

# Heterogeneity Adjustment with Applications to Graphical Model Inference

Jianqing Fan\*, Han Liu, Weichen Wang, and Ziwei Zhu

Department of Operations Research and Financial Engineering, Princeton University

## Abstract

Heterogeneity is an unwanted variation when analyzing aggregated datasets from multiple sources. Though different methods have been proposed for heterogeneity adjustment, no systematic theory exists to justify these methods. In this work, we propose a generic framework named ALPHA (short for Adaptive Low-rank Principal Heterogeneity Adjustment) to model, estimate, and adjust heterogeneity from the original data. Once the heterogeneity is adjusted, we are able to remove the biases of batch effects and to enhance the inferential power by aggregating the homogeneous residuals from multiple sources. Under a pervasive assumption that the latent heterogeneity factors simultaneously affect a large fraction of observed variables, we provide a rigorous theory to justify the proposed framework. Our framework also allows the incorporation of informative covariates and appeals to the ‘Bless of Dimensionality’. As an illustrative application of this generic framework, we consider a problem of estimating high-dimensional precision matrix for graphical model inference based on multiple datasets. We also provide thorough numerical studies on both synthetic datasets and a brain imaging dataset to demonstrate the efficacy of the developed theory and methods.

**Keywords:** Heterogeneity, Batch effect, Graphical model inference, Semiparametric factor model, Principal component analysis, Brain image network.

## 1 Introduction

Aggregating and analyzing heterogeneous data is one of the most fundamental challenges in scientific data analysis. In particular, the intrinsic heterogeneity across multiple data

---

\*Address: Department of ORFE, Sherrerd Hall, Princeton University, Princeton, NJ 08544, USA, e-mail: [jqfan@princeton.edu](mailto:jqfan@princeton.edu), [hanliu@princeton.edu](mailto:hanliu@princeton.edu), [weichenw@princeton.edu](mailto:weichenw@princeton.edu), [ziweiz@princeton.edu](mailto:ziweiz@princeton.edu). This project was supported by National Science Foundation grants DMS-1206464, DMS-1406266 and 2R01-GM072611-12.

sources violates the ideal ‘independent and identically distributed’ sampling assumption and may produce misleading results if it is ignored. For example, in genomics, data heterogeneity is ubiquitous and referred to as either ‘batch effect’ or ‘lab effect’. Microarray gene expression data obtained from different labs at different processing dates may contain systematic variability. More specifically, Leek et al. (2010) analyzed a microarray data from a bladder cancer study and showed that the gene expressions vary significantly across different batches even after data normalization. Furthermore, Leek and Storey (2007) pointed out that heterogeneity across multiple data sources may be caused by unobserved factors that have confounding effects on the variables of interest, generating spurious signals. In finance, it is also known that asset returns are driven by varying market regimes and economy status, which can be regarded as a temporal batch effect. Later in this paper, we will use a brain imaging dataset to show similar heterogeneity effect. Therefore, to properly analyze data aggregated from multiple sources, we need to carefully model and adjust the heterogeneity effect.

Modeling and estimating heterogeneity effect is challenging for two reasons. (i) Typically, we can only access a limited number of samples from an individual group, given the high cost of biological experiment, technological constraint or fast economy regime switching. (ii) The dimensionality can be much larger than the total aggregated number of samples. The past decade has witnessed the development of many methods for adjusting batch effect in high throughput genomics data. See, for example, Sims et al. (2008); Alter et al. (2000); Leek and Storey (2007); Johnson et al. (2007). Though progresses have been made, most of the aforementioned papers focus on the practical side and none of them has a systematic theoretical justification. In fact, most of these methods are developed in a case-by-case fashion and are only applicable to certain problem domains. Thus, there is still a gap that exists between practice and theories.

To bridge this gap, we propose a generic theoretical framework to model, estimate, and adjust heterogeneity across multiple datasets. Formally, we assume the data come from  $m$  different sources: the  $i^{th}$  data source contributes  $n_i$  samples, each having  $p$  measurements such as gene expressions of an individual or stock returns of a day. To explicitly model heterogeneity, we assume that the batch-specific latent factors  $\mathbf{f}_t^i$  influence the observed data  $X_{jt}^i$  in batch  $i$  ( $j$  indexes variables;  $t$  indexes samples) as in the approximate factor model:

$$X_{jt}^i = \boldsymbol{\lambda}_j^i \mathbf{f}_t^i + u_{jt}^i, \quad 1 \leq j \leq p, 1 \leq t \leq n_i, 1 \leq i \leq m, \quad (1.1)$$

where  $\boldsymbol{\lambda}_j^i$  is unknown factor loading for variable  $j$  and  $u_{jt}^i$  is true uncorrupted signals. The linear term  $\boldsymbol{\lambda}_j^i \mathbf{f}_t^i$  models the heterogeneity effect. We assume that  $\mathbf{f}_t^i$  is independent of  $u_{jt}^i$

and  $\mathbf{u}_t^i = (u_{1t}, \dots, u_{pt})'$  shares the same common distribution with mean  $\mathbf{0}$  and covariance  $\Sigma_{p \times p}$  across all data sources. In the matrix-form model, (1.1) can be written as

$$\mathbf{X}^i = \Lambda^i \mathbf{F}^{i'} + \mathbf{U}^i, \quad (1.2)$$

where  $\mathbf{X}^i$  is a  $p \times n_i$  data matrix in the  $i^{th}$  batch,  $\Lambda^i$  is a  $p \times K^i$  factor loading matrix with  $\lambda_j^i$  in the  $j^{th}$  row,  $\mathbf{F}^i$  is an  $n_i \times K^i$  factor matrix and  $\mathbf{U}^i$  is a signal matrix of dimension  $p \times n_i$ . Here, we allow the number of latent factors  $K^i$  to depend on batch  $i$ .

To see how model (1.2) models the heterogeneity, we assume  $\mathbf{f}_t^i \sim N(\mathbf{0}, \mathbf{I})$  and  $\mathbf{u}_t^i \sim N(\mathbf{0}, \Sigma)$ . Then, the  $t^{th}$  sample  $\mathbf{X}_t^i$ , which is the  $t^{th}$  column of  $\mathbf{X}^i$ , follows

$$\mathbf{X}_t^i \sim N(\mathbf{0}, \Lambda^i \Lambda^{i'} + \Sigma). \quad (1.3)$$

Therefore, the heterogeneity effect is modeled as a low rank component  $\Lambda^i \Lambda^{i'}$  of the population covariance matrix of  $\mathbf{X}_t^i$ . Later, we will show that, under a pervasive assumption, the heterogeneity component can be estimated by directly applying principal component analysis (PCA) or Projected-PCA, which is more accurate when there are sufficiently informative covariates  $\mathbf{W}^i$  (Fan et al., 2016). Let  $\widehat{\Lambda}^i \widehat{\mathbf{F}}^{i'}$  be the estimated heterogeneity component. We denote  $\widehat{\mathbf{U}}^i = \mathbf{X}_t^i - \widehat{\Lambda}^i \widehat{\mathbf{F}}^{i'}$  to be the heterogeneity adjusted signal, which can be treated as homogeneous across different datasets and thus can be combined together for downstream statistical analysis. This whole framework of heterogeneity adjustment is termed ALPHA (short for Adaptive Low-rank Principal Heterogeneity Adjustment) and is schematically shown in Figure 1.

The proposed ALPHA framework is fully generic and applicable to almost all kinds of multivariate analysis of the combined, heterogeneity adjusted datasets. As an illustrative example, in this paper, we focus on the problem of Gaussian graphical model inference based on multiple datasets. It is a powerful tool to explore complex dependence structure among variables  $\mathbf{X} = (X_1, \dots, X_p)' \sim N(\mathbf{0}, \Sigma)$ . The sparsity pattern of the precision matrix  $\Omega = \Sigma^{-1}$  encodes the information of an undirected graph  $G = (V, E)$  where  $V$  consists of  $p$  vertices corresponding to  $p$  variables in  $\mathbf{X}$  and  $E$  describes their dependence relationship. To be specific,  $V_i$  and  $V_j$  are linked by an edge if and only if  $\Omega_{ij} \neq 0$ , meaning that  $X_i$  and  $X_j$  are dependent conditioning on the rest variables. For heterogeneous data across  $m$  data sources, we need to first adjust for heterogeneity using the ALPHA framework. The idea of covariate-adjusted precision matrix estimation has been studied by Cai et al. (2012), but the factor model they used assumes observed factors and no heterogeneity issue, i.e.,  $m = 1$ .

A significant amount of literature has focused on the estimation of the precision matrix  $\Omega$  for graphical models for homogeneous data. Yuan and Lin (2007), Banerjee et al. (2008),

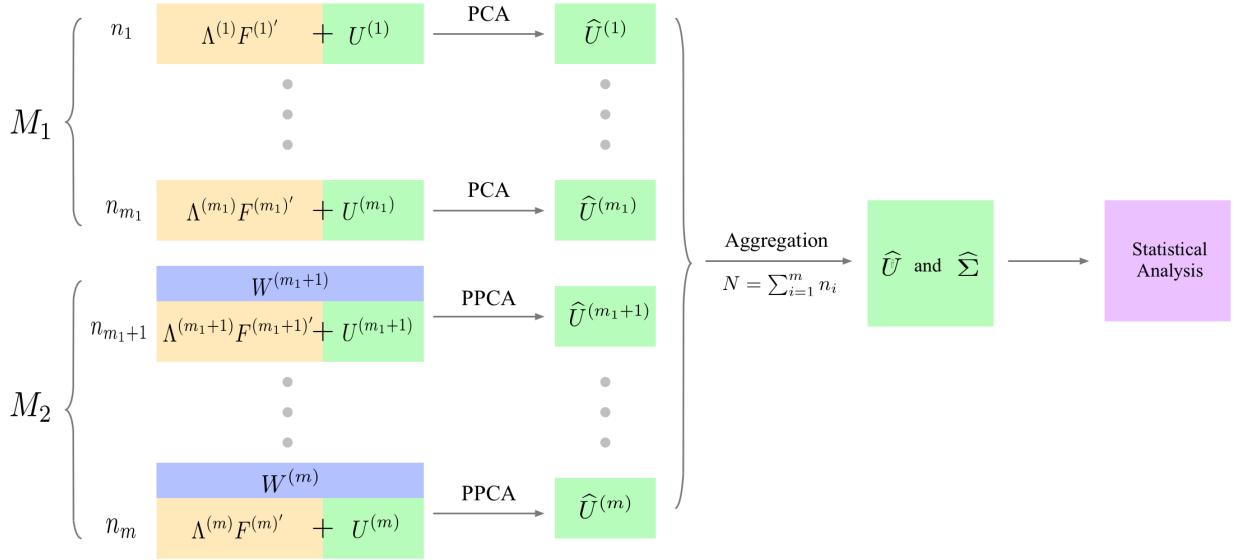


Figure 1: Schematic illustration of ALPHA: Depending whether we can find some sufficiently informative covariates  $\mathbf{W}$ , we implement principal component analysis (PCA) or Projected-PCA (PPCA) methods (labeled respectively  $M_1$  and  $M_2$ ) to remove the heterogeneity effects  $\boldsymbol{\Lambda}\mathbf{F}'$  for each batch of data. This decision was made adaptively by a test statistic. After removing the unwanted variations, the homogeneous data  $\{\mathbf{U}^{(i)}\}_{i=1}^m$  are aggregated for further analysis.

Friedman et al. (2008) developed the Graphical Lasso method using the  $L_1$  penalty and Lam and Fan (2009) and Shen et al. (2012) used a non-convex penalty. Furthermore, Ravikumar et al. (2011) and Loh and Wainwright (2013) studied the theoretical properties under different assumptions. Estimating  $\boldsymbol{\Omega}$  can be equivalently reformulated as a set of node-wise sparse linear regression that utilizes Lasso or Danzig selector for each node (Meinshausen and Bühlmann, 2006; Yuan, 2010; Cai et al., 2011). To relax the assumption of Gaussian data, Liu et al. (2009) and Liu et al. (2012) extend the graphical model to the case of semiparametric Gaussian copula and transelliptical family. Under the ALPHA framework, the adjusted data  $\widehat{\mathbf{U}}^i$  can be combined to construct an estimator for the inverse matrix  $\boldsymbol{\Omega}$  by the above methods.

The rest of the paper is organized as follows. Section 2 lays out a basic problem setup and necessary assumptions. We model the heterogeneity by a semiparametric factor model. Section 3 introduces the ALPHA methodology for heterogeneity adjustment. Two main methods of PCA and Projected-PCA will be introduced for adjusting the factor effects under different settings. A guiding rule of thumb is also proposed to determine which method is more appropriate in the real data analysis. The heterogeneity-adjusted data will be combined to provide valid graph estimation in Section 4. The CLIME method of Cai et al. (2011) is

used for estimating precision matrix, although other related methods are also applicable. Some synthetic simulations and a real dataset are analyzed to demonstrate the proposed framework in Section 5. Section 6 contains some further discussions. All the proofs are relegated to the Appendix.

## 2 Problem Setup

To more efficiently use the external covariate information in removing heterogeneity effect, we first present a semiparametric factor model. Then, based on whether the collected external covariates have explaining power on factor loadings, we discuss two different regimes where PCA or Projected-PCA (PPCA) should be used. We will state the conditions under which these methods can be formally justified.

### 2.1 Semiparametric factor model

We assume that for subgroup  $i$ , we have  $d$  external covariates  $\mathbf{W}_j^i = (W_{j1}^i, \dots, W_{jd}^i)'$  for variable  $j$ . In stock returns, these can be attributes of a firm; in brain imaging, these can be the physical locations of voxels. We assume that these covariates have some explanatory power on the loading parameters  $\boldsymbol{\lambda}_j^i$  in (1.1) so that it can be further modeled as  $\boldsymbol{\lambda}_j^i = \mathbf{g}^i(\mathbf{W}_j^i) + \boldsymbol{\gamma}_j^i$ , where  $\mathbf{g}^i(\cdot)$  is the external covariate effects on  $\boldsymbol{\lambda}_j^i$  and  $\boldsymbol{\gamma}_j^i$  is the part that can not be explained by the covariates. Thus, model (1.1) can be written as

$$X_{jt}^i = \boldsymbol{\lambda}_j^i \mathbf{f}_t^i + u_{jt}^i = (\mathbf{g}^i(\mathbf{W}_j^i) + \boldsymbol{\gamma}_j^i)' \mathbf{f}_t^i + u_{jt}^i. \quad (2.1)$$

Model (2.1) does not put much restriction. If  $\mathbf{W}_j^i$  is not informative (i.e.,  $\boldsymbol{\lambda}_j^i$  does not depend on  $\mathbf{W}_j^i$ ), then  $\mathbf{g}^i(\cdot) = 0$ , the model reduces to a regular factor model. In a matrix form, model (2.1) can be written as

$$\mathbf{X}^i = \boldsymbol{\Lambda}^i \mathbf{F}^{i\prime} + \mathbf{U}^i \text{ where } \boldsymbol{\Lambda}^i = \mathbf{G}^i(\mathbf{W}^i) + \boldsymbol{\Gamma}^i, \quad 1 \leq i \leq m. \quad (2.2)$$

In (2.2),  $\mathbf{G}^i(\mathbf{W}^i)$  and  $\boldsymbol{\Gamma}^i$  are  $p \times K^i$  component matrices of the factor loading  $\boldsymbol{\Lambda}^i$ . More specifically,  $g_k^i(\mathbf{W}_j^i)$  and  $\gamma_{jk}$  are the  $(j, k)^{th}$  element of  $\mathbf{G}^i(\mathbf{W}^i)$  and  $\boldsymbol{\Gamma}^i$  respectively. Expression (2.2) suggests that the observed data can be decomposed into a low-rank heterogeneity term  $\boldsymbol{\Lambda}^i \mathbf{F}^{i\prime}$  and a homogeneous signal term  $\mathbf{U}^i$ .

Letting  $\mathbf{u}_t^i$  be the  $t^{th}$  column of  $\mathbf{U}^i$ , we assume all  $\mathbf{u}_t^i$ 's share the same distribution for any  $t \leq n_i$  and for all subgroups  $i \leq m$  with  $\mathbb{E}[\mathbf{u}_t^i] = \mathbf{0}$ ,  $\text{Var}(\mathbf{u}_t^i) = \boldsymbol{\Sigma}$ . Our goal is to recover  $\mathbf{U}^i$  from the observation  $\mathbf{X}^i$  and combine all the estimated  $\mathbf{U}^i$ 's together to enhance the

inferential power of  $\Sigma$  or  $\Omega = \Sigma^{-1}$ .

There has been a large literature on factor models in econometrics (Bai, 2003; Bai and Ng, 2013; Fan et al., 2013; Stock and Watson, 2002), machine learning (Cai et al., 2013; Negahban and Wainwright, 2011; Candès et al., 2011) and random matrix theories (Johnstone and Lu, 2009; Paul, 2007; Shen et al., 2013; Fan and Wang, 2015). We refer the interested readers to those relevant papers and the references therein. However, none of these models incorporate the external covariate information. The semiparametric factor model (2.1) was first proposed by Connor and Linton (2007) and further investigated by Connor et al. (2012); Fan et al. (2016). Using sufficiently informative external covariates, we are able to more accurately estimate the factors and loadings, and hence yield better heterogeneous adjustment.

## 2.2 Modeling assumptions and general methodology

In this subsection, we explicitly list all the required modeling assumptions. We start with an introduction of the data generating processes.

**Assumption 2.1** (Data Generating Process). (i)  $n_i^{-1}\mathbf{F}^{i'}\mathbf{F}^i = \mathbf{I}$ .

(ii)  $\{\mathbf{u}_t^i\}_{t \leq n_i, i \leq m}$  are independent within and between subgroups.  $\mathbf{u}_t^i$ 's are identically sub-Gaussian distributed with mean zero and variance  $\Sigma$  across all subgroups and are independent of  $\{\mathbf{W}_j^i, \mathbf{f}_t^i\}$ .  $\{\mathbf{f}_t^i\}_{t \leq n_i}$  is a stationary process, but with arbitrary temporal dependency.

(iii) There exists a constant  $C_0 > 0$  such that  $\|\Sigma\|_2 \leq C_0$ .

(iv) The tail of the factors is sub-Gaussian, i.e.,  $\exists C_1, C_2 > 0$  such that for  $k \leq K^i, t \leq n_i$ ,

$$P(|f_{tk}^i| > t) \leq C_1 \exp(-C_2 t^2).$$

The above set of assumptions are commonly used in the literature, see Bai and Ng (2013); Fan et al. (2016). We omit detailed discussions here.

Based on whether the external covariates are informative, we specify two regimes, each of which requires some additional technical conditions.

### 2.2.1 Regime 1: External covariates are not informative

For the case that  $\mathbf{G}^i(\mathbf{W}^i) = \mathbf{0}$ , the external covariates do not have explanatory power on the factor loadings  $\Lambda^i$  and model (2.2) reduces to the traditional factor model, extensively studied in econometrics (Bai, 2003; Stock and Watson, 2002; Onatski, 2012). PCA will be employed in Section 3.1 to estimate the heterogeneous effect. It requires the following assumptions.

**Assumption 2.2.** (i) (*Pervasiveness*) There are two positive constants  $c_{\min}$  and  $c_{\max}$  so that

$$c_{\min} < \lambda_{\min}(p^{-1}\boldsymbol{\Lambda}^{i'}\boldsymbol{\Lambda}^i) < \lambda_{\max}(p^{-1}\boldsymbol{\Lambda}^{i'}\boldsymbol{\Lambda}^i) < c_{\max}, \quad a.s. \quad \forall i.$$

$$(ii) \max_{k \leq K^i, j \leq p} |\lambda_{jk}^i| = O_P(\sqrt{\log p}).$$

The first condition is common and essential in the factor model literature (e.g., Stock and Watson (2002)). It requires the factors to be strong enough such that the covariance matrix  $\boldsymbol{\Lambda}^i \text{cov}(\mathbf{f}_t^i)\boldsymbol{\Lambda}^i + \boldsymbol{\Sigma}$  has spiked eigenvalues. This is trivially true if  $\{\boldsymbol{\lambda}_j^i\}_{j=1}^p$ 's can be regarded as random samples from a population with nondegenerate sample covariance matrix (Fan et al., 2013). The second condition is technical, and a relaxation of bounded requirement in the literature (Fan et al., 2013; Bai and Ng, 2013).

### 2.2.2 Regime 2: External covariates are informative

When covariates are informative, we will employ the PPCA (Fan et al., 2016) to better estimate the heterogeneous effect. It requires the following assumptions.

**Assumption 2.3.** (i) (*Pervasiveness*) There are two positive constants  $c_{\min}$  and  $c_{\max}$  so that

$$c_{\min} < \lambda_{\min}(p^{-1}\mathbf{G}^i(\mathbf{W}^i)' \mathbf{G}^i(\mathbf{W}^i)) < \lambda_{\max}(p^{-1}\mathbf{G}^i(\mathbf{W}^i)' \mathbf{G}^i(\mathbf{W}^i)) < c_{\max}, \quad a.s. \quad \forall i.$$

$$(ii) \max_{k \leq K^i, j \leq p} E g_k(\mathbf{W}_j^i)^2 < \infty.$$

This assumption is parallel to Assumption 2.2 (i). Pervasiveness is trivially satisfied if  $\{\mathbf{W}_j^i\}_{j \leq p}$  are independent and  $\mathbf{G}^i$  is sufficiently smooth.

**Assumption 2.4.** (i)  $E\gamma_{jk}^i = 0$ ,  $\max_{k \leq K^i, j \leq p} |\gamma_{jk}^i| = O_P(\sqrt{\log p})$ .

(ii) Write  $\boldsymbol{\gamma}_j^i = (\gamma_{j1}^i, \dots, \gamma_{jK^i}^i)'$ . We assume  $\{\boldsymbol{\gamma}_j^i\}_{j \leq p}$  are independent of  $\{\mathbf{W}_j^i\}_{j \leq p}$ .

(iii) Define  $\nu_p = \max_{i \leq m} \max_{k \leq K^i} p^{-1} \sum_{j \leq p} \text{var}(\gamma_{jk}^i) < \infty$ . We assume

$$\max_{k \leq K^i, j \leq p} \sum_{j' \leq p} |E\gamma_{j'k}^i \gamma_{jk}^i| = O(\nu_p).$$

Condition (i) is parallel to Assumption 2.2 (ii) whereas Condition (ii) is natural since  $\boldsymbol{\Gamma}^i$  can not be explained by  $\mathbf{W}^i$ . Condition (iii) imposes cross-sectional weak dependence of  $\boldsymbol{\gamma}_j^i$ , which is much weaker than assuming independent and identically distributed  $\{\boldsymbol{\gamma}_j^i\}_{j \leq p}$ . This condition is mild as main serial dependency has been taken care of by  $g_k(\cdot)$ 's.

### 3 The ALPHA Framework

We introduce the ALPHA framework for heterogeneity adjustment. Methodologically, for each sub-dataset we aim to estimate the heterogeneity component and subtract it from the raw data. Theoretically, we aim to obtain the explicit rates of convergence for both the corrected homogeneous signal and its sample covariance matrix. Those rates will be useful when aggregating the homogeneous residuals from multiple sources.

This section covers details for heterogeneity adjustments under both regimes that  $\mathbf{G}^i(\cdot) = 0$  and  $\mathbf{G}^i(\cdot) \neq 0$ : they correspond to estimating  $\mathbf{U}^i$  by either PCA or Projected-PCA. From now on, we drop the superscript  $i$  whenever there is no confusion as we focus on the  $i^{th}$  data source. We will use the notation  $\widehat{\mathbf{F}}$  if  $\mathbf{F}$  is estimated by PCA and  $\widetilde{\mathbf{F}}$  if estimated by PPCA. This convention applies to other related quantities such as  $\widehat{\mathbf{U}}$  and  $\widetilde{\mathbf{U}}$ , the heterogeneity-adjusted estimator. In addition, we use notations such as  $\check{\mathbf{F}}$  and  $\check{\mathbf{U}}$  to denote the final estimators, which are  $\widehat{\mathbf{F}}$  and  $\widehat{\mathbf{U}}$  if PCA is used, and  $\widetilde{\mathbf{F}}$  and  $\widetilde{\mathbf{U}}$  if PPCA is used.

Estimators for latent factors under regimes 1 and 2 satisfy  $n^{-1}\check{\mathbf{F}}'\check{\mathbf{F}} = \mathbf{I}$ , which corresponds to normalization in Assumption 2.1 (i). By the principle of least squares, the residual estimator of  $\mathbf{U}$  then admits the form

$$\check{\mathbf{U}} = \mathbf{X} \left( \mathbf{I} - \frac{1}{n} \check{\mathbf{F}} \check{\mathbf{F}}' \right). \quad (3.1)$$

It possesses the following properties.

**Theorem 3.1.** *For any  $K$  by  $K$  matrix  $\mathbf{H}$  such that  $\|\mathbf{H}\| = O_P(1)$ , if  $\log p = O(n)$ ,*

$$\check{\mathbf{U}} - \mathbf{U} = -\frac{1}{n} \mathbf{U} \mathbf{F} \mathbf{F}' + \boldsymbol{\Pi},$$

where  $\|\boldsymbol{\Pi}\|_{\max} = O_P(\sqrt{\log n}/n \cdot (\|\mathbf{F}'(\check{\mathbf{F}} - \mathbf{F}\mathbf{H})\|_{\max} \|\boldsymbol{\Lambda}\|_{\max} + \|\mathbf{U}(\check{\mathbf{F}} - \mathbf{F}\mathbf{H})\|_{\max}) + \|\check{\mathbf{F}} - \mathbf{F}\mathbf{H}\|_{\max} \|\boldsymbol{\Lambda}\|_{\max} + \sqrt{\log n} \cdot \|\mathbf{H}\mathbf{H}' - \mathbf{I}\|_{\max} \|\boldsymbol{\Lambda}\|_{\max})$ ; and furthermore

$$\check{\mathbf{U}}\check{\mathbf{U}}' - \mathbf{U}\mathbf{U}' = -\frac{1}{n} \mathbf{U} \mathbf{F} \mathbf{F}' \mathbf{U}' + \boldsymbol{\Delta},$$

where  $\|\boldsymbol{\Delta}\|_{\max} = O_P(\|\mathbf{U}(\check{\mathbf{F}} - \mathbf{F}\mathbf{H})\|_{\max} \|\boldsymbol{\Lambda}\|_{\max} + \|\mathbf{U}(\check{\mathbf{F}} - \mathbf{F}\mathbf{H})\|_{\max}^2 + \|\mathbf{F}'(\check{\mathbf{F}} - \mathbf{F}\mathbf{H})\|_{\max} \|\boldsymbol{\Lambda}\|_{\max}^2 + n \|\mathbf{H}\mathbf{H}' - \mathbf{I}\|_{\max} \|\boldsymbol{\Lambda}\|_{\max}^2)$ .

The above theorem states that the error of estimating  $\mathbf{U}$  by  $\check{\mathbf{U}}$  (or estimating  $\mathbf{U}\mathbf{U}'$  by  $\check{\mathbf{U}}\check{\mathbf{U}}'$ ) is decomposed into two parts. The first part is inevitable even when the factor matrix  $\mathbf{F}$  in (3.1) is known in advance. The second part is caused by the uncertainty from estimating  $\mathbf{F}$ . Since the true  $\mathbf{F}$  is identifiable up to an orthonormal transformation  $\mathbf{H}$ , we

need to carefully choose  $\mathbf{H}$  to bound the error  $\boldsymbol{\Pi}$  (or  $\boldsymbol{\Delta}$ ). We will provide explicit rates of convergence for those terms in the following two subsections.

### 3.1 Estimating factors by PCA

In regime 1, we directly use PCA to adjust data heterogeneity. PCA estimates  $\mathbf{F}$  by  $\widehat{\mathbf{F}}$  where the  $k^{th}$  column of  $\widehat{\mathbf{F}}/\sqrt{n}$  is the eigenvector of  $(pn)^{-1}\mathbf{X}'\mathbf{X}$  corresponding to the  $k^{th}$  largest eigenvalue. By the definition of  $\widehat{\mathbf{F}}$ , we have  $(np)^{-1}\mathbf{X}'\mathbf{X}\widehat{\mathbf{F}} = \widehat{\mathbf{F}}\mathbf{K}$ , where  $\mathbf{K}$  is a  $K$  by  $K$  diagonal matrix with top  $K$  eigenvalues of  $(np)^{-1}\mathbf{X}'\mathbf{X}$  in descending order as diagonal elements. Define a  $K$  by  $K$  matrix  $\mathbf{H}$  as in Fan et al. (2013):

$$\mathbf{H} = \frac{1}{np} \boldsymbol{\Lambda}' \boldsymbol{\Lambda} \mathbf{F}' \widehat{\mathbf{F}} \mathbf{K}^{-1}.$$

It has been shown that  $\|\mathbf{K}\|$ ,  $\|\mathbf{K}^{-1}\|$  and  $\|\mathbf{H}\|$ ,  $\|\mathbf{H}^{-1}\|$  are all  $O_P(1)$ . The following theorem provides all the rates of convergences that are needed for downstream analysis.

**Theorem 3.2.** *Under Assumptions 2.1 and 2.2, we have  $\|\boldsymbol{\Lambda}\|_{\max} = O_P(\sqrt{\log p})$  and*

- (i)  $\|\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H}\|_F = O_P(\sqrt{n/p} + 1/\sqrt{n})$  and  $\|\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H}\|_{\max} = O_P(\sqrt{\log n/p} + \sqrt{\log n}/n)$ ;
- (ii)  $\|\mathbf{F}'(\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H})\|_{\max} = O_P(1 + \sqrt{n/p})$ ;
- (iii)  $\|\mathbf{U}(\widehat{\mathbf{F}} - \mathbf{F}\mathbf{H})\|_{\max} = O_P((1 + n/p)\sqrt{\log p} + n\|\boldsymbol{\Sigma}\|_1/p)$ ;
- (iv)  $\|\mathbf{H}\mathbf{H}' - \mathbf{I}\|_{\max} = O_P(1/n + 1/p)$ .

Combining the above results with Theorem 3.1, we have

$$\widehat{\mathbf{U}} - \mathbf{U} = -\frac{1}{n} \mathbf{U} \mathbf{F} \mathbf{F}' + \boldsymbol{\Pi},$$

where  $\|\boldsymbol{\Pi}\|_{\max} = O_P(\sqrt{\log n \log p}(1/\sqrt{p} + 1/n) + \sqrt{\log n}\|\boldsymbol{\Sigma}\|_1/p)$  and additionally

$$\widehat{\mathbf{U}}\widehat{\mathbf{U}}' - \mathbf{U}\mathbf{U}' = -\frac{1}{n} \mathbf{U} \mathbf{F} \mathbf{F}' \mathbf{U}' + \boldsymbol{\Delta},$$

where  $\|\boldsymbol{\Delta}\|_{\max} = O_P((1 + n/p)\log p + n^2\|\boldsymbol{\Sigma}\|_1^2/p^2)$ .

### 3.2 Estimating factors by Projected-PCA

In regime 2, we would like to incorporate the external covariates using the Projected-PCA method proposed by Fan et al. (2016). We now explain this method.

To reduce the curse of dimensionality of  $g_k(\mathbf{W}_j)$ , we assume it takes an additive form:

$$g_k(\mathbf{W}_j) = \sum_{l=1}^d g_{kl}(W_{jl}). \quad (3.2)$$

To model the unknown function  $g_{kl}(\cdot)$ , we adopt a sieve based idea which approximates  $g_{kl}(\cdot)$  by a linear combination of basis functions (e.g., B-spline, Fourier series, polynomial series, wavelets). Let  $\{\phi_1(x), \phi_2(x), \dots\}$  be a set of basis functions. Then for each  $l \leq d$ ,

$$g_{kl}(W_{jl}) = \sum_{\nu=1}^J b_{\nu,kl} \phi_{\nu}(W_{jl}) + R_{kl}(W_{jl}), \quad k \leq K, j \leq p, l \leq d. \quad (3.3)$$

Here  $\{b_{\nu,kl}\}_{\nu=1}^J$  are the sieve coefficients of the  $l^{th}$  additive component of  $g_k(\mathbf{W}_j)$ , corresponding to the  $k^{th}$  factor loading;  $R_{kl}$  is the remainder function representing the approximation error;  $J$  denotes the number of sieve bases which may grow slowly as  $p$  diverges. The basic assumption for sieve approximation is that  $\sup_x |R_{kl}(x)| \rightarrow 0$  as  $J \rightarrow \infty$ . To facilitate notation, we take the same basis functions in (3.3) for all  $k$  and  $l$  though they can be different.

Define, for each  $k \leq K$  and for each  $j \leq p$ ,

$$\begin{aligned} \mathbf{b}'_k &= (b_{1,k1}, \dots, b_{J,k1}, \dots, b_{1,kd}, \dots, b_{J,kd}) \in \mathbb{R}^{Jd}, \\ \phi(\mathbf{W}_j)' &= (\phi_1(W_{j1}), \dots, \phi_J(W_{j1}), \dots, \phi_1(W_{jd}), \dots, \phi_J(W_{jd})) \in \mathbb{R}^{Jd}. \end{aligned}$$

Then, we can write

$$g_k(\mathbf{W}_j) = \phi(\mathbf{W}_j)' \mathbf{b}_k + \sum_{l=1}^d R_{kl}(W_{jl}).$$

Let  $\mathbf{B} = (\mathbf{b}_1, \dots, \mathbf{b}_K)$  be a  $(Jd) \times K$  matrix of sieve coefficients,  $\Phi(\mathbf{W}) = (\phi(\mathbf{W}_1), \dots, \phi(\mathbf{W}_p))'$  be a  $p \times (Jd)$  matrix of basis functions, and  $\mathbf{R}(\mathbf{W})$  be a  $p \times K$  matrix with the  $(j, k)^{th}$  element  $\sum_{l=1}^d R_{kl}(W_{jl})$ . Then the matrix form (2.2) can be written as

$$\mathbf{X} = \Phi(\mathbf{W}) \mathbf{B} \mathbf{F}' + \mathbf{R}(\mathbf{W}) \mathbf{F}' + \boldsymbol{\Gamma} \mathbf{F}' + \mathbf{U}, \quad (3.4)$$

recalling that we drop the data source index  $i$ . Thus the residual term contains three parts: the sieve approximation error  $\mathbf{R}(\mathbf{W}) \mathbf{F}'$ , unexplained loading  $\boldsymbol{\Gamma} \mathbf{F}'$  and true signal  $\mathbf{U}$ .

The idea of Projected-PCA is simple: since the factor loadings are a function of the covariates in (3.4) and  $\mathbf{U}$  and  $\boldsymbol{\Gamma}$  are independent of  $\mathbf{W}$ , if we project (smooth) the observed data onto the space of  $\mathbf{W}$ , the effect of  $\mathbf{U}$  and  $\boldsymbol{\Gamma}$  will be significantly reduced and the problem becomes nearly a noiseless one, recalling that the approximation error  $\mathbf{R}(\mathbf{W})$  is small.

Define  $\mathbf{P}$  as the projection operator onto the space spanned by the basis functions of  $\mathbf{W}$ :

$$\mathbf{P} = \Phi(\mathbf{W})(\Phi(\mathbf{W})'\Phi(\mathbf{W}))^{-1}\Phi(\mathbf{W})'. \quad (3.5)$$

Then, by (3.4),  $\mathbf{P}\mathbf{X} \approx \mathbf{P}\Phi(\mathbf{W})\mathbf{B}\mathbf{F}' \approx \mathbf{G}(\mathbf{W})\mathbf{F}'$ . Thus,  $\mathbf{F}$  can be estimated from the ‘noiseless data’  $\mathbf{P}\mathbf{X}$ , using the traditional PCA. Let the columns of  $\tilde{\mathbf{F}}/\sqrt{n}$  be the eigenvectors corresponding to the top  $K$  eigenvalues of the  $n \times n$  matrix  $\mathbf{X}'\mathbf{P}\mathbf{X}$ , which is the sample covariance matrix of the projected data  $\mathbf{P}\mathbf{X}$ . Then,  $\tilde{\mathbf{F}}$  is the PPCA estimator of  $\mathbf{F}$ . It differs from the conventional PCA in that we use smoothed or projected data  $\mathbf{P}\mathbf{X}$ .

By the definition of  $\tilde{\mathbf{F}}$ , we have  $(np)^{-1}\mathbf{X}'\mathbf{P}\mathbf{X}\tilde{\mathbf{F}} = \tilde{\mathbf{F}}\mathbf{K}$  where  $\mathbf{K}$  is a  $K \times K$  diagonal matrix with the first  $K$  largest eigenvalues of  $(np)^{-1}\mathbf{X}'\mathbf{P}\mathbf{X}$  in descending order as its diagonal elements. Define the  $K$  by  $K$  matrix  $\mathbf{H}$  as in Fan et al. (2016):

$$\mathbf{H} = \frac{1}{np}\mathbf{B}'\Phi(\mathbf{W})'\Phi(\mathbf{W})\mathbf{B}\mathbf{F}'\tilde{\mathbf{F}}\mathbf{K}^{-1}.$$

It has been shown that  $\|\mathbf{K}\|$ ,  $\|\mathbf{K}^{-1}\|$  and  $\|\mathbf{H}\|$ ,  $\|\mathbf{H}^{-1}\|$  are all  $O_P(1)$ . Here we remind that though  $\mathbf{H}$  and  $\mathbf{K}$  are different from those in regime 1, they play essentially the same roles (thus with same notations).

As in Fan et al. (2016), we need the following conditions for the basis functions and accuracy of the sieve approximation.

**Assumption 3.1** (Basis functions). (i) There are  $d_{\min}$  and  $d_{\max} > 0$  so that almost surely,

$$d_{\min} < \lambda_{\min}(p^{-1}\Phi(\mathbf{W})'\Phi(\mathbf{W})) < \lambda_{\max}(p^{-1}\Phi(\mathbf{W})'\Phi(\mathbf{W})) < d_{\max}.$$

$$(ii) \max_{\nu \leq J, j \leq p, l \leq d} E\phi_{\nu}(W_{jl})^2 < \infty.$$

**Assumption 3.2** (Accuracy of sieve approximation). For each  $l \leq d, k \leq K$ ,

(i) The sieve coefficients  $\{b_{\nu,jl}\}_{\nu=1}^J$  satisfy:  $\exists \kappa \geq 4$ , as  $J \rightarrow \infty$ ,

$$\sup_{x \in \mathcal{X}_l} |g_{kl}(x) - \sum_{\nu=1}^J b_{\nu,kl}\phi_{\nu}(x)|^2 = O(J^{-\kappa}),$$

where  $\mathcal{X}_l$  is the support of the  $l^{th}$  element of  $\mathbf{W}_j$ , and  $J$  is the sieve dimension.

$$(ii) \max_{\nu,k,l} |b_{\nu,kl}| < \infty.$$

Condition (i) in Assumption 3.2 is satisfied by most commonly used basis. For example, when  $\{\phi_{\nu}\}$  is polynomial basis or B-splines, it is implied by the condition that smooth curve  $g_{kl}(\cdot)$  belongs to a Hölder class  $\mathcal{G}$ , defined by  $\mathcal{G} = \{g : |g^{(r)}(s) - g^{(r)}(t)| \leq L|s - t|^{\alpha}\}$  for

some  $L > 0$ , with  $\kappa = 2(r + \alpha) \geq 4$  (Lorentz, 2005; Chen, 2007). Another example is step function  $g_{kl}(\cdot)$  with finite many distinct values, which can be expressed exactly as the linear combination of disjoint indicator functions so that  $\kappa$  can be arbitrarily large.

With the above conditions, the following theorem provides all the rates we need, recalling the definition of  $\nu_p$  in Assumption 2.4 (iii).

**Theorem 3.3.** *Choose  $J = (p \min\{n, p, \nu_p^{-1}\})^{1/\kappa}$  and assume  $J^2 \phi_{\max}^2 \log(nJ) = O(p)$  where  $\phi_{\max} = \max_{\nu \leq J} \sup_{x \in \mathcal{X}} \phi_\nu(x)$ . Under Assumptions 2.1, 2.3, 2.4, 3.1 and 3.2, we have  $\|\Lambda\|_{\max} = O_P(J\phi_{\max} + \sqrt{\log p})$  and*

- (i)  $\|\tilde{\mathbf{F}} - \mathbf{FH}\|_F = O_P(\sqrt{n/p})$  and  $\|\tilde{\mathbf{F}} - \mathbf{FH}\|_{\max} = O_P(\sqrt{\log n/p})$ ;
- (ii)  $\|\mathbf{F}'(\tilde{\mathbf{F}} - \mathbf{FH})\|_{\max} = O_P(\sqrt{n/p} + n/p + n\sqrt{\nu_p/p})$ ;
- (iii)  $\|\mathbf{U}(\tilde{\mathbf{F}} - \mathbf{FH})\|_{\max} = O_P(\sqrt{n \log p/p} + nJ\phi_{\max}\|\Sigma\|_1/p)$ ;
- (iv)  $\|\mathbf{HH}' - \mathbf{I}\|_{\max} = O_P(1/p + 1/\sqrt{pn} + \sqrt{\nu_p/p})$ .

Combining the above theorem with Theorem 3.1, we obtain

$$\tilde{\mathbf{U}} - \mathbf{U} = -\frac{1}{n} \mathbf{UFF}' + \boldsymbol{\Pi},$$

where  $\|\boldsymbol{\Pi}\|_{\max} = O_P(\sqrt{\log n/p}(J\phi_{\max} + \sqrt{\log p}) + J\phi_{\max}\|\Sigma\|_1\sqrt{\log n/p})$  and

$$\tilde{\mathbf{U}}\tilde{\mathbf{U}}' - \mathbf{UU}' = -\frac{1}{n} \mathbf{UFF}'\mathbf{U}' + \boldsymbol{\Delta},$$

where  $\|\boldsymbol{\Delta}\|_{\max} = O_P(n\sqrt{\nu_p/p}(J^2\phi_{\max}^2 + \log p) + nJ\phi_{\max}\|\Sigma\|_1(J\phi_{\max} + \sqrt{\log p})/p + n^2J^2\phi_{\max}^2\|\Sigma\|_1^2/p^2)$  if there exists  $C$  s.t.  $\nu_p > C/n$ . We choose to keep  $\|\Sigma\|_1$  terms here although it makes a long presentation of the rate.

### 3.3 Specification test

In this section, we give an adaptive rule to decide whether the covariates  $\mathbf{W}$  are informative enough to use PPCA or just PCA. We test the hypothesis that  $H_0 : \mathbf{G}(\mathbf{W}) = 0$  using the test statistic (Fan et al., 2016)

$$S = \frac{1}{p} \text{tr}(\boldsymbol{\Xi} \hat{\boldsymbol{\Lambda}}' \mathbf{P} \hat{\boldsymbol{\Lambda}}) \text{ where } \boldsymbol{\Xi} = \left( \frac{1}{p} \hat{\boldsymbol{\Lambda}}' \hat{\boldsymbol{\Lambda}} \right)^{-1}.$$

Here, we use PCA estimator  $\hat{\boldsymbol{\Lambda}}$  as PPCA is not applicable under  $H_0$ . If  $\boldsymbol{\Lambda}$  has nothing to do with  $\mathbf{W}$ , then  $\mathbf{P}\hat{\boldsymbol{\Lambda}} \approx 0$  and  $S$  should be quite small after projection. Conversely, if  $\mathbf{G}(\mathbf{W}) \neq 0$ ,  $S$  will be large, hence we reject the null. We showed the following theorem, whose proof is omitted.

**Theorem 3.4.** Under all the assumptions discussed above, if  $\{\mathbf{W}_j, \boldsymbol{\gamma}_j\}_{j \leq p}$  are independent and identically distributed, as  $p, n_i, J \rightarrow \infty$ , we have under  $H_0$  for the  $i^{\text{th}}$  subgroup,

$$\frac{pS - JdK}{\sqrt{2JdK}} \xrightarrow{d} N(0, 1).$$

Based on this result, we can decide whether or not to reject the null hypothesis, namely to use PPCA or PCA. When the test is applied to all  $m$  data sources, it becomes a multiple testing problem. The thresholding can be chosen by using various false discovery rate control methods such as Benjamini-Hochberg method (Benjamini and Hochberg, 1995). If a hypothesis is rejected, we identify the subgroup as regime 2 and use Projected-PCA to obtain  $\tilde{\mathbf{U}}$ ; otherwise, we identify the subgroup as regime 1 and apply regular PCA to get  $\hat{\mathbf{U}}$ .

### 3.4 Estimating number of factors

We now address the problem of estimating the number of factors for two different regimes. Extensive literature has made contributions to this problem in regime 1, i.e. the regular factor model (Bai and Ng, 2002; Hallin and Liška, 2007; Ahn and Horenstein, 2013; Lam and Yao, 2012). Ahn and Horenstein (2013) and Lam and Yao (2012) proposed to use ratio of adjacent eigenvalues of  $\mathbf{X}'\mathbf{X}$  to infer the number of factors. They showed the estimator  $\hat{K} = \arg \max_{k \leq K_{\max}} \lambda_k(\mathbf{X}'\mathbf{X})/\lambda_{k+1}(\mathbf{X}'\mathbf{X})$  correctly identifies  $K$  with probability tending to 1, where  $K_{\max}$  can be a fixed prior upper bound for the number of factors.

For the genuine semiparametric factor model, in the recent work by Fan et al. (2016), they propose  $\tilde{K} = \arg \max_{k \leq K_{\max}} \lambda_k(\mathbf{X}'\mathbf{P}\mathbf{X})/\lambda_{k+1}(\mathbf{X}'\mathbf{P}\mathbf{X})$ . Here  $K_{\max}$  is of the same order as  $Jd$ , say  $K_{\max} = Jd/2$ . It was shown that  $\mathbb{P}(\tilde{K} = K) \rightarrow 1$  under assumptions we omit here. When we have genuine and pervasive covariates,  $\tilde{K}$  typically outperforms  $\hat{K}$ . More details can be found in Fan et al. (2016).

### 3.5 Summary of ALPHA

We now summarize the final procedure and convergence rates. We first divide  $m$  subgroups based on whether the collected covariates have influence on the loadings. Let

$$\mathcal{M}_1 = \{i \leq m \mid \mathbf{G} = 0\}, \quad \mathcal{M}_2 = \{i \leq m \mid \mathbf{G} \neq 0\}.$$

ALPHA consists of the following three steps.

Step 1: **(Preprocessing)** For data source  $i$ , estimate  $K^i$  by  $\hat{K}^i$ ; test  $H_0 : \mathbf{G}^i(\mathbf{X}^i) = 0$  by  $S^i$  and use FDR control to construct two groups. For rejected groups, refine  $\hat{K}^i$  by  $\tilde{K}^i$ .

Step 2: (**Adjustment**) Apply Projected-PCA to estimate  $\Lambda^i \mathbf{F}^{i'}$  if the test is rejected, otherwise use PCA to remove the heterogeneity, resulting in adjusted data  $\check{\mathbf{U}}$ , which is either  $\hat{\mathbf{U}}^i$  or  $\tilde{\mathbf{U}}^i$ .

Step 3: (**Aggregation**) Combine adjusted data  $\{\check{\mathbf{U}}^i\}_{i=1}^m$  to conduct further statistical analysis. For example, estimate sample covariance  $\Sigma$  by  $\widehat{\Sigma} = (N - \sum_i \widehat{K}_i)^{-1} \sum_{i=1}^m \check{\mathbf{U}}^i \check{\mathbf{U}}^{i'}$  where  $N = \sum_i n_i$  is the aggregated sample size; or estimate sparse precision matrix  $\Omega$  by existing graphical model methods say CLIME (Cai et al., 2011).

We summarize the ALPHA procedure in Algorithm 1 given in the appendix.

We also summarize the convergence of  $\hat{\mathbf{U}}^i$  and  $\tilde{\mathbf{U}}^i$  here. To ease presentation, we consider a typical regime in practice:  $n_i < Cp$ ,  $\sum_{i \leq m} K^i < CN$  for some constant  $C$ . Also we focus on the situation of sufficiently smooth curves  $\kappa = \infty$  so that  $J$  diverges very slowly (say with rate  $O(\sqrt{\log p})$ ) and constant  $\phi_{\max}, \nu_p$ . Based on discussions of the previous subsections, for estimation of  $\mathbf{U}$ , we have

$$\check{\mathbf{U}}^i - \mathbf{U}^i = -\mathbf{U}^i \mathbf{F}^i \mathbf{F}^{i'}/n_i + \begin{cases} O_P\left(\sqrt{\log n_i \log p/p} + \sqrt{\log n_i \log p/n_i}\right) & \text{if } i \in \mathcal{M}_1, \\ O_P\left(\sqrt{\log n_i \log p/p}\right) & \text{if } i \in \mathcal{M}_2. \end{cases}$$

Therefore, Projected-PCA dominates PCA as long as the effective covariates are provided. However,  $\mathbf{U}^i \mathbf{F}^i \mathbf{F}^{i'}/n_i$  dominates all the remaining terms so that  $\|\check{\mathbf{U}}^i - \mathbf{U}^i\|_{\max} = O_P(\|\mathbf{U}^i \mathbf{F}^i \mathbf{F}^{i'}/n_i\|_{\max}) = O_P(\sqrt{\log n_i \log p/n_i})$ .

In addition, for estimation of  $\mathbf{U} \mathbf{U}'$ , we have

$$\check{\mathbf{U}}^i \check{\mathbf{U}}^{i'} - \mathbf{U}^i \mathbf{U}^{i'} = -\mathbf{U}^i \mathbf{F}^i \mathbf{F}^{i'} \mathbf{U}^{i'}/n_i + \begin{cases} O_P\left(\log p + \delta\right) & \text{if } i \in \mathcal{M}_1, \\ O_P\left(n_i \log p \sqrt{\nu_p/p} + \delta\right) & \text{if } i \in \mathcal{M}_2, \end{cases} \quad (3.6)$$

where  $\delta = n_i^2 \|\Sigma\|_1^2 \log p/p^2$ , depending on  $\|\Sigma\|_1$ . If we consider a general  $\Sigma$  so that  $\|\Sigma\|_1$  can be as large as  $O(\sqrt{p})$ , then the rate for  $i \in \mathcal{M}_1$  is simplified to  $O_P((1 + n_i^2/p) \log p)$  while the rate for  $i \in \mathcal{M}_2$  is  $O_P((n_i/\sqrt{p} + n_i^2/p) \log p)$ . This illustrates the advantage of Projected-PCA since its convergence is faster. If we only consider very sparse covariance matrix so that  $\|\Sigma\|_1$  is bounded, we can simply drop the term  $\delta$  in both regimes. Then, regime 1 achieves better rate if  $p = O(n_i^2 \nu_p)$ , but regime 2 dominates otherwise.

## 4 Conditional Graphical Model

We have summarized the order of bias caused by adjusting heterogeneity for each data source in Section 3.5. Now we combine the adjusted data together for further statistical analysis. As an example, we study graph estimation under a Gaussian graphical model.

Assume  $\mathbf{u}_t^i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$  and consider the class of the precision matrices:

$$\mathcal{F}(s, R) = \left\{ \boldsymbol{\Omega} : \boldsymbol{\Omega} \succ \mathbf{0}, \|\boldsymbol{\Omega}\|_1 \leq R, \max_{1 \leq i \leq p} \sum_{j=1}^p \mathbb{1}(\Omega_{ij} \neq 0) \leq s \right\}. \quad (4.1)$$

To simplify the analysis, we assume  $R$  is fixed, but all the analysis can be easily extended to include growing  $R$ .

To estimate  $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$  via CLIME, we first need a covariance estimator as the input. We also assume here the number of factors is known, i.e., the exception probability of recovering  $K^i$  has been ignored for ease of discussion. Such an estimator is naturally given by

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{N - \sum_{i \leq m} K^i} \sum_{i=1}^m \check{\mathbf{U}}^i \check{\mathbf{U}}^{i'}. \quad (4.2)$$

Since the number of data sources is large, we focus on the typical case of diverging  $N$  and  $p$ .

### 4.1 Covariance estimation

Denote by  $\boldsymbol{\Sigma}_N$  the oracle sample covariance matrix i.e.  $\boldsymbol{\Sigma}_N = N^{-1} \sum_{i=1}^m \mathbf{U}^i \mathbf{U}^{i'}$ . We consider the difference of our proposed  $\widehat{\boldsymbol{\Sigma}}$  with  $\boldsymbol{\Sigma}_N$  in this subsection. The oracle estimator obviously attains the rate  $\|\boldsymbol{\Sigma}_N - \boldsymbol{\Sigma}\|_{\max} = O_P(\sqrt{\log p/N})$ .

Let  $\boldsymbol{\xi}_k^i = \mathbf{U}^i \bar{\mathbf{f}}_k^i / \sqrt{n_i}$  where  $\bar{\mathbf{f}}_k^i$  is the  $k^{th}$  column of  $\mathbf{F}^i$ . It is Gaussian distributed with mean zero and variance  $\boldsymbol{\Sigma}$ . Note that  $\boldsymbol{\xi}_k^i$  are iid with respect to  $k$  and  $i$ , using the assumption  $\mathbf{F}^{i'} \mathbf{F}^i / n_i = \mathbf{I}$ . By the standard concentration bound,

$$\left\| \sum_{i \leq m} \left( \frac{1}{n_i} \mathbf{U}^i \mathbf{F}^i \mathbf{F}^{i'} \mathbf{U}^{i'} - K^i \boldsymbol{\Sigma} \right) \right\|_{\max} = \left\| \sum_{i \leq m} \sum_{k \leq K^i} \left( \boldsymbol{\xi}_k^i \boldsymbol{\xi}_k^{i'} - \boldsymbol{\Sigma} \right) \right\|_{\max} = O_P\left(\sqrt{K^{tot} \log p}\right),$$

where  $K^{tot} = \sum_{i \leq m} K^i$ . Therefore, by (3.6), we have

$$\begin{aligned}\|\widehat{\Sigma} - \Sigma_N\|_{\max} &= \left\| \frac{N}{N - \sum_{i \leq m} K^i} \frac{1}{N} \sum_{i \leq m} \left( \check{\mathbf{U}}^i \check{\mathbf{U}}^{i'} - \mathbf{U}^i \mathbf{U}^{i'} + K^i \Sigma \right) \right. \\ &\quad \left. + \frac{\sum_{i \in \mathcal{M}} K^i}{N - \sum_{i \in \mathcal{M}} K^i} \left( \frac{1}{N} \sum_{i \leq m} \mathbf{U}^i \mathbf{U}^{i'} - \Sigma \right) \right\|_{\max} \\ &=: O_P(a_{m,N,p}),\end{aligned}\tag{4.3}$$

where  $a_{m,N,p} = \frac{|\mathcal{M}_1| \log p}{N} + \frac{N_2 \log p}{N} \sqrt{\frac{\nu_p}{p}} + \frac{\sqrt{K^{tot} \log p}}{N} + \frac{K^{tot}}{N} \sqrt{\frac{\log p}{N}}$  and  $N_2 = \sum_{i \in \mathcal{M}_2} n_i$ .

We now examine the difference of the ALPHA estimator from the oracle estimator for two specific cases. In the first case, we apply PCA to all data sources, i.e., all  $i \in \mathcal{M}_1$  and  $K^i$  is bounded. We then have  $a_{m,N,p} = m \log p / N$ . This rate is dominated by the oracle error rate  $\sqrt{\log p / N}$  if and only if  $m = O(\sqrt{N / \log p})$ . This means traditional PCA performs optimally for adjusting heterogeneity as long as the number of subgroups grows slower than the order of  $\sqrt{N / \log p}$ .

If we apply PPCA to all data sources, i.e.,  $i \in \mathcal{M}_2$  and  $K^i$  is bounded, then  $a_{m,N,p} = \sqrt{\nu_p / p} \log p + \sqrt{m \log p} / N$ . This rate is of smaller order than rate  $\sqrt{\log p / N}$  if  $p / \log p > CN$  for some constant  $C > 0$ . The advantage of using PPCA is that when  $n_i$  is bound so that  $m \asymp N$ , we can still achieve optimal rate of convergence so long as we have a large enough dimensionality at least of the order  $N$ .

## 4.2 Precision matrix estimation

In order to obtain an estimator for the sparse precision matrix from  $\widehat{\Sigma}$ , we apply the CLIME estimator proposed by Cai et al. (2011). For a given  $\widehat{\Sigma}$ , CLIME solves the following optimization problem:

$$\widehat{\Omega} = \arg \min_{\Omega} \|\Omega\|_{1,1} \quad \text{subject to} \quad \|\widehat{\Sigma} \Omega - \mathbf{I}\|_{\max} \leq \lambda, \tag{4.4}$$

where  $\|\Omega\|_{1,1} = \sum_{i,j \leq p} |\sigma_{ij}|$  and  $\lambda$  is a tuning parameter. Note that (4.4) can be solved column-wisely by linear programming. However, CLIME does not necessarily generate a symmetric matrix. We can simply symmetrize it by taking the one with minimal magnitude of  $\widehat{\sigma}_{ij}$  and  $\widehat{\sigma}_{ji}$ . The resulting matrix after symmetrization, still denoted as  $\widehat{\Omega}$  with a little bit abuse of notation, also attains good rate of convergence. In particular, we consider the sparse precision matrix class  $\mathcal{F}(s, C_0)$  in (4.1). The following lemma provides guarantee for recovering any sparse matrix  $\Omega \in \mathcal{F}(s, C_0)$ .

**Theorem 4.1.** Suppose  $\Omega \in \mathcal{F}(s, C_0)$  and  $\widehat{\Sigma}$  given by (4.2) attains the rate  $\|\widehat{\Sigma} - \Sigma_N\|_{\max} = O_p(a_{m,N,p})$  in (4.3) with  $\Sigma_N$  denoting oracle sample covariance matrix. Letting  $\tau_{m,N,p} = \sqrt{\log p/N} + a_{m,N,p}$  and  $\lambda \asymp \tau_{m,N,p}$ , we have

$$\|\widehat{\Omega} - \Omega\|_{\max} = O_p(\tau_{m,N,p}).$$

Furthermore,

$$\|\widehat{\Omega} - \Omega\|_1 = O_p(s\tau_{m,N,p}) \quad \text{and} \quad \|\widehat{\Omega} - \Omega\|_2 = O_p(s\tau_{m,N,p}).$$

The proof of the theorem can be found in the appendix. The theorem shows that CLIME has strong theoretical guarantee of convergence under different matrix norms. The rate of convergence has two parts, one corresponding to the minimax optimal rate (Yuan, 2010) while the other is due to the error caused by estimating the unknown factors under various situations. The discussions at the end of Section 4.1 suggests that the latter error is often negligible.

## 5 Numerical Studies

In this section, we first validate the theoretical results derived above through Monte Carlo simulations. Our purpose is to show that after heterogeneity adjustment, our proposed aggregated covariance estimator  $\widehat{\Sigma}$  approximates well the oracle sample covariance  $\Sigma_N$ , thereby leading to accurate estimation of the true covariance matrix  $\Sigma$  and precision matrix  $\Omega$ . We also compare the performance of Projected-PCA and regular PCA on heterogeneity adjustments under different asymptotic settings.

In addition, we analyze a real brain image data using the proposed procedure. The dataset to be analyzed is the ADHD-200 data (Biswal et al., 2010). It consists of rs-fMRI images of 688 subjects, of whom 491 are healthy and 197 are diagnosed with ADHD. We dropped 16 subjects (13 healthy, 3 diseased) in our analysis since their data contain missing values. Following Power et al. (2011), we divided the whole brain into 264 regions of interest (ROI,  $p = 264$ ), which are regarded as nodes in our graphical model. Each brain is scanned for multiple times with sample sizes ranging from 76 to 261 ( $76 \leq n_i \leq 261$ ). In each scan, we acquire the blood-oxygen-level dependent (BOLD) signal within each ROI. Here the heterogeneity among subjects arises from the difference in age, gender, handedness and IQ.

## 5.1 Preliminary analysis

To analyze the data, the first question is what external covariates  $\mathbf{W}_j^i$  are for each of the 264 regions. Ideally, we hope these covariates have pervasive power on explaining the batch effect, while bearing no association with the graph structure of  $\mathbf{u}_t$ . For the current data, we can construct such covariates from physical locations of the regions, since the level of batch effect is non-uniform over different locations of the brain when scanned in fMRI machines, and furthermore it has been widely acknowledged in biological study that spatial adjacency does not necessarily imply brain functional connectivity.

Here we simply split the 264 regions into 10 clusters ( $J = 10$ ) by the hierarchy clustering (Ward's minimum variance method) of their physical locations and use the categorical cluster indices as the covariates of the nodes. Note that the healthy and ADHD group share the same physical locations. The clustering result is shown in Figure 2 and the spatial locations of the 264 regions are shown in Figure 6 in 10 different colors. Black (middle), green (left) and blue (right) represent roughly the region of frontal lobe; gray (middle), pink (left) and magenta (right) occupy the region of parietal lobe; red (left) and orange (right) are in the area of occipital lobe; finally yellow (left) and navy (right) provide information about temporal lobe.

Other possible values for  $J$  could also be considered, but we do not want  $J$  too large to overfit the smooth loading functions. Note that here since the covariate  $W$  is one-dimensional ( $d = 1$ ) and discrete, the sieve basis functions are just indicator functions  $\mathbb{1}(w - 0.5 \leq W < w + 0.5)$  for  $w = 1, \dots, 10$ . We use the same covariates for all subjects.

The next question is whether the selected covariates can explain the loadings well. We implemented the specification tests described in Section 5 and find out the p-values for each subject. Most of the p-values are rather small (82.4% subjects in the healthy group and 79.0% subjects in the patient group have p-values smaller than  $10^{-3}$ ). We chose to control the FDR by Benjamini-Hochberg method (Benjamini and Hochberg, 1995) below the level of 1%. We discovered 425 healthy samples (91.4%) and 129 diseased samples (90.2%) rejecting the null, meaning that the selected covariates have significant explanatory powers on factor loadings of most subjects. We identified them as samples in the class  $\mathcal{M}_2$  and used Projected-PCA to estimate the heterogeneity effect. For those whose null hypotheses were not rejected, we classified them as individuals in the class  $\mathcal{M}_1$  and regular PCA was applied.

Based on which class each subject falls into, we employ the corresponding method to estimate the number of factors. We used  $K_{\max} = 5$ . The estimated number of factors for the two groups are summarized in Table 1.

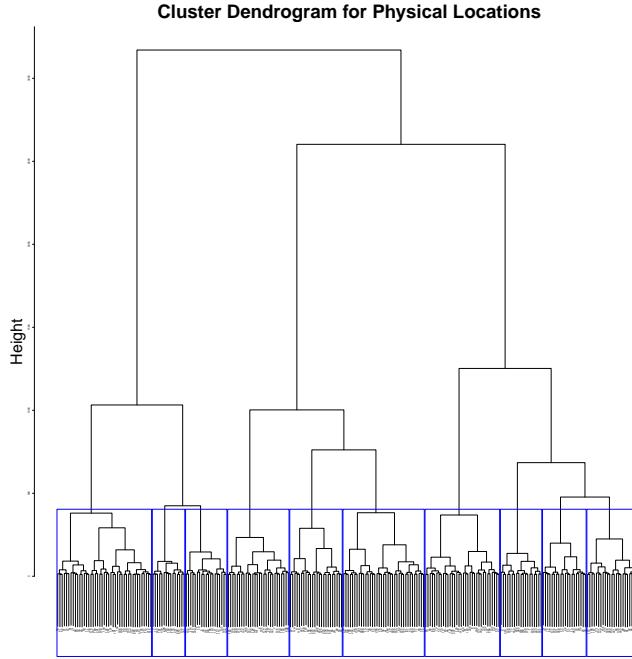


Figure 2: Cluster Dendrogram for  $J = 10$  for physical locations.

Table 1: The distributions of the estimated number of factors for healthy and ADHD groups

$\widehat{K}^i$	1	2	3	4	5
Healthy	227	126	59	31	22
ADHD	67	34	23	12	7

## 5.2 Synthetic datasets

In this simulation study, for stability, we use the first 15 subjects in the healthy group to calibrate the simulation models. The testing results reveal that the external covariates  $\mathbf{W}$  are informative for each of these 15 subjects. We specify four asymptotic settings for our simulation studies:

1.  $m = 500$ ,  $n_i = 10$  for  $i = 1, \dots, m$ ,  $p = 100, 200, \dots, 600$  and  $\mathbf{G}(\mathbf{W}) \neq 0$ ;
2.  $m = 100, 200, \dots, 1000$ ,  $n_i = 10$  for  $i = 1, \dots, m$ ,  $p = 264$  and  $\mathbf{G}(\mathbf{W}) \neq 0$ ;
3.  $m = 100$ ,  $n_i = 10, 20, \dots, 100$  for  $i = 1, \dots, m$ ,  $p = 264$  and  $\mathbf{G}(\mathbf{W}) \neq 0$ ;
4.  $m = 20, 40, \dots, 200$ ,  $n_i = 20, 40, \dots, 200$  for  $i = 1, \dots, m$ ,  $p = 264$  and  $\mathbf{G}(\mathbf{W}) = 0$ .

Here the last setting represents regime 1 with  $\mathbf{G}(\mathbf{W}) = 0$  where we should expect PCA to work well when the number of subjects is of order of square root of the total sample size, that is  $m \asymp \sqrt{N}$ . The first three settings represent regime 2 with informative covariates  $\mathbf{G}(\mathbf{W}) \neq 0$ ; they present asymptotics with growing  $p$ ,  $m$  and  $n_i$  respectively.

### 5.2.1 Model calibration and data generation

We calibrate (estimate) the covariance matrix  $\Sigma$  of  $\mathbf{u}_t$ , which is a 264 by 264 matrix, by our proposed method to the data in the healthy group. Plugging it as input in CLIME solver delivers a sparse precision matrix  $\Omega$ , which will be taken as truth in the simulation. Note that due to the regularization in CLIME,  $\Omega^{-1}$  is not the same as  $\Sigma$ . To obtain the covariance matrix used in setting 1, we also calibrate, using the same method, a sub-model that involves only the first 100 regions. We then copy this  $100 \times 100$  matrix multiple times to form a  $p \times p$  block diagonal matrix and used it for simulations in setting 1. We describe how we calibrate these ‘true models’ and generate data from the models as follows.

1. **(External covariates)** For each  $j \leq p$ , generate the external covariate  $W$  i.i.d. from the multinomial distribution with  $\mathbb{P}(W_j = s) = w_s, s \leq 10$  where  $\{w_s\}_{s=1}^{10}$  are calibrated with the hierarchy clustering results of the real data (Figure 2).
2. **(Calibration)** For the first 15 healthy subjects, obtain estimators for  $\mathbf{F}$ ,  $\mathbf{B}$  and  $\Gamma$  by PPCA, resulting in  $\tilde{\mathbf{F}}$ ,  $\tilde{\mathbf{B}} = n^{-1}(\Phi(\mathbf{W})'\Phi(\mathbf{W}))^{-1}\Phi(\mathbf{W})'\mathbf{X}\tilde{\mathbf{F}}$  and  $\tilde{\Gamma} = n^{-1}(\mathbf{I} - \mathbf{P})\mathbf{X}\tilde{\mathbf{F}}$  according to Fan et al. (2016). Use the rows of the estimated factors to fit a stationary VAR model  $\mathbf{f}_t = \mathbf{Af}_{t-1} + \epsilon_t$ , where  $\epsilon_t \sim N(0, \Sigma_\epsilon)$ , and obtain the estimators  $\tilde{\mathbf{A}}$  and  $\tilde{\Sigma}_\epsilon$ .
3. **(Simulation)** For each subject  $i \leq m$ , pick one of the 15 calibrated models and their associated parameters from above at random and do the following.
  - (a) Generate  $\gamma_{jk}^i$  i.i.d. from  $N(0, \tilde{\sigma}_\gamma^2)$  where  $\tilde{\sigma}_\gamma^2$  is the variance of all entries of  $\tilde{\Gamma}$ . For the first three settings, compute the ‘true’ loading matrix  $\Lambda^i = \Phi(\mathbf{W})\tilde{\mathbf{B}} + \Gamma^i$ . For the last setting, set  $\Lambda^i = \Gamma^i$  since  $\mathbf{G}(\mathbf{W}) = 0$ .
  - (b) Generate factors  $\mathbf{f}_t^i$  from the VAR model  $\mathbf{f}_t^i = \tilde{\mathbf{A}}\mathbf{f}_{t-1}^i + \epsilon_t$  with  $\epsilon_t \sim N(0, \tilde{\Sigma}_\epsilon)$ , where the parameters  $\tilde{\mathbf{A}}$  and  $\tilde{\Sigma}_\epsilon$  are taken from the fitted values in step 2.
  - (c) Finally, generate the observed data  $\mathbf{X}^i = \Lambda^i \mathbf{F}^{i'} + \mathbf{U}^i$ , where each column of  $\mathbf{U}^i$  is randomly sampled from  $N(\mathbf{0}, \Omega^{-1})$ , where  $\Omega$  has been calibrated by the CLIME solver as described at beginning of the section.

### 5.2.2 Estimation of $\Sigma$

In this subsection, we investigate the errors of estimating covariance of  $\mathbf{u}_t$  in max-norm after applying Projected-PCA or regular PCA for heterogeneity adjustment. We also compare them with the estimation errors if we naively pool all the data together without any heterogeneity adjustment, but the estimation errors for the first 3 cases are too large to fit

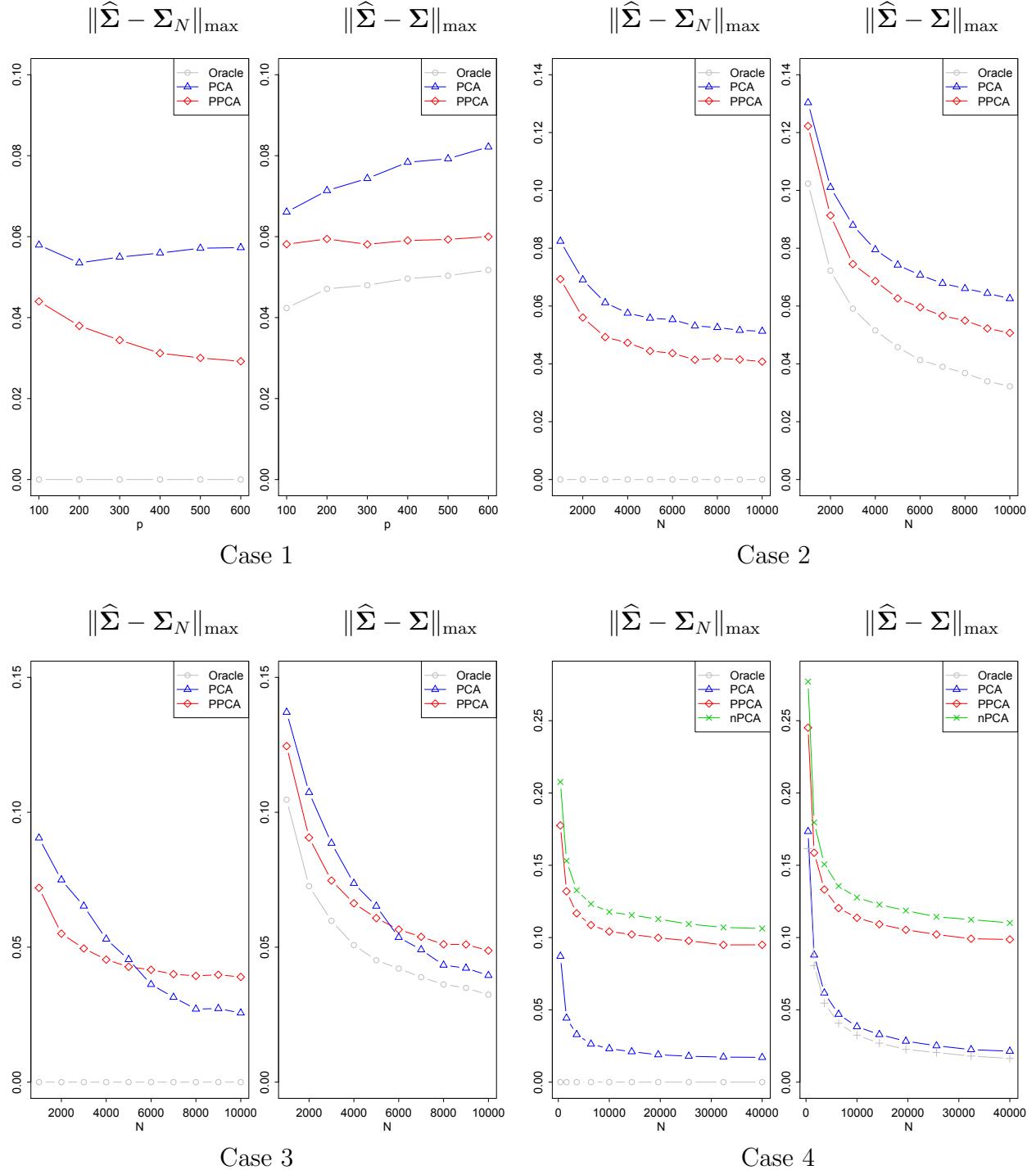


Figure 3: Estimation of  $\Sigma$  by PCA, PPCA and the oracle sample covariance matrix for 4 different settings. Case 1:  $m$  and  $n_i$  are fixed while the dimension  $p$  increases; case 2:  $n_i$  and  $p$  are fixed while  $m$  increases; case 3:  $m$  and  $p$  are fixed while  $n_i$  increases; case 4:  $p$  is fixed, and both  $m$  and  $n_i$  increase and conditions for PPCA are violated.

in the graph. Denote the oracle sample covariance of  $\mathbf{u}_t$  by  $\Sigma_N$  as before. The estimation errors under all four settings are presented in Figure 3, which are based on 100 simulations.

In Case 1,  $m$  and  $n_i$  are fixed while the dimension  $p$  increases. For this setting,  $n_i$  is small and this highlights more the advantages of Projected-PCA over regular PCA. From the left panel, we observe that increase of dimensionality improves the performance of Projected-PCA. This is consistent with the rate we derived in theories. In Case 2,  $n_i$  and  $p$  are fixed while  $m$  increases. Both Projected-PCA and regular PCA benefit from increasing number of subjects. However, since  $n_i$  is small, again Projected-PCA outperforms regular PCA. In Case 3,  $m$  and  $p$  are fixed while  $n_i$  increases. Again both methods achieve better estimation as  $n_i$  increases, but more importantly, regular PCA outperforms Projected-PCA when  $n_i$  is large enough. This is again consistent with our theories. As illustrated by Section 4.1, when  $m$  is fixed, PCA attains the convergence rate  $\|\widehat{\Sigma} - \Sigma\|_{\max} = O_P(\sqrt{\log p/N})$ , while Projected-PCA only achieves  $\|\widehat{\Sigma} - \Sigma\|_{\max} = O_P(\log p/\sqrt{p})$ , which is worse than PCA when  $p/\log p = o(N)$ . In Case 4,  $p$  is fixed, and both  $m$  and  $n_i$  increase. Note that the covariates have no explanation power at all, i.e., Condition 2.3 about pervasiveness does not hold so that PPCA is not applicable. As expected, adjusting by PCA behaves much better than by Projected-PCA, which can sometimes be as bad as ‘nPCA’, corresponding to no heterogeneity adjustment. This is not unexpected as we utilized a noisy external covariates.

### 5.2.3 Estimation of $\Omega$

In this subsection, we focus on estimation error of the precision matrix of  $\mathbf{u}_t$ . We plug  $\widehat{\Sigma}$ , obtained from data after adjusting for heterogeneity, into CLIME to get an estimator  $\widehat{\Omega}$  of  $\Omega$ . In Figure 4,  $\|\widehat{\Omega} - \Omega\|_{\max}$  and  $\|\widehat{\Omega} - \Omega\|_1$  are depicted under the same four asymptotic settings as before. From the plots we see  $\|\widehat{\Omega} - \Omega\|_{\max}$  and  $\|\widehat{\Omega} - \Omega\|_1$  share similar behavior with  $\|\widehat{\Sigma} - \Sigma\|_{\max}$  in all the four settings. In the first three cases, if we do not adjust data heterogeneity,  $\|\widehat{\Omega} - \Omega\|_{\max}$  and  $\|\widehat{\Omega} - \Omega\|_1$  will be too large to be fitted in the current plots.

We also present the ROC curves of our proposed methods in Figure 5, which is of interest to readers concerned with sparsity pattern recovery. The black dashed line is the 45 degree line connecting  $(0, 0)$  and  $(1, 1)$ , representing performance of the random guess. It is obvious from those plots that heterogeneity adjustment very much improves the sparsity recovery of the precision matrix  $\Omega$ . When the sample size of each subject is small, genuine pervasive covariates increase the power of Projected-PCA method while on the other hand if the sample size is relatively large, PCA is sufficiently good in recovering graph structures. Also notice that in all cases, the naive method with no heterogeneity adjustment can still achieve a certain amount of power, but we can improve the performance dramatically by correcting the batch effects.

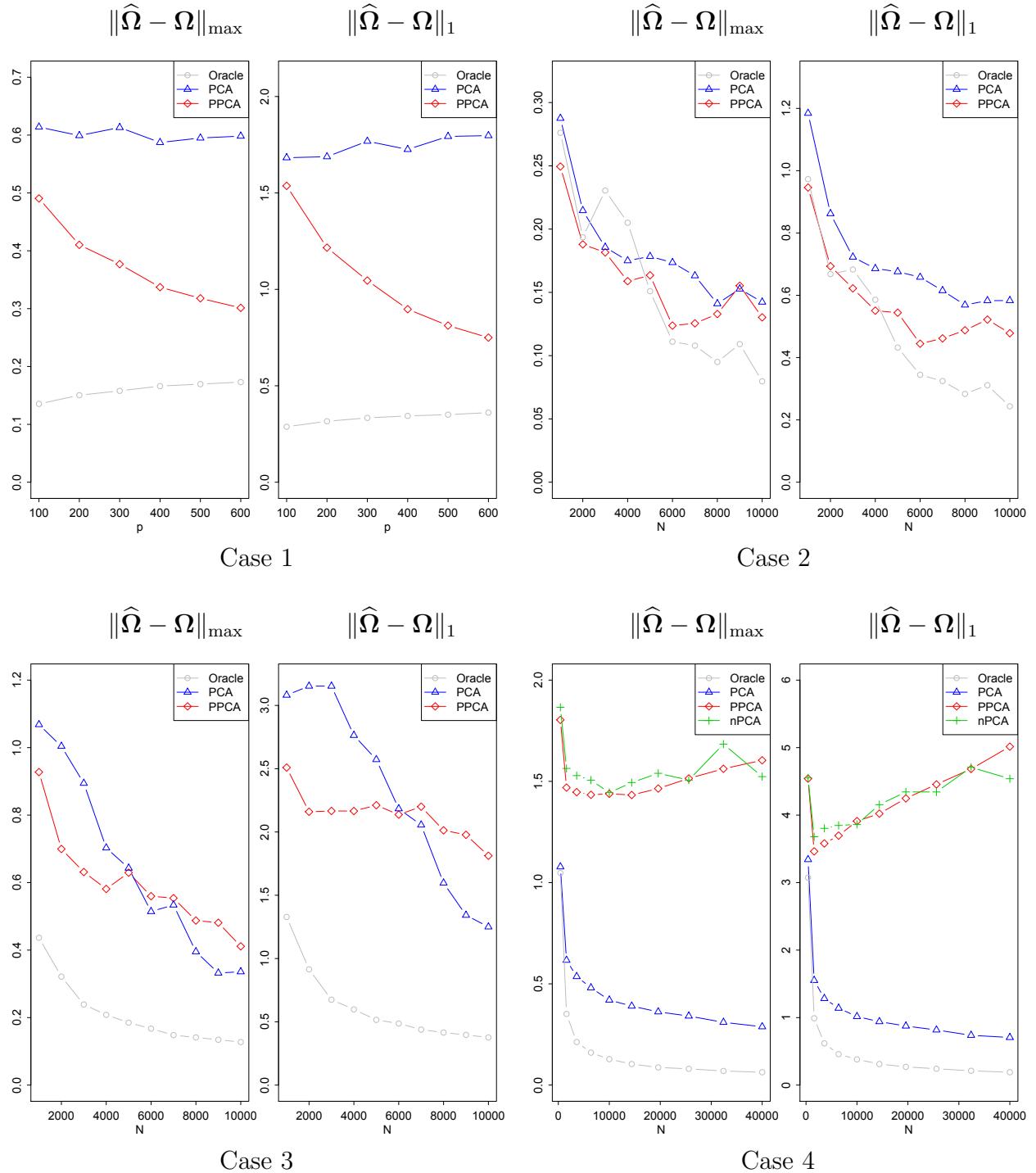


Figure 4: Estimation of  $\Omega$ . Presented are the estimation errors in max-norm and in  $L_1$ -norm for 4 different settings. The same captions in Figure 3 apply.

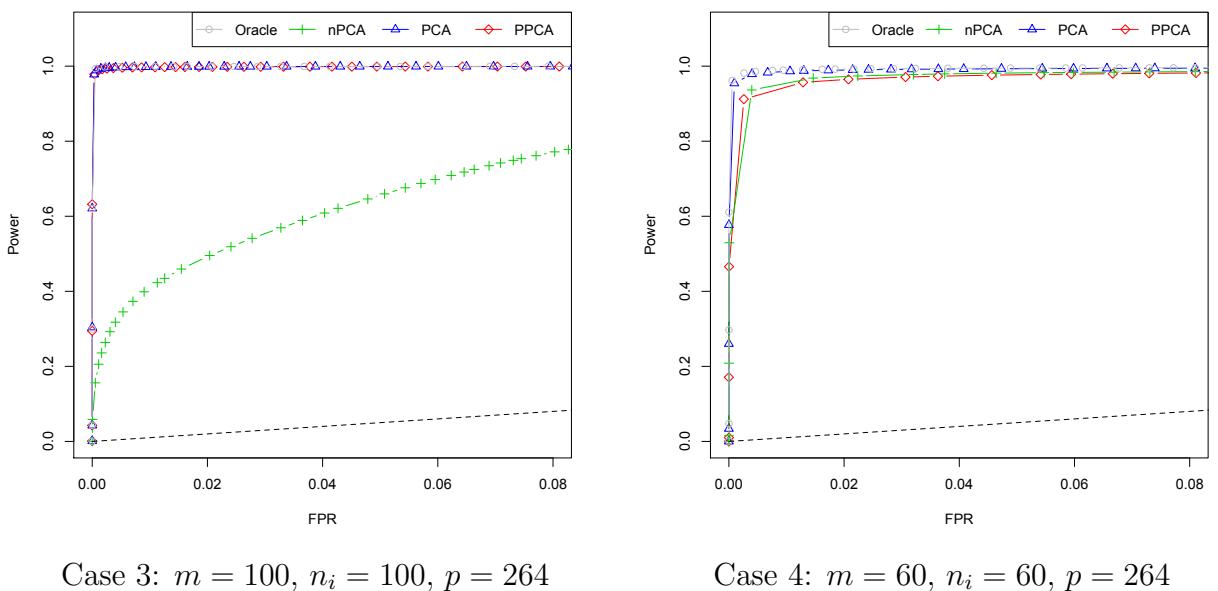
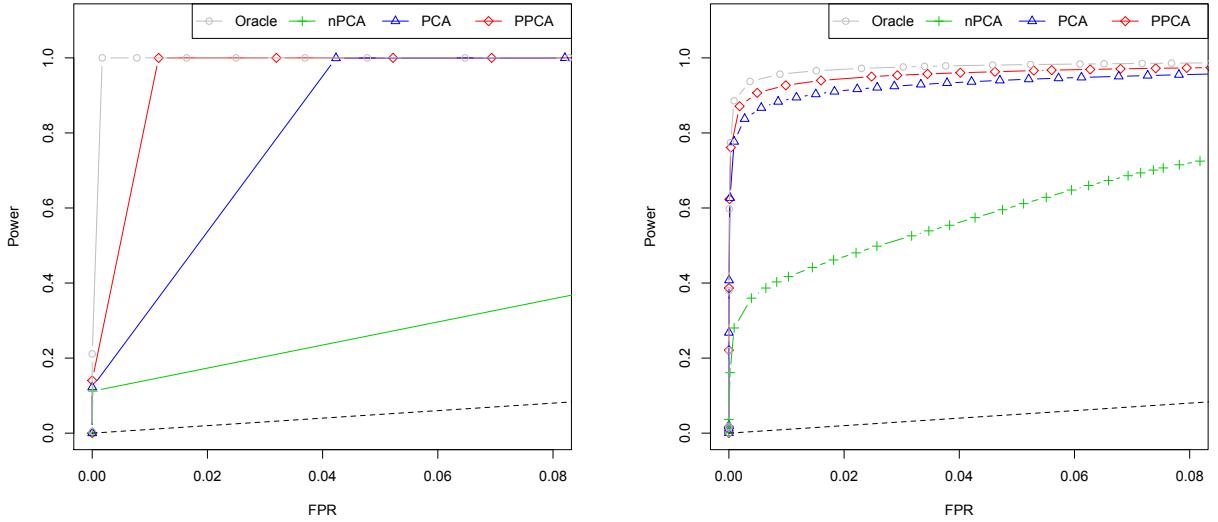


Figure 5: ROC curves for sparsity recovery of  $\Omega$  for 4 different settings. The captions in Figure 3 apply.

### 5.3 Brain image network data

We report the estimated graphs for both the healthy group and the ADHD patient group with batch effects removed using three methods: (1) PPCA using physical locations of ROI as covariates; (2) PCA without using any covariates; (3) no-PCA, which ignores heterogeneity

and naively pool the data from all subjects together. We took various sparsity levels of the networks from 1% to 5% (corresponding to the same set of  $\lambda$ 's for two groups) and selected the common edges, which are stable with respect to tuning, to be depicted.

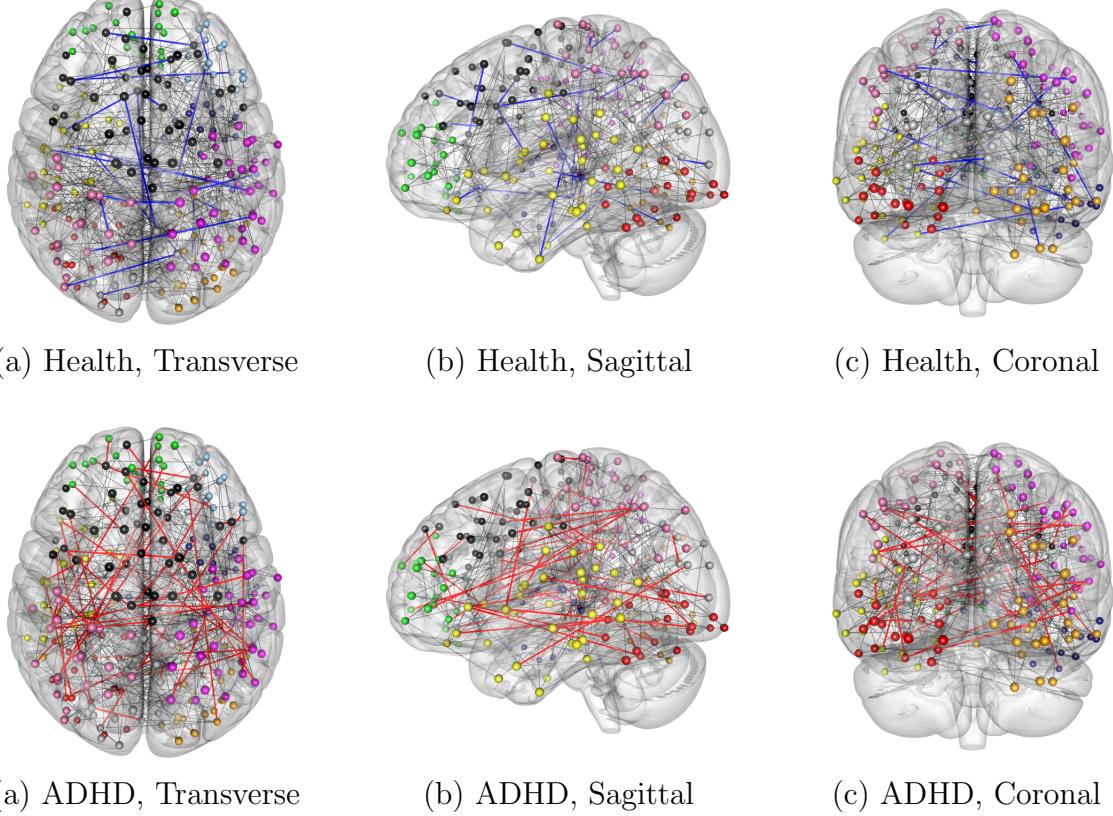


Figure 6: Estimated brain functional connectivity networks using physical locations as covariates to correct heterogeneity. 10 region clusters are labeled in 10 colors. Black, blue and red edges represent respectively common edges, unshared edges in the healthy group and in the ADHD group.

The brain network produced by Method 1 is reported in Figures 6. We omit the networks produced by Methods 2 and 3 since the inferred graphs actually do not differ too much, given that the number of subjects and total sample size are large. All methods give around 90% identical edges for the two networks, while respectively generating 8.6%, 8.0% and 11.6% unshared edges. Therefore, Methods 1 and 2 with adjustments provide more consistent and trustworthy graphs as batch effect brings more non-biological factors that exaggerate the difference. Preferences for Methods 1 and 2 should be based on relationship of  $p, n_i, m$  and whether the collected covariates are influential enough to explain the loadings. From Figure 6, it is obvious that the brain is more connected for the ADHD subjects, but the connections are weaker. Actually, the average correlation reduction for the estimated correlation matrices

Table 2: The degree of unshared edge vertices for each cluster

	red	orange	blue	green	yellow	navy	pink	black	magenta	gray
Health	0	5	4	0	6	2	7	9	4	5
ADHD	10	3	8	7	14	5	8	15	10	10
p-values (%)	2.8	5.8	90.8	6.7	85.6	85.2	20.4	53.1	78.9	89.7

(obtained from  $\widehat{\Omega}^{-1}$ ) of the two groups is 0.001; a paired t-test gives p-value < 2.2e–16. This is consistent with the recent finding that kids with ADHD show weaker interactions among brain networks (Cai et al., 2016).

In addition, we investigate how those unshared edges between two groups of people are distributed across the 10 clusters. We only focus on networks from Method 1. We are interested in which cluster contributes to the difference of the distributions of vertices of the unshared edges the most. We summarized the total degree of unshared edge vertices within each cluster in Table 2. For each column  $j$ , we consider the hypothesis testing for  $p_{jH} = p_{jD}$ , where  $H$  and  $D$  denotes the healthy and diseased group respectively, meaning that the unshared edges within the  $j^{th}$  cluster are found due to the same Bernoulli distribution. A simple chi-square test for the null was carried out for each column and those p-values are reported also in Table 2. The most noteworthy fact is that occipital lobe shows dependency change from right brain (orange) to left brain (red) for ADHD patients. The left frontal lobe (green) and the left parietal lobe (pink) have relatively large change in dependence structure compared with other parts of the brain. These are signs that ADHD is a complex disease that affects many regions of the brain. The general methodology we provide here could be valuable for further understanding the mechanism of the disease.

## 6 Discussions

In this paper, we developed a generic method called ALPHA that can consistently estimate and remove data heterogeneity and lead to effective subsequent statistical analysis on the true signal. The entire analysis relies on the pervasive assumption that most of the dimensions are corrupted by the heterogeneous factors. Future work may relax such pervasive conditions to allow for weaker signal batch effect, thus delivering more flexibility to recover the homogeneous residual.

As we have seen, ALPHA is adaptive to factor structures and is flexible to include external information. For brain image data analysis, previous literature rarely took physical locations into considerations. With the new framework, we can take advantage of external characteristics of the voxels or genes relevant to the batch effect, and consistently estimate

the pervasive heterogeneity term even with very limited samples. However, this advantage of Projected-PCA is accompanied by more assumptions and the practical issue of selecting proper basis functions and the number of them in sieve approximation. On the other hand, if no valuable covariates exist and the sample size is relatively large for each data source, we have shown conventional PCA is still an effective tool. Direct aggregation of less heterogeneous subgroups (say subjects with the same age and gender in the ADHD dataset) might also be helpful to increase the sample size.

Finally, note that after heterogeneity adjustment, the recovered residuals  $\check{\mathbf{U}}$  are not column-wisely independently distributed anymore. Statistical procedures that require assumptions of i.i.d. data cannot be directly applied on  $\check{\mathbf{U}}$ . However, the ALPHA procedure gives theoretical guarantee for  $\|\check{\mathbf{U}} - \mathbf{U}\|_{\max}$  and  $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max}$ , which serve as foundations for establishing the statistical properties of the subsequent procedure. In this sense, our framework is compatible with any statistical procedure that only requires an accurate estimator as the input, for instance, the CLIME procedure. Methods robust to small perturbations on the truth are preferable.

## References

- AHN, S. C. and HORENSTEIN, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica* **81** 1203–1227.
- ALTER, O., BROWN, P. O. and BOTSTEIN, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences* **97** 10101–10106.
- BAI, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica* **71** 135–171.
- BAI, J. and NG, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70** 191–221.
- BAI, J. and NG, S. (2013). Principal components estimation and identification of static factors. *Journal of Econometrics* **176** 18–29.
- BANERJEE, O., EL GHAOUI, L. and d'ASPREMONT, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research* **9** 485–516.

- BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 289–300.
- BISWAL, B. B., MENNES, M., ZUO, X.-N., GOHEL, S., KELLY, C., SMITH, S. M., BECKMANN, C. F., ADELSTEIN, J. S., BUCKNER, R. L. and COLCOMBE, S. (2010). Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences* **107** 4734–4739.
- CAI, T. T., LI, H., LIU, W. and XIE, J. (2012). Covariate-adjusted precision matrix estimation with an application in genetical genomics. *Biometrika* ass058.
- CAI, T. T., LIU, W. and LUO, X. (2011). A constrained  $\ell_1$  minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association* **106** 594–607.
- CAI, T. T., MA, Z. and WU, Y. (2013). Sparse PCA: Optimal rates and adaptive estimation. *The Annals of Statistics* **41** 3074–3110.
- CAI, W., CHEN, T., SZEGLETES, L., SUPEKAR, K. and MENON, V. (2016). Aberrant cross-brain network interaction in children with attention-deficit/hyperactivity disorder and its relation to attention deficits: A multisite and cross-site replication study. *Biological psychiatry (to appear)*.
- CANDÈS, E. J., LI, X., MA, Y. and WRIGHT, J. (2011). Robust principal component analysis? *Journal of the ACM (JACM)* **58** 11.
- CHEN, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics* **6** 5549–5632.
- CONNOR, G., HAGMANN, M. and LINTON, O. (2012). Efficient semiparametric estimation of the fama–french model and extensions. *Econometrica* **80** 713–754.
- CONNOR, G. and LINTON, O. (2007). Semiparametric estimation of a characteristic-based factor model of common stock returns. *Journal of Empirical Finance* **14** 694–717.
- FAN, J., LIAO, Y. and MINCHEVA, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75** 603–680.
- FAN, J., LIAO, Y. and WANG, W. (2016). Projected principal component analysis in factor models. *The Annals of Statistics* **44** 219–254.

- FAN, J. and WANG, W. (2015). Asymptotics of empirical eigen-structure for ultra-high dimensional spiked covariance model. *arXiv preprint arXiv:1502.04733* .
- FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics* **9** 432–441.
- HALLIN, M. and LIŠKA, R. (2007). Determining the number of factors in the general dynamic factor model. *Journal of the American Statistical Association* **102** 603–617.
- HSU, D., KAKADE, S. M. and ZHANG, T. (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab* **17**.
- JOHNSON, W. E., LI, C. and RABINOVIC, A. (2007). Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics* **8** 118–127.
- JOHNSTONE, I. M. and LU, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association* **104** 682–693.
- LAM, C. and FAN, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of Statistics* **37** 4254.
- LAM, C. and YAO, Q. (2012). Factor modeling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics* **40** 694–726.
- LEEK, J. T., SCHARPF, R. B., BRAVO, H. C., SIMCHA, D., LANGMEAD, B., JOHNSON, W. E., GEMAN, D., BAGGERLY, K. and IRIZARRY, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics* **11** 733–739.
- LEEK, J. T. and STOREY, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet* **3** 1724–1735.
- LIU, H., HAN, F. and ZHANG, C.-H. (2012). Transelliptical graphical models. In *Advances in Neural Information Processing Systems*.
- LIU, H., LAFFERTY, J. and WASSERMAN, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *The Journal of Machine Learning Research* **10** 2295–2328.

- LOH, P.-L. and WAINWRIGHT, M. J. (2013). Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. *The Annals of Statistics* **41** 3022–3049.
- LORENTZ, G. G. (2005). *Approximation of functions*, vol. 322. American Mathematical Soc.
- MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics* 1436–1462.
- NEGAHBAN, S. and WAINWRIGHT, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics* 1069–1097.
- ONATSCHI, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics* **168** 244–258.
- PAUL, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica* **17** 1617.
- POWER, J. D., COHEN, A. L., NELSON, S. M., WIG, G. S., BARNES, K. A., CHURCH, J. A., VOGEL, A. C., LAUMANN, T. O., MIEZIN, F. M. and SCHLAGGAR, B. L. (2011). Functional network organization of the human brain. *Neuron* **72** 665–678.
- RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G. and YU, B. (2011). High-dimensional covariance estimation by minimizing 1-penalized log-determinant divergence. *Electronic Journal of Statistics* **5** 935–980.
- RUDELSON, M. and VERSHYNIN, R. (2013). Hanson-wright inequality and sub-gaussian concentration. *Electron. Commun. Probab* **18**.
- SHEN, D., SHEN, H., ZHU, H. and MARRON, J. (2013). Surprising asymptotic conical structure in critical sample eigen-directions. Tech. rep.
- SHEN, X., PAN, W. and ZHU, Y. (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association* **107** 223–232.
- SIMS, A. H., SMETHURST, G. J., HEY, Y., OKONIEWSKI, M. J., PEPPER, S. D., HOWELL, A., MILLER, C. J. and CLARKE, R. B. (2008). The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets—improving meta-analysis and prediction of prognosis. *BMC medical genomics* **1** 42.

- STOCK, J. H. and WATSON, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American statistical association* **97** 1167–1179.
- YUAN, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *The Journal of Machine Learning Research* **11** 2261–2286.
- YUAN, M. and LIN, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika* **94** 19–35.

## APPENDIX

We first give the algorithm for our ALPHA procedure. We then outline the key ideas of the technical proofs of Theorems 3.1—3.3 and 4.1 in respectively the next four sections and leave additional proofs to the technical lemmas in Appendix F.

### A Algorithm for ALPHA

### B Proof of Theorem 3.1

*Proof.* By definition of  $\check{\mathbf{U}}$ ,  $\check{\mathbf{U}} = \mathbf{U}(\mathbf{I} - n^{-1}\mathbf{FF}') + n^{-1}\mathbf{X}(\check{\mathbf{F}}\check{\mathbf{F}}' - \mathbf{FF}')$ . We first look at the converge of  $\check{\mathbf{U}} - \mathbf{U}$ . Obviously  $\boldsymbol{\Pi} = n^{-1}\mathbf{X}(\check{\mathbf{F}}\check{\mathbf{F}}' - \mathbf{FF}') = I + II$  where

$$I = \frac{1}{n}\boldsymbol{\Lambda}\mathbf{F}'(\check{\mathbf{F}}\check{\mathbf{F}}' - \mathbf{FF}'), \quad II = \frac{1}{n}\mathbf{U}(\check{\mathbf{F}}\check{\mathbf{F}}' - \mathbf{FF}').$$

Since  $\mathbf{F}'(\check{\mathbf{F}}\check{\mathbf{F}}' - \mathbf{FF}') = \mathbf{F}'(\check{\mathbf{F}} - \mathbf{FH})\check{\mathbf{F}}' + n\mathbf{H}(\check{\mathbf{F}} - \mathbf{FH})' + n(\mathbf{HH}' - \mathbf{I})\mathbf{F}'$ , we have

$$\|I\|_{\max} = O_P(\|\boldsymbol{\Lambda}\|_{\max}(\|\mathbf{F}'(\check{\mathbf{F}} - \mathbf{FH})\|_{\max}\|\check{\mathbf{F}}/n\|_{\max} + \|\check{\mathbf{F}} - \mathbf{FH}\|_{\max} + \|\mathbf{HH}' - \mathbf{I}\|_{\max}\|\mathbf{F}\|_{\max})).$$

Similarly  $\mathbf{U}(\check{\mathbf{F}}\check{\mathbf{F}}' - \mathbf{FF}') = \mathbf{U}(\check{\mathbf{F}} - \mathbf{FH})\check{\mathbf{F}}' + \mathbf{UFH}(\check{\mathbf{F}} - \mathbf{FH})' + \mathbf{UF}(\mathbf{HH}' - \mathbf{I})\mathbf{F}'$ , so

$$\|II\|_{\max} = O_P(\|\mathbf{U}'(\check{\mathbf{F}} - \mathbf{FH})\|_{\max}\|\check{\mathbf{F}}/n\|_{\max} + \|\mathbf{UF}/n\|_{\max}(\|\check{\mathbf{F}} - \mathbf{FH}\|_{\max} + \|\mathbf{HH}' - \mathbf{I}\|_{\max}\|\mathbf{F}\|_{\max})).$$

According to Lemma F.4 (i),  $\|\mathbf{UF}/n\|_{\max} = O_P(1)$  and noting both  $\|\mathbf{F}\|_{\max}$  and  $\|\check{\mathbf{F}}\|_{\max}$  are  $O_P(\sqrt{n})$ , we conclude the result for  $\|\boldsymbol{\Pi}\|_{\max}$  easily.

---

**Algorithm 1** Algorithm for adaptive low-rank principal heterogeneity adjustment

---

**Input:** Panel  $\mathbf{X}_{p \times n_i}^i$  and  $d$ -dimensional  $\{\mathbf{W}_j^i\}_{j=1}^p$  from  $m$  data sources  
**Output:**  $\check{\mathbf{U}}^i$ , the adjusted estimator for  $\mathbf{U}^i$  and  $\hat{\Sigma}$

```

1: procedure ALPHA
2:   for each subject  $i \leq m$  do
3:      $\widehat{K}^i \leftarrow$  non-projected eigenvalue-ratio method
4:     test  $H_0 : \mathbf{G}^i(\mathbf{X}^i) = 0$  by  $S^i$ 
5:   end for
6:    $\mathcal{M}_1, \mathcal{M}_2 \leftarrow$  control FDR by Benjamini-Hochberg
7:   for each subject  $i \leq m$  do
8:     if  $i \in \mathcal{M}_1$  then
9:        $\widehat{\mathbf{F}}^i / \sqrt{n_i} \leftarrow$  eigenvectors of  $\mathbf{X}^{i'} \mathbf{X}^i$  corresponding to top  $\widehat{K}^i$  eigenvalues
10:       $\widehat{\Lambda}^i \leftarrow \mathbf{X}^i \widehat{\mathbf{F}}^i / n_i$ ,  $\widehat{\mathbf{U}}^i \leftarrow \mathbf{X}^i - \widehat{\Lambda}^i \widehat{\mathbf{F}}^i$  and  $\check{\mathbf{U}}^i \leftarrow \widehat{\mathbf{U}}^i$ 
11:    else
12:       $\widetilde{K}^i \leftarrow$  projected eigenvalue-ratio method
13:       $\mathbf{P}^i \leftarrow \Phi(\mathbf{W}^i)(\Phi(\mathbf{W}^i)' \Phi(\mathbf{W}^i))^{-1} \Phi(\mathbf{W}^i)'$ 
14:       $\widetilde{\mathbf{F}}^i / \sqrt{n_i} \leftarrow$  eigenvectors of  $\mathbf{X}^{i'} \mathbf{P}^i \mathbf{X}^i$  corresponding to top  $\widetilde{K}^i$  eigenvalues
15:       $\widetilde{\Lambda}^i \leftarrow \mathbf{X}^i \widetilde{\mathbf{F}}^i / n_i$ ,  $\widetilde{\mathbf{U}}^i \leftarrow \mathbf{X}^i - \widetilde{\Lambda}^i \widetilde{\mathbf{F}}^i$  and  $\check{\mathbf{U}}^i \leftarrow \widetilde{\mathbf{U}}^i$ 
16:    end if
17:  end for
18:   $\hat{\Sigma} \leftarrow (\sum_i n_i - \sum_i \widehat{K}_i)^{-1} \sum_{i=1}^m \check{\mathbf{U}}^i \check{\mathbf{U}}^{i'}$ 
19:  return  $\{\check{\mathbf{U}}^i\}_{i=1}^m$  and  $\hat{\Sigma}$ 
20: end procedure

```

---

Now we consider  $\check{\mathbf{U}} \check{\mathbf{U}}'$  in the following.

$$\begin{aligned} \check{\mathbf{U}} \check{\mathbf{U}}' &= \mathbf{U}(\mathbf{I} - n^{-1} \mathbf{FF}') \mathbf{U}' + n^{-1} \mathbf{U}(\mathbf{I} - n^{-1} \mathbf{FF}') (\check{\mathbf{F}} \check{\mathbf{F}}' - \mathbf{FF}') \mathbf{X}' + n^{-2} \mathbf{X} (\check{\mathbf{F}} \check{\mathbf{F}}' - \mathbf{FF}')^2 \mathbf{X}' \\ &=: \mathbf{U} \mathbf{U}' - \frac{1}{n} \mathbf{U} \mathbf{F} \mathbf{F}' \mathbf{U}' + III + IV. \end{aligned}$$

So  $\Delta = III + IV$  and it suffices to bound the two terms.

$$\begin{aligned} \|III\|_{\max} &= O_P(\|n^{-1} \mathbf{U}(\mathbf{I} - \mathbf{FF}' / n) \check{\mathbf{F}} \check{\mathbf{F}}' \mathbf{F}\|_{\max} \|\Lambda\|_{\max} + \|n^{-1} \mathbf{U}(\mathbf{I} - \mathbf{FF}' / n) \check{\mathbf{F}} \check{\mathbf{F}}' \mathbf{U}'\|_{\max}) \\ &=: O_P(\|\mathbf{J}_1\|_{\max} \|\Lambda\|_{\max} + \|\mathbf{J}_2\|_{\max}). \end{aligned}$$

Decompose  $\mathbf{J}_1$  by  $\mathbf{J}_1 = n^{-1} \mathbf{U}(\check{\mathbf{F}} - \mathbf{FH}) \check{\mathbf{F}}' \mathbf{F} - n^{-2} \mathbf{UF} \cdot \mathbf{F}' (\check{\mathbf{F}} - \mathbf{FH}) \check{\mathbf{F}}' \mathbf{F}$ . Therefore,

$$\|\mathbf{J}_1\|_{\max} = O_P(\|\mathbf{U}(\check{\mathbf{F}} - \mathbf{FH})\|_{\max} + n^{-1} \|\mathbf{UF}\|_{\max} \|\mathbf{F}'(\check{\mathbf{F}} - \mathbf{FH})\|_{\max}),$$

since  $\|\check{\mathbf{F}}' \mathbf{F} / n\|_{\max} \leq \|\check{\mathbf{F}}' \mathbf{F} / n\|_F \leq \|\check{\mathbf{F}}'\|_F \|\mathbf{F}\|_F / n = K$ . Similar to  $\mathbf{J}_1$ , we decompose  $\mathbf{J}_2$  only

replacing  $\check{\mathbf{F}}'\mathbf{F}$  with  $\check{\mathbf{F}}'\mathbf{U}'$ . According to Lemma F.4 (i),  $\|\check{\mathbf{F}}'\mathbf{U}'/n\|_{\max} = O_P(\|\mathbf{U}\mathbf{F}/n\|_{\max} + \|\mathbf{U}(\check{\mathbf{F}} - \mathbf{F}\mathbf{H})\|_{\max}) = O_P(1 + \|\mathbf{U}(\check{\mathbf{F}} - \mathbf{F}\mathbf{H})\|_{\max})$ , hence  $\|\mathbf{J}_2\|_{\max} = O_P(\|\mathbf{J}_1\|_{\max}(1 + \|\mathbf{U}(\check{\mathbf{F}} - \mathbf{F}\mathbf{H})\|_{\max}))$ . We then conclude that  $\|III\|_{\max} = O_P((\|\mathbf{U}(\check{\mathbf{F}} - \mathbf{F}\mathbf{H})\|_{\max} + n^{-1}\|\mathbf{U}\mathbf{F}\|_{\max}\|\mathbf{F}'(\check{\mathbf{F}} - \mathbf{F}\mathbf{H})\|_{\max})(\|\boldsymbol{\Lambda}\|_{\max} + \|\mathbf{U}(\check{\mathbf{F}} - \mathbf{F}\mathbf{H})\|_{\max}))$ .

Now let us take a look at  $IV$ .  $\|IV\|_{\max} = \|\mathbf{D}_1 + \mathbf{D}_2 + \mathbf{D}'_2 + \mathbf{D}_3\|_{\max}$  where

$$\begin{aligned}\mathbf{D}_1 &= n^{-2}\boldsymbol{\Lambda}\mathbf{F}'(\check{\mathbf{F}}\check{\mathbf{F}}' - \mathbf{F}\mathbf{F}')^2\mathbf{F}\boldsymbol{\Lambda}' = \boldsymbol{\Lambda}(n\mathbf{I} - n^{-1}\mathbf{F}'\check{\mathbf{F}}'\check{\mathbf{F}}\mathbf{F})\boldsymbol{\Lambda}', \\ \mathbf{D}_2 &= n^{-2}\mathbf{U}(\check{\mathbf{F}}\check{\mathbf{F}}' - \mathbf{F}\mathbf{F}')^2\mathbf{F}\boldsymbol{\Lambda}' = -n^{-2}\mathbf{U}\mathbf{F}\mathbf{F}'(\check{\mathbf{F}}\check{\mathbf{F}}' - \mathbf{F}\mathbf{F}')\mathbf{F}\boldsymbol{\Lambda}' \\ \mathbf{D}_3 &= n^{-2}\mathbf{U}(\check{\mathbf{F}}\check{\mathbf{F}}' - \mathbf{F}\mathbf{F}')^2\mathbf{U}'.\end{aligned}$$

By assumption,  $\|\mathbf{H}\|_{\max} \leq \|\mathbf{H}\| = O_P(1)$ . Simple decompositions of  $\mathbf{D}_1$  gives

$$\|\mathbf{D}_1\|_{\max} = O_P((\|\mathbf{F}'(\check{\mathbf{F}} - \mathbf{F}\mathbf{H})\|_{\max} + n\|\mathbf{H}\mathbf{H}' - \mathbf{I}\|_{\max})\|\boldsymbol{\Lambda}\|_{\max}^2).$$

Since  $\mathbf{D}_2 = -n^{-2}\mathbf{U}\mathbf{F}\mathbf{F}'(\check{\mathbf{F}} - \mathbf{F}\mathbf{H})\check{\mathbf{F}}'\mathbf{F}\boldsymbol{\Lambda}' - n^{-1}\mathbf{U}\mathbf{F}\mathbf{H}(\check{\mathbf{F}} - \mathbf{F}\mathbf{H})'\mathbf{F}\boldsymbol{\Lambda}' - \mathbf{U}\mathbf{F}(\mathbf{H}\mathbf{H}' - \mathbf{I})\boldsymbol{\Lambda}'$ , we have

$$\|\mathbf{D}_2\|_{\max} = O_P(\|\mathbf{U}\mathbf{F}/n\|_{\max}\|\mathbf{D}_1\|_{\max}) = O_P(\|\mathbf{D}_1\|_{\max}).$$

It is also not hard to show  $\|\mathbf{D}_3\|_{\max} = O_P(\|III\|_{\max} + \|\mathbf{D}_1\|_{\max})$ . Under both Theorems 3.2 and 3.3 (replacing  $\check{\mathbf{F}}$  by  $\hat{\mathbf{F}}$  for regime 1 and  $\tilde{\mathbf{F}}$  for regime 2), we can check the following relationship holds:

$$n^{-1}\|\mathbf{U}\mathbf{F}\|_{\max}\|\mathbf{U}(\check{\mathbf{F}} - \mathbf{F}\mathbf{H})\|_{\max} = O_P(\|\boldsymbol{\Lambda}\|_{\max}^2).$$

Therefore we have

$$\begin{aligned}\|\boldsymbol{\Delta}\|_{\max} = \|III + IV\|_{\max} &= O_P(\|\mathbf{U}(\check{\mathbf{F}} - \mathbf{F}\mathbf{H})\|_{\max}\|\boldsymbol{\Lambda}\|_{\max} + \|\mathbf{U}(\check{\mathbf{F}} - \mathbf{F}\mathbf{H})\|_{\max}^2 \\ &\quad + \|\mathbf{F}'(\check{\mathbf{F}} - \mathbf{F}\mathbf{H})\|_{\max}\|\boldsymbol{\Lambda}\|_{\max}^2 + n\|\mathbf{H}\mathbf{H}' - \mathbf{I}\|_{\max}\|\boldsymbol{\Lambda}\|_{\max}^2).\end{aligned}$$

□

## C Proof of Theorem 3.2

### C.1 Convergence of factors $\widehat{\mathbf{F}}$

Let  $\mathbf{K}$  denote the  $K \times K$  diagonal matrix consisting of the first  $K$  largest eigenvalues of  $(pn)^{-1}\mathbf{X}'\mathbf{X}$  in descending order. By the definition of eigenvalues, we have

$$\frac{1}{np}(\mathbf{X}'\mathbf{X})\widehat{\mathbf{F}} = \widehat{\mathbf{F}}\mathbf{K}.$$

Recall  $\mathbf{H} = (np)^{-1}\Lambda'\Lambda\mathbf{F}'\widehat{\mathbf{F}}\mathbf{K}^{-1}$ . Substituting  $\mathbf{X} = \Lambda\mathbf{F}' + \mathbf{U}$ , we have,

$$\widehat{\mathbf{F}} - \mathbf{FH} = \left( \sum_{i=1}^3 \mathbf{E}_i \right) \mathbf{K}^{-1}, \quad (\text{C.1})$$

$$\mathbf{E}_1 = \frac{1}{np}\mathbf{F}\Lambda'\mathbf{U}\widehat{\mathbf{F}}, \quad \mathbf{E}_2 = \frac{1}{np}\mathbf{U}'\Lambda\mathbf{F}'\widehat{\mathbf{F}}, \quad \mathbf{E}_3 = \frac{1}{np}\mathbf{U}'\mathbf{U}\widehat{\mathbf{F}}.$$

To bound  $\|\widehat{\mathbf{F}} - \mathbf{FH}\|_{\max}$ , note that there is a constant  $C > 0$ , so that

$$\|\widehat{\mathbf{F}} - \mathbf{FH}\|_{\max} \leq C\|\mathbf{K}^{-1}\|_2 \sum_{i=1}^3 \|\mathbf{E}_i\|_{\max}.$$

Hence we need to bound  $\|\mathbf{E}_i\|_{\max}$  for  $i = 1, 2, 3$  since  $\|\mathbf{K}^{-1}\|_2 = O_P(1)$ . The following lemma gives the stochastic bounds for each individual term.

**Lemma C.1.** (i)  $\|\mathbf{E}_1\|_F = O_P(\sqrt{n/p}) = \|\mathbf{E}_2\|_F$ ,  $\|\mathbf{E}_3\|_F = O_P(1/\sqrt{n} + 1/\sqrt{p} + \sqrt{n/p})$ .  
(ii)  $\|\mathbf{E}_1\|_{\max} = O_P(\sqrt{\log n/p}) = \|\mathbf{E}_2\|_{\max}$ ,  $\|\mathbf{E}_3\|_{\max} = O_P(1/\sqrt{p} + \sqrt{\log n}/n)$ .

*Proof.* (i) Obviously  $\|\mathbf{E}_1\|_F \leq p^{-1}\|\Lambda'\mathbf{U}\|_F = O_P(\sqrt{n/p})$  according to Lemma F.1.  $\|\mathbf{E}_2\|_F$  attains the same rate. In addition,  $\|\mathbf{E}_3\|_F \leq n^{-1/2}p^{-1}\|\mathbf{U}'\mathbf{U}\|_F = O_P(1 + \sqrt{n/p})$  again according to Lemma F.1. So combining the three terms, we have  $\|\widehat{\mathbf{F}} - \mathbf{FH}\|_F = O_P(1 + \sqrt{n/p})$ . We now refine the bound for  $\|\mathbf{E}_3\|_F$ .  $\|\mathbf{E}_3\|_F \leq (np)^{-1}(\|\mathbf{U}'\mathbf{U}\|_F\|\mathbf{H}\|_F + \|\mathbf{U}'\mathbf{U}\|_F\|\widehat{\mathbf{F}} - \mathbf{FH}\|_F) = O_P(1/\sqrt{n} + 1/\sqrt{p} + \sqrt{n/p})$ . Then the refined rate of  $\|\widehat{\mathbf{F}} - \mathbf{FH}\|_F$  is  $O_P(\sqrt{n/p} + 1/\sqrt{n})$ .

(ii) Since  $\|\Lambda'\mathbf{U}\widehat{\mathbf{F}}\|_F = O_P(n\sqrt{p})$  by Lemma F.1,

$$\|\mathbf{E}_1\|_{\max} = O_P((np)^{-1}\|\mathbf{F}\|_{\max}\|\Lambda'\mathbf{U}\widehat{\mathbf{F}}\|_F) = O_P(\sqrt{\log n/p}).$$

$\|\mathbf{E}_2\|_{\max}$  is bounded by  $p^{-1}\|\mathbf{U}'\Lambda\|_{\max} = O_P(\sqrt{\log n/p})$  while  $\|\mathbf{E}_3\|_{\max}$  is bounded by

$$O_P\left((np)^{-1}(\|\mathbf{U}'\mathbf{U}\|_{\max} + \sqrt{n}\|\mathbf{U}'\mathbf{U}\|_{\max}\|\widehat{\mathbf{F}} - \mathbf{FH}\|_F)\right),$$

which based on results of Lemma F.2 and (i) is  $O_P(1/\sqrt{p} + \sqrt{\log n}/n)$ .  $\square$

The final rate of convergence for  $\|\widehat{\mathbf{F}} - \mathbf{FH}\|_{\max}$  and  $\|\widehat{\mathbf{F}} - \mathbf{FH}\|_F$  are summarized as follows.

**Proposition C.1.**

$$\|\widehat{\mathbf{F}} - \mathbf{FH}\|_{\max} = O_P\left(\sqrt{\frac{\log n}{p}} + \frac{\sqrt{\log n}}{n}\right) \text{ and } \|\widehat{\mathbf{F}} - \mathbf{FH}\|_F = O_P\left(\sqrt{\frac{n}{p}} + \frac{1}{\sqrt{n}}\right). \quad (\text{C.2})$$

*Proof.* The results follow from Lemmas C.1.  $\square$

## C.2 Rates of $\|\mathbf{F}'(\widehat{\mathbf{F}} - \mathbf{FH})\|_{\max}$ and $\|\mathbf{HH}' - \mathbf{I}\|_{\max}$

Note first that the two matrices under consideration is both  $K$  by  $K$ , so we do not lose rates bounding them by their Frobenius norm.

Let us find out rate for  $\|\mathbf{F}'(\widehat{\mathbf{F}} - \mathbf{FH})\|_F$ . Basically we need to bound  $\|\mathbf{F}'\mathbf{E}_i\|_F$  for  $i = 1, 2, 3$ . Firstly

$$\|\mathbf{F}'\mathbf{E}_1\|_F = p^{-1}\|\Lambda' \mathbf{U} \widehat{\mathbf{F}}\|_F \leq p^{-1}(\|\Lambda' \mathbf{U} \mathbf{F}\|_F \|\mathbf{H}\|_F + \|\Lambda' \mathbf{U}\|_F \|\widehat{\mathbf{F}} - \mathbf{FH}\|_F).$$

Since  $\|\Lambda' \mathbf{U} \mathbf{F}\|_F = O_P(\sqrt{np})$  and  $\|\Lambda' \mathbf{U}\|_F = O_P(\sqrt{np})$  by Lemma F.1, we have  $\|\mathbf{F}'\mathbf{E}_1\|_F = O_P(\sqrt{n/p} + n/p)$ . Secondly,

$$\|\mathbf{F}'\mathbf{E}_2\|_F \leq p^{-1}\|\mathbf{F}'\mathbf{U}'\Lambda\|_F = O_P(\sqrt{n/p}).$$

Finally,

$$\|\mathbf{F}'\mathbf{E}_3\|_F = O_P\left(\frac{1}{np}\|\mathbf{U} \mathbf{F}\|_F^2 + \frac{1}{np}\|\mathbf{F}'\mathbf{U}'\mathbf{U}\|_F \|\widehat{\mathbf{F}} - \mathbf{FH}\|_F\right) = O_P(1 + \sqrt{n/p}).$$

So combining three terms we have  $\|\mathbf{F}'(\widehat{\mathbf{F}} - \mathbf{FH})\|_{\max} \leq \|\mathbf{F}'(\widehat{\mathbf{F}} - \mathbf{FH})\|_F = O_P(1 + \sqrt{n/p})$ .

Now we bound  $\|\mathbf{HH}' - \mathbf{I}\|_F$ . Since  $\mathbf{H}'\mathbf{H} = n^{-1}(\mathbf{FH} - \widehat{\mathbf{F}})' \mathbf{FH} + n^{-1}\widehat{\mathbf{F}}'(\mathbf{FH} - \widehat{\mathbf{F}}) + \mathbf{I}$ , we have

$$\|\mathbf{H}'\mathbf{H} - \mathbf{I}\|_F = O_P\left(\frac{1}{n}\|\mathbf{F}'(\widehat{\mathbf{F}} - \mathbf{FH})\|_F + \frac{1}{n}\|\widehat{\mathbf{F}} - \mathbf{FH}\|_F^2\right) = O_P\left(\frac{1}{n} + \frac{1}{p}\right).$$

Therefore  $\|\mathbf{HH}' - \mathbf{I}\|_F$  has the same rate since  $\|\mathbf{HH}' - \mathbf{I}\|_F \leq \|\mathbf{H}\|_F \|\mathbf{H}'\mathbf{H} - \mathbf{I}\|_F \|\mathbf{H}^{-1}\|_F$ . So  $\|\mathbf{HH}' - \mathbf{I}\|_{\max} = O_P(1/n + 1/p)$ .

### C.3 Rate of $\|\mathbf{U}(\widehat{\mathbf{F}} - \mathbf{FH})\|_{\max}$

In order to study rate of  $\|\mathbf{U}(\widehat{\mathbf{F}} - \mathbf{FH})\|_{\max}$ , we essentially need to bound  $\|\mathbf{UE}_i\|_{\max}$  for  $i = 1, 2, 3$ . We handle each term separately.

$$\|\mathbf{UE}_1\|_{\max} = O_P\left(\frac{1}{np}\|\mathbf{UF}\|_{\max}\|\Lambda' \mathbf{U}\widehat{\mathbf{F}}\|_F\right) = O_P\left(\frac{1}{n}\|\mathbf{UF}\|_{\max}\|\mathbf{F}'\mathbf{E}_1\|_F\right) = O_P\left(\sqrt{\frac{\log p}{p}} + \frac{\sqrt{n \log p}}{p}\right).$$

By Lemma F.5,  $\|\mathbf{UU}'\Lambda\|_{\max} = O_P(\sqrt{np \log p} + n\|\Sigma\|_1)$ . Therefore,

$$\|\mathbf{UE}_2\|_{\max} = O_P\left(\frac{1}{p}\|\mathbf{UU}'\Lambda\|_{\max}\right) = O_P\left(\frac{n\|\Sigma\|_1}{p} + \sqrt{\frac{n \log p}{p}}\right).$$

From bounding  $\|\mathbf{E}_3\|_F$ , the last term has rate

$$\|\mathbf{UE}_3\|_{\max} = \frac{1}{np}\|\mathbf{UU}'\mathbf{U}\widehat{\mathbf{F}}\|_{\max} \leq \frac{1}{\sqrt{np}}\|\mathbf{U}\|_{\max}\|\mathbf{U}'\mathbf{U}\widehat{\mathbf{F}}\|_F = O_P((1 + n/p)\sqrt{\log p}).$$

So combining three terms, we conclude  $\|\mathbf{U}(\widehat{\mathbf{F}} - \mathbf{FH})\|_{\max} = O_P((1 + n/p)\sqrt{\log p} + n\|\Sigma\|_1/p)$ .

## D Proof of Theorem 3.3

### D.1 Convergence of factors $\widetilde{\mathbf{F}}$

Let  $\mathbf{K}$  denote the  $K \times K$  diagonal matrix consisting of the first  $K$  largest eigenvalues of  $(pn)^{-1}\mathbf{X}'\mathbf{P}\mathbf{X}$  in descending order. By the definition of eigenvalues, we have

$$\frac{1}{np}(\mathbf{X}'\mathbf{P}\mathbf{X})\widetilde{\mathbf{F}} = \widetilde{\mathbf{F}}_-.$$

Recall  $\mathbf{H} = (np)^{-1}\mathbf{B}'\Phi(\mathbf{W})'\Phi(\mathbf{W})\mathbf{BF}'\widetilde{\mathbf{F}}\mathbf{K}^{-1}$ . Substituting  $\mathbf{X} = \Phi(\mathbf{W})\mathbf{BF}' + \mathbf{R}(\mathbf{W})\mathbf{F}' + \mathbf{\Gamma}\mathbf{F}' + \mathbf{U}$ , we have,

$$\widetilde{\mathbf{F}} - \mathbf{FH} = \left(\sum_{i=1}^{15} \mathbf{A}_i\right)\mathbf{K}^{-1} \tag{D.1}$$

where  $\mathbf{A}_i, i \leq 3$  has nothing to do with  $\mathbf{R}(\mathbf{W})$  and  $\mathbf{\Gamma}$ :

$$\mathbf{A}_1 = \frac{1}{np}\mathbf{FB}'\Phi(\mathbf{W})'\mathbf{U}\widetilde{\mathbf{F}}, \quad \mathbf{A}_2 = \frac{1}{np}\mathbf{U}'\Phi(\mathbf{W})\mathbf{BF}'\widetilde{\mathbf{F}}, \quad \mathbf{A}_3 = \frac{1}{np}\mathbf{U}'\mathbf{P}\mathbf{U}\widetilde{\mathbf{F}};$$

$\mathbf{A}_i, 3 \leq i \leq 8$  takes care of terms involving  $\mathbf{R}(\mathbf{W})$ :

$$\begin{aligned}\mathbf{A}_4 &= \frac{1}{np} \mathbf{F} \mathbf{B}' \Phi(\mathbf{W})' \mathbf{R}(\mathbf{W}) \mathbf{F}' \tilde{\mathbf{F}}, \quad \mathbf{A}_5 = \frac{1}{np} \mathbf{F} \mathbf{R}(\mathbf{W})' \Phi(\mathbf{W}) \mathbf{B} \mathbf{F}' \tilde{\mathbf{F}}, \\ \mathbf{A}_6 &= \frac{1}{np} \mathbf{F} \mathbf{R}(\mathbf{W})' \mathbf{P} \mathbf{R}(\mathbf{W}) \mathbf{F}' \tilde{\mathbf{F}}, \quad \mathbf{A}_7 = \frac{1}{np} \mathbf{F} \mathbf{R}(\mathbf{W})' \mathbf{P} \mathbf{U} \tilde{\mathbf{F}}, \quad \mathbf{A}_8 = \frac{1}{np} \mathbf{U}' \mathbf{P} \mathbf{R}(\mathbf{W}) \mathbf{F}' \tilde{\mathbf{F}};\end{aligned}$$

the remaining are terms involving  $\mathbf{\Gamma}$ :

$$\begin{aligned}\mathbf{A}_9 &= \frac{1}{np} \mathbf{F} \mathbf{B}' \Phi(\mathbf{W})' \mathbf{T} \mathbf{F}' \tilde{\mathbf{F}}, \quad \mathbf{A}_{10} = \frac{1}{np} \mathbf{F} \mathbf{\Gamma}' \Phi(\mathbf{W}) \mathbf{B} \mathbf{F}' \tilde{\mathbf{F}}, \quad \mathbf{A}_{11} = \frac{1}{np} \mathbf{F} \mathbf{\Gamma}' \mathbf{P} \mathbf{T} \mathbf{F}' \tilde{\mathbf{F}}, \\ \mathbf{A}_{12} &= \frac{1}{np} \mathbf{F} \mathbf{\Gamma}' \mathbf{P} \mathbf{U} \tilde{\mathbf{F}}, \quad \mathbf{A}_{13} = \frac{1}{np} \mathbf{U}' \mathbf{P} \mathbf{\Gamma} \mathbf{F}' \tilde{\mathbf{F}}, \quad \mathbf{A}_{14} = \frac{1}{np} \mathbf{F} \mathbf{R}' \mathbf{P} \mathbf{\Gamma} \mathbf{F}' \tilde{\mathbf{F}}, \quad \mathbf{A}_{15} = \frac{1}{np} \mathbf{F} \mathbf{\Gamma}' \mathbf{P} \mathbf{R} \mathbf{F}' \tilde{\mathbf{F}}.\end{aligned}$$

To bound  $\|\tilde{\mathbf{F}} - \mathbf{F} \mathbf{H}\|_{\max}$ , as in Theorem 3.2 we only need to bound  $\|\mathbf{A}_i\|_{\max}$  for  $i = 1, \dots, 15$  since again we have  $\|\mathbf{K}^{-1}\|_2 = O_P(1)$ . The following lemma gives the rate for each term.

**Lemma D.1.** (i)  $\|\mathbf{A}_1\|_{\max} = O_P(\sqrt{\log n/p}) = \|\mathbf{A}_2\|_{\max}$ ,  
(ii)  $\|\mathbf{A}_3\|_{\max} = O_P(J \phi_{\max} \sqrt{\log(nJ)/p})$ ,  
(iii)  $\|\mathbf{A}_4\|_{\max} = O_P(J^{-\kappa/2} \sqrt{\log n}) = \|\mathbf{A}_5\|_{\max}$  and  $\|\mathbf{A}_9\|_{\max} = O_P(\sqrt{\nu_p \log n/p}) = \|\mathbf{A}_{10}\|_{\max}$ ,  
(iv)  $\|\mathbf{A}_6\|_{\max} = O_P(J^{-\kappa} \sqrt{\log n})$  and  $\|\mathbf{A}_{11}\|_{\max} = O_P(J \nu_p \sqrt{\log n/p})$ ,  
(v)  $\|\mathbf{A}_7\|_{\max} = O_P(\phi_{\max} \sqrt{p^{-1} J^{1-\kappa} \log(nJ) \log n}) = \|\mathbf{A}_8\|_{\max}$   
and  $\|\mathbf{A}_{12}\|_{\max} = O_P(J \phi_{\max} \sqrt{\nu_p \log(nJ) \log n/p}) = \|\mathbf{A}_{13}\|_{\max}$ ,  
(vi)  $\|\mathbf{A}_{14}\|_{\max} = O_P(\sqrt{p^{-1} J^{1-\kappa} \nu_p \log n}) = \|\mathbf{A}_{15}\|_{\max}$ .

*Proof.* (i) Because  $\|\mathbf{F}\|_{\max} = O_P(\sqrt{\log n})$ ,  $\|\tilde{\mathbf{F}}\|_F = O_P(\sqrt{n})$ . By Lemmas F.3 and F.4,  $\|\mathbf{U}' \Phi(\mathbf{W}) \mathbf{B}\|_F = O_P(\sqrt{pn})$  and  $\|\mathbf{U}' \Phi(\mathbf{W}) \mathbf{B}\|_{\max} = O_P(\sqrt{p \log n})$ . Hence

$$\|\mathbf{A}_1\|_{\max} \leq \frac{\sqrt{K}}{np} \|\mathbf{F}\|_{\max} \|\mathbf{B}' \Phi(\mathbf{W})' \mathbf{U}\|_F \|\tilde{\mathbf{F}}\|_F = O_P(\sqrt{\log n/p}),$$

$$\|\mathbf{A}_2\|_{\max} \leq \frac{\sqrt{K}}{np} \|\mathbf{U}' \Phi(\mathbf{W}) \mathbf{B}\|_{\max} \|\mathbf{F}\|_F \|\tilde{\mathbf{F}}\|_F = O_P(\sqrt{\log n/p}).$$

(ii) We have  $\mathbf{A}_3 = \frac{1}{np} \mathbf{U}' \Phi(\mathbf{W}) (\Phi(\mathbf{W})' \Phi(\mathbf{W}))^{-1} \Phi(\mathbf{W})' \mathbf{U} \tilde{\mathbf{F}}$ . By Lemma F.3 and F.4,  $\|\mathbf{U}' \Phi(\mathbf{W})\|_F = O_P(\sqrt{npJ})$  and  $\|\mathbf{U}' \Phi(\mathbf{W})\|_{\max} = O_P(\phi_{\max} \sqrt{p \log(nJ)})$ . By Assumption 3.1,  $\|(\Phi(\mathbf{W})' \Phi(\mathbf{W}))^{-1}\|_2 = O_P(p^{-1})$ . Note the fact that for matrix  $\mathbf{A}_{m \times n}$ ,  $\mathbf{B}_{n \times n}$ ,  $\mathbf{C}_{n \times r}$ ,  $\|\mathbf{ABC}\|_{\max} = \max_{i \leq m, k \leq r} |\mathbf{a}_i' \mathbf{B} \mathbf{c}_k| \leq \sqrt{n} \|\mathbf{A}\|_{\max} \|\mathbf{B}\|_2 \|\mathbf{C}\|_F$ . So

$$\begin{aligned}\|\mathbf{A}_3\|_{\max} &\leq \frac{\sqrt{Jd}}{np} \|\mathbf{U}' \Phi(\mathbf{W})\|_{\max} \|(\Phi(\mathbf{W})' \Phi(\mathbf{W}))^{-1}\|_2 \|\Phi(\mathbf{W})' \mathbf{U}\|_F \|\tilde{\mathbf{F}}\|_F \\ &= O_P(J \phi_{\max} \sqrt{\log(nJ)/p}).\end{aligned}$$

(iii) Note that  $\|\Phi(\mathbf{W})\mathbf{B}\|_2 \leq \|\mathbf{G}(\mathbf{W})\|_2 + \|\mathbf{R}(\mathbf{W})\|_2 = O_P(\sqrt{p})$ , and  $\|\mathbf{R}(\mathbf{W})\|_{\max} = O_P(J^{-\kappa/2})$ . Hence we have  $\|\mathbf{B}'\Phi(\mathbf{W})'\mathbf{R}(\mathbf{W})\|_{\max} \leq \|\mathbf{B}'\Phi(\mathbf{W})'\|_1 \|\mathbf{R}(\mathbf{W})\|_{\max} \leq \sqrt{p} \|\mathbf{B}'\Phi(\mathbf{W})'\|_2 \|\mathbf{R}(\mathbf{W})\|_{\max} = O_p(pJ^{-\kappa/2})$ . Thus

$$\|\mathbf{A}_4\|_{\max} \leq \frac{K^{3/2}}{np} \|\mathbf{F}\|_{\max} \|\mathbf{B}'\Phi(\mathbf{W})'\mathbf{R}(\mathbf{W})\|_{\max} \|\mathbf{F}\tilde{\mathbf{F}}\|_F = O_P(J^{-\kappa/2} \sqrt{\log n}).$$

Similarly,  $\|\mathbf{A}_5\|_{\max}$  attains the same rate of convergence.

In addition, notice  $\mathbf{A}_9, \mathbf{A}_{10}$  have similar representation as  $\mathbf{A}_4, \mathbf{A}_5$ . The only difference is to replace  $\mathbf{R}$  by  $\mathbf{\Gamma}$ . It is not hard to see  $\|\mathbf{B}'\Phi'\mathbf{\Gamma}\|_{\max} = O_P(\sqrt{p\nu_p})$ . Therefore  $\|\mathbf{A}_9\|_{\max} = O_P(\sqrt{\nu_p \log n/p}) = \|\mathbf{A}_{10}\|_{\max}$ .

(iv) Note that  $\|\mathbf{P}\|_2 = \|(\Phi(\mathbf{W})'\Phi(\mathbf{W}))^{-1/2}\Phi(\mathbf{W})'\Phi(\mathbf{W})(\Phi(\mathbf{W})'\Phi(\mathbf{W}))^{-1/2}\|_2 = 1$  and  $\|\mathbf{R}(\mathbf{W})'\mathbf{P}\mathbf{R}(\mathbf{W})\|_{\max} \leq p \|\mathbf{R}(\mathbf{W})\|_{\max}^2 \|\mathbf{P}\|_2 = O_p(pJ^{-\kappa})$ . Hence

$$\|\mathbf{A}_6\|_{\max} \leq \frac{K}{np} \|\mathbf{F}\|_{\max} \|\mathbf{R}(\mathbf{W})'\mathbf{P}\mathbf{R}(\mathbf{W})\|_{\max} \|\mathbf{F}\tilde{\mathbf{F}}\|_F = O_P(J^{-\kappa} \sqrt{\log n}).$$

$\mathbf{A}_{11}$  has similar representation as  $\mathbf{A}_6$ . Since  $\|\mathbf{\Gamma}'\mathbf{P}\mathbf{\Gamma}\|_{\max} \leq \|\Phi'\mathbf{\Gamma}\|_F^2 \|(\Phi'\Phi)^{-1}\|_2 = O_P(J\nu_p)$ , we have  $\|\mathbf{A}_{11}\|_{\max} = O_P(J\nu_p \sqrt{\log n}/p)$ .

(v) According to Lemma F.4,  $\|\mathbf{U}'\Phi(\mathbf{W})\|_{\max} = O_P(\phi_{\max} \sqrt{p \log(nJ)})$ . Thus

$$\begin{aligned} \|\mathbf{A}_7\|_{\max} &\leq \frac{K}{\sqrt{np}} \|\mathbf{F}\|_{\max} \|\tilde{\mathbf{F}}\|_F \|\mathbf{R}'\Phi(\Phi'\Phi)^{-1}\Phi'\mathbf{U}\|_{\max} \\ &\leq O_p(p^{-1} \sqrt{J \log n}) \|\mathbf{R}'\Phi\|_F \|(\Phi'\Phi)^{-1}\|_2 \|\Phi'\mathbf{U}\|_{\max} = O_p\left(\phi_{\max} \sqrt{\frac{J \log(nJ) \log n}{p J^\kappa}}\right), \end{aligned}$$

since  $\|\mathbf{R}'\Phi\|_F \leq \|\mathbf{R}\|_F \|\Phi\|_2 = O_P(pJ^{-\kappa/2})$ . The rate of convergence for  $\mathbf{A}_8$  can be bounded in the same way. So do  $\mathbf{A}_{12}$  and  $\mathbf{A}_{13}$ . Given that  $\|\mathbf{\Gamma}'\Phi\|_F = O_P(pJ\nu_p)$ , we have  $\|\mathbf{A}_{12}\|_{\max} = O_P(J\phi_{\max} \sqrt{\nu_p \log(nJ) \log n}/p) = \|\mathbf{A}_{13}\|_{\max}$ .

(vi) Obviously,  $\|\mathbf{A}_{14}\|_{\max} = O_P(p^{-1} \sqrt{\log n} \|\mathbf{R}'\mathbf{P}\mathbf{\Gamma}\|_{\max})$  and  $\|\mathbf{R}'\mathbf{P}\mathbf{\Gamma}\|_{\max} \leq \|\mathbf{R}'\Phi\|_F \|(\Phi'\Phi)^{-1}\| \|\Phi'\mathbf{\Gamma}\|_F$ . We conclude  $\|\mathbf{A}_{14}\|_{\max} = O_P(\sqrt{p^{-1} J^{1-\kappa} \nu_p \log n})$ . Same bound holds for  $\mathbf{A}_{15}$ .  $\square$

The final rate of convergence for  $\|\tilde{\mathbf{F}} - \mathbf{FH}\|_{\max}$  and  $\|\tilde{\mathbf{F}} - \mathbf{FH}\|_F$  are summarized as follows.

**Proposition D.1.** Choose  $J = (p \min(n, p, \nu_p^{-1}))^{1/\kappa}$  and assume  $J^2 \phi_{\max}^2 \log(nJ) = O(p)$  and  $\nu_p = O(1)$ ,

$$\|\tilde{\mathbf{F}} - \mathbf{FH}\|_{\max} = O_P\left(\sqrt{\frac{\log n}{p}}\right) \text{ and } \|\tilde{\mathbf{F}} - \mathbf{FH}\|_F = O_P\left(\sqrt{\frac{n}{p}}\right). \quad (\text{D.2})$$

*Proof.* The max norm result follows from Lemmas D.1 and (D.1), while the Frobenius norm result has been shown in Fan et al. (2016).  $\square$

## D.2 Rates of $\|\mathbf{F}'(\tilde{\mathbf{F}} - \mathbf{FH})\|_{\max}$ and $\|\mathbf{HH}' - \mathbf{I}\|_{\max}$

Note first that the two matrices under consideration is both  $K$  by  $K$ , so we do not lose rates bounding them by their Frobenius norm.

It has been proved in Fan et al. (2016) that  $\|\mathbf{F}'(\tilde{\mathbf{F}} - \mathbf{FH})\|_F = O_P(\sqrt{n/p} + n/p + n\sqrt{\nu_p/p} + nJ^{-\kappa/2})$ . By the choice of  $J$ , the last term vanishes. So

$$\|\mathbf{F}'(\tilde{\mathbf{F}} - \mathbf{FH})\|_{\max} \leq \|\mathbf{F}'(\tilde{\mathbf{F}} - \mathbf{FH})\|_F = O_P(\sqrt{n/p} + n/p + n\sqrt{\nu_p/p}).$$

Fan et al. (2016) also showed that  $\|\mathbf{H}'\mathbf{H} - \mathbf{I}\|_F = O_P(1/p + 1/\sqrt{pn} + J^{-\kappa/2} + \sqrt{\nu_p/p})$ . Since  $\|\mathbf{H}\|$  and  $\|\mathbf{H}^{-1}\|$  are both  $O_P(1)$ , we easily show  $\|\mathbf{HH}' - \mathbf{I}\|_{\max} \leq \|\mathbf{HH}' - \mathbf{I}\|_F \leq \|\mathbf{H}\|\|\mathbf{H}'\mathbf{H} - \mathbf{I}\|_F\|\mathbf{H}^{-1}\| = O_P(1/p + 1/\sqrt{pn} + \sqrt{\nu_p/p})$  since  $J^\kappa \geq p/\nu_p$ .

## D.3 Rate of $\|\mathbf{U}(\tilde{\mathbf{F}} - \mathbf{FH})\|_{\max}$

By (D.1), in order to bound  $\|\mathbf{U}(\tilde{\mathbf{F}} - \mathbf{FH})\|_{\max}$ , we essentially need to bound  $\|\mathbf{UA}_i\|_{\max}$  for  $i = 1, \dots, 15$ . We do not bother going into the details of each term again as in Lemma D.1. However, we point out the difference here. All  $\mathbf{A}_i$  are separated into two types: the ones starting with  $\mathbf{F}$  and the ones starting with  $\mathbf{U}$ .

If a term  $\mathbf{A}_i$  starts with  $\mathbf{F}$ , say  $\mathbf{A}_i = \mathbf{FQ}$ , in Lemma D.1, we bound  $\|\mathbf{A}_i\|_{\max}$  using  $\sqrt{K}\|\mathbf{F}\|_{\max}\|\mathbf{Q}\|_F$ . Now we use bound  $\|\mathbf{UA}_i\|_{\max} \leq \sqrt{K}\|\mathbf{UF}\|_{\max}\|\mathbf{Q}\|_F$  so that we obtain all related rates by just changing rate  $\|\mathbf{F}\|_{\max} = O_P(\sqrt{\log n})$  to  $\|\mathbf{UF}\|_{\max} = O_P(\sqrt{n \log p})$ .

Terms starting with  $\mathbf{U}$  includes  $\mathbf{A}_i, i = 2, 3, 8, 13$ . In Lemma D.1, we bound  $\|\mathbf{A}_i\|_{\max}, i = 3, 8, 13$  using  $\|\mathbf{U}'\Phi\|_{\max}$  while we bound  $\|\mathbf{A}_2\|_{\max}$  using  $\|\mathbf{U}'\Phi\mathbf{B}\|_{\max}$ . Correspondingly now we need to control  $\|\mathbf{UU}'\Phi\|_{\max}$  and  $\|\mathbf{UU}'\Phi\mathbf{B}\|_{\max}$  separately to update the rates. The derivation is relegated to Lemma F.5. We have  $\|\mathbf{UU}'\Phi(\mathbf{W})\|_{\max} = O_P(\phi_{\max}(\sqrt{np \log p} + n\|\Sigma\|_1))$  and  $\|\mathbf{UU}'\Phi(\mathbf{W})\mathbf{B}\|_{\max} = O_P(\sqrt{np \log p} + nJ\phi_{\max}\|\Sigma\|_1)$ .

So we replace the corresponding terms in Lemma D.1. It is not hard to see the dominating term is  $\|\mathbf{UA}_2\|_{\max} = O_P(\sqrt{n \log p/p} + nJ\phi_{\max}\|\Sigma\|_1/p)$ . Therefore,  $\|\mathbf{U}(\tilde{\mathbf{F}} - \mathbf{FH})\|_{\max}$  has the same rate.

## E Proof of Theorem 4.1

*Proof.* Denote the empirical covariance matrix as

$$\Sigma_N = \frac{1}{N} \sum_{i=1}^m \mathbf{U}^i \mathbf{U}^{i'}$$

As in Cai et al. (2011), the upper bound on  $\|\widehat{\Omega} - \Omega\|$  is obtained by proving

$$\|(\widehat{\Sigma} - \Sigma_N)\Omega\|_{\max} = O_p(\tau_{N,p}) \quad \text{and} \quad \|(\Sigma_N - \Sigma)\Omega\|_{\max} = O_p(\tau_{N,p}). \quad (\text{E.1})$$

Once the two bounds are established, we proceed by observing

$$\|\mathbf{I}_p - \widehat{\Sigma}\Omega\|_{\max} = \|(\widehat{\Sigma} - \Sigma)\Omega\|_{\max} = O_p(\tau_{N,p}),$$

and then it readily follows that if  $\lambda \asymp \tau_{N,p}$ ,

$$\begin{aligned} \|\widehat{\Omega} - \Omega\|_{\max} &\leq \|\Omega(\mathbf{I}_p - \widehat{\Sigma}\widehat{\Omega})\|_{\max} + \|(\mathbf{I}_p - \widehat{\Sigma}\Omega)'\widehat{\Omega}\|_{\max} \\ &\leq \|\Omega\|_1 \|\mathbf{I}_p - \widehat{\Sigma}\widehat{\Omega}\|_{\max} + \|\mathbf{I}_p - \widehat{\Sigma}\Omega\|_{\max} \|\widehat{\Omega}\|_1 \leq \lambda \|\Omega\|_1 + \tau \|\Omega\|_1 = O_p(\tau_{N,p}), \end{aligned}$$

where the first term of the last inequality uses the constraint of (4.4) while the optimality condition of (4.4) is applied to bound  $\|\widehat{\Omega}\|_1$  by  $\|\Omega\|_1$ . So it remains to find  $\tau_{N,p}$  in (E.1). Since  $\Omega \in \mathcal{F}(s, C_0)$ ,  $\|\Omega\|_1 \leq C_0$ , so we just need to bound  $\|\widehat{\Sigma} - \Sigma_N\|_{\max}$  and  $\|\Sigma_N - \Sigma\|_{\max}$ . Obviously,

$$\|\Sigma_N - \Sigma\|_{\max} = O_p\left(\sqrt{\frac{\log p}{N}}\right).$$

By assumption  $\|\widehat{\Sigma} - \Sigma_N\|_{\max} = O_P(a_{m,N,p})$ . Thus  $\tau = \sqrt{\log p/N} + a_{m,N,p}$ . Similar proof as in Cai et al. (2011) can also reach error bounds under  $\|\cdot\|_1$  and  $\|\cdot\|_2$ , which we omit. The proof is now complete.  $\square$

## F Technical lemmas

- Lemma F.1.** (i)  $\|\Lambda' \mathbf{U}\|_F^2 = O_P(np)$ ,  
(ii)  $\|\mathbf{U}' \mathbf{U}\|_F^2 = O_P(np^2 + pn^2)$ ,  
(iii)  $\|\mathbf{U}' \mathbf{U} \mathbf{F}\|_F^2 = O_P(np^2 + pn^2)$ .

*Proof.* We simply apply Markov inequality to get the rates.

$$\mathbb{E}\|\Lambda' \mathbf{U}\|_F^2 = \mathbb{E}[\text{tr}(\Lambda' \mathbf{U} \mathbf{U}' \Lambda)] = n \cdot \text{tr}(\Lambda' \Sigma \Lambda) \leq n \|\Sigma\| \cdot \text{tr}(\Lambda' \Lambda) = O(np).$$

$$\begin{aligned}\mathbb{E}\|\mathbf{U}'\mathbf{U}\|_F^2 &= \mathbb{E}\left[\sum_{t=1}^n \sum_{t'=1}^n \left(\sum_{j=1}^p u_{jt} u_{jt'}\right)^2\right] = \sum_{j_1, j_2=1}^p \left(\sum_{t=1}^n \mathbb{E}[u_{j_1 t}^2 u_{j_2 t}^2] + \sum_{1 \leq t \neq t_1 \leq n} \sigma_{j_1 j_2}^2\right) \\ &= O_P(np^2 + pn^2),\end{aligned}$$

since  $\sum_{j_1, j_2} \sigma_{j_1 j_2}^2 = \text{tr}(\Sigma^2) \leq \|\Sigma\| \text{tr}(\Sigma) = O(p)$ .

$$\begin{aligned}\mathbb{E}\|\mathbf{U}'\mathbf{U}\mathbf{F}\|_F^2 &= \mathbb{E}\left[\sum_{t=1}^n \sum_{k=1}^K \left(\sum_{t'=1}^n \sum_{j=1}^p u_{jt} u_{jt'} f_{t'k}\right)^2\right] \\ &= \sum_{k=1}^K \sum_{j_1, j_2=1}^p \left(\sum_{t=1}^n \mathbb{E}[u_{j_1 t}^2 u_{j_2 t}^2] f_{tk}^2 + \sum_{1 \leq t \neq t_1 \leq n} \sigma_{j_1 j_2}^2 f_{t_1 k}^2\right) = O_P(np^2 + pn^2).\end{aligned}$$

□

**Lemma F.2.** (i)  $\|\Lambda'\mathbf{U}\|_{\max} = O_P(\sqrt{p \log n})$ .

(ii)  $\|\mathbf{U}'\mathbf{U}\|_{\max} = O_P(p)$ ,

(iii)  $\|\mathbf{U}'\mathbf{U}\mathbf{F}\|_{\max} = O_P(\sqrt{np \log n} + p\sqrt{\log n})$ .

*Proof.* (i)  $\|\Lambda'\mathbf{U}\|_{\max} = \max_{t,k} |\mathbf{u}_t' \boldsymbol{\lambda}_k|$  where  $\boldsymbol{\lambda}_k$  is the  $k^{th}$  column of  $\Lambda$ . Since  $\mathbf{u}_t' \boldsymbol{\lambda}_k$  is mean zero sub-Gaussian with variance proxy  $\boldsymbol{\lambda}_k' \Sigma \boldsymbol{\lambda}_k \leq \|\Sigma\| \|\boldsymbol{\lambda}_k\|^2 = O(p)$ , we have  $\|\Lambda'\mathbf{U}\|_{\max} = O_P(\sqrt{p \log n})$ .

(ii)  $\|\mathbf{U}'\mathbf{U}\|_{\max} = \max_{t,t'} |\mathbf{u}_t' \mathbf{u}_{t'}| \leq \max_{t \neq t'} |\mathbf{u}_t' \mathbf{u}_{t'}| + \max_t |\mathbf{u}_t' \mathbf{u}_t|$ . We need to bound each term separately. The second term is bounded by the upper tail bound of Hanson-Wright inequality for sub-Gaussian vector (Hsu et al., 2012; Rudelson and Vershynin, 2013) i.e.

$$\mathbb{P}(\|\mathbf{u}_t\|^2 > \text{tr}(\Sigma) + 2\sqrt{\text{tr}(\Sigma)s} + 2\|\Sigma\|s) \leq e^{-s}.$$

Choose  $s = \log n$  and apply union bound, we have  $\max_t |\mathbf{u}_t' \mathbf{u}_t| = O_P(\text{tr}(\Sigma) + 2\sqrt{\text{tr}(\Sigma)s}) = O_P(p + \sqrt{p \log n}) = O_P(p)$ . Then we deal with the first term. By Chernoff bound,

$$\mathbb{P}(\max_{t \neq t'} |\mathbf{u}_t' \mathbf{u}_{t'}| > s) \leq 2n^2 e^{-s\theta} \mathbb{E}[\exp(\theta \mathbf{u}_t' \mathbf{u}_{t'})],$$

where  $\mathbb{E}[\exp(\theta \mathbf{u}_t' \mathbf{u}_{t'})] = E[\exp(\theta^2 \mathbf{u}_t' \Sigma \mathbf{u}_t / 2)] \leq E[\exp(C\theta^2 \|\mathbf{u}_t\|^2)]$ . Hsu et al. (2012) showed that

$$\mathbb{E}[\exp(\eta \|\mathbf{u}_t\|^2)] \leq \exp\left(\text{tr}(\Sigma)\eta + \frac{\text{tr}(\Sigma^2)\eta^2}{1 - 2\|\Sigma\|\eta}\right)$$

For  $\eta < 1/(4\|\Sigma\|) \leq \text{tr}(\Sigma)/(4\text{tr}(\Sigma^2))$ , the right hand side is less than  $\exp(3\text{tr}(\Sigma)\eta/2) \leq$

$\exp(Cp\eta)$ . Choose  $\eta = C\theta^2$ , we have

$$\mathbb{P}(\max_{t \neq t'} |\mathbf{u}'_t \mathbf{u}_{t'}| > s) \leq 2n^2 \exp(-s\theta + C\theta^2 p).$$

We minimize the right hand side and choose  $\theta = s/(2Cp)$ , it is easy to check  $\eta < 1/(4\|\Sigma\|)$  and see that  $\max_{t \neq t'} |\mathbf{u}'_t \mathbf{u}_{t'}| = O_P(\sqrt{p \log n})$ . So we conclude that  $\|\mathbf{U}'\mathbf{U}\|_{\max} = O_P(p)$ .

(iii) Let  $\bar{\mathbf{f}}_k$  be the  $k^{th}$  column of  $\mathbf{F}$ .  $\|\mathbf{U}'\mathbf{U}\mathbf{F}\|_{\max} = \max_{t,k} |\mathbf{u}'_t \mathbf{U}\bar{\mathbf{f}}_k| \leq \max_{t,k} |\mathbf{u}'_t \mathbf{U}_{(-t)}\bar{\mathbf{f}}_{k(-t)}| + \max_{t,k} |\mathbf{u}'_t \mathbf{u}_t f_{tk}|$  where  $\mathbf{U}_{(-t)}$ ,  $\bar{\mathbf{f}}_{k(-t)}$  are  $\mathbf{U}$  and  $\bar{\mathbf{f}}_k$  canceling the  $t^{th}$  column and element respectively. From (ii) we know the second term is of order  $O_P(p \max_{t,k} |f_{tk}|) = O_P(p\sqrt{\log n})$ . Define  $\boldsymbol{\xi} = \mathbf{U}_{(-t)}\bar{\mathbf{f}}_{k(-t)} \sim \text{subGaussian}(\mathbf{0}, \Sigma \|\bar{\mathbf{f}}_{k(-t)}\|^2)$ , which is independent with  $\mathbf{u}_t$ . Thus

$$\mathbb{P}(\max_{t,k} |\mathbf{u}'_t \boldsymbol{\xi}| > s) \leq 2nK e^{-s\theta} \mathbb{E}[\exp(\theta \mathbf{u}'_t \boldsymbol{\xi})],$$

where  $\mathbb{E}[\exp(\theta \mathbf{u}'_t \boldsymbol{\xi})] \leq \mathbb{E}[\exp(\theta^2 \mathbf{u}'_t \Sigma \mathbf{u}_t \|\bar{\mathbf{f}}_{k(-t)}\|^2 / 2)] \leq \mathbb{E}[\exp(C\theta^2 n \|\mathbf{u}_t\|^2)]$ . Similar to (ii), we choose  $\eta = C\theta^2 n$  here. It is not hard to see  $\max_{t,k} |\mathbf{u}'_t \boldsymbol{\xi}| = O_P(\sqrt{np \log n})$ . Thus  $\|\mathbf{U}'\mathbf{U}\mathbf{F}\|_{\max} = O_P(\sqrt{np \log n} + p\sqrt{\log n})$ .  $\square$

**Lemma F.3.** (i)  $\|\mathbf{F}'\mathbf{U}'\|_F^2 = O_P(np)$ .

(ii)  $\|\mathbf{U}'\Phi(\mathbf{W})\|_F^2 = O_P(npJ)$ ,  $\|\mathbf{U}'\Phi(\mathbf{W})\mathbf{B}\|_F^2 = O_P(np)$ .

(iii)  $\|\Phi(\mathbf{W})'\mathbf{U}\mathbf{F}\|_F^2 = O_P(npJ)$ ,  $\|\mathbf{B}'\Phi(\mathbf{W})'\mathbf{U}\mathbf{F}\|_F^2 = O_P(np)$ .

*Proof.* This results can be found in the paper of Fan, Liao and Wang (2014). But the conditions they used are a little bit different from our conditions. In particular, we allow no time (sample) dependence and only require bounded  $\|\Sigma\|_2$  instead of  $\|\Sigma\|_1$ . By Markov inequality, it is sufficient to show the expected value of each term attains the corresponding rate of convergence.

$$\mathbb{E}\|\mathbf{F}'\mathbf{U}'\|_F^2 = \mathbb{E}[\text{tr}(\mathbf{F}'\mathbb{E}[\mathbf{U}'\mathbf{U}]\mathbf{F})] = \mathbb{E}[\text{tr}(\mathbf{F}'\text{tr}(\Sigma)\mathbf{F})] = n \cdot \text{tr}(\Sigma) = O(np).$$

$$\begin{aligned} \mathbb{E}\|\mathbf{U}'\Phi(\mathbf{W})\|_F^2 &= \mathbb{E}[\text{tr}(\Phi'\mathbb{E}[\mathbf{U}\mathbf{U}'|\mathbf{W}]\Phi)] = n \cdot \mathbb{E}[\text{tr}(\Phi'\Sigma\Phi)] \leq nJd \cdot \mathbb{E}[\|\Phi'\Sigma\Phi\|_2] \\ &\leq nJdC_0 \mathbb{E}[\|\Phi'\Phi\|_2] = O(npJ). \end{aligned}$$

$$\mathbb{E}\|\Phi(\mathbf{W})'\mathbf{U}\mathbf{F}\|_F^2 = \mathbb{E}[\text{tr}(\Phi'\mathbb{E}[\mathbf{U}\mathbf{F}'\mathbf{U}'|\mathbf{W}]\Phi)] = \mathbb{E}[\text{tr}(\mathbf{F}\mathbf{F}')\text{tr}(\Phi'\Sigma\Phi)] = O(npJ).$$

$\mathbb{E}\|\mathbf{U}'\Phi(\mathbf{W})\mathbf{B}\|_F^2$  and  $\|\mathbf{B}'\Phi(\mathbf{W})'\mathbf{U}\mathbf{F}\|_F^2$  are both  $O(np)$  following the same proof as above. Thus the proof is complete.  $\square$

**Lemma F.4.** (i)  $\|\mathbf{F}'\mathbf{U}'\|_{\max} = O_P(\sqrt{n \log p})$

(ii)  $\|\mathbf{U}'\Phi(\mathbf{W})\|_{\max} = O_P(\phi_{\max} \sqrt{p \log(nJ)})$ ,  $\|\mathbf{U}'\Phi(\mathbf{W})\mathbf{B}\|_{\max} = O_P(\sqrt{p \log n})$ .

(iii)  $\|\Phi(\mathbf{W})'\mathbf{U}\mathbf{F}\|_{\max} = O_P(\phi_{\max}\sqrt{np \log J})$ ,  $\|\mathbf{B}'\Phi(\mathbf{W})'\mathbf{U}\mathbf{F}\|_{\max} = O_P(\sqrt{np})$ .

*Proof.* (i) It is not hard to see  $\|\mathbf{F}'\mathbf{U}'\|_{\max} = \max_{k \leq K, j \leq p} |\sum_{t=1}^n f_{tk} u_{jt}| = O_p(\sqrt{n \log p})$ . The detailed proof by Chernoff bound is given in the following. By union bound and Chernoff bound, we have

$$\mathbb{P}\left(\max_{k \leq K, i \leq p} \left| \sum_{t=1}^n f_{tk} u_{jt} \right| > t\right) \leq 2pK e^{-t\theta} \cdot \mathbb{E}\left[e^{\theta \sum_{t=1}^n f_{tk} u_{jt}}\right].$$

The expectation is calculated by fist conditioning on  $\mathbf{F}$ ,

$$E\left[e^{\theta \sum_{t=1}^n f_{tk} u_{jt}}\right] = \mathbb{E}\left[\mathbb{E}\left[e^{\theta \sum_{t=1}^n f_{tk} u_{jt}} | \mathbf{F}\right]\right] \leq \mathbb{E}\left[e^{\theta^2 \sum_{t=1}^n f_{tk}^2 \sigma_{jj}/2}\right] \leq e^{\frac{1}{2} n C_0 \theta^2},$$

where the second equality uses the sub-Gaussianity of  $u_{jt}$  and the last inequality is from  $n^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}$  and  $\|\Sigma\|_2 \leq C_0$ . Therefore, choosing  $\theta = \frac{t}{nC_0}$ , we have

$$\mathbb{P}\left(\max_{k \leq K, j \leq p} \left| \sum_{t=1}^n f_{tk} u_{jt} \right| > t\right) \leq 2pK e^{-t\theta} e^{\frac{C_0}{2} n \theta^2} = 2pK e^{-\frac{t^2}{2C_0 n}}.$$

Thus  $\|\mathbf{F}'\mathbf{U}'\|_{\max} = O_p(\sqrt{n \log p})$ .

(ii)  $\|\mathbf{U}'\Phi(\mathbf{W})\|_{\max} = \max_{\nu, l, t} |\sum_{j=1}^p u_{jt} \phi_\nu(W_{jl})| = \max_{\nu, l, t} |\bar{\phi}'_{\nu l} \mathbf{u}_t|$ , where  $\bar{\phi}_{\nu l} = (\phi_\nu(\mathbf{W}_{1l}), \dots, \phi_\nu(\mathbf{W}_{pl}))'$ . Consider the tail probability condition on  $\mathbf{W}$ :

$$\mathbb{P}\left(\max_{\nu \leq J, l \leq d, k \leq n} |\bar{\phi}'_{\nu l} \mathbf{u}_k| > t \mid \mathbf{W}\right) \leq 2Jdn \cdot e^{-t\theta} \mathbb{E}[e^{\theta \bar{\phi}'_{\nu l} \mathbf{u}_k} \mid \mathbf{W}] \leq 2Jdn \cdot \exp\left\{-t\theta + \frac{1}{2}\theta^2 \bar{\phi}'_{\nu l} \Sigma \bar{\phi}_{\nu l}\right\}.$$

The right hand side can be further bounded by

$$2Jdn \cdot \exp\left(-t\theta + \frac{1}{2}\theta^2 C_0 \|\bar{\phi}_{\nu l}\|^2\right) \leq 2Jdn \cdot \exp\left(-t\theta + \frac{1}{2}pC_0 \theta^2 \phi_{\max}^2\right).$$

Choose  $\theta$  to minimize the upper bound and take expectation with respect to  $\mathbf{W}$ , we obtain

$$\mathbb{P}\left(\max_{\nu \leq J, l \leq d, k \leq n} |\bar{\phi}'_{\nu l} \mathbf{u}_k| > t\right) \leq 2Jdn \cdot \exp\left\{-\frac{t^2}{2pC_0 \phi_{\max}^2}\right\}.$$

Finally choose  $t \asymp \phi_{\max} \sqrt{p \log(nJ)}$ , the tail probability is arbitrarily small with a proper constant. So  $\|\mathbf{U}'\Phi(\mathbf{W})\|_{\max} = O_P(\phi_{\max} \sqrt{p \log(nJ)})$ . The second part of the results follows similarly. Note  $\|\mathbf{U}'\Phi(\mathbf{W})\mathbf{B}\|_{\max} \leq \|\mathbf{U}'\mathbf{G}(\mathbf{W})\|_{\max} + \|\mathbf{U}'\mathbf{R}(\mathbf{W})\|_{\max}$  and the first term

dominates. So the same derivation gives

$$\mathbb{P}\left(\|\mathbf{U}'\mathbf{G}(\mathbf{W})\|_{\max} > t\right) \leq 2Kn \cdot \exp\left\{-\frac{t^2}{2C_0\|\bar{g}_k\|^2}\right\},$$

where  $\bar{g}_k = (g_k(\mathbf{W}_1), \dots, g_k(\mathbf{W}_p))$ .  $\|\bar{g}_k\|^2 = O_p(p)$  since it is assumed eigenvalues of  $p^{-1}\mathbf{G}(\mathbf{W})'\mathbf{G}(\mathbf{W})$  is bounded almost surely. Hence,  $\|\mathbf{U}'\Phi(\mathbf{W})\mathbf{B}\|_{\max} = O_P(\sqrt{p \log n})$ .

(iii)  $\|\Phi(\mathbf{W})'\mathbf{U}\mathbf{F}\|_{\max} = \max_{\nu \leq J, l \leq d, k \leq K} |\sum_{j=1}^p \sum_{i=1}^n \phi_\nu(W_{jl}) u_{ji} f_{ik}|$ . Using Chernoff bound again, we get

$$\mathbb{P}\left(\max_{\nu \leq J, l \leq d, k \leq K} \left| \sum_{j=1}^p \sum_{i=1}^n \phi_\nu(W_{jl}) u_{ji} f_{ik} \right| > t\right) \leq 2JdK \cdot e^{-t\theta} \cdot \mathbb{E}\left[e^{\theta \sum_{t=1}^n f_{tk} \bar{\phi}'_{\nu l} \mathbf{u}_t}\right].$$

Since  $\sum_{t=1}^n f_{tk} \bar{\phi}'_{\nu l} \mathbf{u}_t | \mathbf{F} \sim \text{sub-Gaussian}(0, \sum_{t=1}^n f_{tk}^2 \bar{\phi}'_{\nu l} \Sigma \bar{\phi}_{\nu l}) = \text{sub-Gaussian}(0, n \bar{\phi}'_{\nu l} \Sigma \bar{\phi}_{\nu l})$ , the right hand side is easy to bound by first conditioning on  $\mathbf{F}$ .

$$\mathbb{E}\left[e^{\theta \sum_{t=1}^n f_{tk} \bar{\phi}'_{\nu l} \mathbf{u}_t}\right] \leq \mathbb{E}\left[\exp\left(\frac{1}{2}n\theta^2 \bar{\phi}'_{\nu l} \Sigma \bar{\phi}_{\nu l}\right)\right] \leq E\left[\exp\left(\frac{1}{2}npC_0\phi_{\max}^2\theta^2\right)\right].$$

Therefore, choosing  $\theta = \frac{t}{npC_0\phi_{\max}^2}$ , we have

$$\mathbb{P}\left(\|\Phi(\mathbf{W})'\mathbf{U}\mathbf{F}\|_{\max} > t\right) \leq 2JdK \cdot \exp\left\{-t\theta + \frac{1}{2}npC_0\phi_{\max}^2\theta^2\right\} = 2JdK \exp\left\{-\frac{t^2}{2npC_0\phi_{\max}^2}\right\}.$$

So we conclude  $\|\Phi(\mathbf{W})'\mathbf{U}\mathbf{F}\|_{\max} = O_p(\phi_{\max}\sqrt{np \log J})$ . By similar derivation as in (ii), we also have  $\|\mathbf{B}'\Phi(\mathbf{W})'\mathbf{U}\mathbf{F}\|_{\max}$  and  $\|\mathbf{G}(\mathbf{W})'\mathbf{U}\mathbf{F}\|_{\max}$  are both of order  $O_P(\sqrt{np})$ .

□

**Lemma F.5.** (i)  $\|\mathbf{U}\mathbf{U}'\Lambda\|_{\max} = O_P(\sqrt{np \log p} + n\|\Sigma\|_1)$ ,  
(ii)  $\|\mathbf{U}\mathbf{U}'\Phi(\mathbf{W})\|_{\max} = O_P(\phi_{\max}(\sqrt{np \log p} + n\|\Sigma\|_1))$  and  $\|\mathbf{U}\mathbf{U}'\Phi(\mathbf{W})\mathbf{B}\|_{\max} = O_P(\sqrt{np \log p} + nJ\phi_{\max}\|\Sigma\|_1)$ .

*Proof.* (i)  $\|\mathbf{U}\mathbf{U}'\Lambda\|_{\max} \leq \max_{j,k} |\sum_{t=1}^n u_{jt} \mathbf{u}_t' \boldsymbol{\lambda}_k - n \sum_{j'=1}^p \sigma_{jj'} \lambda_{j'k}| + n \max_{j,k} \sum_{j'=1}^p |\sigma_{jj'}||\lambda_{j'k}|$ . The second term is  $O(n\|\Sigma\|_1)$ . So it suffices to focus on the first term. Let  $\Sigma = \mathbf{A}\mathbf{A}'$  and  $\mathbf{u}_t = \mathbf{A}\mathbf{v}_t$  so that  $\text{Var}(\mathbf{v}_t) = \mathbf{I}$ . Write  $\mathbf{A}' = (\mathbf{a}_1, \dots, \mathbf{a}_p)$ , so we have  $u_{jt} = \mathbf{a}_j' \mathbf{v}_t$ . Also denote  $\mathbf{d}_k = \mathbf{A}' \boldsymbol{\lambda}_k$ . Thus  $u_{jt} \mathbf{u}_t' \boldsymbol{\lambda}_k = \mathbf{a}_j' \mathbf{v}_t \mathbf{v}_t' \mathbf{d}_k$  and  $\sum_{j'=1}^p \sigma_{jj'} \lambda_{j'k} = \mathbf{a}_j' \mathbf{d}_k$ .

$$\mathbb{P}\left(\max_{j,k} \left| \sum_{t=1}^n (\mathbf{a}_j' \mathbf{v}_t \mathbf{v}_t' \mathbf{d}_k - \mathbf{a}_j' \mathbf{d}_k) \right| > s\right) \leq pK \mathbb{P}\left(\left| \sum_{t=1}^n (\tilde{\mathbf{a}}_j' \mathbf{v}_t \mathbf{v}_t' \tilde{\mathbf{d}}_k - \tilde{\mathbf{a}}_j' \tilde{\mathbf{d}}_k) \right| > \frac{s}{\max_{j,k} \|\mathbf{a}_j\| \|\mathbf{d}_k\|}\right), \quad (\text{F.1})$$

where  $\tilde{\mathbf{a}}_j$  and  $\tilde{\mathbf{d}}_k$  are two unit vectors of dimension  $p$ . We will bound the right hand side

with arbitrary unit vectors  $\tilde{\mathbf{a}}_j$  and  $\tilde{\mathbf{d}}_k$ .

$$\begin{aligned} & \mathbb{P}\left(\left|\sum_{t=1}^n \tilde{\mathbf{a}}_j' \mathbf{v}_t \mathbf{v}_t' \tilde{\mathbf{d}}_k - n \tilde{\mathbf{a}}_j' \tilde{\mathbf{d}}_k\right| > s\right) \\ & \leq \mathbb{P}\left(\left|\sum_{t=1}^n ((\tilde{\mathbf{a}}_j + \tilde{\mathbf{d}}_k)' \mathbf{v}_t)^2 - n \|\tilde{\mathbf{a}}_j + \tilde{\mathbf{d}}_k\|^2\right| > 2s\right) + \mathbb{P}\left(\left|\sum_{t=1}^n ((\tilde{\mathbf{a}}_j - \tilde{\mathbf{d}}_k)' \mathbf{v}_t)^2 - n \|\tilde{\mathbf{a}}_j - \tilde{\mathbf{d}}_k\|^2\right| > 2s\right). \end{aligned}$$

Note that  $(\tilde{\mathbf{a}}_j + \tilde{\mathbf{d}}_k)' \mathbf{v}_t \sim \text{subGaussian}(0, \|\tilde{\mathbf{a}}_j + \tilde{\mathbf{d}}_k\|^2)$  and  $\|\tilde{\mathbf{a}}_j + \tilde{\mathbf{d}}_k\|^2 \leq 4$ . By Bernstein inequality, we have for constant  $C > 0$ ,

$$\mathbb{P}\left(\left|\sum_{t=1}^n (\tilde{\mathbf{a}}_j' \mathbf{v}_t \mathbf{v}_t' \tilde{\mathbf{d}}_k - \tilde{\mathbf{a}}_j' \tilde{\mathbf{d}}_k)\right| > s\right) \leq 2 \exp\left(-C \min(s^2/n, s)\right).$$

Choose  $s = C\sqrt{n \log p} \max_{jk} \|\mathbf{a}_j\| \|\mathbf{d}_k\|$  in (F.1), we can easily show that the exception probability is small as long as  $C$  is large enough. Therefore, noting  $\max_{jk} \|\mathbf{a}_j\| \|\mathbf{d}_k\| \leq C \max_k \|\boldsymbol{\lambda}_k\|$ ,  $\max_{j,k} |\sum_{t=1}^n u_{jt} \mathbf{u}_t' \boldsymbol{\lambda}_k - n \sum_{j'=1}^p \sigma_{jj'} \lambda_{j'k}| = O_P(\sqrt{n \log p} \max_k \|\boldsymbol{\lambda}_k\|) = O_P(\sqrt{np \log p})$ . Finally  $\|\mathbf{U} \mathbf{U}' \boldsymbol{\Lambda}\|_{\max} = O_P(\sqrt{np \log p} + n \|\boldsymbol{\Sigma}\|_1)$ .

(ii) The rates of  $\|\mathbf{U} \mathbf{U}' \Phi(\mathbf{W})\|_{\max}$  and  $\|\mathbf{U} \mathbf{U}' \Phi(\mathbf{W}) \mathbf{B}\|_{\max}$  can be similarly derived as (i). Denote  $\boldsymbol{\Phi}_{vl} = (\phi_v(W_{1l}), \dots, \phi_v(W_{pl}))'$ , so

$$\begin{aligned} \|\mathbf{U} \mathbf{U}' \Phi(\mathbf{W})\|_{\max} & \leq \max_{j,v,l} \left| \sum_{t=1}^n u_{jt} \mathbf{u}_t' \boldsymbol{\Phi}_{vl} - n \sum_{j'=1}^p \sigma_{jj'} \phi_v(W_{j'l}) \right| + n \max_{j,v,l} \sum_{j'=1}^p |\sigma_{jj'}| |\phi_v(W_{j'l})| \\ & = O_P(\sqrt{n \log p} \max_{v,l} \|\boldsymbol{\Phi}_{vl}\| + n \phi_{\max} \|\boldsymbol{\Sigma}\|_1) = O_P(\phi_{\max} (\sqrt{np \log p} + n \|\boldsymbol{\Sigma}\|_1)). \end{aligned}$$

Denote the  $k^{th}$  column of  $\Phi(\mathbf{W}) \mathbf{B}$  by  $(\boldsymbol{\Phi} \mathbf{B})_k$ , we have

$$\begin{aligned} \|\mathbf{U} \mathbf{U}' \Phi(\mathbf{W}) \mathbf{B}\|_{\max} & \leq \max_{j,k} \left| \sum_{t=1}^n u_{jt} \mathbf{u}_t' (\boldsymbol{\Phi} \mathbf{B})_k - n \sum_{j'=1}^p \sigma_{jj'} (\boldsymbol{\Phi} \mathbf{B})_{j'k} \right| + n \max_{j,k} \sum_{j'=1}^p |\sigma_{jj'}| |(\boldsymbol{\Phi} \mathbf{B})_{j'k}| \\ & = O_P(\sqrt{n \log p} \max_k \|(\boldsymbol{\Phi} \mathbf{B})_k\| + n J \phi_{\max} \|\boldsymbol{\Sigma}\|_1) = O_P(\sqrt{np \log p} + n J \phi_{\max} \|\boldsymbol{\Sigma}\|_1), \end{aligned}$$

where we use  $\max_k \|(\boldsymbol{\Phi} \mathbf{B})_k\| \leq \|\boldsymbol{\Phi} \mathbf{B}\|_F = O_P(\sqrt{p})$ .  $\square$