16,172/4}} = 23.3,$$

which has a corresponding P value $\leq 10^{-119}$ under the normal approximation.

4.5.3 A lower bound for the fraction of RR-CpGs mediated by local genotype

Here we describe how we conservatively estimated the fraction of RR-CpGs with a SNP in a 10kB window that were likely mediated by neighboring meQTLs. Fix some network composed of a CpG with methylation M and the SNP whose genotype G was most correlated with M according to the above meQTL regression procedure. Let $\{\text{RR} \to \text{M}\}$ be the event the CpG is an RR-CpG, $\{\text{RR} \to \text{G}\}$ the event RR affects genotype, and $\{\text{G} \to \text{M}\}$ the event G affects M independently of RR (see previous section). We would like to estimate

$$\begin{aligned}\mathbb{P}\left(\text{RR} \to \text{G}, \text{G} \to \text{M} \mid \text{RR} \to \text{M}\right) &= \frac{\mathbb{P}\left(\text{RR} \to \text{G}, \text{G} \to \text{M}, \text{RR} \to \text{M}\right)}{\mathbb{P}\left(\text{RR} \to \text{M}\right)}\\ &= \frac{\mathbb{P}\left(\text{RR} \to \text{M} \mid \text{RR} \to \text{G}, \text{G} \to \text{M}\right)\mathbb{P}\left(\text{RR} \to \text{G}, \text{G} \to \text{M}\right)}{\mathbb{P}\left(\text{RR} \to \text{M}\right)}\end{aligned} \tag{4.8}$$

Define $H_0 = \{\text{RR} \not\to \text{G}\}$. For each SNP we computed a P value for the null hypothesis H_0 using the logistic regression model G $\sim$ RR, where RR was either Black or Hispanic (we assumed a Hardy-Weinberg equilibrium model for the genotypes of all SNPs considered). Let t be the test statistic from the regression and $t_\alpha^* > 0$ be some threshold with significance level α. Then because G and RR are independent under H_0 (regardless of whether or not $\{\text{G} \to \text{M}\}$ or $\{\text{RR} \to \text{M}\}$ hold),

$$
\begin{aligned}
q &= \mathbb{P}\left(H_0 \mid |t| \geq t_\alpha^*, \mathrm{G} \to \mathrm{M}, \mathrm{RR} \to \mathrm{M}\right) \\
&= \frac{\mathbb{P}\left(|t| \geq t_\alpha^* \mid H_0, \mathrm{G} \to \mathrm{M}, \mathrm{RR} \to \mathrm{M}\right) \mathbb{P}\left(H_0 \mid \mathrm{G} \to \mathrm{M}, \mathrm{RR} \to \mathrm{M}\right)}{\mathbb{P}\left(|t| \geq t_\alpha^*, \mathrm{G} \to \mathrm{M}, \mathrm{RR} \to \mathrm{M}\right)} \\
&= \frac{\mathbb{P}\left(|t| \geq t_\alpha^* \mid H_0\right) \mathbb{P}\left(H_0 \mid \mathrm{G} \to \mathrm{M}, \mathrm{RR} \to \mathrm{M}\right)}{\mathbb{P}\left(|t| \geq t_\alpha^*, \mathrm{G} \to \mathrm{M}, \mathrm{RR} \to \mathrm{M}\right)} \\
&\leq \frac{\alpha}{\mathbb{P}\left(|t| \geq t_\alpha^*, \mathrm{G} \to \mathrm{M}, \mathrm{RR} \to \mathrm{M}\right)}
\end{aligned}
$$

where the first equality in the second line comes from the fact that under the null hypothesis and given the rest of the graph, the behavior of G and RR are independent. Therefore, we could upper bound q with qvalue in R using the RR-meQTL p-values from the logistic regression and restricting $\pi_0 = 1$ (this is just the Benjamini-Hochberg procedure (Benjamini & Hochberg 1995) interpreted in a Bayesian framework). We finally established an estimated lower bound for (4.8) by using the following:

$$
\begin{aligned}
&\mathbb{P}\left(\mathrm{RR} \to \mathrm{G}, \mathrm{G} \to \mathrm{M} \mid \mathrm{RR} \to \mathrm{M}\right) = \frac{\#\text{ networks with RR} \to \text{G}, \text{G} \to \text{M}, \text{RR} \to \text{M}}{\#\text{ networks with RR} \to \text{M}} \\
\geq& \frac{\#\text{ networks with RR} \to \text{G}, \text{G} \to \text{M}, \text{RR} \to \text{M and } q \leq 0.2}{\#\text{ of networks with RR} \to \text{M}} \\
\gtrsim& (1-0.2)\frac{\#\text{ of networks with } q \leq 0.2 \text{ among RR-meQTLs}}{\#\text{ RR-CpGs}} = 0.26.
\end{aligned}
$$

CHAPTER 5

ESTIMATION AND INFERENCE IN METABOLOMICS WITH NON-RANDOM MISSING DATA AND LATENT COVARIATES

5.1 Introduction

Metabolomics is the study of tissue- or body fluid-specific small molecule metabolites, and has the potential to lead to new insights into the origin of human disease (Young & Wallace 2009, Sampson et al. 2013, Finkelstein et al. 2015, Reinke et al. 2017) and drug metabolism (Chen et al. 2007, Dubuis et al. 2018). Metabolomics is particularly attractive because the metabolome provides information at the molecular scale, and often responds quicker than the transcriptome, DNA methylome and proteome to changes in the external environment (Herman et al. 2017).

Recent advances in both liquid chromatography (LC) and untargeted mass spectrometry (MS) have made it possible to identify and quantify hundreds to thousands of metabolites per sample (Liu et al. 2014). Similar to high throughput gene expression, proteomic and DNA methylation data, these data contain systematic technical and biological variation whose sources are unobserved by the practitioner (Salerno et al. 2017). However, what makes untargeted LC-MS metabolomic data particularly challenging is the vast amount of missing data, nearly all of which is missing not at random due to an unknown, metabolite-specific missingness mechanism in which more abundant and ionizable analytes are more likely to be observed (Do et al. 2018). For instance, 22% of all observations were missing from our data example in Section 5.8, in which we analyzed $p = 1138$ metabolites quantified by untargeted LC-MS in $n = 661$ samples.

There have been several methods to attempt to account for either the latent systematic variation (De Livera et al. 2012, 2015, Salerno et al. 2017) or the non-random missing data (Chen et al. 2017, Hedeker et al. 2018, O'Brien et al. 2018) when trying to infer the relationship between the metabolome and a covariate of interest. Surprisingly, to the best of our knowledge, Wehrens et al. (2016) is the only work to even acknowledge the challenge of accounting for both. However, they propose simply imputing missing values with an arbitrary, user-specified limit of detection and require the practitioner to have prior knowledge of a set of control metabolites that are unrelated to the covariate of interest.

Given the paucity of methods to analyze metabolomic data, we develop MetabMiss, a statistically rigorous and computationally efficient method to account for both latent covariates and non-random missing data when analyzing metabolomic data. A key component in our procedure is estimating each metabolite-dependent missingness mechanism, which we do using instrumental variable generalized method of moments (IV-GMM) and leveraging the fact that the majority of the variation in metabolimic data can be explained by a small number of factors. Besides being the first method to account for both latent covariates and non-random missing data that avoids erroneously imputing missing data, our method offers the following advantages:

(a) We assume the underlying probability distribution of the missing data is unknown.

(b) We do not require the practitioner to have access to internal standards or prior knowledge of control metabolites to correct for latent covariates.

(c) We modularize our method such that one only needs to estimate the metabolite-dependent missingness mechanisms once per dataset, which makes computation on the order of a metabolome genome wide association study tractable.

And while we do assume the functional form of the missing data mechanism is known, we provide a method to access the veracity of said form for each metabolite.

The remainder of the chapter is organized as follows: we give a mathematical description of the data in Section 5.2 and review IV-GMM in Section 5.3. We then present our method to estimate each metabolite-dependent missingness mechanism in Section 5.4, describe how we estimate and perform inference on the coefficients of interest in a linear model when the latent covariates are observed in Section 5.5, and finally describe our method to estimate the aforementioned latent covariates in Section 5.6. We conclude by illustrating how our method performs in simulated data in Section 5.7 and use it to analyze real metabolomic data in Section 5.8.

5.2 Problem set-up

5.2.1 Notation

In addition to the notation defined in Section 1.1, we let $F_\nu(x)$ be the cumulative distribution function for the t-distribution with $\nu > 0$ degrees of freedom in this chapter.

5.2.2 A description of and model for the data

Define $y_{gi} = [\boldsymbol{Y}]_{gi}$ to be the log-transformed metabolite integrated-intensity for metabolite $g \in [p]$ in sample $i \in [n]$, which may be observed or unobserved. Let $\boldsymbol{X} = (\boldsymbol{x}_1 \cdots \boldsymbol{x}_n)^{\mathrm{T}} \in \mathbb{R}^{n\times d}$ and $\boldsymbol{C} = (\boldsymbol{c}_1 \cdots \boldsymbol{c}_n)^{\mathrm{T}} \in \mathbb{R}^{n\times K}$ be observed and unobserved covariates. For coefficients $\boldsymbol{\beta}_{*_g} \in \mathbb{R}^d$ and $\boldsymbol{\ell}_{*_g} \in \mathbb{R}^K$, we assume

$$[\boldsymbol{Y}]_{gi} = y_{gi} = \mu_{gi} + e_{gi} \quad (i = 1, \dots, n; g = 1, \dots, p) \tag{5.1a}$$

$$\mu_{gi} = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_{*_g} + \boldsymbol{c}_i^{\mathrm{T}} \boldsymbol{\ell}_{*_g} \quad (i = 1, \dots, n; g = 1, \dots, p) \tag{5.1b}$$

$$e_{gi} \dot{\sim} \left(0, \sigma_{*_g}^2\right) \quad (i = 1, \dots, n; g = 1, \dots, p), \tag{5.1c}$$

where the residuals $\{e_{gi}\}_{i\in[n];g\in[p]}$ are independent and $e_{g1}, \dots, e_{gn}$ are identically distributed for each $g \in [p]$. We use the subscript "$*$" throughout this chapter to indicate true parameters. We do not assume an explicit probability distribution for the residuals in order to avoid assuming a distribution for the missing data. The observed covariates $\boldsymbol{X}$ can contain biological factors like disease status, as well as technical factors like observed batch variables or normalizing factors. We assume throughout this work that $\boldsymbol{1}_n \in \operatorname{Im}(\boldsymbol{X})$ and that $\boldsymbol{X}$ is non-random. The unobserved covariates $\boldsymbol{c}_i$ can confound the relationship between $\boldsymbol{x}_i$ and y_{gi}, and also induce dependencies between different metabolites. We will assume throughout the chapter that $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_n$ are independent and are independent of $\{e_{gi}\}_{i\in[n];g\in[p]}$, which implies $y_{g1}, \ldots, y_{gn}$ are independent for all $g \in [p]$. We lastly define $\boldsymbol{y}_g = \left(y_{g1} \cdots y_{gn}\right)^{\mathrm{T}} \in \mathbb{R}^n$ and $\boldsymbol{e}_g = \left(e_{g1} \cdots e_{gn}\right)^{\mathrm{T}} \in \mathbb{R}^n$, which will be used throughout the chapter.

We next define the indicator variable

$$r_{gi} = I\left(y_{gi} \text{ is observed}\right), \tag{5.2}$$

and assume that for some known, increasing function $\Psi : \mathbb{R} \to (0, 1)$ such that $\lim_{x\to-\infty} \Psi(x) = 0$ and $\lim_{x\to\infty} \Psi(x) = 1$,

$$\pi_{gi} = \mathbb{P}\left(r_{gi} = 1 \mid y_{gi}\right) = \Psi\left\{\alpha_{*_g}\left(y_{gi} - \delta_{*_g}\right)\right\} \quad (i = 1, \ldots, n; g = 1, \ldots, p). \tag{5.3}$$

The unknown scale and location parameters $\alpha_{*_g} > 0$ and $\delta_{*_g} \in \mathbb{R}$ make the missingness mechanism metabolite dependent, where $\alpha_{*_g} \searrow 0$ implies the mechanism is missing completely at random and $\alpha_{*_g} \nearrow \infty$ implies metabolite g in sample i is observed if and only if $y_{gi} > \delta_{*_g}$.

Since Ψ is an increasing function, metabolites with smaller intensities are less likely to be observed, which is consistent with previous observations in untargeted mass spectrometry experiments (Karpievitch et al. 2010, Chen et al. 2017, Do et al. 2018, O'Brien et al. 2018, Hedeker et al. 2018). Model (5.3) is also a classic model for missing data in untargeted mass spectrometry data (Chen et al. 2017, O'Brien et al. 2018, Hedeker et al. 2018), where typical values for Ψ include the logistic function, an exponential probabilistic model (Chen et al. 2017, Hedeker et al. 2018) and the cumulative distribution function for the normal distribution (O'Brien et al. 2018). However, we find that letting $\Psi(x) = F_4(x)$ is a more robust option, since its heavy tails make it less sensitive to outliers. This approach has been previously used as a robust alternative to logistic and probit functions (Liu 2005, Kang &

Schafer 2007).

Implicit in this characterization of the missingness mechanism is the assumption that conditional on the underlying intensity data $\{y_{gi}\}_{i\in[n];g\in[p]}$, $\{r_{gi}\}_{i\in[n];g\in[p]}$ are independent. This is likely only approximately true, since other intense analytes can preclude MS/MS fragmentation in data dependent mass spectrometry experiments. However, properly tuning the dynamic exclusion time can mitigate this substantially (Johnson et al. 2013).

5.3 Estimating α_{*_g} and δ_{*_g} when $\boldsymbol{C}$ is observed

5.3.1 A review of instrumental variable generalized method of moments

We use an instrumental variable generalized method of moments (IV-GMM) estimator to estimate the metabolite-dependent missingness mechanism, where IV-GMM is a standard technique in the economics literature (Amemiya 1974, 1977, Gallant 1977, Newey 1990). To review, suppose we observe data $\boldsymbol{w}_i \in \mathbb{R}^r$, $i \in [n]$, and are interested in estimating the parameter $\boldsymbol{\theta}_* \in \mathbb{R}^d$. Assume there exists a sub-vector $\boldsymbol{v}_i$ of $\boldsymbol{w}_i$ such that $\mathbb{E}\left\{f\left(\boldsymbol{\theta}_*, \boldsymbol{w}_i\right) \mid \boldsymbol{v}_i\right\} = \boldsymbol{0}$ for some function $f : \mathbb{R}^d \times \mathbb{R}^r \to \mathbb{R}$. Since the conditional expectation typically does not have an analytic form, we construct instrumental variables $\boldsymbol{A}\left(\boldsymbol{v}_i\right) \in \mathbb{R}^q$, where the law of total expectation implies $\mathbb{E}\left\{\boldsymbol{A}\left(\boldsymbol{v}_i\right) f\left(\boldsymbol{\theta}_*, \boldsymbol{w}_i\right)\right\} = \boldsymbol{0}$. The popular two-step IV-GMM estimator for $\boldsymbol{\theta}_*$, $\hat{\boldsymbol{\theta}}$, is defined as

$$\begin{aligned}
\hat{\boldsymbol{\theta}}^{(1)} &= \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\Theta\subseteq\mathbb{R}^d} \left\{\sum_{i=1}^n \boldsymbol{A}\left(\boldsymbol{v}_i\right) f\left(\boldsymbol{\theta}, \boldsymbol{w}_i\right)\right\}^{\mathrm{T}} \left\{\sum_{i=1}^n \boldsymbol{A}\left(\boldsymbol{v}_i\right) f\left(\boldsymbol{\theta}, \boldsymbol{w}_i\right)\right\} \\
\hat{\boldsymbol{\theta}} &= \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\Theta\subseteq\mathbb{R}^d} \left(\left\{\sum_{i=1}^n \boldsymbol{A}\left(\boldsymbol{v}_i\right) f\left(\boldsymbol{\theta}, \boldsymbol{w}_i\right)\right\}^{\mathrm{T}} \left[n^{-1}\sum_{i=1}^n f\left\{\hat{\boldsymbol{\theta}}^{(1)}, \boldsymbol{w}_i\right\}^2 \boldsymbol{A}\left(\boldsymbol{v}_i\right) \boldsymbol{A}\left(\boldsymbol{v}_i\right)^{\mathrm{T}}\right]^{-1} \right. \\
&\qquad \left. \times \left\{\sum_{i=1}^n \boldsymbol{A}\left(\boldsymbol{v}_i\right) f\left(\boldsymbol{\theta}, \boldsymbol{w}_i\right)\right\}\right). \qquad (5.4)
\end{aligned}$$

This IV-GMM estimator is well understood and is consistent for $\boldsymbol{\theta}_*$, asymptotically normal and achieves the smallest possible asymptotic variance among all IV-GMM estimators with instruments $\boldsymbol{A}(\boldsymbol{v}_i)$ under certain regularity conditions when $q \geq d$ and $\{\boldsymbol{w}_i\}_{i\geq 1}$ is stationary and ergodic (Hansen 1982).

The accuracy of this estimator depends on the choice of instruments $\boldsymbol{A}(\boldsymbol{v}_i)$ (Newey 1990). For example, if $\boldsymbol{A}(\boldsymbol{v}_i)$ is independent of $\boldsymbol{w}_i$, then the population moment

$$\mathbb{E}\{\boldsymbol{A}(\boldsymbol{v}_i) f(\boldsymbol{\theta}_*, \boldsymbol{w}_i)\} = \mathbb{E}\{\boldsymbol{A}(\boldsymbol{v}_i)\} \mathbb{E}\{f(\boldsymbol{\theta}_*, \boldsymbol{w}_i)\}$$

is $\mathbf{0}$ whenever $\mathbb{E}\{f(\boldsymbol{\theta}_*, \boldsymbol{w}_i)\} = \mathbf{0}$. The latter implies $\boldsymbol{\theta}_*$ may not be identifiable if $d > 1$, since $\mathbb{E}\{f(\boldsymbol{\theta}_*, \boldsymbol{w}_i)\}$ could then be $\mathbf{0}$ at an infinite number of points $\boldsymbol{\theta}_*$. Therefore, it is critical the instruments be strongly related to $\boldsymbol{w}_i$. We discuss how we choose appropriate instruments when we present our method in Section 5.4.1.

5.3.2 An instrumental variable generalized method of moments estimator when $\boldsymbol{C}$ is observed

Let $g \in [p]$ be a metabolite with missing data. For $\boldsymbol{\theta}_{*g} = \left(\alpha_{*g}\ \delta_{*g}\right)^{\mathrm{T}}$, we follow Wang et al. (2014), Ai et al. (2018) and define

$$f\left\{\boldsymbol{\theta}_{*g}, \left(\boldsymbol{x}_i, \boldsymbol{c}_i, r_{gi}, y_{gi}\right)\right\} = 1 - r_{gi}\Psi\left\{\alpha_{*g}\left(y_{gi} - \delta_{*g}\right)\right\}^{-1},$$

where $p = 1$ and $\boldsymbol{C}$ is observed in both Wang et al. (2014) and Ai et al. (2018). By Model (5.3),

$$\mathbb{E}\left[f\left\{\boldsymbol{\theta}_{*g}, \left(\boldsymbol{x}_i, \boldsymbol{c}_i, r_{gi}, y_{gi}\right)\right\} \mid \boldsymbol{x}_i, \boldsymbol{c}_i, y_{gi}\right] = 0, \tag{5.5}$$

meaning we need only construct appropriate instruments to estimate $\boldsymbol{\theta}_{*g}$ with (5.4). Since y_{gi} is only observed when $r_{gi} = 1$, we must rely on $\boldsymbol{x}_i$ and $\boldsymbol{c}_i$ to construct the instruments.

In typical metabolomic experiments, the observed covariates in $\boldsymbol{X}$ like disease status only account for a small fraction of the variance in $\boldsymbol{Y}$. The majority, as it turns out, is explainable by the latent covariates $\boldsymbol{C}$ (De Livera et al. 2012, 2015, Salerno et al. 2017). We therefore assume for simplicity that $\boldsymbol{X} = \mathbf{1}_n$. Define $\boldsymbol{L} = \left(\boldsymbol{\ell}_{*_1} \cdots \boldsymbol{\ell}_{*_p}\right)^{\mathrm{T}}$. Since $\boldsymbol{C}$ and $\boldsymbol{L}$ are only identifiable through their product $\boldsymbol{L}\boldsymbol{C}^{\mathrm{T}}$ in (5.1), it suffices to assume that for $\lambda_1 \geq \cdots \geq \lambda_K > 0$, $\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L} = \operatorname{diag}(\lambda_1, \ldots, \lambda_K)$ and $n^{-1}\boldsymbol{C}^{\mathrm{T}} P_{1_n}^{\perp} \boldsymbol{C} = I_K$. That is, $[\boldsymbol{C}]_{*k}$ explains the kth largest proportion of the variance in $\boldsymbol{Y}$. By the discussion in Section 5.3.1, a reasonable choice of instruments for metabolite g is $\boldsymbol{A}_g\left(\boldsymbol{x}_i, \boldsymbol{c}_i, y_{gi}\right) = \left(1, \left[P_{1_n}^{\perp}\boldsymbol{C}\right]_{ij_1}, \ldots, \left[P_{1_n}^{\perp}\boldsymbol{C}\right]_{ij_k}\right)^{\mathrm{T}} \in \mathbb{R}^{k+1}$ if $\boldsymbol{C}$ were observed, where $k \geq 1$ and $j_1, \ldots, j_k$ are distinct elements in $[K]$. We remark that one would typically choose $j_1 = 1, \ldots, j_k = k$. A reasonable choice for k is $k = 2$ for reasons discussed in Section 5.4.3.

Evidently, $\boldsymbol{C}$ is not observed in Model (5.1) and therefore must be estimated. We present an overview of our method to estimate α_{*_g} and δ_{*_g}, as well as $\boldsymbol{C}$, $\boldsymbol{\beta}_{*_g}$ and $\sigma_{*_g}^2$, below in Section 5.4.1.

5.4 Estimating the metabolite-dependent missingness mechanisms

5.4.1 An overview of the estimation and inference procedure

While our ultimate goal is to estimate and do inference on the coefficients of interest $\boldsymbol{\beta}_{*_1}, \ldots, \boldsymbol{\beta}_{*_p}$ in (5.1b), we are also interested in performing dimension reduction, estimating the fraction of variance in metabolite concentration explained by a covariate, conducting a metabolite genome wide association study (GWAS) and estimating the mean and variance of a metabolite's observed integrated intensity. The latter may be useful when optimizing a protocol for a specific class of metabolites. We present an overview of our methodology in Algorithm 5.1 below, which uses the intuition developed in Sections 5.3.1 and 5.3.2 to estimate the missingness mechanisms for each metabolite with missing data.

***Algorithm* 5.1.** *Define the full-rank matrix* $\boldsymbol{Z} = (\boldsymbol{z}_1 \cdots \boldsymbol{z}_n)^{\mathrm{T}} \in \mathbb{R}^{n \times t}$ *to be such that* $t \leq 3$

and $\mathbf{1}_n \in \operatorname{Im}(\boldsymbol{Z}) \subseteq \operatorname{Im}(\boldsymbol{X})$. *Let* $\mathcal{S} \subseteq [p]$ *be such that* $n^{-1}\sum_{i=1}^{n}(1-r_{gi}) > \epsilon$ *for all* $g \in \mathcal{S}$ *for some user-defined* $\epsilon \geq 0$.

(a) Let $\boldsymbol{Y}_{\mathcal{S}^c} \in \mathbb{R}^{|\mathcal{S}^c|\times n}$ *be the submatrix of* $\boldsymbol{Y}$ *restricted to the rows* $g \in \mathcal{S}^c$. *Define* $\hat{\boldsymbol{C}}_\perp$ *to be the maximum likelihood estimate for* $P_Z^\perp \boldsymbol{C}$, *assuming any missing data in* $\boldsymbol{Y}_{\mathcal{S}^c}$ *are missing completely at random and*

$$\boldsymbol{Y}_{\mathcal{S}^c} \sim MN_{|\mathcal{S}^c|\times n}\left\{\tilde{\boldsymbol{L}}\left(P_Z^\perp \boldsymbol{C}\right)^{\mathrm{T}} + \tilde{\boldsymbol{\beta}}\boldsymbol{Z}^{\mathrm{T}}, \sigma^2 I_{|\mathcal{S}^c|}, I_n\right\},$$

where $P_Z^\perp \boldsymbol{C}$, $\tilde{\boldsymbol{L}}$, $\tilde{\boldsymbol{\beta}}$ *and* σ^2 *are unknown parameters and* $\tilde{\boldsymbol{L}}^{\mathrm{T}}\tilde{\boldsymbol{L}}$ *is diagonal with non-increasing elements.*

(b) If $t = 3$, *define* $\hat{\boldsymbol{U}}_g = \boldsymbol{Z}$ *for all* $g \in \mathcal{S}$. *If* $t < 3$:

(i) For each $j \in \left[\hat{K}\right]$, *use ordinary least squares (OLS) to regress* $\boldsymbol{y}_g$ *onto* $\boldsymbol{A}_{g,j} = \left(\boldsymbol{Z}\ \left[\hat{\boldsymbol{C}}_\perp\right]_{*j}\right) \in \mathbb{R}^{n\times(t+1)}$, *where all missing data is assumed missing completely at random, and let* $p_{g,j}$ *be the OLS P value corresponding to the* $t+1$*st column of* $\boldsymbol{A}_{g,j}$.

(ii) For $j \in \left[\hat{K}\right]$, *use* $\left\{p_{g',j}\right\}_{g'\in\mathcal{S}}$ *to determine the corresponding q-values* $\left\{q_{g',j}\right\}_{g'\in\mathcal{S}}$.

(iii) Let $j_1, \ldots, j_{3-t} \in \left[\hat{K}\right]$ *be the indices of the* $3-t$ *smallest q-values* $\left\{q_{g,1}, \ldots, q_{g,\hat{K}}\right\}$. *Define* $\hat{\boldsymbol{U}}_g = \left(\boldsymbol{Z}\ \left[\hat{\boldsymbol{C}}_\perp\right]_{*j_1} \cdots \left[\hat{\boldsymbol{C}}_\perp\right]_{*j_{3-t}}\right) \in \mathbb{R}^{n\times 3}$, *where* $j_1, \ldots, j_{3-t}$ *depend on* g.

(c) Use $\hat{\boldsymbol{U}}_g$ *as instrumental variables in our method **H**ierarchical **B**ayesian **G**eneralize **M**ethod of **M**oments (HB-GMM) to estimate* α_{*g}, δ_{*g} *for each* $g \in \mathcal{S}$.

(d) Use the estimated metabolite-dependent missingness mechanism to estimate $\boldsymbol{C}$ *in Model* (5.1).

(e) Estimate the parameters of interest from Model (5.1).

The set $\mathcal{S}$ contains metabolites with a non-trivial amount of missing data, i.e. all metabolites whose observations are missing in more than $100 \times \epsilon\%$ of the samples. Our software default is to let $\epsilon = 0.05$. In Section 5.3.2, we assumed for simplicity that $\boldsymbol{Z} = \boldsymbol{1}_n$, which is generally the case in real data applications. However, we allow for the possibility that the practitioner has access to additional informative covariates.

Just like $P_{1_n}^{\perp}\boldsymbol{C}$ in Section 5.3.2, the columns of the estimate for $P_Z^{\perp}\boldsymbol{C}$ in Step (a) serve as candidate instrumental variables, where $\hat{\boldsymbol{C}}_{\perp}$ is a scalar multiple of the first $\hat{K}$ right singular vectors of $\boldsymbol{Y}_{\mathcal{S}^c}P_Z^{\perp}$ if $\epsilon = 0$. When $\epsilon > 0$, the assumption that missing data are missing completely at random has a negligible effect on the estimate for $P_Z^{\perp}\boldsymbol{C}$ because of the small number of missing observations for metabolites in $\mathcal{S}^c$.

We use the estimate $\hat{\boldsymbol{U}}_g$ in Step (b) as an instrumental variables for the missingness mechanism, meaning we require $\dim\left\{\mathrm{Im}\left(\hat{\boldsymbol{U}}_g\right)\right\} \geq 2$. We set $\dim\left\{\mathrm{Im}\left(\hat{\boldsymbol{U}}_g\right)\right\} = 3$ for reasons described in Section 5.4.3. It is critical that the ith row of $\hat{\boldsymbol{U}}_g$ be correlated with y_{gi} if $r_{gi} = 0$. The selection Step (b) helps ensure this is the case, and is justified by Theorem 5.1 in Section 5.10.5, in which we prove the P value $p_{g,j}$ is asymptotically uniform under minor regularity conditions and the null hypothesis that $\left[\hat{\boldsymbol{C}}_{\perp}\right]_{*j}$ is independent of $\boldsymbol{y}_g$.

A final important consideration is how the data-dependent instrument selection will impact the accuracy of generalized method of moments. Suppose $t < 3$, and let $T_g = T_g\left(j_1, \ldots, j_{3-t}\right)$ be any non-negative statistic regarding the accuracy of the instrumental variable generalized method of moments estimator whose instruments are $\boldsymbol{Z}$ and columns $j_1, \ldots, j_{3-t} \in \left[\hat{K}\right]$ of $\hat{\boldsymbol{C}}_{\perp}$. For example, T_g might be the mean squared error or the indicator that the estimates for $\alpha_{*_g}, \delta_{*_g}$ lie outside a small ball centered at the true values $\alpha_{*_g}, \delta_{*_g}$. Then for the event

$$E = \left\{\text{any one of the selected columns of } \hat{\boldsymbol{C}}_{\perp} \text{ are independent of } \boldsymbol{y}_g\right\},$$

we have

$$\begin{aligned}
&\sum_{j_1<\cdots<j_{3-t}\in\hat{K}} \mathbb{E}\left\{T_g\left(j_1,\ldots,j_{3-t}\right) I\left(\text{Columns } j_1,\ldots,j_{3-t} \text{ are selected}\right)\right\}\\
\leq &\sum_{\substack{\text{Columns } j_1<\cdots<j_{3-t}\\ \text{are not independent of } \boldsymbol{y}_g}} \mathbb{E}\left\{T_g\left(j_1,\ldots,j_{3-t}\right)\right\} + c_g\,\mathbb{P}\left(E\right),
\end{aligned}$$

where

$$c_g = \mathbb{E}\left\{\sum_{\substack{\text{Column } j_1 \text{ or } \ldots \text{ or } j_{3-t}\\ \text{is independent of } \boldsymbol{y}_g}} T_g\left(j_1,\ldots,j_{3-t}\right) \mid E\right\}.$$

Therefore, if $\mathbb{P}\left(E\right)$ is small and under the assumption that the columns of $P_Z^{\perp}\boldsymbol{C}$ that are not independent of $\boldsymbol{y}_g$ are suitable instruments, it suffices to assume the instruments in $\hat{\boldsymbol{U}}_g$ are given and not selected when we describe and motivate our method throughout the remainder of Section 5.4, as well as Sections 5.5 and 5.6. The assumption that $\mathbb{P}\left(E\right)$ is small tends to hold in real data. For example, in our two data applications, we set $\boldsymbol{Z}=\boldsymbol{1}_n$, $\hat{K}=10$ and found the median smallest and second smallest q-value from Step (b) to be no greater than 10^{-7} and 10^{-3}, respectively.

5.4.2 Initial estimates with generalized method of moments

Here we use $\hat{\boldsymbol{U}}_g$ from Step (b) of Algorithm 5.1 to get initial estimates for α_{*_g} and δ_{*_g} for all $g\in\mathcal{S}$. Fix a $g\in\mathcal{S}$, let $\hat{\boldsymbol{U}}_g=\left(\hat{\boldsymbol{u}}_{g1}\cdots\hat{\boldsymbol{u}}_{gn}\right)^{\mathrm{T}}$ and define

$$w_{gi}\left(\boldsymbol{\theta}_g\right)=\Psi\left\{\alpha_g\left(y_{gi}-\delta_g\right)\right\}^{-1},\quad \boldsymbol{\theta}_g=\left(\alpha_g,\delta_g\right)^{\mathrm{T}}\quad (i\in[n]) \tag{5.6a}$$

$$\boldsymbol{G}_{gi}\left(\boldsymbol{\theta}_g\right)=\hat{\boldsymbol{u}}_{gi}\left\{1-r_{gi}w_{gi}\left(\boldsymbol{\theta}_g\right)\right\}\quad (i\in[n]), \tag{5.6b}$$

Note that $\boldsymbol{G}_{gi}$ is an observable quantity, and is simply $\hat{\boldsymbol{u}}_{gi}$ if the ith observation is missing. If we let $\boldsymbol{\theta}_{g*} = \left(\alpha_{*_g}, \delta_{*_g}\right)^{\mathrm{T}}$, then just like (5.5),

$$\begin{aligned}\mathbb{E}\left\{1 - r_{gi} w_{gi}\left(\boldsymbol{\theta}_{*_g}\right) \mid \hat{\boldsymbol{u}}_{gi}\right\} &= \mathbb{E}\left[\mathbb{E}\left\{1 - r_{gi} w_{gi}\left(\boldsymbol{\theta}_{*_g}\right) \mid \hat{\boldsymbol{u}}_{gi}, y_{gi}\right\} \mid \hat{\boldsymbol{u}}_{gi}\right] \\ &= \mathbb{E}\left[\mathbb{E}\left\{1 - r_{gi} w_{gi}\left(\boldsymbol{\theta}_{*_g}\right) \mid y_{gi}\right\} \mid \hat{\boldsymbol{u}}_{gi}\right] = 0.\end{aligned}$$

The second equality follows because the missingness indicator r_{gi} is independent of $\hat{\boldsymbol{u}}_{gi}$ given y_{gi}, and the last equality follows from (5.3). As discussed in Sections 5.3.1 and 5.3.2, this implies one can use the instruments $\hat{\boldsymbol{U}}_g$ to define $\hat{\boldsymbol{\theta}}_g$, the two-step estimator for $\boldsymbol{\theta}_{*_g}$ defined in (5.4), where

$$f\left\{\boldsymbol{\theta}_{*_g}, \left(\hat{\boldsymbol{u}}_{gi}, y_{gi}, r_{gi}\right)\right\} = 1 - r_{gi} w_{gi}\left(\boldsymbol{\theta}_{*_g}\right), \quad \boldsymbol{A}\left(\hat{\boldsymbol{u}}_{gi}\right) = \hat{\boldsymbol{u}}_{gi}. \tag{5.7}$$

We justify this choice in Section 5.10.7, in which we show that under similar regularity assumptions used in Wang et al. (2014), $\hat{\boldsymbol{\theta}}_g$ is consistent for $\boldsymbol{\theta}_{*_g}$ and

$$\hat{\boldsymbol{V}}_g^{-1/2}\left(\hat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_g\right) \overset{\mathcal{D}}{=} N_2\left(\mathbf{0}, I_2\right) + o_P(1) \tag{5.8a}$$

$$\hat{\boldsymbol{V}}_g = \left[\left\{\sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}_g} \boldsymbol{G}_{gi}\left(\hat{\boldsymbol{\theta}}_g\right)\right\}^{\mathrm{T}} \left\{\sum_{i=1}^{n} \boldsymbol{G}_{gi}\left(\hat{\boldsymbol{\theta}}_g\right) \boldsymbol{G}_{gi}\left(\hat{\boldsymbol{\theta}}_g\right)^{\mathrm{T}}\right\}^{-1} \left\{\sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}_g} \boldsymbol{G}_{gi}\left(\hat{\boldsymbol{\theta}}_g\right)\right\}\right]^{-1} \tag{5.8b}$$

as $n, p \to \infty$. We use the asymptotic distribution of $\hat{\boldsymbol{\theta}}_g$ to refine the estimate for $\boldsymbol{\theta}_g$ by pooling information across metabolites in Section 5.4.4.

5.4.3 Identifying metabolites that do not follow Model (5.3)

The accuracy of $\hat{\boldsymbol{\theta}}_g$, and therefore that of downstream estimators, is contingent on the missing data model being approximately correct. Therefore, we developed a procedure to flag metabolites whose missingness mechanisms may not follow Model (5.3) using the Sargan-

Hansen J-test, which is generally used to test the moment restrictions used in generalized method of moment estimators (Sun & Kim 2012).

Assume $\dim\left\{\mathrm{Im}\left(\hat{\boldsymbol{U}}_g\right)\right\} = 3$ and for each $g \in \mathcal{S}$, define

$$\bar{\boldsymbol{G}}_g(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^{n}\hat{\boldsymbol{u}}_{gi}\left\{1 - r_{gi}w_{gi}(\boldsymbol{\theta})\right\} \tag{5.9a}$$

$$\bar{\boldsymbol{\Gamma}}_g(\boldsymbol{\theta}) = \nabla_\theta \bar{\boldsymbol{G}}_g(\boldsymbol{\theta}) = -n^{-1}\sum_{i=1}^{n} r_{gi}\hat{\boldsymbol{u}}_{gi}\left\{\nabla_\theta w_{gi}(\boldsymbol{\theta})\right\}^{\mathrm{T}} \tag{5.9b}$$

$$\boldsymbol{W}_g^{-1} = n^{-1}\sum_{i=1}^{n}\left[1 - r_{gi}w_{gi}\left\{\hat{\boldsymbol{\theta}}_g^{(1)}\right\}\right]^2 \hat{\boldsymbol{u}}_{gi}\hat{\boldsymbol{u}}_{gi}^{\mathrm{T}}, \tag{5.9c}$$

where $\hat{\boldsymbol{\theta}}_g^{(1)}$ is the first step estimator in (5.4) for $\boldsymbol{A}$ and f defined in (5.7). Then if $\bar{\boldsymbol{\Gamma}}_g\left(\hat{\boldsymbol{\theta}}_g\right)$ is full rank, the estimator $\hat{\boldsymbol{\theta}}_g$ satisfies

$$P_{W_g^{1/2}\bar{\Gamma}_g\left(\hat{\theta}_g\right)}\boldsymbol{W}_g^{1/2}\bar{\boldsymbol{G}}_g\left(\hat{\boldsymbol{\theta}}_g\right) = \boldsymbol{0},$$

where $\bar{\boldsymbol{G}}_g\left(\hat{\boldsymbol{\theta}}_g\right) \in \mathbb{R}^3$ and $P_{W_g^{1/2}\bar{\Gamma}_g\left(\hat{\theta}_g\right)}$ projects vectors in $\mathbb{R}^3$ onto a two dimensional subspace. Therefore, under the null hypothesis that Model (5.3) and the regularity assumptions used to prove (5.8) hold, we should be able to predict how the one degree of freedom quadratic form

$$n\bar{\boldsymbol{G}}_g\left(\hat{\boldsymbol{\theta}}_g\right)^{\mathrm{T}}\boldsymbol{W}_g\bar{\boldsymbol{G}}_g\left(\hat{\boldsymbol{\theta}}_g\right) = n\bar{\boldsymbol{G}}_g\left(\hat{\boldsymbol{\theta}}_g\right)^{\mathrm{T}}\boldsymbol{W}_g^{1/2}P^{\perp}_{W_g^{1/2}\bar{\Gamma}_g\left(\hat{\theta}_g\right)}\boldsymbol{W}_g^{1/2}\bar{\boldsymbol{G}}_g\left(\hat{\boldsymbol{\theta}}_g\right)$$

behaves. On the other hand, the above quadratic form will tend to be large if Model (5.3) is too inaccurate, since $n^{1/2}\bar{\boldsymbol{G}}_g\left(\hat{\boldsymbol{\theta}}_g\right)$ will generally not be stochastically bounded because the population moment $\mathbb{E}\left\{\bar{\boldsymbol{G}}_g(\boldsymbol{\theta})\right\}$ will likely not be $\boldsymbol{0}$ for any $\boldsymbol{\theta}$. We make this rigorous in Corollary 5.2, in which we show that when the null hypothesis

$$H_{0,g} : \mathbb{E}\left\{\bar{\boldsymbol{G}}_g\left(\boldsymbol{\theta}_{*g}\right)\right\} = \boldsymbol{0} \quad (g \in \mathcal{S})$$

and the same regularity conditions used to prove (5.8) hold,

$$J_g = n\bar{\boldsymbol{G}}_g\left(\hat{\boldsymbol{\theta}}_g\right)^{\mathrm{T}}\boldsymbol{W}_g\bar{\boldsymbol{G}}_g\left(\hat{\boldsymbol{\theta}}_g\right)\xrightarrow{\mathcal{D}}\chi_1^2 \quad (g\in\mathcal{S}) \tag{5.10}$$

as $n, p \to \infty$. The J-statistic, J_g, is typically used to check the validity of the assumed moment restrictions (Hansen 1982), and we therefore it to assess the veracity of Model (5.3).

It has been repeatedly observed that using the asymptotic chi-squared distribution to do inference with J_g is anti-conservative in data with moderate, and even large sample sizes (Hall & Horowitz 1996, Hansen & West 2002, Brown & Newey 2002). To circumvent this, we follow Brown & Newey (2002) and develop a bootstrap null distribution for J_g. Since the bootstrap sampling population is $\left\{\left(\hat{\boldsymbol{u}}_{gi}, y_{gi}, r_{gi}\right)\right\}_{i\in[n]}$, we require $H_{0,g}$ to hold at $\hat{\boldsymbol{\theta}}_g$ instead of $\boldsymbol{\theta}_{*g}$ for the bootstrap sampling population. However, this will not be the case if we use uniform resampling to generate boostrapped datasets, because $\bar{\boldsymbol{G}}_g\left(\hat{\boldsymbol{\theta}}_g\right)\neq\mathbf{0}$ when $\dim\left\{\mathrm{Im}\left(\hat{\boldsymbol{U}}_g\right)\right\}>2$. Instead, we define

$$d\hat{F}_g=\sum_{i=1}^n\gamma_{gi}\delta_{\left(\hat{\boldsymbol{u}}_{gi},y_{gi},r_{gi}\right)},\quad \sum_{i=1}^n\gamma_{gi}\boldsymbol{G}_{gi}\left(\hat{\boldsymbol{\theta}}_g\right)=\mathbf{0},\quad \sum_{i=1}^n\gamma_{gi}=1 \quad (g\in\mathcal{S}),$$

where

$$\left\{\gamma_{g1},\ldots,\gamma_{gn}\right\}=\underset{x_1,\ldots,x_n\in[0,1]}{\arg\max}\prod_{i=1}^n x_i,\quad \sum_{i=1}^n x_i=1,\quad \sum_{i=1}^n x_i\boldsymbol{G}_{gi}\left(\hat{\boldsymbol{\theta}}_g\right)=\mathbf{0}\quad (g\in\mathcal{S}).$$

If $\hat{\boldsymbol{\theta}}_g=\left(\hat{\alpha}_g,\hat{\delta}_g\right)^{\mathrm{T}}$ and $(\hat{\boldsymbol{u}}, y, r)\sim\hat{F}_g$, then $\mathbb{E}_{\hat{F}_g}\left(\hat{\boldsymbol{u}}\left[1-r\Psi\left\{\hat{\alpha}_g\left(y-\hat{\delta}_g\right)\right\}^{-1}\right]\right)=\mathbf{0}$, which is exactly what is required. We subsequently generate a bootstrap distribution for J_g by repeatedly sampling from $\hat{F}_g$ for all $g\in\mathcal{S}$, and estimate

$$lfdr_g = P\left[H_{0,g}\mid\left\{\left(\hat{\boldsymbol{u}}_{gi}, y_{gi}, r_{gi}\right)\right\}_{i\in[n]}\right]$$

using q-value (Storey et al. 2015). We then flag any metabolites with an $lfdr_g$ smaller than a user-specified value, which defaults to 0.8 in our software. Unsurprisingly, metabolites with a substantial amount of missing data ($> 40\%$) are typically the metabolites we flag in practice, as Model (5.3) is likely too simple to capture the subtleties of those missingness mechanisms.

5.4.4 Hierarchical Bayesian generalized method of moments

So far we have estimated each metabolite-specific missingness mechanism independently for each metabolite $g \in \mathcal{S}$. While the mechanisms are almost certainly not identical, one might expect them to be relatively similar, and that one should be able to design a better estimator by pooling information across metabolites. Further, constructing an informative prior on the missingness mechanisms allows one to better explore the objective function in (5.4), which could be multimodal (Domínguez & Lobato 2004, Franks et al. 2016). This would also give us better estimates for the uncertainty in our estimators. We therefore developed a Bayesian method to estimate $\boldsymbol{\theta}_{*_g}$ and the weights $w_{gi}\left(\boldsymbol{\theta}_{*_g}\right)$ defined in (5.6a) using the generalized method of moments estimator $\hat{\boldsymbol{\theta}}_g$ and its estimated variance $\hat{\mathbb{V}}\left(\hat{\boldsymbol{\theta}}_g\right)$ from Section 5.4.2. The weights $w_{gi}\left(\boldsymbol{\theta}_{*_g}\right)$ play an important role in estimating the effects of interest in Section 5.5. An important part of our method will be determining the posterior uncertainty of our estimates, as we incorporate them into our downstream estimators in Section 5.5.

To describe our method, let $\hat{\alpha}_g, \hat{\delta}_g$ and $\hat{\boldsymbol{\theta}}_g$ be the generalized method of moment estimates defined in Section 5.4.2 and let

$$\boldsymbol{\eta}_{*_g} = \left(\log\left(\alpha_{*_g}\right), \delta_{*_g}\right)^{\mathrm{T}}, \; \hat{\boldsymbol{\eta}}_g = \left(\log\left(\hat{\alpha}_g\right), \; \hat{\delta}_g\right)^{\mathrm{T}} \quad (g \in \mathcal{S}).$$

We prove in Theorem 5.2 in Section 5.10.6 that under minor regularity conditions on the moments of $\boldsymbol{y}_g$ and the behavior of $\Psi(x)$ as $x \to -\infty$,

$$n^{1/2} \left\{ \hat{\boldsymbol{\Sigma}}_g \left(\boldsymbol{\eta}_{*g} \right) \right\}^{-1/2} \bar{\boldsymbol{G}}_g \left(\boldsymbol{\eta}_{*g} \right) \xrightarrow{\mathcal{D}} N_3 \left(\mathbf{0}, I_3 \right) \quad (g \in \mathcal{S})$$

$$\hat{\boldsymbol{\Sigma}}_g \left(\boldsymbol{\eta}_{*g} \right) = (n-1)^{-1} \sum_{i=1}^{n} \left\{ \boldsymbol{G}_{gi} \left(\boldsymbol{\eta}_{*g} \right) - \bar{\boldsymbol{G}}_g \left(\boldsymbol{\eta}_{*g} \right) \right\} \left\{ \boldsymbol{G}_{gi} \left(\boldsymbol{\eta}_{*g} \right) - \bar{\boldsymbol{G}}_g \left(\boldsymbol{\eta}_{*g} \right) \right\}^{\mathrm{T}} \quad (g \in \mathcal{S}).$$

We therefore approximate the likelihood of $\bar{\boldsymbol{G}}_g \left(\boldsymbol{\eta}_{*g} \right) \mid \boldsymbol{\eta}_{*g}$ as

$$\bar{\boldsymbol{G}}_g \left(\boldsymbol{\eta}_{*g} \right) \mid \boldsymbol{\eta}_{*g} \sim N_3 \left(\mathbf{0}, n^{-1} \hat{\boldsymbol{\Sigma}}_g \left(\boldsymbol{\eta}_{*g} \right) \right) \quad (g \in \mathcal{S}) \tag{5.11}$$

Let $g \neq h \in \mathcal{S}$. Then because $\bar{\boldsymbol{G}}_g \left(\boldsymbol{\eta}_{*g} \right) \mid \left(\boldsymbol{\eta}_{*g}, \boldsymbol{y}_g, \hat{\boldsymbol{U}}_g \right)$ and $\bar{\boldsymbol{G}}_h \left(\boldsymbol{\eta}_{*h} \right) \mid \left(\boldsymbol{\eta}_{*h}, \boldsymbol{y}_h, \hat{\boldsymbol{U}}_h \right)$ are independent and $\mathbb{E} \left\{ \bar{\boldsymbol{G}}_g \left(\eta_{*g} \right) \mid \boldsymbol{\eta}_{*g}, \boldsymbol{y}_g, \hat{\boldsymbol{U}}_g \right\} = \mathbf{0}$,

$$\begin{aligned} \operatorname{Cov} \left\{ \bar{\boldsymbol{G}}_g \left(\boldsymbol{\eta}_{*g} \right), \bar{\boldsymbol{G}}_h \left(\boldsymbol{\eta}_{*h} \right) \mid \boldsymbol{\eta}_{*g}, \boldsymbol{\eta}_{*h} \right\} = \mathbb{E} \Big[& \operatorname{Cov} \Big\{ \bar{\boldsymbol{G}}_g \left(\boldsymbol{\eta}_{*g} \right), \bar{\boldsymbol{G}}_h \left(\boldsymbol{\eta}_{*h} \right) \mid \boldsymbol{\eta}_{*g}, \boldsymbol{\eta}_{*h}, \boldsymbol{y}_g, \boldsymbol{y}_h, \\ & \hat{\boldsymbol{U}}_g, \hat{\boldsymbol{U}}_h \Big\} \mid \boldsymbol{\eta}_{*g}, \boldsymbol{\eta}_{*h} \Big] = \mathbf{0}. \end{aligned}$$

We use this fact to justify modelling $\left\{ \bar{\boldsymbol{G}}_g \left(\boldsymbol{\eta}_{*g} \right) \mid \boldsymbol{\eta}_{*g} \right\}_{g \in \mathcal{S}}$ as independent and jointly normal with marginal distributions given by (5.11). We remark that a similar technique was also used in in Yin (2009), Li & Jiang (2016) to put a distribution on the sample average $\bar{\boldsymbol{G}}_g$, although both articles assume $p = 1$. The latter article gives the sufficient conditions under which the quasi-posterior using the quasi-likelihood in (5.11) and known prior for $\boldsymbol{\eta}_{*g}$ concentrates on the true posterior $\mathrm{pr} \left(\boldsymbol{\eta}_{*g} \mid \hat{\boldsymbol{\eta}}_g \right)$.

Lastly, we assume the prior for $\boldsymbol{\eta}_{*g}$ is

$$\boldsymbol{\eta}_{*g} \mid \left(\boldsymbol{\mu}, \boldsymbol{U} \right) \sim N_2 \left(\boldsymbol{\mu}, \boldsymbol{U} \right) \quad (g \in \mathcal{S})$$

for some $\boldsymbol{\mu} \in \mathbb{R}^2, \boldsymbol{U} \in \mathbb{R}^{2 \times 2}$. Bayes' rule implies

$$\mathrm{pr} \left\{ \boldsymbol{\eta}_{*g} \mid \bar{\boldsymbol{G}}_g \left(\boldsymbol{\eta}_g \right) \right\} \propto \mathrm{pr} \left\{ \bar{\boldsymbol{G}}_g \left(\boldsymbol{\eta}_{*g} \right) \mid \boldsymbol{\eta}_{*g} \right\} \mathrm{pr} \left(\boldsymbol{\eta}_{*g} \mid \boldsymbol{\mu}, \boldsymbol{U} \right) \quad (g \in \mathcal{S}),$$

and we use Markov Chain Monte Carlo to sample from the posterior, which we use to estimate

$$\mathbb{E}\left\{\alpha_{*_g} \mid \bar{\boldsymbol{G}}_g\left(\boldsymbol{\eta}_g\right)\right\}, \ \mathbb{E}\left\{\delta_{*_g} \mid \bar{\boldsymbol{G}}_g\left(\boldsymbol{\eta}_g\right)\right\} \tag{5.12a}$$

$$\mathbb{E}\left\{w_{gi}\left(\boldsymbol{\eta}_{*_g}\right) \mid \bar{\boldsymbol{G}}_g\left(\boldsymbol{\eta}_g\right)\right\}, \ \mathbb{V}\left\{w_{gi}\left(\boldsymbol{\eta}_{*_g}\right) \mid \bar{\boldsymbol{G}}_g\left(\boldsymbol{\eta}_g\right)\right\} \tag{5.12b}$$

for all $g \in \mathcal{S}$ and i such that $r_{gi} = 1$.

It remains to show how we determine the hyper-parameters $\boldsymbol{\mu}$ and $\boldsymbol{U}$. We first estimate $\boldsymbol{\mu}$ as $\hat{\boldsymbol{\mu}} = |\mathcal{S}|^{-1} \sum_{g \in \mathcal{S}} \hat{\boldsymbol{\eta}}_g$, and then estimate $\boldsymbol{U}$ with empirical Bayes, assuming

$$\hat{\boldsymbol{\eta}}_g \mid \boldsymbol{\eta}_{*_g}, \hat{\boldsymbol{V}}_g \sim N_2\left(\boldsymbol{\eta}_{*_g}, \hat{\boldsymbol{W}}_g\right), \quad \hat{\boldsymbol{W}}_g = \operatorname{diag}\left(\left[\hat{\boldsymbol{V}}_g\right]_{11}^{-1}, 1\right) \hat{\boldsymbol{V}}_g \operatorname{diag}\left(\left[\hat{\boldsymbol{V}}_g\right]_{11}^{-1}, 1\right) \quad (g \in \mathcal{S})$$

$$\boldsymbol{\eta}_{*_g} \mid \hat{\boldsymbol{\mu}}, \boldsymbol{U} \sim N_2\left(\hat{\boldsymbol{\mu}}, \boldsymbol{U}\right) \quad (g \in \mathcal{S}),$$

where $\hat{\boldsymbol{V}}_g$ is defined in (5.8). This allows us to run the Markov chains in parallel to speed up computation.

5.5 Inference on the coefficients of interest

5.5.1 Estimating the coefficients of interest

While there are many ways to estimate and perform inference on the coefficients of interest once one has an estimate for the metabolite-dependent missingness mechanism (Liang & Qin 2000, Wang et al. 2014, Chen et al. 2017, O'Brien et al. 2018, Hedeker et al. 2018), we focus on inverse probability weighting (Liang & Qin 2000) in this section because estimates are consistent, it obviates the need to specify a probability model for the missing data and computation is fast enough to perform a metabolite genome wide association study. We will assume the covariates $\boldsymbol{C}$ in Model (5.1) are known and, for notational convenience, re-write the mean model in (5.1b) as

$$\mu_{gi} = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_{*_g} \quad (g = 1, \ldots, p; i = 1, \ldots, n). \tag{5.13}$$

We first show how we estimate $\boldsymbol{\beta}_{*_g}$ and then describe our method to assess the uncertainty in our estimate in Section 5.5.2. We show how to use the method we develop here to estimate $\boldsymbol{C}$ in Section 5.6. Unless otherwise specified, all expectations and variances in this section (Section 5.5) are taken conditional on $\boldsymbol{X} = (\boldsymbol{x}_1 \cdots \boldsymbol{x}_n)^{\mathrm{T}}$.

Define the score function

$$\boldsymbol{s}_{gi}\left(\boldsymbol{\beta}_g\right) = \boldsymbol{x}_i \left(y_{gi} - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_g\right) \quad (i = 1, \ldots, n; g \in \mathcal{S}),$$

where we focus on metabolites with missing data, i.e. $g \in \mathcal{S}$, because estimation and inference with little to no missing data is straightforward. We define

$$\boldsymbol{f}_g\left(\boldsymbol{\beta}_g\right) = \sum_{i=1}^{n} r_{gi} \hat{w}_{gi} \hat{\gamma}_{gi} \boldsymbol{s}_{gi}\left(\boldsymbol{\beta}_g\right) \quad (g \in \mathcal{S}) \tag{5.14a}$$

$$\hat{w}_{gi} = \hat{\mathbb{E}}\left\{w_{gi}\left(\boldsymbol{\eta}_{*_g}\right) \mid \bar{\boldsymbol{G}}_g\left(\boldsymbol{\eta}_g\right)\right\} \quad \left((g,i) \in \left\{(g,i) \in \mathcal{S} \times [n] : r_{gi} = 1\right\}\right) \tag{5.14b}$$

$$\hat{\gamma}_{gi} = \hat{\mathbb{P}}\left(r_{gi} = 1 \mid \boldsymbol{x}_i\right) \quad (g \in \mathcal{S}; i \in [n]), \tag{5.14c}$$

where $\hat{\mathbb{E}}\left\{w_{gi}\left(\boldsymbol{\eta}_{*_g}\right) \mid \bar{\boldsymbol{G}}_g\left(\boldsymbol{\eta}_g\right)\right\}$ is the estimate for $\mathbb{E}\left\{w_{gi}\left(\boldsymbol{\eta}_{*_g}\right) \mid \bar{\boldsymbol{G}}_g\left(\boldsymbol{\eta}_g\right)\right\}$ defined in Section 5.4.4 and $\hat{\gamma}_{gi}$ is an estimate of $\gamma_{gi} = \mathbb{P}\left(r_{gi} = 1 \mid \boldsymbol{x}_i\right)$. If $\hat{w}_{gi} = w_{gi}\left(\boldsymbol{\eta}_{*_g}\right)$ and $\hat{\gamma}_{gi} = \gamma_{gi}$, then

$$\begin{aligned}
\mathbb{E}\left\{\boldsymbol{f}_g\left(\boldsymbol{\beta}_{*_g}\right) \mid \boldsymbol{X}\right\} &= \sum_{i=1}^{n} \gamma_{gi} \mathbb{E}\left\{w_{gi}\left(\boldsymbol{\eta}_{*_g}\right) \boldsymbol{s}_{gi}\left(\boldsymbol{\beta}_{*_g}\right) \mathbb{E}\left(r_{gi} \mid \boldsymbol{X}, \boldsymbol{y}_g\right) \mid \boldsymbol{X}\right\} \\
&= \sum_{i=1}^{n} \gamma_{gi} \mathbb{E}\left\{\boldsymbol{s}_{gi}\left(\boldsymbol{\beta}_{*_g}\right) \mid \boldsymbol{X}\right\} = \boldsymbol{0}.
\end{aligned} \tag{5.15}$$

The above equality holds provided γ_{gi} is not a function of the data $\boldsymbol{y}_g$. We include γ_{gi} to stabilize large weights $\hat{w}_{gi}$ and thereby reduce the variance of our estimates, since γ_{gi} will tend to be small if $\hat{w}_{gi}$ is large. This method of stabilized inverse probability weighting has been successfully applied to data that are missing at random (Xu, Ross, Raebel, Shetterly, Blanchette & Smith 2010), and we estimate γ_{gi} using a logistic regression with the estimated instruments $\hat{\boldsymbol{U}}_g$.

We define our estimate for $\boldsymbol{\beta}_{*g}$ as

$$\hat{\boldsymbol{\beta}}_g = \underset{\boldsymbol{\beta}_g \in \mathbb{R}^d}{\arg\min} \|\boldsymbol{f}_g\left(\boldsymbol{\beta}_g\right)\| = \left(\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{W}}_g\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{W}}_g\boldsymbol{y}_g \quad (g \in \mathcal{S}) \tag{5.16a}$$

$$\hat{\boldsymbol{W}}_g = \operatorname{diag}\left(r_{g1}\hat{w}_{g1}\hat{\gamma}_{g1}, \ldots, r_{gn}\hat{w}_{gn}\hat{\gamma}_{g1}\right) \quad (g \in \mathcal{S}). \tag{5.16b}$$

The accuracy of $\hat{\boldsymbol{\beta}}_g$ depends on the accuracy of the weights $\hat{w}_{gi}$ and because $\mathbb{E}\left\{\boldsymbol{s}_{gi}\left(\boldsymbol{\beta}_{*g}\right)\right\} = \boldsymbol{0}$. In view of the latter, we can replace the score function $\boldsymbol{s}_{gi}\left(\boldsymbol{\beta}_g\right)$ with any M-estimator, like Huber's or Tukey's robust estimators. We remark that we did not need to specify a probability model for the residuals e_{gi} to estimate $\hat{\boldsymbol{\beta}}_g$.

5.5.2 Quantifying the uncertainty of our estimators

Here we describe how we estimate $\mathbb{V}\left(\hat{\boldsymbol{\beta}}_g\right)$, which is important for constructing confidence intervals for and performing inference on $\boldsymbol{\beta}_g$ for $g \in \mathcal{S}$. Our estimator is a finite sample-corrected sandwich variance estimator that also accounts for the uncertainty in the estimated weights $\hat{w}_{gi}$. To describe the estimator, define

$$\boldsymbol{f}_g^*\left(\boldsymbol{\beta}_g\right) = \sum_{i=1}^n r_{gi}w_{gi}\left(\boldsymbol{\eta}_{*g}\right)\gamma_{gi}\boldsymbol{s}_{gi}\left(\boldsymbol{\beta}_g\right).$$

Then under suitable regularity conditions,

$$\boldsymbol{0} = n^{-1}\boldsymbol{f}_g\left(\hat{\boldsymbol{\beta}}_g\right) = n^{-1}\boldsymbol{f}_g^*\left(\boldsymbol{\beta}_{*g}\right) + n^{-1}\nabla_{\boldsymbol{\beta}_g}\boldsymbol{f}_g^*\left(\boldsymbol{\beta}_{*g}\right)\left(\hat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_{*g}\right) + o_P\left(\|\hat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_{*g}\|_2\right)$$

$$\nabla_{\boldsymbol{\beta}_g}\boldsymbol{f}_g^*\left(\boldsymbol{\beta}_{*g}\right) = \sum_{i=1}^n r_{gi}w_{gi}\left(\boldsymbol{\eta}_{*g}\right)\gamma_{gi}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}.$$

Therefore, we can approximate the variance of $\hat{\boldsymbol{\beta}}_g$ as

$$\begin{aligned}
n\,\mathbb{V}\left(\hat{\boldsymbol{\beta}}_g\right) \approx & \left\{n^{-1}\nabla_{\boldsymbol{\beta}_g}\boldsymbol{f}_g^*\left(\boldsymbol{\beta}_{*_g}\right)\right\}^{-1}\left[n^{-1}\mathbb{E}\left\{\sum_{i=1}^{n} r_{gi}w_{gi}\left(\boldsymbol{\eta}_{*_g}\right)^2\gamma_{gi}^2\boldsymbol{s}_{gi}\left(\boldsymbol{\beta}_{*_g}\right)\boldsymbol{s}_{gi}\left(\boldsymbol{\beta}_{*_g}\right)^{\mathrm{T}}\right\}\right] \\
& \times\left\{n^{-1}\nabla_{\boldsymbol{\beta}_g}\boldsymbol{f}_g^*\left(\boldsymbol{\beta}_{*_g}\right)\right\}^{-1} \\
\approx & \left(n^{-1}\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{W}}_g\boldsymbol{X}\right)^{-1}\left[n^{-1}\sum_{i=1}^{n}\gamma_{gi}^2\mathbb{E}\left\{r_{gi}w_{gi}\left(\boldsymbol{\eta}_{*_g}\right)^2\boldsymbol{s}_{gi}\left(\boldsymbol{\beta}_{*_g}\right)\boldsymbol{s}_{gi}\left(\boldsymbol{\beta}_{*_g}\right)^{\mathrm{T}}\right\}\right] \\
& \times\left(n^{-1}\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{W}}_g\boldsymbol{X}\right)^{-1},
\end{aligned}$$

where we have replaced $w_{gi}\left(\boldsymbol{\eta}_{*_g}\right)$ and γ_{gi} with $\hat{w}_{gi}$ and $\hat{\gamma}_{gi}$ to estimate $\nabla_{\boldsymbol{\beta}_g}\boldsymbol{f}_g^*\left(\boldsymbol{\beta}_{*_g}\right)$ in the second approximate equality.

It remains to estimate the middle term. Simply plugging in $\hat{w}_{gi}$ for $w_{gi}\left(\boldsymbol{\eta}_{*_g}\right)$ when $w_{gi}\left(\boldsymbol{\eta}_{*_g}\right)$ is large will typically result in underestimates for $\mathbb{V}\left(\hat{\boldsymbol{\beta}}_g\right)$, since the uncertainty in estimates for $w_{gi}\left(\boldsymbol{\eta}_{*_g}\right)$ increases as $w_{gi}\left(\boldsymbol{\eta}_{*_g}\right)$ increases. Further, simply plugging in $\hat{\boldsymbol{\beta}}_g$ for $\boldsymbol{\beta}_{*_g}$ will underestimate $\mathbb{V}\left(\hat{\boldsymbol{\beta}}_g\right)$ because the plug-in sandwich variance estimator can notably underestimate the variance of $\hat{\boldsymbol{\beta}}_g$ (Wang et al. 2016). We circumvent the former by replacing $w_{gi}\left(\boldsymbol{\eta}_{*_g}\right)^2$ with $\hat{\mathbb{E}}\left\{w_{gi}\left(\boldsymbol{\eta}_{*_g}\right)^2 \mid \bar{\boldsymbol{G}}_g\left(\boldsymbol{\eta}_g\right)\right\}$, where

$$\begin{aligned}
\hat{\mathbb{E}}\left\{w_{gi}\left(\boldsymbol{\eta}_{*_g}\right)^2 \mid \bar{\boldsymbol{G}}_g\left(\boldsymbol{\eta}_g\right)\right\} &= \hat{\mathbb{E}}\left\{w_{gi}\left(\boldsymbol{\eta}_{*_g}\right) \mid \bar{\boldsymbol{G}}_g\left(\boldsymbol{\eta}_g\right)\right\}^2 + \hat{\mathbb{V}}\left\{w_{gi}\left(\boldsymbol{\eta}_{*_g}\right) \mid \bar{\boldsymbol{G}}_g\left(\boldsymbol{\eta}_g\right)\right\} \\
&= \hat{w}_{gi}^2 + \hat{\mathbb{V}}\left\{w_{gi}\left(\boldsymbol{\eta}_{*_g}\right) \mid \bar{\boldsymbol{G}}_g\left(\boldsymbol{\eta}_g\right)\right\} \geq \hat{w}_{gi}^2.
\end{aligned}$$

The above expression holds with equality if and only if $\hat{\mathbb{V}}\left\{w_{gi}\left(\boldsymbol{\eta}_{*_g}\right) \mid \bar{\boldsymbol{G}}_g\left(\boldsymbol{\eta}_g\right)\right\} = 0$, i.e. there is no uncertainty in our estimate for $w_{gi}\left(\boldsymbol{\eta}_{*_g}\right)$. We lastly show how we account for the uncertainty in $\hat{\boldsymbol{\beta}}_g$ in Section 5.10.2, which we use to estimate $\mathbb{V}\left(\hat{\boldsymbol{\beta}}_g\right)$ as

$$\begin{aligned}\hat{\mathbb{V}}\left(\hat{\boldsymbol{\beta}}_g\right) = &\left(\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{W}}_g\boldsymbol{X}\right)^{-1}\left\{\sum_{i=1}^{n} r_{gi}\hat{v}_{gi}\hat{\gamma}_{gi}\left(1-\hat{h}_{gi}\right)^{-2}\boldsymbol{s}_{gi}\left(\hat{\boldsymbol{\beta}}_g\right)\boldsymbol{s}_{gi}\left(\hat{\boldsymbol{\beta}}_g\right)^{\mathrm{T}}\right\}\\ &\times\left(\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{W}}_g\boldsymbol{X}\right)^{-1} \quad (g\in\mathcal{S})\end{aligned} \tag{5.17a}$$

$$\hat{v}_{gi} = \hat{\mathbb{E}}\left\{w_{gi}\left(\boldsymbol{\eta}_{*g}\right)^2 \mid \bar{\boldsymbol{G}}_g\left(\boldsymbol{\eta}_g\right)\right\} \quad (g\in\mathcal{S}; i\in[n]) \tag{5.17b}$$

$$\hat{h}_{gi} = r_{gi}\hat{w}_{gi}\hat{\gamma}_{gi}\boldsymbol{x}_i^{\mathrm{T}}\left(\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{W}}_g\boldsymbol{X}\right)^{-1}\boldsymbol{x}_i \quad (g\in\mathcal{S}; i\in[n]) \tag{5.17c}$$

The term $\left(1-\hat{h}_{gi}\right)^{-2}$ is a finite sample correction, where $\hat{h}_{gi}$ is the ith leverage score of $\hat{\boldsymbol{W}}_g^{1/2}\boldsymbol{X}$. As far as we are aware, this is the first such finite sample variance correction for inverse probability weighted estimators in data with non-random missing data. We use a similar technique in Section 5.10.3 to derive the estimator for σ_{*g}^2, which we define as

$$\hat{\sigma}_g^2 = \operatorname{tr}\left(\hat{\boldsymbol{W}}_g\right)^{-1}\sum_{i=1}^{n} r_{gi}\hat{w}_{gi}\hat{\gamma}_{gi}\left(1-\hat{h}_{gi}\right)^{-2}\left(y_{gi}-\boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_g\right)^2 \quad (g\in\mathcal{S}). \tag{5.18}$$

5.6 Estimating $\boldsymbol{C}$ in the presence of non-random missing data

Here we describe our method for estimating the latent covariates $\boldsymbol{C}$ in Model (5.1), and will assume throughout Section 5.6 that μ_{gi} is as defined in (5.1b), where $\boldsymbol{X}$ is observed. Define $\boldsymbol{X}_{\text{int}}, \boldsymbol{X}_{\text{nuis}}$ such that $\boldsymbol{X} = (\boldsymbol{X}_{\text{int}}\, \boldsymbol{X}_{\text{nuis}})$, where $\boldsymbol{X}_{\text{int}}$ contains the covariates of interest like disease status and $\boldsymbol{X}_{\text{nuis}}$ contains observed nuisance covariates like the intercept and technical factors. Since typical estimates for the coefficients of interest, like the estimator we presented in Section 5.5, only depend on $\boldsymbol{X}_{\text{int}}$ and $\operatorname{Im}(\boldsymbol{X}_{\text{nuis}}) \cup \operatorname{Im}\left(P_{X_{\text{nuis}}}^{\perp}\boldsymbol{C}\right)$, our goal is to estimate $\operatorname{Im}\left(P_{X_{\text{nuis}}}^{\perp}\boldsymbol{C}\right)$. This is quite auspicious because even though $\boldsymbol{C}$ is not identifiable in Model (5.1), $\operatorname{Im}\left(P_{X_{\text{nuis}}}^{\perp}\boldsymbol{C}\right)$ is identifiable under assumptions on the sparsity of the effects of interest in $\boldsymbol{\beta}_{*g}$ (McKennan & Nicolae 2018*a*,*b*). We assume for simplicity that $\boldsymbol{X} = \boldsymbol{X}_{\text{int}}$ in the presentation of our method, and we describe the extension when $\boldsymbol{X} = (\boldsymbol{X}_{\text{int}}\, \boldsymbol{X}_{\text{nuis}})$ in Section 5.10.4.

We use a strategy similar to that in Sun et al. (2012), Wang et al. (2017), Lee et al. (2017), McKennan & Nicolae (2018*a*,*b*) and estimate $\mathrm{Im}\,(\boldsymbol{C})$ by separately estimating $\mathrm{Im}\left(P_X^{\perp}\boldsymbol{C}\right)$ and $\mathrm{Im}\,(P_X\boldsymbol{C})$, and then combine the estimates to recover $\mathrm{Im}\,(\boldsymbol{C})$. In vector form,

$$\boldsymbol{y}_g = \boldsymbol{X}\boldsymbol{\beta}_{*_g} + \boldsymbol{C}\boldsymbol{\ell}_{*_g} + \boldsymbol{e}_g, \quad \boldsymbol{e}_g \dot{\sim} \left(\boldsymbol{0}, \sigma_{*_g}^2 I_n\right) \quad (g = 1, \ldots, p).$$

Then we can also express $\boldsymbol{y}_g$ as

$$\boldsymbol{y}_g = \boldsymbol{X}\tilde{\boldsymbol{\beta}}_{*_g} + \boldsymbol{C}_2\boldsymbol{\ell}_{*_g} + \boldsymbol{e}_g, \quad \tilde{\boldsymbol{\beta}}_{*_g} = \boldsymbol{\beta}_{*_g} + \boldsymbol{\Omega}\boldsymbol{\ell}_{*_g}, \quad (g = 1, \ldots, p) \tag{5.19a}$$

$$\boldsymbol{\Omega} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{C}, \quad \boldsymbol{C}_2 = P_X^{\perp}\boldsymbol{C}. \tag{5.19b}$$

Let $\boldsymbol{R}_g = \mathrm{diag}\left(r_{g1}, \ldots, r_{gn}\right)$ and $\mathcal{S}_1 \subseteq \mathcal{S}$, where our software's default $\mathcal{S}_1$ is all metabolites with less than 25% missing data and with an $lfdr_g$ above 0.8 (see Section 5.4.3). We then estimate $\tilde{\boldsymbol{\beta}}_{*_g}, \boldsymbol{\ell}_{*_g}$ and $\boldsymbol{C}_2$ using the following quasi log-likelihood objective function:

$$\begin{aligned}
&\left(\left\{\hat{\tilde{\boldsymbol{\beta}}}_g\right\}_{g \in \mathcal{S}^c \cup \mathcal{S}_1}, \left\{\hat{\boldsymbol{\ell}}_g\right\}_{g \in \mathcal{S}^c \cup \mathcal{S}_1}, \hat{\boldsymbol{C}}_2\right) \\
&= \underset{\substack{\tilde{\boldsymbol{\beta}}_g \in \mathbb{R}^d, \boldsymbol{\ell}_g \in \mathbb{R}^K, \boldsymbol{C}_2 \in \mathbb{R}^{n \times K} \\ \boldsymbol{C}_2^{\mathrm{T}}\boldsymbol{X} = \boldsymbol{0}, n^{-1}\boldsymbol{C}_2^{\mathrm{T}}\boldsymbol{C}_2 = I_K}}{\arg\max} \left[-\sum_{g \in \mathcal{S}^c} \|\boldsymbol{R}_g\left\{\boldsymbol{y}_g - \left(\boldsymbol{X}\tilde{\boldsymbol{\beta}}_g + \boldsymbol{C}_2\boldsymbol{\ell}_g\right)\right\}\|_2^2\right. \\
&\qquad \left. - \sum_{g \in \mathcal{S}_1} \|\hat{\boldsymbol{W}}_g^{1/2}\left\{\boldsymbol{y}_g - \left(\boldsymbol{X}\tilde{\boldsymbol{\beta}}_g + \boldsymbol{C}_2\boldsymbol{\ell}_g\right)\right\}\|_2^2\right].
\end{aligned} \tag{5.20}$$

The matrix $\boldsymbol{R}_g = I_n$ if ϵ, defined in Algorithm 5.1, is 0. When $\epsilon > 0$, this optimization treats missing data from metabolites with little to no missing data as missing completely at random. Note that for fixed $\boldsymbol{C}_2$, the updates for $\tilde{\boldsymbol{\beta}}_g$ and $\boldsymbol{\ell}_g$ for all $g \in \mathcal{S}_1$ are the stabilized inverse probability weighted estimates given in (5.16). The update for $\boldsymbol{C}_2$ given $\tilde{\boldsymbol{\beta}}_g$ and $\boldsymbol{\ell}_g$ for all $g \in \mathcal{S}^c \cup \mathcal{S}_1$ is similar. We lastly remark that the constraints on $\boldsymbol{C}_2$ define it uniquely up to a $K \times K$ rotation matrix. However, since our goal is only to estimate $\mathrm{Im}\,(\boldsymbol{C}_2)$, it suffices to estimate $\boldsymbol{C}_2$ up to a change of basis of its column space.

It remains to define $\hat{\boldsymbol{\Omega}}$, our estimator for $\boldsymbol{\Omega}$ from (5.19), where our estimate for $\boldsymbol{C}$ will then be

$$\hat{\boldsymbol{C}} = \boldsymbol{X}\hat{\boldsymbol{\Omega}} + \hat{\boldsymbol{C}}_2. \tag{5.21}$$

The estimates for $\tilde{\boldsymbol{\beta}}_{*_g}$ and $\boldsymbol{\ell}_{*_g}$ obtained from (5.20) are simply the inverse probability weighted estimates defined in (5.16), where the design matrix is $\left(\boldsymbol{X}\,\hat{\boldsymbol{C}}_2\right)$ and the weight matrix is $\boldsymbol{R}_g$ for $g \in \mathcal{S}^c$ and $\hat{\boldsymbol{W}}_g$ for $g \in \mathcal{S}_1$. Since $\boldsymbol{X}$ is orthogonal to $\boldsymbol{C}_2$, (5.15) is satisfied when evaluated at $\left(\tilde{\boldsymbol{\beta}}_{*_g}^{\mathrm{T}}, \boldsymbol{\ell}_{*_g}^{\mathrm{T}}\right)^{\mathrm{T}}$ for $g \in \mathcal{S}$. Therefore, we model $\hat{\tilde{\boldsymbol{\beta}}}_g$ as

$$\hat{\tilde{\boldsymbol{\beta}}}_g \,\dot{\sim}\, \left(\boldsymbol{\beta}_{*_g} + \boldsymbol{\Omega}\boldsymbol{\ell}_g, \hat{\boldsymbol{\nu}}_g\right) \quad \left(g \in \mathcal{S}^c \cup \mathcal{S}_1\right)$$

where $\hat{\boldsymbol{\nu}}_g$ is the ordinary least squares estimate for the variance of $\hat{\tilde{\boldsymbol{\beta}}}_g$ using the design matrix $\left(\boldsymbol{X}\,\hat{\boldsymbol{C}}_2\right)$ when $g \in \mathcal{S}^c$, and $\hat{\boldsymbol{\nu}}_g$ is a submatrix of the estimate for the variance defined in (5.17) when $g \in \mathcal{S}_1$. If only a small number of the effects of interest $\boldsymbol{\beta}_{*_1}, \ldots, \boldsymbol{\beta}_{*_p}$ are non-zero, then we can use the estimates for $\boldsymbol{\ell}_{*_g}$ obtained from (5.20) to regress the estimates for $\tilde{\boldsymbol{\beta}}_{*_g}$ onto those of $\boldsymbol{\ell}_{*_g}$ to estimate $\boldsymbol{\Omega}$. We describe this procedure in Algorithm 5.2.

***Algorithm* 5.2** (Estimating $\boldsymbol{\Omega}$). *Let $\epsilon_{\text{q-value}} \in [0, 1]$ and $R \geq 0$ be an integer.*

(0) Define $\hat{\boldsymbol{\Omega}}^{(0)} = \left(\hat{\boldsymbol{\Omega}}_1^{(0)} \cdots \hat{\boldsymbol{\Omega}}_d^{(0)}\right)^{\mathrm{T}}$ such that

$$\hat{\boldsymbol{\Omega}}_j^{(0)} = \left(\sum_{g \in \mathcal{S}^c \cup \mathcal{S}_1} \left[\hat{\boldsymbol{\nu}}_g\right]_{jj}^{-1} \hat{\boldsymbol{\ell}}_g \hat{\boldsymbol{\ell}}_g^{\mathrm{T}}\right)^{-1} \left(\sum_{g \in \mathcal{S}^c \cup \mathcal{S}_1} \left[\hat{\boldsymbol{\nu}}_g\right]_{jj}^{-1} \left[\hat{\tilde{\boldsymbol{\beta}}}_g\right]_j \hat{\boldsymbol{\ell}}_g\right).$$

Define $\hat{\boldsymbol{C}}^{(0)} = \boldsymbol{X}\hat{\boldsymbol{\Omega}}^{(0)} + \hat{\boldsymbol{C}}_2$. If $R = 0$, return $\hat{\boldsymbol{C}} = \hat{\boldsymbol{C}}^{(0)}$.

(1) Let $\hat{\boldsymbol{C}}^{(r)}$ be given. Define $p_{g,j}$ to be the P value for the null hypothesis $H_{0,g,j} : \left[\boldsymbol{\beta}_g\right]_j = 0$ using ordinary least squares if $g \in \mathcal{S}^c$ or the procedure described in Sections 5.5.1 and 5.5.2 if $g \in \mathcal{S}_1$ with the design matrix $\left(\boldsymbol{X}\,\hat{\boldsymbol{C}}^{(r)}\right)$.

(2) Obtain the q-values $\left\{q_{g,j}\right\}_{g\in\mathcal{S}^c\cup\mathcal{S}_1}$ *using the P values* $\left\{p_{g,j}\right\}_{g\in\mathcal{S}^c\cup\mathcal{S}_1}$ *for each* $j\in[d]$. *Define* $\hat{\boldsymbol{\Omega}}^{(r+1)}=\left(\hat{\boldsymbol{\Omega}}_1^{(r+1)}\cdots\hat{\boldsymbol{\Omega}}_d^{(r+1)}\right)^{\mathrm{T}}$ *to be*

$$\hat{\boldsymbol{\Omega}}_j^{(r+1)}=\left\{\sum_{g\in\mathcal{S}^c\cup\mathcal{S}_1} I\left(q_{g,j}>\epsilon_{\text{q-value}}\right)\left[\hat{\boldsymbol{\nu}}_g\right]_{jj}^{-1}\hat{\boldsymbol{\ell}}_g\hat{\boldsymbol{\ell}}_g^{\mathrm{T}}\right\}^{-1}$$
$$\times\left\{\sum_{g\in\mathcal{S}^c\cup\mathcal{S}_1} I\left(q_{g,j}>\epsilon_{\text{q-value}}\right)\left[\hat{\boldsymbol{\nu}}_g\right]_{jj}^{-1}\left[\hat{\tilde{\boldsymbol{\beta}}}_g\right]_j\hat{\boldsymbol{\ell}}_g\right\}.$$

Update $r\leftarrow r+1$ *and define* $\hat{\boldsymbol{C}}^{(r)}=\boldsymbol{X}\hat{\boldsymbol{\Omega}}^{(r)}+\hat{\boldsymbol{C}}_2$.

(3) Repeat Steps (1) and (2) for $r=0,1,\dots,R-1$ *and return* $\hat{\boldsymbol{C}}=\hat{\boldsymbol{C}}^{(R)}$.

Our software's default is $\epsilon_{\text{q-value}}=0.1$ and $R=3$. The estimator in Step (0) is similar to that used in Lee et al. (2017), Wang et al. (2017), McKennan & Nicolae (2018*a*). We refine the estimate for $\boldsymbol{\Omega}$ in Step (2) by attempting to remove metabolites with non-zero coefficients of interest $\boldsymbol{\beta}_g$. This helps alleviate the impact of outliers in the regression estimate for $\boldsymbol{\Omega}$. We subsequently use the estimated design matrix $\left(\boldsymbol{X}\,\hat{\boldsymbol{C}}\right)$ to estimate and perform inference on the coefficients of interest using techniques described in Section 5.5.

5.7 A simulation study

5.7.1 Simulation setup

Here we analyze simulated metabolomic data to compare the performance of our method with other existing methods. We simulated the log-intensities of $p=1,200$ metabolites in $n=600$ individuals, 300 of which were cases and the remaining 300 were controls. The observed design matrix was therefore $\boldsymbol{X}=(\boldsymbol{1}_n\,\boldsymbol{X}_{\text{int}})$, where $\boldsymbol{X}_{\text{int}}=\left(\boldsymbol{1}_{n/2},\boldsymbol{0}_{n/2}\right)^{\mathrm{T}}\in\mathbb{R}^n$. The parameters p and n were chosen to match those from our real data examples in Section 2.4, and we include additional simulation results when $n=100$ and $n=300$ in Section

5.10.1. We set $K = 10$, and for some constant a, simulated data as

$$\log\left(\alpha_{*_g}\right) \sim N_1\left(\mu_\alpha, 0.4^2\right) \quad (g = 1, \ldots, p) \tag{5.22a}$$

$$\delta_{*_g} \sim N_1\left(16, 1.2^2\right) \quad (g = 1, \ldots, p) \tag{5.22b}$$

$$\mu_{*_g} \sim N_1\left(18, 5^2\right) \quad (g = 1, \ldots, p) \tag{5.22c}$$

$$\beta_{*_g} \sim 0.8\delta_0 + 0.2N_1\left(0, 0.4^2\right) \quad (g = 1, \ldots, p) \tag{5.22d}$$

$$\boldsymbol{C} \sim MN_{n\times K}\left\{\left(a\boldsymbol{X}_{\text{int}}\, \boldsymbol{0}_n \cdots \boldsymbol{0}_n\right), I_n, I_K\right\} \tag{5.22e}$$

$$\left[\boldsymbol{\ell}_{*_g}\right]_k \sim \pi_k\delta_0 + (1 - \pi_k)\, N_1\left(0, \tau_k^2\right) \quad (g = 1, \ldots, p; k = 1, \ldots, K) \tag{5.22f}$$

$$\sigma_{*_g}^2 \sim \text{Gamma}\left(0.2^{-2}, 0.2^{-2}\right) \quad (g = 1, \ldots, p) \tag{5.22g}$$

$$y_{gi} \sim N_1\left(\mu_{*_g} + \left[\boldsymbol{X}_{\text{int}}\right]_i \beta_{*_g} + \boldsymbol{c}_i^{\mathrm{T}}\boldsymbol{\ell}_{*_g}, \sigma_{*_g}^2\right) \quad (g = 1, \ldots, p; i = 1, \ldots, n) \tag{5.22h}$$

$$r_{gi} = \text{Bernoulli}\left[\Psi\left\{\alpha_{*_g}\left(y_{gi} - \delta_{*_g}\right)\right\}\right] \quad (g = 1, \ldots, p; i = 1, \ldots, n), \tag{5.22i}$$

where δ_0 is the point mass at 0 and μ_α in (5.22a) was set so that if Z has cumulative distribution function $\Psi\left\{\exp\left(\mu_\alpha\right)x\right\}$, $\mathbb{V}(Z) = 1$. The constant a in (5.22e) was chosen so that $\boldsymbol{C}$ explained 7.5% of the variance in $\boldsymbol{X}_{\text{int}}$ on average across all simulations, and Table 5.1 contains the values of π_k and τ_k^2. These were chosen so that the non-zero eigenvalues $\lambda_1, \ldots, \lambda_K$ of

$$\mathcal{I} = (n-1)^{-1} P_{1_n}^{\perp} \boldsymbol{C}\left(p^{-1}\sum_{g=1}^{p} \sigma_{*_g}^{-2}\boldsymbol{\ell}_{*_g}\boldsymbol{\ell}_{*_g}^{\mathrm{T}}\right)\boldsymbol{C}^{\mathrm{T}} P_{1_n}^{\perp}$$

were 0.61, 0.33, 0.19, 0.14, 0.12, 0.08, 0.07, 0.05, 0.05 and 0.05 on average across all simulated datasets, since these were the first 10 eigenvalues of the estimated $\mathcal{I}$ in our data example in Section 5.8. Similarly, the prior variances for the missingness mechanism parameters in (5.22a) and (5.22b), as well as the mean and variance for the global mean in (5.22c), were set to their estimated equivalents from the first data example in Section 5.8. The distribution of missing data when Ψ was the logistic function is given in Figure 5.1, which is similar to

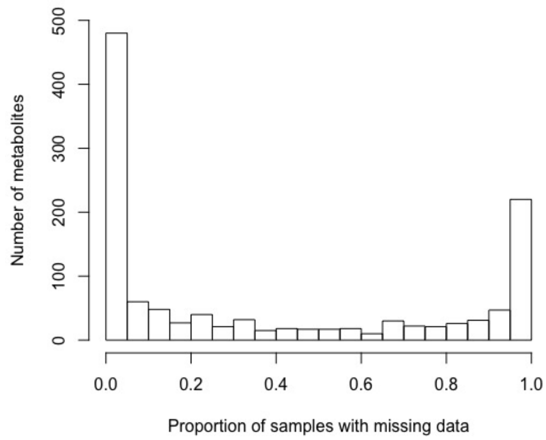


Figure 5.1: The fraction of metabolite-specific missing data $n^{-1} \sum_{i=1}^{n} \left(1 - r_{gi}\right)$, $g \in [p]$, for one simulated dataset with $\Psi(x) = \exp(x) / \left\{1 + \exp(x)\right\}$.

that observed real data from Section 2.4.

Table 5.1: *The* π_k *and* τ_k *values used to simulate* $\boldsymbol{\ell}_{*1}, \ldots, \boldsymbol{\ell}_{*p}$ $(k = 1, \ldots, 10)$.

Factor number (k)	1	2	3	4	5	6	7	8	9	10
π_k	0	0	0.76	0.56	0.48	0.32	0.28	0.20	0.20	0.20
τ_k	0.78	0.57	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5

We used our method detailed in Algorithm 5.1 with $\epsilon = 0.05$, $\boldsymbol{Z} = \mathbf{1}_n$ and $\hat{K} = 5$ to estimate the metabolite-dependent missingness mechanism parameters α_{*g}, δ_{*g}, the latent covariates $\boldsymbol{C}$ and the effects of interest $\beta_{*1}, \ldots, \beta_{*p}$. In each simulation, we removed all metabolites that were missing in more than 50% of the samples, since we find that these metabolites tend to have large J-statistics in real data (see Section 5.4.3). Similar to our real data examples in Section 5.8, the number of potential instruments, $\hat{K}$, was such that at least 90% of all metabolites $g \in \mathcal{S}$ had at least two q-values $q_{g,j}$, defined in Step (b) of Algorithm 5.1, that were less than 0.05 in each simulated dataset. All results were nearly identical when we let $\hat{K}$ be as small as 3 and as large as 10.

Given the complexity of our method, we evaluated its ability to recover α_{*g}, δ_{*g} in Section

5.7.2, as well as estimate and perform inference on the coefficients of interest in Section 5.7.3. This allows one to better understand the sources of variation in our estimators, and also highlights the novel aspects of our methodology.

5.7.2 Accuracy of our estimates for α_{*_g} and δ_{*_g}

We simulated and analyzed 20 datasets according to (5.22) with $\Psi(x) = F_4(x)$ to compare our estimators for α_{*_g} and δ_{*_g}, defined in (5.12a) using empirical Bayes to estimate the prior variance $\boldsymbol{U}$, to the those proposed in Wang et al. (2014). The latter is simply $\hat{\boldsymbol{\theta}}_g$ defined in Section 5.4.2, which is the two-step estimator from (5.4) with instruments $\boldsymbol{A}$ and function f given by (5.7). We could not compare it to the estimators proposed in Chen et al. (2017) or Hedeker et al. (2018), because they assume the missingness mechanisms are the same for each analyte. We remark that the estimators for α_{*_g} and δ_{*_g} these authors proposed in the aforementioned articles rely on the assumption that y_{gi} is normally distributed. We also did not compare our estimators to those proposed in Ai et al. (2018), since their method is nearly identical to that proposed in Wang et al. (2014).

The results are given in Figure 5.2. Our Bayesian estimator (HB-GMM) outperforms the standard two-step generalized method of moments estimator proposed in Wang et al. (2014) (GMM), which illustrates the advantages of pooling information across metabolites.

5.7.3 Inference on the simulated parameters of interest

We simulated 60 datasets according to (5.22) with $\Psi(x) = \exp(x)/\left\{1 + \exp(x)\right\}$ and analyzed them by first using HB-GMM to estimate the posterior means and variances for the sample weights given in (5.12b), and estimated $\boldsymbol{C}$ according to (5.21). We then used the estimated design matrix $\left(\boldsymbol{X}\,\hat{\boldsymbol{C}}\right)$ to calculate $\hat{\beta}_{*_g}$, an estimate for β_{*_g}, and $\hat{\mathbb{V}}\left(\hat{\beta}_{*_g}\right)$, an estimate for $\mathbb{V}\left(\hat{\beta}_{*_g}\right)$, with ordinary least squares if $g \in \mathcal{S}^c$ or the estimators defined in (5.16) and (5.17) if $g \in \mathcal{S}$. We subsequently formed 95% confidence intervals for β_{*_g} and computed P values for the null hypothesis $H_{0,g} : \beta_{*_g} = 0$ assuming $\hat{\beta}_{*_g} \sim N_1\left\{\beta_{*_g}, \hat{\mathbb{V}}\left(\hat{\beta}_{*_g}\right)\right\}$. Since

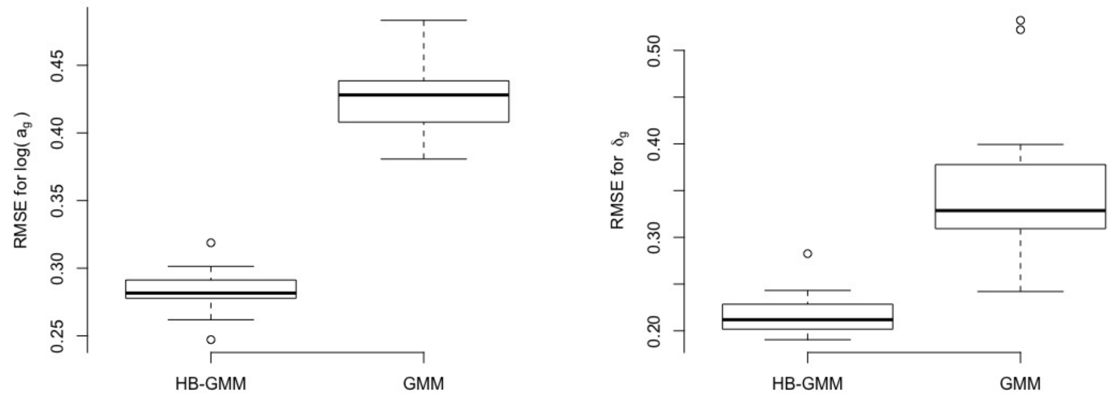


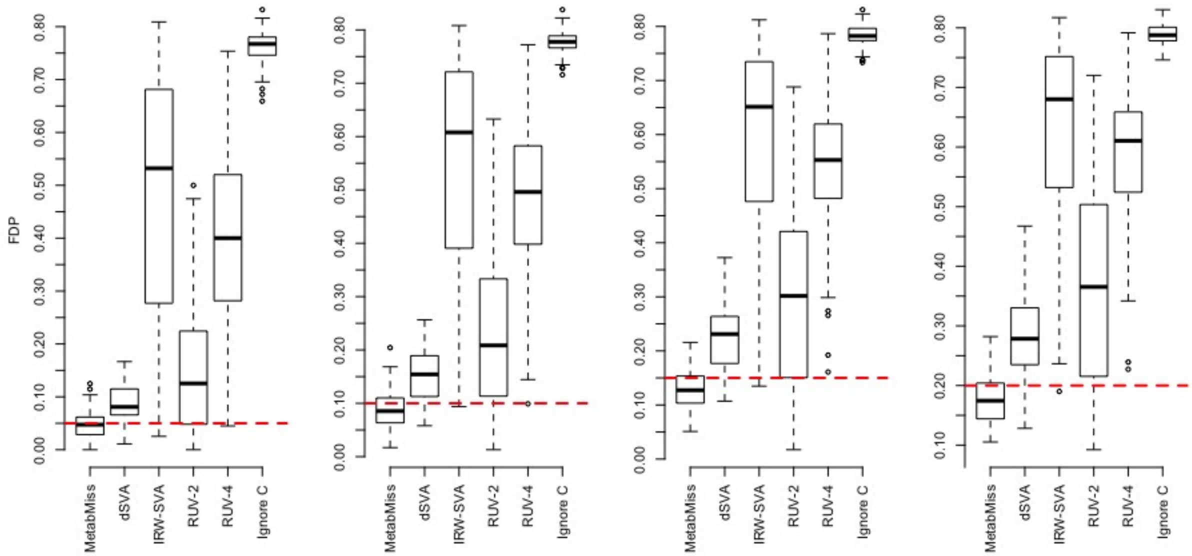
Figure 5.2: The root mean squared error (RMSE) $\left[|\mathcal{S}|^{-1} \sum_{g \in \mathcal{S}} \left\{\log\left(\hat{\alpha}_g\right) - \log\left(\alpha_{*g}\right)\right\}^2\right]^{1/2}$ (left) and $\left\{|\mathcal{S}|^{-1} \sum_{g \in \mathcal{S}} \left(\hat{\delta}_g - \delta_{*g}\right)^2\right\}^{1/2}$ (right) over 20 simulations. The estimates $\log\left(\hat{\alpha}_g\right)$ and $\hat{\delta}_g$ were either our Bayesian estimators proposed in Section 5.4.4 (HB-GMM), or the two-step estimator proposed in Wang et al. (2014) (GMM).

we typically do not know the exact functional form of $\Psi(x)$ in practice, all analyses were done assuming $\Psi(x) = F_4(x)$. We refer to this procedure as "MetabMiss" for the remainder of the chapter.

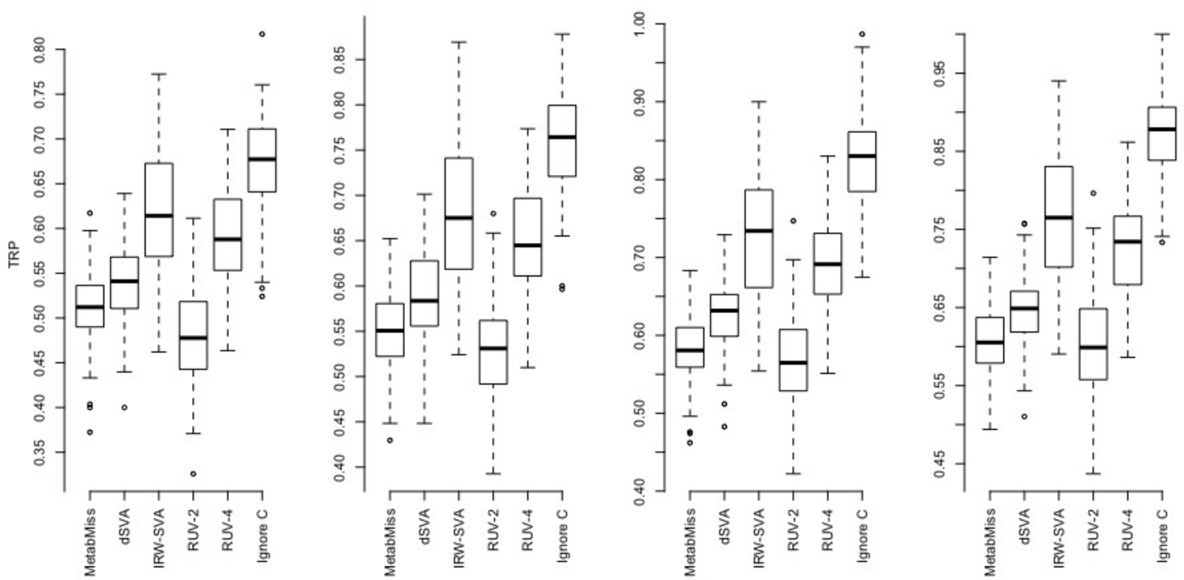
As far as we are aware, there are currently no methods designed to analyze these simulated data. In fact, the current state of the art to account for latent covariates in untargeted metabolomic data is to simply impute missing data with an arbitrarily chosen detection limit (Wehrens et al. 2016, Salerno et al. 2017), which is clearly sub-optimal. Therefore, we could only compare our method to those that account for non-random missing data or $\boldsymbol{C}$, but not both. We did not compare MetabMiss to methods that only account for the former, like those proposed in Chen et al. (2017), Hedeker et al. (2018), O'Brien et al. (2018), because we showed in Chapter 2 that estimators for β_{*g} that ignore $\boldsymbol{C}$ tend to be biased. Instead, we compared MetabMiss to existing methods to estimate $\boldsymbol{C}$ but do not account for the non-random missing data, which included Leek & Storey (2008) (IRW-SVA), Lee et al.

(2017) (dSVA), Gagnon-Bartsch & Speed (2012) (RUV-2) and Gagnon-Bartsch et al. (2013) (RUV-4). We do not report results from the method proposed in Wang et al. (2017), as it performed nearly identically to dSVA in all simulation scenarios. Since none of the of the aforementioned methods can accommodate missing data, we estimated $\boldsymbol{C}$ with each using only metabolites with complete observations, and subsequently estimated and computed confidence intervals for β_{*_g} and computed P values for the null hypothesis $H_{0,g} : \beta_{*_g} = 0$ with ordinary least squares using the estimated design matrix $\left(\boldsymbol{X}\, \hat{\boldsymbol{C}}\right)$, assuming the missing data were missing completely at random. RUV-like methods, which are frequently used when analyzing metabolomic data (De Livera et al. 2012, 2015, Wehrens et al. 2016), assume the practitioner has access to metabolites whose effect of interest $\beta_{*_g} = 0$. We therefore randomly chose 8% of all metabolites with $\beta_{*_g} = 0$ and no missing data to act as control metabolites when using RUV-2 and RUV-4, which is greater than the number typically used in applications (De Livera et al. 2012, 2015, Wehrens et al. 2016). We lastly remark that we could not analyze these simulated data with the methods proposed in De Livera et al. (2012) or Salerno et al. (2017) because both methods rely on a random effects model whose implemented estimators are not amenable to any missing data.

We first evaluated each method's ability to identify metabolites with non-zero effect of interest β_{*_g} while controlling the false discovery rate at a nominal level. The results are given in Figure 5.3, where the only method to suitably control for false discoveries is MetabMiss. The fact that MetabMiss is slightly underpowered compared to the other methods is to be expected, as anti-conservative inference is often more powerful. Lastly, we evaluated the confidence interval coverage for the effects of interest β_{*_g} for each method in Figure 5.4. These results illustrate the consequences of performing inference on estimators that do not properly account for the missing data, and also highlight the fidelity of our finite sample-corrected estimator for the variance given by (5.17).



(a)



(b)

Figure 5.3: From left to right: the false discovery proportion, FDP (a), and true recovery proportion, TRP (b), for metabolites with q-values $\leq$ 0.05, 0.1, 0.15 and 0.2. TRP is as defined in Figure 3.4 from Chapter 3 and q-values were determined using the qvalue package in R (Storey et al. 2015).

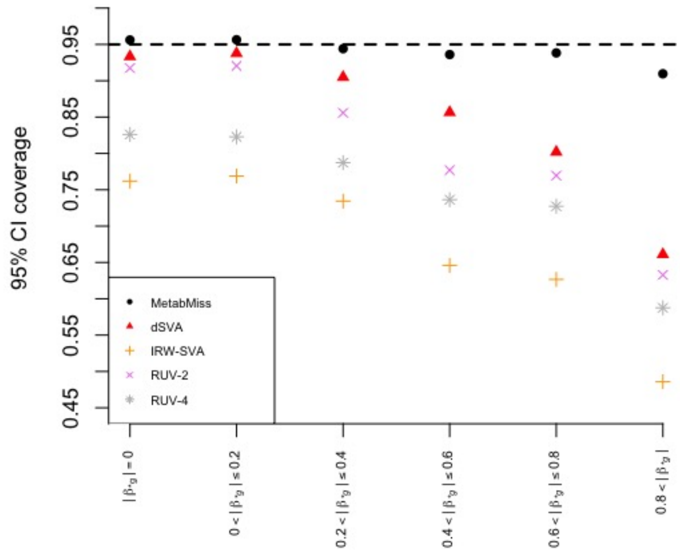


Figure 5.4: The fraction of effects of interest $\left\{\beta_{*g}\right\}_{g\in\mathcal{S}}$ in all 60 simulated datasets that lie in their respective 95% confidence intervals $\hat{\beta}_g \pm 1.96\left\{\hat{\mathbb{V}}\left(\hat{\beta}_g\right)\right\}^{1/2}$, stratified by $|\beta_{*g}|$. For MetabMiss, $\hat{\beta}_g$ and $\hat{\mathbb{V}}\left(\hat{\beta}_g\right)$ were calculated using (5.16) and (5.17). All other methods' estimates were derived using ordinary least squares and ignored the missingness mechanism. We did not include the results when $\boldsymbol{C}$ was ignored, as the coverage was uniformly less than IRW-SVA's.

5.8 Data analysis

We used blood plasma metabolomic data measured at 6 months and 6 years old in $n = 661$ Danish children enrolled in the COPSAC cohort (Bisgaard et al. 2013) to demonstrate the importance of accounting for both missing data and unobserved covariates in untargeted metabolomic data. Table 5.2 provides an overview of the extent of the missing data in each of the $p = 1138$ measured metabolites. Since the metabolites at 6 months and 6 years were quantified via mass spectrometry in two separate batches, we estimated the missingness mechanisms at 6 months and 6 years separately, since missingness patterns tend to depend on the mass spectrometer's performance (Do et al. 2018). We excluded metabolites that were missing in more than 50% of the samples and set $\boldsymbol{Z} = \boldsymbol{1}_n$, $\epsilon = 0.05$

and $\hat{K} = \min\left\{\hat{K}^{(PA)}_{6\text{ month}}, \hat{K}^{(PA)}_{6\text{ years}}\right\}/2 = 10$ when estimating the missingness mechanisms with HB-GMM in both analyses, where $\hat{K}^{(PA)}_{6\text{ month}}$ and $\hat{K}^{(PA)}_{6\text{ years}}$ were parallel analysis' (Buja & Eyuboglu 1992) estimates for K in the 6 month and 6 year datasets. With this value of $\hat{K}$, at least 90% of all metabolites $g \in \mathcal{S}$ in both datasets had at least two q-values $q_{g,j}$, defined in Step (b) of Algorithm 5.1, less than 0.05.

Table 5.2: *The number of metabolites in each missing data bin in the six month and 6 year datasets. Metabolites with more than 50% missing data were excluded from both analyses.*

Data \ Frequency of missing data, f	$f = 0$	$0 < f \leq 0.05$	$0.05 < f \leq 0.5$	$0.5 < f$
Six month	399	257	249	233
6 year	400	256	300	182

Once we estimated the missingness mechanisms, we could easily assess the relationships between the quantified metabolome at six months and 6 years and the many recorded phenotypes using MetabMiss. We were particularly interested in phenotypes related to lung function, and present the results for specific airway resistance (sR_{AW}) and forced vital capacity (FVC) measured at age 6. The former is a measure of airway resistance (Kaminsky 2012) and the latter is the total volume of air exhaled during a forced breath. While these two quantitative traits are ostensibly manifestations of a similar underlying phenomenon, they were completely uncorrelated in these children (P value $= 0.31$). For each of the four analyses (2 metabolomes $\times$ 2 phenotypes), we estimated K with parallel analysis (Leek et al. 2017) and subsequently regressed the quantified metabolites at 6 months and 6 years onto $(\mathbf{1}_n\, \boldsymbol{X}_{\text{int}})$ while accounting for $\boldsymbol{C}$ using MetabMiss. The covariate $\boldsymbol{X}_{\text{int}} \in \mathbb{R}^{n \times 1}$ was either sR_{AW} or FVC. The results of two of the analyses are given in Figures 5.5 and 5.6. The remaining two analyses, like that which is shown in Figure 5.6, did not reveal any metabolome-wide significant results.

As one can see in Figure 5.6, MetabMiss corrects the apparent P value inflation, which is commonly observed in practice when $\boldsymbol{C}$ is ignored (Leek & Storey 2007, Efron 2010).

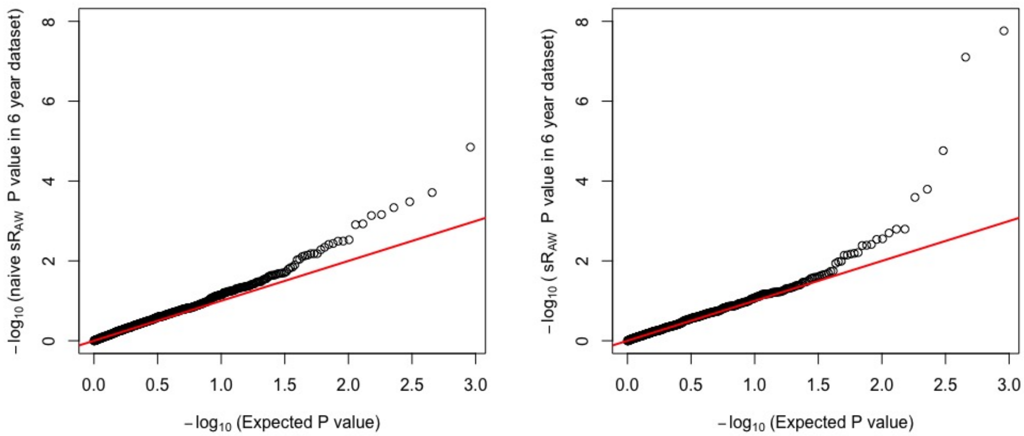


Figure 5.5: P values for the null hypothesis that the effect of $\mathrm{sR_{AW}}$ on a metabolite's intensity at age 6 is 0 when the latent covariates $\boldsymbol{C}$ and the missing data are ignored (left) and when they are both accounted for using MetabMiss (right). "Expected P value" is the expected ordered P value under the null hypothesis, assuming all test statistics are independent.

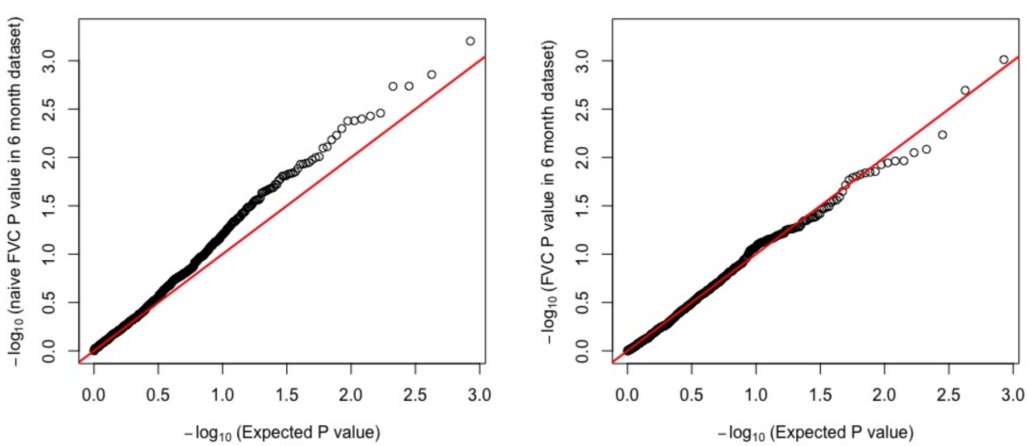


Figure 5.6: P values for the null hypothesis that the effect of FVC on a metabolite's intensity at 6 months is 0 when the latent covariates $\boldsymbol{C}$ and the missing data are ignored (left) and when they are both accounted for using MetabMiss (right).

Further, MetabMiss appears to not only correct the minor P value inflation in the $\mathrm{sR_{AW}}$ analysis in Figure 5.5, but also empowers the analysis by reducing the residual variance. We identified seven metabolites at a a q-value threshold of 0.2 in the $\mathrm{sR_{AW}}$ analysis: two sphingolipids, a benzoate derivative, pyruvate and three derivatives of piperine, which is an alkaloid found in black pepper. A reduction in sphingolipid synthesis tends to increase airway hyperractivity in children (Ono et al. 2015), which is congruent with the estimated sign of the $\mathrm{sR_{AW}}$ effect on the two sphingolipid metabolites' intensities at 6 years of age. Benzoate preservatives have been linked to asthma and lung function-related phenotypes (Balatsinou et al. 2004, Pacor et al. 2004), and pyruvate and lactate (q-value = 0.23) levels have previously been associated with asthma (Xu, Cui, Wang, Yin, Gao, Liu & Yang 2010, Ostroukhova et al. 2012).

The three derivatives of piperine are particularly interesting in the context of our methodology because all three had between 12% and 48% missing data with J-test P values between 0.77 and 0.99 (see Section 5.4.3), suggesting that Model (5.3) is a reasonable model for their missingness mechanisms. Piperine abundance has also been shown to be associated with asthma and asthma-related phenotypes (Kim & Lee 2009, Finkelstein et al. 2015), and one of the mechanisms by which it aids in alleviating airway hypersensitivity in mice has been identified (Reinke et al. 2017). Moreover, its affinity for the TRPV1 cation channel (Premkumar 2014) and the latter's role in respiratory illness in humans (Jia & Lee 2007) provides additional evidence that these three observed associations are in fact biologically meaningful.

5.9 Discussion

We have presented, to the best of our knowledge, the first method to account for both non-random missing data and latent covariates in untargeted metabolomic experiments. Our method leverages the fact that the majority of the variation in metabolomic data can be explained by a small number of latent covariates, which we use as instrumental variables in a generalized method of moments estimator to estimate the metabolite-dependent missingness mechanisms. We then present methodology to accurately and efficiently estimate the effects of interest in a linear model while accounting for both the non-random missing data and latent covariates. One of the most appealing aspects of our procedure is that one only needs to estimate the missingness mechanisms once per data set, since our inverse probability weighted estimators only depend on the estimated sample weights. This makes many-phenotype inference tractable, such as the analysis we presented in Section 5.8 or regressing the metabolome onto single nucleotide polymorphism data.

An important hyperparameter in our procedure is the number of instrumental variables used to estimate each metabolite's missingness mechanism. We currently use three instruments to estimate two parameters, which allows one to evaluate the veracity of Model (5.3) using the J-statistic. An important avenue of future research could be developing a data-driven approach to choose the appropriate number of instruments. Lastly, we remark that while our inverse probability weighted estimator is consistent and its estimated variance is accurate enough to perform valid inference, it could be made more efficient if one were willing to make assumptions about the distribution of missing data. This is currently an area that I am actively investigating.

5.10 Appendix

5.10.1 Additional simulation results

Here we analyze additional data simulated according to (5.22) with $\Psi(x) = \frac{\exp(x)}{1+\exp(x)}$ and $n = 300$ or $n = 100$ to demonstrate our method's performance on data with smaller sample sizes. Just like we did in Section 5.7.3, we analyzed each of the 60 simulated datasets for each value of n with MetabMiss by making the incorrect assumption that $\Psi(x) = F_4(x)$. For simplicity of presentation, we only report each method's potential to estimate and perform inference on β_{*_g} for $g \in \mathcal{S}$. The results for each set of simulations are given in Figures 5.7 and 5.8.

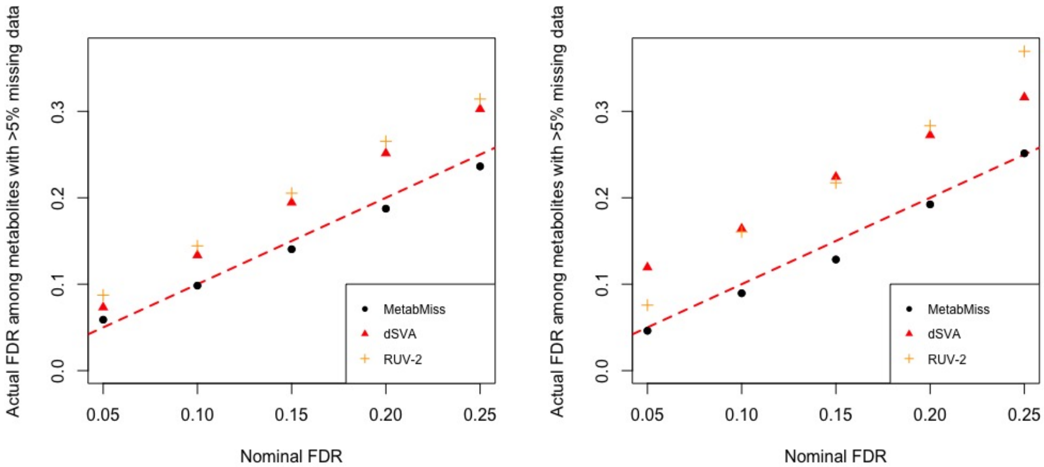


Figure 5.7: The false discovery rate (FDR) among rejected metabolites with $>5\%$ but $\leq 50\%$ missing data as a function of q-value threshold (Nominal FDR) when $n = 300$ (left) and when $n = 100$ (right). IRW-SVA's and RUV-4's performance was uniformly worse that dSVA's in both simulation settings. The dashed red line is the line $y = x$.

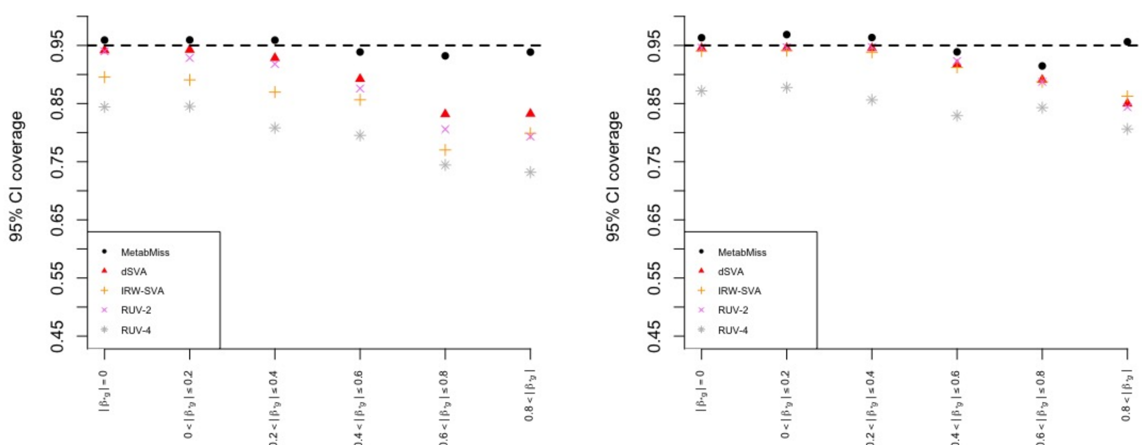


Figure 5.8: The 95% confidence interval coverage when $n = 300$ (left) and when $n = 100$ (right). These plots are analogous to Figure 5.4.

5.10.2 *Justification of the estimate for* $\mathbb{V}\left(\hat{\boldsymbol{\beta}}_g\right)$ *defined in* (5.17)

Here were perform an error analysis to justify using (5.17) to estimate $\mathbb{V}\left(\hat{\boldsymbol{\beta}}_g\right)$. For notational convenience, we assume all covariates are observed and contained in $\boldsymbol{X} = (\boldsymbol{x}_1 \cdots \boldsymbol{x}_n)^{\mathrm{T}} \in$

$\mathbb{R}^{n\times d}$, where μ_{gi} is as defined in (5.13). We also define

$$
\begin{aligned}
w_{gi} &= w_{gi}\left(\boldsymbol{\eta}_{*g}\right) \quad (g \in \mathcal{S}; i \in [n]) \\
\boldsymbol{s}_{gi} &= \boldsymbol{s}_{gi}\left(\beta_{*g}\right) = \boldsymbol{x}_i\left(y_{gi} - \boldsymbol{x}_i\boldsymbol{\beta}_{*g}\right) \quad (g \in \mathcal{S}; i \in [n]) \\
\hat{\boldsymbol{s}}_{gi} &= \boldsymbol{x}_i\left(y_{gi} - \boldsymbol{x}_i\hat{\boldsymbol{\beta}}_g\right) \quad (g \in \mathcal{S}; i \in [n])
\end{aligned}
$$

where $\hat{\boldsymbol{\beta}}_g$ is defined in (5.16). We assume throughout this section that w_{gi} is known. Unless otherwise stated, all expectations and variances are taken conditional on $\boldsymbol{X}$. By the derivation in Section 5.5.2, our goal is to approximate $\gamma_{gi}^2\mathbb{E}\left(r_{gi}w_{gi}^2\boldsymbol{s}_{gi}\boldsymbol{s}_{gi}^{\mathrm{T}} \mid \boldsymbol{X}\right)$ for fixed $g \in \mathcal{S}$ and all $i \in [n]$.

Fix a $g \in \mathcal{S}$ and define $\xi_{gi} = r_{gi}\gamma_{gi}w_{gi}$ and $\boldsymbol{W}_g = \operatorname{diag}\left(\xi_{g1}, \ldots, \xi_{gn}\right)$. When γ_{gi} and w_{gi} are known, the estimator $\hat{\boldsymbol{\beta}}_g$ is such that

$$
\begin{aligned}
\sum_{i=1}^{n} \xi_{gi}\boldsymbol{s}_{gi} &= \boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_g\left(\boldsymbol{y}_g - \boldsymbol{X}\boldsymbol{\beta}_{*g}\right) = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_g\left(\boldsymbol{y}_g - \boldsymbol{X}\hat{\boldsymbol{\beta}}_g\right) + \boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_g\boldsymbol{X}\left(\hat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_{*g}\right) \\
&= \mathbf{0} + \boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_g\boldsymbol{X}\left(\hat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_{*g}\right),
\end{aligned}
$$

where the last equality follows by the definition of $\hat{\boldsymbol{\beta}}_g$. Therefore, for all $i \in [n]$,

$$
\begin{aligned}
\hat{\boldsymbol{s}}_{gi} &= \boldsymbol{x}_i\left(y_{gi} - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_{*g}\right) + \boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\left(\boldsymbol{\beta}_{*g} - \hat{\boldsymbol{\beta}}_g\right) = \boldsymbol{s}_{gi} - \boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_g\boldsymbol{X}\right)^{-1}\sum_{j=1}^{n}\xi_{gj}\boldsymbol{s}_{gj} \\
&= \left\{I_d - \xi_{gi}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_g\boldsymbol{X}\right)^{-1}\right\}\boldsymbol{s}_{gi} - \boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_g\boldsymbol{X}\right)^{-1}\sum_{j\neq i}\xi_{gj}\boldsymbol{s}_{gj} \\
&= \boldsymbol{x}_i\left\{1 - \xi_{gi}\boldsymbol{x}_i^{\mathrm{T}}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_g\boldsymbol{X}\right)^{-1}\boldsymbol{x}_i\right\}e_{gi} - \boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_g\boldsymbol{X}\right)^{-1}\sum_{j\neq i}\xi_{gj}\boldsymbol{s}_{gj} \\
&= \left(1 - h_{gi}\right)\boldsymbol{s}_{gi} - \boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_g\boldsymbol{X}\right)^{-1}\sum_{j\neq i}\xi_{gj}\boldsymbol{s}_{gj}
\end{aligned}
$$

where h_{gi} is the ith leverage score of $\boldsymbol{W}_g^{1/2}\boldsymbol{X}$. For $\boldsymbol{A}_i = \sum_{j\neq i}\xi_{gj}\boldsymbol{s}_{gj}$, the naive plug-in estimator can be expressed as

$$
\begin{aligned}
r_{gi}\gamma_{gi}^{2}w_{gi}^{2}\hat{\boldsymbol{s}}_{gi}\hat{\boldsymbol{s}}_{gi}^{\mathrm{T}} =& \xi_{gi}^{2}\left(1-h_{gi}\right)^{2}\boldsymbol{s}_{gi}\boldsymbol{s}_{gi}^{\mathrm{T}} - \left(1-h_{gi}\right)\boldsymbol{x}_{i}\boldsymbol{x}_{i}^{\mathrm{T}}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_{g}\boldsymbol{X}\right)^{-1}\boldsymbol{A}_{i}\xi_{gi}^{2}\boldsymbol{s}_{gi}^{\mathrm{T}} \\
&- \left\{\left(1-h_{gi}\right)\boldsymbol{x}_{i}\boldsymbol{x}_{i}^{\mathrm{T}}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_{g}\boldsymbol{X}\right)^{-1}\boldsymbol{A}_{i}\xi_{gi}^{2}\boldsymbol{s}_{gi}^{\mathrm{T}}\right\}^{\mathrm{T}} \\
&+ \xi_{gi}^{2}\left(1-h_{gi}\right)^{2}\boldsymbol{x}_{i}\boldsymbol{x}_{i}^{\mathrm{T}}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_{g}\boldsymbol{X}\right)^{-1}\boldsymbol{A}_{i}\boldsymbol{A}_{i}^{\mathrm{T}}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_{g}\boldsymbol{X}\right)^{-1}\boldsymbol{x}_{i}\boldsymbol{x}_{i}^{\mathrm{T}}, \quad (5.23)
\end{aligned}
$$

and the corrected estimator as

$$
\begin{aligned}
\left(1-h_{gi}\right)^{-2}r_{gi}\gamma_{gi}^{2}w_{gi}^{2}\hat{\boldsymbol{s}}_{gi}\hat{\boldsymbol{s}}_{gi}^{\mathrm{T}} =& \xi_{gi}^{2}\boldsymbol{s}_{gi}\boldsymbol{s}_{gi}^{\mathrm{T}} - \left(1-h_{gi}\right)^{-1}\boldsymbol{x}_{i}\boldsymbol{x}_{i}^{\mathrm{T}}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_{g}\boldsymbol{X}\right)^{-1}\boldsymbol{A}_{i}\xi_{gi}^{2}\boldsymbol{s}_{gi}^{\mathrm{T}} \\
&- \left\{\left(1-h_{gi}\right)^{-1}\boldsymbol{x}_{i}\boldsymbol{x}_{i}^{\mathrm{T}}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_{g}\boldsymbol{X}\right)^{-1}\boldsymbol{A}_{i}\xi_{gi}^{2}\boldsymbol{s}_{gi}^{\mathrm{T}}\right\}^{\mathrm{T}} \\
&+ \xi_{gi}^{2}\boldsymbol{x}_{i}\boldsymbol{x}_{i}^{\mathrm{T}}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_{g}\boldsymbol{X}\right)^{-1}\boldsymbol{A}_{i}\boldsymbol{A}_{i}^{\mathrm{T}}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_{g}\boldsymbol{X}\right)^{-1}\boldsymbol{x}_{i}\boldsymbol{x}_{i}^{\mathrm{T}}. \quad (5.24)
\end{aligned}
$$

We first note that $\boldsymbol{A}_i$ is independent of r_{gi} and y_{gi} and $\mathbb{E}\left(\boldsymbol{A}_i\right) = \boldsymbol{0}$. Therefore, if we ignore the uncertainty in $n^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_g\boldsymbol{X}$, the second and third terms in (5.23) and (5.24) will have expectation $\boldsymbol{0}$. The last terms are positive semi-definite, where since $\left(r_{g1}, y_{g1}\right), \ldots, \left(r_{gn}, y_{gn}\right)$ are independent, $\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_g\boldsymbol{X}\right)^{-1}\boldsymbol{A}_i\boldsymbol{A}_i\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_g\boldsymbol{X}\right)^{-1}$ will have eigenvalues that are $O_P\left(n^{-1}\right)$ under suitable regularity conditions. However, the first term in (5.23) will tend to be small and therefore underestimate $\gamma_{gi}^2\mathbb{E}\left(r_{gi}w_{gi}^2\boldsymbol{s}_{gi}\boldsymbol{s}_{gi}^{\mathrm{T}}\right)$, especially when the weights w_{gi} or d are large, since this will increase leverage scores. Further, unlike data whose missing values are missing at random, the leverage scores are correlated with $\boldsymbol{s}_{gi}$, where large values of $\boldsymbol{s}_{gi}$ typically imply the weights will be large. These two facts cause the naive plug-in sandwich estimator to underestimate the variance. The corrected estimator circumvents these issues because the first term in (5.24) is exactly the term whose expectation we are attempting to estimate.

5.10.3 Justification of the estimate for $\sigma_{*_g}^2$ defined in (5.18)

The estimator for $\sigma_{*_g}^2$ defined in (5.18) is a similar finite sample correction as that presented in Section 5.10.2 applied to

$$\operatorname{tr}\left(\hat{\boldsymbol{W}}_g\right)^{-1} \sum_{i=1}^n r_{gi} \hat{w}_{gi} \hat{\gamma}_{gi} \left(y_{gi} - \boldsymbol{x}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}_g\right)^2 \quad (g \in \mathcal{S}).$$

The details have been ommitted.

5.10.4 Estimating $\boldsymbol{C}$ with nuisance covariates

Here we extend our estimator for $\boldsymbol{C}$ defined in (5.21) of Section 5.6 when $\boldsymbol{X} = (\boldsymbol{X}_{\text{int}}\, \boldsymbol{X}_{\text{nuis}})$, where $\boldsymbol{X}_{\text{int}} \in \mathbb{R}^{n \times d_{\text{int}}}$ and $\boldsymbol{X}_{\text{nuis}} \in \mathbb{R}^{n \times d_{\text{nuis}}}$. To do so, let

$$\boldsymbol{\beta}_{*_g} = \begin{pmatrix} \boldsymbol{\beta}_{*_g}^{(\text{int})} \\ \boldsymbol{\beta}_{*_g}^{(\text{nuis})} \end{pmatrix} \quad (g = 1, \ldots, p).$$

For all $g \in [p]$, we can then re-write (5.19) as

$$\begin{aligned} &\boldsymbol{y}_g = P_{X_{\text{nuis}}}^{\perp} \boldsymbol{X}_{\text{int}} \tilde{\boldsymbol{\beta}}_{*_g}^{(\text{int})} + \boldsymbol{X}_{\text{nuis}} \tilde{\boldsymbol{\beta}}_{*_g}^{(\text{nuis})} + \boldsymbol{C}_2 \boldsymbol{\ell}_{*_g} + \boldsymbol{e}_g, \quad \tilde{\boldsymbol{\beta}}_{*_g}^{(\text{int})} = \boldsymbol{\beta}_{*_g}^{(\text{int})} + \boldsymbol{\Omega} \boldsymbol{\ell}_{*_g} \\ &\boldsymbol{\Omega} = \left(\boldsymbol{X}_{\text{int}}^{\mathrm{T}} P_{X_{\text{nuis}}}^{\perp} \boldsymbol{X}_{\text{int}}\right)^{-1} \boldsymbol{X}_{\text{int}}^{\mathrm{T}} P_{X_{\text{nuis}}}^{\perp} \boldsymbol{C}, \quad \boldsymbol{C}_2 = P_X^{\perp} \boldsymbol{C} \end{aligned}$$

where $\tilde{\boldsymbol{\beta}}_{*_g}^{(\text{nuis})}$ is a nuisance parameter. We estimate $\boldsymbol{C}_2$ according to (5.20), where

$$\hat{\tilde{\boldsymbol{\beta}}}_g = \begin{pmatrix} \hat{\tilde{\boldsymbol{\beta}}}_g^{(\text{int})} \\ \hat{\tilde{\boldsymbol{\beta}}}_g^{(\text{nuis})} \end{pmatrix}$$

and $\hat{\tilde{\boldsymbol{\beta}}}_g^{(\text{int})} \in \mathbb{R}^{d_{\text{int}}}$, $\hat{\tilde{\boldsymbol{\beta}}}_g^{(\text{nuis})} \in \mathbb{R}^{d_{\text{nuis}}}$. Using the same reasoning as when $\boldsymbol{X} = \boldsymbol{X}_{\text{int}}$ in Section 5.6, we model $\hat{\tilde{\boldsymbol{\beta}}}_g^{(\text{int})} \dot{\sim} \left(\boldsymbol{\beta}_{*_g}^{(\text{int})} + \boldsymbol{\Omega} \boldsymbol{\ell}_{*_g}, \hat{\boldsymbol{v}}_g\right)$, where $\hat{\boldsymbol{v}}_g$ is a submatrix of the estimate for the variance defined in (5.17) when $g \in \mathcal{S}$ or the ordinary least squares estimator for $\mathbb{V}\left(\hat{\tilde{\boldsymbol{\beta}}}_g^{(\text{int})}\right)$

when $g \in \mathcal{S}^c$ using the design matrix $\left(\boldsymbol{X}_{\text{int}}\, \boldsymbol{X}_{\text{nuis}}\, \hat{\boldsymbol{C}}_2\right)$. We subsequently estimate $\boldsymbol{\Omega}$ with Algorithm 5.2.

5.10.5 A mathematical justification for Step (b) of Algorithm 5.1

In this section, we state and prove a few technical results that help justify the instrumental variable selection step in Step (b) of Algorithm 5.1. Our main result is Theorem 5.1 at the end of the section. We first state the assumptions that we will use throughout this section.

***Assumption* 5.1.** *Suppose $\boldsymbol{y}_g = \boldsymbol{Z}\boldsymbol{\xi}_{*_g} + \boldsymbol{C}\boldsymbol{\ell}_{*_g} + \boldsymbol{e}_g$ for all $g \in [p]$, where $\boldsymbol{Z} \in \mathbb{R}^{n \times t}$ and $\boldsymbol{C} \in \mathbb{R}^{n \times K}$ for some constants $t, K \geq 1$, and that Model (5.3) holds for $g \in \mathcal{S}$. Define $\boldsymbol{L}_{\mathcal{S}^c}$ to be the sub-matrix of $\boldsymbol{L} = \left(\boldsymbol{\ell}_1 \cdots \boldsymbol{\ell}_p\right)^{\mathrm{T}}$, restricted to the rows $g \in \mathcal{S}^c$, and let $\tilde{p} = |\mathcal{S}^c|$. Then following hold for some constant $c_1 > 1$:*

(i) The entries of $\boldsymbol{Z} \in \mathbb{R}^{n \times t}$ are non-random, uniformly bounded and $\lim_{n \to \infty} n^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z} = \boldsymbol{\Sigma}_Z \succeq c_1^{-1} I_t$. Additionally, the eigenvalues $\lambda_1 > \cdots > \lambda_K \geq c_1^{-1}$ of $\mathcal{I} = \tilde{p}^{-1}\boldsymbol{L}_{\mathcal{S}^c}^{\mathrm{T}}\boldsymbol{L}_{\mathcal{S}^c}$ are such that $\lambda_1/\lambda_K \leq c_1$ and $1 - \lambda_{k+1}/\lambda_k \geq c_1^{-1}$ for all $k \in [K]$, where $\lambda_{K+1} = 0$.

(ii) $\boldsymbol{C} = \boldsymbol{Z}\boldsymbol{A} + \boldsymbol{\Xi}$ for some non-random matrix $\boldsymbol{A} \in \mathbb{R}^{t \times K}$. The random matrix $\boldsymbol{\Xi}$ is independent of $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_p$, $[\boldsymbol{\Xi}]_{1*}, \ldots, [\boldsymbol{\Xi}]_{n*}$ are independent and identically distributed and $[\boldsymbol{\Xi}]_{1*} \dot{\sim} (\boldsymbol{0}_K, I_K)$. Further, the entries of $\boldsymbol{\Xi}$ have uniformly bounded sixth moments.

(iii) $\left[\boldsymbol{e}_g\right]_1, \ldots, \left[\boldsymbol{e}_g\right]_n \dot{\sim} \left(0, \sigma_g^2\right)$ are independent and identically distributed with uniformly bounded fourth moments for all $g \in [p]$. Further, $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_p$ are independent.

(iv) The function $\Psi(x) : \mathbb{R} \to (0, 1)$ defined in (5.3) is continuous and strictly increasing, where $\lim_{x \to -\infty} \Psi(x) = 0$ and $\lim_{x \to \infty} \Psi(x) = 1$. Further $\alpha_{*_g} > 0$ for all $g \in \mathcal{S}$.

(v) If $g \in \mathcal{S}^c$, the entries of $\boldsymbol{y}_g$ are missing completely at random. Further, $\tilde{p}/p, p/n \geq c_1^{-1}$.

***Lemma* 5.1.** *Fix a $g \in \mathcal{S}$. If Assumption 5.1 holds, then $\mathbb{P}\left(r_{gi} = 1 \mid [\boldsymbol{Z}]_{i*}\right) \geq c$ for all $i \in [n]$ for some constant $c > 0$ that does not depend on n.*

Proof. Define $\tilde{\boldsymbol{\xi}}_{*g} = \boldsymbol{\xi}_{*g} + \boldsymbol{A}\boldsymbol{\ell}_{*g}$. By Assumption 5.1, $\tilde{e}_{g1} = \left[\boldsymbol{y}_g\right]_1 - [\boldsymbol{Z}]_{1*}^{\mathrm{T}} \tilde{\boldsymbol{\xi}}_{*g}, \ldots, \tilde{e}_{gn} = \left[\boldsymbol{y}_g\right]_n - [\boldsymbol{Z}]_{n*}^{\mathrm{T}} \tilde{\boldsymbol{\xi}}_{*g}$ are identically distributed. Let $\mu_n = \min_{i\in[n]} \left([\boldsymbol{Z}]_{i*}^{\mathrm{T}} \tilde{\boldsymbol{\xi}}_{*g}\right)$. Note that $\mu_n \geq \mu$ for some constant $\mu > -\infty$ because the entries of $\boldsymbol{Z}$ are uniformly bounded. Then because $\Psi(x)$ is strictly increasing and non-zero, for any $n > 0$ and $i \in [n]$,

$$\begin{aligned}
\mathbb{P}\left(r_{gi} = 1 \mid [\boldsymbol{Z}]_{i*}\right) &= \mathbb{E}\left[\Psi\left\{\alpha_{*g}\left([\boldsymbol{Z}]_{i*}^{\mathrm{T}} \tilde{\boldsymbol{\xi}}_{*g} + \tilde{e}_{gi} - \delta_{*g}\right)\right\} \mid [\boldsymbol{Z}]_{i*}\right] \\
&\geq \mathbb{E}\left[\Psi\left\{\alpha_{*g}\left(\mu + \tilde{e}_{gi} - \delta_{*g}\right)\right\}\right] \\
&= \mathbb{E}\left[\Psi\left\{\alpha_{*g}\left(\mu + \tilde{e}_{g1} - \delta_{*g}\right)\right\}\right] > 0.
\end{aligned}$$

□

Lemma 5.2. *Suppose $\epsilon = 0$ and that Assumption 5.1 holds, and let $n^{-1/2}\hat{\boldsymbol{C}}_2$ be the first K right singular vectors of $\boldsymbol{Y}_{\mathcal{S}^c} P_Z^{\perp} = \boldsymbol{L}_{\mathcal{S}}\left(P_Z^{\perp}\boldsymbol{C}\right)^{\mathrm{T}} + \boldsymbol{E}_{\mathcal{S}^c} P_Z^{\perp}$. Define*

$$\boldsymbol{C}_2 = P_Z^{\perp}\boldsymbol{C}\left(n^{-1}\boldsymbol{C}^{\mathrm{T}} P_Z^{\perp}\boldsymbol{C}\right)^{-1/2}\boldsymbol{U}, \quad \tilde{\boldsymbol{L}}_{\mathcal{S}^c} = \boldsymbol{L}_{\mathcal{S}^c}\left(n^{-1}\boldsymbol{C}^{\mathrm{T}} P_Z^{\perp}\boldsymbol{C}\right)^{1/2}\boldsymbol{U}, \tag{5.25}$$

where $\boldsymbol{U}$ is a unitary matrix such that $\tilde{\boldsymbol{L}}_{\mathcal{S}^c}^{\mathrm{T}}\tilde{\boldsymbol{L}}_{\mathcal{S}^c}$ is diagonal with non-increasing elements. Then

$$\hat{\boldsymbol{C}}_2 = \boldsymbol{C}_2\hat{\boldsymbol{v}} + n^{1/2}\boldsymbol{Q}_{P_Z^{\perp}C}\hat{\boldsymbol{w}}, \tag{5.26}$$

where

$$\|\hat{\boldsymbol{v}} - I_K\|_2 = O_P\left\{(n\tilde{p})^{-1/2}\right\}, \quad \|\hat{\boldsymbol{w}}\|_2 = O_P\left(\tilde{p}^{-1/2}\right) \tag{5.27a}$$

$$\|[\hat{\boldsymbol{w}}]_{*k} - (n\lambda_k\tilde{p})^{-1}\boldsymbol{Q}_{P_Z^{\perp}C}^{\mathrm{T}} P_Z^{\perp}\boldsymbol{E}_{\mathcal{S}^c}^{\mathrm{T}}\left(n^{1/2}\boldsymbol{L}_{\mathcal{S}^c} + n^{-1/2}\boldsymbol{E}_{\mathcal{S}^c}P_Z^{\perp}\boldsymbol{C}_2\right)[\hat{\boldsymbol{v}}]_{*k}\|_2 = O_P\left\{(n\tilde{p})^{-1/2}\right\}. \tag{5.27b}$$

Proof. For notational simplicity, we drop the subscript $\mathcal{S}^c$ and redefine $p \leftarrow \tilde{p}$, $\lambda_k \leftarrow n\lambda_k$ for all $k \in [K]$ and $\boldsymbol{C}_2 \leftarrow n^{-1/2}\boldsymbol{Q}_Z^{\mathrm{T}}\boldsymbol{C}_2$. The latter implies $\boldsymbol{Y}\boldsymbol{Q}_Z^{\mathrm{T}} = n^{1/2}\tilde{\boldsymbol{L}}\boldsymbol{C}_2^{\mathrm{T}} + \boldsymbol{E}\boldsymbol{Q}_Z$,

$\boldsymbol{C}_2^{\mathrm{T}}\boldsymbol{C}_2 = I_K$ and $np^{-1}\tilde{\boldsymbol{L}}^{\mathrm{T}}\tilde{\boldsymbol{L}} = \operatorname{diag}(\gamma_1, \ldots, \gamma_K)$, where $\gamma_k = \lambda_k \{1 + o_P(1)\}$ by Assumption 5.1. Define $\boldsymbol{E}_1 = \boldsymbol{E}\boldsymbol{Q}_Z\tilde{\boldsymbol{C}}$, $\boldsymbol{Q} = \boldsymbol{Q}_{\tilde{C}_2}$, $\boldsymbol{E}_2 = \boldsymbol{E}\boldsymbol{Q}_Z\boldsymbol{Q}$ and $\boldsymbol{\Sigma} = \operatorname{diag}\left(\sigma_1^2, \ldots, \sigma_p^2\right)$. The proof is nearly identical to that of Theorem 3.10 from Chapter 3. Define

$$\begin{aligned}\boldsymbol{S} &= (p\lambda_K)^{-1} (\boldsymbol{C}_2\boldsymbol{Q})^{\mathrm{T}} \boldsymbol{Y}^{\mathrm{T}} P_Z^{\perp} \boldsymbol{Y} (\boldsymbol{C}_2\boldsymbol{Q}) \\ &= \begin{pmatrix} (p\lambda_K)^{-1}\left(n^{1/2}\tilde{\boldsymbol{L}} + \boldsymbol{E}_1\right)^{\mathrm{T}}\left(n^{1/2}\tilde{\boldsymbol{L}} + \boldsymbol{E}_1\right) & (p\lambda_K)^{-1}\left(n^{1/2}\tilde{\boldsymbol{L}} + \boldsymbol{E}_1\right)^{\mathrm{T}}\boldsymbol{E}_2 \\ (p\lambda_K)^{-1}\boldsymbol{E}_2^{\mathrm{T}}\left(n^{1/2}\tilde{\boldsymbol{L}} + \boldsymbol{E}_1\right) & (p\lambda_K)^{-1}\boldsymbol{E}_2^{\mathrm{T}}\boldsymbol{E}_2. \end{pmatrix}\end{aligned} \tag{5.28}$$

By Theorem 5.37 of Eldar & Kutyniok (2012), $\|(p\lambda_K)^{-1}\boldsymbol{E}_2^{\mathrm{T}}\boldsymbol{E}_2\|_2 = O_P\left(\lambda_K^{-1}\right)$ under Assumption 5.1. Further, it is easy to see that conditional on $\boldsymbol{C}$,

$$\left[\{n/(p\lambda_K)\}^{1/2}\tilde{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{E}_j\right]_{*k} \dot{\sim} \begin{cases} \left(\boldsymbol{0}_K, n/(p\lambda_K)\,\tilde{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{\Sigma}\tilde{\boldsymbol{L}}\right) & \text{if } j = 1 \\ \left(\boldsymbol{0}_{n-t-K}, n/(p\lambda_K)\,\tilde{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{\Sigma}\tilde{\boldsymbol{L}}\right) & \text{if } j = 2 \end{cases},$$

meaning $\|\{n/(p\lambda_K)\}^{1/2}\tilde{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{E}_1\|_2 = O_P(1)$ and $\|\{n/(p\lambda_K)\}^{1/2}\tilde{\boldsymbol{L}}^{\mathrm{T}}\boldsymbol{E}_2\|_2 = O_P\left(n^{1/2}\right)$. Next, for $\rho = p^{-1}\sum_{g=1}^{p}\sigma_g^2$, $\mathbb{E}\left(p^{-1}\boldsymbol{E}_1^{\mathrm{T}}\boldsymbol{E}_1 \mid \boldsymbol{C}\right) = \rho I_K$ and

$$\begin{aligned}\mathbb{V}\left(\left[p^{-1}\boldsymbol{E}_1^{\mathrm{T}}\boldsymbol{E}_1\right]_{rs} \mid \boldsymbol{C}\right) &= p^{-2}\sum_{g=1}^{p}\mathbb{V}\left([\boldsymbol{Q}_Z\boldsymbol{C}_2]_{*r}^{\mathrm{T}}\boldsymbol{e}_g\,[\boldsymbol{Q}_Z\boldsymbol{C}_2]_{*s}^{\mathrm{T}}\boldsymbol{e}_g \mid \boldsymbol{C}\right) \\ &\leq p^{-2}\sum_{g=1}^{p}\mathbb{E}\left\{\left([\boldsymbol{Q}_Z\boldsymbol{C}_2]_{*r}^{\mathrm{T}}\boldsymbol{e}_g\right)^4 \mid \boldsymbol{C}\right\}^{1/2}\mathbb{E}\left\{\left([\boldsymbol{Q}_Z\boldsymbol{C}_2]_{*s}^{\mathrm{T}}\boldsymbol{e}_g\right)^4 \mid \boldsymbol{C}\right\}^{1/2},\end{aligned}$$

where

$$\begin{aligned}\mathbb{E}\left\{\left([\boldsymbol{Q}_Z\boldsymbol{C}_2]_{*r}^{\mathrm{T}}\boldsymbol{e}_g\right)^4 \mid \boldsymbol{C}\right\} &= \sum_{i,j=1}^{n}\mathbb{E}\left([\boldsymbol{e}_g]_i^2\,[\boldsymbol{e}_g]_j^2\,[\boldsymbol{Q}_Z\boldsymbol{C}_2]_{ir}^2\,[\boldsymbol{Q}_Z\boldsymbol{C}_2]_{jr}^2 \mid \boldsymbol{C}\right) \\ &\leq c\mathbb{E}\left\{\left(\sum_{i=1}^{n}[\boldsymbol{Q}_Z\boldsymbol{C}_2]_{ir}^2\right)^2 \mid \boldsymbol{C}\right\} = c\end{aligned}$$

for $c = \max_{g \in [p]} \mathbb{E}\left([\boldsymbol{e}_g]_1^4 \right)$. This shows that $\|p^{-1}\boldsymbol{E}_1^{\mathrm{T}}\boldsymbol{E}_1 - \rho I_K\|_2 = O_P\left(p^{-1/2}\right)$. Lastly, $\mathbb{E}\left(p^{-1}\boldsymbol{E}_2^{\mathrm{T}}\boldsymbol{E}_1 \mid \boldsymbol{C}\right) = \boldsymbol{0}$ and for $k \in [K], i \in [n-t-K]$,

$$
\begin{aligned}
\mathbb{E}\left\{ \left(p^{-1}\left[\boldsymbol{E}_2^{\mathrm{T}}\boldsymbol{E}_1\right]_{ik}\right)^2 \mid \boldsymbol{C} \right\} &= \mathbb{V}\left(p^{-1}\left[\boldsymbol{E}_2^{\mathrm{T}}\boldsymbol{E}_1\right]_{ik} \mid \boldsymbol{C}\right) \\
&= p^{-2}\sum_{g=1}^{p} \mathbb{V}\left([\boldsymbol{Q}_Z\boldsymbol{Q}]_{*i}^{\mathrm{T}} \boldsymbol{e}_g\boldsymbol{e}_g^{\mathrm{T}} [\boldsymbol{Q}_Z\boldsymbol{C}_2]_{*k} \mid \boldsymbol{C}\right) \leq cp^{-1}
\end{aligned}
$$

for c defined above. Therefore, $\|p^{-1}\boldsymbol{E}_2^{\mathrm{T}}\boldsymbol{E}_1\|_2 = O_P\left\{\left(np^{-1}\right)^{1/2}\right\}$.

Define $\mu_k = \gamma_k/\lambda_K$ and let $\hat{\mu}_k$ be the the kth eigenvalue of (5.28) for $k \in [K]$. By Weyl's inequality, the above work shows that $\hat{\mu}_k = \mu_k + O_P\left(p^{-1/2}\right)$. Next, define

$$
\begin{aligned}
\tilde{\boldsymbol{N}} &= \{n/(p\lambda_K)\}^{1/2}\tilde{\boldsymbol{L}} + (p\lambda_K)^{-1/2}\boldsymbol{E}_1, \\
\boldsymbol{B} &= (p\lambda_K)^{-1}\left(n^{1/2}\tilde{\boldsymbol{L}} + \boldsymbol{E}_1\right)^{\mathrm{T}}\boldsymbol{E}_2, \quad \boldsymbol{D} = (p\lambda_K)^{-1}\boldsymbol{E}_2^{\mathrm{T}}\boldsymbol{E}_2
\end{aligned}
$$

and let $\hat{\boldsymbol{v}} \in \mathbb{R}^{K\times K}, \hat{\boldsymbol{w}} \in \mathbb{R}^{(n-t-K)\times K}$ be such that $\left(\hat{\boldsymbol{v}}^{\mathrm{T}}\,\hat{\boldsymbol{w}}^{\mathrm{T}}\right)^{\mathrm{T}}$ are the first K eigenvectors of (5.28). Then

$$
\begin{aligned}
\hat{\mu}_k [\hat{\boldsymbol{v}}]_{*k} &= \left\{\tilde{\boldsymbol{N}}^{\mathrm{T}}\tilde{\boldsymbol{N}} + \boldsymbol{B}\left(\hat{\mu}_k I_{n-t-K} - \boldsymbol{D}\right)^{-1}\boldsymbol{B}^{\mathrm{T}}\right\}[\hat{\boldsymbol{v}}]_{*k} \quad (k = 1, \ldots, K) \\
[\hat{\boldsymbol{w}}]_{*k} &= \left(\hat{\mu}_k I_{n-t-K} - \boldsymbol{D}\right)^{-1}\boldsymbol{B}^{\mathrm{T}}[\hat{\boldsymbol{v}}]_{*k} \quad (k = 1, \ldots, K),
\end{aligned}
$$

where (5.27) then follows by the proof of Theorem 3.10 in Chapter 3. □

***Corollary* 5.1** (Accuracy of the estimate for $P_Z^{\perp}\boldsymbol{C}$ in Step (a) of Algorithm 5.1)**.** *Under the assumptions of Lemma 5.2,*

$$
\|P_{P_Z^{\perp}C} - P_{\hat{C}_2}\|_F^2 = O_P\left(\tilde{p}^{-1}\right).
$$

Proof. The proof is identical to that of (3.29) and is ommitted. □

***Lemma* 5.3.** *Suppose the assumptions of Lemma 5.2 hold and let $g \in \mathcal{S}$. For some $k \in [K]$,* *define z_g to be the ordinary least squares z-score for the $[t+1]$st regressor from the regression $\boldsymbol{y}_g \sim (\boldsymbol{Z}\ [\boldsymbol{C}_2]_{*k})$, restricted to only the observed values of $\boldsymbol{y}_g$. Then if $[\boldsymbol{C}_2]_{*k}$ is independent of $\boldsymbol{y}_g$, $z_g \xrightarrow{\mathcal{D}} N_1(0,1)$ as $n \to \infty$.*

Proof. For notation purposes, we define $\boldsymbol{y}_g = \boldsymbol{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$, $\boldsymbol{e}_g = \boldsymbol{e} = (e_1, \ldots, e_n)^{\mathrm{T}}$, $\boldsymbol{Z} = (\boldsymbol{z}_1 \cdots \boldsymbol{z}_n)^{\mathrm{T}}$, $\mu_i = \boldsymbol{z}_i^{\mathrm{T}}\left(\boldsymbol{\xi}_{*g} + \boldsymbol{A}\boldsymbol{\ell}_{*g}\right)$, $r_i = r_{gi}$ for all $i \in [n]$ and $\boldsymbol{c}_\perp = [\boldsymbol{C}_\perp]_{*k}$. Define $\boldsymbol{R} = \operatorname{diag}(r_1, \ldots, r_n)$ and let $\hat{\ell}$ be the $t+1$st regression coefficient from the ordinary least squares regression of $\boldsymbol{y}$ onto $(\boldsymbol{Z}, \boldsymbol{c}_\perp)$ that ignores missing data. That is,

$$\left(\hat{\boldsymbol{\beta}}, \hat{\ell}\right)^{\mathrm{T}} = \left\{(\boldsymbol{Z}, \boldsymbol{c}_\perp)^{\mathrm{T}} \boldsymbol{R} (\boldsymbol{Z}, \boldsymbol{c}_\perp)\right\}^{-1} \begin{pmatrix} \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{y} \\ \boldsymbol{c}_\perp^{\mathrm{T}}\boldsymbol{R}\boldsymbol{y} \end{pmatrix}. \tag{5.29}$$

Note that

$$\boldsymbol{c}_\perp = P_Z^\perp \boldsymbol{c}, \quad \boldsymbol{c} = \boldsymbol{\Xi}\left(n^{-1}\boldsymbol{\Xi}^{\mathrm{T}} P_Z^\perp \boldsymbol{\Xi}\right)^{-1/2} \boldsymbol{u}_k,$$

where $\boldsymbol{u}_k \in \mathbb{R}^K$ is the kth column of $\boldsymbol{U}$ defined in the statement of Lemma 5.2. By the assumptions of $\boldsymbol{\Xi}$ in Assumption 5.1, $n^{-1}\boldsymbol{\Xi}^{\mathrm{T}} P_Z^\perp \boldsymbol{\Xi} = I_K + O_P\left(n^{-1/2}\right)$ and $\|\boldsymbol{u}_k - \boldsymbol{a}_k\|_2 = O_P\left(n^{-1/2}\right)$, where $\boldsymbol{a}_k \in \mathbb{R}^K$ is the kth standard basis vector. Since $\boldsymbol{c}_\perp$ is independent of $\boldsymbol{y}$, it is also independent of $\boldsymbol{R}$ by Model (5.3), meaning it suffices to assume $\boldsymbol{\Xi}$ is independent of $\boldsymbol{y}$. Further, we can re-write (5.29) as

$$\left(\hat{\boldsymbol{\beta}}, \hat{\ell}\right)^{\mathrm{T}} = \begin{pmatrix} I_t & \boldsymbol{s} \\ \boldsymbol{0} & I_K \end{pmatrix} \left\{(\boldsymbol{Z}, \boldsymbol{c})^{\mathrm{T}} \boldsymbol{R} (\boldsymbol{Z}, \boldsymbol{c})\right\}^{-1} \begin{pmatrix} \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{y} \\ \boldsymbol{c}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{y} \end{pmatrix}, \quad \boldsymbol{s} = \left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}\right)^{-1} \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{c}.$$

Therefore, to understand the distribution of $\hat{\ell}$, it suffices to replace $\boldsymbol{c}_\perp$ with $\boldsymbol{c}$ in (5.29).

Define the function

$$\boldsymbol{h}(\boldsymbol{\beta}, \ell) = n^{-1} (\boldsymbol{Z}, \boldsymbol{c})^{\mathrm{T}} \boldsymbol{R} (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta} - \boldsymbol{c}\ell)$$

and let

$$\boldsymbol{\beta}_{r=1} = \left\{\mathbb{E}\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{Z}\right)\right\}^{-1}\mathbb{E}\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{y}\right).$$

Note that $\limsup_{n\to\infty}\|\boldsymbol{\beta}_{r=1}\|_2 < \infty$. Let $\boldsymbol{R} = \operatorname{diag}(r_1, \ldots, r_n)$. We start by understanding the asymptotic properties of $\boldsymbol{h}(\boldsymbol{\beta}_{r=1}, 0)$, which we can do by analyzing the following:

(i) $\underline{n^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{y}}$

$$\mathbb{V}\left(n^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{y}\right) = n^{-2}\sum_{i=1}^{n}\boldsymbol{z}_i\boldsymbol{z}_i^{\mathrm{T}}\,\mathbb{V}(r_i y_i) \preceq n^{-2}\sum_{i=1}^{n}\boldsymbol{z}_i\boldsymbol{z}_i^{\mathrm{T}}\mathbb{E}\left(y_i^2\right) \preceq n^{-1}c\left(n^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}\right)$$

where $c > 0$ is a constant. Therefore, $n^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{y} = \mathbb{E}\left(n^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{y}\right) + O_P\left(n^{-1/2}\right)$.

(ii) $\underline{n^{-1}\boldsymbol{c}^{\mathrm{T}}\boldsymbol{R}(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}_{r=1})}$

$$\begin{aligned} n^{-1}\boldsymbol{c}^{\mathrm{T}}\boldsymbol{R}(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}_{r=1}) &= \boldsymbol{a}_k^{\mathrm{T}}\left(n^{-1}\boldsymbol{\Xi}^{\mathrm{T}}P_Z^{\perp}\boldsymbol{\Xi}\right)^{-1/2} n^{-1}\boldsymbol{\Xi}^{\mathrm{T}}\boldsymbol{R}(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}_{r=1}) \\ &= n^{-1}\sum_{i=1}^{n}[\boldsymbol{\Xi}]_{ij}\, r_i\left(y_i - \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\beta}_{r=1}\right) + O_P\left(n^{-1}\right) \end{aligned}$$

Since $\boldsymbol{\Xi}$ is independent of r_i and y_i, $\mathbb{E}\left\{[\boldsymbol{\Xi}]_{ij}\, r_i\left(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_{r=1}\right)\right\} = 0$ and

$$n^{-1}\sum_{i=1}^{n}[\boldsymbol{\Xi}]_{ij}^2\, r_i\left(y_i - \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\beta}_{r=1}\right)^2 = n^{-1}\sum_{i=1}^{n}\mathbb{E}\left\{r_i\left(y_i - \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\beta}_{r=1}\right)^2\right\} + O_P\left(n^{-1/2}\right)$$

by the bounded fourth moment assumptions. Let $\alpha_* = \alpha_{*_g} > 0, \delta_* = \delta_{*_g} \in \mathbb{R}$ and $\gamma_i = \mu_i - \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\beta}_{r=1}$. Note that $\max_{i\in[n]}|\gamma_i|, \max_{i\in[n]}|\mu_i| \leq c$ for some constant $c > 0$ by Assumption 5.1. Let $M > 0$ be large enough so that

$$\max_{i\in[n]}\left[\mathbb{E}\left\{(e_1 + \gamma_i)^2\right\} I(e_1 \geq -M)\right] \geq c_1^{-1}.$$

Then because $e_1, \ldots, e_n$ are identically distributed,

$$\begin{aligned}
\mathbb{E}\left\{r_i\left(y_i - \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\beta}_{r=1}\right)^2\right\} &= \mathbb{E}\left[\Psi\left\{\alpha_*\left(e_i + \mu_i - \delta_*\right)\right\}\left(e_i + \gamma_i\right)^2\right] \\
&\geq \mathbb{E}\left[\Psi\left\{\alpha_*\left(e_i + \mu_i - \delta_*\right)\right\}\left(e_i + \gamma_i\right)^2 I\left(e_i \geq -M\right)\right] \\
&\geq \Psi\left[\alpha_*\left\{-M + \min_{j\in[n]}\left(\mu_j\right) - \delta_*\right\}\right] c_1^{-1},
\end{aligned}$$

where

$$\liminf_{n\to\infty}\left\{\min_{j\in[n]}\left(\mu_j\right)\right\} > -\infty \Rightarrow \liminf_{n\to\infty}\Psi\left[\alpha_*\left\{-M + \min_{j\in[n]}\left(\mu_j\right) - \delta_*\right\}\right] > 0.$$

By the Lindeberg-Feller Central Limit Theorem, we get that

$$\begin{aligned}
&n^{-1/2}v_n^{-1/2}\boldsymbol{c}^{\mathrm{T}}\boldsymbol{R}\left(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}_{r=1}\right) \overset{\mathcal{D}}{=} N\left(0,1\right) + o_P(1) \\
&v_n = n^{-1}\sum_{i=1}^{n}\mathbb{E}\left\{r_i\left(y_i - \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\beta}_{r=1}\right)^2\right\},
\end{aligned}$$

where $\liminf_{n\to\infty} v_n > 0$ by the above work.

(iii) We see that $n^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{Z} = \mathbb{E}\left(n^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{Z}\right) + O_P\left(n^{-1/2}\right)$, where $\mathbb{E}\left(n^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{Z}\right) \succeq cn^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}$ for some constant $c > 0$ by Lemma 5.1. An analysis identical to that in (ii) shows $n^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{c} = O_P\left(n^{-1/2}\right)$.

(iv) $\underline{n^{-1}\boldsymbol{c}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{c}}$

$$\begin{aligned}
n^{-1}\boldsymbol{c}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{c} &= \boldsymbol{a}_k^{\mathrm{T}}\left(n^{-1}\Xi^{\mathrm{T}}P_Z^{\perp}\Xi\right)^{-1/2} n^{-1}\Xi^{\mathrm{T}}\boldsymbol{R}\Xi\left(n^{-1}\Xi^{\mathrm{T}}P_Z^{\perp}\Xi\right)^{-1/2}\boldsymbol{a}_k \\
&= n^{-1}\sum_{i=1}^{n}\mathbb{P}\left(r_i = 1 \mid \boldsymbol{Z}\right) + O_P\left(n^{-1/2}\right) \\
&= n^{-1}\underbrace{\sum_{i=1}^{n} r_i}_{n_{r=1}} + O_P\left(n^{-1/2}\right).
\end{aligned}$$

This shows that

$$n^{1/2}\begin{pmatrix}\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_{r=1}\\ \hat{\ell}\end{pmatrix}=\begin{pmatrix}n^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{Z} & n^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{c}\\ n^{-1}\boldsymbol{c}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{Z} & n^{-1}\boldsymbol{c}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{c}\end{pmatrix}^{-1}\begin{Bmatrix}n^{-1/2}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{R}\left(\boldsymbol{y}-\boldsymbol{Z}\boldsymbol{\beta}_{r=1}\right)\\ n^{-1/2}\boldsymbol{c}^{\mathrm{T}}\boldsymbol{R}\left(\boldsymbol{y}-\boldsymbol{Z}\boldsymbol{\beta}_{r=1}\right)\end{Bmatrix}.$$

The results from (i), (ii), (iii) and (iv) and a little algebra then show that

$$\left\{[\boldsymbol{M}]_{(t+1)(t+1)}\,\tilde{v}_n\right\}^{-1/2}\hat{\ell}\overset{\mathcal{D}}{=}N(0,1)+o_P(1),\quad \boldsymbol{M}=\begin{pmatrix}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{Z} & \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{c}\\ \boldsymbol{c}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{Z} & \boldsymbol{c}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{c}\end{pmatrix}^{-1}$$

$$\tilde{v}_n=\left(n_{r=1}-t-1\right)^{-1}\sum_{i=1}^{n}r_i\left(y_i-\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\beta}_{r=1}\right)^2$$

Lastly, for $\boldsymbol{\delta}=\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_{r=1}$ and $\tilde{n}_{r=1}=n_{r=1}-t-1$,

$$\begin{aligned}\hat{\tilde{v}}_n=&\tilde{n}_{r=1}^{-1}\sum_{i=1}^{n}r_i\left(y_i-\boldsymbol{z}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}-[\boldsymbol{c}]_i\,\hat{\ell}\right)^2=\tilde{v}_n+2\tilde{n}_{r=1}^{-1}\sum_{i=1}^{n}r_i\left(y_i-\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\beta}_{r=1}\right)\left(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\delta}-[\boldsymbol{c}]_i\,\hat{\ell}\right)\\ &+\tilde{n}_{r=1}^{-1}\sum_{i=1}^{n}r_i\left(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\delta}+[\boldsymbol{c}]_i\,\hat{\ell}\right)^2.\end{aligned}\tag{5.30}$$

By (i), (ii), (iii) and (iv), $\|\boldsymbol{\delta}\|_2,\|\hat{\ell}\|_2=O_P\left(n^{-1/2}\right)$, meaning the the last two terms are $o_P(1)$. This completes the proof. □

***Theorem* 5.1.** *Suppose Assumption 5.1 holds, $\epsilon=0$ and let $g\in\mathcal{S}$. For some $k\in[K]$, define $\hat{z}_g$ to be the ordinary least squares z-score for the $[t+1]$st regressor from the regression $\boldsymbol{y}_g\sim\left(\boldsymbol{Z}\ \left[\hat{\boldsymbol{C}}_2\right]_{*k}\right)$, restricted to only the observed values of $\boldsymbol{y}_g$. Then under the null hypothesis that $[\boldsymbol{C}_2]_{*k}$ or $\left[\hat{\boldsymbol{C}}_\perp\right]_{*k}$ is independent of $\boldsymbol{y}_g$, $\hat{z}_g\xrightarrow{\mathcal{D}}N_1(0,1)$ as $n,p\to\infty$.*

Proof. Let z_g, $\boldsymbol{R}$, r_i, $\boldsymbol{y}$ and y_i be as defined in Lemma 5.3. Both $\hat{z}_g$ and z_g can be written as $\frac{\text{"estimate"}}{\text{"standard error of estimate"}}$. Define

$$\hat{\boldsymbol{M}} = \begin{pmatrix} \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{Z} & \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{R}\left[\hat{\boldsymbol{C}}_2\right]_{*k} \\ \left[\hat{\boldsymbol{C}}_2\right]_{*k}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{Z} & \left[\hat{\boldsymbol{C}}_2\right]_{*k}^{\mathrm{T}}\boldsymbol{R}\left[\hat{\boldsymbol{C}}_2\right]_{*k} \end{pmatrix}^{-1}, \quad \boldsymbol{M} = \begin{pmatrix} \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{Z} & \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{R}\left[\boldsymbol{C}_2\right]_{*k} \\ \left[\boldsymbol{C}_2\right]_{*k}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{Z} & \left[\boldsymbol{C}_2\right]_{*k}^{\mathrm{T}}\boldsymbol{R}\left[\boldsymbol{C}_2\right]_{*k} \end{pmatrix}^{-1}.$$

The numerators of $\hat{z}_g$ and z_g are the $[t+1]st$ elements of

$$\hat{\boldsymbol{M}}\begin{pmatrix} \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{y} \\ \left[\hat{\boldsymbol{C}}_2\right]_{*k}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{y} \end{pmatrix}, \quad \boldsymbol{M}\begin{pmatrix} \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{y} \\ \left[\boldsymbol{C}_2\right]_{*k}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{y} \end{pmatrix}$$

and the denominators are

$$\left[\left[\hat{\boldsymbol{M}}\right]_{(t+1)(t+1)}\left\{(n_{r=1}-t-1)^{-1}(\boldsymbol{R}\boldsymbol{y})^{\mathrm{T}}P^{\perp}_{R\left(Z,\left[\hat{C}_2\right]_{*k}\right)}(\boldsymbol{R}\boldsymbol{y})\right\}\right]^{1/2},$$
$$\left[\left[\boldsymbol{M}\right]_{(t+1)(t+1)}\left\{(n_{r=1}-t-1)^{-1}(\boldsymbol{R}\boldsymbol{y})^{\mathrm{T}}P^{\perp}_{R(Z,[C_2]_{*k})}(\boldsymbol{R}\boldsymbol{y})\right\}\right]^{1/2},$$

where $n_{r=1} = \sum_{i=1}^{n} r_i$. Note that for $\hat{\hat{v}}_n$ defined in (5.30),

$$\hat{\hat{v}}_n = (n_{r=1}-t-1)^{-1}(\boldsymbol{R}\boldsymbol{y})^{\mathrm{T}}P^{\perp}_{R(Z,[C_2]_{*k})}(\boldsymbol{R}\boldsymbol{y}).$$

First, if $\left[\hat{\boldsymbol{C}}_{\perp}\right]_{*k}$ is independent of $\boldsymbol{y}$, then it suffices to assume that $\boldsymbol{\Xi}$ is independent of both $\boldsymbol{y}$ and $\boldsymbol{R}$. Therefore, by Lemma 5.3, we simply have to show the following to prove the theorem:

$$\|n\hat{\boldsymbol{M}} - n\boldsymbol{M}\|_2 = o_P\left(n^{-1/2}\right) \tag{5.31a}$$
$$n^{-1}\left[\hat{\boldsymbol{C}}_2\right]_{*k}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{y} = n^{-1}\left[\boldsymbol{C}_2\right]_{*k}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{y} + o_P\left(n^{-1/2}\right) \tag{5.31b}$$
$$(n_{r=1}-t-1)^{-1}(\boldsymbol{R}\boldsymbol{y})^{\mathrm{T}}P^{\perp}_{R\left(Z,\left[\hat{C}_2\right]_{*k}\right)}(\boldsymbol{R}\boldsymbol{y}) = \hat{\hat{v}}_n + o_P(1). \tag{5.31c}$$

We start by showing (5.31a). Define $\hat{\boldsymbol{c}} = \left[\hat{\boldsymbol{C}}_2\right]_{*k}$ and $\boldsymbol{c} = [\boldsymbol{C}_2]_{*k}$. Then by Lemma 5.2

and (5.26),

$$n^{-1}\hat{\boldsymbol{c}}^{\mathrm{T}}\boldsymbol{R}\hat{\boldsymbol{c}} = n^{-1}\left[\hat{\boldsymbol{v}}\right]_{*k}^{\mathrm{T}}\boldsymbol{C}_2^{\mathrm{T}}\boldsymbol{R}\boldsymbol{C}_2\left[\hat{\boldsymbol{v}}\right]_{*k} + 2n^{-1/2}\left[\hat{\boldsymbol{v}}\right]_{*k}^{\mathrm{T}}\boldsymbol{C}_2^{\mathrm{T}}\boldsymbol{R}\boldsymbol{Q}_{C_2}\left[\hat{\boldsymbol{w}}\right]_{*k}$$
$$+ \left[\hat{\boldsymbol{w}}\right]_{*k}^{\mathrm{T}}\boldsymbol{Q}_{C_2}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{Q}_{C_2}\left[\hat{\boldsymbol{w}}\right]_{*k}$$

By (5.27a), the third term is $o_P\left(n^{-1/2}\right)$. Similarly, since $\|n^{-1/2}\boldsymbol{R}\boldsymbol{C}_2\|_2 \leq \|n^{-1/2}\boldsymbol{C}_2\|_2 = 1$, we can use (5.27a) to get

$$n^{-1}\hat{\boldsymbol{c}}^{\mathrm{T}}\boldsymbol{R}\hat{\boldsymbol{c}} = n^{-1}\boldsymbol{c}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{c} + 2n^{-1/2}\boldsymbol{c}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{Q}_{C_2}\left[\hat{\boldsymbol{w}}\right]_{*k} + o_P\left(n^{-1/2}\right).$$

We then use (5.27b) to show that the second term in the above expression is $o_P\left(n^{-1/2}\right)$. The proof of this follows from the fact that $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{R}$ is independent of $\boldsymbol{E}_{\mathcal{S}^c}$. The details are nearly identically to those used to prove Lemma 5.2 and are ommitted. An identical technique can also be used to show that $\|n^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{R}\hat{\boldsymbol{c}} - n^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{c}\|_2 = o_P\left(n^{-1/2}\right)$, which proves (5.31a).

To show (5.31b), we again use (5.26), which shows that

$$n^{-1}\hat{\boldsymbol{c}}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{y} = n^{-1}\left[\hat{\boldsymbol{v}}\right]_{*k}^{\mathrm{T}}\boldsymbol{C}_2^{\mathrm{T}}\boldsymbol{R}\boldsymbol{y} + n^{-1/2}\left[\hat{\boldsymbol{w}}\right]_{*k}^{\mathrm{T}}\boldsymbol{Q}_{C_2}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{y} = n^{-1}\boldsymbol{c}^{\mathrm{T}}\boldsymbol{R}y + n^{-1/2}\left[\hat{\boldsymbol{w}}\right]_{*k}^{\mathrm{T}}\boldsymbol{Q}_{C_2}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{y}$$
$$+ o_P\left(n^{-1/2}\right),$$

where the second equality follows from (5.27a). Again, the proof that $n^{-1/2}\left[\hat{\boldsymbol{w}}\right]_{*k}^{\mathrm{T}}\boldsymbol{Q}_{C_2}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{y} = o_P\left(n^{-1/2}\right)$ follows from (5.27b) the fact that $\boldsymbol{R}\boldsymbol{y}$ is independent of $\boldsymbol{E}_{\mathcal{S}^c}$, and is ommitted.

To show (5.31c), let $\tilde{n} = n_{r=1} - t - 1$. Then

$$\begin{aligned}\hat{x}_n &= \tilde{n}^{-1}\left(\boldsymbol{R}\boldsymbol{y}\right)^{\mathrm{T}} P^{\perp}_{R\left(Z,\left[\hat{C}_2\right]_{*k}\right)}\left(\boldsymbol{R}\boldsymbol{y}\right) \\ &= n^{-1}\left(\tilde{n}^{-1/2}\boldsymbol{R}\boldsymbol{y}\right)^{\mathrm{T}}\left(\boldsymbol{Z},\,\hat{\boldsymbol{c}}\right)\left(n\hat{\boldsymbol{M}}\right)\left(\boldsymbol{Z},\,\hat{\boldsymbol{c}}\right)^{\mathrm{T}}\left(\tilde{n}^{-1/2}\boldsymbol{R}\boldsymbol{y}\right).\end{aligned}$$

First, $\|\tilde{n}^{-1/2}\boldsymbol{R}\boldsymbol{y}\|_2 \leq \|\tilde{n}^{-1/2}\boldsymbol{y}\|_2 = O_P(1)$. Next, because $\|n\hat{\boldsymbol{M}} - n\boldsymbol{M}\| = o_P\left(n^{-1/2}\right)$,

$\|n^{-1/2}\boldsymbol{Z}\|_2 = O(1)$ and $\|n^{-1/2}\hat{\boldsymbol{c}}\|_2 = 1$,

$$\hat{x}_n = n^{-1}\left(\tilde{n}^{-1/2}\boldsymbol{Ry}\right)^{\mathrm{T}}(\boldsymbol{Z},\,\hat{\boldsymbol{c}})\,(n\boldsymbol{M})\,(\boldsymbol{Z},\,\hat{\boldsymbol{c}})^{\mathrm{T}}\left(\tilde{n}^{-1/2}\boldsymbol{Ry}\right) + o_P\left(n^{-1/2}\right).$$

Lastly, (5.26) and (5.27a) imply

$$\hat{x}_n = n^{-1}\left(\tilde{n}^{-1/2}\boldsymbol{Ry}\right)^{\mathrm{T}}(\boldsymbol{Z},\,\boldsymbol{c})\,(n\boldsymbol{M})\,(\boldsymbol{Z},\,\boldsymbol{c})^{\mathrm{T}}\left(\tilde{n}^{-1/2}\boldsymbol{Ry}\right) + o_P(1) = \hat{\tilde{v}}_n + o_P(1),$$

which completes the proof. □

5.10.6 Approximating $\mathrm{pr}\left\{\bar{\boldsymbol{G}}_g\left(\boldsymbol{\eta}_{*_g}\right) \mid \boldsymbol{\eta}_{*_g}\right\}$ *with a normal distribution.*

In this section, we justify approximating the distribution of $\bar{\boldsymbol{G}}\left(\boldsymbol{\eta}_{*g}\right) \mid \boldsymbol{\eta}_{*g}$ with a normal distribution. This helps to justify our hierarchical Bayesian generalized method of moments (HB-GMM) procedure in Section 5.4.4. Our main result is Theorem 5.2.

***Assumption* 5.2.** *$\Psi(x)$ is twice continuously differentiable with bounded first and second derivatives. Further, for some large constants $M_1, M_2 > 0$,*

(i) Either $a|x|^k\Psi(x) = 1 + R(x)$ or $a\exp(k|x|)\Psi(x) = 1 + R(x)$ for all $x \in (-\infty, -M_1)$, where $\lim_{x\to-\infty} R(x) = 0$ and $|dR(x)/dx|, |d^2R(x)/dx^2| \leq M_1$ for some $a, k > 0$.

(ii) $\mathbb{E}\left(\left[\Psi\left\{\left(\alpha_{*_g} + M_1^{-1}\right)\left(\left[\boldsymbol{y}_g\right]_1 - \delta_{*_g}\right)\right\}\right]^{-\left(3+M_2^{-1}\right)}\right) < \infty$ for all $g \in \mathcal{S}$.

***Lemma* 5.4.** *Fix a $g \in \mathcal{S}$ and let M_1, M_2 be as defined in Assumption 5.2. Under Assumptions 5.1 and 5.2,*

$$\limsup_{n\to\infty}\left[\max_{i\in[n]}\left\{\mathbb{E}\left(\left[\Psi\left\{\left(\alpha_{*_g} + M_1^{-1}\right)\left(\left[\boldsymbol{y}_g\right]_i - \delta_{*_g}\right)\right\}\right]^{-\left(3+M_2^{-1}\right)}\right)\right\}\right] < \infty.$$

Proof. Let $\mu_i = [\boldsymbol{Z}]_{i*}^{\mathrm{T}}\left(\boldsymbol{\xi}_{*_g} + \boldsymbol{A}\boldsymbol{\ell}_{*_g}\right)$ and $\tilde{e}_{gi} = \left[\boldsymbol{y}_g\right]_i - \mu_i$ for each $i \in [n]$. Then $\tilde{e}_{g1}, \ldots, \tilde{e}_{gn}$ are identically distributed and $\liminf_{n\to\infty}\min_{i\in[n]}\mu_i \geq \mu > -\infty$ because the entries of $\boldsymbol{Z}$ are uniformly bounded. Therefore, for any $i \in [n]$,

$$\Psi\left\{\left(\alpha_{*_g} + M_1^{-1}\right)\left(\left[\boldsymbol{y}_g\right]_i - \delta_{*_g}\right)\right\} \geq \Psi\left[\left(\alpha_{*_g} + M_1^{-1}\right)\left\{\tilde{e}_{gi} + \mu_1 + (\mu - \mu_1) - \delta_{*_g}\right\}\right]$$
$$\stackrel{\mathcal{D}}{=} \Psi\left[\left(\alpha_{*_g} + M_1^{-1}\right)\left\{\left[\boldsymbol{y}_g\right]_1 + (\mu - \mu_1) - \delta_{*_g}\right\}\right].$$

The result then follows because $\mu - \mu_1$ is finite. □

***Remark* 5.1.** *Under Assumption 5.1, one can show Assumption 5.2 holds for the following values of* $\Psi(x)$*:*

(i) If $\Psi(x) = \exp(x)/\{1 + \exp(x)\}$*, Assumption 5.2 holds if the entries of* $\boldsymbol{\Xi}$ *and* $\boldsymbol{e}_g$ *have a moment generating function that is defined on all of* $\mathbb{R}$*.*

(ii) If $\Psi(x) = F_\nu(x)$*, Assumption 5.2 holds if* $\mathbb{E}\left(|[\boldsymbol{\Xi}]_{1k}|^{3\nu}\right), \mathbb{E}\left(|\boldsymbol{e}_{g1}|^{3\nu}\right) < \infty$ *for all* $k \in K$*.*

***Lemma* 5.5.** *Suppose* $t < 3$*, fix a* $g \in \mathcal{S}$ *and let* $j_1, \ldots, j_{3-t} \in [\hat{K}]$ *where* $\hat{K} \leq K$*. For* $\boldsymbol{C}_2$ *defined in* (5.25) *and* $w_{gi}\left(\boldsymbol{\theta}_g\right) = 1/\Psi\left\{\left[\boldsymbol{\theta}_g\right]_1\left(y_{gi} - \left[\boldsymbol{\theta}_g\right]_2\right)\right\}$*, define*

$$\boldsymbol{u}_{gi} = \left([\boldsymbol{Z}]_i^{\mathrm{T}}, [\boldsymbol{C}_2]_{ij_1}, \cdots, [\boldsymbol{C}_2]_{ij_{3-t}}\right)^{\mathrm{T}} \in \mathbb{R}^3 \quad (i = 1, \ldots, n)$$
$$\bar{\boldsymbol{G}}_g\left(\boldsymbol{\theta}_g\right) = n^{-1}\sum_{i=1}^n \boldsymbol{u}_{gi}\left\{1 - r_{gi}w_{gi}\left(\boldsymbol{\theta}_g\right)\right\}$$
$$\bar{\boldsymbol{\Sigma}}_g\left(\boldsymbol{\theta}_g\right) = n^{-1}\sum_{i=1}^n\left[\boldsymbol{u}_{gi}\left\{1 - r_{gi}w_{gi}\left(\boldsymbol{\theta}_g\right)\right\} - \bar{\boldsymbol{G}}_g\left(\boldsymbol{\theta}_g\right)\right]\left[\boldsymbol{u}_{gi}\left\{1 - r_{gi}w_{gi}\left(\boldsymbol{\theta}_g\right)\right\} - \bar{\boldsymbol{G}}_g\left(\boldsymbol{\theta}_g\right)\right]^{\mathrm{T}}.$$

Then for $\boldsymbol{\theta}_{*_g} = \left(\alpha_{*_g}, \delta_{*_g}\right)^{\mathrm{T}}$ *and if Assumptions 5.1 and 5.2 hold,*

$$\|n\,\mathbb{V}\left\{\bar{\boldsymbol{G}}_g\left(\boldsymbol{\theta}_{*_g}\right)\right\} - \bar{\boldsymbol{\Sigma}}_g\left(\boldsymbol{\theta}_{*_g}\right)\|_2 = o_P(1), \quad \|n\,\mathbb{V}\left\{\bar{\boldsymbol{G}}_g\left(\boldsymbol{\theta}_{*_g}\right)\right\}\|_2, \|\left[n\,\mathbb{V}\left\{\bar{\boldsymbol{G}}_g\left(\boldsymbol{\theta}_{*_g}\right)\right\}\right]^{-1}\|_2 \leq c$$
$$n^{1/2}\left\{\bar{\boldsymbol{\Sigma}}_g\left(\boldsymbol{\theta}_{*_g}\right)\right\}^{-1/2}\bar{\boldsymbol{G}}_g\left(\boldsymbol{\theta}_{*_g}\right) \xrightarrow{\mathcal{D}} N_3\left(\boldsymbol{0}_3, I_3\right)$$

as $n \to \infty$*, where* $c > 0$ *is a constant that does not depend on* n*.*

Proof. Since K is at most finite, it suffices to assume $\boldsymbol{u}_{gi} = \left([\boldsymbol{Z}]_{i*}^{\mathrm{T}}, [\boldsymbol{C}_2]_{i*}^{\mathrm{T}}\right)^{\mathrm{T}}$ to prove the lemma. By Assumption 5.1,

$$\boldsymbol{u}_{gi} = \hat{\boldsymbol{M}}^{\mathrm{T}} \begin{pmatrix} [\boldsymbol{Z}]_{i*} \\ \Xi_{i*} \end{pmatrix}, \quad \hat{\boldsymbol{M}} = \begin{pmatrix} I_t & -\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}\right)^{-1} \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\Xi} \\ \boldsymbol{0} & \left(\boldsymbol{\Xi}^{\mathrm{T}} P_Z^{\perp} \boldsymbol{\Xi}\right)^{-1/2} \boldsymbol{U} \end{pmatrix}$$

for $\boldsymbol{U}$ defined in the statement of Lemma 5.2. And since $\|\hat{\boldsymbol{M}} - I_{t+K}\|_2 = o_P(1)$ by Assumption 5.1, it suffices to further simplify the problem and assume $\boldsymbol{u}_{gi} = \left([\boldsymbol{Z}]_{i*}^{\mathrm{T}}, [\boldsymbol{\Xi}]_{i*}^{\mathrm{T}}\right)^{\mathrm{T}}$, meaning $\bar{\boldsymbol{G}}_g\left(\boldsymbol{\theta}_{*g}\right)$ is an average of independent random variables. Therefore, to prove the lemma, we need only check that the Lindeberg condition hold and that $\|n\,\mathbb{V}\left\{\bar{\boldsymbol{G}}_g\left(\boldsymbol{\theta}_{*g}\right)\right\} - \bar{\boldsymbol{\Sigma}}\left(\boldsymbol{\theta}_{*g}\right)\|_2 = o_P(1)$.

Let $\boldsymbol{v} = \left(\boldsymbol{v}_1^{\mathrm{T}}, \boldsymbol{v}_2^{\mathrm{T}}\right)^{\mathrm{T}} \in \mathbb{R}^{t+K}$ be a unit vector, where $\boldsymbol{v}_1 \in \mathbb{R}^t$ and $\boldsymbol{v}_2 \in \mathbb{R}^{\hat{K}}$. First,

$$\begin{aligned} n\,\mathbb{V}\left\{\boldsymbol{v}^{\mathrm{T}}\bar{\boldsymbol{G}}_g\left(\boldsymbol{\theta}_{*g}\right)\right\} &= n^{-1}\sum_{i=1}^{n} \mathbb{E}\left[\left\{1 - r_{gi}w_{gi}\left(\boldsymbol{\theta}_{*g}\right)\right\}^2 \left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{u}_{gi}\right)^2\right] \\ &\leq n^{-1}\sum_{i=1}^{n} \left(\mathbb{E}\left[\left\{1 - r_{gi}w_{gi}\left(\boldsymbol{\theta}_{*g}\right)\right\}^4\right]\right)^{1/2} \left[\mathbb{E}\left\{\left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{u}_{gi}\right)^4\right\}\right]^{1/2}. \end{aligned}$$

We see that

$$\mathbb{E}\left[\left\{r_{gi}w_{gi}\left(\boldsymbol{\theta}_{*g}\right)\right\}^4\right] = \mathbb{E}\left[\left\{w_{gi}\left(\boldsymbol{\theta}_{*g}\right)\right\}^3\right], \mathbb{E}\left\{\left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{u}_{gi}\right)^4\right\} \leq c$$

for some constant c that does not depend on i or n by Lemma 5.6 and Assumption 5.1, meaning $n\,\mathbb{V}\left\{\boldsymbol{v}^{\mathrm{T}}\bar{\boldsymbol{G}}_g\left(\boldsymbol{\theta}_{*g}\right)\right\}$ exists and is bounded from above. Next, let $M > 0$ be a large constant. Then for $\tilde{e}_{gi}$ as defined in Lemma 5.6, and because $\mathbb{E}\left(\left[\boldsymbol{y}_g\right]_i\right)$ is uniformly bounded from below,

$$\begin{aligned}
n\,\mathbb{V}\left\{\boldsymbol{v}^{\mathrm{T}}\bar{\boldsymbol{G}}_g\left(\boldsymbol{\theta}_{*g}\right)\right\} &= n^{-1}\sum_{i=1}^{n}\mathbb{E}\left[\left\{1-r_{gi}w_{gi}\left(\boldsymbol{\theta}_{*g}\right)\right\}^2\left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{u}_{gi}\right)^2\right] \\
&\geq n^{-1}\sum_{i=1}^{n}\mathbb{E}\left[\left\{1-r_{gi}w_{gi}\left(\boldsymbol{\theta}_{*g}\right)\right\}^2\left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{u}_{gi}\right)^2 I\left(\tilde{e}_{gi}\geq -M\right)\right] \\
&\geq \eta_M\mathbb{E}\left\{\left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{u}_{g1}\right)^2 I\left(\tilde{e}_{g1}\geq -M\right)\right\}
\end{aligned}$$

where $\eta_M > 0$ for all M. And since $\mathbb{E}\left\{\left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{u}_{g1}\right)^2 I\left(\tilde{e}_{g1}\geq -M\right)\right\} \geq c_1^{-1}/2$ for all M large enough (where c_1 is defined in Assumption 5.1), $n\,\mathbb{V}\left\{\boldsymbol{v}^{\mathrm{T}}\bar{\boldsymbol{G}}_g\left(\boldsymbol{\theta}_{*g}\right)\right\} \geq c_1^{-1}\eta_M/2$. This proves that the eigenvalues of $n\,\mathbb{V}\left\{\boldsymbol{v}^{\mathrm{T}}\bar{\boldsymbol{G}}_g\left(\boldsymbol{\theta}_{*g}\right)\right\}$ are uniformly bounded above 0 and below infinity.

To prove that the Lindeberg condition holds, we note that $\boldsymbol{v}^{\mathrm{T}}\boldsymbol{u}_{gi} = \boldsymbol{v}_1^{\mathrm{T}}\left[\boldsymbol{Z}\right]_{i*} + \boldsymbol{v}_2^{\mathrm{T}}\left[\Xi\right]_{i*}$, where for $\|[\boldsymbol{Z}]_{i*}\|_2 \leq c_z$,

$$\left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{u}_{gi}\right)^2 \leq \left(c_z + \boldsymbol{v}_2^{\mathrm{T}}\left[\Xi\right]_{i*}\right)^2 I\left(\boldsymbol{v}_2^{\mathrm{T}}\left[\Xi\right]_{i*}\geq 0\right) + \left(c_z - \boldsymbol{v}_2^{\mathrm{T}}\left[\Xi\right]_{i*}\right)^2 I\left(\boldsymbol{v}_2^{\mathrm{T}}\left[\Xi\right]_{i*} < 0\right).$$

For the remainder of the proof, we let $\tilde{e}_{gi} = \left[\boldsymbol{e}_g\right]_i + \left[\Xi\right]_{i*}^{\mathrm{T}}\boldsymbol{\ell}_{*g}$, μ_i and μ be as defined in the proof of Lemma 5.4, and let $\tilde{\mu} = \limsup_{n\to\infty}\left(\max_{i\in[n]}\mu_i\right)$. For each $i\in[n]$, we define

$$X_i = \left\{1-r_{gi}w_{gi}\left(\boldsymbol{\theta}_{*g}\right)\right\}^2\left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{u}_{gi}\right)^2 = \left(1-r_{gi}\left[\Psi\left\{\alpha_{*g}\left(\mu_i+\tilde{e}_{gi}-\delta_{*g}\right)\right\}\right]^{-1}\right)^2\left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{u}_{gi}\right)^2.$$

Next, define

$$r_{gi}^{(m)}\begin{cases} = 1 & \text{if } r_{gi}=1 \\ \sim \mathrm{Ber}\left[\frac{\Psi\left\{\alpha_{*g}\left(\tilde{\mu}+\tilde{e}_{gi}-\delta_{*g}\right)\right\}-\Psi\left\{\alpha_{*g}\left(\mu_i+\tilde{e}_{gi}-\delta_{*g}\right)\right\}}{1-\Psi\left\{\alpha_{*g}\left(\mu_i+\tilde{e}_{gi}-\delta_{*g}\right)\right\}}\right] & \text{if } r_{gi}=0 \end{cases},$$

where $r_{gi}\leq r_{gi}^{(m)}$ and conditional on $\left[\boldsymbol{e}_g\right]_i$ and $\left[\Xi\right]_{i*}$, $r_{gi}^{(m)} \sim \mathrm{Ber}\left[\Psi\left\{\alpha_{*g}\left(\tilde{\mu}+\tilde{e}_{gi}-\delta_{*g}\right)\right\}\right]$.

Lastly, define

$$X_i^{(m)} = \left(1 - r_{gi}^{(m)} \left[\Psi\left\{\alpha_{*g}\left(\mu + \tilde{e}_{gi} - \delta_{*g}\right)\right\}\right]^{-1}\right)^2 \left\{\left(c_z + \boldsymbol{v}_2^{\mathrm{T}} [\boldsymbol{\Xi}]_{i*}\right)^2 I\left(\boldsymbol{v}_2^{\mathrm{T}} [\boldsymbol{\Xi}]_{i*} \geq 0\right)\right.$$
$$\left. + \left(c_z - \boldsymbol{v}_2^{\mathrm{T}} [\boldsymbol{\Xi}]_{i*}\right)^2 I\left(\boldsymbol{v}_2^{\mathrm{T}} [\boldsymbol{\Xi}]_{i*} < 0\right)\right\}.$$

Clearly, $X_1^{(m)}, \ldots, X_n^{(m)}$ are independent and identically distributed and $X_i \leq X_i^{(m)}$ for all $i \in [n]$. We also see that

$$\begin{aligned}
&\mathbb{E}\left\{\left(r_{g1}^{(m)} \left[\Psi\left\{\alpha_{*g}\left(\mu + \tilde{e}_{g1} - \delta_{*g}\right)\right\}\right]^{-1}\right)^4\right\} \\
=&\mathbb{E}\left(\frac{\Psi\left\{\alpha_{*g}\left(\tilde{\mu} + \tilde{e}_{g1} - \delta_{*g}\right)\right\}}{\Psi\left\{\alpha_{*g}\left(\mu + \tilde{e}_{g1} - \delta_{*g}\right)\right\}} \left[\Psi\left\{\alpha_{*g}\left(\mu + \tilde{e}_{g1} - \delta_{*g}\right)\right\}\right]^{-3}\right) < \infty
\end{aligned}$$

because $\frac{\Psi\left\{\alpha_{*g}\left(\tilde{\mu}+\tilde{e}_{g1}-\delta_{*g}\right)\right\}}{\Psi\left\{\alpha_{*g}\left(\mu+\tilde{e}_{g1}-\delta_{*g}\right)\right\}}$ is bounded from above by Assumption 5.2. This then shows that $E\left\{X_1^{(m)}\right\} < \infty$. Therefore, for any $\eta > 0$,

$$\begin{aligned}
n^{-1} \sum_{i=1}^n \mathbb{E}\left\{X_i I\left(X_i \geq \eta n\right)\right\} &\leq n^{-1} \sum_{i=1}^n \mathbb{E}\left[X_i^{(m)} I\left\{X_i^{(m)} \geq \eta n\right\}\right] \\
&= \mathbb{E}\left[X_1^{(m)} I\left\{X_1^{(m)} \geq \eta n\right\}\right] \to 0 \text{ as } n \to \infty
\end{aligned}$$

by the dominated convergence theorem. This proves that

$$n^{1/2}\left[n\,\mathbb{V}\left\{\bar{\boldsymbol{G}}_g\left(\boldsymbol{\theta}_{*g}\right)\right\}\right]^{-1/2} \bar{\boldsymbol{G}}_g\left(\boldsymbol{\theta}_{*g}\right) \xrightarrow{\mathcal{D}} N_{t+\hat{K}}\left(\boldsymbol{0}_{t+\hat{K}}, I_{t+\hat{K}}\right)$$

as $n \to \infty$.

We use a standard truncation argument to show that $\|\bar{\boldsymbol{\Sigma}}_g\left(\boldsymbol{\theta}_{*g}\right) - n\,\mathbb{V}\left\{\bar{\boldsymbol{G}}_g\left(\boldsymbol{\theta}_{*g}\right)\right\}\|_2 = o_P(1)$. Let $\boldsymbol{v} = \left(\boldsymbol{v}_1^{\mathrm{T}}, \boldsymbol{v}_2^{\mathrm{T}}\right)^{\mathrm{T}} \in \mathbb{R}^{t+K}$ be a unit vector, where $\boldsymbol{v}_1 \in \mathbb{R}^t$ and $\boldsymbol{v}_2 \in \mathbb{R}^{\hat{K}}$, and let

X_i, $r_{gi}^{(m)}$ and $X_i^{(m)}$ be as defined above. We also define

$$Y_i = \{X_i - \mathbb{E}(X_i)\} I \left\{X_i^{(m)} \leq i\right\} \quad (i = 1, \ldots, n).$$

Since $X_1^{(m)}, X_2^{(m)}, \ldots$ are identically distributed and $\mathbb{E}\left\{X_1^{(m)}\right\} < \infty$,

$$\mathbb{P}\left[\bigcap_{N \geq 1} \bigcup_{n \geq N} \left\{X_n^{(m)} > n\right\}\right] = 0 \tag{5.32}$$

by Lemma 4.31 of Lalley (2015). We also have

$$\begin{aligned}|n^{-1} \sum_{i=1}^{n} \{X_i - \mathbb{E}(X_i)\}| \leq & |n^{-1} \sum_{i=1}^{n} I\left\{X_i^{(m)} > i\right\}| + |n^{-1} \sum_{i=1}^{n} \mathbb{E}(Y_i)| \\ & + |n^{-1} \sum_{i=1}^{n} \{Y_i - \mathbb{E}(Y_i)\}|.\end{aligned}$$

By (5.32), the first term is $o_{a.s.}(1)$ as $n \to \infty$. For the second term, we may assume $\mu_1 \leq \cdots \leq \mu_n$ without loss of generality. Define $r_{g1}^{(i)}$ inductively as

$$\begin{aligned}& r_{g1}^{(1)} = r_{g1} \\ & r_{g1}^{(i)} \begin{cases} = 1 & \text{if } r_{g1}^{(i-1)} = 1 \\ \sim \text{Ber}\left[\frac{\Psi\{\alpha_{*g}(\mu_i + \tilde{e}_{g1} - \delta_{*g})\} - \Psi\{\alpha_{*g}(\mu_{i-1} + \tilde{e}_{g1} - \delta_{*g})\}}{1 - \Psi\{\alpha_{*g}(\mu_{i-1} + \tilde{e}_{g1} - \delta_{*g})\}}\right] & \text{if } r_{g1}^{(i-1)} = 0 \end{cases} \quad (i = 2, \ldots, n).\end{aligned}$$

and let

$$\tilde{X}_i = \left(1 - r_{g1}^{(i)}\left[\Psi\left\{\alpha_{*g}\left(\mu_i + \tilde{e}_{g1} - \delta_{*g}\right)\right\}\right]^{-1}\right)^2 \left(\boldsymbol{v}_1^{\mathrm{T}}[\boldsymbol{Z}]_{i*} + \boldsymbol{v}_2^{\mathrm{T}}[\boldsymbol{\Xi}]_{1*}\right)^2.$$

And for $\|[\boldsymbol{Z}]_{i*}\|_2 \leq c_z$, define

$$r_{g1}^{(m)} = \begin{cases} = 1 & \text{if } r_{g1}^{(n)} = 1 \\ \sim \text{Ber}\left[\frac{\Psi\left\{\alpha_{*g}\left(\tilde{\mu}+\tilde{e}_{g1}-\delta_{*g}\right)\right\}-\Psi\left\{\alpha_{*g}\left(\mu_n+\tilde{e}_{g1}-\delta_{*g}\right)\right\}}{1-\Psi\left\{\alpha_{*g}\left(\mu_n+\tilde{e}_{g1}-\delta_{*g}\right)\right\}}\right] & \text{if } r_{g1}^{(n)} = 0 \end{cases}$$

$$\begin{aligned}\tilde{X}_1^{(m)} &= \left(1 - r_{g1}^{(m)}\left[\Psi\left\{\alpha_{*g}\left(\mu + \tilde{e}_{g1} - \delta_{*g}\right)\right\}\right]^{-1}\right)^2 \left\{\left(c_z + \boldsymbol{v}_2^{\mathrm{T}}\left[\boldsymbol{\Xi}\right]_{1*}\right)^2 I\left(\boldsymbol{v}_2^{\mathrm{T}}\left[\boldsymbol{\Xi}\right]_{1*} \geq 0\right)\right. \\ &\quad \left. + \left(c_z - \boldsymbol{v}_2^{\mathrm{T}}\left[\boldsymbol{\Xi}\right]_{1*}\right)^2 I\left(\boldsymbol{v}_2^{\mathrm{T}}\left[\boldsymbol{\Xi}\right]_{1*} < 0\right)\right\}.\end{aligned}$$

Note that

$$X_i I\left\{X_i^{(m)} \leq i\right\} \overset{\mathcal{D}}{=} \tilde{X}_i I\left\{X_1^{(m)} \leq i\right\} \leq X_1^{(m)} \quad (i = 1, \dots, n).$$

Since $X_1^{(m)}$ is integrable and $\mathbb{E}\left(X_i\right)$ is uniformly bounded from above, $|n^{-1} \sum_{i=1}^{n} \mathbb{E}\left(Y_i\right)| \to 0$ as $n \to \infty$ by the dominated convergence theorem. We lastly show $|n^{-1} \sum_{i=1}^{n} \left\{Y_i - \mathbb{E}\left(Y_i\right)\right\}| = o_{a.s.}(1)$ as $n \to \infty$ to complete the proof. By Kroneckers Lemma and the Khintchine-Kolmogorov theorem, it suffices to show $\sum_{n=1}^{\infty} n^{-2}\mathbb{E}\left(Y_n^2\right) < \infty$. And because $\mathbb{E}\left(X_i\right)$ is uniformly bounded and $X_i \leq X_i^{(m)}$, we need only show that

$$\sum_{n=1}^{\infty} n^{-2}\mathbb{E}\left[\left\{X_n^{(m)}\right\}^2 I\left\{X_n^{(m)} \leq n\right\}\right] < \infty.$$

However, this follows from the proof of Theorem 4.30 in Lalley (2015). $\square$

***Theorem* 5.2.** *Fix a $g \in \mathcal{S}$ and suppose Assumptions 5.1 and 5.2 hold for $t < 3$ and $\epsilon = 0$. Let $j_1, \dots, j_{3-t} \in \left[\hat{K}\right]$, where $\hat{K} \leq K$. For $w_{gi}\left(\boldsymbol{\theta}_g\right)$ defined in the statement of Lemma 5.5* *and* $\hat{\boldsymbol{C}}_2$ *defined in* (5.26), *let*

$$\boldsymbol{u}_{gi} = \left([\boldsymbol{Z}]_i^{\mathrm{T}}, \left[\hat{\boldsymbol{C}}_2\right]_{ij_1}, \cdots, \left[\hat{\boldsymbol{C}}_2\right]_{ij_{3-t}}\right)^{\mathrm{T}} \in \mathbb{R}^3 \quad (i = 1, \ldots, n)$$

$$\bar{\boldsymbol{G}}_g\left(\boldsymbol{\theta}_g\right) = n^{-1}\sum_{i=1}^{n} \boldsymbol{u}_{gi}\left\{1 - r_{gi}w_{gi}\left(\boldsymbol{\theta}_g\right)\right\}$$

$$\bar{\boldsymbol{\Sigma}}_g\left(\boldsymbol{\theta}_g\right) = n^{-1}\sum_{i=1}^{n}\left[\boldsymbol{u}_{gi}\left\{1 - r_{gi}w_{gi}\left(\boldsymbol{\theta}_g\right)\right\} - \bar{\boldsymbol{G}}_g\left(\boldsymbol{\theta}_g\right)\right]\left[\boldsymbol{u}_{gi}\left\{1 - r_{gi}w_{gi}\left(\boldsymbol{\theta}_g\right)\right\} - \bar{\boldsymbol{G}}_g\left(\boldsymbol{\theta}_g\right)\right]^{\mathrm{T}}.$$

Then for $\boldsymbol{\theta}_{*g} = \left(\alpha_{*g}, \delta_{*g}\right)^{\mathrm{T}}$,

$$n^{1/2}\left\{\bar{\boldsymbol{\Sigma}}_g\left(\boldsymbol{\theta}_{*g}\right)\right\}^{-1/2}\bar{\boldsymbol{G}}_g\left(\boldsymbol{\theta}_{*g}\right) \xrightarrow{\mathcal{D}} N_3\left(\boldsymbol{0}_3, I_3\right)$$

as $n, p \to \infty$.

Proof. As we did in Lemma 5.5, it suffices to re-define $\boldsymbol{u}_{gi} = \left([\boldsymbol{Z}]_i^{\mathrm{T}}, \left[\hat{\boldsymbol{C}}_2\right]_{i*}\right)^{\mathrm{T}}$. Let

$$\boldsymbol{D} = \operatorname{diag}\left\{1 - r_{g1}w_{g1}\left(\boldsymbol{\theta}_{*g}\right), \ldots, 1 - r_{gn}w_{gn}\left(\boldsymbol{\theta}_{*g}\right)\right\}$$

$$\boldsymbol{d} = \left(1 - r_{g1}w_{g1}\left(\boldsymbol{\theta}_{*g}\right), \ldots, 1 - r_{gn}w_{gn}\left(\boldsymbol{\theta}_{*g}\right)\right)^{\mathrm{T}} \in \mathbb{R}^n.$$

By Lemma 5.5, it suffices to show that

$$\|n^{-1/2}\boldsymbol{d}^{\mathrm{T}}\left(\hat{\boldsymbol{C}}_2 - \boldsymbol{C}_2\right)\|_2 = o_P(1) \tag{5.33a}$$

$$\|n^{-1}\hat{\boldsymbol{C}}_2^{\mathrm{T}}\boldsymbol{D}^2\hat{\boldsymbol{C}}_2 - n^{-1}\boldsymbol{C}_2^{\mathrm{T}}\boldsymbol{D}^2\boldsymbol{C}_2\|_2 = o_P(1) \tag{5.33b}$$

$$\|n^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{D}^2\hat{\boldsymbol{C}}_2 - n^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{D}^2\boldsymbol{C}_2\|_2 = o_P(1) \tag{5.33c}$$

to prove the theorem. By Assumption 5.2 and Lemma 5.4, $\|\boldsymbol{d}\|_2 = O_P\left(n^{1/2}\right)$ and $\|\boldsymbol{D}^2\|_2 = o_P\left(n^{1/2}\right)$. The latter follows from the fact under the assumptions on the left hand tail of

$\Psi(x)$,

$$\mathbb{E}\left[\left\{|1-r_{gi}w_{gi}\left(\boldsymbol{\theta}_{*g}\right)|\right\}^{4+\eta}\right] \leq 1+\mathbb{E}\left[\left\{r_{gi}w_{gi}\left(\boldsymbol{\theta}_{*g}\right)\right\}^{4+\eta}\right] \leq c$$

for $\eta > 0$ small enough and $c > 0$ large enough.

We start by showing (5.33a). By (5.26),

$$\|n^{-1/2}\boldsymbol{d}^{\mathrm{T}}\left(\hat{\boldsymbol{C}}_2 - \boldsymbol{C}_2\right)\|_2 \leq \|\boldsymbol{d}\|_2\|\hat{v} - I_K\|_2 + \|\boldsymbol{d}^{\mathrm{T}}\boldsymbol{Q}_{P_Z^{\perp}C}\hat{\boldsymbol{w}}\|_2.$$

The first term is $o_P(1)$ by (5.27a). And since $\boldsymbol{d}$ is independent of $\boldsymbol{E}_{\mathcal{S}^c}$, the second term is also $o_P(1)$ by (5.27a).

For (5.33b),

$$\begin{aligned}\|n^{-1}\hat{\boldsymbol{C}}_2^{\mathrm{T}}\boldsymbol{D}^2\hat{\boldsymbol{C}}_2 - n^{-1}\boldsymbol{C}_2^{\mathrm{T}}\boldsymbol{D}^2\boldsymbol{C}_2\|_2 \leq & \|n^{-1}\hat{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{C}_2^{\mathrm{T}}\boldsymbol{D}^2\boldsymbol{C}_2\hat{\boldsymbol{v}} - n^{-1}\boldsymbol{C}_2^{\mathrm{T}}\boldsymbol{D}^2\boldsymbol{C}_2\|_2 \\ & + 2\|n^{-1/2}\boldsymbol{C}_2^{\mathrm{T}}\boldsymbol{D}^2\boldsymbol{Q}_{P_Z^{\perp}C}\hat{\boldsymbol{w}}\|_2 \\ & + \|\hat{\boldsymbol{w}}^{\mathrm{T}}\boldsymbol{Q}_{P_Z^{\perp}C}^{\mathrm{T}}\boldsymbol{D}^2\boldsymbol{Q}_{P_Z^{\perp}C}\hat{\boldsymbol{w}}\|_2.\end{aligned}$$

The first and third terms are clearly $o_P(1)$ by (5.27a). And since $\|\boldsymbol{D}^2\|_2 = o_P\left(n^{1/2}\right)$, the second term is also $o_P(1)$. Identical techniques can be used to show (5.33c), which completes the proof. $\square$

5.10.7 The asymptotic distribution of the generalized method of moments estimator

Here we prove that under mild assumptions, the two-step generalized method of moments estimator, $\hat{\boldsymbol{\theta}}$, given by

$$\hat{\boldsymbol{\theta}}_g^{(1)} = \arg\min_{\boldsymbol{\theta}} \bar{\boldsymbol{G}}_g\left(\boldsymbol{\theta}\right)^{\mathrm{T}}\left(n^{-1}\hat{\boldsymbol{U}}_g^{\mathrm{T}}\hat{\boldsymbol{U}}_g\right)^{-1}\bar{\boldsymbol{G}}_g\left(\boldsymbol{\theta}\right) \quad \hat{\boldsymbol{\theta}}_g = \arg\min_{\boldsymbol{\theta}} \bar{\boldsymbol{G}}_g\left(\boldsymbol{\theta}\right)^{\mathrm{T}}\boldsymbol{W}_g\bar{\boldsymbol{G}}_g\left(\boldsymbol{\theta}\right),$$

is consistent and asymptotically normal, where $\bar{\boldsymbol{G}}_g$ and $\boldsymbol{W}_g$ were defined in (5.9). Our results are analogous to those in Wang et al. (2014), which assumes the instruments $\hat{\boldsymbol{U}}_g$ are observed. Our results are also easier to interpret, since the assumptions we make only involve the moments of $\boldsymbol{y}_g$ and the properties of the function $\Psi(x)$. We first make an assumption.

***Assumption* 5.3.** *Define* $\boldsymbol{u}_{gi} = \left([\boldsymbol{Z}]_{i*}, [\Xi]_{ij_1}, \ldots, [\Xi]_{ij_{3-t}} \right)^{\mathrm{T}}$ *and*

$$\boldsymbol{M}_g(\boldsymbol{\theta}) = -\nabla_{\boldsymbol{\theta}} \left(n^{-1} \sum_{i=1}^{n} \mathbb{E}\left[\boldsymbol{u}_{gi} \frac{r_{gi}}{\Psi\left\{ [\boldsymbol{\theta}]_1 \left(y_{gi} - [\boldsymbol{\theta}]_2 \right) \right\}} \right] \right)$$

for each $g \in \mathcal{S}$*. Then* $\boldsymbol{M}\left(\boldsymbol{\theta}_{*_g}\right)^{\mathrm{T}} \boldsymbol{M}\left(\boldsymbol{\theta}_{*_g}\right) \succeq \gamma I_2$ *for some constant* $\gamma > 0$ *that does not depend on* n *or* p*. Further,* ϵ*, defined in the Algorithm 3.1, is 0.*

***Remark* 5.2.** *We prove* $\boldsymbol{M}\left(\boldsymbol{\theta}_{*_g}\right)$ *exists in Lemma 5.9. This assumption on the gradient of the population moment is a standard assumption in the generalized method of moment literature (Hansen 1982, Wang et al. 2014) and helps to guarantee that* $\boldsymbol{\theta}_{*_g}$ *is locally identifiable.*

For notation purposes, we drop the subscript g for the remainder of the section. Next, we reparametrize $\bar{\boldsymbol{G}}(\boldsymbol{\theta})$ as

$$\bar{\boldsymbol{G}}(\boldsymbol{\theta}) = n^{-1} \sum_{i=1}^{n} \hat{\boldsymbol{u}}_i \left[1 - r_i \left\{ \Psi\left([\boldsymbol{\theta}]_1 y_i + [\boldsymbol{\theta}]_2 \right) \right\}^{-1} \right].$$

We also define $\dot{\Psi}(x)$ and $\ddot{\Psi}(x)$ to be the first and second derivatives of $\Psi(x)$. For any weight matrix $\boldsymbol{W}$, the generalized method of moments estimate, $\hat{\boldsymbol{\theta}}$, satisfies

$$\boldsymbol{0} = \boldsymbol{\Gamma}\left(\hat{\boldsymbol{\theta}}\right)^{\mathrm{T}} \boldsymbol{W} \bar{\boldsymbol{G}}\left(\hat{\boldsymbol{\theta}}\right) = \boldsymbol{\Gamma}\left(\hat{\boldsymbol{\theta}}\right)^{\mathrm{T}} \boldsymbol{W} \bar{\boldsymbol{G}}\left(\boldsymbol{\theta}_*\right) + \boldsymbol{\Gamma}\left(\hat{\boldsymbol{\theta}}\right)^{\mathrm{T}} \boldsymbol{W} \boldsymbol{\Gamma}\left(\tilde{\boldsymbol{\theta}}\right)\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*\right) \tag{5.34a}$$

$$\boldsymbol{\Gamma}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \bar{\boldsymbol{G}}(\boldsymbol{\theta}) = n^{-1} \sum_{i=1}^{n} r_i \frac{\dot{\Psi}\left([\boldsymbol{\theta}]_1 y_i + [\boldsymbol{\theta}]_2 \right)}{\Psi\left([\boldsymbol{\theta}]_1 y_i + [\boldsymbol{\theta}]_2 \right)^2} \hat{\boldsymbol{u}}_i \left(y_i, 1 \right), \tag{5.34b}$$

where $\tilde{\boldsymbol{\theta}} = b\boldsymbol{\theta}_* + (1-b)\hat{\boldsymbol{\theta}}$ for some $b \in [0,1]$. Since we have already proven that $\bar{\boldsymbol{G}}(\boldsymbol{\theta}_*)$ is asymptotically normal in Theorem 5.2, proving the asymptotic normality of $\hat{\boldsymbol{\theta}}$ requires understanding the convergence of $\mathbf{\Gamma}\left(\hat{\boldsymbol{\theta}}\right)$ and $\boldsymbol{W}$.

***Lemma* 5.6.** *Under Assumption 5.2, there exists a constant $M > 0$ such that $|\dot{\Psi}(x)/\Psi(x)|$, $|\ddot{\Psi}(x)/\Psi(x)| \leq M$ for all $x \in \mathbb{R}$.*

Proof. Since $|\dot{\Psi}(x)|, |\ddot{\Psi}(x)|$ are uniformly bounded, we need only consider the case when $x \to -\infty$. When $\Psi(x) = |x|^{-k}\{a + R(x)\}$,

$$\begin{aligned}
\dot{\Psi}(x) &= k|x|^{-(k+1)}\{a + R(x)\} + |x|^{-k}\frac{dR(x)}{dx} \\
\ddot{\Psi}(x) &= k(k+1)|x|^{-(k+2)}\{a + R(x)\} + 2k|x|^{-(k+1)}\frac{dR(x)}{dx} + |x|^{-k}\frac{d^2R(x)}{dx^2}
\end{aligned}$$

and when $\Psi(x) = \exp(-k|x|)\{a + R(x)\}$,

$$\begin{aligned}
\dot{\Psi}(x) &= k\exp(-k|x|)\{a + R(x)\} + \exp(-k|x|)\frac{dR(x)}{dx} \\
\ddot{\Psi}(x) &= k^2\exp(-k|x|)\{a + R(x)\} + 2k\exp(-k|x|)\frac{dR(x)}{dx} + \exp(-k|x|)\frac{d^2R(x)}{dx^2}.
\end{aligned}$$

The result then follows by the assumptions on $R(x)$. □

***Lemma* 5.7.** *Let $B(\eta; \boldsymbol{x}) = \{\boldsymbol{x}_0 : \|\boldsymbol{x} - \boldsymbol{x}_0\|_2 < \eta\}$ and suppose Assumptions 5.1 and 5.2 hold. Let $\boldsymbol{u}_i = \left([\boldsymbol{Z}]_{i*}, \Xi_{ij_1}, \ldots, \Xi_{ij_{3-t}}\right)$ for $j_1, \ldots, j_{3-t} \in [K]$ and define*

$$\begin{aligned}
&\tilde{\boldsymbol{G}}(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^{n}\boldsymbol{u}_i\left\{1 - \frac{r_i}{\Psi\left([\boldsymbol{\theta}]_1 y_i + [\boldsymbol{\theta}]_2\right)}\right\}, \quad \tilde{\mathbf{\Gamma}}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\tilde{\boldsymbol{G}}(\boldsymbol{\theta}) \\
&\tilde{\mathbf{\Sigma}}(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^{n}\left\{1 - \frac{r_i}{\Psi\left([\boldsymbol{\theta}]_1 y_i + [\boldsymbol{\theta}]_2\right)}\right\}^2\boldsymbol{u}_i\boldsymbol{u}_i^{\mathrm{T}}
\end{aligned} \tag{5.35}$$

Then for any $\gamma > 0$, there exists a constant $\eta > 0$ such that

$$\mathbb{E}\left\{\sup_{\boldsymbol{\theta}\in B(\eta;\boldsymbol{\theta}_*)} \|\tilde{\boldsymbol{G}}(\boldsymbol{\theta}) - \tilde{\boldsymbol{G}}(\boldsymbol{\theta}_*)\|_2\right\} \leq \gamma \tag{5.36a}$$

$$\mathbb{E}\left\{\sup_{\boldsymbol{\theta}\in B(\eta;\boldsymbol{\theta}_*)} \|\tilde{\boldsymbol{\Gamma}}(\boldsymbol{\theta}) - \tilde{\boldsymbol{\Gamma}}(\boldsymbol{\theta}_*)\|_2\right\} \leq \gamma \tag{5.36b}$$

$$\mathbb{E}\left\{\sup_{\boldsymbol{\theta}\in B(\eta;\boldsymbol{\theta}_*)} \|\tilde{\boldsymbol{\Sigma}}(\boldsymbol{\theta}) - \tilde{\boldsymbol{\Sigma}}(\boldsymbol{\theta}_*)\|_2\right\} \leq \gamma \tag{5.36c}$$

for all $n > 0$.

Proof. Fix $\eta > 0$, let $\boldsymbol{\theta} \in B(\eta;\boldsymbol{\theta}_*)$ and let $\boldsymbol{v} \in \mathbb{R}^3$ be any unit vector. Then for some set of $\tilde{\boldsymbol{\theta}}_i = b_i\boldsymbol{\theta} + (1-b_i)\boldsymbol{\theta}_*$, $b_i \in [0,1]$, and $(\epsilon_1, \epsilon_2)^{\mathrm{T}} = \boldsymbol{\theta} - \boldsymbol{\theta}_*$,

$$\begin{aligned}
|\boldsymbol{v}^{\mathrm{T}}\left\{\tilde{\boldsymbol{G}}(\boldsymbol{\theta}) - \tilde{\boldsymbol{G}}(\boldsymbol{\theta}_*)\right\}| =& |n^{-1}\sum_{i=1}^{n}(\epsilon_1 y_i + \epsilon_2)\frac{\dot{\Psi}\left(\left[\tilde{\boldsymbol{\theta}}_i\right]_1 y_i + \left[\tilde{\boldsymbol{\theta}}_i\right]_2\right)}{\Psi\left(\left[\tilde{\boldsymbol{\theta}}_i\right]_1 y_i + \left[\tilde{\boldsymbol{\theta}}_i\right]_2\right)}\frac{r_i}{\Psi\left(\left[\tilde{\boldsymbol{\theta}}_i\right]_1 y_i + \left[\tilde{\boldsymbol{\theta}}_i\right]_2\right)}\boldsymbol{v}^{\mathrm{T}}\boldsymbol{u}_i| \\
\leq & M|\epsilon_1|n^{-1}\sum_{i=1}^{n}\frac{r_i}{\Psi\left(\left[\tilde{\boldsymbol{\theta}}_i\right]_1 y_i + \left[\tilde{\boldsymbol{\theta}}_i\right]_2\right)}|\boldsymbol{v}^{\mathrm{T}}\boldsymbol{u}_i| \\
& + M|\epsilon_2|n^{-1}\sum_{i=1}^{n}\frac{r_i}{\Psi\left(\left[\tilde{\boldsymbol{\theta}}_i\right]_1 y_i + \left[\tilde{\boldsymbol{\theta}}_i\right]_2\right)}|y_i\boldsymbol{v}^{\mathrm{T}}\boldsymbol{u}_i|
\end{aligned}$$

where $M > 0$ is defined in Lemma 5.6. (5.36a) then follows easily by Assumptions 5.1 and 5.2.

For (5.36b),

$$\begin{aligned}
\tilde{\boldsymbol{\Gamma}}(\boldsymbol{\theta}) - \tilde{\boldsymbol{\Gamma}}(\boldsymbol{\theta}_*) =& n^{-1}\sum_{i=1}^{n} r_i(\epsilon_1 y_i + \epsilon_2)\left\{-2\frac{\dot{\Psi}\left(\left[\tilde{\boldsymbol{\theta}}_i\right]_1 y_i + \left[\tilde{\boldsymbol{\theta}}_i\right]_2\right)^2}{\Psi\left(\left[\tilde{\boldsymbol{\theta}}_i\right]_1 y_i + \left[\tilde{\boldsymbol{\theta}}_i\right]_2\right)^3}\right. \\
& \left. + \frac{\ddot{\Psi}\left(\left[\tilde{\boldsymbol{\theta}}_i\right]_1 y_i + \left[\tilde{\boldsymbol{\theta}}_i\right]_2\right)}{\Psi\left(\left[\tilde{\boldsymbol{\theta}}_i\right]_1 y_i + \left[\tilde{\boldsymbol{\theta}}_i\right]_2\right)^2}\right\}\boldsymbol{u}_i(y_i, 1)
\end{aligned}$$

where ϵ_1, ϵ_2 are defined above and $\tilde{\boldsymbol{\theta}}_i = b_i\boldsymbol{\theta} + (1-b_i)\boldsymbol{\theta}_*$ for some $b_i \in [0,1]$. To prove (5.36b), it suffices to show that

$$n^{-1}\sum_{i=1}^{n} y_i^2 \frac{r_i}{\Psi\left(\left[\tilde{\boldsymbol{\theta}}_i\right]_1 y_i + \left[\tilde{\boldsymbol{\theta}}_i\right]_2\right)} \boldsymbol{u}_i$$

has at most finite expectation by Lemma 5.6. However, this follows because the entries of $\boldsymbol{u}_i$ have uniformly bounded sixth moment.

Using the same notation as above, we can express (5.36c) as

$$\begin{aligned}\tilde{\boldsymbol{\Sigma}}\left(\boldsymbol{\theta}\right) - \tilde{\boldsymbol{\Sigma}}\left(\boldsymbol{\theta}_*\right) =& 2n^{-1}\sum_{i=1}^{n} \left(\epsilon_1 y_i + \epsilon_2\right) r_i \left\{ \frac{\dot{\Psi}\left(\left[\tilde{\boldsymbol{\theta}}_i\right]_1 y_i + \left[\tilde{\boldsymbol{\theta}}_i\right]_2\right)}{\Psi\left(\left[\tilde{\boldsymbol{\theta}}_i\right]_1 y_i + \left[\tilde{\boldsymbol{\theta}}_i\right]_2\right)^2} \right. \\ & \left. - \frac{\dot{\Psi}\left(\left[\tilde{\boldsymbol{\theta}}_i\right]_1 y_i + \left[\tilde{\boldsymbol{\theta}}_i\right]_2\right)}{\Psi\left(\left[\tilde{\boldsymbol{\theta}}_i\right]_1 y_i + \left[\tilde{\boldsymbol{\theta}}_i\right]_2\right)^3} \right\} \boldsymbol{u}_i \boldsymbol{u}_i^{\mathrm{T}}.\end{aligned}$$

Again, by Lemma 5.6, it suffices to show that

$$n^{-1}\sum_{i=1}^{n} \frac{r_i}{\Psi\left(\left[\tilde{\boldsymbol{\theta}}_i\right]_1 y_i + \left[\tilde{\boldsymbol{\theta}}_i\right]_2\right)^2} |y_i| \boldsymbol{u}_i \boldsymbol{u}_i^{\mathrm{T}}$$

has bounded expectation. However, this follows by the bounded sixth moment assumption on the entries of $\boldsymbol{u}_i$ and Assumption 5.2. $\square$

***Lemma* 5.8.** *Suppose the assumptions of Lemma 5.7 hold. Then for $\eta > 0$ small enough,*

$$\sup_{\boldsymbol{\theta}\in B(\eta;\boldsymbol{\theta}_*)} \|\bar{\boldsymbol{G}}\left(\boldsymbol{\theta}\right) - \tilde{\boldsymbol{G}}\left(\boldsymbol{\theta}\right)\|_2, \quad \sup_{\boldsymbol{\theta}\in B(\eta;\boldsymbol{\theta}_*)} \|\boldsymbol{\Gamma}\left(\boldsymbol{\theta}\right) - \tilde{\boldsymbol{\Gamma}}\left(\boldsymbol{\theta}\right)\|_2, \quad \sup_{\boldsymbol{\theta}\in B(\eta;\boldsymbol{\theta}_*)} \|\bar{\boldsymbol{\Sigma}}\left(\boldsymbol{\theta}\right) - \tilde{\boldsymbol{\Sigma}}\left(\boldsymbol{\theta}\right)\|_2 = o_P(1)$$

as $n, p \to \infty$.

Proof. Define

$$
\begin{aligned}
d_1(\boldsymbol{\theta}) &= \left(1 - \frac{r_1}{\Psi\left([\boldsymbol{\theta}]_1 y_1 + [\boldsymbol{\theta}]_2\right)}, \ldots, 1 - \frac{r_n}{\Psi\left([\boldsymbol{\theta}]_1 y_n + [\boldsymbol{\theta}]_2\right)}\right)^{\mathrm{T}} \\
d_2(\boldsymbol{\theta}) &= \left(1 - \frac{r_1 \dot{\Psi}\left([\boldsymbol{\theta}]_1 y_1 + [\boldsymbol{\theta}]_2\right)}{\Psi\left([\boldsymbol{\theta}]_1 y_1 + [\boldsymbol{\theta}]_2\right)^2}, \ldots, 1 - \frac{r_n \dot{\Psi}\left([\boldsymbol{\theta}]_1 y_n + [\boldsymbol{\theta}]_2\right)}{\Psi\left([\boldsymbol{\theta}]_1 y_n + [\boldsymbol{\theta}]_2\right)^2}\right)^{\mathrm{T}} \\
\boldsymbol{D}(\boldsymbol{\theta}) &= \operatorname{diag}\left[\left\{1 - \frac{r_1}{\Psi\left([\boldsymbol{\theta}]_1 y_1 + [\boldsymbol{\theta}]_2\right)}\right\}^2, \ldots, \left\{1 - \frac{r_n}{\Psi\left([\boldsymbol{\theta}]_1 y_n + [\boldsymbol{\theta}]_2\right)}\right\}^2\right].
\end{aligned}
$$

By Assumption 5.2 and Lemma 5.4, there exists a small constant $\eta > 0$ such that

$$
\| \sup_{\boldsymbol{\theta} \in B(\eta; \boldsymbol{\theta}_*)} \boldsymbol{d}_j(\boldsymbol{\theta}) \|_2 = O_P\left(n^{1/2}\right), \quad \| \sup_{\boldsymbol{\theta} \in B(\eta; \boldsymbol{\theta}_*)} \boldsymbol{D}(\boldsymbol{\theta}) \|_2 = o_P\left(n^{1/2}\right)
$$

for $j = 1, 2$. The latter follows from the fact that

$$
n^{-1} \sum_{i=1}^{n} \mathbb{E}\left[|1 - \frac{r_i}{\Psi\left\{\left([\boldsymbol{\theta}_*]_1 + \eta\right) y_i + \left([\boldsymbol{\theta}_*]_2 - \eta\right)\right\}}|^{2+\gamma} I\left(y_i < 0\right)\right] \leq c
$$

for some constants $\eta, \gamma > 0$ small enough and $c > 0$ large enough by Assumption 5.2 and Lemma 5.4. Let $\boldsymbol{a}_j \in \mathbb{R}^n$ be the jth standard basis vector, $\boldsymbol{A} = \left(\boldsymbol{a}_{j_1} \cdots \boldsymbol{a}_{j_{3-t}}\right)$ and $\boldsymbol{M} = (\boldsymbol{Z}\, \boldsymbol{\Xi})$. Then by Lemma 5.2,

$$
\begin{aligned}
\bar{\boldsymbol{G}}(\boldsymbol{\theta}) - \tilde{\boldsymbol{G}}(\boldsymbol{\theta}) =& n^{-1} \boldsymbol{A}^{\mathrm{T}} \boldsymbol{R}^{\mathrm{T}} \boldsymbol{M}^{\mathrm{T}} \boldsymbol{d}_1(\boldsymbol{\theta}) + n^{-1/2} \boldsymbol{A}^{\mathrm{T}} \hat{\boldsymbol{w}}^{\mathrm{T}} \boldsymbol{Q}_{P_Z^{\perp} C}^{\mathrm{T}} \boldsymbol{d}_1(\boldsymbol{\theta}) \\
\boldsymbol{\Gamma}(\boldsymbol{\theta}) - \tilde{\boldsymbol{\Gamma}}(\boldsymbol{\theta}) =& n^{-1} \boldsymbol{A}^{\mathrm{T}} \boldsymbol{R}^{\mathrm{T}} \boldsymbol{M}^{\mathrm{T}} \boldsymbol{d}_2(\boldsymbol{\theta}) + n^{-1/2} \boldsymbol{A}^{\mathrm{T}} \hat{\boldsymbol{w}}^{\mathrm{T}} \boldsymbol{Q}_{P_Z^{\perp} C}^{\mathrm{T}} \boldsymbol{d}_2(\boldsymbol{\theta}) \\
\bar{\boldsymbol{\Sigma}}(\boldsymbol{\theta}) - \tilde{\boldsymbol{\Sigma}}(\boldsymbol{\theta}) =& n^{-1} \boldsymbol{A}^{\mathrm{T}} \boldsymbol{R}^{\mathrm{T}} \boldsymbol{M}^{\mathrm{T}} \boldsymbol{D}(\boldsymbol{\theta}) \boldsymbol{M} \boldsymbol{A} + n^{-1} \boldsymbol{A}^{\mathrm{T}} \boldsymbol{M}^{\mathrm{T}} \boldsymbol{D}(\boldsymbol{\theta}) \boldsymbol{M} \boldsymbol{R} \boldsymbol{A} \\
&+ n^{-1} \boldsymbol{A}^{\mathrm{T}} \boldsymbol{R}^{\mathrm{T}} \boldsymbol{M}^{\mathrm{T}} \boldsymbol{D}(\boldsymbol{\theta}) \boldsymbol{M} \boldsymbol{R} \boldsymbol{A} + n^{-1/2} \boldsymbol{A}^{\mathrm{T}} \hat{\boldsymbol{w}}^{\mathrm{T}} \boldsymbol{Q}_{P_Z^{\perp} C}^{\mathrm{T}} \boldsymbol{D}(\boldsymbol{\theta}) \hat{\boldsymbol{U}}
\end{aligned}
$$

where $\|\boldsymbol{R}\|_2 = o_P(1)$. Since $\|\hat{\boldsymbol{w}}\|_2 = O_P\left(n^{-1/2}\right)$ and $\|n^{-1/2} \boldsymbol{M}\|_2 = O_P(1)$ and by the properties of $\boldsymbol{d}_1(\boldsymbol{\theta})$, $\boldsymbol{d}_2(\boldsymbol{\theta})$ discussed above

$$\sup_{\boldsymbol{\theta}\in B(\eta;\boldsymbol{\theta}_*)} \|\bar{\boldsymbol{G}}(\boldsymbol{\theta}) - \tilde{\boldsymbol{G}}(\boldsymbol{\theta})\|_2, \sup_{\boldsymbol{\theta}\in B(\eta;\boldsymbol{\theta}_*)} \|\boldsymbol{\Gamma}(\boldsymbol{\theta}) - \tilde{\boldsymbol{\Gamma}}(\boldsymbol{\theta})\|_2 = o_P(1).$$

By Lemmas 5.5 and 5.7,

$$\sup_{\boldsymbol{\theta}\in B(\eta;\boldsymbol{\theta}_*)} \|n^{-1}\boldsymbol{M}^{\mathrm{T}}\boldsymbol{D}(\boldsymbol{\theta})\boldsymbol{M}\|_2 = O_P(1).$$

This completes the proof. □

***Lemma* 5.9.** *Suppose the assumptions of Lemma 5.7 hold and let $\boldsymbol{M}(\boldsymbol{\theta}) = \boldsymbol{M}_g(\boldsymbol{\theta})$ be as defined in Assumption 5.3. Then the following hold*

(i) There exists a constant $\eta > 0$ such that $\mathbb{E}\left\{\tilde{\boldsymbol{G}}(\boldsymbol{\theta})\right\}$ and $\boldsymbol{M}(\boldsymbol{\theta})$ exist and are continuous for $\boldsymbol{\theta} \in B(\eta;\boldsymbol{\theta}_*)$.

(ii) For all $\gamma_1, \gamma_2 > 0$, there exists a constant $\eta > 0$ such that

$$\mathbb{P}\left[\sup_{\boldsymbol{\theta}\in B(\eta;\boldsymbol{\theta}_*)} \|\bar{\boldsymbol{G}}(\boldsymbol{\theta}) - \mathbb{E}\left\{\tilde{\boldsymbol{G}}(\boldsymbol{\theta})\right\}\|_2 \geq \gamma_1\right] \leq \gamma_2$$
$$\mathbb{P}\left\{\sup_{\boldsymbol{\theta}\in B(\eta;\boldsymbol{\theta}_*)} \|\boldsymbol{\Gamma}(\boldsymbol{\theta}) - \boldsymbol{M}(\boldsymbol{\theta})\|_2 \geq \gamma_1\right\} \leq \gamma_2$$

for all n and p large enough.

Proof. The existence of $\mathbb{E}\left\{\tilde{\boldsymbol{G}}(\boldsymbol{\theta})\right\}$ follows from Lemma 5.7, and the existence of $\boldsymbol{M}(\boldsymbol{\theta})$ follows by a simple application of the dominated convergence theorem. The continuity of $\boldsymbol{M}(\boldsymbol{\theta})$ also follows from Lemma 5.7.

For (ii), Assumptions 5.1 and 5.2 show that $\mathbb{V}\left\{\tilde{\boldsymbol{\Gamma}_*}(\boldsymbol{\theta})\right\}$ and $\mathbb{V}\left\{\tilde{\boldsymbol{G}}(\boldsymbol{\theta}_*)\right\}$ are $o(1)$ as $n \to \infty$. The remainder of follows from Lemmas 5.7 and 5.8. □

***Theorem* 5.3.** *Suppose Assumptions 5.1, 5.2 and 5.3 hold. Define*

$$f(\boldsymbol{\theta}) = \bar{\boldsymbol{G}}(\boldsymbol{\theta})^{\mathrm{T}} \left(n^{-1}\hat{\boldsymbol{U}}^{\mathrm{T}}\hat{\boldsymbol{U}}\right)^{-1} \bar{\boldsymbol{G}}(\boldsymbol{\theta})$$

and let $\left\{\hat{\boldsymbol{\theta}}_n^{(1)}\right\}_{n\geq 1}$ *be a sequence of minima of* $f(\boldsymbol{\theta})$*. Next, define*

$$\boldsymbol{W}_n^{-1} = \bar{\boldsymbol{\Sigma}}\left\{\hat{\boldsymbol{\theta}}_n^{(1)}\right\} = n^{-1}\sum_{i=1}^{n}\left\{1 - \frac{r_i}{\Psi\left(\left[\hat{\boldsymbol{\theta}}_n^{(1)}\right]_1 y_i + \left[\hat{\boldsymbol{\theta}}_n^{(1)}\right]_2\right)}\right\}^2 \hat{\boldsymbol{u}}_i\hat{\boldsymbol{u}}_i^{\mathrm{T}}.$$

Then there exists a sequence $\left\{\hat{\boldsymbol{\theta}}_n\right\}_{n\geq 1}$ *of minima of* $\bar{\boldsymbol{G}}(\boldsymbol{\theta})^{\mathrm{T}}\boldsymbol{W}_n\bar{\boldsymbol{G}}(\boldsymbol{\theta})$ *such that* (5.8) *holds as* $n, p \to \infty$*.*

Proof. By Assumption 5.3 and (i) of Lemma 5.9,

$$\|\mathbb{E}\left\{\tilde{\boldsymbol{G}}(\boldsymbol{\theta})\right\}\|_2 \geq c\|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|_2$$

for some constant $c > 0$ and $\boldsymbol{\theta} \in B(\eta; \boldsymbol{\theta}_*)$ for some $\eta > 0$. By Lemma 5.9, this implies there exists a minimizer $\hat{\boldsymbol{\theta}}_n^{(1)}$ of f defined above such that $\|\hat{\boldsymbol{\theta}}_n^{(1)} - \boldsymbol{\theta}_*\|_2 = o_P(1)$ as $n, p \to \infty$. Therefore

$$\|\bar{\boldsymbol{\Sigma}}\left\{\hat{\boldsymbol{\theta}}_n^{(1)}\right\} - \mathbb{E}\left\{\tilde{\boldsymbol{\Sigma}}(\boldsymbol{\theta}_*)\right\}\|_2,\ \|\boldsymbol{\Gamma}\left\{\hat{\boldsymbol{\theta}}_n^{(1)}\right\} - \boldsymbol{M}(\boldsymbol{\theta}_*)\|_2 = o_P(1)$$

by Lemmas 5.5, 5.7 and 5.8. By Lemma 5.5, there exists a minimizer $\hat{\boldsymbol{\theta}}_n$ of $\bar{\boldsymbol{G}}(\boldsymbol{\theta})^{\mathrm{T}}\boldsymbol{W}_n\bar{\boldsymbol{G}}(\boldsymbol{\theta})$ such that $\|\hat{\boldsymbol{\theta}}_n^{(1)} - \boldsymbol{\theta}_*\|_2 = o_P(1)$ as $n, p \to \infty$. Then by Lemmas 5.7 and 5.8 and for $\tilde{\boldsymbol{\theta}}$ defined in (5.34),

$$\|\boldsymbol{\Gamma}\left(\tilde{\boldsymbol{\theta}}\right) - \boldsymbol{M}(\boldsymbol{\theta}_*)\|_2 = o_P(1)$$

as $n, p \to \infty$. The reuslt then follows. $\square$

***Corollary* 5.2.** *Suppose the assumptions of Theorem 5.3 hold. Then for $\boldsymbol{W}_n$ and $\hat{\boldsymbol{\theta}}_n$ defined in the statement of Theorem 5.3,*

$$n\bar{\boldsymbol{G}}\left(\hat{\boldsymbol{\theta}}_n\right)^{\mathrm{T}}\boldsymbol{W}_n\bar{\boldsymbol{G}}\left(\hat{\boldsymbol{\theta}}_n\right)\xrightarrow{\mathcal{D}}\chi_1^2$$

as $n,p\to\infty$.

Proof. Let $\tilde{\boldsymbol{W}}_n=\mathbb{E}\left\{\tilde{\boldsymbol{\Sigma}}\left(\boldsymbol{\theta}_*\right)\right\}$. Then for

$$\hat{\boldsymbol{A}}_n=P^{\perp}_{W_n^{1/2}\Gamma\left(\hat{\theta}_n\right)}=I_3-\boldsymbol{W}_n^{1/2}\boldsymbol{\Gamma}\left(\hat{\boldsymbol{\theta}}_n\right)\left\{\boldsymbol{\Gamma}\left(\hat{\boldsymbol{\theta}}_n\right)^{\mathrm{T}}\boldsymbol{W}_n\boldsymbol{\Gamma}\left(\hat{\boldsymbol{\theta}}_n\right)\right\}^{-1}\boldsymbol{\Gamma}\left(\hat{\boldsymbol{\theta}}_n\right)^{\mathrm{T}}\boldsymbol{W}_n^{1/2}$$
$$\tilde{\boldsymbol{A}}_n=P^{\perp}_{\tilde{W}_n^{1/2}M(\theta_*)}=I_3-\tilde{\boldsymbol{W}}_n^{1/2}\boldsymbol{M}\left(\boldsymbol{\theta}_*\right)\left\{\boldsymbol{M}\left(\boldsymbol{\theta}_*\right)^{\mathrm{T}}\tilde{\boldsymbol{W}}_n\boldsymbol{M}\left(\boldsymbol{\theta}_*\right)\right\}^{-1}\boldsymbol{M}\left(\boldsymbol{\theta}_*\right)^{\mathrm{T}}\tilde{\boldsymbol{W}}_n^{1/2},$$

$\|\hat{\boldsymbol{A}}_n-\tilde{\boldsymbol{A}}_n\|_2=o_P(1)$ by Lemmas 5.5, 5.7 and 5.8. Further, $\tilde{\boldsymbol{A}}_n$ is a non-random, rank 1 matrix for all n,p large enough by Assumption 5.3 and Lemma 5.9. We then get that

$$\begin{aligned}n\bar{\boldsymbol{G}}\left(\hat{\boldsymbol{\theta}}_n\right)^{\mathrm{T}}\boldsymbol{W}_n\bar{\boldsymbol{G}}\left(\hat{\boldsymbol{\theta}}_n\right)=&n\bar{\boldsymbol{G}}\left(\hat{\boldsymbol{\theta}}_n\right)^{\mathrm{T}}\boldsymbol{W}_n^{1/2}P_{W_n^{1/2}\Gamma\left(\hat{\theta}_n\right)}\boldsymbol{W}_n^{1/2}\bar{\boldsymbol{G}}\left(\hat{\boldsymbol{\theta}}_n\right)\\&+n\bar{\boldsymbol{G}}\left(\hat{\boldsymbol{\theta}}_n\right)^{\mathrm{T}}\boldsymbol{W}_n^{1/2}\hat{\boldsymbol{A}}_n\boldsymbol{W}_n^{1/2}\bar{\boldsymbol{G}}\left(\hat{\boldsymbol{\theta}}_n\right),\end{aligned}$$

where the first term is zero because $\hat{\boldsymbol{\theta}}_n$ is such that $\boldsymbol{\Gamma}\left(\hat{\boldsymbol{\theta}}_n\right)^{\mathrm{T}}\boldsymbol{W}_n\bar{\boldsymbol{G}}\left(\hat{\boldsymbol{\theta}}_n\right)=\boldsymbol{0}$. For the second term,

$$n^{1/2}\hat{\boldsymbol{A}}_n\boldsymbol{W}_n^{1/2}\bar{\boldsymbol{G}}\left(\hat{\boldsymbol{\theta}}_n\right)=n^{1/2}\hat{\boldsymbol{A}}_n\boldsymbol{W}_n^{1/2}\bar{\boldsymbol{G}}\left(\boldsymbol{\theta}_*\right)+n^{1/2}\hat{\boldsymbol{A}}_n\boldsymbol{W}_n^{1/2}\boldsymbol{\Gamma}\left(\tilde{\boldsymbol{\theta}}\right)\left(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_*\right)$$

for some $\tilde{\boldsymbol{\theta}}=b\boldsymbol{\theta}_*+(1-b)\hat{\boldsymbol{\theta}}_n$, $b\in[0,1]$. The result follows because $n^{1/2}\|\bar{\boldsymbol{G}}\left(\boldsymbol{\theta}_*\right)\|_2, n^{1/2}\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_*\|_2=O_P(1)$, $\|\boldsymbol{\Gamma}\left(\tilde{\boldsymbol{\theta}}\right)-\boldsymbol{\Gamma}\left(\hat{\boldsymbol{\theta}}_n\right)\|_2=o_P(1)$ and $\hat{\boldsymbol{A}}_n\boldsymbol{W}_n^{1/2}\boldsymbol{\Gamma}\left(\hat{\boldsymbol{\theta}}_n\right)=\boldsymbol{0}$. □

SOFTWARE

The R packages and code to implement the methods presented in Chapters 2, 3, 4 and 5 are available from https://github.com/chrismckennan. The repositories include:

- /BCconf, which contains the R package BCconf that implements Algorithm 2.2, as well as the code necessary to reproduce all simulations from Chapter 2.
- /CorrConf, which contains the R package CorrConf that implements Algorithm 3.2 (ICaSE), Algorithm 3.3 (CBCV) and the estimator for $\boldsymbol{C}$ given by (3.12) in Chapter 3.
- /LongitudinalAncestry, which contains the summary statistics and code necessary to estimate $\pi_{(0,0)}, \pi_{(1,0)}^{(1)}, \ldots, \pi_{(1,0)}^{(K)}, \pi_{(0,1)}^{(1)}, \ldots, \pi_{(0,1)}^{(K)}, \pi_{(1,1)}^{(1,1)}, \pi_{(1,1)}^{(S,K)}$ from Model (4.3b), as well as the conserved and discordant sign rates defined given by (4.4) in Chapter 4.
- /MetabMiss, which contains the R package MetabMiss that implements Algorithm 5.1 from Chapter 5.

REFERENCES

Adkins, R. M., Krushkal, J., Tylavsky, F. A. & Thomas, F. (2011), 'Racial differences in gene-specific DNA methylation levels are present at birth', *Birth Defects Research Part A: Clinical and Molecular Teratology* **91**(8), 728–736.

Ai, C., Linton, O. & Zhang, Z. (2018), 'A simple and efficient estimation method for models with nonignorable missing data', *arXiv:1801.04202* .

Amemiya, T. (1974), 'The nonlinear two-stage least-squares estimator', *Journal of Econometrics* **2**(2), 105–110.

Amemiya, T. (1977), 'The maximum likelihood and the nonlinear three-stage least squares estimator in the general nonlinear simultaneous equation model', *Econometrica* **45**(4), 955–968.

Anderson, T. W. & Rubin, H. (1956), Statistical inference in factor analysis, *in* 'Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 5: Contributions to Econometrics, Industrial Research, and Psychometry', University of California Press, Berkeley, California, pp. 111–150.

Auffinger, A. & Tang, S. (2015), 'Extreme eigenvalues of sparse, heavy tailed random matrices', *arXiv:1506.06175v1* .

Bai, J. (2009), 'Panel data models with interactive fixed effects', *Econometrica* **77**(4), 1229–1279.

Bai, J. & Li, K. (2012), 'Statistical analysis of factor models of high dimension', *The Annals of Statistics* **40**(1), 436–465.

Bai, J. & Ng, S. (2002), 'Determining the number of factors in approximate factor models', *Econometrica* **70**(1), 191–221.

Bai, T. R. & Knight, D. A. (2005), 'Structural changes in the airways in asthma: observations and consequences', *Clinical Science* **108**(6), 463.

Balatsinou, L., Gioacchino, G. D., Sabatino, G., Cavallucci, E., Caruso, R., Gabriele, E., Ramondo, S., Giampaolo, L. D., Verna, N. & Gioacchino, M. D. (2004), 'Asthma worsened by benzoate contained in some antiasthmatic drugs', *International Journal of Immunopathology and Pharmacology* **17 2**, 225–6.

Baumgart, M., Priebe, S., Groth, M., Hartmann, N., Menzel, U., Pandolfini, L., Koch, P., Felder, M., Ristow, M., Englert, C., Guthke, R., Platzer, M. & Cellerino, A. (2016), 'Longitudinal RNA-Seq analysis of vertebrate aging identifies mitochondrial complex I as a small-molecule-sensitive modifier of lifespan', *Cell Systems* **2**(2), 122–132.

Bell, J. T., Pai, A. A., Pickrell, J. K., Gaffney, D. J., Pique-Regi, R., Degner, J. F., Gilad, Y. & Pritchard, J. K. (2011), 'DNA methylation patterns associate with genetic and gene expression variation in hapmap cell lines', *Genome Biology* **12**(1), R10.

Benjamini, Y. & Hochberg, Y. (1995), 'Controlling the false discovery rate: A practical and powerful approach to multiple testing', *Journal of the Royal Statistical Society: Series B* **57**(1), 289–300.

Bisgaard, H., Vissing, N. H., Carson, C. G., Bischoff, A. L., Følsgaard, N. V., Kreiner-Møller, E., Chawes, B. L. K., Stokholm, J., Pedersen, L., Bjarnadóttir, E., Thysen, A. H., Nilsson, E., Mortensen, L. J., Olsen, S. F., Schjørring, S., Krogfelt, K. A., Lauritzen, L., Brix, S. & Bønnelykke, K. (2013), 'Deep phenotyping of the unselected copsac2010 birth cohort study', *Clinical and Experimental Allergy* **43**(12), 1384–1394.

Bocklandt, S., Lin, W., Sehl, M. E., Sánchez, F. J., Sinsheimer, J. S., Horvath, S. & Vilain, E. (2011), 'Epigenetic predictor of age', *PLOS ONE* **6**(6), e14821–.

Breton, C. V., Marsit, C. J., Faustman, E., Nadeau, K., Goodrich, J. M., Dolinoy, D. C., Herbstman, J., Holland, N., LaSalle, J. M., Schmidt, R., Yousefi, P., Perera, F., Joubert, B. R., Wiemels, J., Taylor, M., Yang, I. V., Chen, R., Hew, K. M., Freeland, D. M. H., Miller, R. & Murphy, S. K. (2017), 'Small-magnitude effect sizes in epigenetic end points are important in children's environmental health studies: The children's environmental health and disease prevention research center's epigenetics working group', *Environmental Health Perspectives* **125**(4), 511–526.

Brown, B. W. & Newey, W. K. (2002), 'Generalized method of moments, efficient bootstrapping, and improved inference', *Journal of Business & Economic Statistics* **20**(4), 507–517.

Brown, T. A. (2015), *Confirmatory Factor Analysis for Applied Research*, Guilford Publications.

Buja, A. & Eyuboglu, N. (1992), 'Remarks on parallel analysis', *Multivariate Behavioral Research* **27**(4), 509–540.

C., K. D., Michele, A.-W., Andres, H. E., R., K. M. & J., M. C. (2013), 'Differential DNA methylation in umbilical cord blood of infants exposed to low levels of arsenic in utero', *Environmental Health Perspectives* **121**(8), 971–977.

Cangelosi, R. & Goriely, A. (2007), 'Component retention in principal component analysis with application to cDNA microarray data', *Biology Direct* **2**(1), 2.

Chan, M. A., Ciaccio, C. E., Gigliotti, N. M., Rezaiekhaligh, M., Siedlik, J. A., Kennedy, K. & Barnes, C. S. (2017), 'DNA methylation levels associated with race and childhood asthma severity', *Journal of Asthma* **54**(8), 825–832.

Chatterton, Z., Hartley, B. J., Seok, M.-H., Mendelev, N., Chen, S., Milekic, M., Rosoklija, G., Stankov, A., Trencevsja-Ivanovska, I., Brennand, K., Ge, Y., Dwork, A. J. & Haghighi, F. (2017), 'In utero exposure to maternal smoking is associated with DNA methylation alterations and reduced neuronal content in the developing fetal brain', *Epigenetics & Chromatin* **10**, 4; 4–4.

Chen, C., Gonzalez, F. J. & Idle, J. R. (2007), 'LC-MS-based metabolomics in drug metabolism', *Drug Metabolism Reviews* **39**(2-3), 581–597.

Chen, L. S., Wang, J., Wang, X. & Wang, P. (2017), 'A mixed-effects model for incomplete data from labeling-based quantitative proteomics experiments', *The Annals of Applied Statistics* **11**(1), 114–138.

Davey Smith, G., Tilling, K., Suderman, M., Ring, S. M., Gaunt, T. R., Simpkin, A. J., Relton, C. L., Lyttleton, O. & McArdle, W. L. (2015), 'Longitudinal analysis of DNA methylation associated with birth weight and gestational age', *Human Molecular Genetics* **24**(13), 3752–3763.

De Livera, A. M., Dias, D. A., De Souza, D., Rupasinghe, T., Pyke, J., Tull, D., Roessner, U., McConville, M. & Speed, T. P. (2012), 'Normalizing and integrating metabolomics data', *Analytical Chemistry* **84**(24), 10768–10776.

De Livera, A. M., Sysi-Aho, M., Jacob, L., Gagnon-Bartsch, J. A., Castillo, S., Simpson, J. A. & Speed, T. P. (2015), 'Statistical methods for handling unwanted variation in metabolomics data', *Analytical Chemistry* **87**(7), 3606–3615.

Dedeurwaerder, S., Defrance, M., Calonne, E., Denis, H., Sotiriou, C. & Fuks, F. (2011), 'Evaluation of the Infinium Methylation 450K technology', *Epigenomics* **3**(6), 771–784.

Dey, K. K., Hsiao, C. J. & Stephens, M. (2017), 'Visualizing the structure of RNA-seq expression data using grade of membership models', *PLOS Genetics* **13**(3).

Do, K. T., Wahl, S., Raffler, J., Molnos, S., Laimighofer, M., Adamski, J., Suhre, K., Strauch, K., Peters, A., Gieger, C., Langenberg, C., Stewart, I. D., Theis, F. J., Grallert, H., Kastenmüller, G. & Krumsiek, J. (2018), 'Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies', *Metabolomics* **14**(10), 128; 128–128.

Dobriban, E. (2017), 'Factor selection by permutation', *arXiv:1710.00479v1* .

Dobriban, E. & Owen, A. B. (2019), 'Deterministic parallel analysis: an improved method for selecting factors and principal components', *Journal of the Royal Statistical Society: Series B* **81**(1), 163–183.

Domínguez, M. A. & Lobato, I. N. (2004), 'Consistent estimation of models defined by conditional moment restrictions', *Econometrica* **72**(5), 1601–1615.

Douc, R., Moulines, É. & Rydén, T. (2004), 'Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime', *The Annals of Statistics* **32**(5), 2254–2304.

Dubuis, S., Ortmayr, K. & Zampieri, M. (2018), 'A framework for large-scale metabolome drug profiling links coenzyme a metabolism to the toxicity of anti-cancer drug dichloroacetate', *Communications Biology* **1**(1), 101.

Efron, B. (2010), 'Correlated z-values and the accuracy of large-scale statistical estimates', *Journal of the American Statistical Association* **105**(491), 1042–1055.

Eldar, Y. & Kutyniok, G. (2012), *Compressed Sensing: Theory and Applications*, Cambridge University Press.

Fahy, J. V. (2002), 'Goblet cell and Mucin gene abnormalities in asthma', *Chest* **122**(6, Supplement), 320S–326S.

Fan, J. & Han, X. (2017), 'Estimation of the false discovery proportion with unknown dependence', *Journal of the Royal Statistical Society: Series B* **79**(4), 1143–1164.

Ferguson, T. (1996), *A Course in Large Sample Theory*, Chapman & Hall.

Finkelstein, J. L., Pressman, E. K., Cooper, E. M., Kent, T. R., Bar, H. Y. & O'Brien, K. O. (2015), 'Vitamin D status affects serum metabolomic profiles in pregnant adolescents', *Reproductive Sciences* **22**(6), 685–695.

Flutre, T., Wen, X., Pritchard, J. & Stephens, M. (2013), 'A statistical framework for joint eQTL analysis in multiple tissues', *PLOS Genetics* **9**(5), e1003486.

Franks, A. M., Airoldi, E. M. & Rubin, D. B. (2016), 'Non-standard conditionally specified models for non-ignorable missing data'.

Fu, J., Wolfs, M. G. M., Deelen, P., Westra, H.-J., Fehrmann, R. S. N., te Meerman, G. J., Buurman, W. A., Rensen, S. S. M., Groen, H. J. M., Weersma, R. K., van den Berg, L. H., Veldink, J., Ophoff, R. A., Snieder, H., van Heel, D., Jansen, R. C., Hofker, M. H., Wijmenga, C. & Franke, L. (2012), 'Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression', *PLOS Genetics* **8**(1), e1002431–.

Gagnon-Bartsch, J. A., Jacob, L. & Speed, T. P. (2013), Removing unwanted variation from high dimensional data with negative controls, Technical report, UC Berkeley.

Gagnon-Bartsch, J. A. & Speed, T. P. (2012), 'Using control genes to correct for unwanted variation in microarray data', *Biostatistics* **13**(3), 539–552.

Galanter, J. M., Gignoux, C. R., Oh, S. S., Torgerson, D., Pino-Yanes, M., Thakur, N., Eng, C., Hu, D., Huntsman, S., Farber, H. J., Avila, P. C., Brigino-Buenaventura, E., LeNoir, M. A., Meade, K., Serebrisky, D., Rodríguez-Cintrón, W., Kumar, R., Rodríguez-Santana, J. R., Seibold, M. A., Borrell, L. N., Burchard, E. G. & Zaitlen, N. (2017), 'Differential methylation between ethnic sub-groups reflects the effect of genetic ancestry and environmental exposures', *eLife* **6**, e20532.

Gallant, A. R. (1977), 'Three-stage least-squares estimation for a system of simultaneous, nonlinear, implicit equations', *Journal of Econometrics* **5**(1), 71–88.

Gaunt, T. R., Shihab, H. A., Hemani, G., Min, J. L., Woodward, G., Lyttleton, O., Zheng, J., Duggirala, A., McArdle, W. L., Ho, K., Ring, S. M., Evans, D. M., Davey Smith, G. & Relton, C. L. (2016), 'Systematic identification of genetic influences on methylation across the human life course', *Genome Biology* **17**(1), 61.

Gerard, D. & Stephens, M. (2018), 'Empirical Bayes shrinkage and false discovery rate estimation, allowing for unwanted variation', *Biostatistics* .

Gern, J. E., Visness, C. M., Gergen, P. J., Wood, R. A., Bloomberg, G. R., O'Connor, G. T., Kattan, M., Sampson, H. A., Witter, F. R., Sandel, M. T., Shreffler, W. G., Wright, R. J., Arbes, S. J. & Busse, W. W. (2009), 'The Urban Environment and Childhood Asthma (URECA) birth cohort study: design, methods, and study population', *BMC Pulmonary Medicine* **9**(1), 17.

Goodrich, J. M., Meeker, J. D., Dolinoy, D. C., Peterson, K. E., Snchez, B. N., Zhang, Z., Mercado-Garcia, A., Solano-Gonzlez, M., Tllez-Rojo, M. M. & Hu, H. (2016), 'Adolescent epigenetic profiles and environmental exposures from early life through peri-adolescence', *Environmental Epigenetics* **2**(3).

Gosset, W. S. (1908), 'The probable error of a mean', *Biometrika* **6**(1), 1–25.

GTEx Consortium (2017), 'Genetic effects on gene expression across human tissues', *Nature* **550**.

Hall, P. & Horowitz, J. L. (1996), 'Bootstrap critical values for tests based on generalized-method-of-moments estimators', *Econometrica* **64**(4), 891–916.

Hansen, B. E. & West, K. D. (2002), 'Generalized method of moments and macroeconomics', *Journal of Business & Economic Statistics* **20**(4), 460–469.

Hansen, L. P. (1982), 'Large sample properties of generalized method of moments estimators', *Econometrica* **50**(4), 1029–1054.

Hastie, T., Tibshirani, R. & Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Second edition)*, Springer.

Hayton, J. C., Allen, D. G. & Scarpello, V. (2004), 'Factor retention decisions in exploratory factor analysis: a tutorial on parallel analysis', *Organizational Research Methods* **7**(2), 191–205.

Hedeker, D., Wang, J., Chen, L. S. & Wang, P. (2018), 'Using multivariate mixed-effects selection models for analyzing batch-processed proteomics data with non-ignorable missingness', *Biostatistics* .

Herman, S., Emami Khoonsari, P., Aftab, O., Krishnan, S., Strömbom, E., Larsson, R., Hammerling, U., Spjuth, O., Kultima, K. & Gustafsson, M. (2017), 'Mass spectrometry based metabolomics for in vitro systems pharmacology: pitfalls, challenges, and computational solutions', *Metabolomics* **13**(7), 79.

Heyn, H., Moran, S., Hernando-Herraez, I., Sayols, S., Gomez, A., Sandoval, J., Monk, D., Hata, K., Marques-Bonet, T., Wang, L. & Esteller, M. (2013), 'DNA methylation contributes to natural human variation', *Genome Research* **23**(9), 1363–1372.

Horvath, S. (2013), 'DNA methylation age of human tissues and cell types', *Genome Biology* **14**(10), 3156.

Horvath, S., Erhart, W., Brosch, M., Ammerpohl, O., von Schönfels, W., Ahrens, M., Heits, N., Bell, J. T., Tsai, P.-C., Spector, T. D., Deloukas, P., Siebert, R., Sipos, B., Becker, T., Röcken, C., Schafmayer, C. & Hampe, J. (2014), 'Obesity accelerates epigenetic aging of human liver', *Proceedings of the National Academy of Sciences* **111**(43), 15538–15543.

Houseman, E. A., Molitor, J. & Marsit, C. J. (2014), 'Reference-free cell mixture adjustments in analysis of DNA methylation data', *Bioinformatics* **30**(10), 1431–1439.

Jaffe, A. E. & Irizarry, R. A. (2014), 'Accounting for cellular heterogeneity is critical in epigenome-wide association studies', *Genome Biology* **15**(31).

Jia, Y. & Lee, L.-Y. (2007), 'Role of TRPV receptors in respiratory diseases', *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* **1772**(8), 915–927.

Johnson, A. A., Akman, K., Calimport, S. R. G., Wuttke, D., Stolzing, A. & de Magalhães, J. P. (2012), 'The role of DNA methylation in aging, rejuvenation, and age-related disease', *Rejuvenation Research* **15**(5), 483–494.

Johnson, D., Boyes, B., Fields, T., Kopkin, R. & Orlando, R. (2013), 'Optimization of data-dependent acquisition parameters for coupling high-speed separations with LC-MS/MS for protein identifications', *Journal of Biomolecular Techniques* **24**(2), 62–72.

Joubert, B. R., Felix, J. F., Yousefi, P., Bakulski, K. M., Just, A. C., Breton, C., Reese, S. E., Markunas, C. A., Richmond, R. C., Xu, C.-J., Küpers, L. K., Oh, S. S., Hoyo, C., Gruzieva, O., Söderhäll, C., Salas, L. A., Baïz, N., Zhang, H., Lepeule, J., Ruiz, C., Ligthart, S., Wang, T., Taylor, J. A., Duijts, L., Sharp, G. C., Jankipersadsing, S. A., Nilsen, R. M., Vaez, A., Fallin, M. D., Hu, D., Litonjua, A. A., Fuemmeler, B. F., Huen, K., Kere, J., Kull, I., Munthe-Kaas, M. C., Gehring, U., Bustamante, M., Saurel-Coubizolles, M. J., Quraishi, B. M., Ren, J., Tost, J., Gonzalez, J. R., Peters, M. J., Håberg, S. E., Xu, Z., van Meurs, J. B., Gaunt, T. R., Kerkhof, M., Corpeleijn, E., Feinberg, A. P., Eng, C., Baccarelli, A. A., Benjamin Neelon, S. E., Bradman, A., Merid, S. K., Bergström, A., Herceg, Z., Hernandez-Vargas, H., Brunekreef, B., Pinart, M., Heude, B., Ewart, S., Yao, J., Lemonnier, N., Franco, O. H., Wu, M. C., Hofman, A., McArdle, W., Van der Vlies, P., Falahi, F., Gillman, M. W., Barcellos, L. F., Kumar, A., Wickman, M., Guerra, S., Charles, M.-A., Holloway, J., Auffray, C., Tiemeier, H. W., Smith, G. D., Postma, D., Hivert, M.-F., Eskenazi, B., Vrijheid, M., Arshad, H., Antó, J. M., Dehghan, A., Karmaus, W., Annesi-Maesano, I., Sunyer, J., Ghantous, A., Pershagen, G., Holland, N., Murphy, S. K., DeMeo, D. L., Burchard, E. G., Ladd-Acosta, C., Snieder, H., Nystad, W., Koppelman, G. H., Relton, C. L., Jaddoe, V. W. V., Wilcox, A., Melén, E. & London, S. J. (2016), 'DNA methylation in newborns and maternal smoking in pregnancy: Genome-wide consortium meta-analysis', *The American Journal of Human Genetics* **98**(4), 680–696.

Joubert, B. R., Håberg, S. E., Nilsen, R. M., Wang, X., Vollset, S. E., Murphy, S. K., Huang, Z., Hoyo, C., Midttun, Ø., Cupul-Uicab, L. A., Ueland, P. M., Wu, M. C., Nystad, W., Bell, D. A., Peddada, S. D. & London, S. J. (2012), '450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy', *Environmental Health Perspectives* **120**(10), 1425–1431.

Kaminsky, D. A. (2012), 'What does airway resistance tell us about lung function?', *Respiratory Care* **57**(1), 85.

Kang, J. D. Y. & Schafer, J. L. (2007), 'Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data', *Statistical Science* **22**(4), 523–539.

Karpievitch, Y. V., Polpitiya, A. D., Anderson, G. A., Smith, R. D. & Dabney, A. R. (2010), 'Liquid chromatography mass spectrometry-based proteomics: biological and technical aspects', *The Annals of Applied Statistics* **4**(4), 1797–1823.

Kim, S.-H. & Lee, Y.-C. (2009), 'Piperine inhibits eosinophil infiltration and airway hyperresponsiveness by suppressing T cell activity and Th2 cytokine production in the ovalbumin-induced asthma model', *Journal of Pharmacy and Pharmacology* **61**(3), 353–359.

Kippler, M., Engström, K., Mlakar, S. J., Bottai, M., Ahmed, S., Hossain, M. B., Raqib, R., Vahter, M. & Broberg, K. (2013), 'Sex-specific effects of early life cadmium exposure on DNA methylation and implications for birth weight', *Epigenetics* **8**(5), 494–503.

Knight, A. K., Craig, J. M., Theda, C., Bækvad-Hansen, M., Bybjerg-Grauholm, J., Hansen, C. S., Hollegaard, M. V., Hougaard, D. M., Mortensen, P. B., Weinsheimer, S. M., Werge, T. M., Brennan, P. A., Cubells, J. F., Newport, D. J., Stowe, Z. N., Cheong, J. L. Y., Dalach, P., Doyle, L. W., Loke, Y. J., Baccarelli, A. A., Just, A. C., Wright, R. O., Téllez-Rojo, M. M., Svensson, K., Trevisi, L., Kennedy, E. M., Binder, E. B., Iurato, S., Czamara, D., Räikkönen, K., Lahti, J. M. T., Pesonen, A.-K., Kajantie, E., Villa, P. M., Laivuori, H., Hämäläinen, E., Park, H. J., Bailey, L. B., Parets, S. E., Kilaru, V., Menon, R., Horvath, S., Bush, N. R., LeWinn, K. Z., Tylavsky, F. A., Conneely, K. N. & Smith, A. K. (2016), 'An epigenetic clock for gestational age at birth based on blood methylation data', *Genome Biology* **17**(1), 206.

Knowles, D. A., Burrows, C. K., Blischak, J. D., Patterson, K. M., Serie, D. J., Norton, N., Ober, C., Pritchard, J. K., Gilad, Y. & McVean, G. (2018), 'Determining the genetic basis of anthracycline-cardiotoxicity by response QTL mapping in induced cardiomyocytes', *eLife* **7**, e33480.

Ladd-Acosta, C., Hansen, K. D., Briem, E., Fallin, M. D., Kaufmann, W. E. & Feinberg, A. P. (2013), 'Common DNA methylation alterations in multiple brain regions in autism', *Molecular Psychiatry* **19**, 862 EP –.

Lalley, S. (2015), Measure-theoretic probability I. unpublished.

Lee, K. W. K., Richmond, R., Hu, P., French, L., Shin, J., Bourdon, C., Reischl, E., Waldenberger, M., Zeilinger, S., Gaunt, T., McArdle, W., Ring, S., Woodward, G., Bouchard, L., Gaudet, D., Smith, G. D., Relton, C., Paus, T. & Pausova, Z. (2015), 'Prenatal exposure to maternal cigarette smoking and DNA methylation: epigenome-wide association in a discovery sample of adolescents and replication in an independent cohort at birth through 17 years of age', *Environmental Health Perspectives* **123**(2), 193–199.

Lee, S., Sun, W., Wright, F. A. & Zou, F. (2017), 'An improved and explicit surrogate variable analysis procedure by coefficient adjustment', *Biometrika* **104**(2), 303–316.

Leek, J. T., Johnson, W. E., Parker, H. S., Fertig, E. J., Jaffe, A. E., Storey, J. D., Zhang, Y. & Torres, L. C. (2017), *sva: Surrogate Variable Analysis.* R package version 3.26.0.

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K. & Irizarry, R. A. (2010), 'Tackling the widespread and critical impact of batch effects in high-throughput data', *Nature reviews. Genetics* **11**(10).

Leek, J. T. & Storey, J. D. (2007), 'Capturing heterogeneity in gene expression studies by surrogate variable analysis', *PLOS Genetics* **3**(9), 1724–1735.

Leek, J. T. & Storey, J. D. (2008), 'A general framework for multiple testing dependence', *Proceedings of the National Academy of Sciences* **105**(48), 18718–18723.

Levine, M. E. & Crimmins, E. M. (2014), 'Evidence of accelerated aging among African Americans and its implications for mortality', *Social Science & Medicine* **118**, 27–32.

Li, C. & Jiang, W. (2016), 'On oracle property and asymptotic validity of Bayesian generalized method of moments', *Journal of Multivariate Analysis* **145**, 132 – 147.

Li, D., Qian, J. & Su, L. (2016), 'Panel data models with interactive fixed effects and multiple structural breaks', *Journal of the American Statistical Association* **111**(516), 1804–1819.

Liang, K.-Y. & Qin, J. (2000), 'Regression analysis under non-standard situations: A pairwise pseudolikelihood approach', *Journal of the Royal Statistical Society: Series B* **62**(4), 773–786.

Liu, C. (2005), *Robit Regression: A Simple Robust Alternative to Logistic and Probit Regression*, John Wiley & Sons, Ltd, chapter 21, pp. 227–238.

Liu, C., Marioni, R. E., Hedman, Å., Pfeiffer, L., Tsai, P.-C., Reynolds, L. M., Just, A. C., Duan, Q., Boer, C. G., Tanaka, T., Elks, C. E., Aslibekyan, S., Brody, J. A., Kühnel, B., Herder, C., Almli, L. M., Zhi, D., Wang, Y., Huan, T., Yao, C., Mendelson, M. M., Joehanes, R., Liang, L., Love, S.-A., Guan, W., Shah, S., McRae, A. F., Kretschmer, A., Prokisch, H., Strauch, K., Peters, A., Visscher, P. M., Wray, N. R., Guo, X., Wiggins, K. L., Smith, A. K., Binder, E. B., Ressler, K. J., Irvin, M. R., Absher, D. M., Hernandez, D., Ferrucci, L., Bandinelli, S., Lohman, K., Ding, J., Trevisi, L., Gustafsson, S., Sandling, J. H., Stolk, L., Uitterlinden, A. G., Yet, I., Castillo-Fernandez, J. E., Spector, T. D., Schwartz, J. D., Vokonas, P., Lind, L., Li, Y., Fornage, M., Arnett, D. K., Wareham, N. J., Sotoodehnia, N., Ong, K. K., van Meurs, J. B. J., Conneely, K. N., Baccarelli, A. A., Deary, I. J., Bell, J. T., North, K. E., Liu, Y., Waldenberger, M., London, S. J., Ingelsson, E. & Levy, D. (2016), 'A DNA methylation biomarker of alcohol consumption', *Molecular Psychiatry* **23**, 422 EP –.

Liu, X., Ser, Z. & Locasale, J. W. (2014), 'Development and quantitative evaluation of a high-resolution metabolomics technology', *Analytical Chemistry* **86**(4), 2175–2184.

Lu, X. & Su, L. (2016), 'Shrinkage estimation of dynamic panel data models with interactive fixed effects', *Journal of Econometrics* **190**(1), 148–175.

Mahoney, M. W. (2011), *Randomized Algorithms for Matrices and Data*, Now Publishers Inc.

Maksimovic, J., Gagnon-Bartsch, J. A., Speed, T. P. & Oshlack, A. (2015), 'Removing unwanted variation in a differential methylation analysis of illumina humanmethylation450 array data', *Nucleic Acids Research* **43**(16), e106–e106.

Marioni, R. E., Shah, S., McRae, A. F., Chen, B. H., Colicino, E., Harris, S. E., Gibson, J., Henders, A. K., Redmond, P., Cox, S. R., Pattie, A., Corley, J., Murphy, L., Martin, N. G., Montgomery, G. W., Feinberg, A. P., Fallin, M. D., Multhaup, M. L., Jaffe, A. E., Joehanes, R., Schwartz, J., Just, A. C., Lunetta, K. L., Murabito, J. M., Starr, J. M., Horvath, S., Baccarelli, A. A., Levy, D., Visscher, P. M., Wray, N. R. & Deary, I. J. (2015), 'DNA methylation age of blood predicts all-cause mortality in later life', *Genome Biology* **16**(1), 25.

Markunas, C. A., Xu, Z., Harlid, S., Wade, P. A., Lie, R. T., Taylor, J. A. & Wilcox, A. J. (2014), 'Identification of DNA methylation changes in newborns related to maternal smoking during pregnancy', *Environmental Health Perspectives* **122**(10), 1147–1153.

Martin, E. M. & Fry, R. C. (2018), 'Environmental influences on the epigenome: Exposure-associated DNA methylation in human populations', *Annual Review of Public Health* **39**(1), 309–333.

Martino, D., Loke, Y. J., Gordon, L., Ollikainen, M., Cruickshank, M. N., Saffery, R. & Craig, J. M. (2013), 'Longitudinal, genome-scale analysis of DNA methylation in twins from birth to 18 months of age reveals rapid epigenetic change in early life and pair-specific effects of discordance', *Genome Biology* **14**(5), R42.

Maschietto, M., Bastos, L. C., Tahira, A. C., Bastos, E. P., Euclydes, V. L. V., Brentani, A., Fink, G., de Baumont, A., Felipe-Silva, A., Francisco, R. P. V., Gouveia, G., Grisi, S. J. F. E., Escobar, A. M. U., Moreira-Filho, C. A., Polanczyk, G. V., Miguel, E. C. & Brentani, H. (2017), 'Sex differences in DNA methylation of the cord blood are related to sex-bias psychiatric diseases', *Scientific Reports* **7**.

McKennan, C., Naughton, K., Stanhope, C., Kattan, M., O'Connor, G., Sandel, M., Visness, C., Wood, R., Bacharier, L., Beigelman, A., Lovisky-Desir, S., Togias, A., Gern, J., Nicolae, D. & Ober, C. (2018), 'Longitudinal studies at birth and age 7 reveal strong effects of genetic variation on ancestry-associated DNA methylation patterns in blood cells from ethnically admixed children', *bioRxiv* .

McKennan, C. & Nicolae, D. (2018*a*), 'Accounting for unobserved covariates with varying degrees of estimability in high dimensional experimental data', *Biometrika (forthcoming)* .

McKennan, C. & Nicolae, D. (2018*b*), 'Estimating and accounting for unobserved covariates in high dimensional correlated data', *arXiv:1808.05895v1* .

Moen, E. L., Zhang, X., Mu, W., Delaney, S. M., Wing, C., McQuade, J., Myers, J., Godley, L. A., Dolan, M. E. & Zhang, W. (2013), 'Genome-wide variation of cytosine modifications between European and African populations and the implications for complex traits', *Genetics* **194**(4), 987.

Morales, E., Vilahur, N., Salas, L. A., Motta, V., Fernandez, M. F., Murcia, M., Llop, S., Tardon, A., Fernandez-Tardon, G., Santa-Marina, L., Gallastegui, M., Bollati, V., Estivill, X., Olea, N., Sunyer, J. & Bustamante, M. (2016), 'Genome-wide DNA methylation study in human placenta identifies novel loci associated with maternal smoking during pregnancy', *International Journal of Epidemiology* **45**(5), 1644–1655.

Mozhui, K., Smith, A. K. & Tylavsky, F. A. (2015), 'Ancestry dependent DNA methylation and influence of maternal nutrition', *PLOS ONE* **10**(3), e0118466–.

Newey, W. K. (1990), 'Efficient instrumental variables estimation of nonlinear models', *Econometrica* **58**(4), 809–837.

Nguyen, A. B., Moser, R. & Chou, W. Y. (2014), 'Race and health profiles in the united states: an examination of the social gradient through the 2009 CHIS adult survey', *Public Health* **128**(12), 1076–1086.

Nicodemus-Johnson, J., Myers, R. A., Sakabe, N. J., Sobreira, D. R., Hogarth, D. K., Naureckas, E. T., Sperling, A. I., Solway, J., White, S. R., Nobrega, M. A., Nicolae, D. L., Gilad, Y. & Ober, C. (2016), 'DNA methylation in lung cells is associated with asthma endotypes and genetic risk', *JCI Insight* **1**(20), e90151.

Norets, A. (2010), 'Approximation of conditional densities by smooth mixtures of regressions', *The Annals of Statistics* **38**(3), 1733–1766.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R., Stephens, M. & Bustamante, C. D. (2008), 'Genes mirror geography within Europe', *Nature* **456**(7218), 98–101.

O'Brien, J. J., Gunawardena, H. P., Paulo, J. A., Chen, X., Ibrahim, J. G., Gygi, S. P. & Qaqish, B. F. (2018), 'The effects of nonignorable missing data on label-free mass spectrometry proteomics experiments', *The Annals of Applied Statistics* **12**(4), 2075–2095.

O'Connor, G. T., Lynch, S. V., Bloomberg, G. R., Kattan, M., Wood, R. A., Gergen, P. J., Jaffee, K. F., Calatroni, A., Bacharier, L. B., Beigelman, A., Sandel, M. T., Johnson, C. C., Faruqi, A., Santee, C., Fujimura, K. E., Fadrosh, D., Boushey, H., Visness, C. M. & Gern, J. E. (2018), 'Early-life home environment and risk of asthma among inner-city children', *Journal of Allergy and Clinical Immunology* **141**(4), 1468–1475.

Onatski, A. (2010), 'Determining the number of factors from empirical distribution of eigenvalues', *The Review of Economics and Statistics* **92**(4), 1004–1016.

Ono, J. G., Worgall, T. S. & Worgall, S. (2015), 'Airway reactivity and sphingolipids-implications for childhood asthma', *Molecular and Cellular Pediatrics* **2**(1), 13; 13–13.

Ostroukhova, M., Goplen, N., Karim, M. Z., Michalec, L., Guo, L., Liang, Q. & Alam, R. (2012), 'The role of low-level lactate production in airway inflammation in asthma', *American Journal of Physiology - Lung Cellular and Molecular Physiology* **302**(3), L300–L307.

Owen, A. B. & Perry, P. O. (2009), 'Bi-cross-validation of the SVD and the nonnegative matrix factorization', *Annals of Applied Statistics* **3**(2), 564–594.

Owen, A. B. & Wang, J. (2016), 'Bi-cross-validation for factor analysis', *Statistical Science* **31**(1), 119–139.

Pacor, M. L., Di Lorenzo, G., Martinelli, N., Mansueto, P., Rini, G. B. & Corrocher, R. (2004), 'Monosodium benzoate hypersensitivity in subjects with persistent rhinitis', *Allergy* **59**(2), 192–197.

Parets, S. E., Conneely, K. N., Kilaru, V., Fortunato, S. J., Syed, T. A., Saade, G., Smith, A. K. & Menon, R. (2013), 'Fetal DNA methylation associates with early spontaneous preterm birth and gestational age', *PLOS ONE* **8**(6), e67489–.

Paul, D. (2007), 'Asymptotics of sample eigenstructure for a large dimensional spiked covariance model', *Statistica Sinica* **17**, 1617–1642.

Pearson, K. (1901), *On Lines and Planes of Closest Fit to Systems of Points in Space*, University College.

Peixoto, L., Risso, D., Poplawski, S. G., Wimmer, M. E., Speed, T. P., Wood, M. A. & Abel, T. (2015), 'How data analysis affects power, reproducibility and biological insight of RNA-seq studies in complex datasets', *Nucleic Acids Research* **43**(16), 7664–7674.

Premkumar, L. S. (2014), 'Transient receptor potential channels as targets for phytochemicals', *ACS Chemical Neuroscience* **5**(11), 1117–1130.

Rahmani, E., Shenhav, L., Schweiger, R., Yousefi, P., Huen, K., Eskenazi, B., Eng, C., Huntsman, S., Hu, D., Galanter, J., Oh, S. S., Waldenberger, M., Strauch, K., Grallert, H., Meitinger, T., Gieger, C., Holland, N., Burchard, E. G., Zaitlen, N. & Halperin, E. (2017), 'Genome-wide methylation data mirror ancestry information', *Epigenetics & Chromatin* **10**(1), 1.

Reinke, S. N., Gallart-Ayala, H., Gómez, C., Checa, A., Fauland, A., Naz, S., Kamleh, M. A., Djukanović, R., Hinks, T. S. C. & Wheelock, C. E. (2017), 'Metabolomics analysis identifies different metabotypes of asthma severity', *The European Respiratory Journal* **49**(3), 1601740.

Richmond, R. C., Simpkin, A. J., Woodward, G., Gaunt, T. R., Lyttleton, O., McArdle, W. L., Ring, S. M., Smith, A. D. A. C., Timpson, N. J., Tilling, K., Davey Smith, G. & Relton, C. L. (2015), 'Prenatal exposure to maternal smoking and offspring DNA methylation across the lifecourse: findings from the Avon Longitudinal Study of Parents and Children (ALSPAC)', *Human Molecular Genetics* **24**(8), 2201–2217.

Rogers, D. F. (2002), 'Airway goblet cell hyperplasia in asthma: hypersecretory and anti-inflammatory?', *Clinical & Experimental Allergy* **32**(8), 1124–1127.

Rutledge, J. C., Yokoyama, A. S. & Medici, V. (2017), 'DNA methylation alterations in alzheimers disease', *Environmental Epigenetics* **3**(2).

Rzehak, P., Saffery, R., Reischl, E., Covic, M., Wahl, S., Grote, V., Xhonneux, A., Langhendries, J.-P., Ferre, N., Closa-Monasterolo, R., Verduci, E., Riva, E., Socha, P., Gruszfeld, D., Koletzko, B. & group, E. C. O. T. S. (2016), 'Maternal smoking during pregnancy and DNA-methylation in children at age 5.5 years: Epigenome-wide-analysis in the European Childhood Obesity Project (CHOP)-study', *PLOS ONE* **11**(5), e0155554–.

Salerno, Stephen, J., Mehrmohamadi, M., Liberti, M. V., Wan, M., Wells, M. T., Booth, J. G. & Locasale, J. W. (2017), 'RRmix: A method for simultaneous batch effect correction and analysis of metabolomics data in the absence of internal standards', *PLOS ONE* **12**(6), e0179530–.

Sampson, J. N., Boca, S. M., Shu, X. O., Stolzenberg-Solomon, R. Z., Matthews, C. E., Hsing, A. W., Tan, Y. T., Ji, B.-T., Chow, W.-H., Cai, Q., Liu, D. K., Yang, G., Xiang, Y. B., Zheng, W., Sinha, R., Cross, A. J. & Moore, S. C. (2013), 'Metabolomics in epidemiology: sources of variability in metabolite measurements and implications', *Cancer Epidemiology, Biomarkers & Prevention* **22**(4), 631–640.

Schroeder, J. W., Conneely, K. N., Cubells, J. F., Kilaru, V., Newport, D. J., Knight, B. T., Stowe, Z. N., Brennan, P. A., Krushkal, J., Tylavsky, F. A., Taylor, R. N., Adkins, R. M. & Smith, A. K. (2011), 'Neonatal DNA methylation patterns associate with gestational age', *Epigenetics* **6**(12), 1498–1504.

Smith, A. K., Kilaru, V., Kocak, M., Almli, L. M., Mercer, K. B., Ressler, K. J., Tylavsky, F. A. & Conneely, K. N. (2014), 'Methylation quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage, and tissue type', *BMC Genomics* **15**(1), 145.

Stein, M. M., Hrusch, C. L., Gozdz, J., Igartua, C., Pivniouk, V., Murray, S. E., Ledford, J. G., Marques dos Santos, M., Anderson, R. L., Metwali, N., Neilson, J. W., Maier, R. M., Gilbert, J. A., Holbreich, M., Thorne, P. S., Martinez, F. D., von Mutius, E., Vercelli, D., Ober, C. & Sperling, A. I. (2016), 'Innate immunity and asthma risk in Amish and Hutterite farm children', *New England Journal of Medicine* **375**(5), 411–421.

Stephens, M. (2016), 'False discovery rates: a new deal', *Biostatistics* **18**(2), 275–294.

Storey, J. D. (2001), 'A direct approach to false discovery rates', *Journal of the Royal Statistical Society: Series B* **63**(3), 479–498.

Storey, J. D., Bass, A. J., Dabney, A. & Robinson, D. (2015), *qvalue: Q-value estimation for false discovery rate control*. R package version 2.10.0.
URL: *http://github.com/jdstorey/qvalue*

Su, L. & Ju, G. (2018), 'Identifying latent grouped patterns in panel data models with interactive fixed effects', *Journal of Econometrics* **206**(2), 554–573.

Sun, Y. & Kim, M. S. (2012), 'Simple and powerful GMM over-identification tests with accurate size', *Journal of Econometrics* **166**(2), 267–281.

Sun, Y., Zhang, N. R. & Owen, A. B. (2012), 'Multiple hypothesis testing adjusted for latent variables, with an application to the AGEMAP gene expression data', *The Annals of Applied Statistics* **6**(4), 1664–1688.

Taddy, M. (2012), 'On estimation and selection for topic models', *AISTATS* pp. 1184–1193.

Titus, A. J., Gallimore, R. M., Salas, L. A. & Christensen, B. C. (2017), 'Cell-type deconvolution from DNA methylation: a review of recent applications', *Human Molecular Genetics* **26**(R2), R216–R224.

Tung, J., Zhou, X., Alberts, S. C., Stephens, M., Gilad, Y. & Dermitzakis, E. T. (2015), 'The genetic architecture of gene expression levels in wild baboons', *eLife* **4**, e04729.

van Iterson, M., van Zwet, E. W. & Heijmans, B. T. (2017), 'Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution', *Genome Biology* **18**(1), 19.

Wald, A. (1949), 'Note on the consistency of the maximum likelihood estimate', *The Annals of Mathematical Statistics* **20**(4), 595–601.

Wang, J., Zhao, Q., Hastie, T. & Owen, A. B. (2017), 'Confounder adjustment in multiple hypothesis testing', *The Annals of Statistics* **45**(5), 1863–1894.

Wang, M., Kong, L., Li, Z. & Zhang, L. (2016), 'Covariance estimators for generalized estimating equations (GEE) in longitudinal analysis with small samples', *Statistics in medicine* **35**(10), 1706–1721.

Wang, S., Shao, J. & Kim, J. K. (2014), 'An instrumental variable approach for identification and estimation with nonignorable nonresponse', *Statistica Sinica* **24**(3), 1097–1116.

Wehrens, R., Hageman, J. A., van Eeuwijk, F., Kooke, R., Flood, P. J., Wijnker, E., Keurentjes, J. J. B., Lommen, A., van Eekelen, H. D. L. M., Hall, R. D., Mumm, R. & de Vos, R. C. H. (2016), 'Improved batch correction in untargeted MS-based metabolomics', *Metabolomics* **12**, 88; 88–88.

Wu, D., Yang, H., Winham, S. J., Natanzon, Y., Koestler, D. C., Luo, T., Fridley, B. L., Goode, E. L., Zhang, Y. & Cui, Y. (2018), 'Mediation analysis of alcohol consumption, DNA methylation, and epithelial ovarian cancer', *Journal of Human Genetics* **63**(3), 339–348.

Xu, S., Ross, C., Raebel, M. A., Shetterly, S., Blanchette, C. & Smith, D. (2010), 'Use of stabilized inverse propensity scores as weights to directly estimate relative risk and its confidence intervals', *Value in Health* **13**(2), 273–277.

Xu, Y.-D., Cui, J.-M., Wang, Y., Yin, L.-M., Gao, C.-K., Liu, Y.-Y. & Yang, Y.-Q. (2010), 'The early asthmatic response is associated with glycolysis, calcium binding and mitochondria activity as revealed by proteomic analysis in rats', *Respiratory Research* **11**(1), 107.

Yang, I. V., Pedersen, B. S., Liu, A. H., O'Connor, G. T., Pillai, D., Kattan, M., Misiak, R. T., Gruchalla, R., Szefler, S. J., Khurana Hershey, G. K., Kercsmar, C., Richards, A., Stevens, A. D., Kolakowski, C. A., Makhija, M., Sorkness, C. A., Krouse, R. Z., Visness, C., Davidson, E. J., Hennessy, C. E., Martin, R. J., Togias, A., Busse, W. W. & Schwartz, D. A. (2017), 'The nasal methylome and childhood atopic asthma', *Journal of Allergy and Clinical Immunology* **139**(5), 1478–1488.

Yao, C., Li, H., Shen, X., He, Z., He, L. & Guo, Z. (2012), 'Reproducibility and concordance of differential DNA methylation and gene expression in cancer', *PLOS ONE* **7**(1), e29686–.

Yin, G. (2009), 'Bayesian generalized method of moments', *Bayesian Analysis* **4**(2), 191–207.

Young, S. P. & Wallace, G. R. (2009), 'Metabolomic analysis of human disease and its application to the eye', *Journal of Ocular Biology, Diseases, and Informatics* **2**(4), 235–242.

Yousefi, P., Huen, K., Davé, V., Barcellos, L., Eskenazi, B. & Holland, N. (2015), 'Sex differences in DNA methylation assessed by 450K beadchip in newborns', *BMC Genomics* **16**(1).

Zhang, X., Biagini Myers, J. M., Burleson, J., Ulm, A., Bryan, K. S., Chen, X., Weirauch, M. T., Baker, T. A., Butsch Kovacic, M. S. & Ji, H. (2018), 'Nasal DNA methylation is associated with childhood asthma', *Epigenomics* **10**(5), 629–641.

Zhou, X. & Stephens, M. (2014), 'Efficient multivariate linear mixed model algorithms for genome-wide association studies', *Nature Methods* **11**(4), 407–409.

Zuhdi Alimam, M., Piazza, F. M., Selby, D. M., Letwin, N., Huang, L. & Rose, M. C. (2000), 'Muc-5/5ac mucin messenger RNA and protein expression is a marker of goblet cell metaplasia in murine airways', *American Journal of Respiratory Cell and Molecular Biology* **22**(3), 253–260.