Published in final edited form as:
Electron J Stat. 2018 ; 12(2): 3908–3952. doi:10.1214/18-EJS1466.

Heterogeneity adjustment with applications to graphical model inference

Jianqing Fan,
Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA, jqfan@princeton.edu

Han Liu,
Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60208, USA, hanliu@northwestern.edu

Weichen Wang,
Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA, nickweichwang@gmail.com

Ziwei Zhu
Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA, zzw9348ustc@gmail.com

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 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 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

Multiple sourcing; batch effect; 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 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’. As characterized in [29], microarray gene expression data obtained from different labs on different processing dates may contain systematic variability. Furthermore, [30] 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. Therefore, to properly analyze data aggregated from multiple sources, we need to carefully model and adjust the unwanted variations.

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 experiments, 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, [43], [2], [30], and [25]. 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 theory.

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 factor $\mathbf{f}_t^i$ influence the observed data X_{jt}^i in batch i (j indexes variables; t indexes samples) as in the approximate factor model:

$$X_{jt}^i = \boldsymbol{\lambda}_j^{i\prime}\mathbf{f}_t^i + u_{jt}^i,\ 1 \le j \le p, 1 \le t \le n_i, 1 \le i \le m, \tag{1.1}$$

where $\boldsymbol{\lambda}_j^i$ is an unknown factor loading for variable j and u_{jt}^i is a true uncorrupted signal. We consider a random loading $\boldsymbol{\lambda}_j^i$. The linear term $\boldsymbol{\lambda}_j^{i\prime}\mathbf{f}_t^i$ models the heterogeneity effect. We assume that $\mathbf{f}_t^i$ is independent of u_{jt}^i and $\mathbf{u}_t^i = (u_{1t}, \ldots, u_{pt})'$ shares the same common distribution with mean $\mathbf{0}$ and covariance $\boldsymbol{\Sigma}_{p\times p}$ across all data sources. In the matrix-form model, (1.1) can be written as

$$\mathbf{X}^i = \boldsymbol{\Lambda}^i\mathbf{F}^{i\prime} + \mathbf{U}^i, \tag{1.2}$$

where $\mathbf{X}^i$ is a $p\times n_i$ data matrix in the i^{th} batch, $\boldsymbol{\Lambda}^i$ is a $p\times K^i$ factor loading matrix with $\boldsymbol{\lambda}_j^{i\prime}$ 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 $p \times n_i$. We allow the number of latent factors K^i to depend on batch i. We emphasize here that within one batch, our model is homogeneous. Heterogeneity in this paper refers to that the batch effect terms $\left\{\mathbf{\Lambda}^i\mathbf{F}^{i'}\right\}_{i=1}^m$ are different across different groups $i = 1,\ldots,m$, which are unwanted variations in our study.

To see more clearly on how model (1.2) characterizes the heterogeneity, note that for the t^{th} sample $\mathbf{X}_t^i$, which is the t^{th} column of $\mathbf{X}^i$,

$$\mathrm{var}(\mathbf{X}_t^i) = \mathbf{\Lambda}^i \mathrm{var}(\mathbf{f}_t^i)\mathbf{\Lambda}^{i'} + \mathbf{\Sigma}\,. \tag{1.3}$$

Therefore, the heterogeneity is carried by the low-rank component $\mathbf{\Lambda}^i\mathrm{var}(\mathbf{f}_t^i)\mathbf{\Lambda}^{i'}$ in the population covariance matrix of $\mathbf{X}_t^i$. We need to clarify that since we assume both $\mathbf{F}^i$ and $\mathbf{U}^i$ have mean zero, heterogeneity mentioned in this paper is for covariance structure as shown above instead of mean structure. In addition, our model differs from the random/mixed effect regression model studied in the literature [45, 23, 11] in that our models are factor models without any factors observed, while the mixed/random effect model is a regression model that requires covariate matrices to estimate the batch-specific term.

Under a pervasive assumption, the heterogeneity component can be estimated by directly applying principal component analysis (PCA) or Projected-PCA (PPCA), which is more accurate when there are sufficiently informative covariates $\mathbf{W}^i$ [18]. Let $\widehat{\mathbf{\Lambda}^i}\widehat{\mathbf{F}^{i}}{}'$ be the estimated heterogeneity component and $\widehat{\mathbf{U}^i} = \mathbf{X}_t^i - \widehat{\mathbf{\Lambda}^i}\widehat{\mathbf{F}^{i}}{}'$ 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 <u>A</u>daptive <u>L</u>ow-rank <u>P</u>rincipal <u>H</u>eterogeneity <u>A</u>djustment) 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 $\mathbf{X} = (X_1,\ldots,X_p)'$. The sparsity pattern of the precision matrix $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ 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 $\Omega_{ij} \neq 0$ (the $(i,j)^{th}$ element of $\mathbf{\Omega}$), meaning that X_i and X_j are dependent conditioning on the rest of the 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 [7], but they assumed observed factors and no heterogeneity issue, i.e., $m = 1$.

A significant amount of literature has focused on the estimation of the precision matrix $\mathbf{\Omega}$ for graphical models for homogeneous data. [49] and [20] developed the Graphical Lasso method using the L_1 penalty and [27] and [42] used a non-convex penalty. Furthermore, [40] and [33] studied the theoretical properties under different assumptions. Estimating $\mathbf{\Omega}$ can be equivalently reformulated as a set of node-wise sparse linear regression that utilizes Lasso or

Danzig selector for each node [35, 48]. To relax the assumption of Gaussian data, [32] and [31] extend the graphical model to the case of semiparametric Gaussian copula and transelliptical family. Via the ALPHA framework, we can combine the adjusted data $\widehat{\mathbf{U}^i}$ to construct an estimator for the precision matrix $\boldsymbol{\Omega}$ by the above methods. Recent works also focus on joint estimation of multiple Gaussian or discrete graphical models which share some common structure [22, 15, 47, 8, 21]. They are concerned with both the commonality and individual uniqueness of the graphs. In comparison, ALPHA emphasizes more on heterogeneity-adjusted aggregation for one single graph.

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 PCA and PPCA will be introduced for adjusting the factor effects under different regimes. A guiding rule of thumb is also proposed to determine which method is more appropriate. The heterogeneity-adjusted data will be combined to provide valid graph estimation in Section 4. The CLIME method of [9] is applied for precision matrix estimation. Synthetic and real data analyses are carried out to demonstrate the proposed framework in Section 5. Section 6 contains further discussions and 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 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 $\mathbf{W}^i_j = (W^i_{j1}, \dots, W^i_{jd})'$ 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 $\boldsymbol{\lambda}^i_j$ in (1.1) so that it can be further modeled as $\boldsymbol{\lambda}^i_j = \mathbf{g}^i(\mathbf{W}^i_j) + \gamma^i_j$, where $\mathbf{g}^i(\cdot)$ is the external covariate effects on $\boldsymbol{\lambda}^i_j$ and γ^i_j is the part that can not be explained by the covariates. Thus, model (1.1) can be written as

$$X^i_{jt} = \boldsymbol{\lambda}^{i\prime}_j \mathbf{f}^i_t + u^i_{jt} = (\mathrm{g}^i(\mathbf{W}^i_j) + \gamma^i_j)' \mathbf{f}^i_t + u^i_{jt}\,. \tag{2.1}$$

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

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

In (2.2), $\mathbf{G}^i(\mathbf{W}^i)$ and $\boldsymbol{\Gamma}^i$ are $p\times K^i$ matrices. More specifically, $g_k^i(\mathbf{W}_j^i)$ and γ_{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 $t \le n_i$ and for all subgroups $i \le m$ with $\mathbb{E}\left[\mathbf{u}_t^i\right] = \mathbf{0}$, $\mathrm{var}(\mathbf{u}_t^i) = \boldsymbol{\Sigma}$.

There has been a large literature on factor models in econometrics [3, 5, 17, 44], machine learning [10, 36] and random matrix theories [26, 38, 46]. 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 [14] and further investigated by [13] and [18]. Using sufficiently informative external covariates, we are able to more accurately estimate the factors and loadings, and hence yield better adjustment for heterogeneity.

Here we collect some notations of eigenvalues and matrix norms used in the paper. For matrix $\mathbf{M}$, we use $\lambda_{\max}(\mathbf{M})$, $\lambda_{\min}(\mathbf{M})$ and $\lambda_i(\mathbf{M})$ to denote the maximum eigenvalue, the minimum eigenvalue and the ith eigenvalue of $\mathbf{M}$ respectively. We define the quantities $\|\mathbf{M}\|_{\max} = \max_{i,j}\left|M_{i,j}\right|$, $\|\mathbf{M}\|_2 = \lambda_{\max}^{1/2}(\mathbf{M}'\mathbf{M})$($\|\mathbf{M}\|$ for short), $\|\mathbf{M}\|_F = (\sum_{i,j}M_{ij}^2)^{1/2}$, $\|\mathbf{M}\|_1 = \max_j\sum_i\left|M_{ij}\right|$ and $\|\mathbf{M}\|_{1,1} = \sum_i\sum_j\left|M_{ij}\right|$ to be its entry-wise maximum, spectral, Frobenius, induced ℓ_1 and element-wise ℓ_1 norms.

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*) $n_i^{-1}\mathbf{F}^{i\prime}\mathbf{F}^i = \mathbf{I}$.

(ii) $\left\{\mathbf{u}_t^i\right\}_{t \le n_i, i \le m}$ *are independently and identically sub-Gaussian distributed with mean zero and covariance* Σ *within and between subgroups, and independent of* $\left\{\mathbf{W}_j^i, \mathbf{f}_t^i\right\}$. *Let* $\|\boldsymbol{\Sigma}\|_2 = C_0 < \infty$.

(iii) $\left\{\mathbf{f}_t^i\right\}_{t \le n_i}$ *is a stationary process, with arbitrary temporal dependency. The tail of the factors is sub-Gaussian, i.e., there exist* $C_1, C_2 > 0$ *such that for any* $\alpha \in \mathbb{R}^{K^i}$ *and* $s > 0$, $\mathbb{P}(\left|\alpha'\mathbf{f}_t^i\right| > s) \le C_1\exp(-C_2s^2/\left\|\alpha\right\|^2)$.

The above set of assumptions are commonly used in the literature, see [5] and [18]. 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 the external covariates do not have enough explanatory power on the factor loadings $\mathbf{\Lambda}^i$, we ignore the semiparametric structure and model (2.2) reduces to the traditional factor model, extensively studied in econometrics [3, 44, 37]. 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* $c_{\min}, c_{\max} > 0$ *so that*

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

(ii) $\max_{k \le K^i, j \le p} \left|\lambda^i_{jk}\right| = O_p(\sqrt{\log p})$.

The first condition is common and essential in the factor model literature (e.g., [44]). It requires the factors to be strong enough such that the covariance matrix $\mathbf{\Lambda}^i \mathrm{cov}(\mathbf{f}^i_t)\mathbf{\Lambda}^i + \mathbf{\Sigma}$ has spiked eigenvalues. We emphasize here that this condition is actually not so stringent as it looks. Consider a single-factor model $Y_{it} = b_i f_t + u_{it}$, $i = 1,\ldots,p$, $t = 1,\ldots,T$. The pervasive assumptions actually imply that $c_{\min} p \le \sum_{i=1}^p b_i^2 \le c_{\max} p$. Note that since $c_{\min}$ can be a small constant, our pervasive assumption just says that the factors $\{f_t\}_{t=1}^T$ have non-negligible effect on a non-vanishing proportion of outcomes. In addition, this condition is trivially true if $\{\boldsymbol{\lambda}^i_j\}_{j=1}^p$'s can be regarded as random samples from a population with non-degenerate covariance matrix [17]. Practically, in fMRI data analysis for instance, the lab environment (temperature, air pressure, etc.) or the mental status of the subject being scanned may cause the BOLD (Blood-Oxygen-Level Dependent) level to be uniformly higher at certain time t. This means the brain heterogeneity is globally driven by the factors $\{f_t\}_{t=1}^T$. If the batch effect is only limited to a small number of dimensions, we think it is more appropriate to assume sparsity instead of pervasiveness on top eigenvectors, which is quite different from our problem setups and thus beyond the scope of our paper. The second condition holds if the population has a sub-Gaussian tail.

2.2.2. Regime 2: External covariates are informative—

When covariates are informative, we will employ the PPCA [18] to 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*

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

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

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

Assumption 2.4. *(i)* $E\gamma^i_{jk} = 0,\ \max_{k \le K^i, j \le p} \left|\gamma^i_{jk}\right| = O_P(\sqrt{\log p})$.

(ii) Write $\gamma^i_j = (\gamma^i_{j1}, \dots, \gamma^i_{jK})'$. *we assume* $\left\{\gamma^i_j\right\}_{j \le p}$ *are independent of* $\left\{\mathbf{W}^i_j\right\}_{j \le p}$.

(iii) Define $\nu_p = \max_{i \le m} \max_{k \le K^i} p^{-1} \sum_{j \le p} \mathrm{var}(\gamma^i_{jk}) < \infty$. *We assume*

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

Condition (i) is parallel to Assumption 2.2 (ii) whereas Condition (ii) is natural since $\mathbf{\Gamma}^i$ can not be explained by $\mathbf{W}^i$. Condition (iii) imposes cross-sectional weak dependence of γ^i_j, which is much weaker than assuming independent and identically distributed $\left\{\gamma^i_j\right\}_{j \le p}$. 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 the above two regimes: they correspond to estimating $\mathbf{U}^i$ by either PCA or PPCA. From now on, we drop the superscript i whenever there is no confusion as we focus on the i^{th} data source. We 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 $\breve{\mathbf{F}}$ and $\breve{\mathbf{U}}$ 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 PPAC is used.

Estimators for latent factors under regimes 1 and 2 satisfy $n^{-1}\breve{\mathbf{F}}'\breve{\mathbf{F}} = \mathbf{I}$, 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

$$\breve{\mathbf{U}} = \mathbf{X}(\mathbf{I} - \frac{1}{n}\breve{\mathbf{F}}\breve{\mathbf{F}}') . \tag{3.1}$$

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 $\mathbf{X}'\mathbf{X}$ corresponding to the k^{th} largest eigenvalue. We have the following theoretical results.

Theorem 3.1. *Under Assumptions 2.1 and 2.2, we have*

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

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

Note that we do not explicitly assume bounded $\|\mathbf{\Sigma}\|_1$. In some applications it might be natural to assume a sparse covariance so that all terms involving $\|\mathbf{\Sigma}\|_1$ can be eliminated, while in other applications such as the graphical model, it is more natural to impose sparsity structure on the precision matrix. In this case, one may want to keep track of the effect of $\|\mathbf{\Sigma}\|_1$ as it can be as large as $O(\sqrt{p})$ as $\|\mathbf{\Sigma}\|_1 \le \sqrt{p}\|\mathbf{\Sigma}\|_2 = O(\sqrt{p})$.

3.2. Estimating factors by Projected-PCA

In regime 2, we would like to incorporate the external covariates using the Projected-PCA (PPCA) method proposed by [18]. The method applies PCA on the projected data and by projection, covariates information is leveraged to reduce dimensionality. We now briefly introduce the method.

For simplicity, we only consider $d = 1$, that is, we only have a single covariate. The general case can be found in [18]. To model the unknown function $g_k(W_j)$, we adopt a sieve based idea which approximates $g_k(\cdot)$ by a linear combination of basis functions $\{\phi_1(x), \phi_2(x), \cdots\}$ (e.g., B-spline, Fourier series, polynomial series, wavelets). Then

$$g_k(W_j) = \sum_{\nu=1}^{J} b_{\nu,k}\phi_\nu(W_j) + R_k(W_j), \quad k \le K, j \le p. \tag{3.2}$$

Here $\{b_{\nu,k}\}_{\nu=1}^{J}$ are the sieve coefficients of $g_k(W_j)$, corresponding to the k^{th} factor loading; R_k is the remainder function representing the approximation error; J denotes the number of sieve bases which may grow slowly as p diverges. We take the same basis functions in (3.2) for all k though they can be different.

Define $\mathbf{b}_k' = (b_{1,k}, \cdots, b_{J,k}) \in \mathbb{R}^J$ for each $k \le K$, and correspondingly $\phi(W_j)' = (\phi_1(W_j), \cdots, \phi_J(W_j)) \in \mathbb{R}^J$. Then we can write

$$g_k(W_j) = \phi(W_j)'\mathbf{b}_k + R_k(W_j).$$

Let $\mathbf{B}_{J \times K} = (\mathbf{b}_1, \cdots, \mathbf{b}_K)$, $\mathbf{\Phi}(\mathbf{W})_{p \times J} = (\phi(W_1), \cdots, \phi(W_p))'$ and $R_k(W_j)$ be the $(j,k)^{th}$ element of $\mathbf{R}(\mathbf{W})p{\times}K$. The matrix form (2.2) can be written as

$$\mathbf{X} = \Phi(\mathbf{W})\mathbf{B}\mathbf{F}' + \mathbf{R}(\mathbf{W})\mathbf{F}' + \mathbf{\Gamma}\mathbf{F}' + \mathbf{U}, \tag{3.3}$$

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

The idea of PPCA is simple: since the factor loadings are a function of the covariates in (3.3) and $\mathbf{U}$ and $\mathbf{\Gamma}$ are independent of $\mathbf{W}$, if we project (smooth) the observed data onto the space of $\mathbf{W}$, the effect of $\mathbf{U}$ and $\mathbf{\Gamma}$ will be significantly reduced and the problem becomes nearly a noiseless one, given that the approximation error $\mathbf{R}(\mathbf{W})$ is small.

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

$$\mathbf{P} = \Phi(\mathbf{W})(\Phi(\mathbf{W})'\Phi(\mathbf{W}))^{-1}\Phi(\mathbf{W})'. \tag{3.4}$$

By (3.3), $\mathbf{P}\mathbf{X} \approx \mathbf{P}\Phi(\mathbf{W})\mathbf{B}\mathbf{F}' \approx \mathbf{G}(\mathbf{W})\mathbf{F}'$. Thus, $\mathbf{F}$ can be estimated from the 'noiseless projected data' $\mathbf{PX}$, using the conventional PCA. Let the columns of $\widetilde{\mathbf{F}}/\sqrt{n}$ 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 of the projected data. Then, $\widetilde{\mathbf{F}}$ is the PPCA estimator of $\mathbf{F}$. It only differs from PCA in that we use smoothed or projected data $\mathbf{PX}$.

We need the following conditions for basis functions and accuracy of sieve approximation.

Assumption 3.1. *(i) There are* $d_{\min}, d_{\max} > 0$ *s.t.*

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

almost surely and $\max_{\nu \le J, j \le p} E\phi_\nu(W_j)^2 < \infty$.

(ii) There exists $k \ge 4$ *s.t. as* $J \to \infty$, $\sup_{x \in \chi}|g_k(x) - \sum_{\nu=1}^{J} b_{\nu,k}\phi_\nu(x)|^2 = O(J^{-k})$ *where X is the support of* W_j *and* $\max_{\nu,k}|b_{\nu,k}| < \infty$.

Condition (ii) is mild; for instance, when $\{\varphi_\nu\}$ is polynomial basis or Bsplines, it is implied by the condition that smooth curve $g_k(\cdot)$ belongs to a Hölder class $\mathcal{G} = \left\{g : |g^{(r)}(s) - g^{(r)}(t)| \le L|s - t|^\alpha\right\}$ for some $L > 0$, with $k = 2(r + \alpha) \ge 4$ [34,12].

Recalling the definition of ν_p in Assumption 2.4 (iii), we have the following results.

Theorem 3.2. *Choose* $J = (p\min\left\{n, p, \nu_p^{-1}\right\})^{1/k}$ *and assume* $J^2\phi_{\max}^2\log(nJ) = O(p)$ *where* $\phi_{\max} = \max_{\nu \le J}\sup_{x \in X}\phi_\nu(x)$. *Under Assumptions 2.1, 2.3, 2.4 and 3.1,*

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

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

where $\|\mathbf{\Pi}\|_{\max} = O_P(\sqrt{\log n/p}(J\phi_{\max} + \sqrt{\log p}) + J\phi_{\max}\|\mathbf{\Sigma}\|_1\sqrt{\log n}/p)$ *and* $\|\mathbf{\Delta}\|_{\max} = O_P(n\sqrt{\nu_p/p}(J^2\phi_{\max}^2 + \log p) + nJ\phi_{\max}\|\mathbf{\Sigma}\|_1(J\phi_{\max} + \sqrt{\log p})/p + n^2J^2\phi_{\max}^2\|\mathbf{\Sigma}\|_1^2/p^2)$ *if there exists C s.t.* $v_p > C/n$.

3.3. A guiding rule for estimating the number of factors, the number of basis functions and determining regimes

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 [4, 1, 28]. [28] and [1] proposed to use ratio of adjacent eigenvalues of $\mathbf{X}'\mathbf{X}$ to infer the number of factors. They showed the estimator $\widehat{K} = \arg\max_{k \le K_{\max}} \lambda_k(\mathbf{X}'\mathbf{X})/\lambda_{k+1}(\mathbf{X}'\mathbf{X})$ correctly identifies K with probability tending to 1, as long as $K_{\max} \ge K$ and $K_{\max} = O(n_i \wedge p)$.

For the semiparametric factor model, [18] proposed

$$\widetilde{K} = \arg\max_{k \le K_{\max}} \lambda_k(\mathbf{X}'\mathbf{P}\mathbf{X})/\lambda_{k+1}(\mathbf{X}'\mathbf{P}\mathbf{X}).$$

Here $K_{\max}$ is of the same order as Jd. It was shown that $\mathbb{P}(\widetilde{K} = K) \to 1$ under regular assumptions which we omit here. When we have genuine and pervasive covariates, $\widetilde{K}$ typically outperforms $\widehat{K}$. More details can be found in [18].

Once we use $\widehat{K}$ and $\widetilde{K}$ to estimate the number of factors under the regular factor model and semiparametric factor model respectively, we naturally have an adaptive rule to decide whether the covariates $\mathbf{W}$ are informative enough to use PPCA over PCA. We compare two eigen-ratios:

$$\frac{\lambda_{\widehat{K}}(\mathbf{X}'\mathbf{X})}{\lambda_{\widehat{K}+1}(\mathbf{X}'\mathbf{X})} \text{vs} \frac{\lambda_{\widetilde{K}}(\mathbf{X}'\mathbf{P}\mathbf{X})}{\lambda_{\widetilde{K}+1}(\mathbf{X}'\mathbf{P}\mathbf{X})}.$$

If the former is larger we identify the dataset as regime 1 and apply regular PCA to get $\widehat{\mathbf{U}}$; otherwise it is regime 2 and PPCA is used to obtain $\widetilde{\mathbf{U}}$. The intuition behind this comparison is that the maximal eigen-ratios can be perceived as signal-to-noise ratios in terms of estimating the spiky heterogeneity term. Given that $n^{-1}\mathbf{X}\mathbf{X}' \approx \mathbf{G}\mathbf{G}' + \mathbf{\Gamma}\mathbf{\Gamma}' + \mathbf{\Sigma}$ and $n^{-1}\mathbf{P}\mathbf{X}\mathbf{X}'\mathbf{P} \approx \mathbf{G}\mathbf{G}' + \mathbf{P}\mathbf{\Gamma}\mathbf{\Gamma}'\mathbf{P} + \mathbf{P}\mathbf{\Sigma}\mathbf{P}$, the first ratio measures the eigen-gap between $\mathbf{G}\mathbf{G}' + \mathbf{\Gamma}\mathbf{\Gamma}'$ and $\mathbf{\Sigma}$ and the second ratio measures the eigen-gap between $\mathbf{G}\mathbf{G}' + \mathbf{P}\mathbf{\Gamma}\mathbf{\Gamma}'\mathbf{P}$ and $\mathbf{P}\mathbf{\Sigma}\mathbf{P}$. If $\mathbf{G}(\mathbf{W})$ is much more important than $\mathbf{\Gamma}$ in explaining the loading structure, projection preserves signal and reduces error to improve the eigen-gap. Conversely, if $\mathbf{W}$ is weak in providing useful information, projection reduces both noise and signal. Therefore, if projection enlarges the maximum eigen-gap, we prefer PPCA over PCA to estimate the spiky heterogeneity part. Our proposed guiding rule effectively tells whether projection can further contrast spiky and non-spiky parts of covariance.

The above signal-to-noise ratio comparison can be extended to choose the number of basis functions. Notice that we can regard regular PCA as PPCA with number of basis $J = p$ and hence $\mathbf{P} = \mathbf{I}$. In this line of thinking, we can index $\mathbf{P}$ by J and maximize $\lambda_{\widetilde{K}(J)}(\mathbf{X}'\mathbf{P}^J\mathbf{K})/\lambda_{\widetilde{K}(J)+1}(\mathbf{X}'\mathbf{P}^J\mathbf{X})$ over $J \in \{1, 2, \ldots, J_{\max}, p\}$, where $J = p$ corresponds to PCA. Here we use notation $\widetilde{K}(J)$ and $\mathbf{P}^J$ to exhibit their dependency on J. We implement this guiding rule in real data analysis.

In practice, there is still chance of misspecification of the true number of factors K by ALPHA. One might be curious about how this will affect the performance of ALPHA and the subsequent statistical analysis. To clarify this issue, we conduct sensitivity analysis on the number of factors in Section G.3 in the appendix. The take-home message is that the overestimation of K will not hurt, while underestimation of K might mislead subsequent statistical inference.

3.4. Summary of ALPHA

We now summarize the final procedure and convergence rates. We first divide m subgroups into two classes based on whether the collected covariates have significant influence on the loadings.

$$\mathscr{M}_1 = \left\{ i \le m \,\middle|\, \mathbf{W}^i \text{ is not informative} \right\}, \quad \mathscr{M}_2 = \left\{ i \le m \,\middle|\, \mathbf{W}^i \text{ is informative} \right\}.$$

ALPHA consists of the following three steps.

Step 1: (**Preprocessing**) For data source i, determine whether it belongs to $\mathscr{M}_1$ or $\mathscr{M}_2$ according to the guiding rule given in Section 3.3 and correspondingly estimate K by $\breve{K}$, which equals $\widehat{K}$ or $\widetilde{K}$ (and choose J if necessary).

Step 2: (**Adjustment**) Apply Projected-PCA to estimate if $\mathbf{\Lambda}^i\mathbf{F}^{i'}$ if $i \in \mathscr{M}_2$, otherwise use PCA to remove the heterogeneity, resulting in adjusted data $\breve{\mathbf{U}}^i$, which is either $\widehat{\mathbf{U}}^i$ or $\widetilde{\mathbf{U}}^i$.

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

We summarize the ALPHA procedure in Algorithm 1 given in Section A. We also summarize the convergence of $\widehat{\mathbf{U}}^i$ and $\widetilde{\mathbf{U}}^i$ below. To ease presentation, we consider a typical regime in practice: $n_i < C_p$, $\sum_{i \le m} K^i < CN$ for some constant C. We focus on the situation of sufficiently smooth curves $k = \infty$ so that J diverges very slowly (say with rate $O(\sqrt{\log p})$) and bounded $\phi_{\max}$ and ν_p (defined respectively in Theorem 3.2 and Assumption 2.4). Based on discussions of the previous subsections, for estimation of $\mathbf{U}$ in $\| \cdot \|_{\max}$, we have

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

Therefore, PPCA 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

$$||\breve{\mathbf{U}}^i - \mathbf{U}^i||_{\max} = O_P(||\mathbf{U}^i\mathbf{F}^i\mathbf{F}^{i'}/n_i||_{\max})O_P(\sqrt{\log n_i \log p/n_i}).$$

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

$$\breve{\mathbf{U}}^i\breve{\mathbf{U}}^{i'} - \mathbf{U}^i\mathbf{U}^{i'} = -\mathbf{U}^i\mathbf{F}^i\mathbf{F}^{i'}\mathbf{U}^{i'}/n_i + \begin{cases} O_P(\log p + \delta) & \text{if } i \in \mathscr{M}_1, \\ O_P(n_i \log p\sqrt{\nu_p/p} + \delta) & \text{if } i \in \mathscr{M}_2, \end{cases} \tag{3.5}$$

where $\delta = n_i^2 \|\mathbf{\Sigma}\|_1^2 \log p/p^2$, depending on $\|\mathbf{\Sigma}\|_1$. If we consider a very sparse covariance matrix so that $\|\mathbf{\Sigma}\|_1$ is bounded, we can simply drop the term δ in both regimes. Then, regime 1 achieves better rate if $p = O(n_i^2 \nu_p)$ but regime 2 outperforms otherwise.

4. Post-ALPHA inference

We have summarized the order of biases caused by adjusting heterogeneity for each data source in Section 3.4. Now we combine the adjusted data together for further statistical analysis. As an example, we study estimation of the Gaussian graphical model. Assume further $\mathbf{u}_t^i \sim N(\mathbf{0}, \mathbf{\Sigma})$ and consider the following class of the precision matrices:

$$\mathscr{F}(s, R) = \left\{ \mathbf{\Omega} : \mathbf{\Omega} \succ \mathbf{0},\ \|\mathbf{\Omega}\|_1 \le R,\ \max_{1 \le i \le p} \sum_{j=1}^{p} \mathbb{1}(\Omega_{i,j} \ne 0) \le s \right\}. \tag{4.1}$$

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

To estimate $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ via CLIME, we need a covariance estimator as the input. We 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

$$\widehat{\mathbf{\Sigma}} = \frac{1}{N - \sum_{i \le m} K^i} \sum_{i=1}^{m} \breve{\mathbf{U}}^i \breve{\mathbf{U}}^{i'}. \tag{4.2}$$

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

4.1. Covariance estimation

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

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 not hard to verify that $\boldsymbol{\xi}_k^i$ is Gaussian distributed with mean zero and variance $\mathbf{\Sigma}$. Note that $\left\{\boldsymbol{\xi}_k^i\right\}_{1\le i\le m,\,1\le k\le K^i}$ are i.i.d. with respect to k and i, using the assumption $\mathbf{F}^{i'}\mathbf{F}^i/n_i = \mathbf{I}$. By the standard concentration bound (e.g. Lemma 4.2 of [19]),

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

where $K^{tot} = \sum_{i\le m}K^i$. Therefore, by (3.5), we have

$$\left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}_N\right\|_{\max} = \left\|\frac{N}{N-\sum_{i\le m}K^i}\frac{1}{N}\sum_{i\le m}\left(\breve{\mathbf{U}}^i\breve{\mathbf{U}}^{i'} - \mathbf{U}^i\mathbf{U}^{i'} + K^i\mathbf{\Sigma}\right.\right.$$
$$) \qquad + \frac{\sum_{i\in\mathscr{M}}K^i}{N-\sum_{i\in\mathscr{M}}K^i}\left(\frac{1}{N}\sum_{i\le m}\mathbf{U}^i\mathbf{U}^{i'} - \mathbf{\Sigma}\right)\Bigg\|_{\max} =: O_P(a_{m,N,P}), \tag{4.3}$$

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

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 $i\in\mathscr{M}_1$ 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 more slowly than the order of $\sqrt{N/\log p}$.

If we apply PPCA to all data sources, i.e., $i\in\mathscr{M}_2$ 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 $\widehat{\mathbf{\Sigma}}$, we apply the CLIME estimator proposed by [9]. For a given $\widehat{\mathbf{\Sigma}}$, CLIME solves the following optimization problem:

$$\widehat{\mathbf{\Omega}} = \underset{\mathbf{\Omega}}{\operatorname{argmin}} \|\mathbf{\Omega}\|_{1,1} \quad \text{subject to } \|\widehat{\mathbf{\Sigma}}\mathbf{\Omega} - \mathbf{I}\|_{\max} \le \lambda, \tag{4.4}$$

where $\|\mathbf{\Omega}\|_{1,1} = \sum_{i,j \le p} |\mathbf{\Omega}_{ij}|$ and λ 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 $\hat{\sigma}_{ij}$ and $\hat{\sigma}_{ji}$. The resulting matrix after symmetrization, still denoted as $\widehat{\mathbf{\Omega}}$ with a little bit abuse of notation, also attains good rate of convergence. In particular, we consider the sparse precision matrix class $\mathscr{F}(s, C_0)$ in (4.1). The following lemma guarantees recovery of any sparse matrix $\mathbf{\Omega} \in \mathscr{F}(s, C_0)$.

Theorem 4.1. *Suppose* $\mathbf{\Omega} \in \mathscr{F}(s, C_0)$ *and let* $\tau_{m,N,p} = \sqrt{\log p/N} + a_{m,N\,.\,p}$. *Choosing* $\lambda \asymp \tau_{m,N,p}$, *we have*

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

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

Here we stress that we choose CLIME for the precision matrix estimation because it only relies on the max-norm guarantee $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\max}$. The intuition is that for any true $\mathbf{\Omega}$ with bounded, $\|\mathbf{\Omega}\|_1$,

$$\|\mathbf{I} - \widehat{\mathbf{\Sigma}}\mathbf{\Omega}\|_{\max} = \|(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma})\mathbf{\Omega}\|_{\max} \le \|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\max}\|\mathbf{\Omega}\|_1 = O_{\mathbb{P}}(\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\max}) \,.$$

One can see from above that fast convergence of $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\max}$ encourages feasibility of $\mathbf{\Omega}$**,** which is a necessary step for establishing consistency of the resulting M-estimator. Interested readers can refer to the proof of Theorem 4.1 for more details. Other possible methods for precision matrix recovery (e.g. graphical Lasso in [20], graphical Dantzig selector in [48] and graphical neighborhood selection in [35]) can be considered for post-ALPHA inference as well, but their convergence rate needs to be studied in a case-by-case fashion.

Theorem 4.1 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 [48] 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 suggest that the latter error is often negligible.

In addition, we numerically investigate how misspecification of the number of factors K will affect the precision matrix estimation in Section G.3 in the appendix.

5. Numerical studies

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

In addition, we analyze a real brain image data using the proposed procedure. The dataset to be analyzed is the ADHD-200 data [6]. It consists of rs-fMRI images of 608 subjects, of whom 465 are healthy and 143 are diagnosed with ADHD. We dropped subjects with missing values in our analysis. Following [39], we divided the whole brain into 264 regions of interest (ROI, $p = 264$), which are regarded as nodes in our graphical model. Each brain was scanned for multiple times with sample sizes ranging from 76 to 261 ($76 \le n_i \le 261$). In each scan, we acquired the blood-oxygen-level dependent (BOLD) signal within each ROI. Note that subjects have different ages, genders etc., which results in heterogeneity over the covariance structure of the data. We need to remove this unwanted heterogeneity; otherwise it will dilute or corrupt the true biological signal, i.e., the difference in the brain functional network between healthy people and patients due to the disease ADHD.

5.1. Preliminary analysis

To apply our ALPHA framework, we need to first argue the pervasiveness condition Assumption 2.2 holds for the real dataset considered. This is done in Section G.2, together with further discussions on pervasiveness. We also collect the physical locations of the 264 regions as the external covariates. Ideally, we hope these covariates to be pervasive in explaining the batch effect (Assumption 2.3), while bearing no association with the graph structure of $\mathbf{u}_t$. This is reasonably true because: the level of batch effect is non-uniform over different locations of the brain when scanned in fMRI machines; furthermore it has been widely acknowledged in biological studies that spatial adjacency does not necessarily imply brain functional connectivity.

To construct $\mathbf{W}_j^i$ from the physical locations, we simply split the 264 regions into 10 clusters ($J = 10$) by the hierarchy clustering (Ward's minimum variance method) and use the categorical indices as the covariates of the nodes. 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.

Here $J = 10$ is only used to calibrate our synthetic model in the next subsection. In the real data analysis, we will choose J adaptively according to our heuristic guiding rule of the maximal eigen-gap discussed in Section 3.3. 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 $w = 1,\ldots,10$. We use the same external covariates for all subjects in both healthy and diseased groups.

The next question is how to divide the subjects into $\mathscr{M}_1$ and $\mathscr{M}_2$ based on whether the selected covariates explain the loadings effectively. We implemented the method given in Section 3.3 and discovered that 398 healthy (85.6%) and 126 diseased samples (88.1%) prefer PPCA over PCA, meaning that the physical locations indeed have explanatory powers on factor loadings of most subjects. We identified them as subjects in $\mathscr{M}_2$ while the others were classified as in $\mathscr{M}_1$. Based on the class labels, we employed the corresponding method to estimate the number of factors and adjust the heterogeneity. We used $K_{\max} = 3$. The estimated number of factors for the two groups are summarized in Table 1.

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. We specify four asymptotic settings for our simulation studies:

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

Here the last setting represents regime 1, where we should expect PCA to work well when the number of subjects is of order of square root of the total sample size, i.e., $m \asymp \sqrt{N}$. The first three settings represent regime 2 with informative covariates; they present asymptotics with growing p, m and n_i respectively. The details on model calibration and data generation can be found in Section G.1.

We first investigate the errors of estimating covariance of $\mathbf{u}_t$ in max-norm after applying PPCA or PCA for heterogeneity adjustment. We also compare them with the estimation errors if we naively pool all the data together without any heterogeneity adjustment. However, the estimation error of the naively pooled sample covariance is too large to fit in the graph for the first 3 cases, which we thus do not plot. Denote the oracle sample covariance of $\mathbf{u}_t$ by $\mathbf{\Sigma}_N$ as before. The estimation errors, based on 100 simulations, under the four settings are presented in Figure 3.

In Case 1, m and n_i are fixed while dimension p increases. This setting highlights 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 PPCA and PCA benefit from an increasing number of subjects. However, since n_i is small, again PPCA outperforms. In Case 3, m and p are fixed while n_i increases. Both methods achieve better estimation as n_i increases, but more importantly, regular PCA outperforms PPCA 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 $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\max} = O_P(\sqrt{\log p/N})$, while PPCA only achieves $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\max} = O_P(\log p/\sqrt{p})$, 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. Adjusting by PCA behaves much better and PPCA sometimes is as bad as 'nPCA', corresponding to no heterogeneity adjustment. This is not unexpected as we utilize a noisy external covariates.

Now we focus on estimation error of the precision matrix. We plug $\widehat{\mathbf{\Sigma}}$, obtained from data after adjusting for heterogeneity, into CLIME to get the estimator $\widehat{\mathbf{\Omega}}$ of $\mathbf{\Omega}$. In Figure 4, $\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}\|_{\max}$ and $\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}\|_1$ are depicted under the four asymptotic settings. From the plots we see $\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}\|_{\max}$ and $\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}\|_1$ share similar behavior with $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\max}$ shown in Figure 3: in Case 1, n_i is small, so it is advantageous to use PPCA and PPCA behaves better as dimension increases; in Case 2, both PPCA and PCA benefit from an increasing number of subjects and PPCA outperforms PCA; in Case 3, PCA outperforms PPCA when n_i is large enough since m is fixed; in Case 4, the covariates have no explanation power at all so that PPCA does not make sense. In the first three cases, if we do not adjust data heterogeneity, $\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}\|_{\max}$ and $\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}\|_1$ will be too large to fit in the current scale.

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, representing performance of random guess. It is obvious from those plots that heterogeneity adjustment very much improves the sparsity recovery of the precision matrix. When the sample size of each subject is small, genuine pervasive covariates increase the power of PPCA while if the sample size is relatively large, PCA is sufficiently good in recovering graph structures. Also notice that in all cases, the naive method without heterogeneity adjustment can still achieve a certain amount of power, but we can improve the performance dramatically by correcting the batch effects.

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 our ALPHA framework in this subsection. We took various sparsity levels of the networks from 1% to 5% (corresponding to the same set of λ's for two groups) and selected the common edges, which are stable with respect to tuning, to be depicted.

The brain network produced by our proposed method is presented in Figure 6. It gives 90.7% identical edges for the two networks. However if we ignore heterogeneity and naively pool the data from all subjects together, it generates 10.2% unshared edges, roughly 1% more than ALPHA produces. Therefore, by heterogeneity adjustment, we found less difference in brain functional networks between ADHD patients and healthy people. In addition, we investigate how those unshared edges are distributed across the 10 clusters. We summarized the total degree of unshared edge vertices within each cluster in Table 2. As we can see, in the left occipital lobe (red) and the left parietal lobe (pink), there are significant difference in functional connectivity structure between healthy people and patients, although in general the difference is weak. 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

Heterogeneity is usually informed by the domain knowledge of the dataset. In particular, it occurs with high chance when the data come from different sources or subgroups. In the brain image dataset we used in the numerical study, heterogeneity across patients can stem from difference in age, gender, etc. When it is less clear whether heterogeneity exists, we can calculate multiple summary statistics for all the subgroups and see whether they are significantly different. In the case of pervasive heterogeneity, we can test it by the magnitude of dominating eigenvalues of the covariance matrix in each subgroup. A systematic testing method for heterogeneity is important and we leave it for now as a future research topic. Note that even if all the subgroups are actually homogeneous, ALPHA does not hurt the statistical efficiency under appropriate scaling assumptions. Specifically, for the PCA-based ALPHA, we showed in Section 4.1 that as long as the number of subgroups $m = O(\sqrt{N/\log p})$, $\breve{\mathbf{\Sigma}}$ enjoys the oracle max-norm rate. This means that given homogeneous data, when the number of data splits is not large, ALPHA yields the same statistical rate as the full-sample oracle estimator. For the PPCA-based ALPHA, $\breve{\mathbf{\Sigma}}$ enjoys the oracle rate when $p/\log p = \mathbf{\Omega}(\sqrt{N/\log p})$.

As we have seen, ALPHA is adaptive to factor structures and is flexible to include external information. However, this advantage of PPCA is accompanied by more assumptions and the practical issue of selecting proper basis functions and the number of them in sieve approximation. One contribution of the paper lies in seamless integration of PCA and PPCA, which leverages effective external covariates. 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.

Note that our framework is compatible with any statistical procedure that only requires an accurate estimator as the input, like CLIME we illustrate in this work. The ALPHA procedure gives theoretical guarantee for $\|\breve{\mathbf{U}} - \mathbf{U}\|_{\max}$ and $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\max}$, which serve as foundations for establishing the statistical properties of the subsequent procedure. Besides, ALPHA has potential application and in regression analysis. If the residual terms $\{\mathbf{U}^i\}_{i=1}^m$ are true predictors for the response of interest $\{\mathbf{Y}^i\}_{i=1}^m$, we can first apply ALPHA to extract the residuals before the regression procedure. For example, the residual BOLD signal we obtained by ALPHA in the brain functional network analysis (Section 5.3) is potentially useful in predicting whether a person has ADHD. This is a typical logistic regression problem based on ALPHA adjustment. We leave the detailed study of combining ALPHA with regression models for future investigation. One recent work [16] has adopted a method similar to ALPHA that extracts residuals for model selection in high dimensional regression.

Finally, we point out two current limitations of ALPHA. The first limitation lies in its pervasiveness assumption of the heterogeneity terms $\left\{\mathbf{\Lambda}^i\mathbf{F}^{i'}\right\}_{i=1}^{m}$. More specifically, for each subgroup *i*, ALPHA requires the signal strength of the heterogeneous part $\mathbf{\Lambda}^i\mathbf{F}^{i'}$ to overwhelm the homogeneous residual part $\mathbf{U}^i$ so that PCA or PPCA can accurately estimate $\mathbf{\Lambda}^i\mathbf{F}^{i'}$ and remove it. Such requirement can be violated in practice when the heterogeneous term has similar signal strength as the homogeneous term. Additionally, statistical methods that require more than the max-norm error guarantee $(||\breve{\mathbf{U}} - \mathbf{U}||_{\max}, ||\breve{\mathbf{\Sigma}} - \mathbf{\Sigma}||_{\max})$, say in the general non-sparse situation, may be inappropriate for the post-ALPHA inference for now.

Acknowledgments

This project was supported by National Science Foundation grants DMS-1206464, DMS1406266 and 2R01-GM072611-12.

Appendix A:: Algorithm for ALPHA

The pseudo code for the algorithm ALPHA is shown as follows.

Algorithm 1 Algorithm for adaptive low-rank principal heterogeneity adjustment

```
Input: Panel $\mathbf{X}^i_{p \times n_i}$ and $d$-dimensional $\left\{\mathbf{W}^i_j\right\}_{j=1}^p$ from $m$ data sources,
       $J_{\max}, K_{\max}(J_{\max} \geq (K_{\max}+1)/d)$
Output: $\breve{\mathbf{U}}^i, \breve{K}^i$ and $\widehat{\mathbf{\Sigma}}$
procedure ALPHA
  for each subject $i \leq m$ do
    $\widehat{K}^i \leftarrow \text{argmax}_{K \leq K_{max}} \lambda_k(\mathbf{X}^{i\prime}\mathbf{X}^i)/\lambda_{k+1}(\mathbf{X}^{i\prime}\mathbf{X}^i)$
    $\Delta\lambda^i_0 \leftarrow \lambda_{\widehat{K}^i}(\mathbf{X}^{i\prime}\mathbf{X}^i)/\lambda_{\widehat{K}^i+1}(\mathbf{X}^{i\prime}\mathbf{X}^i)$
    for each $(K_{\max}+1)/d \leq J \leq J_{\max}$ do
      $\mathbf{P}^i_J \leftarrow \Phi(\mathbf{W}^i)(\Phi(\mathbf{W}^i)'\Phi(\mathbf{W}^i))^{-1}\Phi(\mathbf{W}^i)'$ for $J$
      $\widetilde{K}^i_J \leftarrow \text{arg}max_{K \leq K_{\max}} \lambda_k(\mathbf{X}^{i\prime}\mathbf{P}^i_J\mathbf{X}^i)/\lambda_{k+1}(\mathbf{X}^{i\prime}\mathbf{P}^i_J\mathbf{X}^i)$
      $\Delta\lambda^i_J \leftarrow \lambda_{\widetilde{K}^i_J}(\mathbf{X}^{i\prime}\mathbf{P}^i_J\mathbf{X}^i)/\lambda_{\widetilde{K}^i_J+1}(\mathbf{X}^{i\prime}\mathbf{P}^i_J\mathbf{X}^i)$
    end for
    $J^i_* \leftarrow \text{argmax}_J \Delta\lambda^i_J$
    $\widetilde{K}^i \leftarrow \widetilde{K}^i_{J^i_*}$

    if $\Delta\lambda^i_0 > \Delta\lambda^i_{J^i_*} (i \in \mathscr{M}_1)$ then
      $\widehat{\mathbf{F}}^i/\sqrt{n_i} \leftarrow$ eigenvectors of $\mathbf{X}^{i\prime}\mathbf{X}^i$ of the top $\widetilde{K}^i$ eigenvalues
      $\widehat{\mathbf{\Lambda}}^i \leftarrow \mathbf{X}^i\widehat{\mathbf{F}}^i/n_i, \widehat{\mathbf{U}}^i \leftarrow \mathbf{X}^i - \widehat{\mathbf{\Lambda}}^i\widehat{\mathbf{F}}^i$
      $\breve{\mathbf{U}}^i \leftarrow \widehat{\mathbf{U}}^i, \breve{K}^i \leftarrow \widehat{K}^i$
    else
      $\widetilde{\mathbf{F}}^i/\sqrt{n_i} \leftarrow$ eigenvectors of $\mathbf{X}^{i\prime}\mathbf{P}^i_{J^i_*}\mathbf{X}^i$ of the top $\widetilde{K}^i$ eigenvalues
      $\widetilde{\mathbf{\Lambda}}^i \leftarrow \mathbf{X}^i\widetilde{\mathbf{F}}^i/n_i, \widetilde{\mathbf{U}}^i \leftarrow \mathbf{X}^i - \widetilde{\mathbf{\Lambda}}^i\widetilde{\mathbf{F}}^i$
      $\breve{\mathbf{U}}^i \leftarrow \widetilde{\mathbf{U}}^i, \breve{K}^i \leftarrow \widetilde{K}^i$
    end if
  end for

  $\widehat{\mathbf{\Sigma}} \leftarrow (\sum_i n_i - \sum_i \breve{K}_i)^{-1} \sum_{i=1}^m \breve{\mathbf{U}}^i \breve{\mathbf{U}}^{i\prime}$
  return $\{\breve{\mathbf{U}}^i\}_{i=1}^m, \{\breve{K}^i\}_{i=1}^m$ and $\widehat{\mathbf{\Sigma}}$
end procedure
```

Appendix B:: A key lemma

Recall that we defined

$$\breve{\mathbf{U}} = \mathbf{X}(\mathbf{I} - \frac{1}{n}\breve{\mathbf{F}}\breve{\mathbf{F}}') . \qquad \text{(B.1)}$$

where we used notations such as $\breve{\mathbf{F}}$ and $\breve{\mathbf{U}}$ 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.

The following lemma holds for $\breve{\mathbf{U}}$ no matter whether PCA or PPCA is applied.

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

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

where $\|\mathbf{\Pi}\|_{\max} = O_P(\sqrt{\log n}/n \cdot (||\mathbf{F}'(\breve{\mathbf{F}} - \mathbf{FH})||_{\max}\|\mathbf{\Lambda}\|_{\max} + ||\mathbf{U}(\breve{\mathbf{F}} - \mathbf{FH})||_{\max}) + ||\breve{\mathbf{F}} - \mathbf{FH}$; *and* $||_{\max}\|\mathbf{\Lambda}\|_{\max} + \sqrt{\log n} \cdot \|\mathbf{HH}' - \mathbf{I}\|_{\max}\|\mathbf{\Lambda}\|_{\max})$

furthermore

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

where

$\|\mathbf{\Delta}\|_{\max} = O_P(||\mathbf{U}(\breve{\mathbf{F}} - \mathbf{FH})||_{\max}\|\mathbf{\Lambda}\|_{\max} + ||\mathbf{U}(\breve{\mathbf{F}} - \mathbf{FH})||^2_{\max} + ||\mathbf{F}'(\breve{\mathbf{F}} - \mathbf{FH})||_{\max}||\mathbf{\Lambda}||^2_{\max} + n|\cdot$ $|\mathbf{HH}' - \mathbf{I}||_{\max}||\mathbf{\Lambda}||^2_{\max})$

The above lemma states that the error of estimating $\mathbf{U}$ by $\breve{\mathbf{U}}$ (or estimating $\mathbf{UU}'$ by $\breve{\mathbf{U}}\breve{\mathbf{U}}'$) 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 $\mathbf{\Pi}$ (or $\mathbf{\Delta}$). We will provide explicit rates of convergence for those terms in the following two sections.

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

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

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

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

Similarly $\mathbf{U}(\breve{\mathbf{F}}\breve{\mathbf{F}}' - \mathbf{F}\mathbf{F}') = \mathbf{U}(\breve{\mathbf{F}} - \mathbf{F}\mathbf{H})\breve{\mathbf{F}}' + \mathbf{U}\mathbf{F}\mathbf{H}(\breve{\mathbf{F}} - \mathbf{F}\mathbf{H})' + \mathbf{U}\mathbf{F}(\mathbf{H}\mathbf{H}' - \mathbf{I})\mathbf{F}'$, so

$$\left\|II\right\|_{\max} = O_P(\left\|\mathbf{U}'(\breve{\mathbf{F}} - \mathbf{F}\mathbf{H})\right\|_{\max}\left\|\breve{\mathbf{F}}/n\right\|_{\max} + \left\|\mathbf{U}\mathbf{F}/n\right\|_{\max}(\left\|\breve{\mathbf{F}} - \mathbf{F}\mathbf{H}\right\|_{\max} + \left\|\mathbf{H}\mathbf{H}' - \mathbf{I}\right\|_{\max}\left\|\mathbf{F}\right\|_{\max}))$$
.

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

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

$$\breve{\mathbf{U}}\breve{\mathbf{U}}' = \mathbf{U}(\mathbf{I} - n^{-1}\mathbf{F}\mathbf{F}')\mathbf{U}' + n^{-1}\mathbf{U}(\mathbf{I} - n^{-1}\mathbf{F}\mathbf{F}')(\breve{\mathbf{F}}\breve{\mathbf{F}}' - \mathbf{F}\mathbf{F}')\mathbf{X}' + n^{-2}\mathbf{X}(\breve{\mathbf{F}}\breve{\mathbf{F}}' - \mathbf{F}\mathbf{F}')^2\mathbf{X}'$$

$$= :\mathbf{U}\mathbf{U}' - \frac{1}{n}\mathbf{U}\mathbf{F}\mathbf{F}'\mathbf{U}' + III + IV\,.$$

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

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

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

$$\left\|\mathbf{J}_1\right\|_{\max} = O_P(\left\|\mathbf{U}(\breve{\mathbf{F}} - \mathbf{F}\mathbf{H})\right\|_{\max} + n^{-1}\left\|\mathbf{U}\mathbf{F}\right\|_{\max}\left\|\mathbf{F}'(\breve{\mathbf{F}} - \mathbf{F}\mathbf{H})\right\|_{\max}),$$

since $\|\breve{\mathbf{F}}'\mathbf{F}/n\|_{\max} \le \|\breve{\mathbf{F}}'\mathbf{F}/n\|_F \le \|\breve{\mathbf{F}}'\|_F\|\mathbf{F}\|_F/n = K$. Similar to $\mathbf{J}_1$, we decompose $\mathbf{J}_2$ only replacing $\breve{\mathbf{F}}'\mathbf{F}$ with $\breve{\mathbf{F}}'\mathbf{U}'$. According to Lemma F.4 (i), $\|\breve{\mathbf{F}}'\mathbf{U}'/n\|_{\max} = O_P(\|\mathbf{U}\mathbf{F}/n\|_{\max} + \|\mathbf{U}(\breve{\mathbf{F}} - \mathbf{F}\mathbf{H})\|_{\max}) = O_P(1 + \|\mathbf{U}(\breve{\mathbf{F}} - \mathbf{F}\mathbf{H})\|_{\max})$, hence $\|\mathbf{J}_2\|_{\max} = O_P((\|\mathbf{J}_1\|_{\max}(1 + \|\mathbf{U}(\breve{\mathbf{F}} - \mathbf{F}\mathbf{H})\|_{\max}))$. We then conclude that $\|III\|_{\max} = O_P((\|\mathbf{U}(\breve{\mathbf{F}} - \mathbf{F}\mathbf{H})\|_{\max} + n^{-1}\|\mathbf{U}\mathbf{F}\|_{\max}\|\mathbf{F}'(\breve{\mathbf{F}} - \mathbf{F}\mathbf{H})\|_{\max})(\|\mathbf{\Lambda}\|_{\max} + \|\mathbf{U}(\breve{\mathbf{F}} - \mathbf{F}\mathbf{H}.)\|_{\max}))$

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

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

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

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

Since $\mathbf{D}_2 = -n^{-2}\mathbf{UFF}'(\breve{\mathbf{F}} - \mathbf{FH})\breve{\mathbf{F}}'\mathbf{F\Lambda}' - n^{-1}\mathbf{UFH}(\breve{\mathbf{F}} - \mathbf{FH})'\mathbf{F\Lambda}' - \mathbf{UF}(\mathbf{HH}' - \mathbf{I})\mathbf{\Lambda}'$, we have

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

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

$$n^{-1}\|\mathrm{UF}\|_{\max}\|\mathbf{U}(\breve{\mathbf{F}} - \mathbf{FH})\|_{\max} = O_P(\|\mathbf{\Lambda}\|_{\max}^2).$$

Therefore we have

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

□

Appendix C:: Proof of Theorem 3.1

Recall that PCA estimates $\mathbf{F}$ by $\widehat{\mathbf{F}}$ where the k^{th} column of $\widehat{\mathbf{F}}/\sqrt{n}$ is the eigenvector of $(pn)^{-1}\mathbf{X}'X$ corresponding to the k^{th} largest eigenvalue. By the definition of $\widehat{\mathbf{F}}$, we have

$$\frac{1}{np}\mathbf{X}'\mathbf{X}\widehat{\mathbf{F}} = \widehat{\mathbf{F}}\mathbf{K},$$

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 [17]:

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

It has been shown that $\|\mathbf{K}\|$, $\|\mathbf{K}^{-1}\|$ and $\|\mathbf{H}\|$, $\|\mathbf{H}^{-1}\|$ are all $O_P(1)$.

The following lemma provides all the rates of convergences that are needed for downstream analysis.

Lemma C.1. *Under Assumptions 2.1 and 2.2, we have* $\|\mathbf{\Lambda}\|_{\max} = O_P(\sqrt{\log p})$ *and* *(i)* $||\widehat{\mathbf{F}} - \mathbf{FH}||_F = O_P(\sqrt{n/p} + 1/\sqrt{n})$ and $||\widehat{\mathbf{F}} - \mathbf{FH}||_{\max} = O_P(\sqrt{\log n/p} + \sqrt{\log n}/n)$;

(ii) $||\mathbf{F}'(\widehat{\mathbf{F}} - \mathbf{FH})||_{\max} = O_P(1 + \sqrt{n/p})$;

(iii) $||\mathbf{U}(\widehat{\mathbf{F}} - \mathbf{FH})||_{\max} = O_P((1 + n/p)\sqrt{\log p} + n||\mathbf{\Sigma}||_1/p)$;

(iv) $||\mathbf{HH}' - \mathbf{I}||_{\max} = O_P(1/n + 1/p)$.

Combining the above results with Lemma B.1, we have

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

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

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

where $||\Delta||_{\max} = O_P((1 + n/P)\log p + n^2||\mathbf{\Sigma}||_1^2/p^2)$. Thus we complete the proof for Theorem 3.1. We are left to check Lemma C.1, which is done in the following three subsections.

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

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

$$\widehat{\mathbf{F}} - \mathbf{FH} = (\sum_{i=1}^{3}\mathbf{E}_i)\mathbf{K}^{-1}, \tag{C.1}$$
$$\mathbf{E}_1 = \frac{1}{np}\mathbf{F}\mathbf{\Lambda}'\mathbf{U}\widehat{\mathbf{F}},\ \mathbf{E}_2 = \frac{1}{np}\mathbf{U}'\mathbf{\Lambda}\mathbf{F}'\widehat{\mathbf{F}},\ \mathbf{E}_3 = \frac{1}{np}\mathbf{U}'\mathbf{U}\widehat{\mathbf{F}}.$$

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

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

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.2. *(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 $\|\mathbf{E}_1\|_F \le p^{-1}\|\mathbf{\Lambda}'\mathbf{U}\|_F = O_P(\sqrt{n/p})$ according to Lemma F.1. $\|\mathbf{E}_2\|_F$ attains the same rate. In addition, $\|\mathbf{E}_3\|_F \le n^{-1/2}p^{-1}\|\mathbf{U}'\mathbf{U}\|_F = O_P(1 + \sqrt{n/p})$ 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 $\|\mathbf{E}_3\|_F . \|\mathbf{E}_3\|_F \le (np)^{-1}(\|\mathbf{U}'\mathbf{UF}\|_F\|\mathbf{H}\|_F + \|\mathbf{U}'\mathbf{U}\|_F\|\widehat{\mathbf{F}} - \mathbf{FH}\|_F) = O_P(1/\sqrt{n} + 1/\sqrt{p} + \sqrt{n}/p)$. Then the refined rate of $\|\widehat{\mathbf{F}} - \mathbf{FH}\|_F$ is $O_P(\sqrt{n/p} + 1/\sqrt{n})$.

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

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

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

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

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

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.

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

Proof. The results follow from Lemmas C.2. □

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

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

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

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

Finally,

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

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

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

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

Therefore $\|\mathbf{HH}' - \mathbf{I}\|_F$ has the same rate since $\|\mathbf{HH}' - \mathbf{I}\|_F \le \|\mathbf{H}\|_F\|\mathbf{H}'\mathbf{H} - \mathbf{I}\|_F\|\mathbf{H}^{-1}\|_F$. 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.

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

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

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

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

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

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

Appendix D:: Proof of Theorem 3.2

Recall that by the definition of $\widetilde{\mathbf{F}}$, we have

$$\frac{1}{np}\mathbf{X}'\mathbf{P}\mathbf{X}\widetilde{\mathbf{F}} = \widetilde{\mathbf{F}}\mathbf{K},$$

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 [18]:

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

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 defined in the previous section, they play essentially the same roles (thus with same notations).

The following lemma provides all the rates of convergences that are needed for downstream analysis.

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

(i) $\|\widetilde{\mathbf{F}} - \mathbf{F}\mathbf{H}\|_F = O_P(\sqrt{n/p})$ *and* $\|\widetilde{\mathbf{F}} - \mathbf{F}\mathbf{H}\|_{\max} = O_P(\sqrt{\log n/p})$;

(ii) $\|\mathbf{F}'(\widetilde{\mathbf{F}} - \mathbf{F}\mathbf{H})\|_{\max} = O_P(\sqrt{n/p} + n/p + n\sqrt{\nu_p/p})$;

(iii) $\|\mathbf{U}(\widetilde{\mathbf{F}} - \mathbf{F}\mathbf{H})\|_{\max} = O_P(\sqrt{n\log p/p} + nJ\phi_{\max}\|\Sigma\|_1/p)$;

(iv) $\|\mathbf{H}\mathbf{H}' - \mathbf{I}\|_{\max} = O_P(1/p + 1/\sqrt{pn} + \sqrt{\nu_p/p})$.

Combining the above lemma with Lemma B.1, we obtain

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

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

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

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

Thus we complete the proof for Theorem 3.2. We are left to check Lemma D.1, which is done in the following three subsections.

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

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

$$\widetilde{\mathbf{F}} - \mathbf{F}\mathbf{H} = (\sum_{i=1}^{15}\mathbf{A}_j)\mathbf{K}^{-1} \tag{D.1}$$

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

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

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

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

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

$$\begin{aligned}&\mathbf{A}_9 = \frac{1}{np}\mathbf{F}\mathbf{B}'\ \Phi\ (\mathbf{W})'\mathbf{\Gamma}\mathbf{F}'\widetilde{\mathbf{F}},\ \mathbf{A}_{10} = \frac{1}{np}\mathbf{F}\mathbf{\Gamma}'\mathbf{\Phi}(\mathbf{W})\mathbf{B}\mathbf{F}'\widetilde{\mathbf{F}},\\ &\mathbf{A}_{11} = \frac{1}{np}\mathbf{F}\mathbf{\Gamma}'\mathbf{P}\mathbf{\Gamma}\mathbf{F}'\widetilde{\mathbf{F}},\ \mathbf{A}_{12} = \frac{1}{np}\mathbf{F}\mathbf{\Gamma}'\mathbf{P}\mathbf{U}\widetilde{\mathbf{F}},\\ &\mathbf{A}_{13} = \frac{1}{np}\mathbf{U}'\mathbf{P}\mathbf{\Gamma}\mathbf{F}'\widetilde{\mathbf{F}},\ \mathbf{A}_{14} = \frac{1}{np}\mathbf{F}\mathbf{R}'\mathbf{P}\mathbf{\Gamma}\mathbf{F}'\widetilde{\mathbf{F}}, \mathbf{A}_{15} = \frac{1}{np}\mathbf{F}\mathbf{\Gamma}'\mathbf{P}\mathbf{R}\mathbf{F}'\widetilde{\mathbf{F}}\,.\end{aligned}$$

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

Lemma D.2. *(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) $\|\mathbf{A}_4\|_{\max} = O_P(J^{-k/2}\sqrt{\log n}) = \|\mathbf{A}_5\|_{\max}$ *and* $\|\mathbf{A}_9\|_{\max} = O_P(\sqrt{\nu_p \log n/p}) = \|\mathbf{A}_{10}\|_{\max}$,

(iv) $\left\|\mathbf{A}_6\right\|_{\max} = O_P(J^{-k}\sqrt{\log n})$ *and* $\left\|\mathbf{A}_{11}\right\|_{\max} = O_P(J\nu_p\sqrt{\log n}/p)$,

(v) $\left\|\mathbf{A}_7\right\|_{\max} = O_P(\phi_{\max}\sqrt{p^{-1}J^{1-k}\log(nJ)\log n}) = \left\|\mathbf{A}_8\right\|_{\max}$ *and* $\left\|\mathbf{A}_{12}\right\|_{\max} = O_P(J\phi_{\max}\sqrt{\nu_p\log(nJ)\log n/p}) = \left\|\mathbf{A}_{13}\right\|_{\max}$,

(vi) $\left\|\mathbf{A}_{14}\right\|_{\max} = O_P(\sqrt{p^{-1}J^{1-k}\nu_p\log n}) = \left\|\mathbf{A}_{15}\right\|_{\max}$.

Proof. (i) Because $\left\|\mathbf{F}\right\|_{\max} = O_P(\sqrt{\log n})$, $\|\widetilde{\mathbf{F}}\|_F = O_P(\sqrt{n})$. By Lemma 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

$$\left\|\mathbf{A}_1\right\|_{\max} \le \frac{\sqrt{K}}{np}\left\|\mathbf{F}\right\|_{\max}\|\mathbf{B}'\ \Phi\ (\mathbf{W})'\mathbf{U}\|_F\|\widetilde{\mathbf{F}}\|_F = O_P(\sqrt{\log n/p}),$$
$$\left\|\mathbf{A}_2\right\|_{\max} \le \frac{\sqrt{K}}{np}\|\mathbf{U}'\ \Phi\ (\mathbf{W})\mathbf{B}\|_{\max}\left\|\mathbf{F}\right\|_F\|\widetilde{\mathbf{F}}\|_F = O_P(\sqrt{\log n/p}).$$

(ii) We have $\mathbf{A}_3 = \frac{1}{np}\mathbf{U}'\ \Phi\ (\mathbf{W})(\ \Phi\ (\mathbf{W})'\ \Phi\ (\mathbf{W}))^{-1}\ \Phi\ (\mathbf{W})'\mathbf{U}\widetilde{\mathbf{F}}$. 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, $\left\|(\ \Phi\ (\mathbf{W})'\ \Phi\ (\mathbf{W}))^{-1}\right\|_2 = O_P(p^{-1})$. Note the fact that for matrix $\mathbf{A}_{m\times n}$, $\mathbf{B}_{n\times n}$, $\mathbf{C}_{n\times r}$, $\left\|\mathbf{ABC}\right\|_{\max} = \max_{i\le m, k\le r}|\mathbf{a}_i'\mathbf{B}\mathbf{c}_k| \le \sqrt{n}\left\|\mathbf{A}\right\|_{\max}\left\|\mathbf{B}\right\|_2\left\|\mathbf{C}\right\|_F$. So

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

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

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

Similarly, $\left\|\mathbf{A}_5\right\|_{\max}$ 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 $\boldsymbol{\Gamma}$. It is not hard to see $\left\|\mathbf{B}'\ \Phi'\ \boldsymbol{\Gamma}\right\|_{\max} = O_P(\sqrt{p\nu_p})$. Therefore $\left\|\mathbf{A}_9\right\|_{\max} = O_P(\sqrt{\nu_p\log n/p}) = \left\|\mathbf{A}_{10}\right\|_{\max}$.

(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 $\|\mathbf{R}(\mathbf{W})'\mathbf{P}\mathbf{R}(\mathbf{W})\|_{\max} \le p\|\mathbf{R}(\mathbf{W})\|_{\max}^2\|\mathbf{P}\|_2 = O_P(pJ^{-\kappa})$. Hence

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

$\mathbf{A}_{11}$ has similar representation as $\mathbf{A}_6$. Since

$$\|\mathbf{\Gamma}'\mathbf{P}\mathbf{\Gamma}\|_{\max} \le \|\Phi'\mathbf{\Gamma}\|_F^2\|(\Phi'\Phi)^{-1}\|_2 = O_P(J\nu_p),$$

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

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

since $\|\mathbf{R}'\Phi\|_F \le \|\mathbf{R}\|_F\|\Phi\|_2 = O_P(pJ^{-k/2})$. 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 $||\mathbf{R}'\mathbf{P}\mathbf{\Gamma}||_{\max} \le ||\mathbf{R}'\Phi||_F||(\Phi'\Phi)^{-1}||\|\Phi'\mathbf{\Gamma}\|_F$. We conclude $\|\mathbf{A}_{14}\|_{\max} = O_P(\sqrt{p^{-1}J^{1-k}\nu_p\log n})$. Same bound holds for $\mathbf{A}_{15}$. □

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

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

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

Proof. The max norm result follows from Lemmas D.2 and (D.1), while the Frobenius norm result has been shown in [18]. □

D.2. Rates of $\|\mathbf{F}'(\widetilde{\mathbf{F}} - \mathbf{FH})\|_{\max}$ 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 [18] that $\|\mathbf{F}'(\widetilde{\mathbf{F}} - \mathbf{FH})\|_F = O_P(\sqrt{n/p} + n/p + n\sqrt{\nu_p/p} + nJ^{-k/2})$. By the choice of J, the last term vanishes. So

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

[18] 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
$\|\mathbf{HH}' - \mathbf{I}\|_{\max} \le \|\mathbf{HH}' - \mathbf{I}\|_F \le \|\mathbf{H}\|\|\mathbf{H}'\mathbf{H} - \mathbf{I}\|_F\|\mathbf{H}^{-1}\| = O_P(1/p + 1/\sqrt{pn} + \sqrt{\nu_p/p})$ since $J^{\kappa} \ge p/\nu_p$.

D.3. Rate of $\|\mathbf{U}(\widetilde{\mathbf{F}} - \mathbf{FH})\|_{\mathbf{m}ax}$

By (D.1), in order to bound $\|\mathbf{U}(\widetilde{\mathbf{F}} - \mathbf{FH})\|_{\max}$ we essentially need to bound $\|\mathbf{UA}_i\|_{\max}$ for $i = 1, \ldots, 15$. We do not bother going into the details of each term again as in Lemma D.2. 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.2, we bound $\|\mathbf{A}_i\|_{\max}$ in using $\sqrt{K}\|\mathbf{F}\|_{\max}\|\mathbf{Q}\|_F$. Now we use bound $\|\mathbf{UA}_i\|_{\max} \le \sqrt{K}\|\mathbf{UF}\|_{\max}\|\mathbf{Q}\|_F$ 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.2, 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 $\|\mathbf{UU}'\ \Phi\ (\mathbf{W})\|_{\max} = O_P(\phi_{\max}(\sqrt{np\log p} + n\|\mathbf{\Sigma}\|_1))$ and $\|\mathbf{UU}'\ \Phi\ (\mathbf{W})\mathbf{B}\|_{\max} = O_P(\sqrt{np\log p} + nJ\phi_{\max}\|\mathbf{\Sigma}\|_1)$.

So we replace the corresponding terms in Lemma D.2. It is not hard to see the dominating term is $\|\mathbf{UA}_2\|_{\max} = O_P(\sqrt{n\log p/p} + nJ\phi_{\max}\|\mathbf{\Sigma}\|_1/p)$. Therefore, $\|\mathbf{U}(\widetilde{\mathbf{F}} - \mathbf{FH})\|_{\max}$ has the same rate.

Appendix E:: Proof of Theorem 4.1

Proof. Denote the oracle empirical covariance matrix as

$$\mathbf{\Sigma}_N = \frac{1}{N}\sum_{i=1}^{m}\mathbf{U}^i\mathbf{U}^{i'} .$$

As in [9] the upper bound on $\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}\|$ is obtained by proving

$$\left\|(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}_N)\ \mathbf{\Omega}\right\|_{\max} = O_p(\tau_{m,N,p}) \text{ and } \left\|(\mathbf{\Sigma}_N - \mathbf{\Sigma})\mathbf{\Omega}\right\|_{\max} = O_p(\tau_{m,N,p}) . \tag{E.1}$$

Once the two bounds are established, we proceed by observing

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

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

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

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 $\|\widehat{\mathbf{\Omega}}\|_1$ by $\|\mathbf{\Omega}\|_1$. So it remains to find $\tau_{m,N,p}$ in (E.1). Since $\mathbf{\Omega} \in \mathscr{F}(s, C_0), \|\mathbf{\Omega}\|_1 \le C_0$, so we just need to bound $\left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}_N\right\|_{\max}$ and $\|\mathbf{\Sigma}_N - \mathbf{\Sigma}\|_{\max}$. Obviously,

$$\|\mathbf{\Sigma}_N - \mathbf{\Sigma}\|_{\max} = O_p(\sqrt{\frac{\log p}{N}}) .$$

We have shown in (4.3) that $\widehat{\mathbf{\Sigma}}$ given by (4.2) attains the rate $\left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}_N\right\|_{\max} = O_P(a_{m,N,p})$. Thus $\tau_{m,N,p} = \sqrt{\log p/N} + a_{m,N,p}$. Similar proof as in [9] can also reach error bounds under $\left\|\ \cdot\ \right\|_1$ and $\left\|\ \cdot\ \right\|_2$, which we omit. The proof is now complete. □

Appendix F:: Technical lemmas

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

(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.

$$\mathbb{E}\left\|\boldsymbol{\Lambda}'\mathbf{U}\right\|_F^2 = \mathbb{E}[\mathrm{tr}(\boldsymbol{\Lambda}'\mathbf{U}\mathbf{U}'\boldsymbol{\Lambda})] = n\cdot\mathrm{tr}(\boldsymbol{\Lambda}'\boldsymbol{\Sigma}\boldsymbol{\Lambda}) \le n\left\|\boldsymbol{\Sigma}\right\|\cdot\mathrm{tr}(\boldsymbol{\Lambda}'\boldsymbol{\Lambda}) = O(np)\,.$$

$$\begin{aligned}
\mathbb{E}\left\|\mathrm{U}'\mathrm{U}\right\|_F^2 &= \mathbb{E}\left[\sum_{t=1}^{n}\sum_{t'=1}^{n}\Big(\sum_{j=1}^{p}u_{jt}u_{jt'}\Big)^2\right]\\
&= \sum_{j_1,j_2=1}^{p}\left(\sum_{t=1}^{n}\mathbb{E}\left[u_{j_1t}^2u_{j_2t}^2\right] + \sum_{1\le t\ne t_1\le n}\sigma_{j_1j_2}^2\right)\\
&= O_P(np^2 + pn^2),
\end{aligned}$$

since $\sum_{j_1,j_2}\sigma_{j_1j_2}^2 = \mathrm{tr}(\boldsymbol{\Sigma}^2) \le \left\|\boldsymbol{\Sigma}\right\|\mathrm{tr}(\Sigma) = O(p)\,.$

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{U}'\mathbf{U}\mathbf{F}\right\|_F^2 &= \mathbb{E}\left[\sum_{t=1}^{n}\sum_{k=1}^{K}\Big(\sum_{t'=1}^{n}\sum_{j=1}^{p}u_{jt}u_{jt'}f_{t'k}\Big)^2\right]\\
&= \sum_{k=1}^{K}\sum_{j_1,j_2=1}^{p}\left(\sum_{t=1}^{n}\mathbb{E}\left[u_{j_1t}^2u_{j_2t}^2\right]f_{tk}^2 + \sum_{1\le t\ne t_1\le n}\sigma_{j_1j_2}^2f_{t_1k}^2\right)\\
&= O_P(np^2 + pn^2)\,.
\end{aligned}$$

□

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

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

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

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

(ii) $\left\|\mathbf{U}'\mathbf{U}\right\|_{\max} = \max_{t,t'}|\mathbf{u}_t'\mathbf{u}_{t'}| \le \max_{t\ne t'}|\mathbf{u}_t'\mathbf{u}_{t'}| + \max_t|\mathbf{u}_t'\mathbf{u}_t|$. We need tobound each term separately. The second term is bounded by the upper tail bound of Hanson-Wright inequality for sub-Gaussian vector [24, 41] i.e.

$$\mathbb{P}(\left\|\mathbf{u}_t\right\|^2 > \mathrm{tr}(\boldsymbol{\Sigma}) + 2\sqrt{\mathrm{tr}(\boldsymbol{\Sigma})s} + 2\left\|\boldsymbol{\Sigma}\right\|s) \le e^{-s}\,.$$

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

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

where $\mathbb{E}\left[\exp(\theta \mathbf{u}_t'\mathbf{u}_{t'})\right] = E\left[\exp(\theta^2 \mathbf{u}_t'\boldsymbol{\Sigma}\mathbf{u}_t/2)\right] \leq E\left[\exp(C\theta^2 \|\mathbf{u}_t\|^2)\right]$. [24] showed that

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

For $\eta < 1/(4\|\boldsymbol{\Sigma}\|) \leq \mathrm{tr}(\boldsymbol{\Sigma})/(4\mathrm{tr}(\boldsymbol{\Sigma}^2))$, the right hand side is less than $\exp(3\mathrm{tr}(\Sigma)\eta/2) \leq \exp(Cp\eta)$. 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 $\eta < 1/(4\|\boldsymbol{\Sigma}\|)$ 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 $O_p(p \max_{tk} |f_{tk}|) = O_p(p\sqrt{\log n})$. Define $\boldsymbol{\xi} = \mathbf{U}_{(-t)}\bar{\mathbf{f}}_{k(-t)} \sim$ subGaussian $(\mathbf{0}, \boldsymbol{\Sigma}\|\bar{\mathbf{f}}_{k(-t)}\|^2)$, which is independent with $\mathbf{u}_t$. Thus

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

where $\mathbb{E}\left[\exp(\theta \mathbf{u}_t'\boldsymbol{\xi})\right] \leq \mathbb{E}\left[\exp(\theta^2 \mathbf{u}_t'\boldsymbol{\Sigma}\mathbf{u}_t \|\bar{\mathbf{f}}_{k(-t)}\|^2/2)\right] \leq \mathbb{E}\left[\exp(C\theta^2 n \|\mathbf{u}_t\|^2)\right]$. 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 $\|\mathbf{U}'\mathbf{U}\mathbf{F}\|_{\max} = O_p(\sqrt{np \log n} + p\sqrt{\log n})$. □

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 $\|\mathbf{\Sigma}\|_2$ instead of $\|\mathbf{\Sigma}\|_1$. By Markov inequality, it is sufficient to show the expected value of each term attains the corresponding rate of convergence.

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

$$\mathbb{E}\|\mathbf{U}'\ \Phi\ (\mathbf{W})\|_F^2 = \mathbb{E}[\mathrm{tr}(\ \Phi'\ \mathbb{E}[\mathbf{U}\mathbf{U}'|\mathbf{W}]\ \Phi)] = n \cdot \mathbb{E}[\mathrm{tr}(\ \Phi'\ \mathbf{\Sigma}\ \Phi)] \le nJd \cdot \mathbb{E}\left[\|\ \Phi'\ \mathbf{\Sigma}\ \Phi\|_2\right]$$
$$\le nJdC_0\mathbb{E}\left[\|\ \Phi'\ \ \Phi\|_2\right] = O(npJ).$$

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

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

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

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

$$E\left[e^{\theta\sum_{t=1}^n f_{tk}u_{jt}}\right] = \mathbb{E}\left[\mathbb{E}\left[e^{\theta\sum_{t=1}^n f_{tk}u_{jt}}\middle|\mathbf{F}\right]\right] \le \mathbb{E}\left[e^{\theta^2\sum_{t=1}^n f_{tk}^2\sigma_{jj}/2}\right] \le e^{\frac{1}{2}nC_0\theta^2},$$

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 $\|\mathbf{\Sigma}\|_2 \le C_0$. Therefore, choosing $\theta = \frac{t}{nC_0}$, we have

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

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 $\bar{\phi}_{\nu l} = (\phi_\nu(\mathbf{W}_{1l}), \ldots, \phi_\nu(\mathbf{W}_{pl}))'$. Consider the tail probability condition on $\mathbf{W}$:

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

The right hand side can be further bounded by

$$2Jdn \cdot \exp(-t\theta + \frac{1}{2}\theta^2 C_0 \|\bar{\phi}_{\nu l}\|^2) \le 2Jdn \cdot \exp(-t\theta + \frac{1}{2} p C_0 \theta^2 \phi^2_{\max}).$$

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

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

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 $\|\mathbf{U}' \Phi(\mathbf{W})\mathbf{B}\|_{\max} \le \|\mathbf{U}'\mathbf{G}(\mathbf{W})\|_{\max} + \|\mathbf{U}'R(\mathbf{W})\|_{\max}$ and the first term dominates. So the same derivation gives

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

where $\bar{g}_k = (g_k(\mathbf{W}_1), \ldots, g_k(\mathbf{W}_p))$. $\|\bar{g}_k\|^2 = O_p(p)$ 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) $\|\Phi(\mathbf{W})'\mathbf{UF}\|_{\max} = \max_{\nu \le J, l \le d, k \le K} |\sum_{j=1}^{p} \sum_{i=1}^{n} \phi_\nu(W_{jl}) u_{ji} f_{ik}|$. Using Chernoff bound again, we get

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

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

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

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

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

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

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

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

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

$$\begin{aligned}&\mathbb{P}(\max_{j,k}|\sum_{t=1}^{n}(\mathbf{a}'_j\mathbf{v}_t\mathbf{v}'_t\mathbf{d}_k - \mathbf{a}'_j\mathbf{d}_k)| > s)\\ &\le pK\mathbb{P}(|\sum_{t=1}^{n}(\tilde{\mathbf{a}}'_j\mathbf{v}_t\mathbf{v}'_t\tilde{\mathbf{d}}_k - \tilde{\mathbf{a}}'_j\tilde{\mathbf{d}}_k)| > \frac{s}{\max_{j,k}\|\mathbf{a}_j\|\|\mathbf{d}_k\|}),\end{aligned} \tag{F.1}$$

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 $\tilde{\mathbf{d}}_k$.

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

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

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

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 $\max_{jk}\|\mathbf{a}_j\|\|\mathbf{d}_k\| \le C\max_k\|\boldsymbol{\lambda}_k\|$, $\max_{j,k}|\sum_{t=1}^{n}u_{jt}\mathbf{u}_t'\boldsymbol{\lambda}_k - n\sum_{j'=1}^{p}\sigma_{jj'}\lambda_{j',k}| = O_P(\sqrt{n\log p}\max_k\|\boldsymbol{\lambda}_k\|) = O_P(\sqrt{np\log p})$. 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 $\boldsymbol{\Phi}_{\nu l} = (\phi_\nu(W_{1l}),\ldots,\phi_\nu(W_{pl}))'$, so

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

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

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

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

Appendix G:: More details on synthetic data analysis

G.1. Model calibration and data generation

We calibrate (estimate) the 264 by 264 covariance matrix $\widehat{\boldsymbol{\Sigma}}$ of $\mathbf{u}_t$ by our proposed method to the data in the healthy group. Plugging it as input in CLIME solver delivers a sparse precision matrix $\boldsymbol{\Omega}$, which will be taken as truth in the simulation. Note that after regularization in CLIME, $\boldsymbol{\Omega}^{-1}$ is not the same as $\widehat{\boldsymbol{\Sigma}}$, and we set the true covariance $\boldsymbol{\Sigma} = \boldsymbol{\Omega}^{-1}$. 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 × 100 matrix multiple times to form a $p \times p$ block diagonal matrix and use 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 $j \le p$, generate the external covariate W i.i.d. from the multinomial distribution with $\mathbb{P}(W_j = s) = w_s, s \le 10$ where $\{w_s\}_{s=1}^{10}$ are calibrated with the hierarchy clustering results of the real data.

2. **(Calibration)** For the first 15 healthy subjects, obtain estimators for $\mathbf{F}$, $\mathbf{B}$ and $\boldsymbol{\Gamma}$ by PPCA, resulting in $\widetilde{\mathbf{F}}, \widetilde{\mathbf{B}} = n^{-1}(\Phi(\mathbf{W})' \Phi(\mathbf{W}))^{-1} \Phi(\mathbf{W})'\mathbf{X}\widetilde{\mathbf{F}}$ and $\widetilde{\boldsymbol{\Gamma}} = n^{-1}(\mathbf{I} - \mathbf{P})\mathbf{X}\widetilde{\mathbf{F}}$ according to [18]. Use the rows of the estimated factors to fit a stationary VAR model $\mathbf{f}_t = \mathbf{A}\mathbf{f}_{t-1} + \epsilon_t$, where $\epsilon_t \sim N(0, \Sigma_\epsilon)$, and obtain the estimators $\widetilde{\mathbf{A}}$ and $\widetilde{\boldsymbol{\Sigma}}_\epsilon$.

3. **(Simulation)** For each subject $i \le m$, pick one of the 15 calibrated models and their associated parameters from above at random and do the following.

   (a) Generate γ_{jk}^i (entries of $\boldsymbol{\Gamma}^i$) i.i.d. from $N(0, \tilde{\sigma}_\gamma^2)$ where $\tilde{\sigma}_\gamma^2$ is the sample variance of all entries of $\widetilde{\boldsymbol{\Gamma}}$. For the first three settings, compute the 'true' loading matrix $\boldsymbol{\Lambda}^i = \Phi(\mathbf{W})\widetilde{\mathbf{B}} + \boldsymbol{\Gamma}^i$. For the last setting, set $\boldsymbol{\Lambda}^i = \boldsymbol{\Gamma}^i$ since $\mathbf{G}(\mathbf{W}) = 0$.

   (b) Generate factors $\mathbf{f}_t^i$ from the VAR model $\mathbf{f}_t^i = \widetilde{\mathbf{A}}\mathbf{f}_{t-1}^i + \in_t$ with $\in_t \sim N(0, \widetilde{\boldsymbol{\Sigma}}_\in)$ where the parameters $\widetilde{\mathbf{A}}$ and $\widetilde{\boldsymbol{\Sigma}}_\in$ are taken from the fitted values in step 2.

   (c) Finally, generate the observed data $\mathbf{X}^i = \boldsymbol{\Lambda}^i\mathbf{F}^{i'} + \mathbf{U}^i$, where each column of $\mathbf{U}^i$ is randomly sampled from $N(\mathbf{0}, \boldsymbol{\Omega}^{-1})$, where $\boldsymbol{\Omega}$ has been calibrated by the CLIME solver as described at beginning of the section.

G.2. More on pervasiveness

In this subsection, we discuss the pervasive assumption, which requires the spikes to grow with order p, and present numerical performance of ALPHA for different levels of $c_{\min}$ and

$c_{\max}$ (defined in Assumption 2.3). The readers will have a rough idea about how the spikiness (or the constant in front of the rate) affects the performance. We particularly consider the cases when $c_{\max}$ is small or $c_{\min}$ is large. As a threshold matter, we verify that the real data is consistent with the pervasive assumption.

Denote the maximum and minimum eigenvalues of the matrix $\mathbf{\Lambda}'\mathbf{\Lambda}/p$ by $\lambda_{\max}$ and $\lambda_{\min}$ respectively, and denote the maximum eigenvalue of the matrix $\mathbf{U}'\mathbf{U}/p$ by $\lambda^{u}_{\max}$. We first investigate the magnitude of $\lambda_{\min}$, $\lambda_{\max}$ and $\lambda^{u}_{\max}$ derived from the real data. Following exactly the same data generation procedure as in the original simulation study, we randomly generate 1,000 subjects. We find that $\lambda_{\max}$ has mean 15.352 and standard deviation 4.918, $\lambda_{\min}$ has mean 10.069 and standard deviation 5.416 and $\lambda^{u}_{\max}$ has mean 1.317 and standard deviation 0.119. We also investigate the signal-to-noise ratio $\lambda_{\min}/\lambda^{u}_{\max}$, which has mean 7.711 and standard deviation 4.230. Therefore, our real data demonstrates a spiked covariance structure while the spikes are not extremely spiky.

Then we manipulate the data generation process correspondent to two different cases. One is to multiply the original loading matrix $\mathbf{\Lambda}$ by 3, called Modified (a), while the other is to divide $\mathbf{\Lambda}$ by 3, called Modified (b). Note that in the case of Modified (b), $\lambda_{\min}$ will be 1/9 of the original $\lambda_{\min}$ and thus smaller than $\lambda^{u}_{\max}$, so we do not see a clear eigen-gap in this case. Table 3 compares the performance of recovering the precision matrix $\mathbf{\Omega}$ under the original and modified setting when $n_i = 100$.

We can see from the table above that the performance of ALPHA in the case of Modified (a) is slightly better than that in the original case. Note that increasing $c_{\min}$ makes the heterogeneity part more spiky. Larger $c_{\min}$ allows PCA or PPCA to distinguish the spiky heterogeneity term more easily. In contrast, decreasing $c_{\max}$ makes the original spiky heterogeneity term hard to detect. We also tend to miss several heterogeneity factors while extracting them. Therefore, in Modified (b), the estimation error becomes significantly larger compared with the original case.

G.3. Sensitivity analysis on the number of factors

In this section, we study how the estimated number of factors affects the recovery of the Gaussian graphical model through simulations. The specification of the number of factors is critical to the validity of our ALPHA method, which inspires us to assess the performance of $\hat{K}$ and $\tilde{K}$ on our simulated datasets in the first place. Recall that

$$\hat{K} = \arg\max_{k \le K_{\max}} \lambda_k(\mathbf{X}'\mathbf{X})/\lambda_{k+1}(\mathbf{X}'\mathbf{X}),$$
$$\tilde{K} = \arg\max_{k \le K_{\max}} \lambda_k(\mathbf{X}'\mathbf{P}\mathbf{X})/\lambda_{k+1}(\mathbf{X}'\mathbf{P}\mathbf{X}),$$

where $\mathbf{P}$ is the projection operator defined in (3.5) in the main text. The final estimator of the number of factors, denote by $\check{K}$, comes from the heuristic strategy we developed for choosing between PCA or PPCA. We choose PCA if

$\lambda_{\widehat{K}}(\mathbf{X}'\mathbf{X})/\lambda_{\widehat{K}+1}(\mathbf{X}'\mathbf{X}) \geq \lambda_{\widetilde{K}}(\mathbf{X}'\mathbf{PX})/\lambda_{\widetilde{K}+1}(\mathbf{X}'\mathbf{PX})$ and choose PPCA vice versa. The intuition is that we favor the method that yields larger eigen-ratio between the spiked and non-spiked part of the covariance.

Analogous to the simulation study in our paper, we generate $m = 1{,}000$ people's BOLD data based on calibrated "true" data. We investigate the accuracy of the proposed $\widehat{K}$, $\widetilde{K}$ and $\check{K}$ for two cases: (i) $n_i = 20, p = 264$ and (ii) $n_i = 100$, $p = 264$, presented in Table 4. As we can see from the table, when n_i is small, $\widetilde{K}$ outperforms $\widehat{K}$, and when n_i is large, $\widehat{K}$ is better. Note also that our heuristic estimator $\check{K}$ has great performance in both cases of large and small n_i.

Given the performance of our proposed estimators of the factor number, we now artificially enlarge this estimation error and see how it affects the Gaussian graphical model analysis. Let η be a random perturbation with $P(\eta = 0) = 1/2$, $P(\eta = 1) = 1/3$ and $P(\eta = 2) = 1/6$. Define $K^{+} := K + \eta$ and $K^{-} := \max(K - \eta, 0)$, where K is the true number of factors. As the notations indicate, K^{+} overestimates the factor number while K^{-} underestimates it. Since $P(\eta \neq 0) = 1/2$, their estimation accuracy is only 50%, worse than that of $\widehat{K}$ and $\widetilde{K}$ as presented. We use K^{+} and K^{-} as the estimators of the number of factors respectively to recover the precision matrix of $\mathbf{U}$ and compare their performance with that of $\check{K}$. The results are presented in Table 5.

"Oracle" above means that we directly use the generated noise $\mathbf{U}$ to calculate its sample covariance and plug it in CLIME to recover the precision matrix. K^{o} means we know the true number of pervasive factors, and use PCA or Projected-PCA (choosing the method that yields larger eigen-ratio) to adjust factors. As we can see from the table above, K^{+} is nearly as good as K^{o}, which means that overestimating the number of factors does not hurt the recovery accuracy. In contrast, underestimating the number factors will seriously increase the estimation error of $\mathbf{\Omega}$, as shown by K^{-}, because the unadjusted pervasive factors heavily corrupt the covariance of $\mathbf{U}$. Nevertheless, both K^{+} and K^{-} uses partial information of the true number of factors. In comparison, our procedure $\check{K}$, without any prior knowledge about the number of factors, have a great performance in recovering $\mathbf{\Omega}$.

References

[1]. Ahn SC and Horenstein AR (2013). Eigenvalue ratio test for the number of factors. Econometrica 81 1203–1227. MR3064065

[2]. Alter O, Brown PO 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.

[3]. Bai J (2003). Inferential theory for factor models of large dimensions. Econometrica 71 135–171. MR1956857

[4]. Bai J and Ng S (2002). Determining the number of factors in approximate factor models. Econometrica 70 191–221. MR1926259

[5]. Bai J and Ng S (2013). Principal components estimation and identification of static factors. Journal of Econometrics 176 18–29. MR3067022

[6]. Biswal BB, Mennes M, Zuo X-N, Gohel S, Kelly C, Smith SM, Beckmann CF, Adelstein JS, Buckner RL and Colcombe S (2010). Toward discovery science of human brain function. Proceedings of the National Academy of Sciences 107 4734–4739.

[7]. Cai TT, Li H, Liu W and Xie J (2012). Covariate-adjusted precision matrix estimation with an application in genetical genomics. Biometrika ass058 MR3034329

[8]. Cai TT, Li H, Liu W and Xie J (2015). Joint estimation of multiple high-dimensional precision matrices. The Annals of Statistics 38 2118–2144. MR3497754

[9]. Cai TT, Liu W and Luo X (2011). A constrained 1 minimization approach to sparse precision matrix estimation. Journal of the American Statistical Association 106 594–607. MR2847973

[10]. Cai TT, Ma Z and Wu Y (2013). Sparse PCA: Optimal rates and adaptive estimation. The Annals of Statistics 41 3074–3110. MR3161458

[11]. Chen C, Grennan K, Badner J, Zhang D, Gershon E, Jin L and Liu C (2011). Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. PloS one 6 e17238. [PubMed: 21386892]

[12]. Chen X (2007). Large sample sieve estimation of semi-nonparametric models. Handbook of Econometrics 6 5549–5632.

[13]. Connor G, Hagmann M and Linton O (2012). Efficient semiparametric estimation of the fama–french model and extensions. Econometrica 80 713–754. MR2951947

[14]. 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.

[15]. Danaher P, Wang P and Witten DM (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76 373–397. MR3164871 [PubMed: 24817823]

[16]. Fan J, Ke Y and Wang K (2016). Decorrelation of covariates for high dimensional sparse regression. arXiv preprint arXiv:1612.08490

[17]. 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. MR3091653

[18]. Fan J, Liao Y and Wang W (2016). Projected principal component analysis in factor models. The Annals of Statistics 44 219–254. MR3449767 [PubMed: 26783374]

[19]. Fan J, Rigollet P and Wang W (2015). Estimation of functionals of sparse covariance matrices. Annals of statistics 43 2706 MR3405609 [PubMed: 26806986]

[20]. Friedman J, Hastie T and Tibshirani R (2008). Sparse inverse covariance estimation with the graphical Lasso. Biostatistics 9 432–441. [PubMed: 18079126]

[21]. Guo J, Cheng J, Levina E, Michailidis G and Zhu J (2015). Estimating heterogeneous graphical models for discrete data with an application to roll call voting. The annals of applied statistics 9 821 MR3371337

[22]. Guo J, Levina E, Michailidis G and Zhu J (2011). Joint estimation of multiple graphical models. Biometrika asq060 MR2804206

[23]. Higgins J, Thompson SG and Spiegelhalter DJ (2009). A reevaluation of random-effects meta-analysis. Journal of the Royal Statistical Society: Series A (Statistics in Society) 172 137–159. MR2655609

[24]. Hsu D, Kakade SM and Zhang T (2012). A tail inequality for quadratic forms of subgaussian random vectors. Electron. Commun. Probab 17 MR2994877

[25]. Johnson WE, Li C and Rabinovic A (2007). Adjusting batch effects in microarray expression data using empirical bayes methods. Biostatistics 8 118–127. [PubMed: 16632515]

[26]. Johnstone IM and Lu AY (2009). On consistency and sparsity for principal components analysis in high dimensions. Journal of the American Statistical Association 104 682–693. MR2751448 [PubMed: 20617121]

[27]. Lam C and Fan J (2009). Sparsistency and rates of convergence in large covariance matrix estimation. Annals of Statistics 37 4254 MR2572459 [PubMed: 21132082]

[28]. 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. MR2933663

[29]. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K and Irizarry RA (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. Nature Reviews Genetics 11 733–739.

[30]. Leek JT and Storey JD (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet 3 1724–1735. [PubMed: 17907809]

[31]. Liu H, Han F and Zhang C. h. (2012). Transelliptical graphical models. In Advances in Neural Information Processing Systems

[32]. 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. MR2563983

[33]. Loh P-L and Wainwright MJ (2013). Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. The Annals of Statistics 41 3022–3049. MR3161456

[34]. Lorentz GG (2005). Approximation of functions, vol. 322 American Mathematical Soc. MR0213785

[35]. Meinshausen N and Buhlmann P (2006). High-dimensional graphs and variable selection with the lasso. The Annals of Statistics 1436–1462. MR2278363

[36]. Negahban S and Wainwright MJ (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. The Annals of Statistics 1069–1097. MR2816348

[37]. Onatski A (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. Journal of Econometrics 168 244–258. MR2923766

[38]. Paul D (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. Statistica Sinica 17 1617 MR2399865

[39]. Power JD, Cohen AL, Nelson SM, Wig GS, Barnes KA, Church JA, Vogel AC, Laumann TO, Miezin FM and Schlaggar BL (2011). Functional network organization of the human brain. Neuron 72 665–678. [PubMed: 22099467]

[40]. Ravikumar P, Wainwright MJ, Raskutti G and Yu B (2011). High-dimensional covariance estimation by minimizing 1-penalized log-determinant divergence. Electronic Journal of Statistics 5 935–980. MR2836766

[41]. Rudelson M and Vershynin R (2013). Hanson-wright inequality and sub-gaussian concentration. Electron. Commun. Probab 18 MR3125258

[42]. Shen X, Pan W and Zhu Y (2012). Likelihood-based selection and sharp parameter estimation. Journal of the American Statistical Association 107 223–232. MR2949354 [PubMed: 22736876]

[43]. Sims AH, Smethurst GJ, Hey Y, Okoniewski MJ, Pepper SD, Howell A, Miller CJ and Clarke RB (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. [PubMed: 18803878]

[44]. Stock JH and Watson MW (2002). Forecasting using principal components from a large number of predictors. Journal of the American statistical association 97 1167–1179. MR1951271

[45]. Verbeke G and Lesaffre E (1996). A linear mixed-effects model with heterogeneity in the random-effects population. Journal of the American Statistical Association 91 217–221.

[46]. Wang W and Fan J (2017). Asymptotics of empirical eigenstructure for high dimensional spiked covariance. Annals of statistics 45 1342–1374. MR3662457 [PubMed: 28835726]

[47]. Yang S, Lu Z, Shen X, Wonka P and Ye J (2015). Fused multiple graphical lasso. SIAM Journal on Optimization 25 916–943. MR3343365

[48]. Yuan M (2010). High dimensional inverse covariance matrix estimation via linear programming. The Journal of Machine Learning Research 11 2261–2286. MR2719856

[49]. Yuan M and Lin Y (2007). Model selection and estimation in the gaussian graphical model. Biometrika 94 19–35. MR2367824

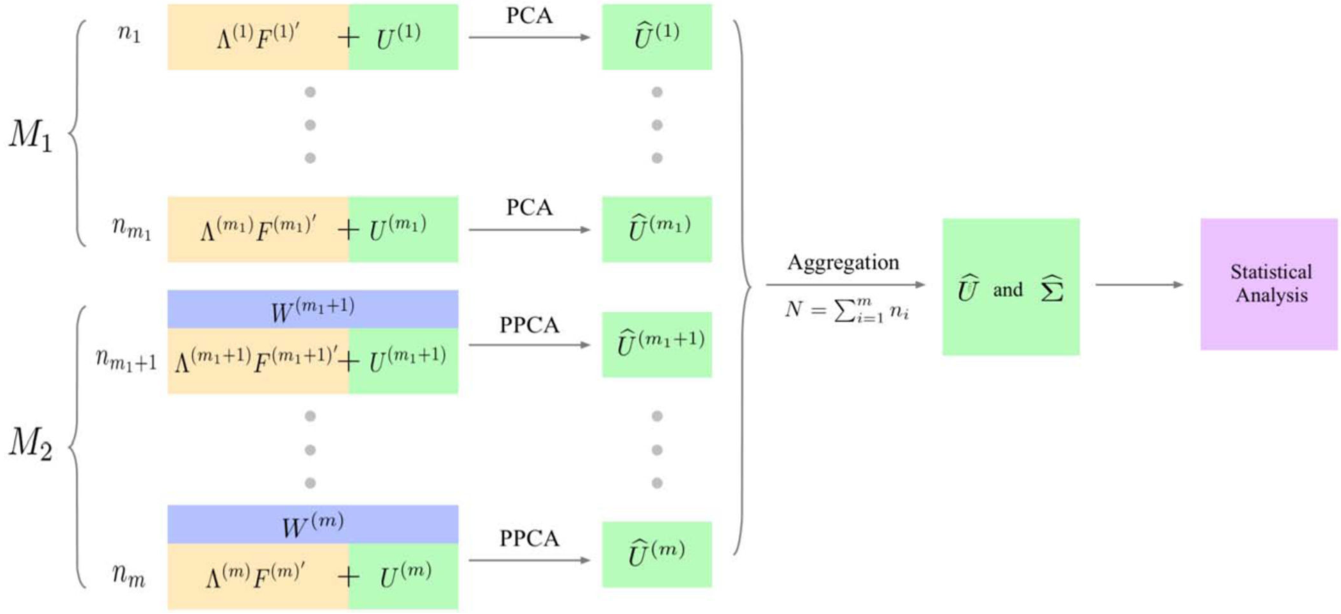


Fig 1.

Schematic illustration of ALPHA: Depending on 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 $\mathbf{\Lambda F'}$ for each batch of data. This decision was made adaptively by a heuristic method. After removing the unwanted variations, the homogeneous data $\left\{\mathbf{U}^{(i)}\right\}_{i=1}^{m}$ are aggregrated for further analysis.

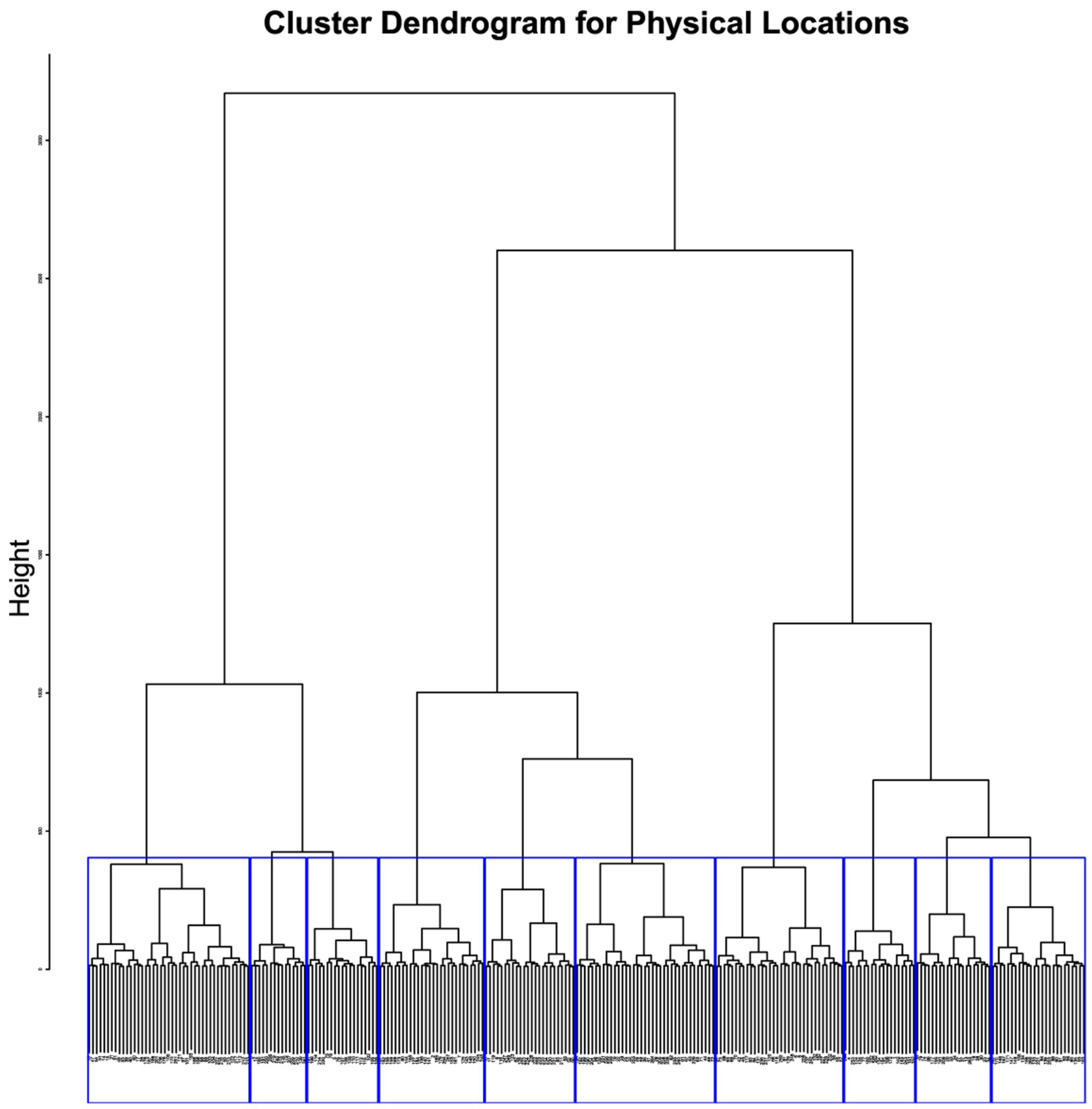


Fig 2.

Cluster Dendrogram for physical locations with J = 10.

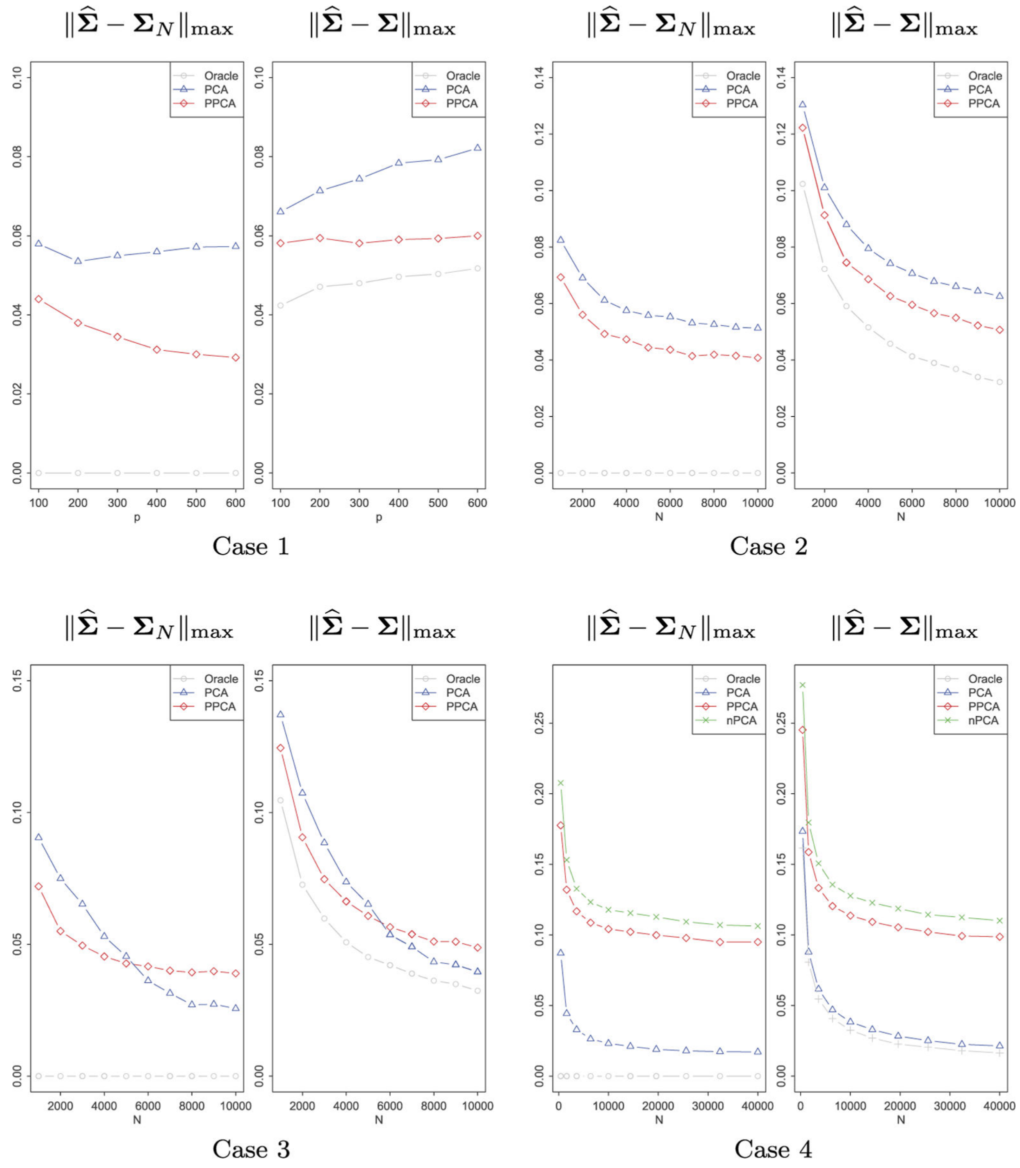


Fig 3.

Estimation of Σ 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.

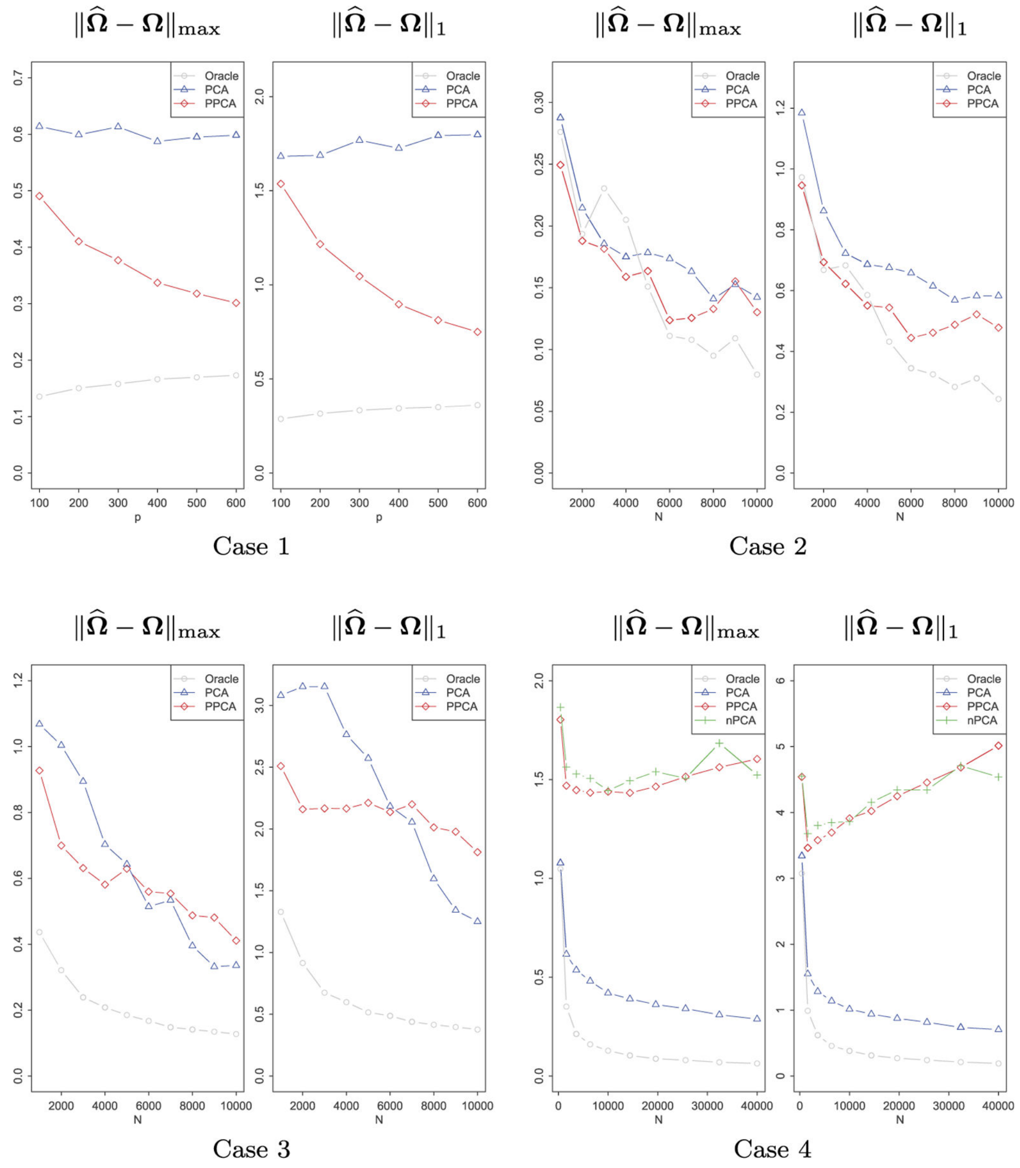
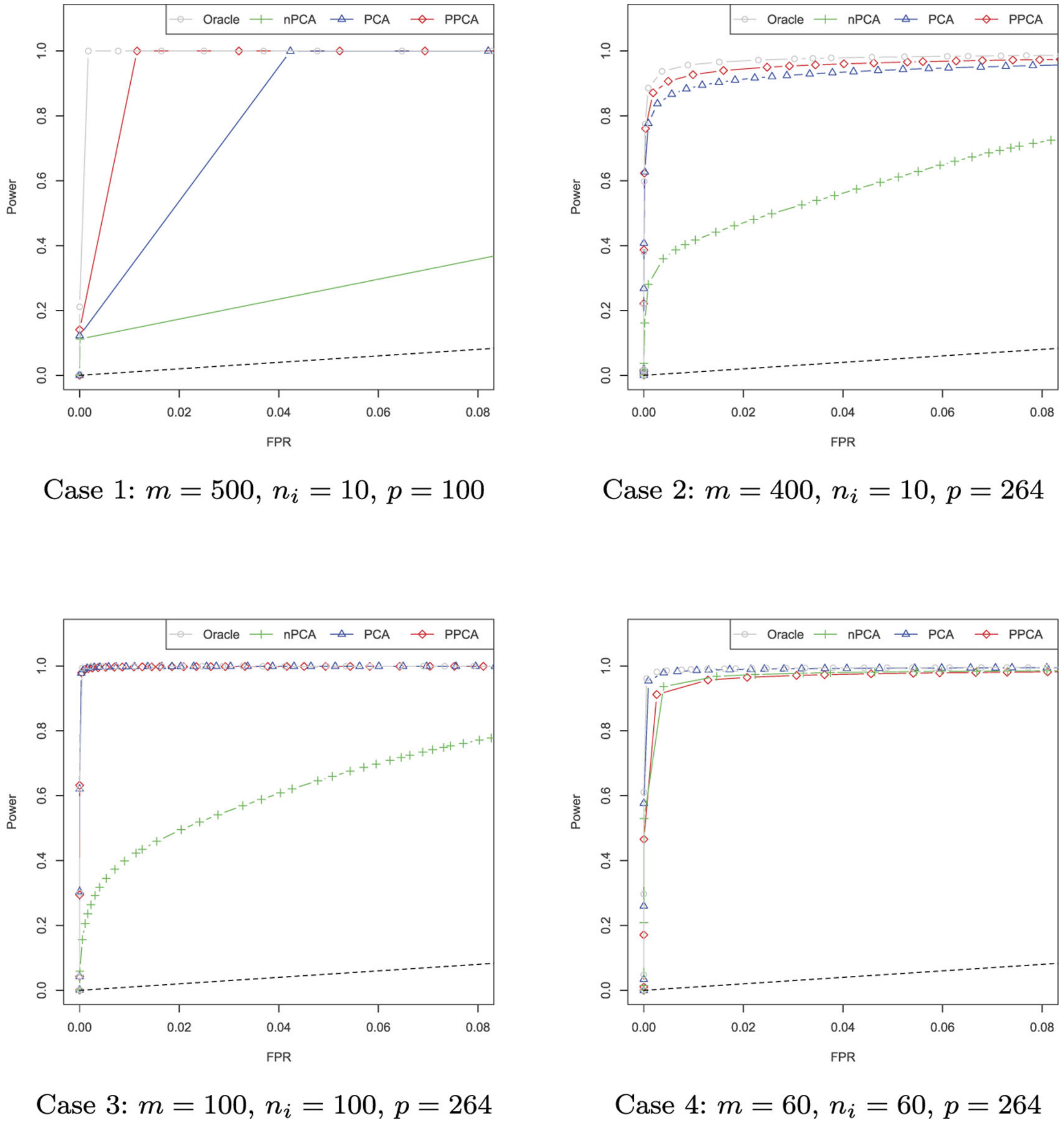


Fig 4.

Estimation of Ω. Presented are the estimation errors in max-norm and L_1-norm for 4 different settings. In Case 4, nPCA refers to no PCA, i.e., we do not adjust heterogeneity.



Case 1: $m = 500$, $n_i = 10$, $p = 100$

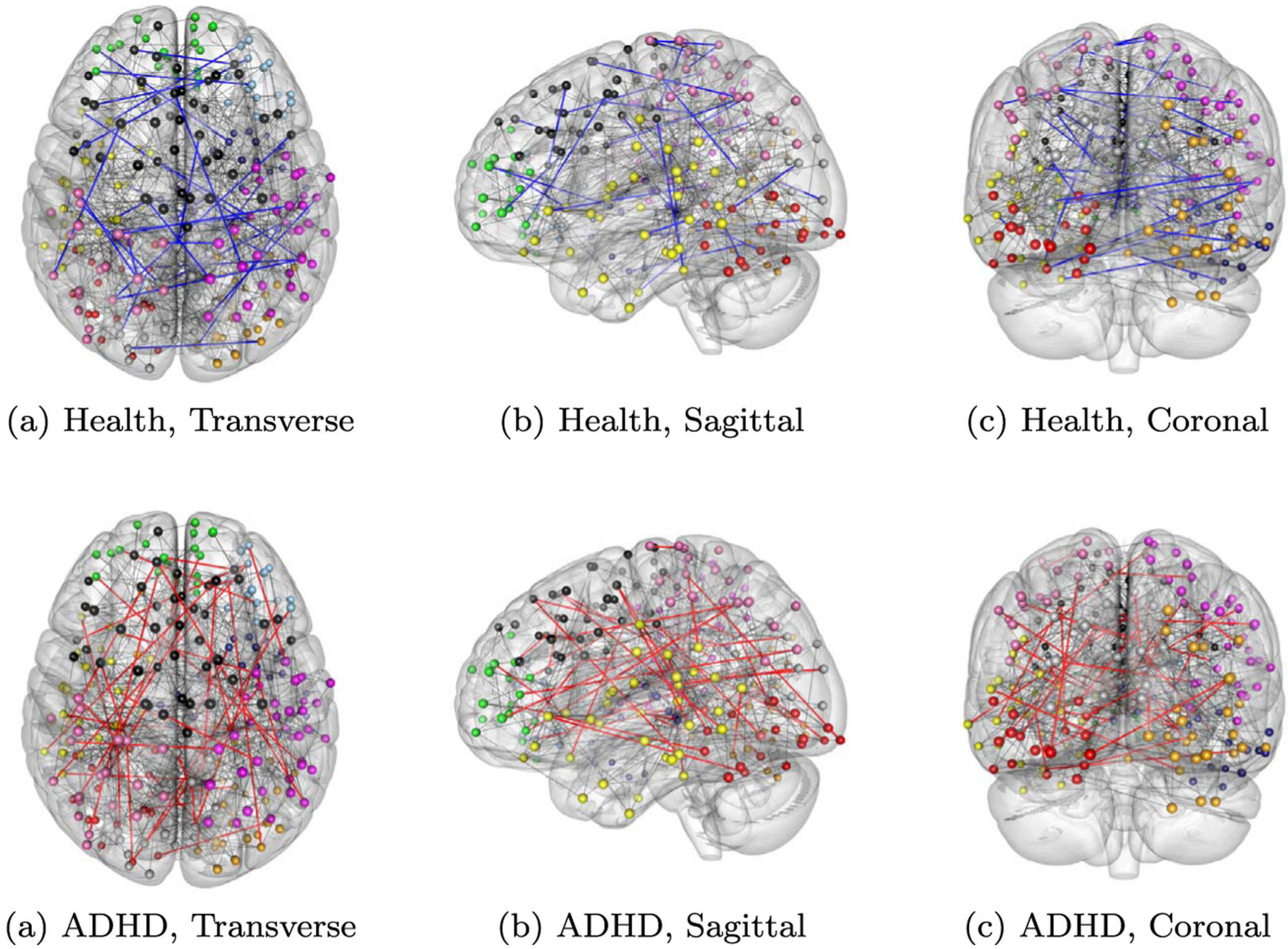
Case 2: $m = 400$, $n_i = 10$, $p = 264$

Case 3: $m = 100$, $n_i = 100$, $p = 264$

Case 4: $m = 60$, $n_i = 60$, $p = 264$

Fig 5.

ROC curves for sparsity recovery of Ω for 4 different settings.



(a) Health, Transverse (b) Health, Sagittal (c) Health, Coronal

(a) ADHD, Transverse (b) ADHD, Sagittal (c) ADHD, Coronal

Fig 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.

Table 1

Distribution of estimated number of factors for healthy and ADHD groups

$\breve{K}^{i}$	**1**	**2**	**3**
Healthy	253	148	64
ADHD	78	40	25

Table 2

The degree of unshared edge vertices for each cluster

	red	orange	blue	green	yellow	navy	pink	black	magenta	gray
Health	3	4	3	2	7	6	10	12	11	6
ADHD	9	6	7	5	12	5	6	15	9	10

Table 3

Gaussian Graphical Model Analysis

	$\|\|\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}\|\|_{\max}$	$\|\|\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}\|\|_1$	$\|\|\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}\|\|_2$
Original	0.564	3.445	1.188
Modified (a)	0.524	3.052	1.066
Modified (b)	0.749	4.914	1.719

Table 4

Accuracy of $\widehat{K}$, $\widetilde{K}$ and $\check{K}$

	n_i = 20			n_i = 100		
	TotErr	**OverEst**	**UnderEst**	**TotErr**	**OverEst**	**UnderEst**
$\widehat{K}$	38.7%	0%	38.7%	0.7%	0%	0.7%
$\widetilde{K}$	29.7%	6.8%	22.9%	4.7%	2.7%	2.0%
$\check{K}$	29.7%	6.8%	22.9%	3.5%	2.3%	1.2%

Table 5

Gaussian Graphical Model Analysis

	$\boldsymbol{n_i}$ **= 20**			$\boldsymbol{n_i}$ **= 100**		
	$\lVert\widehat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}\rVert_{\max}$	$\lVert\widehat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}\rVert_1$	$\lVert\widehat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}\rVert_2$	$\lVert\widehat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}\rVert_{\max}$	$\lVert\widehat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}\rVert_1$	$\lVert\widehat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}\rVert_2$
Oracle	0.687	4.131	1.311	0.335	2.018	0.695
K^o	0.873	2.824	1.351	0.536	2.006	2.017
$\check{K}$	1.156	8.581	2.950	0.564	3.445	1.188
K^+	0.771	3.27	1.49	0.586	2.154	1.074
K^-	1.618	11.384	4.062	1.84	15.133	4.941