Topics in False Discovery Rate Control in Multiple Regression Models with Unobserved Confounders

by

Taiyu Ye

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Statistics)

at the

UNIVERSITY OF WISCONSIN–MADISON

2024

Date of final oral examination: 5/20/2024

The dissertation is approved by the following members of the Final Oral Committee:

Chunming Zhang, Professor, Statistics

Xiaoxia Shi, Professor, Economics

Hyunseung Kang, Associate Professor, Statistics

Yiqiao Zhong, Assistant Professor, Statistics

Zhengjun Zhang, Professor, Statistics

To my family

Acknowledgments

I am deeply indebted to Prof. Chunming Zhang for her exceptional mentorship, unwavering support, and insightful guidance throughout my Ph.D. journey. Her vast knowledge of statistical insights and methodological development, coupled with her dedication to academic excellence, has profoundly influenced the direction and quality of this research. Prof. Zhang has set a standard of rigor and precision that has greatly enriched the quality of this research. I am truly fortunate to have had the privilege of working under her supervision.

My sincere thanks also extend to my committee members: Prof. Xiaoxia Shi, Prof. Hyunseung Kang, Prof. Yiqiao Zhong, and Prof. Zhengjun Zhang. Their collective expertise significantly contributed to the success of this research endeavor. Prof. Xiaoxia Shi's insights in Economics provided invaluable interdisciplinary perspectives, enriching the depth of this study. I am grateful to Prof. Hyunseung Kang not only for sharing a seminal paper that guided me along the path of this research but also for many constructive conversations. I also extend my heartfelt gratitude to Prof. Yiqiao Zhong and Prof. Zhengjun Zhang for their enormous support during my STAT 709 TA experiences. Prof. Yiqiao Zhong's guidance in statistical learning analysis was instrumental in shaping key findings and conclusions. Prof. Zhengjun Zhang's profound understanding of statistical modeling broadened my horizons and enriched the depth of my research. I miss joyful conversations with Prof. Zhengjun Zhang. Although it rarely happened now after his retirement, his influence continues to resonate in my work and thinking.

I also want to extend my thanks to professors who contributed in my PhD jounrey: Bret Larget, Brian Yandell, Cécile Ané, David Anderson, Eric Bach, Garvesh Raskutti, Timo

Seppäläinen, Jun Shao, Jun Zhu, Stephen Wright, Yazhen Wang, Yingyu Liang.

I was fortunate to participate in a insightful research group held by Prof. Chunming Zhang. I want to thank for all the speakers for their thorough preparations and insightful understanding, which greatly enlarged my knowledge base. Especially, I am extremely grateful to Muhong Gao, Yongsu Lee, Yongfeng Wu, Yanbo Shen and Bowen Zhang for their valuable contributions and enriching discussions, and most importantly, their consistent support throughout my PhD journey.

Furthermore, I am grateful to my graduate fellows and dear friends such as Ben Teo, Chi-Shian Dai, Chanwoo Lee, Hanying Jiang, Jinyi Wang, Jingcheng Xu, Kwangmoon Park, Nathaniel Pritchard, Peng Yu, Susan Glenn, Yue Gao for their endless encouragement, listening ears, and uplifting spirits during moments of triumphs and challenges moments. Their camaraderie and unwavering belief in my aspirations have made this journey more meaningful and enjoyable.

Lastly, I am deeply grateful to my girlfriend, Dr. Xiangqi Bai, and my family for their constant love, encouragement, and understanding throughout this challenging yet rewarding academic journey. Their great support and belief in my abilities have strengthened and motivated me. I am especially thankful to my parents for instilling in me a passion for learning and for always believing in my potential to achieve my goals.

This work was supported in part by the University of Wisconsin-Madison Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation.

Contents

List of Tables

List of Figures

Abstract

This thesis introduces a new method for controlling false discovery rate (FDR) in large-dimensional cross-sectional datasets, such as those encountered in genome or brain imaging studies. These datasets can typically be modeled by multiple regression models, where traditional multiple testing procedures can be compromised by strong dependence among test statistics, often due to batch effects or unobserved confounders. Using a probabilistic representation technique, we establish that the unobserved effects can be characterized by low-dimensional variables via a class of potentially nonlinear functions. Furthermore, we prove the existence of a uniform basis approximation to these nonlinear functions when they originate from a common Reproducing Kernel Hilbert Space (RKHS). This enhances the philosophy of modeling the unobserved effects using factor models in the literature. We propose a novel two-step method to mitigate the common dependence caused by the unobserved effects. In the first step, we obtain an estimation of Negative Control (NC) set. This aims to control the false negative rate while allowing for some sacrifice in FDR. In the second step, we extract useful information about the unobserved effects from the estimated NC set, and construct pairwise asymptotically independent test statistics. We provide theoretical guarantees for both steps. Additionally, our theoretical analysis enhances the existing literature by relaxing a common technical assumption that limits theoretical analysis for FDR control. Numerical experiments and theoretical derivations demonstrate that our method's false discovery proportion (FDP) estimations converge to the oracle values that utilize information from the unobserved variables, providing a guarantee for FDR control.

Chapter 1

Introduction

Multiple hypothesis testing finds extensive applications across different research fields such as biology, medicine, genetics, neuroscience, econometrics and finance. Traditional methods for performing multiple testing with false discovery rate (FDR) control typically rely on the assumption of independence or weak dependence among test statistics (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001; Storey et al., 2004). However, in many practical applications, variables are often correlated, making the independence or weak dependence assumption highly unrealistic. For instance, batch effects are common in biological experiments, where data are affected by non-biological or non-scientific variables such as laboratory conditions, technicians adeptness (Leek et al., 2010). These effects can result in correlations between outcomes of interest. In medicine, k-locus analysis is used to detect the causative genes for complex diseases such as diabetes, schizophrenia and cancer (Li and Ji, 2005). When two different k-loci with shared locations are of interest, the tests for these loci are inherently correlated due to shared genetic information in the overlap region. In finance, conducting performance analysis on all hedge fund managers to distinguish their skills from luck (Romano and Wolf, 2005) can lead to correlated tests, as these managers operate in the same macro-econometric environment. The source variables responsible for these correlations may not be included or only partially included in the final datasets, complicating the task of adjusting for correlations.

As emphasized in Efron (2007), accounting for correlation is crucial in deciding which hypotheses should be rejected. Otherwise, the accuracy of multiple testing procedures can be compromised, especially in high-correlation situations. Correlations among test statistics are often reflected in histograms of p-values. Ideally, the histogram of null p-values should be uniformly distributed from 0 to 1. Distorted null distribution histograms may have peaks at either end of the range. If null p-values are concentrated around 0, this indicates a failure in controlling the type I error. Conversely, if null p-values are concentrated around 1, the alternative p-values are likely to become larger due to correlations, leading to increased type II error and reduced power.

Factor models have been successful in capturing the interconnections across various subjects and measured variables in various applications (Almlund et al., 2011; Fan et al., 2013; Grotzinger et al., 2019). Therefore, Factor models are well-suited for addressing correlations arising from unobserved variables. Factor models also find applications in multiple testing procedures, particularly within multiple responses linear regression models (Leek and Storey, 2008; Friguet et al., 2009; Fan et al., 2012; Gagnon-Bartsch et al., 2013; Du and Zhang, 2017; Lee et al., 2017; Wang et al., 2017; McKennan and Nicolae, 2019; Bing et al., 2023). We emphasize that factor models can have different names, such as surrogate variable analysis, batch effects, unwanted variation, latent factor structure, unmeasured confounders, or hidden variables. Regardless of the terminology, their shared objective is to correct distorted null distributions caused by the unobserved effect.

This thesis will focus on addressing a class of multiple testing problems within this framework. We propose a new method for estimating the latent factors, which aids in disentangling correlations among different measured responses, thereby enabling construction of asymptotically independent test statistics.

A multiple responses linear regression model have the following form:

$$Y_{i,j} = \boldsymbol{b}_{i,\boldsymbol{.}}^{\top} \boldsymbol{X}_{\boldsymbol{.},j} + u_{i,j}. \tag{1.1}$$

Here, $Y_{i,j}$ is the measured response variable for the j-th ($j = 1, \ldots, n$) subject in the i-th ($i = 1, \ldots, m$) linear regression model. $\boldsymbol{b}_{i,\boldsymbol{\cdot}} = (b_{i,1}, \ldots, b_{i,p})^\top \in \mathbb{R}^p$ is the vector of regression coefficient, $\boldsymbol{X}_{\boldsymbol{\cdot},j} = (X_{1,j}, \ldots, X_{p,j})^\top \in \mathbb{R}^p$ is the explanatory variable, $u_{i,j}$ is the remainder. Only $Y_{i,j}$ and $\boldsymbol{X}_{\boldsymbol{\cdot},j}$ are observable, while $u_{i,j}$ is unobserved. $u_{i,j}$ contains correlation information across different measured responses, which is the target we aim to address.

Our multiple testing problem involves evaluating the following hypotheses for a user-specific $c \times p$ constraint matrix A with $\mathrm{rank}(A) = c \leq p$:

$$H_{0i} : A\boldsymbol{b}_{i,\boldsymbol{\cdot}} = \mathbf{0} \quad \text{v.s.} \quad H_{1i} : A\boldsymbol{b}_{i,\boldsymbol{\cdot}} \neq \mathbf{0} \quad i = 1, \ldots, m. \tag{1.2}$$

Upon eliminating the effect of $\boldsymbol{X}_{\boldsymbol{\cdot},1}, \ldots, \boldsymbol{X}_{\boldsymbol{\cdot},n}$, the relationship between the i-th and i'-th regressions can be characterized by the correlation between $u_{i,j}$, $u_{i',j}$. This dependence is commonly described using the latent factor structure:

$$u_{i,j} = (\boldsymbol{\gamma}_{i,\boldsymbol{\cdot}}^{(L)})^\top \boldsymbol{g}_{\boldsymbol{\cdot},j}^{(L)} + \epsilon_{i,j}, \tag{1.3}$$

where $\boldsymbol{g}_{\boldsymbol{\cdot},j}^{(L)}$ is an unobserved latent factor that governs the majority of dependence, $\boldsymbol{\gamma}_{i,\boldsymbol{\cdot}}^{(L)}$ is a factor loading, $\epsilon_{i,j}$ is an idiosyncratic error term. Combining Eq.(1.1) and Eq.(1.3), the model can be written in a matrix form:

$$\mathbf{Y} = \mathbf{B}\mathbf{X} + \boldsymbol{\Gamma}_L \mathbf{G}_L + \mathbf{E}. \tag{1.4}$$

However, it is common to assume $\mathbf{G}_L$ to have exact low rank. Consequently, the covariance structures considered in the existing works tend to be restrictive.

1.1. Main contributions

In this thesis, we initiate the modeling of $u_{i,j}$ in a genuinely general sense where no structural assumptions are imposed on the covariance of $u_{1,j}, \ldots, u_{m,j}$. Instead of assuming a

generative model as in Eq.(1.3), we demonstrate, from a representative perspective, that $u_{i,j}$ can be represented by a low-dimensional latent factor $\boldsymbol{\alpha}_{\boldsymbol{\cdot},j}$ through a nonrandom nonlinear function f_i:

$$u_{i,j} = f_i(\boldsymbol{\alpha}_{\boldsymbol{\cdot},j}) + \epsilon_{i,j}, \tag{1.5}$$

where $f_i(\boldsymbol{\alpha}_{\boldsymbol{\cdot},j})$ is the nonlinear latent component, and $\epsilon_{i,j}$ is the idiosyncratic error term.

Eq.(1.5) can be termed as nonlinear factor models, and there are relatively few papers that make efforts in estimating it. Recently, Feng (2020) proposed a local PCA method that calculates a neighborhood for each observation $Y_{i,j}$ where the Taylor expansion is valid for a monomial basis approximation around $\boldsymbol{\alpha}_{\boldsymbol{\cdot},j}$. But in the numerical experiments, the author only used functions that could be expressed as finite linear combinations of some basis functions, essentially reducing Eq.(1.5) to Eq.(1.3). This motivates us to answer a pertinent question: what kind of systems of nonlinear functions $f_1, \ldots, f_m$ can be linearly approximated with suitable choice of basis functions? We demonstrate that if $f_1, \ldots, f_m$ come from a common Reproducing Kernel Hilbert Space (RKHS), then they can be uniformly approximated by eigenfunctions of the reproducing kernel. The approximation error is controlled by the order of the eigenvalue to which the eigen-expansions end and the maximum norm of $f_1, \ldots, f_m$ in the RKHS.

This reveals that the nonlinear representation in Eq.(1.5) reduces to Eq.(1.3) up to the approximation error, and model Eq.(1.1) can always be represented in the following matrix form:

$$\mathbf{Y} = \mathbf{B}\mathbf{X} + \boldsymbol{\Gamma}\mathbf{G} + \mathbf{F} + \mathbf{E}, \tag{1.6}$$

where $\mathbf{G}$ [1] is the latent basis, $\boldsymbol{\Gamma}$ is the matrix of expansion coefficients, $\mathbf{F}$ is the approximation error matrix. The generative model Eq.(1.3) holds true when $\mathbf{F} = \mathbf{0}$. While Eq.(1.6) shares similarities with Eq.(1.4), a notable difference arises in terms of the dimensionality of $\mathbf{G}$. This difference stems from potential existence of infinite positive eigenvalues for a reproducing kernel. Furthermore, the approximation error can be made small only as the dimensionality

[1] Here, we use $\mathbf{G}$ to distinguish from $\mathbf{G}_L$ in the pure linear case. Same for $\boldsymbol{\Gamma}$.

approaches infinity.

To address the latent effect, our method aims to explicitly estimate the latent factor $\mathbf{G}_L$ in linear case, or the latent basis $\mathbf{G}$ in nonlinear case. The main heuristic relies on partial observation of the unobserved component: if many rows of $\mathbf{B}$ are zeroes, the submatrix of $\mathbf{Y}$ that consists of these rows resembles factor models (Bai, 2003; Bai and Ng, 2002), and $\mathbf{G}_L$ or $\mathbf{G}$ can be estimated using PCA method. A collection of zero rows of $\mathbf{B}$ is referred to as a Negative Control (NC) set [2]. This idea draws inspiration from Leek and Storey (2008), who proposed an iteratively re-weighted algorithm for NC set estimation. However, they did not provide a theoretical guarantee for the NC set estimation, and their method is significantly degraded, especially when $\mathbf{G}_L$ is confounding with $\mathbf{X}$ (Wang et al., 2017). Another approach relies on prior knowledge of the NC set, as introduced by Gagnon-Bartsch et al. (2013). However, such prior knowledge is often scarce in many applications. In contrast, we propose a new method for estimating the NC set, which offers a theoretical guarantee to control false negatives even under potential confounding effect between $\mathbf{X}$ and the latent component. This method relies on the "beta-min" condition, which assumes the magnitudes of nonzero rows are not too small.

After adjusting for the dependence caused by the latent component, constructing asymptotically independent test statistics is straightforward. This construction enables the application of traditional multiple testing procedures. Two commonly employed approaches are Benjamini-Hochberg's procedure (Benjamini and Hochberg, 1995) and Storey's procedure (Storey, 2002). Benjamini-Hochberg's method is fundamentally a sequential approach where p-values are arranged from the smallest to the largest. A data-driven cutoff point is then calculated to reject only the most significant hypotheses while controlling FDR throughout the process. Storey's approach involves fixing a threshold value t, estimating false discovery proportion (FDP), where the expectation FDP is FDR, for each choice of t, and then determining the threshold t such that the estimated FDP does not exceed a predefined significance level α. Although these approaches have different motivations,

[2] More precisely, a NC set is a collection of rows that are near to satisfy the null hypothesis in Eq.(1.2). Zero rows correspond to the case when contrast matrix $A = \mathbf{I}_p$.

their equivalence was derived in Storey et al. (2004). In this thesis, we will adopt Storey's approach. Our estimation of the FDP converges to the oracle FDP which utilizes information from the latent component.

Our theoretical contribution encompasses additional two aspects. Firstly, we explicitly assume a linear sparsity condition on the number of true nonzero rows of the coefficient matrix $\mathbf{B}$: the fraction of nonzero rows, no matter how small, does not vanish as m and n grow larger. Consequently, we depart from the common assumption that $\mathcal{P}_{\mathbf{\Gamma}_L}\mathbf{B} = o(1)$, as frequently employed in recent works (Lee et al., 2017; Wang et al., 2017; McKennan and Nicolae, 2019; Bing et al., 2023), where $\mathcal{P}_{\mathbf{\Gamma}_L}$ is the orthogonal projection matrix onto the column space of $\mathbf{\Gamma}_L$. The linear sparsity condition enables us to analyze FDR controlling property theoretically, a feature absent in the aforementioned works. We emphasize that the linear sparsity condition aligns with common beliefs in practice, even in genomic studies where the proportion of significant findings is very small and around $0.1\% \sim 1\%$. Secondly, extending existing analysis to account for the presence of the remainder term $\mathbf{F}$ is non-trivial. This challenge arises especially when the nonlinear latent component is not independent of, and thus confounded with $\mathbf{X}$.

The main contribution of this thesis is summarized as follows.

1. We present a nonlinear decomposition for the remainder term in the multiple responses linear regression model. Also, we establish a connection between the nonlinear decomposition and linear scenarios in terms of basis approximation.

2. We introduce a new method for estimating an NC set, taking into account the presence of confounding effects between $\mathbf{X}$ and the latent component. The NC set is useful in constructing pairwise asymptotically independent test statistics, thereby enabling us to achieve our goal of FDR control.

3. We relax the common assumption that $\mathcal{P}_{\mathbf{\Gamma}_L}\mathbf{B} = o(1)$. Our method can be extended to encompass certain classes of nonlinear latent components.

1.2. Related literature

First of all, the nonlinear factor representation in Eq.(1.5) is compatible with the linear factor representation in Eq.(1.3). Theoretically, many existing works have focused solely on dealing with the linear representation. Thus, we may be among the first to extend this idea to the nonlinear case.

The majority of relevant literature in this domain stems from the introduction of surrogate variable analysis (SVA), a concept pioneered by Jeffrey T. Leek in his P.h.D. thesis and subsequent series of papers (Leek, 2007; Leek and Storey, 2007, 2008; Leek et al., 2010). SVA proves valuable for modeling large-scale noise dependence caused by unmeasured or unmodeled factors. The authors also proposed an iteratively re-weighted algorithm (IRW-SVA), which utilizes bootstrap estimation of the probability of $\boldsymbol{b}_{i,\cdot}$ belonging to the null hypothesis in Eq.(1.2) to estimate an NC set and obtain estimations of surrogate variables. Likewise, Gagnon-Bartsch et al. (2013) proposed using existing NC sets. In practice, such knowledge of NC sets is typically transferred from other similar studies.

By around 2017, researchers shifted their focus towards methods that do not rely on NC set information. Instead, statisticians began to realize that estimating $\mathbf{\Gamma}_L$ is much easier than estimating $\mathbf{G}_L$. By initially identifying the column space of $\mathbf{\Gamma}_L$, it becomes feasible to recover $\mathbf{B}$ from $\mathcal{P}^{\perp}_{\mathbf{\Gamma}_L}\mathbf{B}$ provided that $\mathcal{P}_{\mathbf{\Gamma}_L}\mathbf{B} = o(1)$, where $\mathcal{P}^{\perp}_{\mathbf{\Gamma}_L} = \mathbf{I}_m - \mathcal{P}_{\mathbf{\Gamma}_L}$. Lee et al. (2017) proposed a coefficient adjustment procedure for estimating $\mathbf{\Gamma}_L$ and then recovered $\mathbf{B}$, assuming that the true $\mathbf{B}$ and $\mathbf{\Gamma}_L$ are asymptotically orthogonal after mean centering. Wang et al. (2017) first estimated $\mathbf{\Gamma}_L$ through a rotation technique, and then estimated $\mathbf{B}$ through robust regression, demonstrating its effectiveness in mitigating the confounding effect when $\mathbf{G}_L$ and $\mathbf{X}$ follows a structural linear equation model. McKennan and Nicolae (2019) extended this method and theoretical analysis to scenarios when the observed data lacks sufficient informativeness about the latent factors. Bing et al. (2022, 2023) offered non-asymptotic estimation error bounds and statistical inference results, addressing cases involving ultra high-dimensional $\mathbf{X}$ and sparse $\mathbf{\Gamma}$.

While effective, this framework is not without cost, particularly in the context of FDR control. On one hand, these processes typically yields correlated estimations of B, resulting in correlations between testing statistics derived from these estimations. However, mitigating these correlations is crucial for achieving the goal of controlling FDR in traditional multiple testing procedures. On the other hand, in order to satisfy the assumption that $\mathcal{P}_{\mathbf{\Gamma}_L}\mathbf{B} = o(1)$, many of the above works essentially require small magnitudes of $\mathbf{B}$ such as $\|\mathbf{B}\|_{1,2}\sqrt{n}/m = o(1)$, where $\|\cdot\|_{1,2}$ is the sum of ℓ_2 norms for row vectors. Consequently, the number of statistically significant nonzero rows with magnitudes at least $O(1/\sqrt{n})$ must be of order $o(m)$. This restricts the applicability of traditional FDR control framework, which assumes the number of alternative hypotheses grows linearly with m, thereby resulting in a lack of theoretical guarantee for FDR control.

The remaining parts of this thesis are organized as follows. Chapter 2 delves into latent factor representation for general covariance dependence, exploring both linear and nonlinear approaches. Chapter 3 outlines the models under consideration, addressing their identifiability issues. Chapter 4 introduces the procedure of our proposed method, detailing parameter estimations and hypothesis testings. Chapter 5 outlines technical assumptions as well as provides theoretical guarantees for each step proposed in Section 4. Chapter 6 encompasses numerical experiments designed to validate the effectiveness of our multiple testing procedure. Chapter 7 gives a brief discussion. All technical details and derivations are relegated to Appendices A and B.

Chapter 2

Latent effects modeling

In this chapter, we systematically describe our philosophy in modeling general covariance structure in unobserved effects. The main purpose is to provide theoretical justification for using the latent factor structure to handle correlations among different regressions. We will begin with a general representation theorem, which asserts that arbitrary covariance can always be viewed as generating from a class of nonlinear functions. We then conclude by addressing why relatively few linear latent factors can effectively capture most of the covariance structure in many situations.

We can rewrite model Eq.(1.1) into a matrix form as

$$\mathbf{Y} = \mathbf{B}\mathbf{X} + \mathbf{U}, \tag{2.1}$$

where

$$\mathbf{Y} = \begin{pmatrix} Y_{1,1} & \dots & Y_{1,n} \\ \vdots & \dots & \vdots \\ Y_{m,1} & \dots & Y_{m,n} \end{pmatrix} \in \mathbb{R}^{m\times n} = \left(\boldsymbol{Y}_{\boldsymbol{\cdot},1}, \dots, \boldsymbol{Y}_{\boldsymbol{\cdot},n}\right) = \begin{pmatrix} \boldsymbol{Y}_{1,\boldsymbol{\cdot}}^{\top} \\ \vdots \\ \boldsymbol{Y}_{m,\boldsymbol{\cdot}}^{\top} \end{pmatrix},$$

$$\mathbf{B} = \begin{pmatrix} b_{1,1} & \dots & b_{1,p} \\ \vdots & \dots & \vdots \\ b_{m,1} & \dots & b_{m,p} \end{pmatrix} \in \mathbb{R}^{m\times p} = \left(\boldsymbol{b}_{\boldsymbol{\cdot},1}, \dots, \boldsymbol{b}_{\boldsymbol{\cdot},p}\right) = \begin{pmatrix} \boldsymbol{b}_{1,\boldsymbol{\cdot}}^{\top} \\ \vdots \\ \boldsymbol{b}_{m,\boldsymbol{\cdot}}^{\top} \end{pmatrix},$$

$$\mathbf{X} = \begin{pmatrix} X_{1,1} & \dots & X_{1,n} \\ \vdots & \dots & \vdots \\ X_{p,1} & \dots & X_{p,n} \end{pmatrix} \in \mathbb{R}^{p\times n} = \Big(\boldsymbol{X}_{\boldsymbol{\cdot},1}, \dots, \boldsymbol{X}_{\boldsymbol{\cdot},n}\Big) = \begin{pmatrix} \boldsymbol{X}_{1,\boldsymbol{\cdot}}^\top \\ \vdots \\ \boldsymbol{X}_{p,\boldsymbol{\cdot}}^\top \end{pmatrix},$$

$$\mathbf{U} = \begin{pmatrix} u_{1,1} & \dots & u_{1,n} \\ \vdots & \dots & \vdots \\ u_{m,1} & \dots & u_{m,n} \end{pmatrix} \in \mathbb{R}^{m\times n} = \Big(\boldsymbol{u}_{\boldsymbol{\cdot},1}, \dots, \boldsymbol{u}_{\boldsymbol{\cdot},n}\Big) = \begin{pmatrix} \boldsymbol{u}_{1,\boldsymbol{\cdot}}^\top \\ \vdots \\ \boldsymbol{u}_{m,\boldsymbol{\cdot}}^\top \end{pmatrix}. \tag{2.2}$$

Alternatively, model Eq.(2.1) is a collection of m different linear regression models:

$$\begin{aligned} & \mathbf{Y}(i,:) = \mathbf{B}(i,:)\mathbf{X} + \mathbf{U}(i,:), \\ \iff \quad & \boldsymbol{Y}_{i,\boldsymbol{\cdot}}^\top = \boldsymbol{b}_{i,\boldsymbol{\cdot}}^\top \mathbf{X} + \boldsymbol{u}_{i,\boldsymbol{\cdot}}^\top, \\ \iff \quad & \boldsymbol{Y}_{i,\boldsymbol{\cdot}} = \mathbf{X}^\top \boldsymbol{b}_{i,\boldsymbol{\cdot}} + \boldsymbol{u}_{i,\boldsymbol{\cdot}}, \quad i = 1, \dots, m. \end{aligned} \tag{2.3}$$

Throughout this paper, we will assume that $(\boldsymbol{Y}_{\boldsymbol{\cdot},j},\ \boldsymbol{X}_{\boldsymbol{\cdot},j},\ \boldsymbol{u}_{\boldsymbol{\cdot},j})$, $j = 1, \dots, n$, are i.i.d. random vectors and they have finite covariances. $\mathbf{B}$ is the coefficient matrix. We emphasize that only $\mathbf{Y}$ and $\mathbf{X}$ are observable. Latent effects, and dependence among different regression models are quantified in $\mathbf{U}$.

2.1. Factor representation for general covariance dependence

For the model in Eq.(2.1), test statistics constructed directly from the observed $\mathbf{Y}$, $\mathbf{X}$ could be strongly correlated in applications due to (a) insufficiency in data collection to include important explanatory variables that explain variability in $\mathbf{Y}$, (b) unmeasured confounders that affect both $\mathbf{X}$ and $\mathbf{Y}$, and (c) intrinsic but unknown mechanism yielding the observations $\mathbf{Y}$.

We seek a direct decomposition of $\mathbf{U}$ into a sum of latent but structural part and a random error part, in the hope of removing any source of row-dependence by estimating

the latent part. An intuitive way of decomposition is to write

$$\mathbf{U} - \mathbb{E}[\mathbf{U}] = \mathbf{\Gamma}_L \mathbf{G}_L + \mathbf{E}, \tag{2.4}$$

where

$$\mathbf{\Gamma}_L = \begin{pmatrix} \gamma_{1,1}^{(L)} & \dots & \gamma_{1,r}^{(L)} \\ \vdots & \dots & \vdots \\ \gamma_{m,1}^{(L)} & \dots & \gamma_{m,r}^{(L)} \end{pmatrix} \in \mathbb{R}^{m\times r} = \left(\boldsymbol{\gamma}_{\boldsymbol{\cdot},1}^{(L)}, \dots, \boldsymbol{\gamma}_{\boldsymbol{\cdot},r}^{(L)}\right) = \begin{pmatrix} (\boldsymbol{\gamma}_{1,\boldsymbol{\cdot}}^{(L)})^\top \\ \vdots \\ (\boldsymbol{\gamma}_{m,\boldsymbol{\cdot}}^{(L)})^\top \end{pmatrix},$$

$$\mathbf{G}_L = \begin{pmatrix} g_{1,1}^{(L)} & \dots & g_{1,n}^{(L)} \\ \vdots & \dots & \vdots \\ g_{r,1}^{(L)} & \dots & g_{r,n}^{(L)} \end{pmatrix} \in \mathbb{R}^{r\times n} = \left(\boldsymbol{g}_{\boldsymbol{\cdot},1}^{(L)}, \dots, \boldsymbol{g}_{\boldsymbol{\cdot},n}^{(L)}\right) = \begin{pmatrix} (\boldsymbol{g}_{1,\boldsymbol{\cdot}}^{(L)})^\top \\ \vdots \\ (\boldsymbol{g}_{r,\boldsymbol{\cdot}}^{(L)})^\top \end{pmatrix},$$

$$\mathbf{E} = \begin{pmatrix} \epsilon_{1,1} & \dots & \epsilon_{1,n} \\ \vdots & \dots & \vdots \\ \epsilon_{m,1} & \dots & \epsilon_{m,n} \end{pmatrix} \in \mathbb{R}^{m\times n} = \left(\boldsymbol{\epsilon}_{\boldsymbol{\cdot},1}, \dots, \boldsymbol{\epsilon}_{\boldsymbol{\cdot},n}\right) = \begin{pmatrix} \boldsymbol{\epsilon}_{1,\boldsymbol{\cdot}}^\top \\ \vdots \\ \boldsymbol{\epsilon}_{m,\boldsymbol{\cdot}}^\top \end{pmatrix}.$$

$\mathbf{\Gamma}_L\mathbf{G}_L$ is interpreted as low-rank part, and $\mathbf{E}$ as idiosyncratic part.

Apparently, the decomposition Eq.(2.4) is always possible and not unique. For our purpose, a useful decomposition should have the following two properties: (i) $\boldsymbol{\epsilon}_{i,\boldsymbol{\cdot}}$ are uncorrelated with each other, ensuring that all levels of row-dependence between different regression models are quantified by the low-rank part; (ii) there is a possible way to estimate $\mathbf{\Gamma}_L$ and $\mathbf{G}_L$. For the first property, Leek and Storey (2008) presented a decomposition where rows of $\mathbf{E}$ are independent. But it is not known whether their decomposition yields an identifiable low-rank component. To satisfy the second property, we may consider an easier problem where $\mathbf{U}$ is observable to us. In this case, Eq.(2.4) resembles factor models considered in Bai (2003), enabling technical conditions outlined therein to describe a useful decomposition. For example, we may also consider the following additional conditions:

$$\mathbb{E}[\boldsymbol{\epsilon}_{\boldsymbol{\cdot},j}] = \mathbf{0}, \quad \mathrm{Cov}(\boldsymbol{g}_{\boldsymbol{\cdot},j}^{(L)}, \boldsymbol{\epsilon}_{\boldsymbol{\cdot},j}) = \mathbf{0},$$

$$\mathrm{Cov}(\boldsymbol{g}_{\boldsymbol{\cdot},j}^{(L)}) = \mathbf{I}_r, \quad \mathbf{D}_{\mathbf{E}} := \mathrm{Cov}(\boldsymbol{\epsilon}_{\boldsymbol{\cdot},j}) \text{ is diagonal,}$$

$$\boldsymbol{\Sigma}_{\mathbf{U}} := \mathrm{Cov}(\boldsymbol{u}_{\boldsymbol{\cdot},j}) = \boldsymbol{\Gamma}_L \boldsymbol{\Gamma}_L^\top + \mathbf{D}_{\mathbf{E}}. \tag{2.5}$$

The following theorem extends Proposition 1 in Leek and Storey (2008), provided that our decomposition satisfies Eq.(2.4), property (i) as well as the moment conditions outlined in Eq.(2.5). We also explore the feasibility of nonlinear factor decomposition beyond the linear counterpart.

Theorem 2.1. [1] *For model* (2.1)*, there exist latent factor matrix* $\mathbf{G}_L \in \mathbb{R}^{r\times n}$*, an error matrix* $\mathbf{E} \in \mathbb{R}^{m\times n}$ *such that each column of* $\mathbf{G}_L$*,* $\mathbf{E}$ *are i.i.d., achieved by suitably enlarging underlying probability space. Also,*

(1) $\mathbb{E}[\boldsymbol{g}_{\boldsymbol{\cdot},j}^{(L)}] = \mathbf{0}$*,* $\mathbb{E}[\boldsymbol{\epsilon}_{\boldsymbol{\cdot},j}] = \mathbf{0}$*,* $\mathrm{Cov}(\boldsymbol{g}_{\boldsymbol{\cdot},j}^{(L)}, \boldsymbol{\epsilon}_{\boldsymbol{\cdot},j}) = \mathbf{0}$*,* $\mathrm{Cov}(\boldsymbol{g}_{\boldsymbol{\cdot},j}^{(L)}) = \mathbf{I}_r$*,* $\mathbf{D}_{\boldsymbol{E}} = \mathrm{Cov}(\boldsymbol{\epsilon}_{\boldsymbol{\cdot},j})$ *is diagonal,*

$$\mathbf{U} - \mathbb{E}[\mathbf{U}] = \boldsymbol{\Gamma}_L \mathbf{G}_L + \mathbf{E} \quad a.s. \tag{2.6}$$

where $\boldsymbol{\Gamma}_L \in \mathbb{R}^{m\times r}$ *is non-random and* $\mathrm{rank}(\boldsymbol{\Gamma}_L) = r \leq m-1$*. The distribution of* $\boldsymbol{\epsilon}_{\boldsymbol{\cdot},j}$ *is non-degenerate if and only if* $\exists\, i = 1, \ldots, m$ *such that* $\boldsymbol{e}_i \in \mathrm{Col}(\boldsymbol{\Sigma}_{\mathbf{U}})$*, the column space of* $\boldsymbol{\Sigma}_{\mathbf{U}}$*, where* $\boldsymbol{e}_i$ *is the standard basis for* $\mathbb{R}^m$*.*

(2) There exists a vector-valued Borel function $\boldsymbol{f} = (f_1, \ldots, f_m)^\top \in \mathbb{R}^m$*, and a latent factor matrix* $\mathbf{A} = (\boldsymbol{\alpha}_{\boldsymbol{\cdot},1}, \ldots, \boldsymbol{\alpha}_{\boldsymbol{\cdot},n})^\top \in \mathbb{R}^{k\times n}$*, such that* $\boldsymbol{\alpha}_{\boldsymbol{\cdot},j}$ *are i.i.d.,* $\mathbb{E}\big[f_i^2(\boldsymbol{\alpha}_{\boldsymbol{\cdot},j})\big]$ *is finite* $\forall i$ *and*

$$\boldsymbol{\Gamma}_L \mathbf{G}_L + \mathbb{E}[\mathbf{U}] = \boldsymbol{f}(\mathbf{A}) \quad a.s. \tag{2.7}$$

$$\mathbf{U} = \boldsymbol{f}(\mathbf{A}) + \mathbf{E} \quad a.s. \tag{2.8}$$

[1] Variables and functions constructed in this proposition depend on m in general. We will suppress their superscripts (m) for simplicity.

where

$$\boldsymbol{f}(\mathbf{A}) = \begin{pmatrix} f_1(\boldsymbol{\alpha}_{\cdot,1}) & \dots & f_1(\boldsymbol{\alpha}_{\cdot,n}) \\ \vdots & \dots & \vdots \\ f_m(\boldsymbol{\alpha}_{\cdot,1}) & \dots & f_m(\boldsymbol{\alpha}_{\cdot,n}) \end{pmatrix} \in \mathbb{R}^{m\times n} = \begin{pmatrix} f_1(\mathbf{A})^\top \\ \vdots \\ f_m(\mathbf{A})^\top \end{pmatrix}.$$

In fact, the latent factor $\mathbf{A}$ *in the nonlinear representation Eq.*(2.8) *can be chosen such that it has only one row* $(k = 1)$ *and* $\boldsymbol{\alpha}_{\cdot,j}$ *are* i.i.d. $\mathrm{Uniform}(0, 1)$.

Theorem 2.1 shows that $\boldsymbol{u}_{\cdot,j}$ can be decomposed in both linear and nonlinear ways. Both Eq.(2.6) and Eq.(2.8) depict general dependence in the sense that no additional restrictions were imposed on $\mathbf{U}$ beyond basic distributional assumptions.

Eq.(2.6) is referred to as an *exact factor model* since the idiosyncratic error $\mathbf{E}$ does not introduce any correlation among different measurements. In many applications, researchers often believe that the number of latent components $r = \mathrm{rank}(\mathbf{G}_L)$ is small. Finding the smallest number r such that Eq.(2.5) and Eq.(2.6) hold has a long-standing history and is known as the *minimum rank factor problem*. Ledermann (1937) made a notable attempt at solving this problem, but it remains unresolved to this day. If entries of $\boldsymbol{\Sigma}_{\mathbf{U}}$ are generated according to some continuous distributions such as Gaussian, the minimum rank is no less than the *Ledermann's bound* $(\approx \Theta(m))$ with probability 1 (Shapiro, 1982). This implies that an exact low-rank assumption on $\mathrm{rank}(\mathbf{G}_L)$ cannot describe more general dependence. Conversely, if the smallest r is of order $\Theta(m)$, precise estimation of latent factors becomes unlikely due to oversized parameter space. Therefore, a more practical approach is to find a relatively small r such that Eq.(2.5) and Eq.(2.6) hold approximately.

Eq.(2.8) represents general dependence from another angle. Different entries in $\boldsymbol{u}_{\cdot,j}$ are dependent only through a low-dimensional latent variable $\boldsymbol{\alpha}_{\cdot,j}$, and a class of nonlinear functions. Viewed as data generation processes, Eq.(2.8) could be preferred because the latent variable $\boldsymbol{\alpha}_{\cdot,j}$ could have the same distribution for varying m, while in general, there is no such guarantee for $\boldsymbol{g}^{(L)}_{\cdot,j}$. This brings benefit in justifying why relatively few components can approximate general covariance dependence. For example, if $\boldsymbol{\alpha}_{\cdot,j} \in [0, 1]$, then we can use any basis from $L^2[0, 1]$ to approximate $f_1, \dots, f_m$ simultaneously. When these functions

possess some "nice" properties, the approximation can be uniform. This connection between approximately linear and nonlinear factor representations is crucial and will be explored further in Section 2.2.

2.2. Uniform approximation using RKHS

Equip Eq.(2.1) with the nonlinear representation:

$$\boldsymbol{u}_{\boldsymbol{\cdot},j} = \boldsymbol{f}(\boldsymbol{\alpha}_{\boldsymbol{\cdot},j}) + \boldsymbol{\epsilon}_{\boldsymbol{\cdot},j},$$

where $\boldsymbol{\alpha}_{\boldsymbol{\cdot},j}$ could be k-dimensional, $f_1, \ldots, f_m$ are functions from $\mathbb{R}^k \to \mathbb{R}$. Assuming that $f_i(\boldsymbol{\alpha}_{\boldsymbol{\cdot},j})$ has finite variance for all $i = 1, \ldots, m$. Let $\widetilde{f}_i$ be any approximation to f_i, and $\widetilde{Y}_{i,j} = Y_{i,j} - \widetilde{f}_i(\boldsymbol{\alpha}_{\boldsymbol{\cdot},j})$. The OLS estimation of $\mathbf{B}$ based on regression of $\widetilde{\mathbf{Y}}$ onto $\mathbf{X}$ is consistent provided that

$$\max_{1 \le i \le m} \mathbb{E}\big[\widetilde{f}_i(\boldsymbol{\alpha}_{\boldsymbol{\cdot},j}) - f_i(\boldsymbol{\alpha}_{\boldsymbol{\cdot},j})\big]^2 \to 0.$$

It is thus natural to consider space $L^2(\mathbb{P}^{\boldsymbol{\alpha}})$, where $\mathbb{P}^{\boldsymbol{\alpha}}$ is the distribution law of $\boldsymbol{\alpha}_{\boldsymbol{\cdot},j}$ on $\mathbb{R}^k$. Let $\{\zeta_v(\boldsymbol{s})\}_{v=1}^{\infty}$ be any orthogonal basis in this space. Expand f_i in $L^2(\mathbb{P}^{\boldsymbol{\alpha}})$ w.r.t. the basis up to the K-th order:

$$f_i(\boldsymbol{\alpha}_{\boldsymbol{\cdot},j}) = \sum_{v=1}^{K} \langle f_i, \zeta_v \rangle_{L^2(\mathbb{P}^{\boldsymbol{\alpha}})} \zeta_v(\boldsymbol{\alpha}_{\boldsymbol{\cdot},j}) + f_i^{(K)}(\boldsymbol{\alpha}_{\boldsymbol{\cdot},j}) = \text{ linear part + remainder}, \tag{2.9}$$

where $\langle \cdot, \cdot \rangle_{L^2(\mathbb{P}^{\boldsymbol{\alpha}})}$ is the inner product in the Hilbert space $L^2(\mathbb{P}^{\boldsymbol{\alpha}})$. The uniform approximation error in L^2 is

$$\xi_K := \max_{1 \le i \le m} \mathbb{E}\Big[\big(f_i^{(K)}(\boldsymbol{\alpha}_{\boldsymbol{\cdot},j})\big)^2\Big]. \tag{2.10}$$

In general, ξ_K does not necessarily converge to zero, but this could be true if $f_1, \ldots, f_m$ reside in a suitable subset. The subset we choose is a Reproducing Kernel Hilbert Space (RKHS). The following proposition is adapted from Wahba (1990).

Proposition 2.2. *Let $\mathcal{T}$ the support of $\mathbb{P}^{\boldsymbol{\alpha}}$. Suppose $\mathcal{T}$ is compact, and $R(\boldsymbol{s},\boldsymbol{t})$ is a symmetric, continuous, positive-definite and square-integrable bivariate function on $\mathcal{T}^2$. The RKHS $\mathcal{H}_R$ corresponding to R is associated with the following inner product:*

$$\langle h_1, h_2\rangle_{\mathcal{H}_R} = \sum_{v=1}^{\mathrm{rank}(R)} \frac{\langle h_1, \Phi_v\rangle_{L^2(\mathbb{P}^{\boldsymbol{\alpha}})} \cdot \langle h_2, \Phi_v\rangle_{L^2(\mathbb{P}^{\boldsymbol{\alpha}})}}{\lambda_v}. \tag{2.11}$$

Here, λ_v represents the v-th largest eigenvalue, Φ_v is the associated eigenfunction, and $rank(R)$ denotes the number of eigenvalues of R. $\mathcal{H}_R$ consists of any function $h \in L^2(\mathbb{P}^{\boldsymbol{\alpha}})$ such that $\|h\|_{\mathcal{H}_R} := \sqrt{\langle h, h\rangle_{\mathcal{H}_R}} < \infty$, and h belongs to the closure of the vector space spanned by $\{\Phi_v\}_{v=1}^{\mathrm{rank}(R)}$.

For the remainder of this section, we will use the notations and conditions given in Proposition 2.2. Proposition 2.2 implies that $\mathcal{H}_R$ is a proper subset of $L^2(\mathbb{P}^{\boldsymbol{\alpha}})$. Reconsider Eq.(2.9). When $f_1, \ldots, f_m \in \mathcal{H}_R$, and $\zeta_v = \Phi_v$ for $v = 1, \ldots, \mathrm{rank}(R)$,

$$\mathbb{E}\Big[\big(f_i^{(K)}(\boldsymbol{\alpha}_{\boldsymbol{\cdot},j})\big)^2\Big] = \sum_{v=K+1}^{\mathrm{rank}(R)} \langle f_i, \Phi_v\rangle^2_{L^2(\mathbb{P}^{\boldsymbol{\alpha}})} \le \lambda_{K+1} \sum_{v=K+1}^{\mathrm{rank}(R)} \frac{\langle f_i, \Phi_v\rangle^2_{L^2(\mathbb{P}^{\boldsymbol{\alpha}})}}{\lambda_v} \le \lambda_{K+1} \max_{1\le i\le m} \|f_i\|^2_{\mathcal{H}_R}. \tag{2.12}$$

Note that Eq.(2.12) holds uniformly over $i = 1, \ldots, m$. If $\mathrm{rank}(R)$ is infinite, it is known that $\lim_{K\to\infty} \lambda_{K+1} = 0$. The uniform L^2 approximation error $\xi_K \to 0$ as $K \to \infty$ provided that $\max_{1\le i\le m} \|f_i\|^2_{\mathcal{H}_R}$ is bounded. If $\mathrm{rank}(R)$ is finite, $\xi_K = 0$ as long as $K = \mathrm{rank}(R)$.

Using a similar argument, pointwise approximation error can also be uniformly controlled.

$$\begin{aligned}
\big(f_i^{(K)}(\boldsymbol{\alpha}_{\boldsymbol{\cdot},j})\big)^2 &= \Big(\sum_{v=K+1}^{\mathrm{rank}(R)} \langle f_i, \Phi_v\rangle_{L^2(\mathbb{P}^{\boldsymbol{\alpha}})} \Phi_v(\boldsymbol{\alpha}_{\boldsymbol{\cdot},j})\Big)^2 \\
&\le \sum_{v=K+1}^{\mathrm{rank}(R)} \frac{\langle f_i, \Phi_v\rangle^2_{L^2(\mathbb{P}^{\boldsymbol{\alpha}})}}{\lambda_v} \cdot \sum_{v=K+1}^{\mathrm{rank}(R)} \lambda_v \Phi_v^2(\boldsymbol{\alpha}_{\boldsymbol{\cdot},j}) \\
&\le a_{K+1} \max_{1\le i\le m} \|f_i\|^2_{\mathcal{H}_R}.
\end{aligned} \tag{2.13}$$

Here, $a_{K+1} := \max_{\boldsymbol{s}\in\mathcal{T}} \sum_{v=K+1}^{\mathrm{rank}(R)} \lambda_v \Phi_v^2(\boldsymbol{s})$ is the uniform convergence rate for the function series $\sum_{v=1}^{K} \lambda_v \Phi_v^2(\boldsymbol{s})$, whose limit function is $R(\boldsymbol{s},\boldsymbol{s})$. If $\mathrm{rank}(R)$ is infinite, $\lim_{K\to\infty} a_{K+1} = 0$

is guaranteed by Mercer's theorem. For more details about RKHS and the associated eigen-decomposition, we refer readers to Dunford and Schwartz (1965); Sun (2005); Wahba (1990).

In summary, Eq.(2.12) and Eq.(2.13) demonstrate the potential to uniformly approximate a system of nonlinear functions using relatively few linear components, i.e. $K \ll m$. Eq.(2.12) indicates that the corresponding covariance matrix of $\left(f_1^{(K)}(\boldsymbol{\alpha}_{\bullet,j}), \ldots, f_m^{(K)}(\boldsymbol{\alpha}_{\bullet,j})\right)^\top$ is weakly dependent, as defined in Fan et al. (2012). This helps explain why a few linear latent factors suffice for capturing general covariance dependence caused by potential nonlinear factors, as observed in a numerical example in Fan et al. (2012).

Additionally, our results align well with the philosophy behind the use of *approximate factor models* (Bai and Ng, 2002; Bai, 2003), which capture majority of sample covariance using a few approximate factors, and allow for some correlations in the idiosyncratic error terms. In our decomposition, the orthogonal basis can be seen as the approximate factors. If $\widetilde{\epsilon}_{i,j} := f_i^{(K)}(\boldsymbol{\alpha}_{\bullet,j}) + \epsilon_{i,j}$ is considered as the true idiosyncratic error term, then the approximation error $f_i^{(K)}(\boldsymbol{\alpha}_{\bullet,j})$ is exactly the source that incurs correlations among $\widetilde{\epsilon}_{i,j}$ over $i = 1, \ldots, m$. Even though the exact factor structure could be large-dimensional, assuming an approximate, relatively low-dimensional structure is often justifiable in many situations. Therefore, our results illustrate the power of using latent factor structure to deal with general covariance.

Chapter 3

Models and identification

The main objective of this chapter is to discuss the identifiability of the parameters involved in our nonlinear representation model presented in Chapter 2. This may appear impossible at first glance since the latent component is not observable. However, by carefully considering the relationships between observed and unobserved variables, we can devise strategies to infer these parameters, thereby revealing the underlying structure of the model.

The nonlinear representation model has the following form:

$$\mathbf{Y} = \mathbf{B}\mathbf{X} + \boldsymbol{f}(\mathbf{A}) + \mathbf{E}, \tag{3.1}$$

where $\mathbf{Y}$, $\mathbf{X}$, $\mathbf{A}$, $\mathbf{E}$ are random matrices with i.i.d. columns. Hereafter, we will assume that $\boldsymbol{\epsilon}_{\bullet,j}$ is independent of $\boldsymbol{X}_{\bullet,j}$ and $\boldsymbol{\alpha}_{\bullet,j}$, and $\mathrm{Cov}(\boldsymbol{\epsilon}_{\bullet,j}) = \mathbf{D}_{\mathbf{E}} = diag(\sigma_1^2, \ldots, \sigma_m^2)$ is a diagonal matrix. When $f_1, \ldots, f_m$ originate from a good subset in $L^2(\mathbb{P}^{\boldsymbol{\alpha}})$, $\boldsymbol{f}(\mathbf{A})$ can be uniformly approximated using linear combinations of some orthogonal basis functions.

In light of Eq.(2.12) and Eq.(2.13), let

$$\boldsymbol{\gamma}_{i,v} = \langle f_i, \Phi_v \rangle_{L^2(\mathbb{P}^{\boldsymbol{\alpha}})},\ i = 1, \ldots, m,\ v = 1, \ldots, K,$$

$$\boldsymbol{g}_{v,j} = \Phi_v(\boldsymbol{\alpha}_{\bullet,j}),\ v = 1, \ldots, K,\ j = 1, \ldots, n,$$

$$\mathbf{\Gamma} = \begin{pmatrix} \gamma_{1,1} & \dots & \gamma_{1,K} \\ \vdots & \dots & \vdots \\ \gamma_{m,1} & \dots & \gamma_{m,K} \end{pmatrix} \in \mathbb{R}^{m\times K} = \Big(\boldsymbol{\gamma}_{\boldsymbol{\cdot},1}, \dots, \boldsymbol{\gamma}_{\boldsymbol{\cdot},K}\Big) = \begin{pmatrix} \boldsymbol{\gamma}_{1,\boldsymbol{\cdot}}^\top \\ \vdots \\ \boldsymbol{\gamma}_{m,\boldsymbol{\cdot}}^\top \end{pmatrix},$$

$$\mathbf{G} = \begin{pmatrix} g_{1,1} & \dots & g_{1,n} \\ \vdots & \dots & \vdots \\ g_{K,1} & \dots & g_{K,n} \end{pmatrix} \in \mathbb{R}^{K\times n} = \Big(\boldsymbol{g}_{\boldsymbol{\cdot},1}, \dots, \boldsymbol{g}_{\boldsymbol{\cdot},n}\Big) = \begin{pmatrix} \boldsymbol{g}_{1,\boldsymbol{\cdot}}^\top \\ \vdots \\ \boldsymbol{g}_{K,\boldsymbol{\cdot}}^\top \end{pmatrix},$$

$$\mathbf{F} = \boldsymbol{f}(\mathbf{A}) - \mathbf{\Gamma}\mathbf{G} = \begin{pmatrix} f_1^{(K)}(\boldsymbol{\alpha}_{\boldsymbol{\cdot},1}) & \dots & f_1^{(K)}(\boldsymbol{\alpha}_{\boldsymbol{\cdot},n}) \\ \vdots & \dots & \vdots \\ f_m^{(K)}(\boldsymbol{\alpha}_{\boldsymbol{\cdot},1}) & \dots & f_m^{(K)}(\boldsymbol{\alpha}_{\boldsymbol{\cdot},n}) \end{pmatrix} \in \mathbb{R}^{m\times n} = \Big(\boldsymbol{f}_{\boldsymbol{\cdot},1}, \dots, \boldsymbol{f}_{\boldsymbol{\cdot},n}\Big) = \begin{pmatrix} \boldsymbol{f}_{1,\boldsymbol{\cdot}}^\top \\ \vdots \\ \boldsymbol{f}_{m,\boldsymbol{\cdot}}^\top \end{pmatrix}.^{1}$$

Then, Eq.(3.1) becomes

$$\mathbf{Y} = \mathbf{B}\mathbf{X} + \mathbf{\Gamma}\mathbf{G} + \mathbf{F} + \mathbf{E}. \tag{3.2}$$

Once again, we emphasize that $\mathbf{G}$, $\mathbf{F}$, $\mathbf{E}$ are not observable. It is worth noting that this model representation encompasses pure linear case when approximation error term $\mathbf{F} = \mathbf{0}$. In this thesis, we assume that $p = \operatorname{rank}(\mathbf{X})$ is fixed or low-dimensional, i.e. $p \ll n$. The number of latent basis K is known or pre-chosen, which is a common assumption in the factor model literature. K viewed as a function of m, n can diverge to ∞. By transformation, we can always assume that $\operatorname{Cov}(\boldsymbol{g}_{\boldsymbol{\cdot},j}) = \mathbf{I}_K$. We can also assume that $\operatorname{rank}(\mathbf{\Gamma}) = K$ since otherwise, the corresponding components in $\boldsymbol{g}_{\boldsymbol{\cdot},j}$ can be dropped.

Before diving into the discussion of identification, a few properties about $\mathbf{F}$ should be noted. As a collection of remainder terms in $L^2(\mathbb{P}^{\boldsymbol{\alpha}})$, $\mathbf{F}$ automatically satisfies that $\operatorname{Cov}(\boldsymbol{f}_{\boldsymbol{\cdot},j}, \boldsymbol{g}_{\boldsymbol{\cdot},j}) = \mathbf{0}$, $\forall i = 1, \dots, m$. Next, we illustrate that $\operatorname{Cov}(\boldsymbol{f}_{\boldsymbol{\cdot},j}, \boldsymbol{X}_{\boldsymbol{\cdot},j}) = \mathbf{0}$ is also true if latent factor $\boldsymbol{g}_{\boldsymbol{\cdot},j}$ is augmented with $h_i(\boldsymbol{\alpha}_{\boldsymbol{\cdot},j})$, the conditional expectation of $X_{i,j}$ given $\boldsymbol{\alpha}_{\boldsymbol{\cdot},j}$, $i = 1, \dots, p$. Let $\mathcal{V}$ be the linear subspace of $L^2(\mathbb{P}^{\boldsymbol{\alpha}})$ spanned by $h_1, \dots, h_p$, then

$$L^2(\mathbb{P}^{\boldsymbol{\alpha}}) = \mathcal{V} \oplus \mathcal{V}^\perp,$$

[1] We slightly abuse the notation $\boldsymbol{f}$ here. We emphasize that whenever $\boldsymbol{f}$ is associated with double subscripts such as $\boldsymbol{f}_{\boldsymbol{\cdot},i}$ or $\boldsymbol{f}_{j,\boldsymbol{\cdot}}$, it refers to the corresponding row vector or column vector of $\mathbf{F}$ respectively.

where $\oplus$ represents direct sum of two mutually orthogonal Hilbert spaces. Let $\boldsymbol{h}(\boldsymbol{\alpha}_{\boldsymbol{\cdot},j})$ be the vector stacked by $h_i(\boldsymbol{\alpha}_{\boldsymbol{\cdot},j})$ over i. This implies that

$$\boldsymbol{f}(\boldsymbol{\alpha}_{\boldsymbol{\cdot},j}) = \boldsymbol{\Gamma}^* \boldsymbol{h}(\boldsymbol{\alpha}_{\boldsymbol{\cdot},j}) + \boldsymbol{f}^*(\boldsymbol{\alpha}_{\boldsymbol{\cdot},j}) \tag{3.3}$$

for some $m \times p$ loading matrix $\boldsymbol{\Gamma}^*$ and $\boldsymbol{f}^* = (f_1^*, \ldots, f_m^*)^\top$ where $f_i^* \in \mathcal{V}^\perp$. Apply the uniform approximation scheme outlined in Section 2.2 to $\boldsymbol{f}^*$, and let $\boldsymbol{g}_{\boldsymbol{\cdot},j}$ in Eq.(3.2) include $\boldsymbol{h}(\boldsymbol{\alpha}_{\boldsymbol{\cdot},j})$ as well as first K orthogonal basis functions in $\mathcal{V}^\perp$. The approximation remainder $\boldsymbol{f}_{\boldsymbol{\cdot},j}$ must satisfy $\mathrm{Cov}\big(\boldsymbol{f}_{\boldsymbol{\cdot},j}, \boldsymbol{h}(\boldsymbol{\alpha}_{\boldsymbol{\cdot},j})\big) = \mathbf{0}$. Then, $\mathrm{Cov}(\boldsymbol{f}_{\boldsymbol{\cdot},j}, \boldsymbol{X}_{\boldsymbol{\cdot},j}) = \mathbf{0}$ follows because $\boldsymbol{X}_{\boldsymbol{\cdot},j} - \boldsymbol{h}(\boldsymbol{\alpha}_{\boldsymbol{\cdot},j})$ is uncorrelated with any functions of $\boldsymbol{\alpha}_{\boldsymbol{\cdot},j}$. The above argument justifies that the approximation error term $\mathbf{F}$ does not introduce endogeneity to the idiosyncratic error, even if $\mathbf{X}$ is confounded with $\boldsymbol{f}(\mathbf{A})$ as quantified by $\boldsymbol{\Gamma}^*$ in Eq.(3.3). This provides the foundation for the identification of $\mathbf{B}$.

To facilitate further discussions, we will decompose $\mathbf{BX}$ into two parts: the first part, which is our primary focus for hypothesis testing, and the second part, which is a nuisance component. Since the constraint matrix A is of full rank, we can always find a $(p-c) \times p$ matrix $A_\perp$ such that each row of $A_\perp$ is orthogonal to $\mathrm{Col}(A^\top)$, the column space of $A^\top$, and that $(A^\top, A_\perp^\top)$ is invertible. Let $\mathbf{M}_* = (A^\top, A_\perp^\top)$, and

$$\mathbf{M}_*^{-1}\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix},$$

where $\mathbf{X}_1$ is $c \times n$, $\mathbf{X}_2$ is $(p-c) \times n$. Then, for $\mathbf{B}_1 = \mathbf{B}A^\top$, $\mathbf{B}_2 = \mathbf{B}A_\perp^\top$,

$$\mathbf{BX} = \mathbf{B}_1\mathbf{X}_1 + \mathbf{B}_2\mathbf{X}_2. \tag{3.4}$$

Let

$$\mathbf{B}_1 = \begin{pmatrix} b_{1,1}^{(1)} & \ldots & b_{1,c}^{(1)} \\ \vdots & \ldots & \vdots \\ b_{m,1}^{(1)} & \ldots & b_{m,c}^{(1)} \end{pmatrix} \in \mathbb{R}^{m \times c} = \left(\boldsymbol{b}_{\boldsymbol{\cdot},1}^{(1)}, \ldots, \boldsymbol{b}_{\boldsymbol{\cdot},c}^{(1)}\right) = \begin{pmatrix} (\boldsymbol{b}_{1,\boldsymbol{\cdot}}^{(1)})^\top \\ \vdots \\ (\boldsymbol{b}_{m,\boldsymbol{\cdot}}^{(1)})^\top \end{pmatrix},$$

$$\mathbf{B}_2 = \begin{pmatrix} b_{1,1}^{(2)} & \dots & b_{1,p-c}^{(2)} \\ \vdots & \dots & \vdots \\ b_{m,1}^{(2)} & \dots & b_{m,p-c}^{(2)} \end{pmatrix} \in \mathbb{R}^{m\times(p-c)} = \left(\boldsymbol{b}_{\boldsymbol{\cdot},1}^{(2)}, \dots, \boldsymbol{b}_{\boldsymbol{\cdot},p-c}^{(2)}\right) = \begin{pmatrix} (\boldsymbol{b}_{1,\boldsymbol{\cdot}}^{(2)})^\top \\ \vdots \\ (\boldsymbol{b}_{m,\boldsymbol{\cdot}}^{(2)})^\top \end{pmatrix}.$$

The original testing problem Eq.(1.2) is equivalent to testing the following:

$$H_{0i} : \boldsymbol{b}_{i,\boldsymbol{\cdot}}^{(1)} = \mathbf{0} \quad \text{v.s.} \quad H_{1i} : \boldsymbol{b}_{i,\boldsymbol{\cdot}}^{(1)} \neq \mathbf{0} \quad i = 1, \dots, m.$$

Therefore, the collection of parameters is $\boldsymbol{\Theta} = (\mathbf{B}_1, \mathbf{B}_2, \boldsymbol{\Gamma}, \mathbf{D}_\mathbf{E})$. However, we do not aim to identify the exact values of all parameters. Specifically, the nuisance parameter $\mathbf{B}_2$ is not of interest. Because any left multiplication of orthogonal matrix to $\mathbf{G}$ does not affect its covariance, the exact value of $\boldsymbol{\Gamma}$ is ambiguous and hard to identify. We will only consider identifying its column space $\mathrm{Col}(\boldsymbol{\Gamma})$, which can be quantified by its associated projection matrix $\mathcal{P}_{\boldsymbol{\Gamma}}$.

3.1. Identifiability problem

This section discusses the identifiability of $\mathrm{Col}(\boldsymbol{\Gamma})$, $\mathbf{D}_\mathbf{E}$ and $\mathbf{B}_1$. The identification procedures presented in Section 3.1.1 and 3.1.2 draw inspiration from Wang et al. (2017); Bing et al. (2023). In section 3.1.3, we will talk about a difficulty encountered of applying these procedures with an inexact $\mathrm{Col}(\boldsymbol{\Gamma})$. This issue motivates the aforementioned works to impose the technical condition $\mathcal{P}_{\boldsymbol{\Gamma}}\mathbf{B}_1 = o(1)$.

3.1.1. Identifiability for $\mathrm{Col}(\boldsymbol{\Gamma})$ and $\mathbf{D}_\mathbf{E}$

Consider the column-wise representation of Eq.(3.2):

$$\boldsymbol{Y}_{\boldsymbol{\cdot},j} = \mathbf{B}\boldsymbol{X}_{\boldsymbol{\cdot},j} + \boldsymbol{\Gamma}\boldsymbol{g}_{\boldsymbol{\cdot},j} + \boldsymbol{f}_{\boldsymbol{\cdot},j} + \boldsymbol{\epsilon}_{\boldsymbol{\cdot},j}. \tag{3.5}$$

Let $\mathbf{\Phi} = \mathbb{E}\big[\boldsymbol{g}_{\boldsymbol{\cdot},j}\boldsymbol{X}_{\boldsymbol{\cdot},j}^{\top}\big] \cdot \mathbb{E}\big[\boldsymbol{X}_{\boldsymbol{\cdot},j}\boldsymbol{X}_{\boldsymbol{\cdot},j}^{\top}\big]^{-1}$, we can decompose $\boldsymbol{g}_{\boldsymbol{\cdot},j}$ as:

$$\boldsymbol{g}_{\boldsymbol{\cdot},j} = \mathbf{\Phi}\boldsymbol{X}_{\boldsymbol{\cdot},j} + \boldsymbol{W}_{\boldsymbol{\cdot},j}, \tag{3.6}$$

where $\mathrm{Cov}(\boldsymbol{W}_{\boldsymbol{\cdot},j}, \boldsymbol{X}_{\boldsymbol{\cdot},j}) = \mathbf{0}$.

By finding the best linear projection of $\boldsymbol{Y}_{\boldsymbol{\cdot},j}$ onto $\boldsymbol{X}_{\boldsymbol{\cdot},j}$, the marginal effect of $\boldsymbol{X}_{\boldsymbol{\cdot},j}$ on $\boldsymbol{Y}_{\boldsymbol{\cdot},j}$, $\mathbf{B} + \mathbf{\Gamma\Phi}$ can be identified. Then,

$$\mathrm{Cov}\big(\boldsymbol{Y}_{\boldsymbol{\cdot},j} - (\mathbf{B} + \mathbf{\Gamma\Phi})\boldsymbol{X}_{\boldsymbol{\cdot},j}\big) = \mathbf{\Gamma}\mathbf{\Sigma}_{\mathbf{W}}\mathbf{\Gamma}^{\top} + \mathbf{\Sigma}_{\mathbf{F}} + \mathbf{D}_{\mathbf{E}}, \tag{3.7}$$

where $\mathbf{\Sigma}_{\mathbf{W}} = \mathbb{E}\big[\boldsymbol{W}_{\boldsymbol{\cdot},j}\boldsymbol{W}_{\boldsymbol{\cdot},j}^{\top}\big]$, $\mathbf{\Sigma}_{\mathbf{F}} = \mathbb{E}\big[\boldsymbol{f}_{\boldsymbol{\cdot},j}\boldsymbol{f}_{\boldsymbol{\cdot},j}^{\top}\big]$. Here, we use pairwise uncorrelatedness among $\boldsymbol{W}_{\boldsymbol{\cdot},j}$, $\boldsymbol{f}_{\boldsymbol{\cdot},j}$ and $\boldsymbol{\epsilon}_{\boldsymbol{\cdot},j}$. Suppose $\mathbf{\Sigma}_{\mathbf{W}}$ is invertible.

(i) In the linear case, $\boldsymbol{f}_{\boldsymbol{\cdot},j} = \mathbf{0}$. Applying Theorem 5.1 in Anderson and Rubin (1956), we can ascertain the identification of $\mathrm{Col}\big(\mathbf{\Gamma}\mathbf{\Sigma}_{\mathbf{W}}^{\frac{1}{2}}\big)$ and $\mathbf{D}_{\mathbf{E}}$ if there remains two disjoint submatrices of rank K after deleting any row of $\mathbf{\Gamma}\mathbf{\Sigma}_{\mathbf{W}}^{\frac{1}{2}}$. Consequently, we establish an exact identification for $\mathrm{Col}(\mathbf{\Gamma})$ and $\mathbf{D}_{\mathbf{E}}$.

(ii) In the nonlinear case, there is a lack of established methods to exactly recover $\mathbf{\Gamma}$ from Eq.(3.7). But $\mathrm{Col}(\mathbf{\Gamma})$ can be identified asymptotically by PCA method provided that the last two terms are relatively small compared to the first term (Fan et al., 2013). More precisely, let $\Lambda_{\min}(\cdot)$ be the smallest eigenvalue function, $\|\cdot\|_2$ be the spectral norm of matrices. By Weyl's inequality and $\sin\theta$ theorem (Bhatia, 2013), the first K eigenvector matrix $\widetilde{\mathbf{\Gamma}}$ of $\mathrm{Cov}\big(\boldsymbol{Y}_{\boldsymbol{\cdot},j} - (\mathbf{B} + \mathbf{\Gamma\Phi})\boldsymbol{X}_{\boldsymbol{\cdot},j}\big)$ satisfies

$$\big\|\mathcal{P}_{\widetilde{\mathbf{\Gamma}}} - \mathcal{P}_{\mathbf{\Gamma}}\big\|_2 \leq \frac{\pi}{2} \cdot \frac{\|\mathbf{\Sigma}_{\mathbf{F}}\|_2 + \|\mathbf{D}_{\mathbf{E}}\|_2}{\Lambda_{\min}(\mathbf{\Gamma}\mathbf{\Sigma}_{\mathbf{W}}\mathbf{\Gamma}^{\top}) - \|\mathbf{\Sigma}_{\mathbf{F}}\|_2 - \|\mathbf{D}_{\mathbf{E}}\|_2}.$$

However, simply removing the largest K components from the spectral decomposition of $\mathrm{Cov}\big(\boldsymbol{Y}_{\boldsymbol{\cdot},j} - (\mathbf{B} + \mathbf{\Gamma\Phi})\boldsymbol{X}_{\boldsymbol{\cdot},j}\big)$ cannot ensure the convergence of diagonal elements to those of $\mathbf{D}_{\mathbf{E}}$. We will give a consistent estimator for $\mathbf{D}_{\mathbf{E}}$ in Chapter 4.

3.1.2. Identifiability for $\mathbf{B}_1$ with known $\mathrm{Col}(\mathbf{\Gamma})$

Eq.(3.1) can be viewed as multiple semi-parametric models. From Robinson (1988), a sufficient and necessary condition for identifying $\mathbf{B}_1$ when $\boldsymbol{\alpha}_{\bullet,j}$ is observed is

$$\mathbb{E}\Big[\big(\boldsymbol{X}_{\bullet,j} - \mathbb{E}[\boldsymbol{X}_{\bullet,j}|\boldsymbol{\alpha}_{\bullet,j}]\big)\big(\boldsymbol{X}_{\bullet,j} - \mathbb{E}[\boldsymbol{X}_{\bullet,j}|\boldsymbol{\alpha}_{\bullet,j}]\big)^\top\Big] \succ 0.$$

We emphasize that this condition is implied by $\mathbf{\Sigma_W} \succ 0$, which is already used in identifying $\mathrm{Col}(\mathbf{\Gamma})$. To see this, recall that we have assumed $\boldsymbol{g}_{\bullet,j}$ is augmented with $\mathbb{E}[\boldsymbol{X}_{\bullet,j}|\boldsymbol{\alpha}_{\bullet,j}]$ in the nonlinear case. Thus, $\boldsymbol{W}_{\bullet,j}$ contains the remainder of the best linear projection of $\mathbb{E}[\boldsymbol{X}_{\bullet,j}|\boldsymbol{\alpha}_{\bullet,j}]$ onto $\boldsymbol{X}_{\bullet,j}$. Also, $\boldsymbol{X}_{\bullet,j} - \mathbb{E}[\boldsymbol{X}_{\bullet,j}|\boldsymbol{\alpha}_{\bullet,j}]$ is the remainder of the best linear projection of $\boldsymbol{X}_{\bullet,j}$ onto $\mathbb{E}[\boldsymbol{X}_{\bullet,j}|\boldsymbol{\alpha}_{\bullet,j}]$. By Schur's complement argument, the positive definiteness of these two remainder terms is equivalent.

To identify $\mathbf{B}_1$, the Negative Control condition and the Sparsity condition are used:

- (Negative Control): $(\mathbf{B}_1)_{\mathcal{S}} = \mathbf{0}$ and $\mathrm{rank}(\mathbf{\Gamma}_{\mathcal{S}}) = K$ [1] for a known index set $\mathcal{S} \subseteq \{1, \ldots, m\}$ with $|\mathcal{S}| \geq K$. A set $\mathcal{S}$ satisfies this property is called a Negative Control (NC) set.

- (Sparsity): The number of nonzero rows in $\mathbf{B}_1$ is no greater than $\lfloor (m-s)/2 \rfloor$, and $\mathrm{rank}(\mathbf{\Gamma}_{\mathcal{S}}) = K$ for any s-subset $\mathcal{S} \subseteq \{1, \ldots, m\}$, where s is an integer such that $K \leq s \leq m$.

The two conditions were proposed by Wang et al. (2017) where the authors assume a structural equation model between $\mathbf{X}$ and $\mathbf{G}$. Interestingly, these conditions remain effective even in our nonlinear setup, without assuming any functional relationship between $\mathbf{X}$ and $\mathbf{G}$.

Proposition 3.1. *Suppose* $\mathrm{Col}(\mathbf{\Gamma})$ *is known exactly. Then,* $\mathbf{B}_1$ *is identifiable under either the Negative Control condition or the Sparsity condition.*

[1] Unless stated otherwise, a matrix associated with an index set $\mathcal{S}$ refers to the submatrix formed by selecting rows in $\mathcal{S}$. The same convention applies subsequently.

Proof: Note that $\mathbf{B} + \mathbf{\Gamma\Phi}$ is identified, $\mathbf{B}_1 + \mathbf{\Gamma\Phi}A^\top$ is identified as well. Since $\mathrm{Col}(\mathbf{\Gamma})$ is known, its projection matrix $\mathcal{P}_{\mathbf{\Gamma}}$ becomes available. This indicates that $\mathcal{P}_{\mathbf{\Gamma}}^{\perp}\mathbf{B}_1$ can be identified. Suppose $\mathbf{B}_1^*$ is another matrix such that

$$\mathcal{P}_{\mathbf{\Gamma}}^{\perp}\mathbf{B}_1 = \mathcal{P}_{\mathbf{\Gamma}}^{\perp}\mathbf{B}_1^*. \tag{3.8}$$

For the Negative control condition, we choose index set $\mathcal{S}$ as specified in the condition. For the Sparsity condition, we set $\mathcal{S}$ to be the collection of rows where both $\mathbf{B}_1$ and $\mathbf{B}_1^*$ are zeros. In either case, we can ascertain that $|\mathcal{S}| \geq K$ and $\mathrm{rank}(\mathbf{\Gamma}_{\mathcal{S}}) = K$. By rearranging rows, we can assume that

$$\mathbf{\Gamma} = \begin{pmatrix} \mathbf{\Gamma}_{\mathcal{S}} \\ \mathbf{\Gamma}_{\mathcal{S}^c} \end{pmatrix}.$$

Then, Eq.(3.8) implies that

$$\begin{pmatrix} -\mathbf{\Gamma}_{\mathcal{S}}(\mathbf{\Gamma}_{\mathcal{S}}^\top\mathbf{\Gamma}_{\mathcal{S}} + \mathbf{\Gamma}_{\mathcal{S}^c}^\top\mathbf{\Gamma}_{\mathcal{S}^c})^{-1}\mathbf{\Gamma}_{\mathcal{S}^c}^\top \\ \mathbf{I}_{m-|\mathcal{S}|} - \mathbf{\Gamma}_{\mathcal{S}^c}(\mathbf{\Gamma}_{\mathcal{S}}^\top\mathbf{\Gamma}_{\mathcal{S}} + \mathbf{\Gamma}_{\mathcal{S}^c}^\top\mathbf{\Gamma}_{\mathcal{S}^c})^{-1}\mathbf{\Gamma}_{\mathcal{S}^c}^\top \end{pmatrix} (\mathbf{B}_1 - \mathbf{B}_1^*)_{\mathcal{S}^c} = \mathbf{0}.$$

Due to full rankness of $\mathbf{\Gamma}_{\mathcal{S}}$, it shows that $(\mathbf{B}_1 - \mathbf{B}_1^*)_{\mathcal{S}^c} = \mathbf{0}$. Since $(\mathbf{B}_1)_{\mathcal{S}} = (\mathbf{B}_1^*)_{\mathcal{S}} = \mathbf{0}$, we can conclude that $\mathbf{B}_1 = \mathbf{B}_1^*$, thereby establishing the identifiability of $\mathbf{B}_1$. ■

In the proof of Proposition 3.1, the key is to recover $\mathbf{B}_1$ from $\mathcal{P}_{\mathbf{\Gamma}}^{\perp}\mathbf{B}_1$, which can be seen as an instance of *compressed sensing problem*. The identification with the Sparsity condition is equivalent to solve:

$$\min \|\mathbf{\Theta}\|_{1,0} \quad s.t. \quad \mathcal{P}_{\mathbf{\Gamma}}^{\perp}\mathbf{\Theta} = \mathcal{P}_{\mathbf{\Gamma}}^{\perp}\mathbf{B}_1, \tag{3.9}$$

where $\|\mathbf{\Theta}\|_{1,0} = \sum_{i=1}^{m} 1(\Theta_{i,\cdot} \neq \mathbf{0})$ is the number of nonzero rows of $\mathbf{\Theta}$. This problem is known to be NP-hard and cannot be practically solved (Foucart and Rauhut, 2013). Consider a convex relaxation to Eq.(3.9):

$$\min \|\mathbf{\Theta}\|_{1,2} \quad s.t. \quad \mathcal{P}_{\mathbf{\Gamma}}^{\perp}\mathbf{\Theta} = \mathcal{P}_{\mathbf{\Gamma}}^{\perp}\mathbf{B}_1, \tag{3.10}$$

where $\|\mathbf{\Theta}\|_{1,2} = \sum_{i=1}^{m} \|\Theta_{i,\cdot}\|_2$ is the row-wise sum of ℓ_2 norms of $\mathbf{\Theta}$. For any $s = 1, \ldots, m$, define

$$\delta(s, \mathbf{\Gamma}) = \min \left\{ t \in [0,1] : \|\mathcal{P}_{\mathbf{\Gamma}} \boldsymbol{v}\|_2^2 \leq t \|\boldsymbol{v}\|_2^2, \|\boldsymbol{v}\|_0 \leq s \right\}. \tag{3.11}$$

$\delta(s, \mathbf{\Gamma})$ is the *s-sparse* Restricted Isometry (RIP) constant for $\mathcal{P}_{\mathbf{\Gamma}}^{\perp}$.

Proposition 3.2. *Suppose* $\mathrm{Col}(\mathbf{\Gamma})$ *is known exactly and* $\delta\big(\|\mathbf{B}_1\|_{1,0}, \mathbf{\Gamma}\big) < \frac{1}{3}$*, then* $\mathbf{B}_1$ *can be identified by the minimization problem in Eq.*(3.10)*.*

Proposition 3.2 demonstrates the potential for recovering $\mathbf{B}_1$ when the RIP constant is not too large. The condition required for identification using Eq.(3.10) is stronger than that for Eq.(3.9), representing a trade-off made to reduce computational costs.

3.1.3. Identification for $\mathbf{B}_1$ with inexact $\mathrm{Col}(\mathbf{\Gamma})$

In practice, estimating $\mathrm{Col}(\mathbf{\Gamma})$ introduces the potential bias in identifying $\mathbf{B}_1$. In this section, we will explore the robustness of the identification methods outlined in Section 3.1.2 when only an approximate $\mathrm{Col}(\mathbf{\Gamma})$ is available. Because identification through Eq.(3.9) is not practical, we will focus on identification through NC set or solving the minimization problem in Eq.(3.10). Suppose we have identified an approximation $\widetilde{\mathbf{B}}_1$, the robustness is measured in terms of $\|\widetilde{\mathbf{B}}_1 - \mathbf{B}_1\|_{\infty,2}$, where $\|\cdot\|_{\infty,2}$ is the maximum of row-wise ℓ_2 norms. This measurement aligns well with our goal of performing hypothesis testings for rows of $\mathbf{B}_1$. Let $\widetilde{\mathbf{\Gamma}}$ be any estimate of $\mathbf{\Gamma}$ such that $\|\widetilde{\mathbf{\Gamma}} - \mathbf{\Gamma}\mathbf{H}\|_{\infty,2} = o(1)$ for a matrix $\mathbf{H}$.

If an NC set $\mathcal{S}$ is known, then there is another way to identify $\mathbf{B}_1$, different from the proof presented in Proposition 3.1. Let $\mathbf{M} = \mathbf{B}_1 + \mathbf{\Gamma}\mathbf{\Phi}A^{\top}$. Since we know $\mathbf{M}$ is identified and $(\mathbf{B}_1)_{\mathcal{S}} = \mathbf{0}$,

$$\mathbf{M}_{\mathcal{S}} = \mathbf{\Gamma}_{\mathcal{S}}\mathbf{\Phi}A^{\top}.$$

This implies that

$$\mathbf{B}_1 = \mathbf{M} - \mathbf{\Gamma}(\mathbf{\Gamma}_{\mathcal{S}}^{\top}\mathbf{\Gamma}_{\mathcal{S}})^{-1}\mathbf{\Gamma}_{\mathcal{S}}^{\top}\mathbf{M}_{\mathcal{S}}. \tag{3.12}$$

The plug-in estimator is

$$\widetilde{\mathbf{B}}_1 = \mathbf{M} - \widetilde{\mathbf{\Gamma}}(\widetilde{\mathbf{\Gamma}}_{\mathcal{S}}^\top \widetilde{\mathbf{\Gamma}}_{\mathcal{S}})^{-1}\widetilde{\mathbf{\Gamma}}_{\mathcal{S}}^\top \mathbf{M}_{\mathcal{S}}. \tag{3.13}$$

Under mild assumption on $\mathbf{\Phi}$,

$$\|\widetilde{\mathbf{B}}_1 - \mathbf{B}_1\|_{\infty,2} = o(1).$$

Thus, identification through NC sets is robust to inexact $\mathrm{Col}(\mathbf{\Gamma})$.

Without NC set information, our best hope is to recover $\mathbf{B}_1$ from $\mathcal{P}_{\widetilde{\mathbf{\Gamma}}}^\perp \mathbf{B}_1 + \mathcal{P}_{\widetilde{\mathbf{\Gamma}}}^\perp \mathbf{\Gamma}\mathbf{\Phi}A^\top$, which is essentially $\mathcal{P}_{\widetilde{\mathbf{\Gamma}}}^\perp \mathbf{B}_1$ with small perturbations at each entry. To account for these small perturbations, consider a generalized version of Eq.(3.10):

$$\min \|\mathbf{\Theta}\|_{1,2} \quad s.t. \quad \left\|\mathcal{P}_{\widetilde{\mathbf{\Gamma}}}^\perp \mathbf{\Theta} - \mathcal{P}_{\widetilde{\mathbf{\Gamma}}}^\perp \mathbf{B}_1 - \mathcal{P}_{\widetilde{\mathbf{\Gamma}}}^\perp \mathbf{\Gamma}\mathbf{\Phi}A^\top\right\|_{\infty,2} \le \eta \tag{3.14}$$

Suppose $\widetilde{\mathbf{B}}_1$ is the minimizer of Eq.(3.14). It is generally challenging to ensure $\|\widetilde{\mathbf{B}}_1 - \mathbf{B}_1\|_{\infty,2} = o(1)$, unless imposing conditions such as $\mathcal{P}_{\widetilde{\mathbf{\Gamma}}}^\perp$ is close to $\mathbf{I}_m$ entry-wisely (Rosenbaum and Tsybakov, 2010). To resolve this issue, previous works (Lee et al., 2017; Wang et al., 2017; McKennan and Nicolae, 2019; Bing et al., 2023) impose conditions such that $\|\mathcal{P}_{\mathbf{\Gamma}}\mathbf{B}_1\|_{\infty,2} \approx \|\mathcal{P}_{\widetilde{\mathbf{\Gamma}}}\mathbf{B}_1\|_{\infty,2} = o(1)$. This implies that $\|\mathcal{P}_{\widetilde{\mathbf{\Gamma}}}^\perp \mathbf{B}_1 + \mathcal{P}_{\widetilde{\mathbf{\Gamma}}}^\perp \mathbf{\Gamma}\mathbf{\Phi}A^\top - \mathbf{B}_1\|_{\infty,2} = o(1)$, and $\mathbf{B}_1$ is identified asymptotically. Indeed, $\mathbf{B}_1$ cannot be recovered if $\mathbf{B}_1 \in \mathrm{Col}(\mathbf{\Gamma})$. But there are doubts about whether $\|\mathcal{P}_{\mathbf{\Gamma}}\mathbf{B}_1\|_{\infty,2} = o(1)$ is necessary since it rules out the worst-case scenario by only focusing on the best-case scenario and leave no room in between. Therefore, we are motivated to find intermediate results without this assumption. To achieve this, we will need the following lemma.

Lemma 3.3. *Suppose $\mathcal{S}_0 \subseteq \{1, 2, \ldots, m\}$ is any index set with cardinality s, $\delta := \delta\big(2s, \mathbf{\Gamma}\big)$. For any matrix $\mathbf{V} \in \mathbb{R}^{m\times c}$, let $\mathcal{S}_1$ be the s largest rows of $\mathbf{V}_{\mathcal{S}_0^c}$ in terms of ℓ_2 norm. Let $\mathcal{S}_{01} = \mathcal{S}_0 \cup \mathcal{S}_1$. Then,*

$$\|\mathbf{V}_{\mathcal{S}_{01}}\|_F \le \frac{1}{1-\delta}\big\|(\mathcal{P}_{\mathbf{\Gamma}}^\perp \mathbf{V})_{\mathcal{S}_{01}}\big\|_F + \frac{\delta}{(1-\delta)\sqrt{s}}\|\mathbf{V}_{\mathcal{S}_0^c}\|_{1,2},$$

$$\|\mathbf{V}\|_F^2 \leq \|\mathbf{V}_{\mathcal{S}_{01}}\|_F^2 + \frac{1}{s}\|\mathbf{V}_{\mathcal{S}_0^c}\|_{1,2}^2.$$

Let $\mathcal{S}_0$ be the collection of true nonzero rows of $\mathbf{B}_1$, $\mathbf{\Gamma}$ be replaced by $\widetilde{\mathbf{\Gamma}}$, $\mathbf{V} = \widetilde{\mathbf{B}}_1 - \mathbf{B}_1$ in Lemma 3.3. If η in Eq.(3.14) is chosen such that $\|\mathcal{P}_{\widetilde{\mathbf{\Gamma}}}^{\perp}\mathbf{\Gamma}\mathbf{\Phi}A^\top\|_{\infty,2} \leq \eta$, then

$$\left\|\mathcal{P}_{\widetilde{\mathbf{\Gamma}}}^{\perp}\mathbf{V}\right\|_{\infty,2} \leq 2\eta. \tag{3.15}$$

By triangular inequality and Cauchy-Schwartz inequality, $\|\mathbf{B}_1 + \mathbf{V}\|_{1,2} \leq \|\mathbf{B}_1\|_{1,2}$ implies that $\|\mathbf{V}_{\mathcal{S}_0^c}\|_{1,2} \leq \|\mathbf{V}_{\mathcal{S}_0}\|_{1,2} \leq \sqrt{s}\|\mathbf{V}_{\mathcal{S}_0}\|_F$. Apply Lemma 3.3,

$$\|\mathbf{V}_{\mathcal{S}_{01}}\|_F \leq \frac{\sqrt{2s}\cdot 2\eta}{1-\delta} + \frac{\delta}{1-\delta}\|\mathbf{V}_{\mathcal{S}_0}\|_F,$$

which implies that

$$\|\mathbf{V}_{\mathcal{S}_{01}^c}\|_{\infty,2} \leq \frac{1}{\sqrt{s}}\|\mathbf{V}_{\mathcal{S}_{01}}\|_F \leq \frac{2\sqrt{2}}{1-2\delta}\eta. \tag{3.16}$$

Eq.(3.16) implies that if $\delta = \delta(2s, \widetilde{\mathbf{\Gamma}}) < \frac{1}{2}$ and $\eta = o(1)$, then zero rows in $\mathcal{S}_{01}^c$ can be uniformly estimated. If the number of false negatives can be controlled, this indicates that the identification of an approximate NC set remains robust.

To conclude, we demonstrate that the identification for $\mathbf{B}_1$ through an NC set is robust, and the identification of an NC set using Eq.(3.14) is robust. This observation motivates the main idea of our method proposed in Chapter 4: we will solve Eq.(3.14) to estimate an approximate NC set, and then use this NC set to improve our estimation for $\mathbf{B}_1$.

3.2. Semi-strong factors

In a genuinely nonlinear case, where the latent component does not have a finite rank, the approximation remainder term $\mathbf{F}$ diminishes only if $K \to \infty$. Because of this, we allow a possibly diverging number of latent components. This is in sharp contrast to many existing works about factor models or factor augmented regression models Bai (2003); Bai and Wang (2016).

The major challenge in our linear approximation approach to nonlinear factor models arises from the disparity between the commonly used strong factors assumption in factor models literature and potential existence of an infinite number of factors. In terms of our notations, the strong factors assumption requires

$$\Lambda_{\min}(\mathbf{\Gamma}^\top\mathbf{\Gamma}) = \Omega(m) \qquad \text{and} \qquad \Lambda_{\max}(\mathbf{\Gamma}^\top\mathbf{\Gamma}) = O(m), \tag{3.17}$$

which implies that $\|\mathbf{\Gamma}\|_F^2 = \Theta(mK)$, where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. However, $\|\mathbf{\Gamma}\|_F^2 = \operatorname{tr}\big(\mathbf{\Gamma}^\top\mathbf{\Gamma}\big) \leq \sum_{i=1}^m \mathbb{E}[f_i^2(\boldsymbol{\alpha})] \leq m \max_{1\leq i\leq m} \mathbb{E}[f_i^2(\boldsymbol{\alpha})] = O(m)$, if the L^2 norm of f_i is uniformly bounded over i. This observation implies that the number of strong factors cannot grow infinitely in nonlinear case.

Instead, we consider relaxing the strong factors assumption to a semi-strong form. Assuming further that the RKHS norm $\|f_i\|_{\mathcal{H}_R}$ is uniformly bounded, a similar calculation, as shown in Eq.(2.12), implies that

$$\|\boldsymbol{\gamma}_{\cdot,K}\|_2^2 = O(m\lambda_K),$$

and

$$\Lambda_{\min}(\mathbf{\Gamma}^\top\mathbf{\Gamma}) = O(m\lambda_K).$$

Therefore, if $\mathbf{\Gamma}$ is not rank-deficient, a more realistic assumption for the factor loadings generated by eigenfunction expansions is

$$\Lambda_{\min}(\mathbf{\Gamma}^\top\mathbf{\Gamma}) = \Omega(mr_{m,n}) \qquad \text{and} \qquad \Lambda_{\max}(\mathbf{\Gamma}^\top\mathbf{\Gamma}) = O(m), \tag{3.18}$$

where $r_{m,n} = O(\lambda_K)$. As $\lim_{K\to\infty}\lambda_K = 0$, Eq.(3.18) is a form of semi-strong factors assumption. Note that this assumption is consistent with the linear scenario, in which case the number of nonnegative eigenvalues K is finite and λ_K is a positive constant.

We conjecture that $r_{m,n} = \Theta(\lambda_K^a)$ for some $a \geq 1$, which can be verified in the following

case. Suppose entries of $\mathbf{\Gamma}$ are randomly generated such that

$$\gamma_{i,v} = \lambda_v \zeta_{i,v},$$

where $\zeta_{i,v}$ are i.i.d. bounded random variables, $i = 1, \ldots, m$ and $v = 1, \ldots, K$. It can be verified that, if $\sum_{v=1}^{\infty} \lambda_v < \infty$, the corresponding functions f_i satisfy

$$\|f_i\|_{\mathcal{H}_R}^2 = \sum_{v=1}^{\infty} \lambda_v \zeta_{i,v}^2 < \infty, \forall i.$$

In this case, $\mathbf{\Gamma}$ can be seen as the product of a random matrix with i.i.d. entries and a diagonal matrix consisting of the first K eigenvalues of the reproducing kernel. Under mild assumptions from random matrix theory, it can be shown that

$$\Lambda_{\min}(\mathbf{\Gamma}^\top \mathbf{\Gamma}) = \Omega_P(m\lambda_K^2).$$

This verifies our conjecture with $a = 2$.

Chapter 4

Methodology

In this chapter, we describe our full strategy of conducting multiple testings for Eq.(1.2) with FDR control. We begin by introducing a regularization-based method for estimating an NC set. Secondly, we utilize this NC set to estimate the latent factor $\mathbf{G}$ using PCA method. The final estimation of $\mathbf{B}_1$ is obtained after adjusting for $\mathbf{G}$, and construct testing statistics based on this estimation. We conclude with a detailed description of our FDR controlling procedure.

4.1. Estimation of NC set

Addressing the estimation of NC set is challenging due to the inherent difficulty in isolating the effects between observed and unobserved parts. An exception is the IRW-SVA algorithm proposed in Leek and Storey (2008), which has proven to be useful in various applications. However, IRW-SVA lacks theoretical guarantee, and Wang et al. (2017) demonstrated that it could diverge when $\mathbf{X}$ and $\mathbf{G}$ are confounded.

IRW-SVA begins with an initial guess for $\mathbf{G}$ using PCA method. Let $\mathcal{P}_{\mathbf{X}}$ be the projection matrix onto $\mathrm{Col}(\mathbf{X}^\top)$, $\mathcal{P}_{\mathbf{X}}^\perp := \mathbf{I}_n - \mathcal{P}_{\mathbf{X}}$ be the projection onto the orthogonal complement of $\mathrm{Col}(\mathbf{X}^\top)$. Right multiplying $\mathcal{P}_{\mathbf{X}}^\perp$ in Eq.(3.2),

$$\mathbf{Y}\mathcal{P}_{\mathbf{X}}^\perp = \mathbf{\Gamma}\mathbf{G}\mathcal{P}_{\mathbf{X}}^\perp + (\mathbf{F}+\mathbf{E})\mathcal{P}_{\mathbf{X}}^\perp.$$

Let

$$\mathbf{Y}\mathcal{P}_{\mathbf{X}}^{\perp} = \mathbf{U}_{1:K}\mathbf{D}_{1:K}\mathbf{V}_{1:K}^{\top} + \text{remainder} \tag{4.1}$$

be the singular value decomposition (SVD) of $\mathbf{Y}\mathcal{P}_{\mathbf{X}}^{\perp}$, where $\mathbf{U}_{1:K}\mathbf{D}_{1:K}\mathbf{V}_{1:K}^{\top}$ contains the top K components and the diagonal entries in $\mathbf{D}_{1:K}$ is listed in descending order. IRW-SVA uses $\sqrt{n}$-times $\mathbf{V}_{1:K}$ as the initial estimator for $\mathbf{G}$. Intuitively, this initial estimator is convergent to $\mathbf{G}\mathcal{P}_{\mathbf{X}}^{\perp}$. The OLS estimation for $\mathbf{B}_1$ using $\mathbf{G}\mathcal{P}_{\mathbf{X}}^{\perp}$ has bias $\mathbf{\Gamma}\mathbf{G}\mathbf{X}^{\top}(\mathbf{X}\mathbf{X}^{\top})^{-1}A^{\top}$. This bias becomes more significant when the confounding effect between $\mathbf{X}$ and $\mathbf{G}$ is more severe, which partly explains why IRW-SVA fails in this case. Nonetheless, the information of $\mathbf{\Gamma}$ is intact. Let $\hat{\mathbf{\Gamma}} = \frac{\mathbf{U}_{1:K}\mathbf{D}_{1:K}}{\sqrt{n}}$ for further discussions.

We propose to to solve the following minimization problem:

$$\underset{\mathbf{\Theta}}{\text{minimize}} \left\|\mathcal{P}_{\hat{\mathbf{\Gamma}}}^{\perp}(\mathbf{Y} - \mathbf{\Theta}\mathbf{X}_1)\mathcal{P}_{\mathbf{X}_2}^{\perp}\right\|_F^2 + \eta \cdot \|\mathbf{\Theta}\|_{1,2}, \tag{4.2}$$

where $\eta > 0$ is a tunning parameter. Eq.(4.2) connects to Eq.(3.14) in the sense that $\mathcal{P}_{\hat{\mathbf{\Gamma}}}^{\perp}(\mathbf{Y} - \mathbf{\Theta}\mathbf{X}_1)\mathcal{P}_{\mathbf{X}_2}^{\perp}$ is a sample analogue to $\mathcal{P}_{\tilde{\mathbf{\Gamma}}}^{\perp}\mathbf{B}_1 + \mathcal{P}_{\tilde{\mathbf{\Gamma}}}^{\perp}\mathbf{\Gamma}\mathbf{\Phi}A^{\top} - \mathcal{P}_{\tilde{\mathbf{\Gamma}}}^{\perp}\mathbf{\Theta}$. By vectorization, Eq.(4.2) can be viewed as a high-dimensional regression problem with the group lasso penalty. This problem is complicated by the structural rank-deficiency of its design matrix $(\mathcal{P}_{\mathbf{X}_2}^{\perp}\mathbf{X}_1^{\top}) \otimes \mathcal{P}_{\hat{\mathbf{\Gamma}}}^{\perp}$, where $\otimes$ stands for Kronecker product.

Recovering exact sparsistency through lasso or group lasso generally requires stringent assumptions (Zou, 2006; Lounici et al., 2010). Therefore, we only seek for a subset estimation, which is sufficient for estimation of $\mathbf{B}_1$ as observed in Section 3.1. Let $\hat{\mathbf{B}}_1$ be a solution to Eq.(4.2).

$$\hat{\mathcal{S}}(\theta) = \left\{i = 1, \ldots, m : \left\|\hat{\boldsymbol{b}}_{i,\cdot}^{(1)}\right\|_2 \leq \frac{\theta}{\sqrt{n}}\right\}. \tag{4.3}$$

is an estimation for NC set. $\theta > 0$ is a chosen threshold. Here is the main heuristic: If $|\hat{\mathcal{S}}(\theta)|/m \geq p$ and the true positive proportion is s, then at least $\lfloor \tau - s \rfloor m$ zero rows of $\hat{\mathbf{B}}_1(\hat{\eta}(\tau))$ must be true negatives. When s is relatively small compared to τ, most of these zero rows are true negatives, forming an approximate NC set.

The NC set in Eq.(4.3) may not be sufficiently large. To resolve this, a re-weighted version of Eq.(4.2) can be used:

$$\underset{\mathbf{\Theta}}{\text{minimize}} \left\| \mathcal{P}^{\perp}_{\hat{\mathbf{\Gamma}}} (\mathbf{Y} - \mathbf{\Theta}\mathbf{X}_1) \mathcal{P}^{\perp}_{\mathbf{X}_2} \right\|_F^2 + \eta \cdot \|\mathbf{I}_{\hat{\mathcal{S}}} \mathbf{\Theta}\|_{1,2}, \tag{4.4}$$

where $\hat{\mathcal{S}}$ is any estimated NC set, $\mathbf{I}_{\hat{\mathcal{S}}}$ is $m \times m$ diagonal matrix with i-th diagonal being 1 if $i \in \hat{\mathcal{S}}$ and 0 otherwise. Then, Eq.(4.3) is applied again to enlarge the NC set. This scheme proposed can be repeated until new NC set is not significantly larger than the old one.

4.2. Estimation of $\mathbf{G}$ and $\mathbf{B}_1$

Let $\mathcal{S}$ be our final choice of NC set. With the help of $\mathcal{S}$, the estimations for $\mathbf{G}$ and $\mathbf{B}_1$ are straightforward. Perform SVD decomposition for $\mathbf{Y}_{\mathcal{S}} \mathcal{P}^{\perp}_{\mathbf{X}_2}$:

$$\mathbf{Y}_{\mathcal{S}} \mathcal{P}^{\perp}_{\mathbf{X}_2} = \mathbf{U}^{\mathcal{S}}_{1:K} \mathbf{D}^{\mathcal{S}}_{1:K} \big(\mathbf{V}^{\mathcal{S}}_{1:K}\big)^{\top} + \text{remainder}, \tag{4.5}$$

where $\mathbf{U}^{\mathcal{S}}_{1:K} \mathbf{D}^{\mathcal{S}}_{1:K} \big(\mathbf{V}^{\mathcal{S}}_{1:K}\big)^{\top}$ contains the top K components, $\mathbf{D}^{\mathcal{S}}_{1:K}$ is diagonal matrix with entries listed in the descending order. Similar to Bai and Ng (2002); Bai (2003), our estimation of $\mathbf{G}$ is

$$\hat{\mathbf{G}} = \sqrt{n} \big(\mathbf{V}^{\mathcal{S}}_{1:K}\big)^{\top}, \tag{4.6}$$

where $\mathbf{V}_{1:K}$ is the first K columns of $\mathbf{V}$. Our final estimation of $\mathbf{B}_1$ is obtained by ordinary least squares for each row-wise regression:

$$\hat{\mathbf{B}}_{1,\text{final}} = \underset{\mathbf{\Theta}}{\text{argmin}} \left\| (\mathbf{Y} - \mathbf{\Theta}\mathbf{X}_1) \mathcal{P}^{\perp}_{\mathbf{X}_2, \hat{\mathbf{G}}} \right\|_F^2 = \mathbf{Y} \mathcal{P}^{\perp}_{\mathbf{X}_2, \hat{\mathbf{G}}} \mathbf{X}_1^{\top} \big(\mathbf{X}_1 \mathcal{P}^{\perp}_{\mathbf{X}_2, \hat{\mathbf{G}}} \mathbf{X}_1^{\top}\big)^{-1}, \tag{4.7}$$

where $\mathcal{P}^{\perp}_{\mathbf{X}_2, \hat{\mathbf{G}}}$ is the orthogonal projection matrix onto the complement of both $\text{Col}(\mathbf{X}_2)$ and $\text{Col}(\hat{\mathbf{G}})$. The idiosyncratic variances can be estimated through:

$$\hat{\sigma}_i^2 = \frac{1}{n - p - K} \boldsymbol{Y}_{i,\cdot}^{\top} \mathcal{P}^{\perp}_{\mathbf{X}, \hat{\mathbf{G}}} \boldsymbol{Y}_{i,\cdot}, \quad i = 1, \ldots, m. \tag{4.8}$$

4.3. FDP estimation in multiple testings

Traditional multiple testing procedures are based on a sequence of p-values for a family of hypotheses, and each hypothesis is rejected if its p-value is below or equal to a pre-determined threshold t. Define the following empirical processes for any $t \in [0, 1]$:

$$V(t) = \big|\{i \in \mathcal{I}_0 : P_i \leq t\}\big|,$$

$$S(t) = \big|\{i \in \mathcal{I}_1 : P_i \leq t\}\big|,$$

$$R(t) = \big|\{i = 1, \ldots, m : P_i \leq t\}\big|. \tag{4.9}$$

Then $V(t)$, $S(t)$, $R(t)$ are the number of falsely rejected hypotheses, the number of correctly rejected hypotheses, and the total number of rejected hypotheses. The false discovery proportion w.r.t. the threshold t is defined as

$$\mathrm{FDP}(t) = \frac{V(t)}{R(t) \vee 1} \text{ with } R(t) \vee 1 = \max\{R(t), 1\}. \tag{4.10}$$

Based on different p-values, different FDPs can be defined. The one we are specifically interested in is the FDP derived from the oracle test statistics, which utilize the information from the unobserved component, for simultaneously testing hypotheses in Eq.(1.2). We consider two different sets of oracle test statistics, based on different oracle estimators for $\mathbf{B}_1$. The first set of Wald's type oracle test statistics is based on

$$\hat{\mathbf{B}}_{1,\mathrm{ora}} = \mathbf{Y}\mathcal{P}^{\perp}_{\mathbf{X}_2,\mathbf{G}}\mathbf{X}_1^{\top}\big(\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\mathbf{G}}\mathbf{X}_1^{\top}\big)^{-1}, \tag{4.11}$$

the OLS estimator of $\mathbf{Y}$ regressing onto $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{G}$. The second set of is based on

$$\hat{\mathbf{B}}_{1,\mathrm{ora}} = \big(\mathbf{Y} - \boldsymbol{f}(\mathbf{A})\big)\mathcal{P}^{\perp}_{\mathbf{X}_2}\mathbf{X}_1^{\top}\big(\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2}\mathbf{X}_1^{\top}\big)^{-1}, \tag{4.12}$$

the OLS estimator of $\mathbf{Y} - \boldsymbol{f}(\mathbf{A})$ regressing onto $\mathbf{X}_1$, $\mathbf{X}_2$. The key difference between the two

estimators lies in their efficiency: the second type has smaller asymptotic variance, and thus more efficient.

For $\hat{\mathbf{B}}_{1,\text{ora}}$ in Eq.(4.11), the oracle test statistics are

$$T_{i,\text{ora}} = \frac{(\hat{\boldsymbol{b}}_{i,\text{ora}}^{(1)})^\top \mathbf{X}_1 \mathcal{P}^{\perp}_{\mathbf{X}_2,\mathbf{G}} \mathbf{X}_1^\top \hat{\boldsymbol{b}}_{i,\text{ora}}^{(1)}}{\sigma_i^2}, \tag{4.13}$$

For $\hat{\mathbf{B}}_{1,\text{ora}}$ in Eq.(4.12), the oracle test statistics are

$$T_{i,\text{ora}} = \frac{(\hat{\boldsymbol{b}}_{i,\text{ora}}^{(1)})^\top \mathbf{X}_1 \mathcal{P}^{\perp}_{\mathbf{X}_2} \mathbf{X}_1^\top \hat{\boldsymbol{b}}_{i,\text{ora}}^{(1)}}{\sigma_i^2}, \tag{4.14}$$

In either form, $\hat{\boldsymbol{b}}_{i,\text{ora}}^{(1)}$ is the i-th row vector of $\hat{\mathbf{B}}_{1,\text{ora}}$. The oracle p-values are

$$P_{i,\text{ora}} = 1 - \chi_c^2(T_{i,\text{ora}}, 0), \tag{4.15}$$

$\chi_c^2(t, 0)$ is the c.d.f. for chi-squared distribution with degree of freedom c and non-centrality 0. The main difference between these two oracle test statistics lies in their efficiency. Since $\mathbf{X}_1 \mathcal{P}^{\perp}_{\mathbf{X}_2} \mathbf{X}_1^\top \succeq \mathbf{X}_1 \mathcal{P}^{\perp}_{\mathbf{X}_2,\mathbf{G}} \mathbf{X}_1^\top$, the testing procedure based on Eq.(4.14) is more powerful than Eq.(4.13).

Let $\hat{\boldsymbol{b}}_{i,\text{final}}^{(1)}$ be our final estimate of $\boldsymbol{b}_{i,\bullet}^{(1)}$, $\hat{\sigma}_i$ be the estimated standard error for the i-th regression model, $\forall i = 1, \ldots, m$ as shown in Section 4.2. The Wald's type test statistics and the associated p-values are:

$$\hat{T}_i = \frac{(\hat{\boldsymbol{b}}_{i,\text{final}}^{(1)})^\top \mathbf{X}_1 \mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}} \mathbf{X}_1^\top \hat{\boldsymbol{b}}_{i,\text{final}}^{(1)}}{\hat{\sigma}_i^2},$$

$$\hat{P}_i = 1 - \chi_c^2(\hat{T}_i, 0). \tag{4.16}$$

Note that our $\hat{T}_i$ is more similar to the statistics in Eq.(4.13) than that in Eq.(4.14), so we can expect our tests to behave more like the first set of oracle Wald's tests.

As in Eq.(4.9), $\{V_{\text{ora}}(t), \hat{V}(t)\}$, $\{S_{\text{ora}}(t), \hat{S}(t)\}$, $\{R_{\text{ora}}(t), \hat{R}(t)\}$ can be defined accordingly.

Similarly, define the oracle false discovery proportion and the estimation of it:

$$\mathrm{FDP}_{\mathrm{ora}}(t) = \frac{V_{\mathrm{ora}}(t)}{R_{\mathrm{ora}}(t) \vee 1}, \ \widehat{\mathrm{FDP}}(t) = \frac{\hat{V}(t)}{\hat{R}(t) \vee 1}. \tag{4.17}$$

Note however, $\widehat{\mathrm{FDP}}(t)$ is still not observable due to unknown $\mathcal{I}_0$, the subset of true null hypotheses for Eq.(1.2).

Since $\hat{P}_i$, $i = 1, \ldots, m$ have accounted for the covariance dependence caused by the $\mathbf{G}$, they can be viewed as p-values with weak dependence. Therefore, the following classical estimation of $\widehat{\mathrm{FDP}}(t)$ can be applied (Storey et al., 2004):

$$\widehat{\mathrm{FDP}}_{\lambda}(t) = \frac{m\hat{\pi}_0(\lambda)t}{\hat{R}(t) \vee 1}, \tag{4.18}$$

where $\hat{\pi}_0(\lambda) = \sum_{i=1}^{m} 1(\hat{P}_i > \lambda)/m(1-\lambda)$ is an estimation of the proportion of null hypothesis $|\mathcal{I}_0|/m$. $\lambda \in (0, 1)$ is generally chosen to be very close to 1.

Based on Eq.(4.18), Storey's procedure can be applied to control FDR. Suppose the target tolerance of FDR is $\alpha \in (0, 1)$. Then, threshold t is chosen as:

$$\hat{t} = \sup\left\{t \in [0, 1] : \widehat{\mathrm{FDP}}_{\lambda}(t) \leq \alpha\right\}. \tag{4.19}$$

For each $i = 1, \ldots, m$, hypothesis H_{0i} in Eq.(1.2) is rejected if and only if the corresponding p-value $\hat{P}_i \leq \hat{t}$.

Chapter 5

Theoretical results

In this chapter, we provide theoretical guarantees for the procedure introduced in Chapter 4. Section 5.1 contains our main assumptions. Section 5.2 begins with asymptotic results for estimation of $\mathbf{\Gamma}$. Then, we establish oracle inequalities for the minimization problems in Eq.(4.2). These inequalities help in controlling the number of false negatives in our NC set estimation. Section 5.3 presents asymptotic results for the estimations of $\mathbf{B}_1$ and $\mathbf{D_E}$. Notably, we demonstrate that our estimation for $\mathbf{B}_1$ is as efficient as if we had observed the latent factor $\mathbf{G}$. Lastly, Section 5.4 establishes consistency of our FDP estimator.

5.1. Assumptions

We consider the setting where m, n can both diverge to infinity. The number of latent basis K is assumed known and considered as a function of m, n. K could diverge to infinity in the nonlinear scenario. All asymptotic limits in this paper are taken simultaneously for both m and n (Bai, 2003). We assume that $n = o(m)$, $\log m = o(n)$, $p = o(n^{1/3})$, the dimension of primary variable $\mathbf{X}_1$, i.e. c is fixed. The underlying dimension k for nonlinear factor $\boldsymbol{\alpha}_{\cdot,j}$ is fixed. We begin with basic distributional assumptions.

Assumption 5.1. *$\boldsymbol{Y}_{\cdot,j}, \boldsymbol{X}_{\cdot,j}, \boldsymbol{\alpha}_{\cdot,j}, \boldsymbol{\epsilon}_{\cdot,j}$ are* i.i.d. *random vectors. $\boldsymbol{\epsilon}_{\cdot,j} \sim N(\mathbf{0}, \mathbf{D_E})$ is an independent error, where $\mathbf{D_E} = diag(\sigma_1^2, \ldots, \sigma_m^2)$. Latent basis $\boldsymbol{g}_{\cdot,j}$ and approximation remainder $\boldsymbol{f}_{\cdot,j}$ are*

considered as Borel functions of $\boldsymbol{\alpha}_{\boldsymbol{\cdot},j}$*. Entries of* $\mathbf{X}_1$ *and* $\mathbf{X}_2$ *are sub-gaussian random variables and their sub-gaussian norms are uniformly bounded. Lastly,* $\mathbb{E}\big[\boldsymbol{g}_{\boldsymbol{\cdot},j}\boldsymbol{g}_{\boldsymbol{\cdot},j}^\top\big] = \mathbf{I}_K$, $\mathrm{Cov}(\boldsymbol{X}_{\boldsymbol{\cdot},j}, \boldsymbol{f}_{\boldsymbol{\cdot},j}) = \mathbf{0}$, $\mathrm{Cov}(\boldsymbol{g}_{\boldsymbol{\cdot},j}, \boldsymbol{f}_{\boldsymbol{\cdot},j}) = \mathbf{0}$.

The following regularity conditions on $\boldsymbol{\Sigma}_{\mathbf{X},\mathbf{G}}$, the covariance matrix of $(\boldsymbol{X}_{\boldsymbol{\cdot},j}^\top, \boldsymbol{g}_{\boldsymbol{\cdot},j}^\top)$, and $\boldsymbol{\Gamma}$, $\boldsymbol{\Sigma}_{\mathbf{F}}$, $\mathbf{D}_{\mathbf{E}}$ are required for our analysis.

Assumption 5.2. *Assume the following conditions hold:*

(*a*) $0 < \inf_{m,n} \Lambda_{\min}(\boldsymbol{\Sigma}_{\mathbf{X},\mathbf{G}}) \le \limsup_{m,n} \Lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{X},\mathbf{G}}) < \infty$.

(*b*) $0 < \liminf_{m,n} \Lambda_{\min}\Big(\frac{1}{mr_{m,n}}\boldsymbol{\Gamma}^\top\boldsymbol{\Gamma}\Big) \le \limsup_{m,n} \Lambda_{\max}\Big(\frac{1}{m}\boldsymbol{\Gamma}^\top\boldsymbol{\Gamma}\Big) < \infty$.

(*c*) $\|\boldsymbol{\Sigma}_{\mathbf{F}}\|_2 = O\Big(\frac{m}{n}\Big)$.

(*d*) $0 < \inf_{m,n} \min_{1\le i\le m} \sigma_i^2 \le \sup_{m,n} \max_{1\le i\le m} \sigma_i^2 < \infty$.

Assumption 5.2 contains common conditions in factor models literature (Bai and Ng, 2002; Bai, 2003). Similar conditions were also used in the recent work by Bing et al. (2023). Our assumption is slightly different in (b) and (c), where we have accounted for the loss of factor strength and the remainder term due to the linear approximation to the nonlinear component.

Assumption 5.3. *The support for* $\boldsymbol{\alpha}_{\boldsymbol{\cdot},j}$ *is compact. Nonlinear functions* $f_1, \ldots, f_m$ *are in a RKHS generated by kernel function* $R(\boldsymbol{s}, \boldsymbol{t})$*. Assume that* $n\lambda_K^2 \to \infty$, $K \vee p = o(n\lambda_K^2)$*, and* $Ma_{K+1}\log m = O(1)$*, where* λ_K *is the* K*-th eigenvalue of* R*,* a_{K+1} *is defined in Eq.*(2.13)*,* $M = \sup_{1\le i\le m} \|f_i\|_{\mathcal{H}_R}^2$.

Assumption 5.3 has immediate consequences that are worth mentioned. Firstly, the uniform L^2 approximation error $M\lambda_{K+1}$ is controlled because $\lambda_{K+1} \le a_{K+1}$. Secondly, it implies that $\boldsymbol{g}_{\boldsymbol{\cdot},j}$ is bounded with ℓ_2 norm of order $O\Big(\frac{1}{\sqrt{\lambda_K}}\Big)$:

$$\sum_{v=1}^K g_{v,j}^2 = \sum_{v=1}^K \Phi_v(\boldsymbol{\alpha}_{\boldsymbol{\cdot},j})^2 \le \frac{1}{\lambda_K}\sum_{v=1}^K \lambda_v \Phi_v^2(\boldsymbol{\alpha}_{\boldsymbol{\cdot},j}) \le \frac{1}{\lambda_K} R(\boldsymbol{\alpha}_{\boldsymbol{\cdot},j}, \boldsymbol{\alpha}_{\boldsymbol{\cdot},j}) = O\Big(\frac{1}{\lambda_K}\Big).$$

Lastly, row vectors in the factor loading matrix $\mathbf{\Gamma}$ are uniformly bounded in the sense that $\sup_{1\leq i\leq m}\|\boldsymbol{\gamma}_{i,\boldsymbol{\cdot}}\|_2^2 \leq M\lambda_1$, see Eq.(2.12).

Assumption 5.4. (*Linear sparsity*) *Let $\mathcal{I}_0, \mathcal{I}_1 \subset \{1,\ldots,m\}$ be the index set* [1] *where null hypotheses and alternative hypotheses hold respectively,*

$$\lim_{m\to\infty}\frac{|\mathcal{I}_0|}{m} = \pi_0 \in (0,1),\ |\mathcal{I}_0| + |\mathcal{I}_1| = m.$$

Assumption 5.5. *Suppose $\mathcal{I}_1$ in assumption 5.4 is further decomposed as $\mathcal{I}_1 = \mathcal{I}_{11}\cup\mathcal{I}_{12}\cup\mathcal{I}_{13}$. $|\mathcal{I}_{1j}|\to\infty$ for each $j = 1,2,3$. The asymptotic proportions are $\pi_{11},\pi_{12},\pi_{13}$, where π_{11},π_{12} could be zeroes, $\pi_{13}>0$, $\pi_{11}+\pi_{12}+\pi_{13} = 1-\pi_0$. Also, the following conditions hold:*

(*a*) $\limsup_{m,n}\sqrt{n}\max_{i\in\mathcal{I}_{11}}\|\boldsymbol{b}_{i,\boldsymbol{\cdot}}^{(1)}\|_2 \to 0$.

(*b*) *There exists a vector sequence $\{\boldsymbol{c}_i\}_{i=1}^{\infty}\subset\mathbb{R}^c$ such that*

$$\lim_{m,n}\max_{i\in\mathcal{I}_{12}}\|\sqrt{n}\boldsymbol{b}_{i,\boldsymbol{\cdot}}^{(1)} - \boldsymbol{c}_i\|_2 = 0,\quad \sup_{1\leq i<\infty}\|\boldsymbol{c}_i\|_2 < \infty.$$

(*c*) $\liminf_{m,n}\sqrt{\frac{n}{\log m}}\min_{i\in\mathcal{I}_{13}}\|\boldsymbol{b}_{i,\boldsymbol{\cdot}}^{(1)}\|_2\to\infty$.

Assumption 5.4 is assumed to align with traditional multiple testing settings (Storey, 2002). Assumption 5.5 divides alternative hypotheses into weak, medium and strong parts based on their signal strengths. When $\pi_{11}| = |\pi_{12}| = 0$, assumption 5.5 becomes the "Beta-min" condition, which has been widely used in high-dimensional regression literature to differentiate nonzero coefficients to zeroes (van de Geer et al., 2011).

Assumption 5.6. *The following generative model holds for $\mathbf{X}_1$:*

$$\mathbf{X}_1 = \mathbf{B}_{2\to1}\mathbf{X}_2 + \mathbf{\Gamma}_{\mathbf{X}}\mathbf{G} + \mathbf{E}_{\mathbf{X}},$$

$$\|\mathbf{\Gamma}_{\mathbf{X}}\|_F = O(1).$$

[1] $\mathcal{I}_0$ and $\mathcal{I}_1$ are non-decreasing set sequences indexed by m. The subscript m is omitted for simplification.

Conditional on $\boldsymbol{X}^{(2)}_{\cdot,j}$ *and* $\boldsymbol{\alpha}_{\cdot,j}$, $(\mathbf{E}_{\mathbf{X}})_{i,j}$ *is a zero-mean sub-gaussian random variable. The sub-gaussian norms across different* i *are uniformly bounded away from zero and upper bounded by a bounded function of* $\boldsymbol{X}^{(2)}_{\cdot,j}$ *and* $\boldsymbol{\alpha}_{\cdot,j}$.

Assumption 5.6 is only used in nonlinear case for showing $\sqrt{n}$-consistency of our final estimator $\hat{\mathbf{B}}_{1,\text{final}}$. If the approximation remainder term $\mathbf{F} = \mathbf{0}$, this condition is not needed. Essentially, it assumes that $\mathbb{E}\big[\boldsymbol{X}^{(1)}_{\cdot,j}|\boldsymbol{X}^{(2)}_{\cdot,j}, \boldsymbol{\alpha}_{\cdot,j}\big]$ is in a partial linear form, where the parametric part is linear in $\mathbf{X}_2$ and the nonparametric part is a nonlinear function of $\boldsymbol{\alpha}_{\cdot,j}$. Because $\mathbf{G}$ contains $\mathbb{E}\big[\boldsymbol{X}^{(1)}_{\cdot,j}|\boldsymbol{\alpha}_{\cdot,j}\big]$ in nonlinear case, it can be shown that the nonparametric part must be represented by $\mathbf{G}$. Therefore, it is not restricted to assume that $\mathbf{X}_1$ is linear in $\mathbf{X}_2$ and $\mathbf{G}$.

5.2. Asymptotic for $\hat{\mathbf{\Gamma}}$ and oracle inequalities for $\hat{\mathbf{B}}_1$

In this section, we will establish oracle inequalities for the minimizers of Eq.(4.2), which hold with probability tending to 1. Leveraging the "Beta-min" condition, these inequalities enable control over the number of false negatives selected by Eq.(4.2), thus offering a guarantee for our NC set estimation. The following proposition provides asymptotic results for the estimation of $\mathbf{\Gamma}$.

Proposition 5.7. *Let* $\hat{\mathbf{\Gamma}} = (\hat{\boldsymbol{\gamma}}_{1,\cdot}, \ldots, \hat{\boldsymbol{\gamma}}_{m,\cdot})^\top = \frac{\mathbf{U}_{1:K}\mathbf{D}_{1:K}}{\sqrt{n}}$ *as in Eq.*(4.1). *Under assumptions 5.1-5.3, if* $nr^2_{m,n} \to \infty$, *then the following conclusions hold:*

(1) $\big\|\mathcal{P}_{\hat{\mathbf{\Gamma}}} - \mathcal{P}_{\mathbf{\Gamma}}\big\|_2 = O_P\Big(\frac{1}{\sqrt{nr^2_{m,n}}}\Big)$.

(2) *There exists a* $K \times K$ *matrix* $\mathbf{H}$ *such that* $\mathbf{H}$ *is invertible with probability tending to* 1, *and*

$$\|\mathbf{H}\|_2 = O_P(1),\ \|\mathbf{H}^{-1}\|_2 = O_P(1),\ \big\|\hat{\mathbf{\Gamma}} - \mathbf{\Gamma}\mathbf{H}\big\|_2 = O_P\Big(\sqrt{\frac{m}{n}}\Big).$$

(3) $\max_{1\le i\le m}\big\|\hat{\boldsymbol{\gamma}}_{i,\cdot} - \mathbf{H}\boldsymbol{\gamma}_{i,\cdot}\big\| = O_P\Big(\sqrt{\frac{K\vee\log m+p}{n}} + \sqrt{\frac{Ma_{K+1}}{n\lambda_K}} + \frac{1}{\sqrt{nr_{m,n}}}\Big)$.

Similar results to Proposition 5.7 can be found throughout factor models literature. Our contributions lie in several aspects: (1) allowing the number of latent factors K to

diverge; (2) accommodating the difference of strength between the smallest and the largest eigenvalues in factor loadings (assumption 5.2(b)); (3) taking into account the remainder term $\mathbf{F}$ resulting from linear approximation. If K is finite and $\Lambda_{\min}\big(\mathbf{\Gamma}^\top\mathbf{\Gamma}\big)$ and $\Lambda_{\max}\big(\mathbf{\Gamma}^\top\mathbf{\Gamma}\big)$ are of the same order, we can show that $\sqrt{n}\big(\hat{\gamma}_{i,\cdot} - \mathbf{H}\boldsymbol{\gamma}_{i,\cdot}\big) = O_P(1)$ with a slight modification to our proof, aligning with the findings in Bai (2003); McKennan and Nicolae (2019).

Theorem 5.8. *Suppose assumptions 5.1-5.3 hold, and the $|\mathcal{I}_1|$-sparse RIP constant for $\hat{\mathbf{\Gamma}}$ satisfies that $\delta := \delta\big(2|\mathcal{I}_1|, \hat{\mathbf{\Gamma}}\big) < \frac{1}{4}$. For any $0 < \epsilon < 1$, the following holds simultaneously with probability greater than $1 - \epsilon$:*

(1) *The tunning parameter can be chosen such that*

$$\eta \geq C_\epsilon \sqrt{n}\Big(\sqrt{K \vee \log m + p} + \sqrt{\frac{nMa_{K+1}}{\lambda_K}} + \frac{1}{\sqrt{r_{m,n}}}\Big)\sqrt{\frac{M}{r_{m,n}}},$$

where $C_\epsilon > 0$ is a constant.

(2) *For the minimizer $\hat{\mathbf{B}}_1 = (\hat{\boldsymbol{b}}_{1,\cdot}^{(1)}, \ldots, \hat{\boldsymbol{b}}_{m,\cdot}^{(1)})^\top$ in Eq.(4.2):*

$$\|\hat{\mathbf{B}}_1 - \mathbf{B}_1\|_F \leq \frac{\sqrt{20 \cdot |\mathcal{I}_1|}}{1 - 4\delta} \frac{\eta}{\Lambda_{\min}(\mathbf{X}_1 \mathcal{P}_{\mathbf{X}_2}^\perp \mathbf{X}_1^\top)}$$

and

$$\max_{1 \leq i \leq m} \|\hat{\boldsymbol{b}}_{i,\cdot}^{(1)} - \boldsymbol{b}_{i,\cdot}^{(1)}\|_2 \leq \frac{d_{m,n}\eta}{\big(1 - \delta(1, \hat{\mathbf{\Gamma}})\big)\Lambda_{\min}(\mathbf{X}_1 \mathcal{P}_{\mathbf{X}_2}^\perp \mathbf{X}_1^\top)},$$

where

$$d_{m,n} = \frac{3}{2} + \frac{\sqrt{20 \cdot |\mathcal{I}_1|}}{1 - 4\delta} \frac{\Lambda_{\max}(\mathbf{X}_1 \mathcal{P}_{\mathbf{X}_2}^\perp \mathbf{X}_1^\top)}{\Lambda_{\min}(\mathbf{X}_1 \mathcal{P}_{\mathbf{X}_2}^\perp \mathbf{X}_1^\top)} \eta \max_{1 \leq i \leq m} \sqrt{(\mathcal{P}_{\hat{\mathbf{\Gamma}}})_{ii}}.$$

$(\mathcal{P}_{\hat{\mathbf{\Gamma}}})_{ii}$ is the i-th element in the main diagonal of $\mathcal{P}_{\hat{\mathbf{\Gamma}}}$.

(3) *Let $b_{\min} = \min_{i \in \mathcal{I}_1} \|\boldsymbol{b}_{i,\cdot}^{(1)}\|_2$. For the NC set estimation in Eq.(4.3) with $\theta < \frac{\sqrt{n} b_{\min}}{2}$,*

$$|\hat{\mathcal{S}}(\theta) \cap \mathcal{I}_1| \leq \frac{4\|\hat{\mathbf{B}}_1 - \mathbf{B}_1\|_F^2}{b_{\min}^2}.$$

Theorem 5.8 derives error bounds for the minimizer $\hat{\mathbf{B}}_1$ in Eq.(4.2). It shows that the average ℓ_2 loss as well as maximum ℓ_2 loss are approximately of order $\sqrt{n}$ up to some factors. If $Ma_{K+1}/\lambda_K \leq \log m$ and $nr_{m,n}^2 \to \infty$, Theorem 5.8(2) shows that $\sqrt{\frac{n}{m}}\|\hat{\mathbf{B}}_1 - \mathbf{B}_1\|_F = O_P\Big(\sqrt{r_{m,n}M + K\log m} \cdot \sqrt{\frac{|\mathcal{I}_1|}{m}}\Big)$ and $\sqrt{n}\|\hat{\mathbf{B}}_1 - \mathbf{B}_1\|_{1,2} = O_P\Big((r_{m,n}M + K\log m) \cdot \sqrt{\frac{M|\mathcal{I}_1|}{m}}\Big)$. The bound can be improved especially when $|\mathcal{I}_1|$ is small, due to either small $|\mathcal{I}_1|$ or increased $\delta(2|\mathcal{I}_1|, \hat{\mathbf{\Gamma}})$. Our results do not contain $\sqrt{n}$-consistency for $\hat{\boldsymbol{b}}_{i,\cdot}^{(1)}$. This is due to the theoretical difficulty when we do not assume $\|\mathcal{P}_{\hat{\mathbf{\Gamma}}}^{\perp}\mathbf{B}_1\|_{1,2} = o_P(1)$, as mentioned in Section 3.1.

From the error bounds, it is natural to select an NC set by thresholding. This approach is essentially the thresholded lasso, which is known to share similarities to the adaptive lasso (van de Geer et al., 2011). Assuming the "Beta-min" condition in assumption 5.5, there exists a possibility to select all true negatives without any false negatives with probability tending to 1, provided that $(r_{m,n}M + K\log m) \cdot \sqrt{\frac{M|\mathcal{I}_1|}{m}} = o(\theta)$. Therefore, this thresholded approach is ideal for NC set estimation. As a rule of thumb, we recommend selecting a threshold such that around half of rows are included in an NC set. This recommendation is grounded on the belief that $|\mathcal{I}_1|$ is less than $m/2$, which coincides with the identification condition in the Sparsity scenario, see Section 3.1.

In practice, this approach could be compromised due to lack of "Beta-min" condition. However, the error bounds do exclude the inclusion of those strong $\boldsymbol{b}_{i,\cdot}^{(1)}$ with large ℓ_2 norms. In Section 5.3, we will illustrate that the inclusion of some false negatives with small ℓ_2 norms has a mild effect on the estimation of the latent basis $\mathbf{G}$. Therefore, the estimated NC set is robust to the inclusion of weak $\boldsymbol{b}_{i,\cdot}^{(1)}$.

5.3. Asymptotic for $\hat{\mathrm{G}}$ and $\hat{\mathbf{B}}_{1,\mathrm{final}}$

Based on our NC set estimation, the estimation of the latent basis $\mathbf{G}$ and the final estimation $\hat{\mathbf{B}}_{1,\mathrm{final}}$ are obtained. Unlike the minimizer $\hat{\mathbf{B}}_1$ in Eq.(4.2), the final estimation for $\mathbf{B}_1$ can be proven to be $\sqrt{n}$-consistent, laying the groundwork for subsequent hypothesis testing and inference procedures.

Proposition 5.9. *Let $\mathcal{S}$ be a chosen NC set such that $s = |\mathcal{S}| \to \infty$, and $\Lambda_{\min}\Big(\frac{1}{s}\mathbf{\Gamma}_{\mathcal{S}}^{\top}\mathbf{\Gamma}_{\mathcal{S}}\Big) = \Omega(r_{m,n})$, $\Lambda_{\max}\Big(\frac{1}{s}\mathbf{\Gamma}_{\mathcal{S}}^{\top}\mathbf{\Gamma}_{\mathcal{S}}\Big) = O(1)$. Let $\mathcal{F} \subset \mathcal{S}$ be the collection of false negatives. Under assumptions 5.1-5.3, if $\frac{s}{m} = \Omega_P(1)$ and $\|\mathbf{B}_1\|_{\infty,2} = O(1)$, then the following conclusions hold for $\hat{\mathbf{G}}$ in Eq.*(4.6)*:*

$$\big\|\mathcal{P}_{\hat{\mathbf{G}}} - \mathcal{P}_{\mathbf{G}\mathcal{P}_{\mathbf{X}_2}^{\perp}}\big\|_2 = O_P(\xi_{m,n}),$$

$$\Big\|\Big(\frac{\mathbf{G}\mathcal{P}_{\mathbf{X}_2}^{\perp}\mathbf{G}^{\top}}{n}\Big)^{1/2}\mathbf{O}\hat{\mathbf{G}} - \mathbf{G}\mathcal{P}_{\mathbf{X}_2}^{\perp}\Big\|_2 = O_P(\sqrt{n}\xi_{m,n}),$$

where $\xi_{m,n} = \frac{m}{sr_{m,n}}\Big(\frac{\sqrt{KM}}{n} + \frac{\|(\mathbf{B}_1)_{\mathcal{F}}\|_F}{\sqrt{m}}\Big)$, $\mathbf{O}$ is a $K \times K$ orthogonal matrix.

In Proposition 5.9, we make the assumption that the chosen NC set is nonrandom for simplification purposes. This assumption can be easily satisfied through a sampling splitting procedure: we use the first half of data for NC set estimation, and reserve the second half for latent basis estimation. Therefore, this assumption does not impose a significant restriction on our methodology.

It is worth taking a moment to compare Proposition 5.7 with Proposition 5.9. We can observe that the convergence rate for $\hat{\mathbf{\Gamma}}$ is approximately proportional to $\frac{1}{\sqrt{n}}$, while that for $\hat{\mathbf{G}}$ is $\frac{1}{\sqrt{m}} + \frac{1}{n}$. In $n = o(m)$ scenario, this suggests that estimation of $\mathbf{G}$ is more accurate compared to that of $\mathbf{\Gamma}$, highlighting the advantage of obtaining an explicit latent basis estimation. This phenomenon is also known as the "blessing of dimensionality" in factor models literature Bai (2003); Bing et al. (2022).

Another notable aspect in Proposition 5.9 is that we consider the inclusion of false negatives in our NC set. As mentioned after Theorem 5.8, it is unlikely to include strong rows; only those with ℓ_2 norms less than $\frac{c}{\sqrt{n}}$ for some $c > 0$ such that $c = O_P\Big((r_{m,n}M + K\log m)\cdot\sqrt{\frac{M|\mathcal{I}_1|}{m}}\Big)$ can be included. To ensure that $\sqrt{n}\xi_{m,n} = o_P(1)$, it is sufficient to have $c = o_P(r_{m,n})$. Therefore, the inclusion of weak false negatives poses no issue for our NC set estimation.

Theorem 5.10. *Suppose assumptions 5.1-5.3 hold, and assumption 5.6 hold if $\mathbf{F} \neq \mathbf{0}$. If $\xi_{m,n} =$*

$o\Big(\frac{1}{\sqrt{nM}}\Big)$, *where* $\xi_{m,n}$ *is the convergence rate of* $\mathcal{P}_{\hat{\mathbf{G}}}$ *to* $\mathcal{P}^{\perp}_{\mathbf{G}}\mathcal{P}_{\mathbf{X}_2}$, *then*

$$\sqrt{n}(\hat{\boldsymbol{b}}^{(1)}_{i,\text{final}} - \boldsymbol{b}^{(1)}_{i,\cdot}) \xrightarrow{d} N_c\big(\mathbf{0}, \sigma_i^2 \boldsymbol{\Sigma}^{-1}_{\mathbf{X}_1|\mathbf{X}_2,\mathbf{G}}\big), \quad \forall i = 1, \ldots, m,$$

where $\boldsymbol{\Sigma}_{\mathbf{X}_1|\mathbf{X}_2,\mathbf{G}}$ *is the covariance of remainder after removing the best linear projection of* $\mathbf{X}_1$ *onto* $\mathbf{X}_2$ *and* $\mathbf{G}$ *from* $\mathbf{X}_1$. *Additionally,*

$$\max_{1\le i\le m} \big|\hat{\sigma}_i^2 - \sigma_i^2\big| \xrightarrow{P} 0.$$

Theorem 5.10 establishes asymptotic distributions for our final estimator. The asymptotic variance $\sigma_i^2 \boldsymbol{\Sigma}^{-1}_{\mathbf{X}_1|\mathbf{X}_2,\mathbf{G}}$ is the same as that obtained if we regress $\boldsymbol{Y}_{i,\cdot}$ onto $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{G}$, demonstrating the efficiency of our final estimator. In the nonlinear case, assumption 5.6 holds and $\mathbf{E}_{\mathbf{X}} = \mathbf{X}_1 - \mathbb{E}\big[\mathbf{X}_1|\mathbf{A}\big]$. This implies that $\mathrm{Cov}(\mathbf{E}_{\mathbf{X}}) = \boldsymbol{\Sigma}_{\mathbf{X}_1|\mathbf{X}_2,\mathbf{G}}$ and $\hat{\boldsymbol{b}}^{(1)}_{i,\text{final}}$ is semi-parametrically efficient.

5.4. Consistency of FDP estimation

Theorem 5.11. *Suppose assumptions 5.1-5.5 hold, and assumption 5.6 hold if* $\mathbf{F} \neq \mathbf{0}$. *If* $\xi_{m,n} = o\Big(\frac{1}{\sqrt{nM\log m}}\Big)$, $Ma_{K+1}\log^2 m = o(1)$, *then the following conclusions hold:*

(1)

$$\hat{T}_i \xrightarrow{d} \begin{cases} \chi_c^2 & i \in \mathcal{I}_0 \cup \mathcal{I}_{11} \\ \chi_c^2(\cdot, \lambda_i) & i \in \mathcal{I}_{12} \\ \infty & i \notin \mathcal{I}_{13} \end{cases},$$

where $\chi_c^2(\cdot, \lambda_i)$ *is a non-central chi-squared distribution with non-centrality parameter* $\lambda_i = \frac{\boldsymbol{c}_i^\top \boldsymbol{\Sigma}_{\mathbf{X}_1|\mathbf{X}_2,\mathbf{G}} \boldsymbol{c}_i}{\sigma_i^2}$. *Furthermore, for* $\forall i \neq i'$, $\hat{T}_i$ *and* $\hat{T}_{i'}$ *are asymptotically independent.*

(2) *Suppose* $\hat{\mathbf{B}}_{1,\mathrm{ora}}$ *in Eq.*(4.11) *is used in constructing the oracle test statistics, then*

$$\sup_{0\le t\le 1}\left|\widehat{\mathrm{FDP}}(t)-\mathrm{FDP}_{\mathrm{ora}}(t)\right|\xrightarrow{P}0,$$

$$\sup_{0\le t\le 1}\left|\widehat{\mathrm{FDP}}_{\lambda}(t)-c\widehat{\mathrm{FDP}}(t)\right|\xrightarrow{P}0,$$

where $c=1+\frac{\pi_{11}}{\pi_0}+\frac{\pi_{12}(1-F_1(\lambda))}{\pi_0(1-\lambda)}$, $F_1(\lambda)=\lim_{m\to\infty}\frac{1}{m}\sum_{i=1}^{m}1-\chi_c^2\Big((\chi_c^2)^{-1}(1-\lambda,0),\lambda_i\Big)$.

(3) *Suppose* $\hat{\mathbf{B}}_{1,\mathrm{ora}}$ *in Eq.*(4.12) *is used in constructing the oracle test statistics, then for any* $\epsilon>0$,

$$\inf_{0\le t\le 1}\widehat{\mathrm{FDP}}(t)-\mathrm{FDP}_{\mathrm{ora}}(t)\ge-\epsilon$$

with probability tending to 1, *and*

$$\sup_{0\le t\le 1}\left|\widehat{\mathrm{FDP}}_{\lambda}(t)-c\widehat{\mathrm{FDP}}(t)\right|\xrightarrow{P}0.$$

Theorem 5.11 asserts that our constructed testing statistics are asymptotically independent, justifying the power of factor model idea in addressing general covariance dependence. $\widehat{\mathrm{FDP}}(t)$ is consistent to the oracle FDP in (2). Our estimation of $\widehat{\mathrm{FDP}}(t)$ is $\widehat{\mathrm{FDP}}_{\lambda}(t)$. $\widehat{\mathrm{FDP}}_{\lambda}(t)$ is consistent only when $\pi_{11}=0$ and λ is chosen such that $F_1(\lambda)=1$. But we observe that c is always greater than or equal to 1, indicating that the procedure based on $\widehat{\mathrm{FDP}}_{\lambda}(t)$ is slightly conservative but remains valid for the purpose of FDR control. For the oracle FDP in (3), we no longer have consistent estimation. Instead, $\widehat{\mathrm{FDP}}(t)$ is guaranteed to be a conservative substitute to the oracle FDP. Therefore, $\widehat{\mathrm{FDP}}_{\lambda}(t)$ is a doubly conservative substitute for $\mathrm{FDP}_{\mathrm{ora}}(t)$, and is sufficient for FDR control.

Chapter 6

Numerical experiments

6.1. Simulations

In this section, we assess the performance of our proposed multiple testing procedure, and compare it with other methods in the literature. Because not all methods were proposed for FDP estimation procedure, for fair comparison, we will also use metrics for general testing purpose: (1) Type-I error at $t = 0.05$, (2) test power at $t = 0.05$, (3) FDP at $t = 0.05$. Still, we are interested in (4) the "QQ-plots" of $\widehat{\mathrm{FDP}}_\lambda(t)$ versus $\mathrm{FDP}_{\mathrm{ora}}(t)$ as t changes from 0 to 1 based on p-values calculated from different methods.

We set our large dimensional multiple regression model to be

$$Y_{i,j} = b_{i,1}X_{1,j} + b_{i,2}X_{2,j} + u_{i,j},$$

$$u_{i,j} = f_i(\boldsymbol{\alpha}_{\cdot,j}) + \epsilon_{i,j},$$

$$i = 1, \ldots, m,\ j = 1, \ldots, n.$$

Here are some basic setups we use to generate our model:

(a) $m = 10000$, $n = 100$, $p = k = 2$, $\pi_0 = |\mathcal{I}_0|/m = 0.70$ or 0.95.

(b) $\epsilon_{i,j} \sim 80 \cdot N(0, \sigma_i^2)$ where σ_i are sampled from an inverse Gamma distribution with

parameters $(10, 9)$, so the standard deviation of the random error is around 1.

(c) $X_{1,j} \sim N(0, 1)$. We emphasize that we do not include independence or any special dependence structure assumption on $\mathbf{X}$ and $\boldsymbol{\alpha}$ for theoretical derivations, which makes a great improvement over many existing works performing multiple testings with latent variables. To simulate the situations where $\boldsymbol{\alpha}$ are unobserved confounders, we allow $X_{2,j}$ to be dependent on $\boldsymbol{\alpha}$. The dependence is described by either of the two following equations:

(c1) $X_{2,j} = 5\alpha_{1,j} - 3\alpha_{2,j} + \eta_j,\ \alpha_{i,j} \sim \text{Uniform}(0, 2\pi),\ \eta_j \sim N(0, 1)$;

(c2) $\alpha_{i,j} = X_{i,j} + \eta_{i,j}$, $X_{2,j} \sim \text{Uniform}(-2, 2)$, $\eta_{i,j} \sim N(0, 1)$.

The second case was the dependence structure considered in the confounder adjustment method proposed by Wang et al. (2017). Since their method was proposed specifically for linear factor models, we are specifically interested in comparing the performance of our method with theirs in linear (c2) case.

(d) $b_{i,1}$, $b_{i,2}$ are i.i.d. $N(0, 2.5^2)$ for $i = 1, \ldots, 3000$. $b_{i,1} = b_{i,2} \sim N(3.5, 1)$. To simulate weak signals, we also consider rescaling these coefficients by $1/\sqrt{n}$. It can be seen that our constraint matrix is $A = (1, -1)$, the testing problem Eq.(1.2) becomes

$$H_{0i} : b_{i,1} = b_{i,2} \quad \text{v.s.} \quad H_{1i} : b_{i,1} \neq b_{i,2} \quad i = 1, \ldots, 10000.$$

The null hypothesis is true for $i = 3001, \ldots, 10000$.

We consider two types of confounding mechanisms:

(i) linear case, $f_i(\boldsymbol{\alpha}_{\cdot,j}) = 1 + \boldsymbol{\gamma}_{i,\cdot}^\top \boldsymbol{\alpha}_{\cdot,j}$, where we set $\boldsymbol{\gamma}_{i,\cdot} \sim N_2(\mathbf{0}, 2.5^2\mathbf{I}_2)$;

(ii) nonlinear case, $f_i(\boldsymbol{\alpha}_{\cdot,j}) = \exp(\boldsymbol{\varsigma}_i^\top \boldsymbol{\alpha}_{\cdot,j})$, where $\boldsymbol{\varsigma}_i$ is 2-dimensional with i.i.d. $\text{Uniform}(0, 1)$ entries.

For comparison of different methods, we used the same estimated number of latent components: $\hat{K} = 3$ for linear cases, and $\hat{K} = 20$ for nonlinear cases. Each setup is repeated for 100 times. The methods considered are as follows:

- Unadjusted: This approach treats the multiple testing problem as if no unobserved confounders are present.
- NC-Adjusted: This method is proposed by Du and Zhang (2017). It assumes the availability of a well-chosen NC set. We use our NC set selection algorithm for this method.
- IRW-SVA: The iterative re-weighted surrogate variable analysis algorithm was proposed by Leek and Storey (2008). In each iteration, it computes a bootstrap estimation of $\mathbb{P}(b_{i,1} = b_{i,2}, f_i(\mathbf{A}) \neq 0 | \mathbf{Y}, \mathbf{X}, \hat{\mathbf{G}})$ as the weight used for weighted SVD.
- CATE: The confounder adjustment multiple testing procedure proposed by Wang et al. (2017). Different to χ^2 tests used in other methods, their method uses asymptotic Z-tests instead and does not consider addressing FDR control problem.
- NC-CASVA: Our proposed NC set selection method and multiple testing procedure with confounder adjustment using surrogate variables. It uses the same confounder adjustment technique as IRW-SVA but with a different NC set selection algorithm.
- Oracle: This method assumes knowledge of the unobserved latent components. In the linear case, we directly access the latent confounder $\mathbf{G}$, and the baseline will be based on the oracle estimator in Eq.(4.11). In nonlinear case, the baseline corresponds to the estimator in Eq.(4.12).

The results are summarized in Figure 6.1-6.4. Our proposed method behaves nearly identically to the oracle method across all evaluation metrics. The nominal size is controlled at level 0.05, and the estimated FDP is close to the oracle's. In the linear case, the power is also very similar to the oracle method. In the nonlinear case, the power is slightly lower due

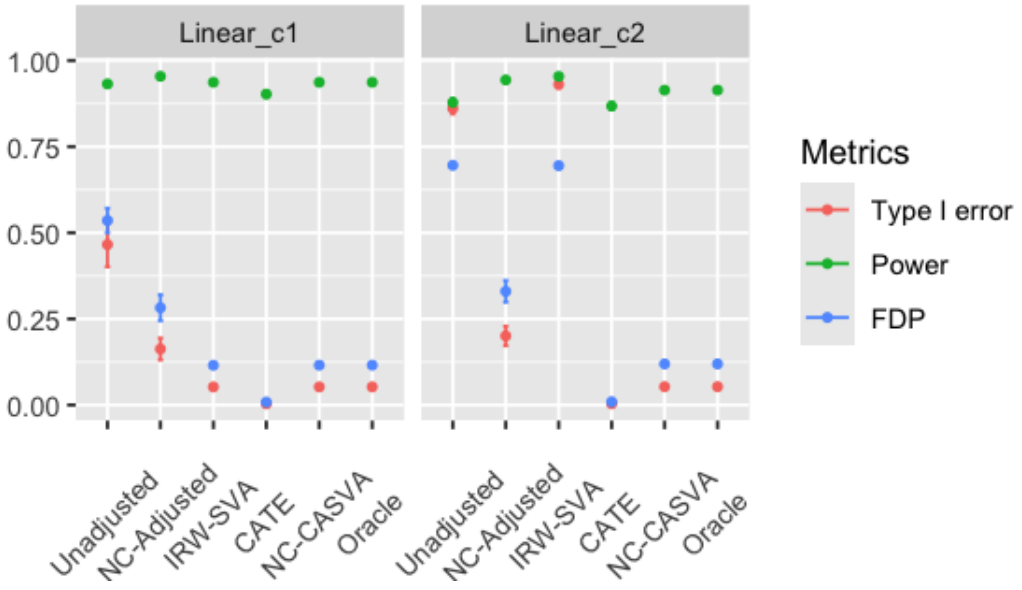

Figure 6.1: Compare the performance of different methods under linear latent confounding effects. Error bars are one standard deviations calculated based on 100 repeated simulations.

to the loss of estimation efficiency, as discussed below Eq.(4.12). These results demonstrate the effectiveness of our proposed testing procedure as a surrogate for the oracle method, making it suitable for mitigating linear and nonlinear confounding effects.

In our simulations, the latent components are confounding with the primary variables. Since NC-Adjusted and IRW-SVA were proposed for scenarios without confounding effects, it is unsurprising that their performance is mediocre.

Compared to our method, CATE effectively controls Type-I error and FDP, though with a slight reduction in test power. It is worth noting that the actual size of CATE method is not consistent to the nominal size, even in the linear case (c2), for which it was specifically designed. This inconsistency may arise from certain technical conditions needed for theoretical analysis of the CATE method. When the proportion of null hypotheses increases, their conditions are better satisfied, resulting in more consistent tests, as shown in Figure 6.3.

It could be valuable if CATE proves to be conservative across many other scenarios. However, without a solid theoretical understanding of CATE, its power may be limited in multiple testing contexts. Assume a researcher faces a multiple testing problem and

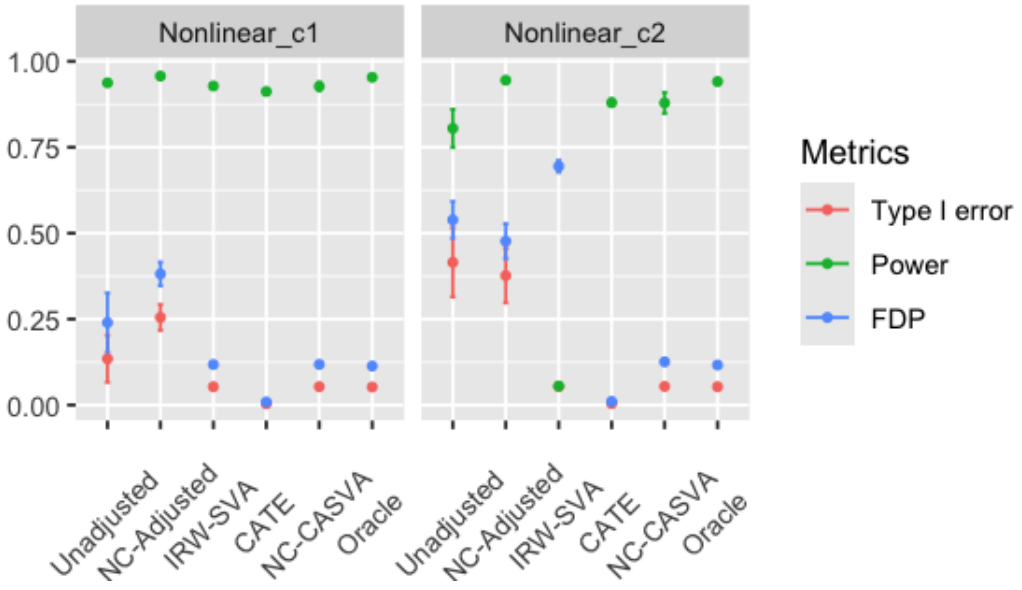

Figure 6.2: Compare the performance of different methods under nonlinear latent confounding effects. Error bars are one standard deviations calculated based on 100 repeated simulations.

applies the Storey's procedure to the p-values obtained from the CATE method to provide uncertainty quantification. Suppose the estimated FDP is twice as large as the actual FDP, as shown in Figure 6.4, and $\mathrm{FDP}_{\mathrm{ora}}(t) = 0.10$. If the researcher wants to control FDR at level 0.10, he would select a smaller threshold than t due to inconsistency in FDP estimation. This would result in even more reduced test power.

In summary, the simulation results align with our theoretical foundations. For the purpose of multiple testing with FDR control, our proposed method stands as a strong competitor among the available approaches.

6.2. Synthetic dataset

In this section,we use *Splatter* package in R to simulate a single-cell dataset with two groups and four different batches, which can be visualized in Figure 6.5. The responses are gene expression counts and we use standard log-normal transformation to preprocess the data. Technically, the counts are generated according to a Poisson-Gamma distribution, which

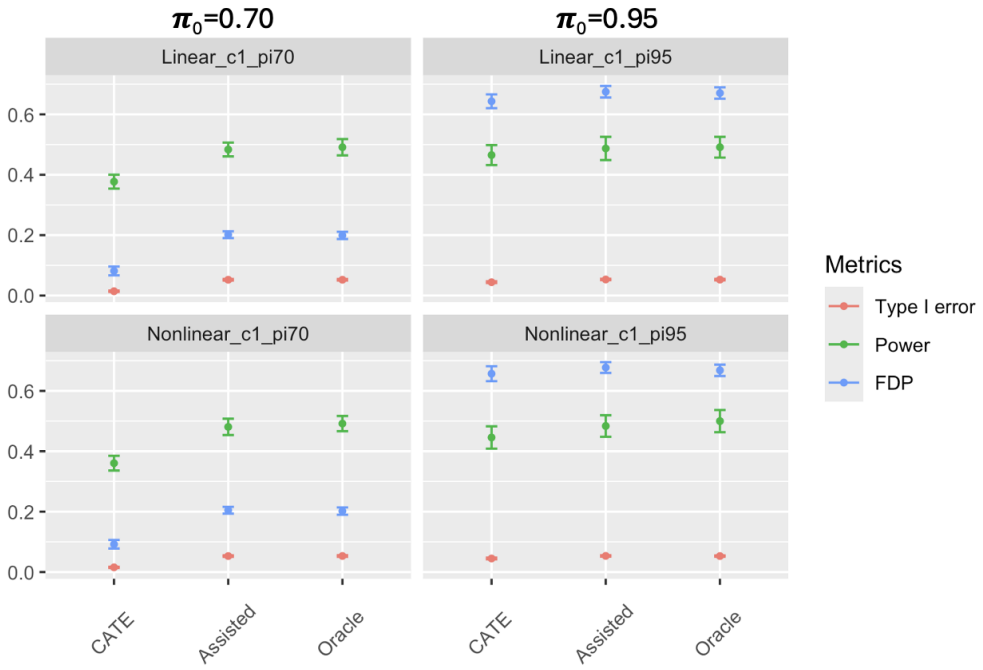

Figure 6.3: Comparisons with CATE in weak signal scenarios for linear and nonlinear models, and varying null hypotheses proportion. Error bars are one standard deviations calculated based on 100 repeated simulations.

does not align with our theoretical framework. The objective is to evaluate the robustness of our method in this scenario.

Since CATE stands out as a strong competitor in the previous section, it is the only method for comparison in this setup. To emphasize the differences between CATE and our proposed method in real applications, we adopt different rejection thresholds: for the CATE method, we use rule-of-thumb threshold of $\alpha = 0.05$; for our method, we focus on FDR control, and select a threshold at which the true FDP is approximately 0.05. The result is shown in Figure 6.6. Our method proves effective in approximating the true FDP even when the model is misspecified. In this case, our method controls the FDR with a slight sacrifice in test power compared to CATE. The advantage of using our method is that it provides an accurate estimation of the underlying FDP in practical applications.

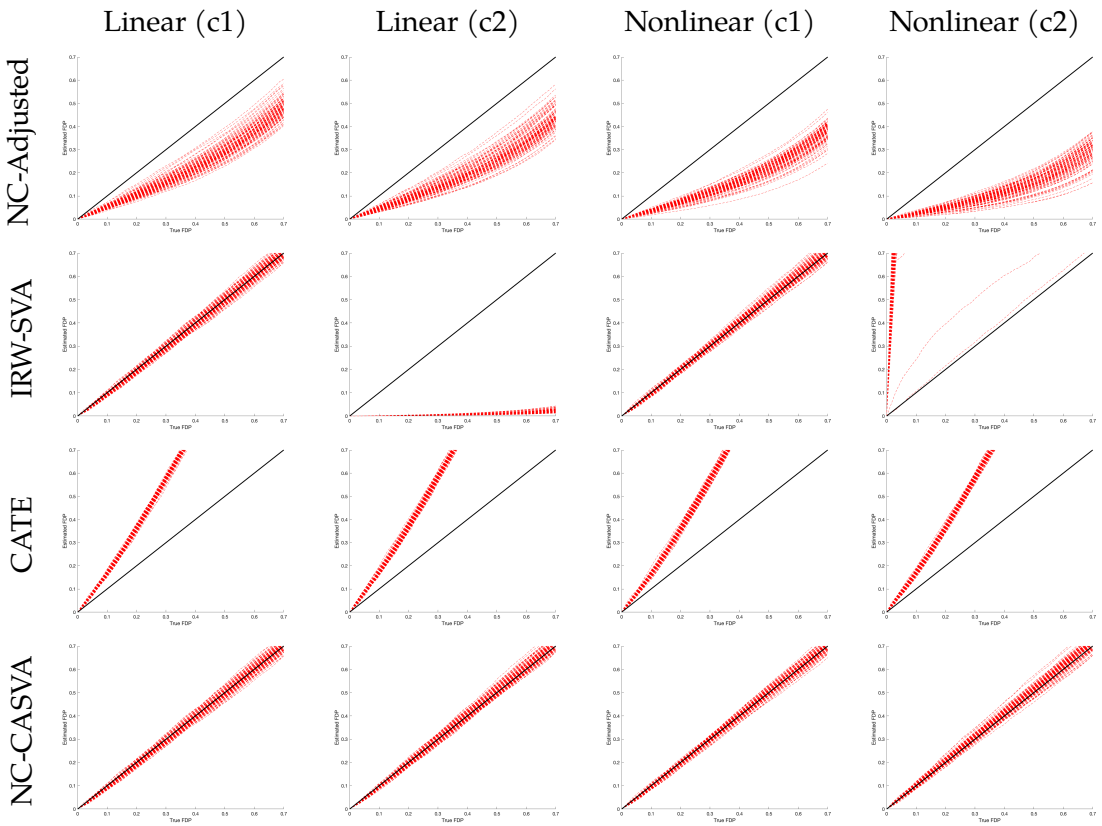

Figure 6.4: "QQ-plots" of estimated FDP versus oracle FDP under three different models and four different methods (from top to bottom): NC-Adjusted, IRW-SVA, CATE and our method NC-CASVA. The black straight line is $y = x$. The performance is better if dashed red lines are closer to the black line.

6.3. A real data application

The dataset we chose was repeatedly used in the literature, and its original version can be downloaded from GEO database (Series GSE2164). It was used in Vawter et al. (2004) for an investigation of gender differences in gene expression to known biological differences due to sex chromosome linked genes. Samples are consisted of 10 individuals, 5 men and 5 women. Each individual was sampled three times for three different brain regions. Each sample was analyzed by three different labs. In total, there should be $10 \times 3 \times 3 = 90$ samples. But 6 of them have missing information, resulting in 84 samples in the final dataset. The response variables consist of expression data from 12600 different genes. Gender is our primary interest. Brain regions and labs information are nuisance variables. Since it is

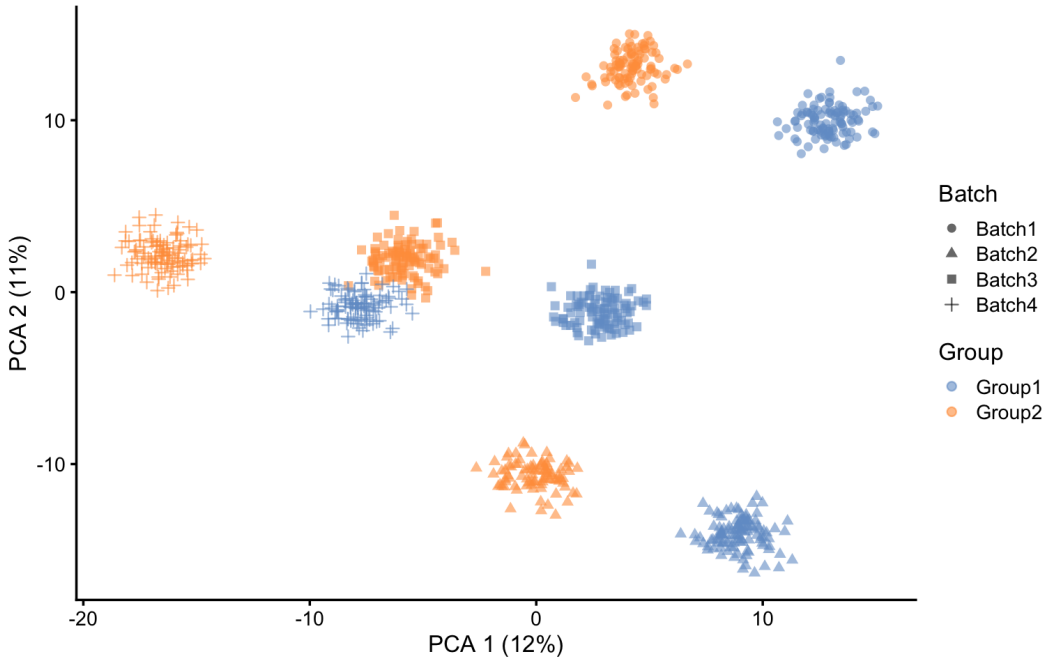

Figure 6.5: Standard PCA visualization of the group and batch effects in the simulated single-cell dataset.

generally believed that batch effects are unobserved and ubiquitous in data obtained from microarray-based experiment, latent factor models could be useful in isolating these effects. We follow the gold standard proposed by Gagnon-Bartsch, Jacob and Speed Gagnon-Bartsch et al. (2013), and use the ratio of significant genes coming from X/Y chromosomes genes as a benchmark in evaluating the performance of our multiple testing procedure. There are 488 such genes in the dataset.

Table 6.1 summarizes our results for this dataset. We chose the threshold in the multiple testing procedure by making our estimated FDP around 0.05. When $\hat{K} = 0$, the largest p-value is 0.0504, and almost all genes are significant at significance level 0.05. This is highly impossible in a genomic data study. The improvement is huge after we introduce latent factor components. To obtain the highest X/Y proportion, we may choose $\hat{K} = 1$. In this case, there are 11 genes coming from X/Y chromosomes among 14 significant genes. By the gold standard we mentioned, the result is very convincing. We expect there is only $14 \times 0.05 \approx 1$ gene to be false positive. However, with only one latent component it is insufficient to

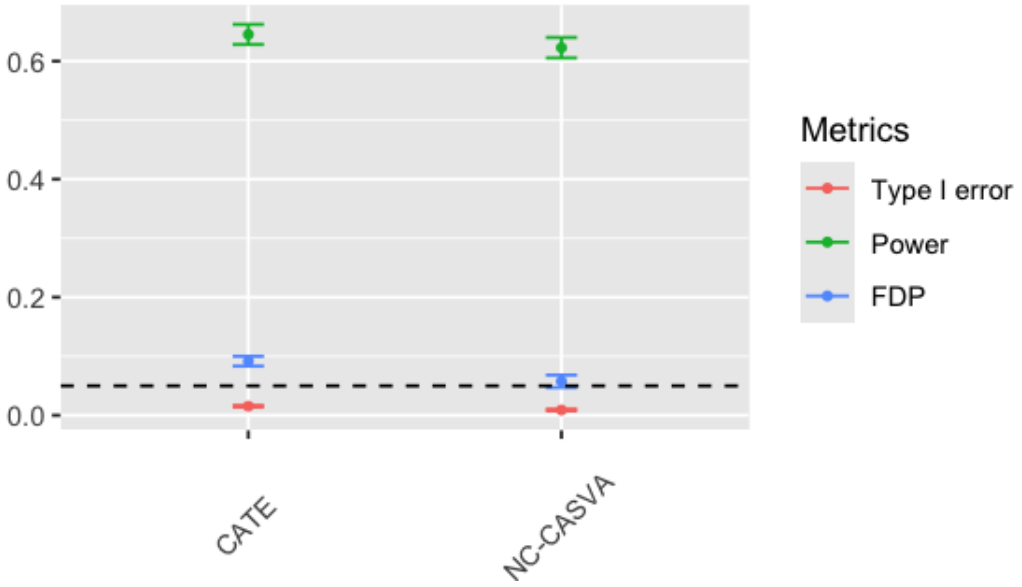

Figure 6.6: Comparison with CATE in the simulated single-cell dataset. For CATE, we select $\alpha = 0.05$ as commonly used in practice. For our method, we apply our FDR control procedure to select the significance level.

capture weak signals that are omnipresent in genomic datasets. Also, it was found that a total of 37% gene expression of all genes exhibit sex-biased expression in at least one tissue Oliva et al. (2020). Significant genes should not be restrictive to X/Y chromosomes.

The choice of $\hat{K}$ is a challenging topic. If we use the number of significant X/Y genes as a standard for choosing $\hat{K}$, we should choose a number that is no less than 10 as the number of significant X/Y genes becomes more stable. In Wang et al. (2017), they chose $\hat{K} = 25$ for this dataset, and they found 27 significant genes on X/Y chromosomes in top 100 genes. Given $\hat{K} = 25$, our method displays a similar result, with one more significant X/Y gene. Our testing procedure is expected to have $111 \times 0.05 \approx 5$ false positive genes.

Since our results are quite comparable to previous analysis of the same dataset, we claim that our approach provides a reliable way to choose the threshold in multiple testings, and to get a consistent estimation of FDP. This could be invaluable to scientific findings in which there may not have a gold standard like this to give an intuitive evaluation of false findings.

$\hat{K}$	threshold ($\times 10^{-4}$)	sig.	X/Y sig.	X/Y ratio in sig.	X/Y ratio in top 100
0	1.00×10^4	12600	17	0.0013	0.17
1	0.61	14	11	0.7857	0.23
2	0.72	18	14	0.7778	0.26
3	1.11	26	17	0.6538	0.25
4	1.14	28	18	0.6429	0.26
5	1.83	43	22	0.5116	0.27
10	6.71	140	27	0.1929	0.27
15	8.24	169	28	0.1657	0.28
20	3.59	85	27	0.3176	0.27
25	4.58	111	28	0.2523	0.28
30	5.65	133	27	0.2030	0.27

Table 6.1: Summary of findings in the GEO dataset with different numbers of latent components. The thresholds are chosen such that the estimated FDPs are around 0.05. The third to the last columns are: number of significant genes with given threshold, number of significant genes coming from X/Y chromosomes with given threshold, ratio of significant X/Y genes to significant genes, ratio of X/Y genes in the top 100 significant genes. When $\hat{K} = 0$ (no adjustment), p-values are all close to 0 and there is no meaningful choice for the threshold.

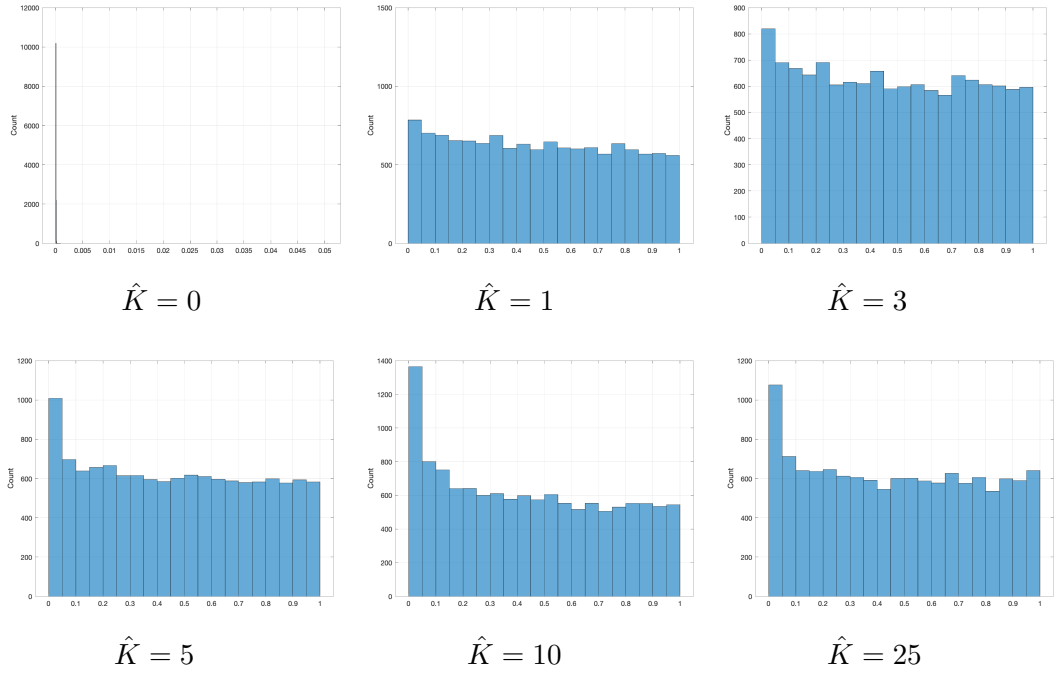

Figure 6.7: Histograms of p-values under different numbers of latent components for GEO dataset.

Chapter 7

Discussions

In this thesis, a new method is proposed to perform multiple testing for multiple regression models with unobserved confounders with FDR control. The cornerstone of our method lies in the application of a latent factor structure to account for covariance dependence. We demonstrate through a representation theorem about the power of using the approximate factor structure for covariance approximation. When the latent confounders are generated through nonlinear functions in a common Reproducing Kernel Hilbert Space (RKHS), the approximation is uniform. Subsequently, we assume the observed dataset is generated from multiple semi-parametric regression models. Even under the most challenging situation where the latent nonlinear component could be confounded with the variables of interest, we succeed in finding an intuitive method for making inference about regression coefficients. Our method involves solving a group lasso typo minimization problem and use the minimizer for NC set estimation. The NC set is guaranteed to rule out strong false positives with high probability, and is robust to weak false positives. Consequently, the latent basis, or unobserved confounders estimated by the NC set are natural "surrogate" for the latent effects and can be used to eliminate covariance dependence across different regression models. Our final estimator for the coefficient matrix is efficient, and asymptotically independent statistics can be constructed, thereby laying the ground for FDR control.

However, our method and theoretical analysis can be further improved in the following

aspects. Firstly, our method does not account for heterogeneity, which may impact the accuracy of our NC set estimation. Improvement can be achieved by proposing the estimation for idiosyncratic variances before completing NC set estimation. Secondly, our method heavily relies on full rankness of $\mathbf{\Gamma}_{\mathcal{S}}$ for the estimated NC set $\mathcal{S}$. If $\mathbf{\Gamma}$ is sparse, we may need to take into account its sparsity when obtaining the NC set estimation. Lastly, our analysis for NC set control is limited in two ends: either weak signals or strong signals. The analysis for the middle level where ℓ_2 norms for regression coefficient are of order $\Theta\left(\frac{1}{\sqrt{n}}\right)$ is challenging. As this could be of an independent research interest, we leave this for future research.

In conclusion, while our proposed method shows promise in addressing multiple testing of multiple regression models with unobserved confounders, we acknowledge that our numerical experiments are currently limited in space. We recognize the need for more extensive empirical validation across a wider range of scenarios and datasets. In future work, we plan to conduct additional numerical experiments to further evaluate the performance and robustness of our method under diverse conditions. These experiments will enable us to gain deeper insights into the strengths and limitations of our approach, ultimately enhancing its applicability and effectiveness in practical settings.

Appendix A

Preliminaries for the main proofs

A.1. Notations

For a matrix $\mathbf{X}$, $\mathrm{Col}(\mathbf{X})$ is the linear space spanned by columns of $\mathbf{X}$, $\mathrm{Null}(\mathbf{X})$ is the null space of $\mathbf{X}$, $\mathrm{rank}(\mathbf{X})$ is the rank of $\mathbf{X}$. We use $\|\mathbf{X}\|_2$ to denote the spectral (operator) norm of $\mathbf{X}$, and $\|\mathbf{X}\|_F$ for the Frobenius norm. When $\mathbf{X}$ becomes a square matrix, $\Lambda_{\min}(\mathbf{X})$ and $\Lambda_{\max}(\mathbf{X})$ are the smallest and largest eigenvalues of $\mathbf{X}$. $\mathrm{tr}(\mathbf{X})$ is the trace of $\mathbf{X}$. For two symmetric square matrices $\mathbf{X}, \mathbf{Y}$, $\mathbf{X} \succeq \mathbf{Y}$ if and only if $\Lambda_{\min}(\mathbf{X} - \mathbf{Y}) \geq 0$.

For a given row index set $\mathcal{S}$, $\mathbf{X}_{\mathcal{S}}$ can have two interpretations based on context. Primarily, it denotes the submatrix of $\mathbf{X}$ formed by selecting the corresponding rows in $\mathcal{S}$. Alternatively, it can be referred to a matrix with the identical dimension as $\mathbf{X}$, where the values align with $\mathbf{X}$ for the corresponding rows in $\mathcal{S}$, and are zeroes otherwise.

For asymptotic arguments, we use $\overset{a.s.}{\to}$, $\overset{P}{\to}$ and $\overset{d}{\to}$ to denote "converges almost surely", "converges in probability" and "converges in distribution" respectively. $a_n = o_P(b_n)$ if $\frac{a_n}{b_n} \overset{P}{\to} 0$. $a_n = O_P(b_n)$ if $|a_n| \leq C_\epsilon |b_n|$ for a constant C_ϵ dependent on $\epsilon > 0$ with probability at least $1 - \epsilon$. $a_n = \Omega_P(b_n)$ if $|a_n| \geq C_\epsilon |b_n|$ for a constant C_ϵ dependent on $\epsilon > 0$ with probability at least $1 - \epsilon$. $a_n = \Theta_P(b_n)$ if $a_n = O_P(b_n)$ and $a_n = \Omega_P(b_n)$. We say a given sequence of events $\{\mathcal{A}_n\}_{n=1}^{\infty}$ happens with high probability if $1_{\mathcal{A}_n^c} = o_P(1)$. $a_n \lesssim b_n$ implies $\exists C > 0$ such that $a_n \leq C b_n$ for large n, or equivalently, $a_n = O(b_n)$.

A.2. Some useful lemmas

To facilitate theoretical derivations of the main results, we outline some useful lemmas and intermediate conclusions below. Lemma A.1 can be found in Boucheron et al. (2003). Lemma A.2 is Weyl's inequality. Lemma A.3 is a more original version of Davis-Kahan $\sin\theta$ theorem. For statistical applications, we generalize the results in Yu et al. (2015) from Frobenius norm to spectral norm. Both lemmas are adapted from Bhatia (2013). Lemma A.4, Lemma A.5 are some concentration results for random matrices, and are directly borrowed from Vershynin (2018). Proposition A.6 establishes asymptotic results that will be repeatedly referenced in Appendix B.

Lemma A.1. *Suppose $X_1, \ldots, X_N$ have chi-squared distribution with p degrees of freedom, then*

$$\mathbb{E}\Big[\max_{i=1,\ldots,N} X_i - p\Big] \leq 2\sqrt{p\log N} + 2\log N.$$

Lemma A.2. *For symmetric matrices $\mathbf{M}_1$, $\mathbf{M}_2$,*

$$\Lambda_{\min}(\mathbf{M}_1 - \mathbf{M}_2) \leq \Lambda_i(\mathbf{M}_1) - \Lambda_i(\mathbf{M}_2) \leq \Lambda_{\max}(\mathbf{M}_1 - \mathbf{M}_2),\ \forall i,$$

where $\Lambda_i(\cdot)$ is the i-th largest eigenvalue of a matrix.

Lemma A.3. *Let $\mathbf{M}_1$, $\mathbf{M}_2$ be symmetric matrices, $\mathcal{S}_1$, $\mathcal{S}_2$ be two subsets in $\mathbb{R}$ that are separated at least $\delta > 0$ away from each other, $\mathcal{P}_{\mathbf{M}_1}(\mathcal{S}_1)$ be the projection matrix onto the space spanned by the eigenvectors of $\mathbf{M}_1$ with corresponding eigenvalues in $\mathcal{S}_1$. $\mathcal{P}_{\mathbf{M}_2}(\mathcal{S}_2^c)$ is defined similarly. Then if $\operatorname{rank}\big(\mathcal{P}_{\mathbf{M}_1}(\mathcal{S}_1)\big) = \operatorname{rank}\big(\mathcal{P}_{\mathbf{M}_2}(\mathcal{S}_2^c)\big)$,*

$$\|\mathcal{P}_{\mathbf{M}_1}(\mathcal{S}_1) - \mathcal{P}_{\mathbf{M}_2}(\mathcal{S}_2^c)\|_2 \leq \frac{\pi}{2\delta}\|\mathbf{M}_1 - \mathbf{M}_2\|_2.$$

If $\mathbf{X}$, $\mathbf{Y}$ are matrices with orthonormal columns such that $\mathcal{P}_{\mathbf{M}_1}(\mathcal{S}_1) = \mathbf{X}\mathbf{X}^\top$, $\mathcal{P}_{\mathbf{M}_2}(\mathcal{S}_2^c) = \mathbf{Y}\mathbf{Y}^\top$,

then there exists an orthogonal matrix $\mathbf{O}$ *such that*

$$\|\mathbf{XO} - \mathbf{Y}\|_2 \leq \frac{\sqrt{2}\pi}{2\delta}\|\mathbf{M}_1 - \mathbf{M}_2\|_2.$$

Proof: The first conclusion is from Bhatia (2013) and we only need to prove the second conclusion. Let $\mathbf{X}$, $\mathbf{Y}$ be of dimension $n \times l$. We assume that $n \geq 2l$, the case when $n < 2l$ can be proved similarly. By Theorem VII.1.8 in Bhatia (2013), there exist $n \times n$ orthogonal matrix $\mathbf{Q}$, $l \times l$ orthogonal matrices $\mathbf{O}_1$, $\mathbf{O}_2$ such that

$$\mathbf{QXO}_1 = \begin{pmatrix} \mathbf{I}_l \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix},$$

$$\mathbf{QYO}_2 = \begin{pmatrix} \mathbf{C} \\ \mathbf{S} \\ \mathbf{0} \end{pmatrix}$$

where $\mathbf{C}$, $\mathbf{S}$ are $l \times l$ nonnegative diagonal matrices such that $\mathbf{C}^2 + \mathbf{S}^2 = \mathbf{I}_l$. The diagonal elements in $\mathbf{S}$ are well-known in defining the $\sin\theta$ distance. Since the spectral norm is invariant to orthogonal transformations (left and right), the spectral norm of $\mathbf{XX}^\top - \mathbf{YY}^\top$ is that of

$$\begin{pmatrix} \mathbf{I}_l - \mathbf{C}^2 & -\mathbf{CS} \\ -\mathbf{SC} & -\mathbf{S}^2 \end{pmatrix}.$$

Also note that

$$\begin{pmatrix} \mathbf{S} & \mathbf{C} \\ -\mathbf{C} & \mathbf{S} \end{pmatrix}$$

is an orthonormal matrix and

$$\begin{pmatrix} \mathbf{I}_l - \mathbf{C}^2 & -\mathbf{CS} \\ -\mathbf{SC} & -\mathbf{S}^2 \end{pmatrix} \cdot \begin{pmatrix} \mathbf{S} & \mathbf{C} \\ -\mathbf{C} & \mathbf{S} \end{pmatrix} = \begin{pmatrix} \mathbf{S} & \mathbf{0} \\ \mathbf{0} & -\mathbf{S} \end{pmatrix},$$

we conclude that

$$\|\mathbf{S}\|_2 = \|\mathbf{X}\mathbf{X}^\top - \mathbf{Y}\mathbf{Y}^\top\|_2 = \|\mathcal{P}_{\mathbf{M}_1}(\mathcal{S}_1) - \mathcal{P}_{\mathbf{M}_2}(\mathcal{S}_2^c)\|_2.$$

Finally, choose $\mathbf{O} = \mathbf{O}_1\mathbf{O}_2^\top$, we have

$$\|\mathbf{X}\mathbf{O} - \mathbf{Y}\|_2^2 = \|\mathbf{Q}\mathbf{X}\mathbf{O}_1 - \mathbf{Q}\mathbf{Y}\mathbf{O}_2\|_2^2 \le \|\mathbf{I}_l - \mathbf{C}\|_2^2 + \|\mathbf{S}\|_2^2 \le 2\|\mathbf{S}\|_2^2,$$

and the result follows. ■

Lemma A.4. *Let* $\mathbf{M}$ *be an* $m \times n$ *matrix,* $m \ge n$*, whose row vectors* $\mathbf{M}_{i,\cdot}$ *are independent, mean-zero, sub-gaussian isotropic random vectors in* $\mathbb{R}^n$*. Then, for any* $t \ge 0$*, we have*

$$\sqrt{m} - CK^2(\sqrt{n} + t) \le s_n(\mathbf{M}) \le s_1(\mathbf{M}) \le \sqrt{m} + CK^2(\sqrt{n} + t)$$

with probability at least $1 - 2\exp(-t^2)$*. Here,* $C > 0$ *is a general constant,* $K = \max_i \|\mathbf{M}_{i,\cdot}\|_{\psi_2}$ *is the largest sub-gaussian norm for rows of* $\mathbf{M}$*,* $s_i(\cdot)$ *is the* i*-th largest singular value of a matrix.*

Lemma A.5. (*Matrix Bernstein's inequality*) *Let* $\mathbf{M}_1, \ldots, \mathbf{M}_n$ *be independent, mean-zero,* $m \times m$ *symmetric random matrices, such that* $\|\mathbf{M}_i\|_2 \le K$ *almost surely for all* i*. Then,*

(*1*) $\mathbb{P}\Big(\Big\|\sum_{i=1}^n \mathbf{M}_i\Big\|_2 \ge t\Big) \le 2m\exp\Big(-\frac{t^2/2}{\|\sum_{i=1}^n \mathbb{E}\mathbf{M}_i^2\|_2 + Kt/3}\Big)$.

(*2*) $\mathbb{E}\Big\|\sum_{i=1}^n \mathbf{M}_i\Big\|_2 \lesssim \Big\|\sum_{i=1}^n \mathbb{E}\mathbf{M}_i^2\Big\|_2^{1/2}\sqrt{1+\log m} + K(1+\log m)$.

Proposition A.6. *Suppose assumptions 5.1-5.3 hold for model* (3.1)*, then*

(*1*) $\max_{1\le i\le m}\|\boldsymbol{\epsilon}_{i,\cdot}\|_2^2 = O_P(n), \|\mathbf{E}\|_2 = O_P(\sqrt{m})$.

(*2*) $\max_{1\le i\le m}\|\boldsymbol{f}_{i,\cdot}\|_2^2 = o(n), \|\mathbf{F}\|_2 = O_P(\sqrt{m})$.

(*3*) $\Big\|\frac{1}{n}\mathbf{X}\mathbf{X}^\top - \mathbb{E}\big[\boldsymbol{X}_{\cdot,1}\boldsymbol{X}_{\cdot,1}^\top\big]\Big\|_2 \xrightarrow{P} 0$, $\big\|\frac{1}{n}\mathbf{G}\mathbf{G}^\top - \mathbf{I}_K\big\|_2 \xrightarrow{P} 0$, *and* $\Big\|\frac{1}{n}\mathbf{X}\mathbf{G}^\top - \mathbb{E}\big[\boldsymbol{X}_{\cdot,1}\boldsymbol{g}_{\cdot,1}^\top\big]\Big\|_2 \xrightarrow{P} 0$.

(*4*) $\Omega_P(n) = \Lambda_{\min}\Big(\mathbf{X}\mathcal{P}_{\mathbf{G}}^\perp\mathbf{X}^\top\Big) \le \Lambda_{\max}\Big(\mathbf{X}\mathcal{P}_{\mathbf{G}}^\perp\mathbf{X}^\top\Big) = O_P(n)$.

(5) $\Omega_P(n) = \Lambda_{\min}\Big(\mathbf{G}\mathcal{P}_{\mathbf{X}}^{\perp}\mathbf{G}^{\top}\Big) \leq \Lambda_{\max}\Big(\mathbf{G}\mathcal{P}_{\mathbf{X}}^{\perp}\mathbf{G}^{\top}\Big) = O_P(n)$.

(6) $\mathbb{E}\|\mathbf{M}\boldsymbol{\epsilon}_{i,\boldsymbol{\cdot}}\|_2^2 = \sigma_i^2\mathbb{E}\|\mathbf{M}\|_F^2$ *for any random matrix* $\mathbf{M}$ *independent of* $\boldsymbol{\epsilon}_{i,\boldsymbol{\cdot}}$, $i = 1, \ldots, m$.

Proof: (1) Let $\mathbf{D}$ be a $m \times m$ diagonal matrix where $\mathbf{D}_{ii} = 1/\sigma_i$. Then $\mathbf{D} \cdot \mathbf{E}$ is a $m \times n$ random matrix with i.i.d. $N(0,1)$ entries. $\frac{\|\boldsymbol{\epsilon}_{i,\boldsymbol{\cdot}}\|_2^2}{\sigma_i^2}$ follows a chi-squared distribution with degree of freedom n. From Lemma A.1,

$$\max_{1\leq i\leq m}\|\boldsymbol{\epsilon}_{i,\boldsymbol{\cdot}}\|_2^2 \leq \max_{1\leq i\leq m}\sigma_i^2 \max_{1\leq i\leq m}\frac{\|\boldsymbol{\epsilon}_{i,\boldsymbol{\cdot}}\|_2^2}{\sigma_i^2} = O_P\big(n + \sqrt{n\log m} + \log m\big) = O_P(n).$$

Additionally, Lemma A.4 implies that $\|\mathbf{DE}\|_2 = O_P(\sqrt{m})$. Hence

$$\|\mathbf{E}\|_2 \leq \|\mathbf{D}^{-1}\|_2 \cdot \|\mathbf{DE}\|_2 = O_P(\sqrt{m}).$$

(2) We will apply Lemma A.5(2) to $\mathbf{M}_j = \boldsymbol{f}_{\boldsymbol{\cdot},j}\boldsymbol{f}_{\boldsymbol{\cdot},j}^{\top} - \boldsymbol{\Sigma}_{\mathbf{F}}$, $j = 1, \ldots, n$. Note that

$$f_{i,j}^2 \leq Ma_{K+1}.$$

Then,

$$\|\boldsymbol{f}_{i,\boldsymbol{\cdot}}\|_2^2 \leq Ma_{K+1}n,$$

$$\|\boldsymbol{f}_{\boldsymbol{\cdot},j}\|_2^2 \leq Ma_{K+1}m,$$

$$\|\mathbf{M}_j\|_2 \leq \|\boldsymbol{f}_{\boldsymbol{\cdot},j}\|_2^2 + \|\boldsymbol{\Sigma}_{\mathbf{F}}\|_2,$$

$$\mathbb{E}\mathbf{M}_j^2 = \mathbb{E}(\boldsymbol{f}_{\boldsymbol{\cdot},j}\boldsymbol{f}_{\boldsymbol{\cdot},j}^{\top})^2 - \boldsymbol{\Sigma}_{\mathbf{F}}^2 \preceq \mathbb{E}\big[\|\boldsymbol{f}_{\boldsymbol{\cdot},j}\|_2^2\boldsymbol{f}_{\boldsymbol{\cdot},j}\boldsymbol{f}_{\boldsymbol{\cdot},j}^{\top}\big] \preceq Ma_{K+1}m\boldsymbol{\Sigma}_{\mathbf{F}}.$$

From Lemma A.5,

$$\mathbb{E}\Big\|\sum_{j=1}^{n}\mathbf{M}_j\Big\|_2 \lesssim \sqrt{Ma_{K+1}mn\log m\|\boldsymbol{\Sigma}_{\mathbf{F}}\|_2} + \Big(Ma_{K+1} + \frac{1}{n}\Big)m\log m. \tag{A.1}$$

Eq.(A.1) implies that

$$\|\mathbf{F}\mathbf{F}^\top\|_2 \le \Big\|\sum_{j=1}^{n}\mathbf{M}_j\Big\|_2 + n\|\mathbf{\Sigma_F}\|_2 = O_P(m) + O(m) = O_P(m),$$

whence the conclusion follows.

(3) Let $\mathbf{M}_{\bullet,j} = (\boldsymbol{X}_{\bullet,j}^\top, \boldsymbol{g}_{\bullet,j}^\top)^\top$ and $\mathbf{M} = (\mathbf{M}_{\bullet,1}, \ldots, \mathbf{M}_{\bullet,n})^\top$. $\widetilde{\mathbf{M}}_{\bullet,j} := \mathbf{\Sigma}_{\mathbf{X},\mathbf{G}}^{-1/2}\mathbf{M}_{\bullet,j}$ is an isotropic sub-gaussian random vector with sub-gaussian norm of order $O\Big(\sqrt{p}+\frac{1}{\sqrt{\lambda_K}}\Big)$. Apply Lemma A.4 to $\widetilde{\mathbf{M}} = (\widetilde{\mathbf{M}}_{\bullet,1}, \ldots, \widetilde{\mathbf{M}}_{\bullet,n})^\top$:

$$1 - \frac{Ca_{p,K,t}}{\sqrt{n}} \le s_{p+K}\Big(\frac{1}{\sqrt{n}}\widetilde{\mathbf{M}}\Big) \le s_1\Big(\frac{1}{\sqrt{n}}\widetilde{\mathbf{M}}\Big) \le 1 + \frac{Ca_{p,K,t}}{\sqrt{n}}$$

with probability greater than $1 - 2\exp(-t^2)$, where $a_{p,K,t} = \Big(p + \frac{1}{\lambda_K}\Big)(\sqrt{p+K} + t)$. Then, by Assumption 5.3,

$$\Big\|\frac{1}{n}\widetilde{\mathbf{M}}^\top\widetilde{\mathbf{M}} - \mathbf{I}_{p+K}\Big\|_2 = \max\Big(\Big|1 - s_{p+K}^2\Big(\frac{1}{\sqrt{n}}\widetilde{\mathbf{M}}\Big)\Big|, \Big|s_1^2\Big(\frac{1}{\sqrt{n}}\widetilde{\mathbf{M}}\Big) - 1\Big|\Big) \xrightarrow{P} 0.$$

By Assumption 5.2(a),

$$\Big\|\frac{1}{n}\mathbf{M}^\top\mathbf{M} - \mathbf{\Sigma}_{\mathbf{X},\mathbf{G}}\Big\|_2 \le \|\mathbf{\Sigma}_{\mathbf{X},\mathbf{G}}^{1/2}\|_2\Big\|\frac{1}{n}\widetilde{\mathbf{M}}^\top\widetilde{\mathbf{M}} - \mathbf{I}_{p+k}\Big\|_2\|\mathbf{\Sigma}_{\mathbf{X},\mathbf{G}}^{1/2}\|_2 \xrightarrow{P} 0.$$

The conclusion follows by considering corresponding blocks in $\frac{1}{n}\mathbf{M}^\top\mathbf{M} - \mathbf{\Sigma}_{\mathbf{X},\mathbf{G}}$.

(4) Partition $\mathbf{\Sigma}_{\mathbf{X},\mathbf{G}}$ with dimensions compatible to $\mathbf{X}$ and $\mathbf{G}$ as follows:

$$\mathbf{\Sigma}_{\mathbf{X},\mathbf{G}} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix}. \tag{A.2}$$

From (3),

$$\frac{\mathbf{X}\mathcal{P}_{\mathbf{G}}^{\perp}\mathbf{X}^\top}{n} = \frac{\mathbf{X}\mathbf{X}^\top}{n} - \frac{\mathbf{X}\mathbf{G}^\top}{n}\Big(\frac{\mathbf{G}\mathbf{G}^\top}{n}\Big)^{-1}\frac{\mathbf{G}\mathbf{X}^\top}{n} \xrightarrow{P} \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}.$$

Note that $\mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}$ is the inverse matrix of the upper-left block of $\mathbf{\Sigma}_{\mathbf{X},\mathbf{G}}^{-1}$. Since $\mathbf{\Sigma}_{\mathbf{X},\mathbf{G}} \succ \mathbf{0}$ and $\mathbf{\Sigma}_{22} \succ \mathbf{0}$, we conclude that $\mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21} \succ \mathbf{0}$ by Schur's complement. Furthermore, $\inf_{m,n} \Lambda_{\min}\big(\mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}\big) > 0$, and the conclusion follows.

(5) The proof is similar to (4).

(6) Consider the j-th row of $\mathbf{M}\boldsymbol{\epsilon}_{i,\cdot}$. Conditional on $\mathbf{M}$, it follows $N\big(\mathbf{0}, \sigma_i^2\|(\mathbf{M})_{j,\cdot}\|_2^2\big)$. Therefore,

$$\mathbb{E}\|\mathbf{M}\boldsymbol{\epsilon}_{i,\cdot}\|_2^2 = \sigma_i^2 \sum_j \mathbb{E}\|(\mathbf{M})_{j,\cdot}\|_2^2 = \sigma_i^2 \mathbb{E}\|\mathbf{M}\|_F^2.$$

■

Appendix B

Proofs of main results

B.1. Proof of Theorem 2.1

(1) Firstly, we show the "if and only if" condition for the existence of non-degenerate $\boldsymbol{\epsilon}_{\cdot,j}$. Any positive semi-definite matrix can be written as $\boldsymbol{\Gamma}_L\boldsymbol{\Gamma}_L^\top$ with appropriate rank by diagonalization. Thus, we only need to consider non-negative diagonal matrix $\mathbf{D_E}$ such that $\boldsymbol{\Sigma}_\mathbf{U} - \mathbf{D_E} \succeq \mathbf{0}$. $\mathbf{D_E}$ can be written as

$$\mathbf{D_E} = \sum_{i=1}^{m} d_i \boldsymbol{e}_i \boldsymbol{e}_i^\top, d_i \geq 0.$$

Necessity. If $\mathbf{D_E}$ is nonzero and $\boldsymbol{\Sigma}_\mathbf{U} - \mathbf{D_E} \succeq \mathbf{0}$, there exists some i' such that $d_{i'} > 0$ and $\boldsymbol{\Sigma}_\mathbf{U} - d_{i'}\boldsymbol{e}_{i'}\boldsymbol{e}_{i'}^\top \succeq \mathbf{0}$. For any $\boldsymbol{x} \in \mathrm{Null}(\boldsymbol{\Sigma}_\mathbf{U})$ we must have $\boldsymbol{e}_{i'}^\top \boldsymbol{x} = \mathbf{0}$ since otherwise $\boldsymbol{x}^\top(\boldsymbol{\Sigma}_\mathbf{U} - d_{i'}\boldsymbol{e}_{i'}\boldsymbol{e}_{i'}^\top)\boldsymbol{x} < 0$. $\boldsymbol{e}_{i'}$ is thus contained in $\mathrm{Col}(\boldsymbol{\Sigma}_\mathbf{U})$.

Sufficiency. Pick an arbitrary $\boldsymbol{e}_i \in \mathrm{Col}(\boldsymbol{\Sigma}_\mathbf{U})$, then there is a unique $\boldsymbol{y}_i \in \mathrm{Col}(\boldsymbol{\Sigma}_\boldsymbol{U})$ such that $\boldsymbol{e}_i = \boldsymbol{\Sigma}_\mathbf{U}\boldsymbol{y}_i$. For any $\boldsymbol{x} \in \mathrm{Null}(\boldsymbol{\Sigma}_\mathbf{U})$, $\boldsymbol{e}_i^\top \boldsymbol{x} = \boldsymbol{y}_i^\top \boldsymbol{\Sigma}_\mathbf{U}\boldsymbol{x} = \mathbf{0}$. Let

$$d_{\max,i} = \sup\{t \geq 0 : \boldsymbol{\Sigma}_\mathbf{U} - t\boldsymbol{e}_i\boldsymbol{e}_i^\top \succeq \mathbf{0}\}.$$

Any $\boldsymbol{x} \in \mathbb{R}^m$ can be decomposed as $\boldsymbol{x}_0 + \boldsymbol{x}_1$ for $\boldsymbol{x}_0 \in \mathrm{Null}(\boldsymbol{\Sigma}_\mathbf{U})$ and $\boldsymbol{x}_1 \in \mathrm{Col}(\boldsymbol{\Sigma}_\mathbf{U})$. Choose

a $\delta > 0$ such that δ is smaller than the smallest nonzero eigenvalue of $\mathbf{\Sigma_U}$. Then

$$\boldsymbol{x}^\top(\mathbf{\Sigma_U} - \delta \boldsymbol{e}_i\boldsymbol{e}_i^\top)\boldsymbol{x} = \boldsymbol{x}_1^\top(\mathbf{\Sigma_U} - \delta \boldsymbol{e}_i\boldsymbol{e}_i^\top)\boldsymbol{x}_1 \geq \boldsymbol{x}_1^\top \mathbf{\Sigma_U}\boldsymbol{x}_1 - \delta\|\boldsymbol{x}_1\|_2^2 > 0.$$

The above shows that $d_{\max,i} \geq \delta > 0$. Then sufficiency follows from $\mathbf{\Sigma_U} - d_{\max,i}\boldsymbol{e}_i\boldsymbol{e}_i^\top = \lim_{t \uparrow d_{\max,i}} \mathbf{\Sigma_U} - t\boldsymbol{e}_i\boldsymbol{e}_i^\top \succeq \mathbf{0}$. Furthermore, with a more detailed argument it can be shown that $d_{\max,i} = \frac{1}{\boldsymbol{e}_i^\top \boldsymbol{y}_i} = \frac{1}{\boldsymbol{y}_i^\top \mathbf{\Sigma_U} \boldsymbol{y}_i}$ and $\mathrm{rank}(\mathbf{\Sigma_U} - d_{\max,i}\boldsymbol{e}_i\boldsymbol{e}_i^\top) = \mathrm{rank}(\mathbf{\Sigma_U}) - 1$.

Now consider the construction of the decomposition in Eq.(2.6). By i.i.d. assumption, it suffices to consider decomposition of $\boldsymbol{u}_{\bullet,1}$. Assume WLOG that $\mathbb{E}[\boldsymbol{u}_{\bullet,1}] = \mathbf{0}$. We can always choose a non-negative diagonal matrix $\mathbf{D_E}$ such that $r = \mathrm{rank}(\mathbf{\Sigma_U} - \mathbf{D_E}) \leq m - 1$. To see this, if $\mathbf{\Sigma_U}$ is full rank, $\boldsymbol{e}_i \in \mathrm{Col}(\mathbf{\Sigma_U})$ for any i and we can choose $\mathbf{D_E}$ as in the proof of sufficiency; or otherwise, it holds with $\mathbf{D_E} = \mathbf{0}$. When $\mathbf{\Sigma_U}$ is non-singular, the proof is straightforward. In case of any singularity of $\mathbf{\Sigma_U}$, suppose $\mathrm{rank}(\mathbf{\Sigma_U}) = k \leq m$. Through diagonalization of $\mathbf{\Sigma_U}$, there exists an orthogonal matrix $\mathbf{O} = (\mathbf{O}_1, \mathbf{O}_2)$ where $\mathbf{O}_1$ is $m \times k$, $\mathbf{O}_2$ is $m \times (m-k)$ such that

$$\mathbf{D} := \mathrm{Cov}(\mathbf{O}_1^\top \boldsymbol{u}_{\bullet,1}) \text{ is a } k \times k \text{ positive diagonal matrix,}$$

and

$$\mathbf{O}_1\mathbf{D}\mathbf{O}_1^\top = \mathbf{\Sigma_U}, \quad \mathbf{O}_2^\top \mathbf{\Sigma_U}\mathbf{O}_2 = \mathbf{0}, \quad \mathbf{O}_2^\top \boldsymbol{u}_{\bullet,1} = \mathbf{0}\ a.s.$$

Let $\mathbf{\Sigma_U} - \mathbf{D_E} = \mathbf{\Gamma}_L\mathbf{\Gamma}_L^\top$ for a $m \times r$ matrix $\mathbf{\Gamma}_L$. Since $\mathbf{\Sigma_U} \succeq \mathbf{\Gamma}_L\mathbf{\Gamma}_L^\top$,

$$\mathbf{0} = \mathbf{O}_2^\top \mathbf{\Sigma_U}\mathbf{O}_2 \succeq \mathbf{O}_2^\top \mathbf{\Gamma}_L\mathbf{\Gamma}_L^\top\mathbf{O}_2 \succeq \mathbf{0},$$

we get

$$\mathbf{O}_2^\top \mathbf{\Gamma}_L\mathbf{\Gamma}_L^\top\mathbf{O}_2 = \mathbf{0}.$$

Let $\widetilde{\boldsymbol{g}}_1$ be sampled independently to $\boldsymbol{u}_{\bullet,1}$, and

$$\widetilde{\boldsymbol{g}}_1 \sim N_r(\mathbf{0}, \mathbf{I}_r - \boldsymbol{\Gamma}_L^\top \mathbf{O}_1 \mathbf{D}^{-1} \mathbf{O}_1^\top \boldsymbol{\Gamma}_L).^1$$

Then for $\boldsymbol{g}_{\bullet,1}^{(L)} = \widetilde{\boldsymbol{g}}_1 + \boldsymbol{\Gamma}_L^\top \mathbf{O}_1 \mathbf{D}^{-1} \mathbf{O}_1^\top \boldsymbol{u}_{\bullet,1}$,

$$\mathrm{Cov}(\mathbf{O}_1^\top \boldsymbol{u}_{\bullet,1}, \boldsymbol{g}_{\bullet,1}^{(L)}) = \begin{pmatrix} \mathbf{D} & \mathbf{O}_1^\top \boldsymbol{\Gamma}_L \\ \boldsymbol{\Gamma}_L^\top \mathbf{O}_1 & \mathbf{I}_r \end{pmatrix}.$$

$\boldsymbol{u}_{\bullet,1} = \mathbf{O}\mathbf{O}^\top \boldsymbol{u}_{\bullet,1} = \mathbf{O}_1 \mathbf{O}_1^\top \boldsymbol{u}_{\bullet,1} + \mathbf{O}_2 \mathbf{O}_2^\top \boldsymbol{u}_{\bullet,1} = \mathbf{O}_1 \mathbf{O}_1^\top \boldsymbol{u}_{\bullet,1}$ a.s.,

$$\mathrm{Cov}(\boldsymbol{u}_{\bullet,1}, \boldsymbol{g}_{\bullet,1}^{(L)}) = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{U}} & \mathbf{O}_1 \mathbf{O}_1^\top \boldsymbol{\Gamma}_L \\ \boldsymbol{\Gamma}_L^\top \mathbf{O}_1 \mathbf{O}_1^\top & \mathbf{I}_r \end{pmatrix}. \tag{B.1}$$

Note that

$$\mathbf{O}_2 \mathbf{O}_2^\top \boldsymbol{\Gamma}_L (\mathbf{O}_2 \mathbf{O}_2^\top \boldsymbol{\Gamma}_L)^\top = \mathbf{O}_2 \mathbf{O}_2^\top \boldsymbol{\Gamma}_L \boldsymbol{\Gamma}_L^\top \mathbf{O}_2 \mathbf{O}_2^\top = \mathbf{0},$$

$$\mathbf{O}_2 \mathbf{O}_2^\top \boldsymbol{\Gamma}_L = \mathbf{0}.$$

Then

$$\mathbf{O}_1 \mathbf{O}_1^\top \boldsymbol{\Gamma}_L = \mathbf{O}_1 \mathbf{O}_1^\top \boldsymbol{\Gamma}_L + \mathbf{O}_2 \mathbf{O}_2^\top \boldsymbol{\Gamma}_L = \mathbf{O}\mathbf{O}^\top \boldsymbol{\Gamma}_L = \boldsymbol{\Gamma}_L.$$

From Eq.(B.1), $\boldsymbol{\epsilon}_{\bullet,1} = \boldsymbol{u}_{\bullet,1} - \boldsymbol{\Gamma}_L \boldsymbol{g}_{\bullet,1}^{(L)}$ and $\boldsymbol{g}_{\bullet,1}^{(L)}$ satisfy all the required properties.

(2) Let v be sampled from $\mathrm{Uniform}(0,1)$. Digits in dyadic representation of v are known to be i.i.d. Bernouli$(1/2)$. Initialize $u_1, \ldots, u_m$ to be zeroes. Add $\frac{v_{km+r}}{2^{k+1}}$ to u_{r+1}, where v_{km+r} is the $(km+r)$-th digit in dyadic representation of v, $k \in \mathbb{N}, r = 0, \ldots, m-1$. Because $u_1, \ldots, u_m$ use different sets of digits in v, $u_1, \ldots, u_m$ are i.i.d. $\mathrm{Uniform}(0,1)$.

Note that $(u_1, \ldots, u_m)$ can be viewed as a Borel function of v. Apply Theorem 2(a) in Rüschendorf and de Valk (1993) for $u_1, \ldots, u_m$, there exists a Borel function $\boldsymbol{f} \in \mathbb{R}^m$ such that $\boldsymbol{f}(v) \overset{d}{=} \boldsymbol{\Gamma}_L \boldsymbol{g}_{\bullet,1}^{(L)}$. $\boldsymbol{f}$ is a measurable map from $\left(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P}_{uniform}\right)$ to $\left(\mathbb{R}^m, \mathcal{B}(\mathbb{R}^m)\right)$,

[1] Using Schur's complement, $\mathbf{I}_r - \boldsymbol{\Gamma}^\top \mathbf{O}_1 \mathbf{D}^{-1} \mathbf{O}_1^\top \boldsymbol{\Gamma}$ can be shown to be positive semi-definite.

$\mathbf{\Gamma}_L \boldsymbol{g}_{\boldsymbol{\cdot},1}^{(L)}$ is a measurable map from $(\Omega, \mathcal{F}, P)$ to $\left(\mathbb{R}^m, \mathcal{B}(\mathbb{R}^m)\right)$. $\boldsymbol{f}$ and $\mathbf{\Gamma}_L \boldsymbol{g}_{\boldsymbol{\cdot},1}^{(L)}$ share the same induced measure as shown:

$$\mathbb{P}_{uniform}^{\boldsymbol{f}} = \mathbb{P}^{\mathbf{\Gamma}_L \boldsymbol{g}_{\boldsymbol{\cdot},1}^{(L)}}.$$

Apply Proposition 1 in Rachev and Ruschendorf (1991), there exists a random variable $\boldsymbol{\alpha}_{\boldsymbol{\cdot},1} : \Omega \to \mathbb{R}$ such that

$$\boldsymbol{f}(\boldsymbol{\alpha}_{\boldsymbol{\cdot},1}) = \mathbf{\Gamma}_L \boldsymbol{g}_{\boldsymbol{\cdot},1}^{(L)} \; a.s.[\mathbb{P}] \text{ and } \mathbb{P}^{\boldsymbol{\alpha}_{\boldsymbol{\cdot},1}} = \mathbb{P}_{uniform}.$$

That is, $\boldsymbol{\alpha}_{\boldsymbol{\cdot},1} \sim \text{Uniform}(0,1)$ as required. ■

B.2. Proof of Proposition 2.2

The main proof is detailed in Wahba (1990). The original claim states that any function $h \in L^2(\mathbb{P}^{\boldsymbol{\alpha}})$ with $\|h\|_{\mathcal{H}_R} < \infty$ will belong to $\mathcal{H}_R$. However, this assertion is not rigorous since h must reside in the linear space spanned by eigenfunctions, as we will demonstrate below.

Let $\mathcal{V}$ be the linear subspace in $L^2(\mathbb{P}^{\boldsymbol{\alpha}})$ spanned by $\{\Phi_v\}_{v=1}^{\text{rank}(R)}$. Suppose $h \in L^2(\mathbb{P}^{\boldsymbol{\alpha}})$ with $\|h\|_{\mathcal{H}_R} < \infty$. By orthogonal projection of h onto $\mathcal{V}$,

$$h = \sum_{v=1}^{\text{rank}(R)} \langle h, \Phi_v \rangle_{L^2(\mathbb{P}^{\boldsymbol{\alpha}})} \Phi_v + h^*,$$

where h^* is orthogonal to $\mathcal{V}$. If $h \in \mathcal{H}_R$, then $h^* \in \mathcal{H}_R$. By the reproducing property,

$$h^*(t) = \left\langle h^*, R(t,\cdot) \right\rangle_{\mathcal{H}_R} = \left\langle h^*, \sum_{v=1}^{\text{rank}(R)} \lambda_v \Phi_v(t) \Phi_v(\cdot) \right\rangle_{\mathcal{H}_R} = 0.$$

for any $t \in \mathcal{T}$, which implies that h must be spanned by $\{\Phi_v\}_{v=1}^{\text{rank}(R)}$. ■

B.3. Proof of Proposition 3.2

This proof extends Theorem 6.9 in Foucart and Rauhut (2013) from ℓ_1 minimization to the group lasso penalty. For simplicity, let $\mathcal{S}$ be the collection of row index i such that $\boldsymbol{b}_{i,\cdot}^{(1)} \neq \mathbf{0}$, and $s = |\mathcal{S}| = \|\mathbf{B}_1\|_{1,0}$, $\delta_s = \delta\big(\|\mathbf{B}_1\|_{1,0}, \mathbf{\Gamma}\big)$. Suppose $\mathbf{B}_1 + \mathbf{V}$ is a minimizer to Eq.(3.10). Then, $\mathcal{P}_{\mathbf{\Gamma}}^{\perp}\mathbf{V} = \mathbf{0}$ and

$$\|(\mathbf{B}_1 + \mathbf{V})_{\mathcal{S}}\|_{1,2} + \|\mathbf{V}_{\mathcal{S}^c}\|_{1,2} = \|\mathbf{B}_1 + \mathbf{V}\|_{1,2} \leq \|\mathbf{B}_1\|_{1,2} = \|(\mathbf{B}_1)_{\mathcal{S}}\|_{1,2}. \tag{B.2}$$

Note that $\|(\mathbf{B}_1 + \mathbf{V})_{\mathcal{S}}\|_{1,2} \geq \|(\mathbf{B}_1)_{\mathcal{S}}\|_{1,2} - \|\mathbf{V}_{\mathcal{S}}\|_{1,2}$, Eq.(B.2) implies that

$$\|\mathbf{V}_{\mathcal{S}}\|_{1,2} \geq \|\mathbf{V}_{\mathcal{S}^c}\|_{1,2}. \tag{B.3}$$

We will establish a contradiction to Eq.(B.3) if $\mathbf{V} \neq \mathbf{0}$, thereby showing that $\mathbf{B}_1$ is the only minimizer to Eq.(3.10).

Let $\mathcal{S}_0 \cup \mathcal{S}_1 \ldots$ be a partition of the index set $\{1, \ldots, m\}$. Here, $\mathcal{S}_0$ is the collection of s largest rows of $\mathbf{V}$ in terms of ℓ_2 norm, $\mathcal{S}_1$ is the collection of s largest rows in $\mathbf{V}_{\mathcal{S}_0^c}$ in terms of ℓ_2 norm, and so on. Since $\mathcal{P}_{\mathbf{\Gamma}}^{\perp}\mathbf{V} = \mathbf{0}$, we have $\mathcal{P}_{\mathbf{\Gamma}}^{\perp}\mathbf{V}_{\mathcal{S}_0} = -\sum_{k\geq 1}\mathcal{P}_{\mathbf{\Gamma}}^{\perp}\mathbf{V}_{\mathcal{S}_k}$, and

$$\begin{aligned}
\|\mathbf{V}_{\mathcal{S}_0}\|_F^2 &\leq \frac{1}{1-\delta_s}\|\mathcal{P}_{\mathbf{\Gamma}}^{\perp}\mathbf{V}_{\mathcal{S}_0}\|_F^2 \\
&= \frac{1}{1-\delta_s}\mathrm{tr}\Big(\mathbf{V}_{\mathcal{S}_0}^{\top}\mathcal{P}_{\mathbf{\Gamma}}^{\perp}\sum_{k\geq 1}(-\mathbf{V}_{\mathcal{S}_k})\Big) \\
&= \frac{1}{1-\delta_s}\mathrm{tr}\Big(\mathbf{V}_{\mathcal{S}_0}^{\top}\mathcal{P}_{\mathbf{\Gamma}}\sum_{k\geq 1}(-\mathbf{V}_{\mathcal{S}_k})\Big) \\
&\leq \frac{1}{1-\delta_s}\|\mathcal{P}_{\mathbf{\Gamma}}\mathbf{V}_{\mathcal{S}_0}\|_F\|\mathcal{P}_{\mathbf{\Gamma}}\mathbf{V}_{\mathcal{S}_k}\|_F \\
&\leq \frac{\delta_s}{1-\delta_s}\|\mathbf{V}_{\mathcal{S}_0}\|_F\|\mathbf{V}_{\mathcal{S}_k}\|_F,
\end{aligned} \tag{B.4}$$

where we use the definition in Eq.(3.11) for the first line and the last line. The second line

uses the fact that $\mathbf{V}_{\mathcal{S}_0}$ and $\mathbf{V}_{\mathcal{S}_k}$ have disjoint supports. Note that by our construction of $\mathcal{S}_k$,

$$\|\mathbf{V}_{\mathcal{S}_k}\|_F \leq \frac{1}{\sqrt{s}}\|\mathbf{V}_{\mathcal{S}_{k-1}}\|_{1,2}, \quad k \geq 1. \tag{B.5}$$

Combining this with Eq.(B.4),

$$\|\mathbf{V}_{\mathcal{S}_0}\|_{1,2} \leq \sqrt{s}\|\mathbf{V}_{\mathcal{S}_0}\|_F \leq \frac{\delta_s}{1-\delta_s}\sum_{k\geq 0}\|\mathbf{V}_{\mathcal{S}_k}\|_{1,2} < \frac{1}{2}\|\mathbf{V}\|_{1,2}.$$

Since $\mathcal{S}_0$ contains the largest s rows,

$$2\|\mathbf{V}_{\mathcal{S}}\|_{1,2} \leq 2\|\mathbf{V}_{\mathcal{S}_0}\|_{1,2} < \|\mathbf{V}\|_{1,2} = \|\mathbf{V}_{\mathcal{S}}\|_{1,2} + \|\mathbf{V}_{\mathcal{S}^c}\|_{1,2}.$$

This contradicts Eq.(B.3) if $\mathbf{V} \neq \mathbf{0}$. Thus, $\mathbf{B}_1$ must be the only minimizer to Eq.(3.10). ■

B.4. Proof of Lemma 3.3

Let $\mathbf{P}_{01}$ be the submatrix of $\mathcal{P}_{\mathbf{\Gamma}}^{\perp}$ formed by the corresponding columns in $\mathcal{S}_{01}$, $\mathcal{P}_{\mathbf{P}}$ be the projection matrix onto $\mathrm{Col}(\mathbf{P}_{01})$. Then, $\mathbf{I}_{2s} - \mathbf{P}_{01}^{\top}\mathbf{P}_{01}$ is the lower right $2s \times 2s$ submatrix of $\mathcal{P}_{\mathbf{\Gamma}}$. By the definition of δ,

$$\|\mathbf{I}_{2s} - \mathbf{P}_{01}^{\top}\mathbf{P}_{01}\|_2 \leq \delta,$$

which implies that

$$(1-\delta)\|\mathbf{P}_{01}\mathbf{C}\|_F^2 \leq \|\mathbf{P}_{01}^{\top}\mathbf{P}_{01}\mathbf{C}\|_F^2 \tag{B.6}$$

for any matrix $\mathbf{C}$. Since $\mathcal{P}_{\mathbf{P}}\mathcal{P}_{\mathbf{\Gamma}}^{\perp}\mathbf{V} = \mathbf{P}_{01}\mathbf{C}$ for some $\mathbf{C}$, and $\mathcal{P}_{\mathbf{P}}\mathbf{P}_{01} = \mathbf{P}_{01}$, we have

$$\sqrt{1-\delta}\left\|\mathcal{P}_{\mathbf{P}}\mathcal{P}_{\mathbf{\Gamma}}^{\perp}\mathbf{V}\right\|_F \leq \left\|\mathbf{P}_{01}^{\top}\mathcal{P}_{\mathbf{\Gamma}}^{\perp}\mathbf{V}\right\|_F = \left\|\mathbf{P}_{01}^{\top}\mathbf{V}\right\|_F = \left\|(\mathcal{P}_{\mathbf{\Gamma}}^{\perp}\mathbf{V})_{\mathcal{S}_{01}}\right\|_F. \tag{B.7}$$

Let $\mathcal{S}_0 \cup \mathcal{S}_1 \cup \mathcal{S}_2 \ldots$ be a partition of the index set $\{1, \ldots, m\}$, where $\mathcal{S}_k$ $(k \geq 2)$ is defined

in the same way as in the proof of Proposition 3.2. Decompose $\mathcal{P}_{\mathbf{P}}\mathcal{P}_{\mathbf{\Gamma}}^{\perp}\mathbf{V}$ as:

$$\mathcal{P}_{\mathbf{P}}\mathcal{P}_{\mathbf{\Gamma}}^{\perp}\mathbf{V} = \mathcal{P}_{\mathbf{\Gamma}}^{\perp}\mathbf{V}_{\mathcal{S}_{01}} + \sum_{k\geq 2}\mathcal{P}_{\mathbf{P}}\mathcal{P}_{\mathbf{\Gamma}}^{\perp}\mathbf{V}_{\mathcal{S}_k}. \tag{B.8}$$

We know $\mathcal{P}_{\mathbf{P}}\mathcal{P}_{\mathbf{\Gamma}}^{\perp}\mathbf{V}_{\mathcal{S}_k} = \mathcal{P}_{\mathbf{\Gamma}}^{\perp}\mathbf{C}_k$ for some matrix $\mathbf{C}_k$ supported on $\mathcal{S}_{01}$, $k \geq 2$. Then,

$$\begin{aligned}
\left\|\mathcal{P}_{\mathbf{P}}\mathcal{P}_{\mathbf{\Gamma}}^{\perp}\mathbf{V}_{\mathcal{S}_k}\right\|_F^2 &= \mathrm{tr}\left(\mathbf{V}_{\mathcal{S}_k}^{\top}\mathcal{P}_{\mathbf{\Gamma}}^{\perp}\mathcal{P}_{\mathbf{P}}\mathcal{P}_{\mathbf{\Gamma}}^{\perp}\mathbf{V}_{\mathcal{S}_k}\right) = \mathrm{tr}\left(\mathbf{C}_k^{\top}\mathcal{P}_{\mathbf{\Gamma}}^{\perp}\mathbf{V}_{\mathcal{S}_k}\right) = \mathrm{tr}\left(\mathbf{C}_k^{\top}\mathcal{P}_{\mathbf{\Gamma}}\mathbf{V}_{\mathcal{S}_k}\right) \\
&\leq \|\mathcal{P}_{\mathbf{\Gamma}}\mathbf{C}_k\|_F\|\mathcal{P}_{\mathbf{\Gamma}}\mathbf{V}_{\mathcal{S}_k}\|_F \\
&\leq \delta\|\mathbf{C}_k\|_F\|\mathbf{V}_{\mathcal{S}_k}\|_F \\
&\leq \frac{\delta}{\sqrt{1-\delta}}\|\mathcal{P}_{\mathbf{\Gamma}}^{\perp}\mathbf{C}_k\|_F\|\mathbf{V}_{\mathcal{S}_k}\|_F \\
&= \frac{\delta}{\sqrt{1-\delta}}\left\|\mathcal{P}_{\mathbf{P}}\mathcal{P}_{\mathbf{\Gamma}}^{\perp}\mathbf{V}_{\mathcal{S}_k}\right\|_F\|\mathbf{V}_{\mathcal{S}_k}\|_F,
\end{aligned}$$

which implies that

$$\left\|\mathcal{P}_{\mathbf{P}}\mathcal{P}_{\mathbf{\Gamma}}^{\perp}\mathbf{V}_{\mathcal{S}_k}\right\|_F \leq \frac{\delta}{\sqrt{1-\delta}}\|\mathbf{V}_{\mathcal{S}_k}\|_F. \tag{B.9}$$

Note that $\mathbf{V}_{\mathcal{S}_k}$ also satisfies the inequality in Eq.(B.5) for $k \geq 2$, thus

$$\left\|\sum_{k\geq 2}\mathcal{P}_{\mathbf{P}}\mathcal{P}_{\mathbf{\Gamma}}^{\perp}\mathbf{V}_{\mathcal{S}_k}\right\|_F \leq \frac{\delta}{\sqrt{(1-\delta)s}}\sum_{k\geq 1}\|\mathbf{V}_{\mathcal{S}_k}\|_{1,2} = \frac{\delta}{\sqrt{(1-\delta)s}}\|\mathbf{V}_{\mathcal{S}_0^c}\|_{1,2}. \tag{B.10}$$

Combining Eq.(B.7), Eq.(B.8) and Eq.(B.10), the first conclusion will follow from

$$\sqrt{1-\delta}\|\mathbf{V}_{\mathcal{S}_{01}}\|_F - \frac{\delta}{\sqrt{(1-\delta)s}}\|\mathbf{V}_{\mathcal{S}_0^c}\|_{1,2} \leq \frac{1}{\sqrt{1-\delta}}\left\|(\mathcal{P}_{\mathbf{\Gamma}}^{\perp}\mathbf{V})_{\mathcal{S}_{01}}\right\|_F. \tag{B.11}$$

To prove the second conclusion, note that the k-th largest ℓ_2 norm row in $\mathbf{V}_{\mathcal{S}_0^c}$ is bounded by $\frac{\|\mathbf{V}_{\mathcal{S}_0^c}\|_{1,2}}{k}$. Therefore,

$$\|\mathbf{V}_{\mathcal{S}_{01}^c}\|_F^2 \leq \|\mathbf{V}_{\mathcal{S}_0^c}\|_{1,2}^2\sum_{k\geq s+1}\frac{1}{k^2} \leq \frac{1}{s}\|\mathbf{V}_{\mathcal{S}_0^c}\|_{1,2}^2.$$

The conclusion follows. ■

B.5. Proof of Proposition 5.7

(1) Let

$$\frac{1}{mn}\mathbf{Y}\mathcal{P}_{\mathbf{X}}^{\perp}\mathbf{Y}^{\top} = \mathbf{M}_1 + \mathbf{M}_2 + \mathbf{M}_3,$$

where

$$\mathbf{M}_1 = \frac{1}{mn}\mathbf{\Gamma}\mathbf{G}\mathcal{P}_{\mathbf{X}}^{\perp}\mathbf{G}^{\top}\mathbf{\Gamma}^{\top},$$

$$\mathbf{M}_2 = \frac{1}{mn}\Big(\mathbf{\Gamma}\mathbf{G}\mathcal{P}_{\mathbf{X}}^{\perp}(\mathbf{F}+\mathbf{E})^{\top} + (\mathbf{F}+\mathbf{E})\mathcal{P}_{\mathbf{X}}^{\perp}\mathbf{G}^{\top}\mathbf{\Gamma}^{\top}\Big),$$

$$\mathbf{M}_3 = \frac{1}{mn}(\mathbf{F}+\mathbf{E})\mathcal{P}_{\mathbf{X}}^{\perp}(\mathbf{F}+\mathbf{E})^{\top}. \tag{B.12}$$

Since $\mathbf{\Gamma}(\mathbf{\Gamma}^{\top}\mathbf{\Gamma})^{-1/2}$ is orthogonal and of rank K, and

$$\mathbf{M}_1 = \mathbf{\Gamma}(\mathbf{\Gamma}^{\top}\mathbf{\Gamma})^{-1/2}\Big\{\frac{1}{mn}(\mathbf{\Gamma}^{\top}\mathbf{\Gamma})^{1/2}\mathbf{G}\mathcal{P}_{\mathbf{X}}^{\perp}\mathbf{G}^{\top}(\mathbf{\Gamma}^{\top}\mathbf{\Gamma})^{1/2}\Big\}(\mathbf{\Gamma}^{\top}\mathbf{\Gamma})^{-1/2}\mathbf{\Gamma}^{\top},$$

by assumption 5.2 and Proposition A.6, we conclude that

$$\Omega_P(r_{m,n}) = \Lambda_K(\mathbf{M}_1) \leq \Lambda_{\max}(\mathbf{M}_1) = O_P(1) \tag{B.13}$$

and

$$\|\mathbf{M}_2\|_2 = O_P\Big(\frac{1}{\sqrt{n}}\Big),\ \|\mathbf{M}_3\|_2 = O_P\Big(\frac{1}{n}\Big). \tag{B.14}$$

Since $\hat{\mathbf{\Gamma}}$ is scaled from the top K eigenvectors of $\frac{1}{mn}\mathbf{Y}\mathcal{P}_{\mathbf{X}}^{\perp}\mathbf{Y}^{\top}$ and $\mathbf{\Gamma}(\mathbf{\Gamma}^{\top}\mathbf{\Gamma})^{-1/2}$ consists of the top K eigenvectors of $\mathbf{M}_1$, by Lemma A.2 and Lemma A.3,

$$\big\|\mathcal{P}_{\hat{\mathbf{\Gamma}}} - \mathcal{P}_{\mathbf{\Gamma}}\big\|_2 = O_P\Big(\frac{1}{\sqrt{nr_{m,n}^2}}\Big)$$

(2) To prove this, we can find an implicit $\mathbf{H}$ using (1) and Lemma A.3. For convenience of later proofs, we find an explicit expression for $\mathbf{H}$. We will use $\mathbf{U}$, $\mathbf{D}$, $\mathbf{V}$ to simplify $\mathbf{U}_{1:K}$,

$\mathbf{D}_{1:K}$, $\mathbf{V}_{1:K}$ in Eq.(4.1). Eq.(4.1) implies that

$$\mathbf{UD} = \mathbf{\Gamma G}\mathcal{P}_{\mathbf{X}}^{\perp}\mathbf{V} + (\mathbf{F}+\mathbf{E})\mathcal{P}_{\mathbf{X}}^{\perp}\mathbf{V}. \tag{B.15}$$

Then,

$$\left\|\mathbf{UD} - \mathbf{\Gamma G}\mathcal{P}_{\mathbf{X}}^{\perp}\mathbf{V}\right\|_2 = O_P(\sqrt{m}), \tag{B.16}$$

which is equivalent to

$$\left\|\hat{\mathbf{\Gamma}} - \frac{\mathbf{\Gamma G}\mathcal{P}_{\mathbf{X}}^{\perp}\mathbf{V}}{\sqrt{n}}\right\|_2 = O_P\Big(\sqrt{\frac{m}{n}}\Big).$$

Let $\mathbf{H} = \frac{\mathbf{G}\mathcal{P}_{\mathbf{X}}^{\perp}\mathbf{V}}{\sqrt{n}}$, then

$$\left\|\hat{\mathbf{\Gamma}} - \mathbf{\Gamma H}\right\|_2 = O_P\Big(\sqrt{\frac{m}{n}}\Big).$$

It is clear that $\|\mathbf{H}\|_2 = O_P(1)$. It remains to prove that $\mathbf{H}$ is invertible with high probability and $\|\mathbf{H}^{-1}\|_2 = O_P(1)$. From Eq.(B.13) and Eq.(B.14),

$$\Omega_P(r_{m,n}) = \Lambda_{\min}\Big(\frac{\mathbf{D}^2}{mn}\Big) \le \Lambda_{\max}\Big(\frac{\mathbf{D}^2}{mn}\Big) = O_P(1).$$

Left multiplying $\mathbf{D}^{-1}\mathbf{U}^\top$ to the matrix in Eq.(B.16):

$$\left\|\mathbf{I}_K - \sqrt{n}\mathbf{D}^{-1}\mathbf{U}^\top\mathbf{\Gamma H}\right\|_2 = O_P\Big(\sqrt{m}\|\mathbf{D}^{-1}\|_2\Big) = O_P\Big(\frac{1}{\sqrt{nr_{m,n}}}\Big). \tag{B.17}$$

Since $nr_{m,n} \to \infty$, we will finish the proof by showing that

$$\Omega_P(1) \le \Lambda_{\min}\big(n\mathbf{D}^{-1}\mathbf{U}^\top\mathbf{\Gamma\Gamma}^\top\mathbf{UD}^{-1}\big) \le \Lambda_{\max}\big(n\mathbf{D}^{-1}\mathbf{U}^\top\mathbf{\Gamma\Gamma}^\top\mathbf{UD}^{-1}\big) \le O_P(1). \tag{B.18}$$

Using Eq.(B.12) and the fact that $\mathbf{U}$ consists of eigenvectors of $\frac{1}{mn}\mathbf{Y}\mathcal{P}_{\mathbf{X}}^{\perp}\mathbf{Y}^\top$,

$$\frac{\mathbf{D}^2}{mn} = \mathbf{U}^\top\mathbf{\Gamma}\frac{\mathbf{G}\mathcal{P}_{\mathbf{X}}^{\perp}\mathbf{G}^\top}{n}\frac{\mathbf{\Gamma}^\top\mathbf{U}}{m} + \mathbf{U}^\top\mathbf{M}_2\mathbf{U} + \mathbf{U}^\top\mathbf{M}_3\mathbf{U}. \tag{B.19}$$

Eq.(B.14) together with Eq.(B.19) imply that

$$\Big\|\mathbf{I}_K - \mathbf{D}^{-1}\mathbf{U}^\top\mathbf{\Gamma}\mathbf{G}\mathcal{P}_{\mathbf{X}}^{\perp}\mathbf{G}^\top\mathbf{\Gamma}^\top\mathbf{U}\mathbf{D}^{-1}\Big\|_2 = O_P\Big(m\sqrt{n}\|\mathbf{D}^{-2}\|_2\Big) = O_P\Big(\frac{1}{\sqrt{nr_{m,n}^2}}\Big). \tag{B.20}$$

Since $nr_{m,n}^2 \to \infty$, $\mathbf{D}^{-1}\mathbf{U}^\top\mathbf{\Gamma}\mathbf{G}\mathcal{P}_{\mathbf{X}}^{\perp}\mathbf{G}^\top\mathbf{\Gamma}^\top\mathbf{U}\mathbf{D}^{-1}$ converges in probability to $\mathbf{I}_K$ with respect to the spectral norm. Also, $\mathbf{D}^{-1}\mathbf{U}^\top\mathbf{\Gamma}$ must be of full rank with high probability. Eq.(B.20) implies that

$$\Lambda_{\min}\big(\mathbf{G}\mathcal{P}_{\mathbf{X}}^{\perp}\mathbf{G}^\top\big)\Lambda_{\max}\big(\mathbf{D}^{-1}\mathbf{U}^\top\mathbf{\Gamma}\mathbf{\Gamma}^\top\mathbf{U}\mathbf{D}^{-1}\big) \le 1 + O_P\Big(\frac{1}{\sqrt{nr_{m,n}}}\Big),$$

and

$$\Lambda_{\max}\big(\mathbf{G}\mathcal{P}_{\mathbf{X}}^{\perp}\mathbf{G}^\top\big)\Lambda_{\min}\big(\mathbf{D}^{-1}\mathbf{U}^\top\mathbf{\Gamma}\mathbf{\Gamma}^\top\mathbf{U}\mathbf{D}^{-1}\big) \ge 1 - O_P\Big(\frac{1}{\sqrt{nr_{m,n}}}\Big).$$

Note that $\Omega_P(n) = \Lambda_{\min}\big(\mathbf{G}\mathcal{P}_{\mathbf{X}}^{\perp}\mathbf{G}^\top\big) \le \Lambda_{\max}\big(\mathbf{G}\mathcal{P}_{\mathbf{X}}^{\perp}\mathbf{G}^\top\big) = O_P(n)$, Eq.(B.18) follows from the above two inequalities.

(3) From Eq.(B.15),

$$\sqrt{n}\big(\hat{\boldsymbol{\gamma}}_{i,\cdot} - \mathbf{H}\boldsymbol{\gamma}_{i,\cdot}\big) = \mathbf{V}^\top\mathcal{P}_{\mathbf{X}}^{\perp}(\boldsymbol{f}_{i,\cdot} + \boldsymbol{\epsilon}_{i,\cdot}). \tag{B.21}$$

Similar to Eq.(B.16), we can show that

$$\big\|\mathbf{V}^\top - \mathbf{D}^{-1}\mathbf{U}^\top\mathbf{\Gamma}\mathbf{G}\mathcal{P}_{\mathbf{X}}^{\perp}\big\|_2 = \big\|\mathbf{D}^{-1}\mathbf{U}^\top(\mathbf{F}+\mathbf{E})\mathcal{P}_{\mathbf{X}}^{\perp}\big\|_2 = O_P\Big(\frac{1}{\sqrt{nr_{m,n}}}\Big).$$

From Proposition A.6, $\max_{1\le i\le m}\|\boldsymbol{f}_{i,\cdot} + \boldsymbol{\epsilon}_{i,\cdot}\|_2 = O_P(\sqrt{n})$. Together with Eq.(B.18), we have:

$$\begin{aligned}\sqrt{n}\max_{1\le i\le m}\big\|\hat{\boldsymbol{\gamma}}_{i,\cdot} - \mathbf{H}\boldsymbol{\gamma}_{i,\cdot}\big\| &= \max_{1\le i\le m}\big\|\mathbf{D}^{-1}\mathbf{U}^\top\mathbf{\Gamma}\mathbf{G}\mathcal{P}_{\mathbf{X}}^{\perp}(\boldsymbol{f}_{i,\cdot} + \boldsymbol{\epsilon}_{i,\cdot})\big\| + O_P\Big(\frac{1}{\sqrt{r_{m,n}}}\Big)\\ &\le \max_{1\le i\le m}\Big\|\frac{\mathbf{G}\mathcal{P}_{\mathbf{X}}^{\perp}}{\sqrt{n}}(\boldsymbol{f}_{i,\cdot} + \boldsymbol{\epsilon}_{i,\cdot})\Big\| + O_P\Big(\frac{1}{\sqrt{r_{m,n}}}\Big).\end{aligned} \tag{B.22}$$

Conditional on $\mathbf{X}$ and $\mathbf{G}$, $\mathbf{G}\mathcal{P}_{\mathbf{X}}^{\perp}\boldsymbol{\epsilon}_{i,\cdot}$ follows $N_K\big(0, \sigma_i^2\mathbf{G}\mathcal{P}_{\mathbf{X}}^{\perp}\mathbf{G}^\top\big)$. Then,

$$\big\|\big(\mathbf{G}\mathcal{P}_{\mathbf{X}}^{\perp}\mathbf{G}^\top\big)^{-1/2}\mathbf{G}\mathcal{P}_{\mathbf{X}}^{\perp}\boldsymbol{\epsilon}_{i,\cdot}\big\|^2 \sim \sigma_i^2\chi_K^2.$$

Apply Lemma A.1,

$$\max_{1\le i\le m}\big\|\mathbf{G}\mathcal{P}_{\mathbf{X}}^{\perp}\boldsymbol{\epsilon}_{i,\cdot}\big\|^2 \le \Lambda_{\max}\big(\mathbf{G}\mathcal{P}_{\mathbf{X}}^{\perp}\mathbf{G}^\top\big)\max_{1\le i\le m}\sigma_i^2\cdot(K+\sqrt{2K\log m}+2\log m) = O_P\big(n(K\vee\log m)\big).$$

Then,

$$\max_{1\le i\le m}\big\|\mathbf{G}\mathcal{P}_{\mathbf{X}}^{\perp}\boldsymbol{\epsilon}_{i,\cdot}\big\|_2 = O_P(\sqrt{n(K\vee\log m)}). \tag{B.23}$$

Consider

$$\mathbf{G}\mathcal{P}_{\mathbf{X}}^{\perp}\boldsymbol{f}_{i,\cdot} = \mathbf{G}\boldsymbol{f}_{i,\cdot} - \mathbf{G}\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\boldsymbol{f}_{i,\cdot}. \tag{B.24}$$

For the first term, $\mathbf{G}\boldsymbol{f}_{i,\cdot} = \sum_{j=1}^n f_{i,j}\boldsymbol{g}_{\cdot,j}$ is a sum of n i.i.d. random vectors where each summand is bounded by $\sqrt{\frac{Ma_{K+1}}{\lambda_K}}$, see Eq.(2.13) and the comments below assumption 5.3. Let

$$\mathbf{M}_{i,j} = \begin{pmatrix} 0 & f_{i,j}\boldsymbol{g}_{\cdot,j}^\top \\ f_{i,j}\boldsymbol{g}_{\cdot,j} & \mathbf{0}\end{pmatrix} \tag{B.25}$$

be a $(K+1)\times(K+1)$ symmetric matrix, $j=1,\dots,n$. Note that $\{\mathbf{M}_{i,j}\}_{j=1}^n$ are mean-zero, i.i.d. random matrices, and $\|\mathbf{M}_{i,j}\|_2 \le \|f_{i,j}\boldsymbol{g}_{\cdot,j}\|_2 \le \sqrt{\frac{Ma_{K+1}}{\lambda_K}}$, $\big\|\mathbb{E}\mathbf{M}_{i,j}^2\big\|_2 \le \frac{Ma_{K+1}}{\lambda_K}$. Apply Proposition A.6 to $\{\mathbf{M}_{i,j}\}_{j=1}^n$,

$$\mathbb{P}\Big(\Big\|\sum_{j=1}^n\mathbf{M}_{i,j}\Big\|_2 \ge t\Big) \le 2(K+1)\exp\Big(-\frac{\lambda_K t^2/2}{nMa_{K+1}+t\sqrt{M\lambda_K a_{K+1}}/3}\Big).$$

By union bound,

$$\begin{aligned}\mathbb{P}\Big(\max_{1\le i\le m}\Big\|\sum_{j=1}^n\mathbf{M}_{i,j}\Big\|_2 \ge t\Big) &\le 2m(K+1)\exp\Big(-\frac{\lambda_K t^2/2}{nMa_{K+1}+t\sqrt{M\lambda_K a_{K+1}}/3}\Big)\\ &\lesssim \exp\Big(2\log m - \frac{\lambda_K t^2/2}{nMa_{K+1}+t\sqrt{M\lambda_K a_{K+1}}/3}\Big).\end{aligned} \tag{B.26}$$

From Eq.(B.26), there exist general constants $c_1, c_2 > 0$ such that if $t \geq c_1\sqrt{\frac{nMa_{K+1}}{\lambda_K}}$,

$$\max_{1\leq i\leq m} \Big\| \sum_{j=1}^{n} \mathbf{M}_{i,j} \Big\|_2 \leq t$$

with probability higher than $1 - m^{-c_2}$. The above inequality implies that

$$\max_{1\leq i\leq m} \big\|\mathbf{G}\boldsymbol{f}_{i,\boldsymbol{\cdot}}\big\|_2 = O_P\Big(\sqrt{\frac{nMa_{K+1}}{\lambda_K}}\Big). \tag{B.27}$$

For the second term in Eq.(B.24),

$$\max_{1\leq i\leq m} \big\|\mathbf{G}\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\boldsymbol{f}_{i,\boldsymbol{\cdot}}\big\|_2 \leq O_P(n)\cdot O_P\Big(\frac{1}{n}\Big)\cdot \max_{1\leq i\leq m} \big\|\mathbf{X}\boldsymbol{f}_{i,\boldsymbol{\cdot}}\big\|_2 = O_P\Big(\max_{1\leq i\leq m} \big\|\mathbf{X}\boldsymbol{f}_{i,\boldsymbol{\cdot}}\big\|_2\Big).$$

To bound $\max_{1\leq i\leq m} \big\|\mathbf{X}\boldsymbol{f}_{i,\boldsymbol{\cdot}}\big\|_2$, we apply the General Hoeffding's inequality (Theorem 2.6.2 in Vershynin (2018)) coordinate-wisely:

$$\mathbb{P}\Big(\Big|\sum_{j=1}^{n} f_{i,j}X_{l,j}\Big|_2 \geq t\Big) \leq 2\exp\Big(-\frac{c_3t^2}{nMa_{K+1}}\Big),$$

where $c_3 > 0$ is a general constant. Also,

$$\begin{aligned}\mathbb{P}\Big(\max_{1\leq i\leq m}\max_{1\leq l\leq p}\Big|\sum_{j=1}^{n} f_{i,j}X_{l,j}\Big|_2 \geq t\Big) &\leq 2mp\cdot\exp\Big(-\frac{c_3t^2}{nMa_{K+1}}\Big)\\ &\lesssim \exp\Big(\log m - \frac{c_3t^2}{nMa_{K+1}}\Big).\end{aligned} \tag{B.28}$$

Similar to Eq.(B.27), Eq.(B.28) implies that

$$\max_{1\leq i\leq m} \big\|\mathbf{X}\boldsymbol{f}_{i,\boldsymbol{\cdot}}\big\|_2 = \sqrt{p}\max_{1\leq i\leq m}\max_{1\leq l\leq p}\Big|\sum_{j=1}^{n} f_{i,j}X_{l,j}\Big|_2 = O_P\Big(\sqrt{npMa_{K+1}\log m}\Big) = O_P(\sqrt{np}). \tag{B.29}$$

Combining Eq.(B.22), Eq.(B.23), Eq.(B.27), Eq.(B.29) we conclude that

$$\max_{1\leq i\leq m} \big\|\hat{\boldsymbol{\gamma}}_{i,\boldsymbol{\cdot}} - \mathbf{H}\boldsymbol{\gamma}_{i,\boldsymbol{\cdot}}\big\| = O_P\Big(\sqrt{\frac{K\vee\log m + p}{n}} + \sqrt{\frac{Ma_{K+1}}{n\lambda_K}} + \frac{1}{\sqrt{nr_{m,n}}}\Big).$$

■

B.6. Proof of Theorem 5.8

(1) The loss function in Eq.(4.2) can be written as:

$$L(\mathbf{\Theta}) := \left\|\mathcal{P}^{\perp}_{\hat{\mathbf{\Gamma}}}(\mathbf{Y} - \mathbf{\Theta}\mathbf{X}_1)\mathcal{P}^{\perp}_{\mathbf{X}_2}\right\|_F^2 + \eta \cdot \sum_{i=1}^{m} \|\boldsymbol{\theta}_{i,\bullet}\|_2. \tag{B.30}$$

Because $L(\hat{\mathbf{B}}_1) \leq L(\mathbf{B}_1^*)$ for any $m \times c$ matrix $\mathbf{B}_1^* = (\boldsymbol{b}_{1,\bullet}^{(1*)}, \ldots, \boldsymbol{b}_{m,\bullet}^{(1*)})^\top$, by rearranging terms in Eq.(B.30) we deduce that

$$\mathrm{tr}\Big(\widetilde{\mathbf{X}}_1^\top \hat{\mathbf{V}}^\top \mathcal{P}^{\perp}_{\hat{\mathbf{\Gamma}}} \hat{\mathbf{V}} \widetilde{\mathbf{X}}_1\Big) \leq 2 \cdot \mathrm{tr}\Big(\hat{\mathbf{V}}\widetilde{\mathbf{X}}_1(\widetilde{\mathbf{Y}} - \mathbf{B}_1^*\widetilde{\mathbf{X}}_1)^\top \mathcal{P}^{\perp}_{\hat{\mathbf{\Gamma}}}\Big) + \eta \cdot \sum_{i=1}^{m} \Big(\|\boldsymbol{b}_{i,\bullet}^{(1*)}\|_2 - \|\hat{\boldsymbol{b}}_{i,\bullet}^{(1)}\|_2\Big) \tag{B.31}$$

where $\widetilde{\mathbf{X}}_1 = \mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2}$, $\widetilde{\mathbf{Y}} = \mathbf{Y}\mathcal{P}^{\perp}_{\mathbf{X}_2}$ and $\hat{\mathbf{V}} = \hat{\mathbf{V}}(\mathbf{B}^*) := \hat{\mathbf{B}}_1 - \mathbf{B}_1^*$.

The first term on the right-hand side of Eq.(B.31) is the sum of m inner products of rows of $\hat{\mathbf{V}}$ and columns of $\widetilde{\mathbf{X}}_1(\widetilde{\mathbf{Y}} - \mathbf{B}_1^*\widetilde{\mathbf{X}}_1)^\top \mathcal{P}^{\perp}_{\hat{\mathbf{\Gamma}}}$, where

$$\begin{aligned}\widetilde{\mathbf{X}}_1(\widetilde{\mathbf{Y}} - \mathbf{B}_1^*\widetilde{\mathbf{X}}_1)^\top \mathcal{P}^{\perp}_{\hat{\mathbf{\Gamma}}} &= \Big(\widetilde{\mathbf{X}}_1\widetilde{\mathbf{X}}_1^\top(\mathbf{B}_1 - \mathbf{B}_1^*)^\top + \widetilde{\mathbf{X}}_1\mathbf{G}^\top\mathbf{\Gamma}^\top + \widetilde{\mathbf{X}}_1(\mathbf{E} + \mathbf{F})^\top\Big)\mathcal{P}^{\perp}_{\hat{\mathbf{\Gamma}}} \\ &:= \mathbf{M}_1 + \mathbf{M}_2 + \mathbf{M}_3. \end{aligned} \tag{B.32}$$

Firstly, consider $\mathbf{M}_2$ and $\mathbf{M}_3$, and we will postpone the discussion of $\mathbf{M}_1$ for different choices of $\mathbf{B}_1^*$.

For $\mathbf{M}_2$,

$$\mathbf{M}_2 = \widetilde{\mathbf{X}}_1\mathbf{G}^\top(\mathbf{\Gamma} - \hat{\mathbf{\Gamma}}\mathbf{H}^{-1})^\top \mathcal{P}^{\perp}_{\hat{\mathbf{\Gamma}}} \tag{B.33}$$

Let $\mathbf{\Xi} = (\mathbf{\Xi}_{j,i})_{1\leq j\leq m, 1\leq i\leq m} = \mathcal{P}^{\perp}_{\hat{\mathbf{\Gamma}}}$. Consider the i-th column of $\mathbf{M}_2$,

$$\begin{aligned}\|(\mathbf{M}_2)_{\bullet,i}\|_2 &\leq \|\widetilde{\mathbf{X}}_1\mathbf{G}^\top\|_2\|(\mathbf{\Gamma} - \hat{\mathbf{\Gamma}}\mathbf{H}^{-1})^\top\mathbf{\Xi}_{\bullet,i}\|_2 \\ &= \|\widetilde{\mathbf{X}}_1\mathbf{G}^\top\|_2\Big\|\sum_{j=1}^{m}(\boldsymbol{\gamma}_{j,\bullet} - \mathbf{H}^{-1}\hat{\boldsymbol{\gamma}}_{j,\bullet})\mathbf{\Xi}_{j,i}\Big\|_2\end{aligned}$$

$$\leq \|\widetilde{\mathbf{X}}_1\mathbf{G}^\top\|_2\|\Big(\|\boldsymbol{\gamma}_{i,\cdot} - \mathbf{H}^{-1}\hat{\boldsymbol{\gamma}}_{i,\cdot}\|_2 + \Big\|\sum_{j\neq i}(\boldsymbol{\gamma}_{j,\cdot} - \mathbf{H}^{-1}\hat{\boldsymbol{\gamma}}_{j,\cdot})\boldsymbol{\Xi}_{j,i}\Big\|_2\Big)$$

$$\leq \|\widetilde{\mathbf{X}}_1\mathbf{G}^\top\|_2\Big(1 + \sqrt{m(1-\boldsymbol{\Xi}_{i,i})}\Big)\cdot \max_{1\leq j\leq m}\|\boldsymbol{\gamma}_{j,\cdot} - \mathbf{H}^{-1}\hat{\boldsymbol{\gamma}}_{j,\cdot}\|_2, \tag{B.34}$$

where the lase line uses Cauchy-Schawartz inequality and the fact that $\boldsymbol{\Xi}_{ii} \leq 1$, $\sum_{j\neq i}\boldsymbol{\Xi}_{ji}^2 = \boldsymbol{\Xi}_{ii}(1-\boldsymbol{\Xi}_{ii})$. Eq.(B.34) implies that

$$\mathrm{tr}(\hat{\mathbf{V}}\mathbf{M}_2) \leq a_{m,n}\|\widetilde{\mathbf{X}}_1\mathbf{G}^\top\|_2\sum_{i=1}^m\|\hat{\mathbf{V}}_{i,\cdot}\|_2, \tag{B.35}$$

where

$$a_{m,n} = \Big(1 + \max_{1\leq i\leq m}\sqrt{m(1-\boldsymbol{\Xi}_{i,i})}\Big)\cdot\max_{1\leq i\leq m}\|\boldsymbol{\gamma}_{i,\cdot} - \mathbf{H}^{-1}\hat{\boldsymbol{\gamma}}_{i,\cdot}\|_2.$$

For $\mathbf{M}_3$,

$$\|(\mathbf{M}_3)_{\cdot,i}\|_2 = \Big\|\sum_{j=1}^m\widetilde{\mathbf{X}}_1(\boldsymbol{\epsilon}_{j,\cdot} + \boldsymbol{f}_{j,\cdot})\boldsymbol{\Xi}_{j,i}\Big\|_2$$

$$\leq \big\|\widetilde{\mathbf{X}}_1(\boldsymbol{\epsilon}_{i,\cdot} + \boldsymbol{f}_{i,\cdot})\big\|_2 + \max_{j\neq i}\big\|\widetilde{\mathbf{X}}_1(\boldsymbol{\epsilon}_{j,\cdot} + \boldsymbol{f}_{j,\cdot})\big\|_2\sum_{j\neq i}|\boldsymbol{\Xi}_{j,i}|$$

$$\leq \big\|\widetilde{\mathbf{X}}_1(\boldsymbol{\epsilon}_{i,\cdot} + \boldsymbol{f}_{i,\cdot})\big\|_2 + \max_{j\neq i}\big\|\widetilde{\mathbf{X}}_1(\boldsymbol{\epsilon}_{j,\cdot} + \boldsymbol{f}_{j,\cdot})\big\|_2\sqrt{(m-1)\boldsymbol{\Xi}_{i,i}(1-\boldsymbol{\Xi}_{i,i})}$$

$$\leq \Big(1 + \sqrt{m(1-\boldsymbol{\Xi}_{i,i})}\Big)\cdot\max_{1\leq j\leq m}\big\|\widetilde{\mathbf{X}}_1(\boldsymbol{\epsilon}_{j,\cdot} + \boldsymbol{f}_{j,\cdot})\big\|_2. \tag{B.36}$$

Eq.(B.36) implies that

$$\mathrm{tr}(\hat{\mathbf{V}}\mathbf{M}_3) \leq b_{m,n}\sum_{i=1}^m\|\hat{\mathbf{V}}_{i,\cdot}\|_2, \tag{B.37}$$

where

$$b_{m,n} = \Big(1 + \max_{1\leq i\leq m}\sqrt{m(1-\boldsymbol{\Xi}_{i,i})}\Big)\cdot\max_{1\leq i\leq m}\big\|\widetilde{\mathbf{X}}_1(\boldsymbol{\epsilon}_{i,\cdot} + \boldsymbol{f}_{i,\cdot})\big\|_2.$$

We need a more precise order for each term in $a_{m,n}$ and $b_{m,n}$. For $a_{m,n}$, $\|\widetilde{\mathbf{X}}_1\mathbf{G}^\top\|_2 = O_P(n)$, and

$$\max_{1\leq i\leq m}1-\boldsymbol{\Xi}_{i,i} = \max_{1\leq i\leq m}\hat{\boldsymbol{\gamma}}_{i,\cdot}^\top(\hat{\boldsymbol{\Gamma}}^\top\hat{\boldsymbol{\Gamma}})^{-1}\hat{\boldsymbol{\gamma}}_{i,\cdot} = O_P\Big(\frac{\max_{1\leq i\leq m}\|\boldsymbol{\gamma}_{i,\cdot}\|_2^2}{mr_{m,n}}\Big), \tag{B.38}$$

where we use Proposition 5.7(3) and the fact that $\hat{\mathbf{\Gamma}}^\top \hat{\mathbf{\Gamma}} = \frac{\mathbf{D}^2}{n}$. Similar to the proof of Eq.(B.24) and Eq.(B.29), we can show that

$$\max_{1\le i\le m} \left\|\mathbf{X}_1 \mathcal{P}^{\perp}_{\mathbf{X}_2} \boldsymbol{\epsilon}_{i,\cdot}\right\|_2 = O_P(\sqrt{n}), \quad \max_{1\le i\le m} \left\|\mathbf{X}_1 \mathcal{P}^{\perp}_{\mathbf{X}_2} \boldsymbol{f}_{i,\cdot}\right\|_2 = O_P(\sqrt{np}). \tag{B.39}$$

Therefore,

$$c_{m,n} = O_P\Big(\Big(n \max_{1\le i\le m} \|\boldsymbol{\gamma}_{i,\cdot} - \mathbf{H}^{-1}\hat{\boldsymbol{\gamma}}_{i,\cdot}\|_2 + \sqrt{np}\Big) \frac{\max_{1\le i\le m} \|\boldsymbol{\gamma}_{i,\cdot}\|_2}{\sqrt{r_{m,n}}}\Big), \tag{B.40}$$

where $c_{m,n} = 2\big(a_{m,n}\|\widetilde{\mathbf{X}}_1 \mathbf{G}^\top\|_2 + b_{m,n}\big)$.

From Proposition 5.7(3), we can choose a constant $C_\delta > 0$ such that

$$2c_{m,n} \le C_\delta \sqrt{n}\Big(\sqrt{K \vee \log m + p} + \sqrt{\frac{nMa_{K+1}}{\lambda_K}} + \frac{1}{\sqrt{r_{m,n}}}\Big) \frac{\max_{1\le i\le m} \|\boldsymbol{\gamma}_{i,\cdot}\|_2}{\sqrt{r_{m,n}}} \le \eta \tag{B.41}$$

with probability greater than $1-\delta$ for sufficiently large m, n. The derivations below will be conditioned on the event when Eq.(B.41) holds.

(2) Firstly, we establish an upper bound for $\left\|\hat{\mathbf{B}}_1 - \mathbf{B}_1\right\|_F$. Return to Eq.(B.31) with the choice of $\mathbf{B}_1^* = \mathbf{B}_1$. In this scenario, $\mathbf{M}_1 = \mathbf{0}$, then

$$\begin{aligned} 0 \le \mathrm{tr}\Big(\widetilde{\mathbf{X}}_1^\top \hat{\mathbf{V}}^\top \mathcal{P}^{\perp}_{\hat{\mathbf{\Gamma}}^\top} \hat{\mathbf{V}} \widetilde{\mathbf{X}}_1\Big) &\le c_{m,n} \sum_{i=1}^m \|\hat{\mathbf{V}}_{i,\cdot}\|_2 + \eta \cdot \sum_{i=1}^m \Big(\|\boldsymbol{b}_{i,\cdot}^{(1*)}\|_2 - \|\hat{\boldsymbol{b}}_{i,\cdot}^{(1)}\|_2\Big) \\ &\le c_{m,n} \sum_{i=1}^m \|\hat{\mathbf{V}}_{i,\cdot}\|_2 + \eta \cdot \|\hat{\mathbf{V}}_{\mathcal{I}_1}\|_{1,2} - \eta \cdot \|\hat{\mathbf{V}}_{\mathcal{I}_0}\|_{1,2}. \end{aligned} \tag{B.42}$$

From Eq.(B.42) and Eq.(B.41),

$$\|\hat{\mathbf{V}}_{\mathcal{I}_0}\|_{1,2} \le 3\|\hat{\mathbf{V}}_{\mathcal{I}_1}\|_{1,2}. \tag{B.43}$$

Taking derivative w.r.t. $\mathbf{\Theta}$ for $L(\mathbf{\Theta})$ in Eq.(B.30), we know

$$\mathcal{P}_{\hat{\mathbf{\Gamma}}}^{\perp}\hat{\mathbf{B}}_1\widetilde{\mathbf{X}}_1\widetilde{\mathbf{X}}_1^{\top} - \mathcal{P}_{\hat{\mathbf{\Gamma}}}^{\perp}\widetilde{\mathbf{Y}}\widetilde{\mathbf{X}}_1^{\top} = -\frac{\eta}{2}\begin{pmatrix} \mathrm{sgn}(\hat{\boldsymbol{b}}_{1,\boldsymbol{\cdot}}^{(1)})^{\top} \\ \vdots \\ \mathrm{sgn}(\hat{\boldsymbol{b}}_{m,\boldsymbol{\cdot}}^{(1)})^{\top} \end{pmatrix}, \tag{B.44}$$

where $\mathrm{sgn}(\cdot)$ denotes the subgradient of ℓ_2 norm. From Eq.(B.44),

$$\big\|\mathcal{P}_{\hat{\mathbf{\Gamma}}}^{\perp}\hat{\mathbf{V}}\widetilde{\mathbf{X}}_1\widetilde{\mathbf{X}}_1^{\top} - \mathcal{P}_{\hat{\mathbf{\Gamma}}}^{\perp}\mathbf{\Gamma}\mathbf{G}\widetilde{\mathbf{X}}_1\widetilde{\mathbf{X}}_1^{\top} - \mathcal{P}_{\hat{\mathbf{\Gamma}}}^{\perp}(\mathbf{F}+\mathbf{E})\widetilde{\mathbf{X}}_1^{\top}\big\|_{\infty,2} \leq \frac{\eta}{2}. \tag{B.45}$$

Note that

$$\big\|\mathcal{P}_{\hat{\mathbf{\Gamma}}}^{\perp}\mathbf{\Gamma}\mathbf{G}\widetilde{\mathbf{X}}_1\widetilde{\mathbf{X}}_1^{\top} + \mathcal{P}_{\hat{\mathbf{\Gamma}}}^{\perp}(\mathbf{F}+\mathbf{E})\widetilde{\mathbf{X}}_1^{\top}\big\|_{\infty,2} = \big\|\mathbf{M}_2^{\top} + \mathbf{M}_3^{\top}\big\|_{\infty,2} \leq \frac{c_{m,n}}{2},$$

we have

$$\big\|\mathcal{P}_{\hat{\mathbf{\Gamma}}}^{\perp}\hat{\mathbf{V}}\widetilde{\mathbf{X}}_1\widetilde{\mathbf{X}}_1^{\top}\big\|_{\infty,2} \leq \eta,$$

and thus

$$\big\|\mathcal{P}_{\hat{\mathbf{\Gamma}}}^{\perp}\hat{\mathbf{V}}\big\|_{\infty,2} \leq \frac{\eta}{\Lambda_{\min}(\widetilde{\mathbf{X}}_1\widetilde{\mathbf{X}}_1^{\top})}. \tag{B.46}$$

We are now ready to apply Lemma 3.3 to $\hat{\mathbf{V}}$. Let $\mathcal{S}_0 = \mathcal{I}_1$, $s = |\mathcal{I}_1|$. $\mathcal{S}_1$, $\mathcal{S}_{01}$ are defined in the same way as in Lemma 3.3. Then,

$$\begin{aligned}
\|\hat{\mathbf{V}}_{\mathcal{S}_{01}}\|_F &\leq \frac{1}{1-\delta}\big\|(\mathcal{P}_{\hat{\mathbf{\Gamma}}}^{\perp}\hat{\mathbf{V}})_{\mathcal{S}_{01}}\big\|_F + \frac{\delta}{(1-\delta)\sqrt{s}}\|\hat{\mathbf{V}}_{\mathcal{I}_0}\|_{1,2} \\
&\leq \frac{\sqrt{2s}}{1-\delta}\frac{\eta}{\Lambda_{\min}(\widetilde{\mathbf{X}}_1\widetilde{\mathbf{X}}_1^{\top})} + \frac{3\delta}{(1-\delta)\sqrt{s}}\|\hat{\mathbf{V}}_{\mathcal{I}_1}\|_{1,2} \\
&\leq \frac{\sqrt{2s}}{1-\delta}\frac{\eta}{\Lambda_{\min}(\widetilde{\mathbf{X}}_1\widetilde{\mathbf{X}}_1^{\top})} + \frac{3\delta}{(1-\delta)}\|\hat{\mathbf{V}}_{\mathcal{I}_1}\|_F \\
&\leq \frac{\sqrt{2s}}{1-\delta}\frac{\eta}{\Lambda_{\min}(\widetilde{\mathbf{X}}_1\widetilde{\mathbf{X}}_1^{\top})} + \frac{3\delta}{(1-\delta)}\|\hat{\mathbf{V}}_{\mathcal{S}_{01}}\|_F,
\end{aligned}$$

which implies that

$$\|\hat{\mathbf{V}}_{\mathcal{S}_{01}}\|_F \leq \frac{\sqrt{2s}}{1-4\delta}\frac{\eta}{\Lambda_{\min}(\widetilde{\mathbf{X}}_1\widetilde{\mathbf{X}}_1^\top)}. \tag{B.47}$$

The second conclusion in Lemma 3.3 together with Eq.(B.43) and Eq.(B.47) imply that

$$\|\hat{\mathbf{B}}_1 - \mathbf{B}_1\|_F = \|\hat{\mathbf{V}}\|_F \leq \sqrt{10}\|\hat{\mathbf{V}}_{\mathcal{S}_{01}}\|_F \leq \frac{\sqrt{20s}}{1-4\delta}\frac{\eta}{\Lambda_{\min}(\widetilde{\mathbf{X}}_1\widetilde{\mathbf{X}}_1^\top)}. \tag{B.48}$$

with probability greater than $1-\delta$ for sufficiently large m, n.

Again, consider Eq.(B.31). For each $i = 1, \ldots, m$, we now choose $\mathbf{B}_1^*$ such that $\boldsymbol{b}_{j,\cdot}^{(1*)}$ equals to $\hat{\boldsymbol{b}}_{j,\cdot}^{(1*)}$ for all $j \neq i$, and $\boldsymbol{b}_{i,\cdot}^{(1*)} = \boldsymbol{b}_{i,\cdot}^{(1)}$. Now $\mathbf{M}_1$ is not a zero matrix, but it suffices to consider the i-th column of $\mathbf{M}_1$ since $\hat{\mathbf{V}}_{j,\cdot}$ is zero except for $j = i$. Note that

$$\mathbf{M}_1 = \widetilde{\mathbf{X}}_1\widetilde{\mathbf{X}}_1^\top(\mathbf{B}_1 - \mathbf{B}_1^*)^\top\mathcal{P}_{\hat{\mathbf{\Gamma}}}^\perp \tag{B.49}$$

Then,

$$\begin{aligned}
\|(\mathbf{M}_1)_{\cdot,i}\|_2 &= \Big\| \sum_{j\neq i} \widetilde{\mathbf{X}}_1\widetilde{\mathbf{X}}_1^\top\big(\boldsymbol{b}_{j,\cdot}^{(1)} - \hat{\boldsymbol{b}}_{j,\cdot}^{(1)}\big)\mathbf{\Xi}_{j,i}\Big\|_2 \\
&\leq \Lambda_{\max}(\widetilde{\mathbf{X}}_1\widetilde{\mathbf{X}}_1^\top)\cdot\|\hat{\mathbf{B}}_1 - \mathbf{B}_1\|_F \cdot \sqrt{\sum_{j\neq i}\mathbf{\Xi}_{j,i}^2} \\
&= \Lambda_{\max}(\widetilde{\mathbf{X}}_1\widetilde{\mathbf{X}}_1^\top)\cdot\|\hat{\mathbf{B}}_1 - \mathbf{B}_1\|_F \cdot \sqrt{\mathbf{\Xi}_{i,i}(1-\mathbf{\Xi}_{i,i})}.
\end{aligned} \tag{B.50}$$

Also,

$$\mathrm{tr}\Big(\widetilde{\mathbf{X}}_1^\top\hat{\mathbf{V}}^\top\mathcal{P}_{\hat{\mathbf{\Gamma}}^\top}^\perp\hat{\mathbf{V}}\widetilde{\mathbf{X}}_1\Big) \geq \big(1-\delta(1,\hat{\mathbf{\Gamma}})\big)\Lambda_{\min}(\widetilde{\mathbf{X}}_1\widetilde{\mathbf{X}}_1^\top)\|\hat{\boldsymbol{b}}_{i,\cdot}^{(1)} - \boldsymbol{b}_{i,\cdot}^{(1)}\|_2^2. \tag{B.51}$$

Putting Eq.(B.35), Eq.(B.37), Eq.(B.48), Eq.(B.50) and Eq.(B.51) together:

$$\big(1-\delta(1,\hat{\mathbf{\Gamma}})\big)\Lambda_{\min}(\widetilde{\mathbf{X}}_1\widetilde{\mathbf{X}}_1^\top)\|\hat{\boldsymbol{b}}_{i,\cdot}^{(1)} - \boldsymbol{b}_{i,\cdot}^{(1)}\|_2^2 \leq d_{m,n}\eta\|\hat{\boldsymbol{b}}_{i,\cdot}^{(1)} - \boldsymbol{b}_{i,\cdot}^{(1)}\|_2, \tag{B.52}$$

where

$$d_{m,n} = \frac{3}{2} + \frac{\sqrt{20\cdot|\mathcal{I}_1|}}{1-4\delta}\frac{\Lambda_{\max}(\widetilde{\mathbf{X}}_1\widetilde{\mathbf{X}}_1^\top)}{\Lambda_{\min}(\widetilde{\mathbf{X}}_1\widetilde{\mathbf{X}}_1^\top)}\eta\max_{1\leq i\leq m}\sqrt{1-\mathbf{\Xi}_{i,i}}.$$

Note that Eq.(B.52) holds uniformly over $i = 1, \ldots, m$, the conclusion follows.

(3) The proof is straightforward from Eq.(B.48):

$$|\hat{\mathcal{S}}(\theta) \cap \mathcal{I}_1| \Big(b_{\min} - \frac{\theta}{\sqrt{n}}\Big)^2 \leq \|\hat{\mathbf{B}}_1 - \mathbf{B}_1\|_F^2. \tag{B.53}$$

■

B.7. Proof of Proposition 5.9

Let $\widetilde{\mathbf{G}} = \mathbf{G}\mathcal{P}_{\mathbf{X}_2}^{\perp}$, $\widetilde{\mathbf{X}}_1 = \mathbf{X}_1\mathcal{P}_{\mathbf{X}_2}^{\perp}$. From Eq.(3.2) and Eq.(4.6),

$$\mathbf{Y}_{\mathcal{S}}\mathcal{P}_{\mathbf{X}_2}^{\perp} = (\mathbf{B}_1)_{\mathcal{S}}\widetilde{\mathbf{X}}_1 + \mathbf{\Gamma}_{\mathcal{S}}\widetilde{\mathbf{G}} + (\mathbf{F}_{\mathcal{S}} + \mathbf{E}_{\mathcal{S}})\mathcal{P}_{\mathbf{X}_2}^{\perp}. \tag{B.54}$$

Then, $\frac{\hat{\mathbf{G}}}{\sqrt{n}}$ is clearly the top K eigenvectors of $\frac{1}{mn}\mathcal{P}_{\mathbf{X}_2}^{\perp}\mathbf{Y}_{\mathcal{S}}^{\top}\mathbf{Y}_{\mathcal{S}}\mathcal{P}_{\mathbf{X}_2}^{\perp}$, where

$$\frac{1}{mn}\mathcal{P}_{\mathbf{X}_2}^{\perp}\mathbf{Y}_{\mathcal{S}}^{\top}\mathbf{Y}_{\mathcal{S}}\mathcal{P}_{\mathbf{X}_2}^{\perp} = \mathbf{M}_1 + \mathbf{M}_2 + \mathbf{M}_3 + \mathbf{M}_4 + \mathbf{M}_5,$$

$$\mathbf{M}_1 = \frac{1}{mn}\widetilde{\mathbf{G}}^{\top}\mathbf{\Gamma}_{\mathcal{S}}^{\top}\mathbf{\Gamma}_{\mathcal{S}}\widetilde{\mathbf{G}},$$

$$\mathbf{M}_2 = \frac{1}{mn}\Big(\widetilde{\mathbf{G}}^{\top}\mathbf{\Gamma}_{\mathcal{S}}^{\top}(\mathbf{F}_{\mathcal{S}} + \mathbf{E}_{\mathcal{S}})\mathcal{P}_{\mathbf{X}_2}^{\perp} + \mathcal{P}_{\mathbf{X}_2}^{\perp}(\mathbf{F}_{\mathcal{S}} + \mathbf{E}_{\mathcal{S}})^{\top}\mathbf{\Gamma}_{\mathcal{S}}\widetilde{\mathbf{G}}\Big),$$

$$\mathbf{M}_3 = \frac{1}{mn}\mathcal{P}_{\mathbf{X}_2}^{\perp}(\mathbf{F}_{\mathcal{S}} + \mathbf{E}_{\mathcal{S}})^{\top}(\mathbf{F}_{\mathcal{S}} + \mathbf{E}_{\mathcal{S}})\mathcal{P}_{\mathbf{X}_2}^{\perp},$$

$$\mathbf{M}_4 = \frac{1}{mn}\widetilde{\mathbf{X}}_1^{\top}(\mathbf{B}_1)_{\mathcal{S}}^{\top}(\mathbf{B}_1)_{\mathcal{S}}\widetilde{\mathbf{X}}_1,$$

$$\mathbf{M}_5 = \frac{1}{mn}\widetilde{\mathbf{X}}_1^{\top}(\mathbf{B}_1)_{\mathcal{S}}^{\top}\Big(\mathbf{\Gamma}_{\mathcal{S}}\widetilde{\mathbf{G}} + (\mathbf{F}_{\mathcal{S}} + \mathbf{E}_{\mathcal{S}})\mathcal{P}_{\mathbf{X}_2}^{\perp}\Big) + (\mathbf{B}_1)_{\mathcal{S}}\widetilde{\mathbf{X}}_1\Big(\mathbf{\Gamma}_{\mathcal{S}}\widetilde{\mathbf{G}} + (\mathbf{F}_{\mathcal{S}} + \mathbf{E}_{\mathcal{S}})\mathcal{P}_{\mathbf{X}_2}^{\perp}\Big)^{\top}.$$

Note that $\mathbf{G}\mathbf{G}^{\top} \succeq \widetilde{\mathbf{G}}\widetilde{\mathbf{G}}^{\top} \succeq \mathbf{G}\mathcal{P}_{\mathbf{X}}^{\perp}\mathbf{G}^{\top}$, Proposition A.6(5) and assumption 5.2(a) imply that

$$\Omega_P(1) \leq \Lambda_{\min}\Big(\frac{1}{n}\widetilde{\mathbf{G}}\widetilde{\mathbf{G}}^{\top}\Big) \leq \Lambda_{\max}\Big(\frac{1}{n}\widetilde{\mathbf{G}}\widetilde{\mathbf{G}}^{\top}\Big) \leq O_P(1). \tag{B.55}$$

From Eq.(B.55),

$$\Omega_P(snr_{m,n}) \leq \Lambda_{\min}\Big((\widetilde{\mathbf{G}}\widetilde{\mathbf{G}}^\top)^{1/2}\mathbf{\Gamma}_{\mathcal{S}}^\top\mathbf{\Gamma}_{\mathcal{S}}(\widetilde{\mathbf{G}}\widetilde{\mathbf{G}}^\top)^{1/2}\Big) \leq \Lambda_{\max}\Big((\widetilde{\mathbf{G}}\widetilde{\mathbf{G}}^\top)^{1/2}\mathbf{\Gamma}_{\mathcal{S}}^\top\mathbf{\Gamma}_{\mathcal{S}}(\widetilde{\mathbf{G}}\widetilde{\mathbf{G}}^\top)^{1/2}\Big) \leq O_P(sn).$$

Since $\widetilde{\mathbf{G}}^\top(\widetilde{\mathbf{G}}\widetilde{\mathbf{G}}^\top)^{-1/2}$ is a rank K orthogonal matrix and

$$\mathbf{M}_1 = \widetilde{\mathbf{G}}^\top(\widetilde{\mathbf{G}}\widetilde{\mathbf{G}}^\top)^{-1/2}\Big\{\frac{1}{mn}(\widetilde{\mathbf{G}}\widetilde{\mathbf{G}}^\top)^{1/2}\mathbf{\Gamma}_{\mathcal{S}}^\top\mathbf{\Gamma}_{\mathcal{S}}(\widetilde{\mathbf{G}}\widetilde{\mathbf{G}}^\top)^{1/2}\Big\}(\widetilde{\mathbf{G}}\widetilde{\mathbf{G}}^\top)^{-1/2}\widetilde{\mathbf{G}},$$

we conclude that

$$\Lambda_K(\mathbf{M}_1) = \Omega_P\Big(\frac{sr_{m,n}}{m}\Big). \tag{B.56}$$

Consider the first term in $\mathbf{M}_2$,

$$\begin{aligned}
\|\widetilde{\mathbf{G}}^\top\mathbf{\Gamma}_{\mathcal{S}}^\top(\mathbf{F}_{\mathcal{S}} + \mathbf{E}_{\mathcal{S}})\mathcal{P}_{\mathbf{X}_2}^\perp\|_2 &= \|\mathcal{P}_{\mathbf{X}_2}^\perp\mathbf{G}^\top\mathbf{\Gamma}_{\mathcal{S}}^\top(\mathbf{F}_{\mathcal{S}} + \mathbf{E}_{\mathcal{S}})\mathcal{P}_{\mathbf{X}_2}^\perp\|_2 \\
&\leq \|\mathbf{G}^\top\mathbf{\Gamma}_{\mathcal{S}}^\top(\mathbf{F}_{\mathcal{S}} + \mathbf{E}_{\mathcal{S}})\|_2 \\
&= \|\mathbf{\Gamma}_{\mathcal{S}}^\top(\mathbf{F}_{\mathcal{S}} + \mathbf{E}_{\mathcal{S}})\mathbf{G}^\top\|_2.
\end{aligned}$$

$\mathbf{\Gamma}_{\mathcal{S}}^\top(\mathbf{F}_{\mathcal{S}} + \mathbf{E}_{\mathcal{S}})\mathbf{G}^\top$ is a $K \times K$ matrix. For its (i,j)th entry,

$$(\mathbf{\Gamma}_{\mathcal{S}}^\top(\mathbf{F}_{\mathcal{S}} + \mathbf{E}_{\mathcal{S}})\mathbf{G}^\top)_{i,j} = \sum_{k_2=1}^{n}\sum_{k_1\in\mathcal{S}}\gamma_{k_1,i}g_{j,k_2}(f_{k_1,k_2} + e_{k_1,k_2}).$$

Note that $\sum_{k_1\in\mathcal{S}}\gamma_{k_1,i}g_{j,k_2}(f_{k_1,k_2} + e_{k_1,k_2})$ has mean zero and is independent over n,

$$\begin{aligned}
\mathbb{E}\big[(\mathbf{\Gamma}_{\mathcal{S}}^\top(\mathbf{F}_{\mathcal{S}} + \mathbf{E}_{\mathcal{S}})\mathbf{G}^\top)_{i,j}^2\big] &= n\mathbb{E}\Big[\sum_{k_1\in\mathcal{S}}\gamma_{k_1,i}g_{j,1}(f_{k_1,1} + e_{k_1,1})\Big]^2 \\
&\leq 2n\mathbb{E}\Big[\sum_{k_1\in\mathcal{S}}\gamma_{k_1,i}g_{j,1}f_{k_1,1}\Big]^2 + 2n\mathbb{E}\Big[\sum_{k_1\in\mathcal{S}}\gamma_{k_1,i}g_{j,1}e_{k_1,1}\Big]^2 \\
&= 2n\mathbb{E}\Big[\sum_{k_1\in\mathcal{S}}\gamma_{k_1,i}g_{j,1}f_{k_1,1}\Big]^2 + 2n\sum_{k_1\in\mathcal{S}}\gamma_{k_1,i}^2\sigma_{k_1}^2 \\
&\leq O(n)\boldsymbol{\gamma}_{\boldsymbol{\cdot},i}^\top\mathbb{E}\big[f_{\boldsymbol{\cdot},1}f_{\boldsymbol{\cdot},1}^\top\big]\boldsymbol{\gamma}_{\boldsymbol{\cdot},i} + O(n)\|\boldsymbol{\gamma}_{\boldsymbol{\cdot},i}\|_2^2 = O(m)\|\boldsymbol{\gamma}_{\boldsymbol{\cdot},i}\|_2^2.
\end{aligned}$$

Sum over i, j, we conclude that

$$\|\mathbf{\Gamma}_{\mathcal{S}}^{\top}(\mathbf{F}_{\mathcal{S}}+\mathbf{E}_{\mathcal{S}})\mathbf{G}^{\top}\|_2 \leq \|\mathbf{\Gamma}_{\mathcal{S}}^{\top}(\mathbf{F}_{\mathcal{S}}+\mathbf{E}_{\mathcal{S}})\mathbf{G}^{\top}\|_F = O_P(\sqrt{Km})\cdot\|\mathbf{\Gamma}\|_F = O_P(m\sqrt{KM}).$$

The spectral norm of the second term in $\mathbf{M}_2$ is of the same order by a similar proof.

By Proposition A.6, $\|\mathbf{M}_3\|_2 = O_P\Big(\frac{1}{n}\Big)$. Thus far, we have shown that

$$\Lambda_K(\mathbf{M}_1) = \Omega_P\Big(\frac{sr_{m,n}}{m}\Big),\ \|\mathbf{M}_2\|_2 = O_P\Big(\frac{\sqrt{KM}}{n}\Big),\ \|\mathbf{M}_3\|_2 = O_P\Big(\frac{1}{n}\Big). \tag{B.57}$$

If there were no false negatives, Eq.(B.57) is sufficient for deriving our conclusions. Note that $\|(\mathbf{B}_1)_{\mathcal{S}}\widetilde{\mathbf{X}}_1\|_2 = \|(\mathbf{B}_1)_{\mathcal{F}}\widetilde{\mathbf{X}}_1\|_2 = O_P\Big(\sqrt{n}\|(\mathbf{B}_1)_{\mathcal{F}}\|_F\Big)$, then

$$\|\mathbf{M}_4\|_2 = O_P\Big(\frac{\|(\mathbf{B}_1)_{\mathcal{F}}\|_F^2}{m}\Big),\ \|\mathbf{M}_5\|_2 = O_P\Big(\frac{\|(\mathbf{B}_1)_{\mathcal{F}}\|_F}{\sqrt{m}}\Big). \tag{B.58}$$

Finally, by diagonalization of $\frac{1}{mn}(\widetilde{\mathbf{G}}\widetilde{\mathbf{G}}^{\top})^{1/2}\mathbf{\Gamma}_{\mathcal{S}}^{\top}\mathbf{\Gamma}_{\mathcal{S}}(\widetilde{\mathbf{G}}\widetilde{\mathbf{G}}^{\top})^{1/2}$ we see that $\widetilde{\mathbf{G}}^{\top}(\widetilde{\mathbf{G}}\widetilde{\mathbf{G}}^{\top})^{-1/2}$ is up to an $K\times K$ orthogonal matrix the top K eigenvectors of $\mathbf{M}_1$. The orthogonal transformation does not affect the projection matrix. By identifying $\mathbf{M}_1$, $\mathbf{M}_2$, $\mathcal{S}_1$, $\mathcal{S}_2$ in Lemma A.2, Lemma A.3 with $\mathbf{M}_1$, $\mathbf{M}_1+\mathbf{M}_2+\mathbf{M}_3+\mathbf{M}_4+\mathbf{M}_5$, $\big[\Lambda_K(\mathbf{M}_1),\infty\big)$, $\big(-\infty,\Lambda_{K+1}(\mathbf{M}_1+\mathbf{M}_2+\mathbf{M}_3+\mathbf{M}_4+\mathbf{M}_5)\big]$, Eq.(B.57) and Eq.(B.58) imply that

$$\big\|\mathcal{P}_{\hat{\mathbf{G}}}-\mathcal{P}_{\widetilde{\mathbf{G}}}\big\|_2 = O_P\Big(\frac{m}{sr_{m,n}}\Big(\frac{\sqrt{KM}}{n}+\frac{\|(\mathbf{B}_1)_{\mathcal{F}}\|_F}{\sqrt{m}}\Big)\Big),$$

and the first result follows. The second result follows from Lemma A.3. ■

B.8. Proof of Theorem 5.10

Since

$$\mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}}-\mathcal{P}^{\perp}_{\mathbf{X}_2,\mathbf{G}} = \mathcal{P}^{\perp}_{\mathbf{X}_2}\Big(\mathcal{P}^{\perp}_{\hat{\mathbf{G}}\mathcal{P}^{\perp}_{\mathbf{X}_2}}-\mathcal{P}^{\perp}_{\mathbf{G}\mathcal{P}^{\perp}_{\mathbf{X}_2}}\Big),$$

We know $\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}}\mathbf{X}_1^{\top}$ is invertible with probability tending to 1 by Proposition A.6(4) and Proposition 5.9.

For each $i = 1, \ldots, m$,

$$\hat{\boldsymbol{b}}^{(1)}_{i,\text{final}} - \boldsymbol{b}^{(1)}_{i,\boldsymbol{\cdot}} = \big(\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}}\mathbf{X}_1^{\top}\big)^{-1}\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}}\big(\mathbf{G}^{\top}\boldsymbol{\gamma}_{i,\boldsymbol{\cdot}} + \boldsymbol{f}_{i,\boldsymbol{\cdot}} + \boldsymbol{\epsilon}_{i,\boldsymbol{\cdot}}\big). \tag{B.59}$$

Before the main proof, we prove the following lemma.

Lemma B.1. *Under the same conditions as in Theorem 5.10:*

(1) $\max_{1\le i\le m}\|\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}}\mathbf{G}^{\top}\boldsymbol{\gamma}_{i,\boldsymbol{\cdot}}\|_2 = O_P(n\xi_{m,n}\sqrt{M})$.

(2) $\max_{1\le i\le m}\|\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}}\boldsymbol{f}_{i,\boldsymbol{\cdot}}\|_2 = O_P\Big(\sqrt{n}\big(\sqrt{n}\xi_{m,n}\sqrt{Ma_{K+1}} + \sqrt{Ma_{K+1}\log m}\big)\Big)$.

(3) $\max_{1\le i\le m}\|\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}}\boldsymbol{\epsilon}_{i,\boldsymbol{\cdot}}\|_2 = O_P(\sqrt{n\log m})$.

Proof. (1) Since $\mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}}\mathbf{G}^{\top} = \mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}}\Big(\mathbf{G}\mathcal{P}^{\perp}_{\mathbf{X}_2} - \Big(\frac{\mathbf{G}\mathcal{P}^{\perp}_{\mathbf{X}_2}\mathbf{G}^{\top}}{n}\Big)^{1/2}\mathbf{O}\hat{\mathbf{G}}\Big)^{\top}$, where $\mathbf{O}$ is from Proposition 5.9.

$$\max_{1\le i\le m}\|\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}}\mathbf{G}^{\top}\boldsymbol{\gamma}_{i,\boldsymbol{\cdot}}\|_2 \le O_P(\sqrt{n})O_P(\xi_{m,n}\sqrt{n})\max_{1\le i\le m}\|\boldsymbol{\gamma}_{i,\boldsymbol{\cdot}}\|_2 = O_P(n\xi_{m,n}\sqrt{M}).$$

(2) In this case, we need assumption 5.6.

$$\begin{aligned}\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}}\boldsymbol{f}_{i,\boldsymbol{\cdot}} &= \boldsymbol{\Gamma}_{\mathbf{X}}\Big(\mathbf{G}\mathcal{P}^{\perp}_{\mathbf{X}_2} - \Big(\frac{\mathbf{G}\mathcal{P}^{\perp}_{\mathbf{X}_2}\mathbf{G}^{\top}}{n}\Big)^{1/2}\mathbf{O}\hat{\mathbf{G}}\Big)\mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}}\boldsymbol{f}_{i,\boldsymbol{\cdot}} + \mathbf{E}_{\mathbf{X}}\mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}}\boldsymbol{f}_{i,\boldsymbol{\cdot}}\\ &= O_P(n\xi_{m,n}\sqrt{Ma_{K+1}}) + \mathbf{E}_{\mathbf{X}}\mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}}\boldsymbol{f}_{i,\boldsymbol{\cdot}}\\ &= O_P(n\xi_{m,n}\sqrt{Ma_{K+1}}) + \big(1 + O_P(\xi_{m,n})\big)\mathbf{E}_{\mathbf{X}}\mathcal{P}^{\perp}_{\mathbf{X}_2,\mathbf{G}}\boldsymbol{f}_{i,\boldsymbol{\cdot}}.\end{aligned}$$

Let $\boldsymbol{z}_i = \mathcal{P}^{\perp}_{\mathbf{X}_2,\mathbf{G}}\boldsymbol{f}_{i,\boldsymbol{\cdot}}$. Note that z_i is a function of $\mathbf{X}_2$ and $\mathbf{A}$. Apply the General Hoeffding's inequality to each row of $\mathbf{E}_{\mathbf{X}}\boldsymbol{z}_i$,

$$\mathbb{P}\Big(\big|(\mathbf{E}_{\mathbf{X}})^{\top}_{j,\boldsymbol{\cdot}}\boldsymbol{z}_i\big| \ge t\big|\mathbf{X}_2, \mathbf{A}\Big) \le 2\exp\Big(-\frac{ct^2}{K^2\|\boldsymbol{z}_i\|_2^2}\Big), \tag{B.60}$$

where c, K are constants depending on the conditional distribution of $\mathbf{E}_{\mathbf{X}}$ given $\mathbf{X}_2$, $\mathbf{A}$. By union bound,

$$\begin{aligned}\mathbb{P}\Big(\max_{1\le j\le c}\max_{1\le i\le m}\big|(\mathbf{E}_{\mathbf{X}})_{j,\cdot}^{\top}\boldsymbol{z}_i\big|\ge t\big|\mathbf{X}_2,\mathbf{A}\Big) &\le 2mc\exp\Big(-\frac{ct^2}{K^2\|\boldsymbol{z}_i\|_2^2}\Big)\\ &\lesssim \exp\Big(\log m-\frac{ct^2}{K^2\|\boldsymbol{z}_i\|_2^2}\Big)\\ &\le \exp\Big(\log m-\frac{ct^2}{K^2nMa_{K+1}}\Big).\end{aligned}\tag{B.61}$$

The last line is independent of $\mathbf{X}_2$ and $\mathbf{A}$, therefore, the upper bound also holds unconditionally. Eq.(B.61) implies that

$$\max_{1\le j\le c}\max_{1\le i\le m}\big|(\mathbf{E}_{\mathbf{X}})_{j,\cdot}^{\top}\boldsymbol{z}_i\big|=O_P\big(\sqrt{nMa_{K+1}\log m}\big),\tag{B.62}$$

and

$$\max_{1\le i\le m}\|\mathbf{E}_{\mathbf{X}}\mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}}\boldsymbol{f}_{i,\cdot}\|_2\le\sqrt{c}\max_{1\le j\le c}\max_{1\le i\le m}\big|(\mathbf{E}_{\mathbf{X}})_{j,\cdot}^{\top}\boldsymbol{z}_i\big|=O_P\big(\sqrt{nMa_{K+1}\log m}\big).\tag{B.63}$$

(3) Firstly, $\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}}\boldsymbol{\epsilon}_{i,\cdot}=\big(1+O_P(\xi_{m,n})\big)\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\mathbf{G}}\boldsymbol{\epsilon}_{i,\cdot}$. Conditional on $\mathbf{X}$ and $\mathbf{G}$, the j-th entry of $\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\mathbf{G}}\boldsymbol{\epsilon}_{i,\cdot}$ follows $N\big(0,\sigma_i^2\|(\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\mathbf{G}})_{j,\cdot}\|_2^2\big)$. Then,

$$\big|(\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\mathbf{G}}\boldsymbol{\epsilon}_{i,\cdot})_j\big|^2\sim\sigma_i^2\|(\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\mathbf{G}})_{j,\cdot}\|_2^2\cdot\chi_1^2.$$

Apply Lemma A.1,

$$\max_{1\le i\le m}\big|(\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}}\boldsymbol{\epsilon}_{i,\cdot})_j\big|^2\le\max_{1\le i\le m}\sigma_i^2\|\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\mathbf{G}}\|_2^2\cdot(1+\sqrt{2\log m}+2\log m)=O_P(n\log m).$$

Summing over $j=1,\dots,c$,

$$\max_{1\le i\le m}\big\|(\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}}\boldsymbol{\epsilon}_{i,\cdot})_j\big\|_2^2=O_P(n\log m).$$

□

Apply Lemma B.1 to Eq.(B.59):

$$\sqrt{n}(\hat{\boldsymbol{b}}^{(1)}_{i,\text{final}} - \boldsymbol{b}^{(1)}_{i,\boldsymbol{\cdot}}) = \Big(\frac{\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}}\mathbf{X}_1^{\top}}{n}\Big)^{-1}\frac{\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}}\boldsymbol{\epsilon}_{i,\boldsymbol{\cdot}}}{\sqrt{n}} + o_P(1). \tag{B.64}$$

Because $\frac{\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}}\mathbf{X}_1^{\top}}{n} = \frac{\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\mathbf{G}}\big(1+O_P(\xi_{m,n})\big)\mathbf{X}_1^{\top}}{n} \xrightarrow{P} \boldsymbol{\Sigma}_{\mathbf{X}_1|\mathbf{X}_2,\mathbf{G}}$ and $\Lambda_{\min}(\boldsymbol{\Sigma}_{\mathbf{X}_1|\mathbf{X}_2,\mathbf{G}}) > 0$,

$$\begin{aligned}
\sqrt{n}(\hat{\boldsymbol{b}}^{(1)}_{i,\text{final}} - \boldsymbol{b}^{(1)}_{i,\boldsymbol{\cdot}}) &= \Big(\frac{\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\mathbf{G}}\mathbf{X}_1^{\top}}{n}\Big)^{-1}\frac{\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\mathbf{G}}\boldsymbol{\epsilon}_{i,\boldsymbol{\cdot}}}{\sqrt{n}} + o_P(1)\\
&= \Big(\frac{\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\mathbf{G}}\mathbf{X}_1^{\top}}{n}\Big)^{-1} N_c\Big(\mathbf{0}, \sigma_i^2\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\mathbf{G}}\mathbf{X}_1^{\top}\Big) + o_P(1)\\
&= \boldsymbol{\Sigma}^{-1/2}_{\mathbf{X}_1|\mathbf{X}_2,\mathbf{G}} N_c(\mathbf{0}, \sigma_i^2\mathbf{I}_c) + o_P(1)\\
&\xrightarrow{d} N_c\big(\mathbf{0}, \sigma_i^2\boldsymbol{\Sigma}^{-1}_{\mathbf{X}_1|\mathbf{X}_2,\mathbf{G}}\big).
\end{aligned}$$

To prove the uniform consistency of $\hat{\sigma}_i^2$, note that

$$\boldsymbol{Y}_{i,\boldsymbol{\cdot}}^{\top}\mathcal{P}^{\perp}_{\mathbf{X},\hat{\mathbf{G}}}\boldsymbol{Y}_{i,\boldsymbol{\cdot}} = \boldsymbol{\gamma}_{i,\boldsymbol{\cdot}}^{\top}\mathbf{G}\mathcal{P}^{\perp}_{\mathbf{X},\hat{\mathbf{G}}}\mathbf{G}^{\top}\boldsymbol{\gamma}_{i,\boldsymbol{\cdot}} + 2\boldsymbol{\gamma}_{i,\boldsymbol{\cdot}}^{\top}\mathbf{G}\mathcal{P}^{\perp}_{\mathbf{X},\hat{\mathbf{G}}}(\boldsymbol{f}_{i,\boldsymbol{\cdot}} + \boldsymbol{\epsilon}_{i,\boldsymbol{\cdot}}) + (\boldsymbol{f}_{i,\boldsymbol{\cdot}} + \boldsymbol{\epsilon}_{i,\boldsymbol{\cdot}})^{\top}\mathcal{P}^{\perp}_{\mathbf{X},\hat{\mathbf{G}}}(\boldsymbol{f}_{i,\boldsymbol{\cdot}} + \boldsymbol{\epsilon}_{i,\boldsymbol{\cdot}}).$$

Recall that $\big\|\mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}}\mathbf{G}^{\top}\big\|_2 = O_P(\sqrt{n}\xi_{m,n})$, then from Proposition A.6,

$$\max_{1\le i\le m}\big\|\boldsymbol{Y}_{i,\boldsymbol{\cdot}}^{\top}\mathcal{P}^{\perp}_{\mathbf{X},\hat{\mathbf{G}}}\boldsymbol{Y}_{i,\boldsymbol{\cdot}} - \boldsymbol{\epsilon}_{i,\boldsymbol{\cdot}}^{\top}\mathcal{P}^{\perp}_{\mathbf{X},\hat{\mathbf{G}}}\boldsymbol{\epsilon}_{i,\boldsymbol{\cdot}}\big\|_2 = O_P\big(n\xi_{m,n}^2M + n\xi_{m,n}\sqrt{M} + nMa_{K+1} + n\sqrt{Ma_{K+1}}\big),$$

which implies that

$$\max_{1\le i\le m}\Big\|\hat{\sigma}_i^2 - \frac{1}{n-p-K}\boldsymbol{\epsilon}_{i,\boldsymbol{\cdot}}^{\top}\mathcal{P}^{\perp}_{\mathbf{X},\hat{\mathbf{G}}}\boldsymbol{\epsilon}_{i,\boldsymbol{\cdot}}\Big\|_2 = o_P(1).$$

From Lemma A.1,

$$\max_{1\le i\le m}\Big|\frac{1}{n-p-K}\boldsymbol{\epsilon}_{i,\boldsymbol{\cdot}}^{\top}\boldsymbol{\epsilon}_{i,\boldsymbol{\cdot}} - \sigma_i^2\Big| = O_P\Big(\sqrt{\frac{\log m}{n}}\Big).$$

Also, from a similar proof to Eq.(B.23),

$$\max_{1\le i\le m}\left\|\boldsymbol{\epsilon}_{i,\boldsymbol{\cdot}}^{\top}(\mathbf{X}^{\top},\mathbf{G}^{\top})\right\|_2 = O_P(\sqrt{n(p+K)\vee\log m}).$$

The above two results show that

$$\max_{1\le i\le m}\left|\frac{1}{n-p-K}\boldsymbol{\epsilon}_{i,\boldsymbol{\cdot}}^{\top}\mathcal{P}^{\perp}_{\mathbf{X},\hat{\mathbf{G}}}\boldsymbol{\epsilon}_{i,\boldsymbol{\cdot}}-\sigma_i^2\right|\xrightarrow{P}0,$$

which establishes that

$$\max_{1\le i\le m}|\hat{\sigma}_i^2-\sigma_i^2|\xrightarrow{P}0.$$

B.9. Proof of Theorem 5.11

(1) Let

$$\hat{T}_i^{ad}(\boldsymbol{v}_i)=\frac{\big(\hat{\boldsymbol{b}}_{i,\text{final}}^{(1)}-\boldsymbol{v}_i\big)^{\top}\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}}\mathbf{X}_1^{\top}\big(\hat{\boldsymbol{b}}_{i,\text{final}}^{(1)}-\boldsymbol{v}_i\big)}{\hat{\sigma}_i^2}.$$

Then,

$$\begin{aligned}\hat{T}_i^{ad}(\boldsymbol{v}_i)&=(\boldsymbol{Y}_{i,\boldsymbol{\cdot}}-\mathbf{X}_1^{\top}\boldsymbol{v}_i)^{\top}\mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}}\mathbf{X}_1^{\top}\big(\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}}\mathbf{X}_1^{\top}\big)^{-1}\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}}(\boldsymbol{Y}_{i,\boldsymbol{\cdot}}-\mathbf{X}_1^{\top}\boldsymbol{v}_i)/\hat{\sigma}_i^2\\&:=S_{i1}+S_{i2}+S_{i3}+2S_{i4}+2S_{i5}+2S_{i6},\end{aligned}\tag{B.65}$$

where

$$S_{i1}=\big(\mathbf{X}_1^{\top}(\boldsymbol{b}_{i,\boldsymbol{\cdot}}^{(1)}-\boldsymbol{v}_i)+\boldsymbol{\epsilon}_{i,\boldsymbol{\cdot}}\big)^{\top}\mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}}\mathbf{X}_1^{\top}\big(\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}}\mathbf{X}_1^{\top}\big)^{-1}\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}}\big(\mathbf{X}_1^{\top}(\boldsymbol{b}_{i,\boldsymbol{\cdot}}^{(1)}-\boldsymbol{v}_i)+\boldsymbol{\epsilon}_{i,\boldsymbol{\cdot}}\big)/\hat{\sigma}_i^2,$$

$$S_{i2}=\boldsymbol{f}_{i,\boldsymbol{\cdot}}^{\top}\mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}}\mathbf{X}_1^{\top}\big(\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}}\mathbf{X}_1^{\top}\big)^{-1}\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}}\boldsymbol{f}_{i,\boldsymbol{\cdot}}/\hat{\sigma}_i^2,$$

$$S_{i3}=\boldsymbol{\gamma}_{i,\boldsymbol{\cdot}}^{\top}\mathbf{G}\mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}}\mathbf{X}_1^{\top}\big(\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}}\mathbf{X}_1^{\top}\big)^{-1}\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}}\mathbf{G}^{\top}\boldsymbol{\gamma}_{i,\boldsymbol{\cdot}}/\hat{\sigma}_i^2,$$

$$S_{i4}=\boldsymbol{f}_{i,\boldsymbol{\cdot}}^{\top}\mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}}\mathbf{X}_1^{\top}\big(\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}}\mathbf{X}_1^{\top}\big)^{-1}\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}}\big(\mathbf{X}_1^{\top}(\boldsymbol{b}_{i,\boldsymbol{\cdot}}^{(1)}-\boldsymbol{v}_i)+\boldsymbol{\epsilon}_{i,\boldsymbol{\cdot}}\big)/\hat{\sigma}_i^2,$$

$$S_{i5}=\boldsymbol{\gamma}_{i,\boldsymbol{\cdot}}^{\top}\mathbf{G}\mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}}\mathbf{X}_1^{\top}\big(\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}}\mathbf{X}_1^{\top}\big)^{-1}\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}}\big(\mathbf{X}_1^{\top}(\boldsymbol{b}_{i,\boldsymbol{\cdot}}^{(1)}-\boldsymbol{v}_i)+\boldsymbol{\epsilon}_{i,\boldsymbol{\cdot}}\big)/\hat{\sigma}_i^2,$$

$$S_{i6} = \boldsymbol{\gamma}_{i,\bullet}^\top \mathbf{G} \mathcal{P}^\perp_{\mathbf{X}_2,\hat{\mathbf{G}}} \mathbf{X}_1^\top \big(\mathbf{X}_1 \mathcal{P}^\perp_{\mathbf{X}_2,\hat{\mathbf{G}}} \mathbf{X}_1^\top\big)^{-1} \mathbf{X}_1 \mathcal{P}^\perp_{\mathbf{X}_2,\hat{\mathbf{G}}} \boldsymbol{f}_{i,\bullet}/\hat{\sigma}_i^2.$$

We set $\boldsymbol{v}_i = \mathbf{0}$ for $i \in \mathcal{I}_0 \cup \mathcal{I}_{11} \cup \mathcal{I}_{12}$, or $\boldsymbol{v}_i = \boldsymbol{b}_{i,\bullet}^{(1)}$ if $i \in \mathcal{I}_{13}$. Note that

$$\max_{1\le i\le m} \big\|\mathbf{X}_1^\top(\boldsymbol{b}_{i,\bullet}^{(1)} - \boldsymbol{v}_i)\big\|_2 = O_P(1)$$

and

$$\big\|\big(\mathbf{X}_1 \mathcal{P}^\perp_{\mathbf{X}_2,\hat{\mathbf{G}}} \mathbf{X}_1^\top\big)^{-1}\big\|_2 = O_P\Big(\frac{1}{n}\Big).$$

Apply Lemma B.1 and the uniform consistency of $\hat{\sigma}_i^2$:

$$\max_{1\le i\le m} |S_{i2}| = O_P\big(n\xi_{m,n}^2 M a_{K+1} + M a_{K+1}\log m\big),$$

$$\max_{1\le i\le m} |S_{i3}| = O_P\big(n\xi_{m,n}^2 M\big),$$

$$\max_{1\le i\le m} |S_{i4}| = O_P\Big(\sqrt{\log m}\big(\sqrt{n}\xi_{m,n}\sqrt{M a_{K+1}} + \sqrt{M a_{K+1}\log m}\big)\Big),$$

$$\max_{1\le i\le m} |S_{i5}| = O_P\big(\sqrt{n}\xi_{m,n}\sqrt{M\log m}\big),$$

$$\max_{1\le i\le m} |S_{i6}| = O_P\Big(\sqrt{n}\xi_{m,n}\sqrt{M}\big(\sqrt{n}\xi_{m,n}\sqrt{M a_{K+1}} + \sqrt{M a_{K+1}\log m}\big)\Big). \tag{B.66}$$

Therefore,

$$\max_{1\le i\le m} \big|\hat{T}_i^{ad}(\boldsymbol{v}_i) - S_{i1}\big| \xrightarrow{P} 0. \tag{B.67}$$

Let

$$S_{i1}^* = \begin{cases} \dfrac{\big(\mathbf{X}_1 \mathcal{P}^\perp_{\mathbf{X}_2,\mathbf{G}} \mathbf{X}_1^\top\big)^{-1/2} \mathbf{X}_1 \mathcal{P}^\perp_{\mathbf{X}_2,\mathbf{G}} \big(\mathbf{X}_1^\top \boldsymbol{b}_{i,\bullet}^{(1)} + \boldsymbol{\epsilon}_{i,\bullet}\big)}{\sigma_i} & i \in \mathcal{I}_0 \cup \mathcal{I}_{11} \cup \mathcal{I}_{12} \\ \dfrac{\big(\mathbf{X}_1 \mathcal{P}^\perp_{\mathbf{X}_2,\mathbf{G}} \mathbf{X}_1^\top\big)^{-1/2} \mathbf{X}_1 \mathcal{P}^\perp_{\mathbf{X}_2,\mathbf{G}} \boldsymbol{\epsilon}_{i,\bullet}}{\sigma_i} & i \in \mathcal{I}_{13} \end{cases}.$$

From Proposition 5.9 and Lemma B.1, it can be shown that

$$\max_{1\le i\le m} \big|\|S_{i1}^*\|_2^2 - S_{i1}\big| \xrightarrow{P} 0. \tag{B.68}$$

Because $\boldsymbol{\epsilon}_{i,\bullet}$ is independent of $\mathbf{X}$ and $\mathbf{G}$, conditional on $\mathbf{X}$ and $\mathbf{G}$,

$$S_{i1}^{*}\big|\mathbf{X},\mathbf{G}\sim N_c\Big(\big(\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\mathbf{G}}\mathbf{X}_1^{\top}\big)^{1/2}\boldsymbol{b}^{(1)}_{i,\bullet}/\sigma_i,\mathbf{I}_c\Big) \tag{B.69}$$

for $i\in\mathcal{I}_0\cup\mathcal{I}_{11}\cup\mathcal{I}_{12}$, and

$$S_{i1}^{*}\big|\mathbf{X},\mathbf{G}\sim N_c\Big(\mathbf{0},\mathbf{I}_c\Big), \tag{B.70}$$

for $i\in\mathcal{I}_{13}$. Let

$$S_{i1}^{**}=\begin{cases}S_{i1}^{*}-\frac{\big(\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\mathbf{G}}\mathbf{X}_1^{\top}\big)^{1/2}\boldsymbol{b}^{(1)}_{i,\bullet}}{\sigma_i} & i\in\mathcal{I}_0\cup\mathcal{I}_{11}\cup\mathcal{I}_{12}\\ S_{i1}^{*} & i\in\mathcal{I}_{13}\end{cases}. \tag{B.71}$$

From Eq.(B.69) and Eq.(B.70), S_{i1}^{**} are i.i.d. $N_c(\mathbf{0},\mathbf{I}_c)$ random vectors.

For $i\in\mathcal{I}_0\cup\mathcal{I}_{11}\cup\mathcal{I}_{12}$, $\hat{T}_i=\hat{T}_i^{ad}(\mathbf{0})$. If $i\in\mathcal{I}_0$, $S_{i1}^{**}=S_{i1}^{*}$. From Eq.(B.67) and Eq.(B.68), $\hat{T}_i\xrightarrow{d}\chi^2_c$. If $i\in\mathcal{I}_{11}$, $\max_{i\in\mathcal{I}_{11}}\big\|(\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\mathbf{G}}\mathbf{X}_1^{\top})^{1/2}\boldsymbol{b}_{i,\bullet}\big\|_2=o_P(1)$. Then, $\big|S_{i1}^{**}-S_{i1}^{*}\big|=o_P(1)$, $\hat{T}_i\xrightarrow{d}\chi^2_c$. If $i\in\mathcal{I}_{12}$, $\max_{i\in\mathcal{I}_{12}}\big\|(\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\mathbf{G}}\mathbf{X}_1^{\top})^{1/2}\boldsymbol{b}_{i,\bullet}-\boldsymbol{\Sigma}^{1/2}_{\mathbf{X}_1|\mathbf{X}_2,\mathbf{G}}\mathbf{c}_i\big\|_2=o_P(1)$. $\hat{T}_i\xrightarrow{d}\Big\|S_{i1}^{**}+\frac{\boldsymbol{\Sigma}^{1/2}_{\mathbf{X}_1|\mathbf{X}_2,\mathbf{G}}\mathbf{c}_i}{\sigma_i}\Big\|_2^2\sim\chi^2_c(\cdot,\lambda_i)$. For $i\in\mathcal{I}_{13}$, by the triangle inequality,

$$\sqrt{\hat{T}_i}\geq\sqrt{\frac{(\boldsymbol{b}^{(1)}_{i,\bullet})^{\top}\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}}\mathbf{X}_1^{\top}\boldsymbol{b}^{(1)}_{i,\bullet}}{\hat{\sigma}_i^2}}-\sqrt{\hat{T}_i^{ad}(\boldsymbol{b}^{(1)}_{i,\bullet})}=O_P(\sqrt{n})\|\boldsymbol{b}^{(1)}_{i,\bullet}\|_2-O_P(1)\xrightarrow{d}\infty. \tag{B.72}$$

The asymptotic independence between $\hat{T}_i$ and $\hat{T}_{i'}$, $\forall i,i'\in\mathcal{I}_0\cup\mathcal{I}_{11}\cup\mathcal{I}_{12}$ follows easily from the independence between S_{i1}^{**} and $S_{i'1}^{**}$. If either i or i' belongs to $\mathcal{I}_{13}$, the asymptotic independence follows from the fact that an almost surely infinite random variable is independent of any other random variable.

(2) In this case, $T_{i,\mathrm{ora}}=\|S_{i1}^{*}\|_2^2$ for $i\in\mathcal{I}_0\cup\mathcal{I}_{11}\cup\mathcal{I}_{12}$. It suffices to show that

$$\frac{\hat{V}(t)-V_{\mathrm{ora}}(t)}{m}\xrightarrow{P}0, \tag{B.73}$$

$$\frac{\hat{R}(t)-R_{\mathrm{ora}}(t)}{m}\xrightarrow{P}0. \tag{B.74}$$

Let $X_{i,m,n} = \hat{T}_i$, $Y_{i,m,n} = T_{i,\text{ora}}$ for $i \notin \mathcal{I}_{13}$, $Z_{m,n} = (\mathbf{X}, \mathbf{G})$ and

$$W_i = \begin{cases} \|S_{i1}^{**}\|_2^2 & i \in \mathcal{I}_0 \cup \mathcal{I}_{11} \\ \left\| S_{i1}^{**} + \frac{\mathbf{\Sigma}_{\mathbf{X}_1|\mathbf{X}_2,\mathbf{G}}^{1/2} \mathbf{c}_i}{\sigma_i} \right\|_2^2 & i \in \mathcal{I}_{12}. \end{cases}$$

It can be verified that the conditions in Lemma B.3 are satisfied with these sequences. Then,

$$\frac{\hat{V}(t)}{m} = \frac{1}{m} \sum_{i \in \mathcal{I}_0} 1(\hat{P}_i \le t) = \frac{1}{m} \sum_{i \in \mathcal{I}_0} 1(\hat{T}_i \ge C) = \frac{1}{m} \sum_{i \in \mathcal{I}_0} \mathbb{P}(\|S_{i1}^{**}\|_2^2 \ge C) + o_P(1) \xrightarrow{P} \pi_0 t. \tag{B.75}$$

Similarly,

$$\frac{\hat{V}_{\text{ora}}(t)}{m} \xrightarrow{P} \pi_0 t. \tag{B.76}$$

Eq.(B.73) follows.

Consider Eq.(B.74),

$$\frac{\hat{R}(t)}{m} = \frac{\hat{V}(t)}{m} + \frac{1}{m} \sum_{i \notin \mathcal{I}_0} 1\big(\hat{T}_i \ge C\big),$$

$$\frac{R_{\text{ora}}(t)}{m} = \frac{V_{\text{ora}}(t)}{m} + \frac{1}{m} \sum_{i \notin \mathcal{I}_0} 1\big(T_i \ge C\big).$$

From Lemma B.3,

$$\frac{1}{m} \sum_{i \in \mathcal{I}_{11}} 1\big(\hat{T}_i \ge C\big) = \frac{1}{m} \sum_{i \in \mathcal{I}_{11}} \mathbb{P}\big(\|S_{i1}^{**}\|_2^2 \ge C\big) + o_P(1) \xrightarrow{P} \pi_{11} t \tag{B.77}$$

and

$$\begin{aligned} \frac{1}{m} \sum_{i \in \mathcal{I}_{12}} 1\big(\hat{T}_i \ge C\big) &= \frac{1}{m} \sum_{i \in \mathcal{I}_{12}} \mathbb{P}\Big(\Big\| S_{i1}^{**} + \frac{\mathbf{\Sigma}_{\mathbf{X}_1|\mathbf{X}_2,\mathbf{G}}^{1/2} \mathbf{c}_i}{\sigma_i} \Big\|_2^2 \ge C\Big) + o_P(1) \\ &\xrightarrow{P} \pi_{12} \lim_{m \to \infty} \frac{1}{|\mathcal{I}_{12}|} \sum_{i \in \mathcal{I}_{12}} 1 - \chi_c^2(C, \lambda_i) \\ &= \pi_{12} F_1(t). \end{aligned} \tag{B.78}$$

Similarly,

$$\frac{1}{m}\sum_{i\in\mathcal{I}_{11}} 1\big(T_{i,\mathrm{ora}} \geq C\big) \xrightarrow{P} \pi_{11}t \tag{B.79}$$

and

$$\frac{1}{m}\sum_{i\in\mathcal{I}_{12}} 1\big(T_{i,\mathrm{ora}} \geq C\big) \xrightarrow{P} \pi_{12}F_1(t). \tag{B.80}$$

From Eq.(B.72),

$$\begin{aligned}\min_{i\in\mathcal{I}_{13}}\sqrt{\hat{T}_i} &\geq \sqrt{\frac{\Lambda_{\min}(\mathbf{X}_1\mathcal{P}^{\perp}_{\mathbf{X}_2,\hat{\mathbf{G}}}\mathbf{X}_1^{\top})\min_{i\in\mathcal{I}_{13}}\|\boldsymbol{b}^{(1)}_{i,\bullet}\|_2^2}{\max_{i\in\mathcal{I}_{13}}\hat{\sigma}_i^2}} - \max_{i\in\mathcal{I}_{13}}\sqrt{\|S^*_{i1}\|_2^2 + \left|\hat{T}_i^{ad} - \|S^*_{i1}\|_2^2\right|} \\ &= O_P(\sqrt{n})\min_{i\in\mathcal{I}_{13}}\|\boldsymbol{b}^{(1)}_{i,\bullet}\|_2 - O_P(\sqrt{\log m}) \xrightarrow{P} \infty.\end{aligned} \tag{B.81}$$

Then,

$$\frac{1}{m}\sum_{i\in\mathcal{I}_{13}} 1\big(\hat{T}_i \geq C\big) \xrightarrow{P} \pi_{13}. \tag{B.82}$$

Similarly,

$$\frac{1}{m}\sum_{i\in\mathcal{I}_{13}} 1\big(T_{i,\mathrm{ora}} \geq C\big) \xrightarrow{P} \pi_{13}. \tag{B.83}$$

Combining Eq.(B.75) to Eq.(B.83),

$$\frac{\hat{R}(t)}{m} \xrightarrow{P} (\pi_0+\pi_{11})t + \pi_{12}F_1(t) + \pi_{13}, \quad \frac{R_{\mathrm{ora}}(t)}{m} \xrightarrow{P} (\pi_0+\pi_{11})t + \pi_{12}F_1(t) + \pi_{13}. \tag{B.84}$$

Then, Eq.(B.73) and Eq.(B.74) are verified for each fixed t. The uniform convergence result follows from Lemma B.2.

For the second conclusion,

$$\begin{aligned}\big|\widehat{\mathrm{FDP}}_\lambda(t) - \widehat{\mathrm{FDP}}(t)\big| &= \frac{\hat{\pi}_0(\lambda)t}{(\hat{R}(t)\vee 1)/m} - c\frac{\hat{V}(t)/m}{(\hat{R}(t)\vee 1)/m} \\ &= \frac{\hat{\pi}_0(\lambda)t - c\pi_0 t}{(\pi_0+\pi_{11})t + \pi_{12}F_1(t) + \pi_{13} \vee \frac{1}{m}} + o_P(1).\end{aligned}$$

Note that

$$\hat{\pi}_0(\lambda) = \frac{m - \hat{R}(\lambda)}{m(1-\lambda)} = \frac{1 - \hat{R}(\lambda)/m}{1-\lambda} \xrightarrow{P} \frac{1 - (\pi_0 + \pi_{11})\lambda - \pi_{12}F_1(\lambda) - \pi_{13}}{1-\lambda} = c\pi_0,$$

the conclusion follows.

(3) The proof is similar to (2). The only difference lies in Eq.(B.80). In this case,

$$\frac{1}{m}\sum_{i\in\mathcal{I}_{12}} 1\big(T_{i,\text{ora}} \geq C\big) \xrightarrow{P} \pi_{12} \lim_{m\to\infty} \frac{1}{|\mathcal{I}_{12}|} \sum_{i\in\mathcal{I}_{12}} 1 - \chi_c^2\Big(C, \frac{\boldsymbol{c}_i^\top \boldsymbol{\Sigma}_{\mathbf{X}_1|\mathbf{X}_2}\boldsymbol{c}_i}{\sigma_i^2}\Big) \geq \pi_{12}F_1(t). \quad \text{(B.85)}$$

This results in

$$\frac{\hat{R}(t)}{m} - \frac{R_{\text{ora}}(t)}{m} \leq \epsilon$$

with probability tending to 1. ■

Lemma B.2. (*Uniform convergence for random c.d.f.s*) *Suppose $F_n(t)$, $F(t)$ are c.d.f.s and $F_n(t) \xrightarrow{P} F(t)$, $F_n(t-) \xrightarrow{P} F(t-)$ for every $t \in \mathbb{R}$. Then,*

$$\sup_t |F_n(t) - F(t)| \xrightarrow{P} 0.$$

Proof. We prove this by verifying the definition of convergence in probability. Let $k > 0$ be any fixed positive integer. For $1 \leq j \leq k-1$, let $x_{j,k} = \inf\{t : F(t) \geq j/k\}$. This implies that $F(x_{j,k}-) - F(x_{j-1,k}) \leq k^{-1}$, where $x_{0,k} = -\infty$ and $x_{k,k} = \infty$. The convergence of $F_n(t)$ and $F_n(t-)$ for each of $x_{j,k}$ imply that we can find a positive integer N such that whenever $n \geq N$,

$$\mathbb{P}\Big(\cup_{j=1}^{k-1} \big\{ \max\big(|F_n(x_{j,k}) - F(x_{j,k})|, |F_n(x_{j,k}-) - F(x_{j,k}-)|\big) > k^{-1}\big\}\Big) \leq \epsilon$$

for an arbitrary $\epsilon > 0$. We restrict our further discussions in the event $\cap_{j=1}^{k-1}\big\{ \max\big(|F_n(x_{j,k}) - F(x_{j,k})|, |F_n(x_{j,k}-) - F(x_{j,k}-)|\big) \leq k^{-1}\big\}$. For any $t \in (x_{j-1,k}, x_{j,k})$ with $1 \leq j \leq k$,

$$F_n(t) \leq F_n(x_{j,k}-) \leq F(x_{j,k}-) + k^{-1} \leq F(x_{j-1,k}) + 2k^{-1} \leq F(x) + 2k^{-1},$$

$$F_n(t) \geq F_n(x_{j-1,k}) \geq F(x_{j-1,k}) - k^{-1} \geq F(x_{j,k}-) - 2k^{-1} \geq F(x) - 2k^{-1}.$$

This proves that for $n \geq N$

$$\mathbb{P}\Big(\sup_t |F_n(t) - F(t)| > 2k^{-1} \Big) \leq \epsilon,$$

and the conclusion follows. □

Lemma B.3. *Let $\{X_{i,m,n}\}_{i=1}^m$, $\{Y_{i,m,n}\}_{i=1}^m$ be two random arrays such that*

$$\max_{1 \leq i \leq m} |X_{i,m,n} - Y_{i,m,n}| \xrightarrow{P} 0$$

as m, n tend to infinity. Suppose $Y_{i,m,n}|Z_{m,n}$ are conditionally independent over $i = 1, \ldots, m$ with conditional Lebesgue c.d.f. $p(t, \lambda_i(Z_n))$. $p(t, \theta)$ is parametrized by $\theta \in \mathbf{\Theta}$ and is continuous w.r.t. t and θ. $\lambda_i(Z_{m,n}) \in \mathbf{\Theta}$ and $\max_{1 \leq i \leq m} |\lambda_i(Z_{m,n}) - c_i| \xrightarrow{P} 0$ for a deterministic bounded sequence $\{c_i\}_{i=1}^\infty \subset \mathbf{\Theta}$. Also, suppose W_i, $i = 1, 2 \ldots$ are independent random variables such that W_i has c.d.f. $p(t, c_i)$. Then, the following conclusions hold:

(1) *For any $C_1, C_2 \in \mathbb{R}$,*

$$\max_{i \neq i'} \big| \mathbb{P}(Y_{i,m,n} \geq C_1, Y_{i',m,n} \geq C_2) - \mathbb{P}(W_i \geq C_1, W_{i'} \geq C_2) \big| \to 0,$$

$$\max_{i \neq i'} \big| \mathbb{P}(X_{i,m,n} \geq C_1, X_{i',m,n} \geq C_2) - \mathbb{P}(W_i \geq C_1, W_{i'} \geq C_2) \big| \to 0.$$

(2) *For any $C \in \mathbb{R}$,*

$$\frac{1}{m} \sum_{i=1}^m 1(Y_{i,m,n} \geq C) - \frac{1}{m} \sum_{i=1}^m \mathbb{P}(W_i \geq C) \xrightarrow{P} 0,$$

$$\frac{1}{m} \sum_{i=1}^m 1(X_{i,m,n} \geq C) - \frac{1}{m} \sum_{i=1}^m \mathbb{P}(W_i \geq C) \xrightarrow{P} 0.$$

Proof. (1) Note that

$$\mathbb{P}(Y_{i,m,n} \geq C_1, Y_{i',m,n} \geq C_2) - \mathbb{P}(W_i \geq C_1, W_{i'} \geq C_2)$$

$$= \int p(C_1, \lambda_i(Z_{m,n}))p(C_2, \lambda_{i'}(Z_{m,n}))d\mathbb{P}^{Z_{m,n}} - p(C_1, c_i)p(C_2, c_{i'})$$
$$= \int \Big(p(C_1, \lambda_i(Z_{m,n})) - p(C_1, c_i)\Big)p(C_2, \lambda_{i'}(Z_{m,n}))d\mathbb{P}^{Z_{m,n}}$$
$$+ \int p(C_1, \lambda_i(Z_{m,n}))\Big(p(C_2, \lambda_{i'}(Z_{m,n})) - p(C_2, c_{i'})\Big)d\mathbb{P}^{Z_{m,n}}.$$

Then,

$$\max_{i \neq i'} \big|\mathbb{P}(Y_{i,m,n} \geq C_1, Y_{i',m,n} \geq C_2) - \mathbb{P}(W_i \geq C_1, W_{i'} \geq C_2)\big| \leq I_1 + I_2$$

where

$$I_1 = \int \max_i \Big|p(C_1, \lambda_i(Z_{m,n})) - p(C_1, c_i)\Big|d\mathbb{P}^{Z_{m,n}},$$

$$I_2 = \int \max_{i'} \Big|p(C_2, \lambda_{i'}(Z_{m,n})) - p(C_2, c_{i'})\Big|d\mathbb{P}^{Z_{m,n}}.$$

Since $p(\cdot, \theta)$ is uniformly continuous w.r.t. θ in any bounded region, $\max_i \Big|p(C_1, \lambda_i(Z_{m,n})) - p(C_1, c_i)\Big| \xrightarrow{P} 0$ and $\max_{i'} \Big|p(C_2, \lambda_{i'}(Z_{m,n})) - p(C_2, c_{i'})\Big| \xrightarrow{P} 0$. By the dominated convergence theorem, we can conclude that $I_1, I_2 \to 0$, and the first conclusion follows.

For the second conclusion, it suffices to show that

$$\max_{i \neq i'} \big|\mathbb{P}(X_{i,m,n} \geq C_1, X_{i',m,n} \geq C_2) - \mathbb{P}(Y_i \geq C_1, Y_{i'} \geq C_2)\big| \to 0.$$

Fix any $\epsilon > 0$, let $\mathcal{E} = \Big\{ \max_{1 \leq i \leq m} |X_{i,m,n} - Y_{i,m,n}| \geq \epsilon \Big\}$. We know

$$\mathbb{P}(\mathcal{E}) \to 0.$$

Then,

$$\mathbb{P}\big(X_{i,m,n} \geq C_1, X_{i',m,n} \geq C\big) \leq \mathbb{P}\big(X_{i,m,n} \geq C_1, X_{i',m,n} \geq C_2, \mathcal{E}^c\big) + \mathbb{P}(\mathcal{E})$$
$$\leq \mathbb{P}\big(Y_{i,m,n} > C_1 - \epsilon, Y_{i',m,n} > C_2 - \epsilon, \mathcal{E}^c\big) + \mathbb{P}(\mathcal{E})$$
$$\leq \mathbb{P}\big(Y_{i,m,n} > C_1 - \epsilon, Y_{i',m,n} > C_2 - \epsilon\big) + \mathbb{P}(\mathcal{E}).$$

and

$$\begin{aligned}
&\mathbb{P}\big(Y_{i,m,n} > C_1 - \epsilon, Y_{i',m,n} > C_2 - \epsilon\big) - \mathbb{P}\big(Y_{i,m,n} \geq C_1, Y_{i',m,n} \geq C_2\big) \\
&\leq \mathbb{P}\big(C_1 - \epsilon < Y_{i,m,n} \leq C_1\big) + \mathbb{P}\big(C_2 - \epsilon < Y_{i',m,n} \leq C_2\big) \\
&= \int p(C_1, \lambda_i(Z_{m,n})) - p(C_1 - \epsilon, \lambda_i(Z_{m,n})) + p(C_2, \lambda_{i'}(Z_{m,n})) - p(C_2 - \epsilon, \lambda_{i'}(Z_{m,n}))d\mathbb{P}^{Z_{m,n}} \\
&\leq \int w_\epsilon(C_1, Z_{m,n}) + w_\epsilon(C_2, Z_{m,n})d\mathbb{P}^{Z_{m,n}},
\end{aligned}$$

where $w_\epsilon(t, Z_{m,n}) = \max_{1\leq i\leq n} \sup_{t-\epsilon\leq s\leq t+\epsilon} \big|p(t, \lambda_i(Z_{m,n})) - p(s, \lambda_i(Z_{m,n}))\big|$. Similarly,

$$\begin{aligned}
\mathbb{P}\big(X_{i,m,n} \geq C_1, X_{i',m,n} \geq C_2\big) &\geq \mathbb{P}\big(X_{i,m,n} \geq C_1, X_{i',m,n} \geq C_2, \mathcal{E}^c\big) \\
&\geq \mathbb{P}\big(Y_{i,m,n} \geq C_1 + \epsilon, Y_{i',m,n} \geq C_2 + \epsilon, \mathcal{E}^c\big) \\
&\geq \mathbb{P}\big(Y_{i,m,n} \geq C_1 + \epsilon, Y_{i',m,n} \geq C_2 + \epsilon\big) - \mathbb{P}(\mathcal{E}),
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{P}\big(Y_{i,m,n} \geq C_1, Y_{i',m,n} \geq C_2\big) - \mathbb{P}\big(Y_{i,m,n} \geq C_1 + \epsilon, Y_{i',m,n} \geq C_2 + \epsilon\big) \\
&\leq \int w_\epsilon(C_1, Z_{m,n}) + w_\epsilon(C_2, Z_{m,n})d\mathbb{P}^{Z_{m,n}}.
\end{aligned}$$

The above show that

$$\max_{i\neq i'} \big|\mathbb{P}(X_{i,m,n} \geq C_1, X_{i',m,n} \geq C_2) - \mathbb{P}(Y_{i,m,n} \geq C_1, Y_{i',m,n} \geq C_2)\big|$$

can be bounded by

$$\int w_\epsilon(C_1, Z_{m,n}) + w_\epsilon(C_2, Z_{m,n})d\mathbb{P}^{Z_{m,n}} + \mathbb{P}(\mathcal{E}),$$

which can be made arbitrarily small for small enough $\epsilon > 0$ and large enough m, n by uniform continuity of p and the dominated convergence theorem. The conclusion follows.

(2) We only prove the first conclusion, the second conclusion follows using the same

argument. By Chebyshev's inequality,

$$\mathbb{P}\Big(\Big|\frac{1}{m}\sum_{i=1}^{m} 1(Y_{i,m,n} \geq C) - \mathbb{P}(Y_{i,m,n} \geq C)\Big| \geq \epsilon\Big) \leq \frac{1}{4m\epsilon^2} + S,\ \forall \epsilon > 0,$$

where

$$S = \frac{1}{m^2\epsilon^2}\sum_{i \neq i'} \mathrm{Cov}\big(1(Y_{i,m,n} \geq C), 1(Y_{i',m,n} \geq C)\big).$$

From (1), we know S can be made arbitrarily small for large m, n. Also,

$$\Big|\frac{1}{m}\sum_{i=1}^{m} \mathbb{P}(Y_{i,m,n} \geq C) - \frac{1}{m}\sum_{i=1}^{m} \mathbb{P}(W_i \geq C)\Big| \leq \max_{1 \leq i \leq m} \big|\mathbb{P}(Y_{i,m,n} \geq C) - \mathbb{P}(W_i \geq C)\big| \to 0,$$

then the conclusion follows. □

Bibliography

Almlund, M., A. L. Duckworth, J. Heckman, and T. Kautz (2011). Personality psychology and economics. In *Handbook of the Economics of Education*, Volume 4, pp. 1–181. Elsevier.

Anderson, T. and H. Rubin (1956). Statistical inference in factor analysis. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability: Held at the Statistical Laboratory, University of California, December, 1954, July and August, 1955*, Volume 1, pp. 111. Univ of California Press.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica 71*(1), 135–171.

Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models. *Econometrica 70*(1), 191–221.

Bai, J. and P. Wang (2016). Econometric analysis of large factor models. *Annual Review of Economics 8*, 53–80.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological) 57*(1), 289–300.

Benjamini, Y. and D. Yekutieli (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics 29*(4), 1165–1188.

Bhatia, R. (2013). *Matrix analysis*, Volume 169. Springer Science & Business Media.

Bing, X., W. Cheng, H. Feng, and Y. Ning (2023). Inference in high-dimensional multivariate response regression with hidden variables. *Journal of the American Statistical Association 119*(547), 2066–2077.

Bing, X., Y. Ning, and Y. Xu (2022). Adaptive estimation in multivariate response regression with hidden variables. *The annals of statistics 50*(2), 640–672.

Boucheron, S., G. Lugosi, and O. Bousquet (2003). Concentration inequalities. In *Summer school on machine learning*, pp. 208–240. Springer.

Du, L. and C. Zhang (2017). Estimation of false discovery proportion in multiple testing: From normal to chi-squared test statistics. *Electronic Journal of Statistics 11*(1), 1048–1091.

Dunford, N. and J. T. Schwartz (1965). Linear operators. part ii. spectral theory. *Bull. Amer. Math. Soc 2*(9904), 11348–9.

Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association 102*(477), 93–103.

Fan, J., X. Han, and W. Gu (2012). Estimating false discovery proportion under arbitrary covariance dependence. *Journal of the American Statistical Association 107*(499), 1019–1035.

Fan, J., Y. Liao, and M. Mincheva (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society. Series B, Statistical methodology 75*(4), 603–680.

Feng, Y. (2020). Causal inference in possibly nonlinear factor models. *arXiv preprint arXiv:2008.13651*.

Foucart, S. and H. Rauhut (2013). *A Mathematical Introduction to Compressive Sensing*. Birkhäuser Basel.

Friguet, C., M. Kloareg, and D. Causeur (2009). A factor model approach to multiple testing under dependence. *Journal of the American Statistical Association 104*(488), 1406–1415.

Gagnon-Bartsch, J. A., L. Jacob, and T. P. Speed (2013). Removing unwanted variation from high dimensional data with negative controls. *Berkeley: Tech Reports from Dep Stat Univ California 64*, 1–112.

Grotzinger, A. D., M. Rhemtulla, de Vlaming, et al. (2019). Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nature human behaviour 3*(5), 513–525.

Ledermann, W. (1937). On the rank of the reduced correlational matrix in multiple-factor analysis. *Psychometrika 2*, 85–93.

Lee, S., W. Sun, F. A. Wright, and F. Zou (2017). An improved and explicit surrogate variable analysis procedure by coefficient adjustment. *Biometrika 104*(2), 303–316.

Leek, J. T. (2007). *Surrogate variable analysis*. University of Washington.

Leek, J. T., R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly, and R. A. Irizarry (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics 11*(10), 733–739.

Leek, J. T. and J. D. Storey (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics 3*(9), 1724–1735.

Leek, J. T. and J. D. Storey (2008). A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences 105*(48), 18718–18723.

Li, J. and L. Ji (2005). Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity 95*(3), 221–227.

Lounici, K., M. Pontil, A. Tsybakov, and S. A. van de Geer (2010). Oracle inequalities and optimal inference under group sparsity. *Annals of Statistics 39*, 2164–2204.

McKennan, C. and D. Nicolae (2019). Accounting for unobserved covariates with varying degrees of estimability in high-dimensional biological data. *Biometrika 106*(4), 823–840.

Oliva, M., M. Muñoz-Aguirre, S. Kim-Hellmuth, V. Wucher, A. D. Gewirtz, D. J. Cotter, P. Parsana, S. Kasela, B. Balliu, A. Viñuela, et al. (2020). The impact of sex on gene expression across human tissues. *Science 369*(6509), eaba3066.

Rachev, S. T. and L. Ruschendorf (1991). A transformation property of minimal metrics. *Theory of Probability and Its Applications 35*(1), 110–117.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society 56*(4), 931–954.

Romano, J. P. and M. Wolf (2005). Stepwise multiple testing as formalized data snooping. *Econometrica 73*(4), 1237–1282.

Rosenbaum, M. and A. B. Tsybakov (2010). Sparse recovery under matrix uncertainty. *The Annals of Statistics 38*(5), 2620 – 2651.

Rüschendorf, L. and V. de Valk (1993). On regression representations of stochastic processes. *Stochastic Processes and their Applications 46*(2), 183–198.

Shapiro, A. (1982). Rank-reducibility of a symmetric matrix and sampling theory of minimum trace factor analysis. *Psychometrika 47*, 187–199.

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

Storey, J. D., J. E. Taylor, and D. Siegmund (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66*(1), 187–205.

Sun, H. (2005). Mercer theorem for RKHS on noncompact sets. *Journal of Complexity 21*(3), 337–349.

van de Geer, S., P. Bühlmann, and S. Zhou (2011). The adaptive and the thresholded Lasso for potentially misspecified models (and a lower bound for the Lasso). *Electron. J. Stat. 5*, 688–749.

Vawter, M. P., S. Evans, P. Choudary, H. Tomita, J. Meador-Woodruff, M. Molnar, J. Li, J. F. Lopez, R. Myers, D. Cox, et al. (2004). Gender-specific gene expression in post-mortem human brain: localization to sex chromosomes. *Neuropsychopharmacology 29*(2), 373–384.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, Volume 47. Cambridge university press.

Wahba, G. (1990). *Spline models for observational data*. SIAM.

Wang, J., Q. Zhao, T. Hastie, and A. B. Owen (2017). Confounder adjustment in multiple hypothesis testing. *Annals of statistics 45*(5), 1863.

Yu, Y., T. Wang, and R. J. Samworth (2015). A useful variant of the davis–kahan theorem for statisticians. *Biometrika 102*(2), 315–323.

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