Reading: Heterogeneity Adjustment with Application to Graphical Model Inference

Meilei Jiang

Department of Statistics and Operations Research

University of North Carolina at Chapel Hill

February 24, 2016

Overview

Heterogeneity Effects

Heterogeneity is an unwanted variation when analyzing aggregated datasets from multiple sources.

Challenge of modeling and estimating heterogeneity effect:

1. We can only access a limited number of samples from an individual group, given the high cost of biological experiment, technological constraint or fast economy regime switching.
2. The dimensionality can be much larger than the total aggregated number of samples.

Model Settings of Data Heterogeneity

- Assume data come from m different sources
  - the ith data source contributes n_i samples.
  - Each sample having p measurements.
- Assume the batch-specific latent factors f_t^i influence the observed data X_{jt}^i in batch i (j indexes variables; t indexes samples).
  - $X_{jt}^i = \lambda_j^{i\prime} f_t^i + u_{jt}^i, 1 \le j \le p, 1 \le t \le n_i, 1 \le i \le m$
  - where λ_j^i is unknown factor loading for variable j and u_{jt}^i is true uncorrupted signals.
- Assume that f_t^i is independent of u_{jt}^i.
- Assume $f_t^i \sim N(\mathbf{0}, \mathbf{I})$ and $\mathbf{u}_t^i = (u_{1t}, \cdots, u_{pt})'$ shares the common normal distribution $N(\mathbf{0}, \mathbf{\Sigma}_{p \times p})$.

Model Settings of Data Heterogeneity

The matrix form model can be written as: $\mathbf{X}^i = \mathbf{\Lambda}^i \mathbf{F}^{i'} + \mathbf{U}^i$.

- $\mathbf{X}^i$ is a $p \times n_i$ data matrix in the ith batch, $\mathbf{\Lambda}^i$ is a $p \times K^i$ factor loading matrix with λ_j^i in the jth row, $\mathbf{F}^i$ is an $n_i \times K^i$ factor matrix and $\mathbf{U}^i$ is a $p \times n_i$ signal matrix.
- $X_t^i \sim N(\mathbf{0}, \mathbf{\Lambda}^i \mathbf{\Lambda}^{i'} + \mathbf{\Sigma})$.
- The heterogeneity effect is modeled as a low rank component $\mathbf{\Lambda}^i \mathbf{\Lambda}^{i'}$ of the population covariance matrix of $\mathbf{X}_t^i$.

Semiparametric Factor Model

- For subgroup i, we have d external covariates $\mathbf{W}_j^i = (W_{j1}^i, \cdots, W_{jd}^i)'$ for variable j.
- Assume that these covariates have some explanatory power on the loading parameters λ_j^i: $\lambda_j^i = g^i(\mathbf{W}_j^i) + \gamma_j^i$.
- $X_{jt}^i = \lambda_j^{i\prime} f_t^i + u_{jt}^i = (g^i(\mathbf{W}_j^i) + \gamma_j^i)' t^i + u_{jt}^i$.
    - If $\mathbf{W}_j^i$ is not informative, then $g^i(.) = 0$.
- $\mathbf{X}^i = \mathbf{\Lambda}^i \mathbf{F}^{i\prime} + \mathbf{U}^i$, where $\mathbf{\Lambda}^i = \mathbf{G}^i(\mathbf{W}^i) + \mathbf{\Gamma}^i, 1 \leq i \leq m$.
    - $\mathbf{G}^i(\mathbf{W}^i)$ and $\mathbf{\Gamma}^i$ are $p \times K^i$ component matrices of $\mathbf{\Lambda}^i$.

Modeling Assumptions And General Methodology

Data Generating Process:

i $n_i \mathbf{F}^{i'} \mathbf{F}^i = \mathbf{I}$.

ii $\{\mathbf{u}_t^i\}$ are independent within and between subgroups. $\{f_t^i\}_{t \leq n_i}$ is a stationary process, but with arbitrary temporal dependency.

iii $\exists C_0 > 0$, such that $\|\mathbf{\Sigma}\|_2 < C_0$.

iv The tail of the factors is sub-Gaussian.

Modeling Assumptions And General Methodology

Regime 1: External covariates are not informative

- (Pervasiveness) $\exists c_{\min}, c_{\max} > 0$, so that $c_{\min} < \lambda_{\min}(p^{-1}\mathbf{\Lambda}^i\mathbf{\Lambda}^{i'}) < \lambda_{\max}(p^{-1}\mathbf{\Lambda}^i\mathbf{\Lambda}^{i'}) < c_{\max}$.
- $\max_{k\leq K^i, j\leq p} |\lambda^i_{jk}| = O_P(\sqrt{\log p})$.

Regime 2: External covariates are informative

- (Pervasiveness) $\exists c_{\min}, c_{\max} > 0$, so that $c_{\min} < \lambda_{\min}(p^{-1}\mathbf{G}^i(\mathbf{W}^i)\mathbf{G}^i(\mathbf{W}^i)') < \lambda_{\max}(p^{-1}\mathbf{G}^i(\mathbf{W}^i)\mathbf{G}^i(\mathbf{W}^i)') < c_{\max}$.
- $\max_{k\leq K^i, j\leq p} E_{g_k}(W^i_j)^2 \leq \infty$.
- $\max_{k\leq K^i, j\leq p} |\gamma^i_{jk}| = O_P(\sqrt{\log p})$.

The ALPHA Framework

This section covers details for heterogeneity adjustments under both regimes that $G_i() = 0$ and $G_i() \neq 0$: they correspond to estimating U_i by either PCA or Projected-PCA.
From now on, we drop the superscript i whenever there is no confusion as we focus on the ith data source. We will use the notation $(\hat{\mathbf{F}})$ if $\mathbf{F}$ is estimated by PCA and $\tilde{\mathbf{F}}$ if estimated by PPCA. This convention applies to other related quantities such as $(\hat{\mathbf{U}})$ and $\tilde{\mathbf{U}}$, the heterogeneity adjusted estimator. In addition, we use notations such as $\check{\mathbf{F}}$ and $\check{\mathbf{U}}$ to denote the final estimators.
By the priciple of least squre, the residul estimator of $\mathbf{U}$ admits the form $\check{\mathbf{U}} = \mathbf{X}(\mathbf{I} - \frac{1}{n}\check{\mathbf{F}}\check{\mathbf{F}}')$.

Estimating factors by PCA

Estimating factors by Projected-PCA

References

Jianqing Fan, Han Liu, Weichen Wang and Ziwei Zhu (2012)
Heterogeneity Adjustment with Applications to Graphical Model Inference
Journal of the American Statistical Association Under review.

The End