Paper Review:
A Pervasive Theory of Heterogeneity Adjustment, with Applications to Graphical Model Inference

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

Presented by Ching-Wei Cheng

February 17, 2016
(BaT Group Meeting)

Outline

Introduction

- Heterogeneity is a fundamental challenge when analyzing aggregated datasets from multiple sources
  - Violation of the ideal “iid” sampling assumption and may produce misleading results
  - Batch/lab effect in genomics
  - In finance, varying market regime and economy status can be viewed as a temporal batch effect
- Modeling and estimating heterogeneity effect is extremely challenging
  - Limited sample sizes are accessible from an individual homogeneous distribution (experiment)
  - High dimensionality (Even much than total sample size)
- Existing batch-effect adjustment methods
  - are more on the practical side and none of them has a systematic theoretical justification
  - are developed in a case-by-case fashion and are only applicable to certain problem domains
- This paper proposes a generic theoretical framework to model, estimate, and adjust heterogeneity across multiple datasets

Introduction

Overview of the proposed heterogeneity adjustment via factor models

Assume the panel data from i^{th} batch/lab, $1 \leq i \leq m$ (fixed), follows an approximate factor model

$$X_{jt}^i = \boldsymbol{\lambda}_j^{i\mathsf{T}} \boldsymbol{f}_t^i + u_{jt}^i, \quad 1 \leq j \leq p,\ 1 \leq t \leq n_i \tag{1}$$

- p-dimensional data with sample size n_i
- Low-rank term $\boldsymbol{\lambda}_j^{i\mathsf{T}} \boldsymbol{f}_t^i$ models the heterogeneity effect
- $\boldsymbol{\lambda}_j^i$ are factor loadings
- $\boldsymbol{f}_t^i$ are the unobserved factors
  - number of factors $K^i = \dim(\boldsymbol{f}_t^i)$, assumed to be fixed
  - independent of u_{jt}^i
- $\boldsymbol{u}_t^i = (u_{1t}^i, \ldots, u_{pt}^i)^\mathsf{T}$ shares the same common distribution with $\mathsf{E}[\boldsymbol{u}_t^i] = \mathbf{0}$ and $\mathsf{Cov}(\boldsymbol{u}_t^i) = \boldsymbol{\Sigma}$ for all $i = 1, \ldots, m$

Introduction

Overview of the proposed heterogeneity adjustment via factor models

Matrix representation:

$$\underset{(p\times n_i)}{\boldsymbol{X}^i} = \underset{(p\times K^i)}{\boldsymbol{\Lambda}^i}\underset{(K^i\times n_i)}{\boldsymbol{F}^{i\mathsf{T}}} + \underset{(p\times n_i)}{\boldsymbol{U}^i} \tag{2}$$

(Rows and columns represent dimension and observation, respectively)

Example 1

If $\boldsymbol{f}_t^i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and $\boldsymbol{u}_t^i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, then the t^{th} observation from i^{th} data

$$\boldsymbol{X}_t^i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}^i\boldsymbol{\Lambda}^{i\mathsf{T}} + \boldsymbol{\Sigma})$$

- Heterogeneity effect is modeled by the low-rank component $\boldsymbol{\Lambda}^i\boldsymbol{\Lambda}^{i\mathsf{T}}$
- Heterogeneity adjusted signal $\widehat{\boldsymbol{U}}^i = \boldsymbol{X}^i - \widehat{\boldsymbol{\Lambda}}^i\widehat{\boldsymbol{F}}^{i\mathsf{T}}$, treated as homogeneous across data sources
- Speaking of estimation...
  - PCA can consistently estimate $\boldsymbol{F}^i$ and $\boldsymbol{\Lambda}^i$ when n_i is large
  - When n_i is small, external covariate information $\boldsymbol{W}_j^i$ (associated with the j^{th} dimension) may help to recover $\boldsymbol{\lambda}_j^i$
    E.g., $\boldsymbol{\lambda}_j^i = (g_1^i(\boldsymbol{W}_j^i), \ldots, g_{K^i}^i(\boldsymbol{W}_j^i))^{\mathsf{T}}$

Introduction

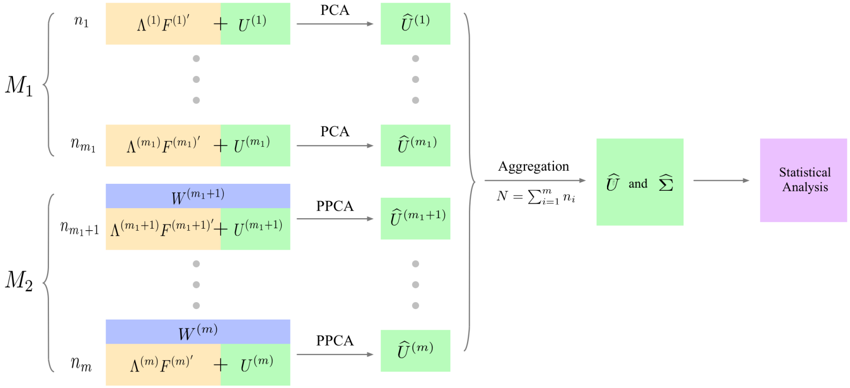
High-dimensional Gaussian graphical model for common covariance structure

After the heterogeneity is removed...

- We can combine $\widehat{\boldsymbol{U}}^i$'s for a Gaussian graphical model inference
- For a typical p-dimensional random vector $\boldsymbol{u} = (u_1, \ldots, u_p) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$
  - The sparsity pattern of the precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ encodes the information of an undirected graph $G = (V, E)$
    - V consists of p vertices corresponding to p dimensions in $\boldsymbol{u}$
    - E describes the dependence relationship between those p variables, i.e., $\Omega_{i,j} \neq 0$ iff X_i and X_j are linked/independent
- Estimate $\boldsymbol{\Omega}$ by using the CLIME method of Cai *et al.* (2011)

Introduction

ALPHA (Adaptive Low-rank Principal Heterogeneity Adjustment)



Goal: Recover $\boldsymbol{U}^i$ from the observation $\boldsymbol{X}^i$ and combine all the estimated $\boldsymbol{U}^i$'s together to enhance the inferential power of $\boldsymbol{\Sigma}$ or $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$

Problem Setup

A semiparametric factor model

$$X_{jt}^i = (\boldsymbol{g}^i(\boldsymbol{W}_j^i) + \boldsymbol{\gamma}_j^i)^\mathsf{T} \boldsymbol{f}_t^i + u_{jt}^i, \quad 1 \le j \le p,\ 1 \le t \le n_i,\ 1 \le i \le m \tag{3}$$

- $\boldsymbol{W}_j^i = (W_{j1}^i, \ldots, W_{jd}^i)$ are the extra covariates for dimension j
- $\boldsymbol{g}^i : \mathbb{R}^d \to \mathbb{R}^{K^i}$
- $\boldsymbol{\gamma}_j^i$ is the loading vector that is invariant of $\boldsymbol{W}_j^i$

Matrix representation:

$$\boldsymbol{X}^i = \boldsymbol{\Lambda}^i \boldsymbol{F}^{i\mathsf{T}} + \boldsymbol{U}^i \;\text{ where } \boldsymbol{\Lambda}^i = \boldsymbol{G}^i(\boldsymbol{W}^i) + \boldsymbol{\Gamma}^i, \quad 1 \le i \le m \tag{4}$$

- $\boldsymbol{G}^i(\boldsymbol{W}^i)$ and $\boldsymbol{\Gamma}^i$ are $(p \times K^i)$ nonparamatric and parametric factor loadings, where $g_k^i(\boldsymbol{W}_j^i)$ and γ_{jk} are the $(j,k)^{\text{th}}$ elements
- $\boldsymbol{U}^i$ is the homogeneous signal matrix of dimension $p \times n_i$ with u_{jt}^i the $(j,k)^{\text{th}}$ element

Problem Setup

Modeling assumptions

Assumption 2.1 (Data Generating Process)

(i) $n_i^{-1} \boldsymbol{F}^{i\top} \boldsymbol{F}^i = \boldsymbol{I}$.

(ii) $\{\boldsymbol{u}_t^i\}_{t \leq n_i, i \leq m}$ are independent within and between subgroups. $\boldsymbol{u}_t^i$'s are sub-Gaussian with $\mathsf{E}[\boldsymbol{u}_t^i] = \boldsymbol{0}$ and $\mathsf{Cov}(\boldsymbol{u}_t^i) = \boldsymbol{\Sigma}$ across all subgroups and are independent of $\{\boldsymbol{W}_j^i, \boldsymbol{f}_t^i\}$. $\{\boldsymbol{f}_t^i\}_{t \leq n_i}$ is a stationary process, but with arbitrary temporal dependency.

(iii) There exists a constant $C_0 > 0$ s.t. $\|\boldsymbol{\Sigma}\|_2 \leq C_0$.

(iv) The tail of the factors is sub-Gaussian, i.e., $\exists\, C_1 > 0$ s.t. for $j \leq K^i, t \leq n_i$, $P(|f_{jt}^i| > t) \leq \exp(-C_1 t^2)$.

- Typical assumptions for factor models in literature
- Factors $\boldsymbol{F}^i$ are identifiable up to an orthonormal trasformation $\boldsymbol{H}^i$
  - We need to choose $\boldsymbol{H}^i$ carefully

Regime 1: $\boldsymbol{G}^i(\boldsymbol{W}^i) = \mathbf{0}$ a.s.

Modeling assumptions

- $\boldsymbol{X}^i = \boldsymbol{\Lambda}^i \boldsymbol{F}^{i\mathsf{T}} + \boldsymbol{U}^i$, reduced to traditional factor models
- Find $\widehat{\boldsymbol{F}}^i$ first using PCA
- $\widehat{\boldsymbol{\Lambda}}^i = n_i^{-1} \boldsymbol{X}^i \boldsymbol{F}^i$ and $\widehat{\boldsymbol{U}}^i = \boldsymbol{X}^i - \widehat{\boldsymbol{\Lambda}}^i \widehat{\boldsymbol{F}}^{i\mathsf{T}}$

Assumption 2.2 (General Loadings)

(i) (Pervasiveness) $\exists\, c_{\min}, c_{\max} > 0$ s.t.

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

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

- The notion of random loadings λ_{jk}^i is natural for providing a unified theoretical treatment regime 1 and regime 2

Regime 2: $\boldsymbol{G}^i(\boldsymbol{W}^i) \neq \mathbf{0}$ a.s.

Modeling assumptions

- $\boldsymbol{X}^i = (\boldsymbol{G}^i(\boldsymbol{W}^i) + \boldsymbol{\Gamma}^i)\boldsymbol{F}^{i\mathsf{T}} + \boldsymbol{U}^i$, need to leverage effectsof external covariates and provide better estimates for the low-rank structure

Assumption 2.3 (Covariate-related Loadings)

(i) (Pervasiveness) $\exists\, c_{\min}, c_{\max} > 0$ s.t.

$$\begin{aligned} c_{\min} &< \lambda_{\min}(p^{-1}\boldsymbol{G}^i(\boldsymbol{W}^i)^{\mathsf{T}}\boldsymbol{G}^i(\boldsymbol{W}^i)) \\ &< \lambda_{\max}(p^{-1}\boldsymbol{G}^i(\boldsymbol{W}^i)^{\mathsf{T}}\boldsymbol{G}^i(\boldsymbol{W}^i)) < c_{\max},\ a.s.\ \forall\, i. \end{aligned}$$

(ii) $\max_{k \le K^i, j \le p} \mathsf{E}[g_k(\boldsymbol{W}_j^i)^2] < \infty$.

- Semiparametric factor models can be better estimated by Projected-PCA (Fan *et al.*, 2016) if Assumption 2.3 holds
  - Estimate $\boldsymbol{f}_t^i$ first by projecting $\boldsymbol{X}^i$ onto the covariate space of $\boldsymbol{W}^i$ to reduce the magnitude of $\boldsymbol{u}_t^i$
  - Apply PCA on the projected data

Regime 2: $\boldsymbol{G}^i(\boldsymbol{W}^i) \neq \mathbf{0}$ a.s.

Modeling assumptions

Assumption 2.4 (Covariate-free Loadings)

(i) $\mathsf{E}[\gamma^i_{jk}] = 0$, $\max_{k \leq K^i, j \leq p} |\gamma^i_{jk}| = O_P(\sqrt{\log p})$.

(ii) Write $\boldsymbol{\gamma}^i_j = (\gamma^i_{j1}, \ldots, \gamma^i_{jK^i})^\mathsf{T}$. Assume $\{\boldsymbol{\gamma}^i_j\}_{j \leq p}$ are independent of $\{\boldsymbol{W}^i_j\}_{j \leq p}$.

(iii) Define $\nu^i_p = \max_{k \leq K^i} p^{-1} \sum_{j \leq p} \mathsf{Var}(\gamma^i_{jk}) < \infty$. Assume

$$\max_{k \leq K^i, j \leq p} \sum_{j\prime \leq p} \left| \mathsf{E}\left[\gamma^i_{j\prime k} \gamma^i_{jk}\right] \right| = O(\nu_p).$$

The ALPHA Framework

- 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
  - Use $\widehat{\cdot}$ and $\widetilde{\cdot}$ to represent estimates by PCA and Projected-PCA, resp.
  - Temporarily forget the subgroup index i in Theorem 3.1–3.3

Overview of the theorems:

- Theorem 3.1 provides generic asymptotic representations for $\widehat{\boldsymbol{U}}$ and $\widehat{\boldsymbol{U}}\widehat{\boldsymbol{U}}^{\mathsf{T}}$, where the detailed rates provided in Theorem 3.2 for regime 1 and Theorem 3.3 for regime 2
- From $\widehat{\boldsymbol{U}}\widehat{\boldsymbol{U}}^{\mathsf{T}}$, we can have convergence rate for $\widehat{\boldsymbol{\Sigma}}$
- Theorem 4.1 gives theoretical guarantee for the CLIME solver $\widehat{\boldsymbol{\Omega}}$, which takes $\widehat{\boldsymbol{\Sigma}}$ as input

The ALPHA Framework

Theorem 3.1

For any $K \times K$ *matrix* $\boldsymbol{H}$ *s.t.* $\|\boldsymbol{H}\| = O_P(1)$, *if* $\log p = O(n)$,

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

where

$$\|\boldsymbol{\Pi}\|_{\max} = O_P\Big[\frac{\sqrt{\log n}}{n}\Big(\|\boldsymbol{F}^\top(\widehat{\boldsymbol{F}} - \boldsymbol{F}\boldsymbol{H})\|_{\max}\|\boldsymbol{\Lambda}\|_{\max} + \|\boldsymbol{U}(\widehat{\boldsymbol{F}} - \boldsymbol{F}\boldsymbol{H})\|_{\max}\Big) + \|\widehat{\boldsymbol{F}} - \boldsymbol{F}\boldsymbol{H}\|_{\max}\|\boldsymbol{\Lambda}\|_{\max} + \sqrt{\log n}\|\boldsymbol{H}\boldsymbol{H}^\top - \boldsymbol{I}\|_{\max}\|\boldsymbol{\Lambda}\|_{\max}\Big];$$

and furthermore

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

where

$$\|\boldsymbol{\Delta}\|_{\max} = O_P\Big[\|\boldsymbol{U}(\widehat{\boldsymbol{F}} - \boldsymbol{F}\boldsymbol{H})\|_{\max}\|\boldsymbol{\Lambda}\|_{\max} + \|\boldsymbol{U}(\widehat{\boldsymbol{F}} - \boldsymbol{F}\boldsymbol{H})\|_{\max}^2 + \|\boldsymbol{F}^\top(\widehat{\boldsymbol{F}} - \boldsymbol{F}\boldsymbol{H})\|_{\max}\|\boldsymbol{\Lambda}\|_{\max}^2 + n\|\boldsymbol{H}\boldsymbol{H}^\top - \boldsymbol{I}\|_{\max}\|\boldsymbol{\Lambda}\|_{\max}^2\Big].$$

The ALPHA Framework by PCA for Regime 1

- Columns of $\widehat{\boldsymbol{F}}/\sqrt{n}$ are the top K eigenvectors of $\boldsymbol{X}^\top \boldsymbol{X}$
- Denote by $\boldsymbol{K}$ the $K \times K$ diagonal matrix of top K eigenvalues of $(np)^{-1}\boldsymbol{X}^\top \boldsymbol{X}$
- Define

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

- It has been shown that $\|\boldsymbol{H}\|, \|\boldsymbol{H}^{-1}\| = O_P(1)$

The ALPHA Framework by PCA for Regime 1

Theorem 3.2 (When $\boldsymbol{G}(\boldsymbol{W}) = \boldsymbol{0}$ a.s.)

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

(i) $\|\widehat{\boldsymbol{F}} - \boldsymbol{F}\boldsymbol{H}\|_F = O_P(\sqrt{n/p} + 1/\sqrt{n})$ *and*
$\|\widehat{\boldsymbol{F}} - \boldsymbol{F}\boldsymbol{H}\|_{\max} = O_P(\sqrt{\log n/p} + \sqrt{\log n}/n)$;

(ii) $\|\boldsymbol{F}^{\mathsf{T}}(\widehat{\boldsymbol{F}} - \boldsymbol{F}\boldsymbol{H})\|_{\max} = O_P(1 + \sqrt{n}/p)$;

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

(iv) $\|\boldsymbol{H}\boldsymbol{H}^{\mathsf{T}} - \boldsymbol{I}\|_{\max} = O_P(1/n + 1/p)$.

As a result,

$$\|\boldsymbol{\Pi}\|_{\max} = O_P\left(\sqrt{\log n \log p}(1/\sqrt{p} + 1/n) + \sqrt{n}\|\boldsymbol{\Sigma}\|_1/p\right)$$
$$\|\boldsymbol{\Delta}\|_{\max} = O_P\left((1 + n/p)\log p + \sqrt{n}\|\boldsymbol{\Sigma}\|_1/p + \sqrt{n}^2\|\boldsymbol{\Sigma}\|_1^2/p^2\right)$$

The ALPHA Framework by Projected-PCA for Regime 2

- Factor loadings $\boldsymbol{\Lambda} = \boldsymbol{G}(\boldsymbol{W}) + \boldsymbol{\Gamma}$
  - A function of covariates $\boldsymbol{W}$, which is independent of $\boldsymbol{\Gamma}$ and $\boldsymbol{U}$
  - Sieve approximation $\boldsymbol{G}(\boldsymbol{W}) \approx \Phi(\boldsymbol{W})\boldsymbol{B}$
    - $\Phi(\boldsymbol{W})$ is a $p \times (Jd)$ matrix of basis functions
    - $\boldsymbol{B}$ is a $(Jd) \times K$ matrix of sieve coefficients
    - J is the sieve dimension
  - The idea of Projected-PCA: $\boldsymbol{P}\boldsymbol{X} \approx \boldsymbol{P}\Phi(\boldsymbol{W})\boldsymbol{B}\boldsymbol{F}^\mathsf{T} \approx \boldsymbol{G}(\boldsymbol{W})\boldsymbol{F}^\mathsf{T}$
- Define the projection matrix

$$\boldsymbol{P} = \Phi(\boldsymbol{W}) \left[\Phi(\boldsymbol{W})\Phi(\boldsymbol{W})^\mathsf{T}\right]^{-1} \Phi(\boldsymbol{W})^\mathsf{T}$$

- Columns of $\widetilde{\boldsymbol{F}}/\sqrt{n}$ are the top K eigenvectors of $\boldsymbol{X}^\mathsf{T}\boldsymbol{P}\boldsymbol{X}$
- Define by $\boldsymbol{K}$ the $K \times K$ diagonal matrix of top K eigenvalues of $(np)^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{P}\boldsymbol{X}$
- Define

$$\boldsymbol{H} = \frac{1}{np}\boldsymbol{B}^\mathsf{T}\Phi(\boldsymbol{W})^\mathsf{T}\Phi(\boldsymbol{W})\boldsymbol{B}\boldsymbol{F}^\mathsf{T}\widetilde{\boldsymbol{F}}\boldsymbol{K}^{-1}$$

- Similarly, $\|\boldsymbol{H}\|, \|\boldsymbol{H}^{-1}\| = O_P(1)$

The ALPHA Framework by Projected-PCA for Regime 2

Theorem 3.3 (When $\boldsymbol{G}(\boldsymbol{W}) \neq \mathbf{0}$ a.s.)

Choose $J = (p\min(n, p, \nu_p^{-1}))^{1/\kappa}$ *and assume* $J^2\phi_{\max}^2 \log(nJ) = O(p)$, *where* $\phi_{\max} = \max_{\nu \le J} \sup_{x \in \mathcal{X}} \phi_\nu(x)$. *Under Assumptions 2.1, 2.3, 2.4, 3.1 and 3.2 (for basis functions and sieve approximation), as* $p, J \to \infty$, n *can be either divergent or bounded, we have* $\|\boldsymbol{\Lambda}\|_{\max} = O_P(J\phi_{\max} + \sqrt{\log p})$ *and*

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

(ii) $\|\boldsymbol{F}^\mathsf{T}(\widetilde{\boldsymbol{F}} - \boldsymbol{F}\boldsymbol{H})\|_{\max} = O_P(\sqrt{n/p} + n/p + n\sqrt{\nu_p/p})$;

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

(iv) $\|\boldsymbol{H}\boldsymbol{H}^\mathsf{T} - \boldsymbol{I}\|_{\max} = O_P(1/p + 1/\sqrt{pn} + \sqrt{\nu_p/p})$.

As a result,

$$\|\boldsymbol{\Pi}\|_{\max} = O_P\left(\sqrt{\log n \log p/p} + \sqrt{\log n}\|\boldsymbol{\Sigma}\|_1/p\right)$$

$$\|\boldsymbol{\Delta}\|_{\max} = O_P\Big(n\sqrt{\nu_p/p}(J^2\phi_{\max}^2 + \log p) + nJ\phi_{\max}\|\boldsymbol{\Sigma}\|_1/p(J\phi_{\max} + \sqrt{\log p}) + n^2J^2\phi_{\max}^2\|\boldsymbol{\Sigma}\|_1^2/p^2\Big)$$

if there exists C s.t. $\nu_p > C/n$

The ALPHA Framework

Specification test

To test $H_0^i : \boldsymbol{G}^i(\boldsymbol{W}^i) = \mathbf{0}$ a.s., Fan *et al.* (2016) proposed a testing statistic

$$S^i = \frac{1}{p}\operatorname{tr}\left(\boldsymbol{\Xi}^i \widehat{\boldsymbol{\Lambda}}^{i\mathsf{T}} \boldsymbol{P}^i \widehat{\boldsymbol{\Lambda}}^i\right) \quad \text{where} \quad \boldsymbol{\Xi}^i = \left(\frac{1}{p}\widehat{\boldsymbol{\Lambda}}^{i\mathsf{T}}\widehat{\boldsymbol{\Lambda}}^i\right)^{-1}$$

Theorem 3.4 (Specification test)

Under all assumptions above, if additionally $\{\boldsymbol{W}_j^i, \boldsymbol{\gamma}_j^i\}_{j \le p}$ are iid, as $p, n^i, J \to \infty$, we have under $H_0^i : \boldsymbol{G}^i(\boldsymbol{W}^i) = \mathbf{0}$ a.s.,

$$\frac{pS^i - JdK^i}{\sqrt{2JdK^i}} \xrightarrow{D} \mathcal{N}(0,1).$$

- Under H_0, $\boldsymbol{\Lambda}^i$ has nothing to do with $\boldsymbol{W}^i$ and so S^i should be close to 0 after projection
- If H_0^i is rejected, we identify $\boldsymbol{X}^i$ as regime 2 and apply Projected-PCA

The ALPHA Framework

Estimating number of factors

- We have assumed observed K^i for each subgroup, but practically it needs to be estimated
- For regime 1...
  - Define $\widehat{K}^i = \arg\max_{k \leq K_{\max}} \lambda_k(\boldsymbol{X}^{i\mathsf{T}}\boldsymbol{X}^i)/\lambda_{k+1}(\boldsymbol{X}^{i\mathsf{T}}\boldsymbol{X}^i)$
  - $P(\widehat{K}^i = K^i) \to 1$
- For regime 2...
  - Define $\widetilde{K}^i = \arg\max_{k \leq K_{\max}} \lambda_k(\boldsymbol{X}^{i\mathsf{T}}\boldsymbol{P}^i\boldsymbol{X}^i)/\lambda_{k+1}(\boldsymbol{X}^{i\mathsf{T}}\boldsymbol{P}^i\boldsymbol{X}^i)$
  - $P(\widetilde{K}^i = K^i) \to 1$
- By slightly altering the original assumptions of Ahn and Horenstein (2013) and Fan *et al.* (2016), we have $P(\widehat{K}^i = K^i, \forall\, i \leq m) \to 1$ and $P(\widetilde{K}^i = K^i, \forall\, i \leq m) \to 1$
- Given $\widehat{K}^i$ for regime 1 and $\widetilde{K}^i$ for regime 2, we can treat the problem as if the number of factors for all subgroups are already known to us

Conditional Graphical Model

- Assume $\boldsymbol{u}_t^i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ with $\|\boldsymbol{\Sigma}\|_1$ bounded
- Let

$$\widehat{\boldsymbol{V}}^i = \begin{cases} \widehat{\boldsymbol{U}}^i & \text{for regime 1} \\ \widetilde{\boldsymbol{U}}^i & \text{for regime 2} \end{cases}$$

  and estimate $\boldsymbol{\Sigma}$ by

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{N - K^{tot}} \sum_{i=1}^{m} \widehat{\boldsymbol{V}}^i \widehat{\boldsymbol{V}}^{i\mathsf{T}} \tag{5}$$

  where $K^{tot} = \sum_{i=1}^{m} K^i$
- Define
  - $\mathcal{M}_1 = \{i \leq m : \boldsymbol{G}^i(\boldsymbol{W}^i) = \mathbf{0} \; a.s.\}$
  - $\mathcal{M}_2 = \{i \leq m : \boldsymbol{G}^i(\boldsymbol{W}^i) \neq \mathbf{0} \; a.s.\}$

Conditional Graphical Model

Covariance estimation

- The oracle $\boldsymbol{\Sigma}_N = N^{-1}\sum_{i=1}^m \boldsymbol{U}^i\boldsymbol{U}^{i\mathsf{T}}$ attains the rate $\|\boldsymbol{\Sigma}_N - \boldsymbol{\Sigma}\|_{\max} = O_P(\sqrt{\log p/N})$
- By standard concentration bound,

$$\left\|\sum_{i=1}^m \left(\frac{1}{n_i}\boldsymbol{U}^i\boldsymbol{F}^i\boldsymbol{F}^{i\mathsf{T}}\boldsymbol{U}^{i\mathsf{T}} - K^i\boldsymbol{\Sigma}\right)\right\|_{\max} = O_P\left(\sqrt{K^{tot}\log p}\right)$$

- Therefore

$$\begin{aligned}
&\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_N\|_{\max} \\
&\quad\le O_P\left(\frac{|\mathcal{M}_1|\log p}{N} + \frac{N_2\log p}{N}\sqrt{\frac{\nu_p}{p}} + \frac{\sqrt{K^{tot}\log p}}{N} + \frac{K^{tot}}{N}\sqrt{\frac{\log p}{N}}\right) \\
&\quad =: O_P(a_{m,N,p})
\end{aligned} \tag{6}$$

Conditional Graphical Model

Covariance estimation

- If all $i \in \mathcal{M}_1$ and $K^i \leq K_{\max} = O(1)$...
  - $a_{m,N,p} = m \log p / N$
  - Dominated by oracle rate $\sqrt{\log p / N} \iff m = O(\sqrt{N / \log p})$
  - PCA works optimally when m does not grow too quickly
- If all $i \in \mathcal{M}_2$ and $K^i \leq K_{\max} = O(1)$...
  - $a_{m,N,p} = \sqrt{\nu_p / p} \log p + \sqrt{m \log p} / N$
  - Smaller than $\sqrt{\log p / N}$ if $p / \log p > CN$ for some $C > 0$
  - Good convergence can still be achieved even when $m \asymp N$ as long as p is large enough (Blessing of dimensionality)

Conditional Graphical Model

Precision estimation

- For a given $\widehat{\boldsymbol{\Sigma}}$, CLIME solves the optimization problem

$$\widehat{\boldsymbol{\Omega}} = \arg\min_{\boldsymbol{\Omega}} \|\boldsymbol{\Omega}\|_{1,1} \quad \text{subject to } \|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Omega} - \boldsymbol{I}\|_{\max} \le \lambda$$

  where $\|\boldsymbol{\Omega}\|_{1,1} = \sum_{i,j \le p} |\sigma_{i,j}|$ and λ is a tuning parameter
- Given C_0 and s, consider the sparse precision matrix class

$$\mathcal{F}(s, C_0) = \left\{ \boldsymbol{\Omega} : \boldsymbol{\Omega} \succ \mathbf{0}, \|\boldsymbol{\Omega}\|_1 \le C_0, \max_{1 \le i \le p} \sum_{j=1}^{p} \mathbb{1}(\Omega_{i,j} \ne 0) \le s \right\}$$

Theorem 4.1

Suppose $\mathbf{\Omega} \in \mathcal{F}(s, C_0)$ *and* $\widehat{\mathbf{\Sigma}}$ *given by* (5) *attains the rate* $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}_N\|_{\max} = O_P(a_{m,N,p})$ *in* (6). *Letting* $\tau_{m,N,p} = \sqrt{\log p/N} + a_{m,N,p}$ *and* $\lambda \asymp \tau_{m,N,p}$, *we have*

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

Furthermore,

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

Numerical Studies

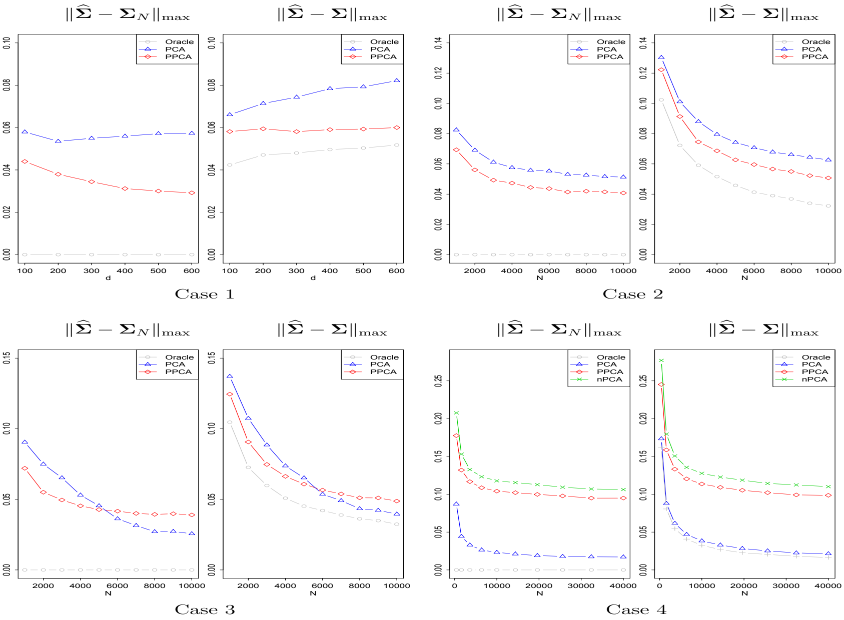
- Brain image network data ADHD-200
  - Contains rs-fMRI images of 688 subjects (491 healthy, 197 diseased)
  - 16 (13 healthy, 3 diseased) were dropped due to missing values
  - $m = 672$ subjects in this analysis
- Divided the whole brain into $p = 264$ seed regions
- Each brain was scanned multiple times ($76 \leq n_i \leq 261$)
- Physical locations of the brain as covariates W ($d = 1$ and discrete)
  - The level of batch effect is non-uniform over different locations of the brain when scanned in fMRI machine
  - Spatial adjacency does not necessarily imply brain functional connectivity (graph structure)
  - Split 264 regions into $J = 10$ clusters by hierarchy clustering
  - Sieve basis are $\mathbb{1}(w - 0.5 \leq W < w + 0.5), w = 1, \ldots, 10$
- $K_{\max} = 5$

Numerical Studies

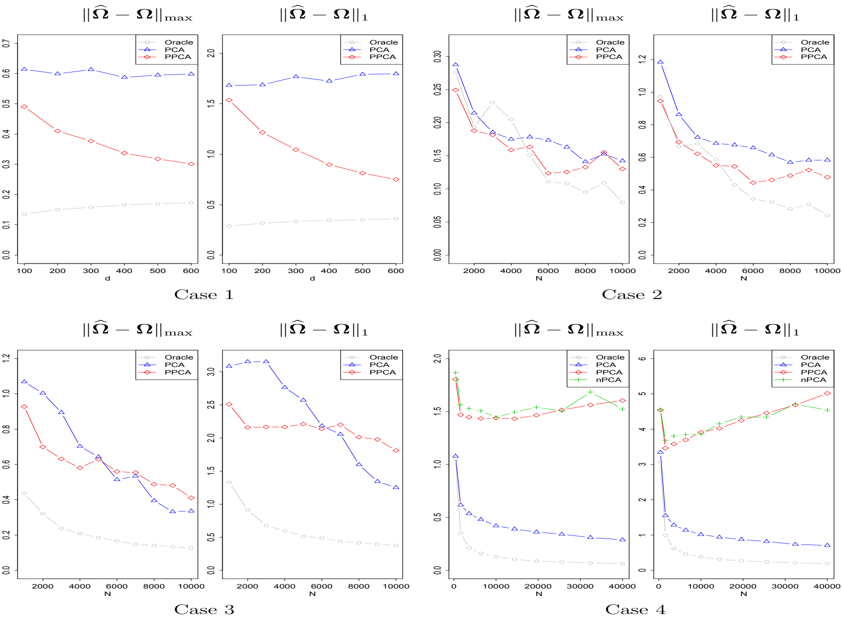
Synthetic datasets

- Simulation settings:
  1. $m = 500, n_i = 10, p = 100, 200, \ldots, 600$ and $\boldsymbol{G}(\boldsymbol{W}) \neq \mathbf{0}$
  2. $m = 100, 200, \ldots, 1000, n_i = 10, p = 264$ and $\boldsymbol{G}(\boldsymbol{W}) \neq \mathbf{0}$
  3. $m = 100, n_i = 10, 20, \ldots, 100, p = 264$ and $\boldsymbol{G}(\boldsymbol{W}) \neq \mathbf{0}$
  4. $m = 20, 40, \ldots, 200, n_i = 20, 40, \ldots, 200, p = 264$ and $\boldsymbol{G}(\boldsymbol{W}) = \mathbf{0}$
- Model calibration and data generation:
  1. For $j \leq p$, generate iid covariates from multinomial distribution $P(W_j = s) = w_s, s = 1, \ldots, 10$, where $\{w_s\}$ are calibrated with the hierarchy clustering results of the real data
  2. Calibrate the parameters (e.g., $\boldsymbol{\Sigma}, \boldsymbol{f}_t^i, \boldsymbol{\Lambda}^i$, etc.) from the first 15 subjects in the healthy group

Numerical Studies



Numerical Studies



Numerical Studies

- Estimation of $\boldsymbol{\Sigma}$
  1. Blassing of dimensionality
  2. Blessing of increasing sample sizes
  3. PCA ourperforms Projected-PCA when n_i is large enough (when $p/\log p = O(N)$)
     - For fixed m, $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max} = O_P(\sqrt{\log p/N})$ and $\|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max} = O_P(\sqrt{\log p/p})$
  4. PCA is much better since covariates have no explanation power at all
     - "nPCA" corresponds to no heterogeneity adjustment
- Estimation of $\boldsymbol{\Omega}$: similar results

Discussion

- A generic methodology ALPHA for heterogeneity adjustment
  - Consistently estimate and remove data heterogeneity
  - Flexible to include external information
- Future work
  - Pervasive conditions may be relaxed to allow for weaker signal batch effect
  - Finding practical interpretations of the estimated factors

Selected References

Ahn, S. C. and Horenstein, A. R. (2013) Eigenvalue ratio test for the number of factors. *Econometrica*, **81**, 1203–1227. URL http://dx.doi.org/10.3982/ECTA8968.

Cai, T., Liu, W. and Luo, X. (2011) A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, **106**, 594–607. URL http://dx.doi.org/10.1198/jasa.2011.tm10155.

Fan, J., Liao, Y. and Wang, W. (2016) Projected principal component analysis in factor models. *The Annals of Statistics*, **44**, 219–254. URL http://dx.doi.org/10.1214/15-AOS1364.