Two-Layer Heterogeneity Model for Massive Data

Ching-Wei Cheng
(Ongoing joint work with Prof. Guang Cheng)

August 9, 2016
(BaT Group Meeting)

Heterogeneity is a promising yet challenging feature in Big Data

Example 1 (US Big-Data Health Network Launches Aspirin Study)

- Initiation of a 10-million pilot study that aim to investigate the use of aspirin to prevent heart disease
- Various healthcare data (e.g., insurance claims, blood tests, medical histories, etc.) will be collected from as many as 30 million people in the United States through a large healthcare network

---

http://www.nature.com/news/us-big-data-health-network-launches-aspirin-study-1.15675

- To fix terminology...
  - Observation $\subseteq$ Unit $\subseteq$ Cluster/subgroup $\subseteq$ Full data (massive)
- Different data units may be endowed with different features (**Heterogeneity**)
  - However, some data units may be similar enough to treated homogeneous
  - Data similarity implies potential clustering effects among the data units, and thus encourages the development of multi-layer heterogeneity models

An Illustration of Nestedly Heterogeneous Data

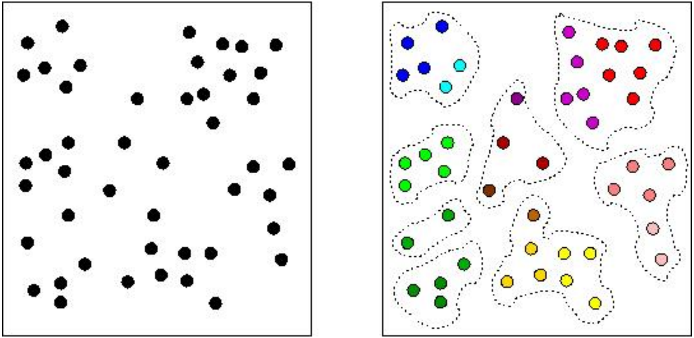


Figure 1 : Every dot is a data unit; colors and dashed lines indicate possible nested clustering structure

Graph is downloaded from http://tcs.legacy.ics.tkk.fi/~satu/mja/

In this talk, we focus on linear model family to illustrate our idea

- Linear models $\mathcal{Y} = \mathcal{X}^\top \boldsymbol{\beta} + \epsilon$
  - Simple and easy to interpret, but homogeneous
- Linear fixed-effect models $\mathcal{Y} = \mathcal{X}^\top \boldsymbol{\beta} + \mathcal{Z}^\top \boldsymbol{\theta}_i + \epsilon$
  - Fixed effects $\boldsymbol{\theta}_i$ indicates (finite) heterogeneity among units
- Linear mixed-effects models (LMMs) $\mathcal{Y} = \mathcal{X}^\top \boldsymbol{\beta} + \mathcal{Z}^\top \boldsymbol{\vartheta}_i + \epsilon$ with $\mathsf{E}[\boldsymbol{\vartheta}_i] = \boldsymbol{\theta}_i$ and $\mathsf{Cov}(\boldsymbol{\vartheta}_i) = \sigma_u^2 \boldsymbol{I}$
  - A well-known powerful tool for grouped (e.g., longitudinal, panel, cross-sectional) data
  - Random effects $\boldsymbol{\vartheta}_i$ accounts for unit heterogeneity
  - Relax independence assumption
- Write $\boldsymbol{\vartheta}_i = \boldsymbol{\theta}_i + \boldsymbol{u}_i$
  - Fixed $\boldsymbol{\theta}_i$ (between-unit heterogeneity)
  - Random $\boldsymbol{u}_i$ with $\mathsf{E}[\boldsymbol{u}_i] = \mathbf{0}$ and $\mathsf{Cov}(\boldsymbol{u}_i) = \sigma_u^2 \boldsymbol{I}$
- We then consider some of $\boldsymbol{\theta}_i$'s are identical
  - Resulting in, if unit i belongs to subgroup s,

$$\begin{aligned}\mathcal{Y} &= \mathcal{X}^\top \boldsymbol{\beta} + \mathcal{Z}^\top \boldsymbol{\theta}_i + \mathcal{Z}^\top \boldsymbol{u}_i + \epsilon \\ &= \mathcal{X}^\top \boldsymbol{\beta} + \mathcal{Z}^\top \boldsymbol{\alpha}_s + \mathcal{Z}^\top \boldsymbol{u}_i + \epsilon\end{aligned}$$

Outline

The Model

- Two-Layer heterogeneity model (THM):

$$\mathcal{Y} = \mathcal{X}^\top \boldsymbol{\beta} + \mathcal{Z}^\top \boldsymbol{\theta}_i + \mathcal{Z}^\top \boldsymbol{u}_i + \epsilon \tag{1}$$

  - $\boldsymbol{\beta}$ is the p-vector of common fixed effects across all units
  - $\boldsymbol{\theta}_i$ is the q-vector of unit-specific fixed effects of unit i
  - $\boldsymbol{u}_i$ is the q-vector of unit-specific random effects with $\mathsf{E}[\boldsymbol{u}_i] = \mathbf{0}$ and $\mathsf{Cov}(\boldsymbol{u}_i) = \sigma_u^2 \boldsymbol{I}$
  - $\boldsymbol{\epsilon}_i$ is the error n_i-vector with $\mathsf{E}[\boldsymbol{\epsilon}_i] = \mathbf{0}$ and $\mathsf{Cov}(\boldsymbol{\epsilon}_i) = \sigma_\epsilon^2$
- Massive data $\left\{(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{z}_i, L_i)\right\}_{i=1}^M$, aggregated from M data units
  - $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})^\top$ is a n_i-vector of responses
  - $\boldsymbol{x}_i$ ($\boldsymbol{z}_i$) is an $n_i \times p$ ($n_i \times q$) design matrx
  - L_i is the (latent) subgroup label for unit i
  - Total sample size $N = \sum_{i=1}^M n_i$

The Model

Practical Model Formulation

- **Unit level:**

$$\boldsymbol{y}_i = \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{\theta}_i + \boldsymbol{z}_i\boldsymbol{u}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \dots, M \tag{2}$$

  - Define $\boldsymbol{W}_i = \mathsf{Cov}(\boldsymbol{y}_i)^{-1} = (\sigma_\epsilon^2 \boldsymbol{I}_{n_i} + \sigma_u^2 \boldsymbol{z}_i\boldsymbol{z}_i^\top)^{-1}$

- **Full-data level:**

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\Theta} + \boldsymbol{Z}\boldsymbol{U} + \boldsymbol{\mathcal{E}} \tag{3}$$

  - $\boldsymbol{Y}_{(N\times 1)} = (\boldsymbol{y}_1^\top, \dots, \boldsymbol{y}_M^\top)^\top$
  - $\boldsymbol{X}_{(N\times p)} = (\boldsymbol{x}_1^\top, \dots, \boldsymbol{x}_M^\top)^\top$
  - $\boldsymbol{Z}_{(N\times Mq)} = \mathrm{bdiag}(\boldsymbol{z}_1, \dots, \boldsymbol{z}_M)$
  - $\boldsymbol{\Theta}_{(Mq\times 1)} = (\boldsymbol{\theta}_1^\top, \dots, \boldsymbol{\theta}_M^\top)^\top$
  - $\boldsymbol{U}_{(Mq\times 1)} = (\boldsymbol{u}_1^\top, \dots, \boldsymbol{u}_M^\top)^\top$
  - $\boldsymbol{\mathcal{E}}_{(N\times 1)} = (\boldsymbol{\epsilon}_1^\top, \dots, \boldsymbol{\epsilon}_M^\top)^\top$
  - Define $\boldsymbol{W} = \mathsf{Cov}(\boldsymbol{Y})^{-1} = \mathrm{bdiag}(\boldsymbol{W}_i)$

The Model

Oracular Model Representations

When subgroup labels L_i's are known, model (1) can be expressed as

$$\begin{aligned} \mathcal{Y} &= \mathcal{X}^\top \boldsymbol{\beta} + \mathcal{Z}^\top \boldsymbol{\theta}_i + \mathcal{Z}^\top \boldsymbol{u}_i + \epsilon \\ &= \mathcal{X}^\top \boldsymbol{\beta} + \mathcal{Z}^\top \boldsymbol{\alpha}_s + \mathcal{Z}^\top \boldsymbol{u}_i + \epsilon, \end{aligned} \tag{4}$$

where $\boldsymbol{\alpha}_s$ is the q-vector of group-specific fixed effects of subgroup s

- $\boldsymbol{\alpha}_{(Sq\times 1)} = (\boldsymbol{\alpha}_1^\top, \ldots, \boldsymbol{\alpha}_S^\top)^\top$
- Define

$$\mathcal{M}_{\mathcal{G}} = \left\{ \boldsymbol{\Theta} \in \mathbb{R}^{Mq} : \boldsymbol{\theta}_i = \boldsymbol{\theta}_j \text{ for any } L_i = L_j \right\}$$

- For each $\boldsymbol{\Theta} \in \mathcal{M}_{\mathcal{G}}$, we have $\boldsymbol{\Theta} = \boldsymbol{A}\boldsymbol{\alpha}$, where $\boldsymbol{A}$ is a $(Mq) \times (Sq)$ matrix with the (i, s) block being $\boldsymbol{I}_q$ if $L_i = s$, and $\boldsymbol{O}_q$ otherwise.
  - An example of matrix $\boldsymbol{A}$ will be delivered later...

The Model

Oracular Model Representations

- **Unit level:**

$$\boldsymbol{y}_i = \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{\alpha}_s + \boldsymbol{z}_i\boldsymbol{u}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, M. \tag{5}$$

- **Full-data level:** (Recall that $\boldsymbol{\Theta} = \boldsymbol{A}\boldsymbol{\alpha}$)

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{A}\boldsymbol{\alpha} + \boldsymbol{Z}\boldsymbol{U} + \boldsymbol{\mathcal{E}}, \tag{6}$$

  - Oracle estimator, defined via GLS:

$$\begin{pmatrix} \widehat{\boldsymbol{\beta}}_{\mathrm{OR}} \\ \widehat{\boldsymbol{\alpha}}_{\mathrm{OR}} \end{pmatrix} = \left[(\boldsymbol{X}, \boldsymbol{Z}\boldsymbol{A})^\top \boldsymbol{W} (\boldsymbol{X}, \boldsymbol{Z}\boldsymbol{A})\right]^{-1} (\boldsymbol{X}, \boldsymbol{Z}\boldsymbol{A})^\top \boldsymbol{W}\boldsymbol{Y} \tag{7}$$

The Model

Oracular Model Representations

Example 2

Suppose $M = 5, S = 2$, $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 = \boldsymbol{\alpha}_1$ and $\boldsymbol{\theta}_3 = \boldsymbol{\theta}_4 = \boldsymbol{\theta}_5 = \boldsymbol{\alpha}_2$. Then we have

$$\boldsymbol{\Theta}_{(5q)\times 1} = \begin{pmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \\ \boldsymbol{\theta}_3 \\ \boldsymbol{\theta}_4 \\ \boldsymbol{\theta}_5 \end{pmatrix}_{(5q)\times 1} = \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \\ \boldsymbol{\alpha}_2 \\ \boldsymbol{\alpha}_2 \end{pmatrix}_{(5q)\times 1} = \begin{bmatrix} \boldsymbol{I}_q & \\ \boldsymbol{I}_q & \\ & \boldsymbol{I}_q \\ & \boldsymbol{I}_q \\ & \boldsymbol{I}_q \end{bmatrix}_{(5q)\times(2q)} \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \end{pmatrix}_{(2q)\times 1}$$

and

$$\boldsymbol{Z}\boldsymbol{A} = \begin{bmatrix} \boldsymbol{z}_1 & & & & \\ & \boldsymbol{z}_2 & & & \\ & & \boldsymbol{z}_3 & & \\ & & & \boldsymbol{z}_4 & \\ & & & & \boldsymbol{z}_5 \end{bmatrix} \begin{bmatrix} \boldsymbol{I}_q & \\ \boldsymbol{I}_q & \\ & \boldsymbol{I}_q \\ & \boldsymbol{I}_q \\ & \boldsymbol{I}_q \end{bmatrix} = \begin{bmatrix} \boldsymbol{z}_1 & \\ \boldsymbol{z}_2 & \\ & \boldsymbol{z}_3 \\ & \boldsymbol{z}_4 \\ & \boldsymbol{z}_5 \end{bmatrix},$$

The Model

A CD Fusion Approach for Massive Data

When dealing with massive data, GLS may be computationally unavailable...

- Due to massive sample size, direct estimation towards the full-data model (3) is usually unavailable
  - GLS approach can be applied for LMMs, but may be computationally infeasible due to massive sample size
- Starting with unit GLS estimates

$$\begin{pmatrix} \widehat{\boldsymbol{\beta}}_i \\ \widehat{\boldsymbol{\theta}}_i \end{pmatrix} = \left[(\boldsymbol{x}_i, \boldsymbol{z}_i)^\top \boldsymbol{W}_i (\boldsymbol{x}_i, \boldsymbol{z}_i)\right]^{-1} (\boldsymbol{x}_i, \boldsymbol{z}_i)^\top \boldsymbol{W}_i \boldsymbol{y}_i, \quad i = 1, \ldots, M \tag{8}$$

  - If σ_ϵ^2 and σ_u^2 are unknown, they can be consistently estimated by restricted maximum likelihood (REML) method
  - Throughout this talk, we assume σ_ϵ^2 and σ_u^2 are known for simplicity
- We adopt the confidence distribution (CD) concept to merge the unit estimates
- To discover potential subgroups, we consider pairwise concave fusion penalty as in Ma and Huang (2016)

Confidence Distribution (CD)

- A CD can be viewed as
  - "A sample-dependent distribution function that can represent confidence intervals of all levels for a parameter of interest"
  - "The frequentist distribution estimator of a parameter"
- See Xie and Singh (2013) for a comprehensive review for the CD development
- The CD concept has been shown effective for combining information from independent sources (unit estimates)
- We are going to adopt the CD approach proposed by Liu *et al.* (2015) for multi-parameter estimates

The Model

A CD Fusion Approach for Massive Data

A CD fusion estimator can be formed as follows:

- Given $(\widehat{\boldsymbol{\beta}}_i^\top, \widehat{\boldsymbol{\theta}}_i^\top)^\top \xrightarrow{D} \mathcal{N}\big((\boldsymbol{\beta}_0^\top, \boldsymbol{\theta}_{i,0}^\top)^\top, [(\boldsymbol{x}_i, \boldsymbol{z}_i)^\top \boldsymbol{W}_i (\boldsymbol{x}_i, \boldsymbol{z}_i)]^{-1}\big)$
- Following the CD concept, the CD densities $h_i(\boldsymbol{\beta}, \boldsymbol{\theta}_i)$ can be defined as the density function of $\mathcal{N}\big((\widehat{\boldsymbol{\beta}}_i^\top, \widehat{\boldsymbol{\theta}}_i^\top)^\top, [(\boldsymbol{x}_i, \boldsymbol{z}_i)^\top \boldsymbol{W}_i (\boldsymbol{x}_i, \boldsymbol{z}_i)]^{-1}\big)$
- The combined CD density can be defined by

$$h(\boldsymbol{\beta}, \boldsymbol{\Theta}) = \prod_{i=1}^{M} h_i(\boldsymbol{\beta}, \boldsymbol{\theta}_i)$$

- A CD estimator of $(\boldsymbol{\beta}, \boldsymbol{\Theta})$ is defined as $\arg\max_{\boldsymbol{\beta}, \boldsymbol{\Theta}} \log h(\boldsymbol{\beta}, \boldsymbol{\Theta})$ (yet not what we want)
- Recall that some underlying values of $\boldsymbol{\theta}_i$'s are assumed identical, and so we incorporate a pairwise concave fusion penalty into the objective function for a fusion estimation

The Model

A CD Fusion Approach for Massive Data

We propose a **CD fusion estimator** defined by

$$\begin{pmatrix} \breve{\boldsymbol{\beta}}(\lambda) \\ \breve{\boldsymbol{\Theta}}(\lambda) \end{pmatrix} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{\Theta} \in \mathbb{R}^{Mq}}{\arg\min} Q_N^{\mathrm{CD}}(\boldsymbol{\beta}, \boldsymbol{\Theta}), \tag{9}$$

where

$$\begin{aligned} Q_N^{\mathrm{CD}}(\boldsymbol{\beta}, \boldsymbol{\Theta}) = \frac{1}{2} \sum_{i=1}^{M} \begin{pmatrix} \widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta} \\ \widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i \end{pmatrix}^{\top} (\boldsymbol{x}_i, \boldsymbol{z}_i)^{\top} \boldsymbol{W}_i (\boldsymbol{x}_i, \boldsymbol{z}_i) \begin{pmatrix} \widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta} \\ \widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i \end{pmatrix} \\ + \sum_{1 \le i < j \le M} p_{\gamma} \left(\| \boldsymbol{\theta}_i - \boldsymbol{\theta}_j \| ; \lambda \right), \end{aligned} \tag{10}$$

- First term of the R.H.S. comes from the simplified $-\log h(\boldsymbol{\beta}, \boldsymbol{\Theta})$ by omitting additive constant terms, due to asymptotic normality of unit GLS estimates
- $p_{\gamma}(t; \lambda)$ is a concave penalty function with a tuning parameter $\lambda > 0$ and a parameter $\gamma > 0$ which is associated with the concavity of the penalty

The Model

Remarks on the CD fusion estimator:

- The objective function (10) can be rewritten as

$$\begin{aligned} Q_N^{\mathrm{CD}}(\boldsymbol{\beta}, \boldsymbol{\Theta}) = &\frac{1}{2}\sum_{i=1}^{M}(\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta})^\top \boldsymbol{x}_i^\top \boldsymbol{W}_i \boldsymbol{x}_i (\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}) \\ &+ \frac{1}{2}\sum_{i=1}^{M}(\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)^\top \boldsymbol{z}_i^\top \boldsymbol{W}_i \boldsymbol{z}_i (\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i) \\ &+ \sum_{1 \le i < j \le M} p_\gamma \left(\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\| \,; \lambda \right). \end{aligned} \tag{11}$$

  - Minimizer of $\boldsymbol{\beta}$ is actually free of $\boldsymbol{\Theta}$, resulting in

$$\breve{\boldsymbol{\beta}}(\lambda) \equiv \breve{\boldsymbol{\beta}} = \left(\sum_{i=1}^{M} \boldsymbol{x}_i^\top \boldsymbol{W}_i \boldsymbol{x}_i \right)^{-1} \left(\sum_{i=1}^{M} \boldsymbol{x}_i^\top \boldsymbol{W}_i \boldsymbol{x}_i \widehat{\boldsymbol{\beta}}_i \right). \tag{12}$$

  - To find $\breve{\boldsymbol{\Theta}}(\lambda)$, we reformulate the objective function as an augmented Lagrangian problem, and establish an alternating direction method of multipliers (ADMM, Boyd *et al.* (2010)) algorithm

The Model

Remarks on $p_\gamma(t;\lambda)$:

- L_1 or Lasso penalty produces biased estimates and thus may not correctly form the clusters
  - Tends to result in either a large number of subgroups or no subgroup on the soplution path
- MCP and SCAD penalty are more appropriate
  - Enjoy sparsity (on pairwise distances) as the L_1 penalty so that they automatically fuse some of $\boldsymbol{\theta}_i$'s together
  - Due to their unbiasedness peoperty, they do not shrink large estimated parameters
  - Remain unbiased in ADMM iterations for concave optimization solvers

Theoretical Properties

An Intermediate Estimator

- We introduce an intermediate estimator:

$$\begin{pmatrix} \widehat{\boldsymbol{\beta}}(\lambda) \\ \widehat{\boldsymbol{\Theta}}(\lambda) \end{pmatrix} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p, \boldsymbol{\Theta}\in\mathbb{R}^{Mq}}{\arg\min}\ Q_N(\boldsymbol{\beta}, \boldsymbol{\Theta}), \tag{13}$$

  where

$$\begin{aligned} Q_N(\boldsymbol{\beta}, \boldsymbol{\Theta}) = &\frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{\Theta})^\top \boldsymbol{W}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{\Theta}) \\ &+ \sum_{1\le i<j\le M} p_\gamma(\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\|; \lambda), \end{aligned} \tag{14}$$

  - A GLS version of the estimator proposed by Ma and Huang (2016)
- Our analysis strategy:
  - Prove the CD fusion estimator is equivalent to the intermediate estimator
  - Show oracle properties of the proposed CD fusion estimator through the intermediate estimator

Theoretical Properties

An Intermediate Estimator

Theorem 2.1 (Equivalence between the proposed CD fusion estimator and the intermediate estimator)

Let

$$\begin{pmatrix} \breve{\boldsymbol{\beta}}(\lambda) \\ \breve{\boldsymbol{\Theta}}(\lambda) \end{pmatrix} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p, \boldsymbol{\Theta}\in\mathbb{R}^{Mq}}{\arg\min} Q_N^{\mathrm{CD}}(\boldsymbol{\beta}, \boldsymbol{\Theta})$$

and

$$\begin{pmatrix} \widehat{\boldsymbol{\beta}}(\lambda) \\ \widehat{\boldsymbol{\Theta}}(\lambda) \end{pmatrix} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p, \boldsymbol{\Theta}\in\mathbb{R}^{Mq}}{\arg\min} Q_N(\boldsymbol{\beta}, \boldsymbol{\Theta}),$$

where the objective functions $Q_N^{\mathrm{CD}}(\boldsymbol{\beta}, \boldsymbol{\Theta})$ *and* $Q_N(\boldsymbol{\beta}, \boldsymbol{\Theta})$ *are defined in* (10) *and* (14)*, respectively. Then we have*

$$P\left(\begin{pmatrix} \breve{\boldsymbol{\beta}} \\ \breve{\boldsymbol{\Theta}}(\lambda) \end{pmatrix} = \begin{pmatrix} \widehat{\boldsymbol{\beta}}(\lambda) \\ \widehat{\boldsymbol{\Theta}}(\lambda) \end{pmatrix}\right) = 1.$$

Remarks on Theorem 2.1:

- The intermediate estimator uses the full data at once, and thus can be treated as the IPD (individual participant data) estimator, which is taken as the gold standard in meta-analysis
- Liu *et al.* (2015) showed that the CD estimator is asymptotically equivalent to the IPD estimator
- Our equivalence result is stronger (non-asymptotic)
  - May result from the LMM structure and GLS estimation approach

When the true subgroup membership L_i's are known, define the following:

- $\mathcal{G}_s = \{(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{z}_i, L_i) : L_i = s\}, s = 1, \ldots, S$
- $M_s = |\mathcal{G}_s|$, the number of units in subgroup s
  - $M_{\min} = \min_{1 \leq s \leq S} M_s$
  - $M_{\max} = \max_{1 \leq s \leq S} M_s$
- $g_s = \sum_{i:L_i=s} n_i$, the number of observations in subgroup s
  - $g_{\min} = \min_{1 \leq s \leq S} g_s$
  - $g_{\max} = \max_{1 \leq s \leq S} g_s$
- The oracle estimator is defined from model (6) by

$$\begin{pmatrix} \widehat{\boldsymbol{\beta}}_{\mathrm{OR}} \\ \widehat{\boldsymbol{\alpha}}_{\mathrm{OR}} \end{pmatrix} = \left[(\boldsymbol{X}, \boldsymbol{Z}\boldsymbol{A})^\top \boldsymbol{W} (\boldsymbol{X}, \boldsymbol{Z}\boldsymbol{A})\right]^{-1} (\boldsymbol{X}, \boldsymbol{Z}\boldsymbol{A})^\top \boldsymbol{W}\boldsymbol{Y} \tag{15}$$

Theoretical Properties

Oracle Properties

Theorem 2.2 (Asymptotics for the oracle estimator)

Assume regularity conditions. If $g_{\min} \gg \sqrt{(p+Sq)N\log N}$ *and* $p+Sq = o(N^{1\wedge\tau})$, *we have*

$$\left\| \begin{pmatrix} \widehat{\boldsymbol{\beta}}_{\mathrm{OR}} - \boldsymbol{\beta}_0 \\ \widehat{\boldsymbol{\Theta}}_{\mathrm{OR}} - \boldsymbol{\Theta}_0 \end{pmatrix} \right\| \le \phi_N = O_P\left(g_{\min}^{-1} \sqrt{(p+Sq)N\log N} \right), \tag{16}$$

$$\left\| \widehat{\boldsymbol{\beta}}_{\mathrm{OR}} - \boldsymbol{\beta}_0 \right\| = O_P\left(\sqrt{\frac{p\log N}{N}} \right), \tag{17}$$

$$\left\| \widehat{\boldsymbol{\alpha}}_{\mathrm{OR}} - \boldsymbol{\alpha}_0 \right\| = O_P\left(\sqrt{\frac{Sq\log g_{\min}}{g_{\min}}} \right). \tag{18}$$

Moreover, for any sequence of $(p+Sq)$*-vectors* $\{\boldsymbol{a}_N\}$, *we have*

$$\sigma_N^{-1}(\boldsymbol{a}_N)\boldsymbol{a}_N^\top \begin{pmatrix} \widehat{\boldsymbol{\beta}}_{\mathrm{OR}} - \boldsymbol{\beta}_0 \\ \widehat{\boldsymbol{\alpha}}_{\mathrm{OR}} - \boldsymbol{\alpha}_0 \end{pmatrix} \xrightarrow{D} \mathcal{N}(0,1), \tag{19}$$

where $\sigma_N(\boldsymbol{a}_N) = \left(\boldsymbol{a}_N^\top \left[(\boldsymbol{X}, \boldsymbol{Z}\boldsymbol{A})^\top \boldsymbol{W} (\boldsymbol{X}, \boldsymbol{Z}\boldsymbol{A}) \right]^{-1} \boldsymbol{a}_N \right)^{1/2}$.

Theoretical Properties

Oracle Properties

- For $S \geq 2$, let

$$\Delta_N = \min_{L_i = s, L_j = s', s \neq s'} \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\| = \min_{s \neq s'} \|\boldsymbol{\alpha}_s - \boldsymbol{\alpha}_{s'}\|$$

denote the minimal signal measure for the unit-specific fixed effects

Theorem 2.3 (Oracle property)

Suppose regularity conditions and $S \geq 2$. If $\Delta_N \gg a\phi_N$, $\lambda \gg \phi_N$, where a is defined in the assumption for penalty functions and ϕ_N is given in Theorem 2.2, then there exists a local minimizer $(\breve{\boldsymbol{\beta}}^\top, \breve{\boldsymbol{\Theta}}(\lambda)^\top)^\top$ of the objective function $Q_N^{\mathrm{CD}}(\boldsymbol{\beta}, \boldsymbol{\Theta})$ given in (10) satisfying

$$P\left(\begin{pmatrix} \breve{\boldsymbol{\beta}} \\ \breve{\boldsymbol{\Theta}}(\lambda) \end{pmatrix} = \begin{pmatrix} \widehat{\boldsymbol{\beta}}_{\mathrm{OR}} \\ \widehat{\boldsymbol{\Theta}}_{\mathrm{OR}} \end{pmatrix}\right) \to 1.$$

Corollary 2.1

Under the conditions in Theorem 2.3, we have for any $(p+Sq)$*-vector* $\boldsymbol{a}_N$*,*

$$\sigma_N^{-1}(\boldsymbol{a}_N)\boldsymbol{a}_N^\top \begin{pmatrix} \breve{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \\ \breve{\boldsymbol{\alpha}}(\lambda) - \boldsymbol{\alpha}_0 \end{pmatrix} \xrightarrow{D} \mathcal{N}(0,1),$$

where $\sigma_N(\boldsymbol{a}_N)$ *is given in Theorem 2.2. Moreover, we have for any* $\boldsymbol{a}_{N1} \in \mathbb{R}^p$ *and* $\boldsymbol{a}_{N2} \in \mathbb{R}^{Sq}$*,*

$$\sigma_{N1}^{-1}(\boldsymbol{a}_{N1})\boldsymbol{a}_{N1}^\top\left(\breve{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) \xrightarrow{D} \mathcal{N}(0,1),$$
$$\sigma_{N2}^{-1}(\boldsymbol{a}_{N2})\boldsymbol{a}_{N2}^\top\left(\breve{\boldsymbol{\alpha}}(\lambda) - \boldsymbol{\alpha}_0\right) \xrightarrow{D} \mathcal{N}(0,1),$$

where

$$\sigma_{N1}(\boldsymbol{a}_{N1}) = \left\{\boldsymbol{a}_{N1}^\top \left(\boldsymbol{X}^\top \boldsymbol{Q}_{\boldsymbol{ZA}} \boldsymbol{X}\right)^{-1} \boldsymbol{a}_{N1}\right\}^{1/2},$$
$$\sigma_{N2}(\boldsymbol{a}_{N2}) = \left\{\boldsymbol{a}_{N2}^\top \left[(\boldsymbol{ZA})^\top \boldsymbol{Q}_{\boldsymbol{X}} (\boldsymbol{ZA})\right]^{-1} \boldsymbol{a}_{N2}\right\}^{1/2}.$$

Remarks on Theorem 2.3 and Corollary 2.1:

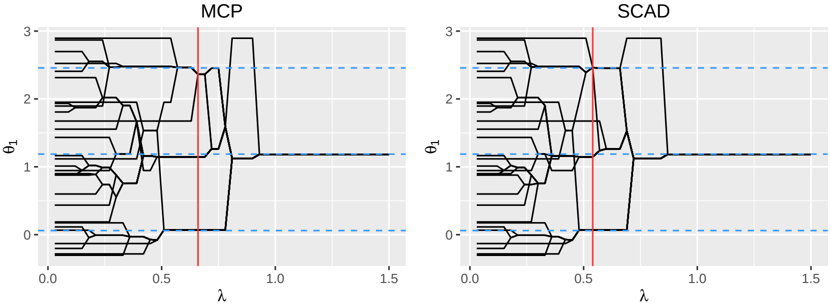
- As $\boldsymbol{\theta}_i$'s can be consistently estimated by $\widehat{\boldsymbol{\theta}}_i(\lambda)$'s, the subgroup membership L_i's and the cluster-specific fixed effects $\boldsymbol{\alpha}_s$'s can be obtained consistently as well
- The asymptotic variances $\sigma_{N1}(\boldsymbol{a}_{N1})$ and $\sigma_{N2}(\boldsymbol{a}_{N2})$ are derived from $\mathsf{Cov}\big[(\widehat{\boldsymbol{\beta}}_{\mathrm{OR}}^\top, \widehat{\boldsymbol{\alpha}}_{\mathrm{OR}}^\top)^\top\big]$
  - $(\widehat{\boldsymbol{\beta}}_{\mathrm{OR}}^\top, \widehat{\boldsymbol{\alpha}}_{\mathrm{OR}}^\top)^\top$ is formed via GLS, which produces best linear unbiased estimator (BLUE)
  - $\sigma_{N1}(\boldsymbol{a}_{N1})$ and $\sigma_{N2}(\boldsymbol{a}_{N2})$ are the smallest asymptotic variances of $\boldsymbol{a}_{N1}^\top \breve{\boldsymbol{\beta}}(\lambda)$ and $\boldsymbol{a}_{N2}^\top \breve{\boldsymbol{\alpha}}(\lambda)$ for any $\boldsymbol{a}_{N1} \in \mathbb{R}^p$ and $\boldsymbol{a}_{N2} \in \mathbb{R}^{Sq}$
  - In summary, Corollary 2.1 suggests that the optimal inference be made by the proposed CD fusion estimator

- Tuning parameter selection:
  - To select λ, we apply the modified BIC (Wang *et al.*, 2009) for diverging parameter dimension
    - $\breve{\lambda} = \arg\min_\lambda \mathrm{BIC}(\lambda)$
  - $\gamma = 3$
- **Solution path analysis** for $\theta_{s,1}$:
  - Blue dashed lines are $\widehat{\alpha}_{\mathrm{OR},s,1}$'s
  - Red solid line indicates $\breve{\lambda}$ (via grid search)

Simulation

Solution Path Analysis

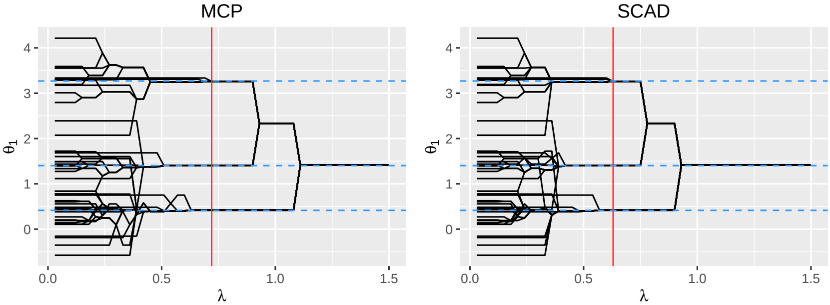
$$n_i \equiv n = 256, M = 30, S = 3; p = 5, q = 4$$



Simulation

Solution Path Analysis

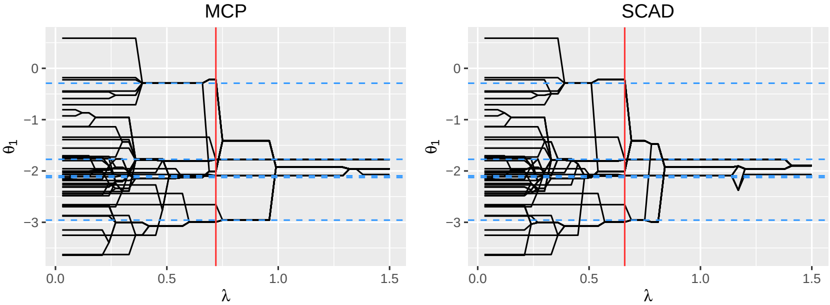
$$n_i \equiv n = 256, M = 50, S = 3; p = 5, q = 4$$



Simulation

Solution Path Analysis

$$n_i \equiv n = 256, M = 50, S = 5; p = 5, q = 4$$



Discussion

- The theoretical results are based on fixed design
  - Currently trying to move to random design regime, which is more appropriate for Big Data
  - Need to deal with lots of probability statements (e.g., bounded columns, eigenvalue bounds, projection matrices, etc.)
- The exhaustive pairwise fusion penalty is too costly when M is large (i.e., $M(M-1)/2$ pairs)
  - Have tried vectorization, i.e., there exists a sparse matrix $\boldsymbol{B}$ s.t. $\boldsymbol{B}^\top \boldsymbol{\Theta}$ includes all pairs $\boldsymbol{\theta}_i - \boldsymbol{\theta}_j$, but seems not good enough
  - May consider the aCARDS proposed by Ke *et al.* (2015)
    - Make a rough segmentation (e.g., $M^{0.7}$ segments)
    - Use a hybrid pairwise fusion penalty to consider between-segment and within-segment penalties
- More evaluation measures for clustering and estimation in simulation

Selected References

Boyd, S., Parikh, N., Chu, E., Peleato, B. and Eckstein, J. (2010) Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, **3**, 1–122. URL http://dx.doi.org/10.1561/2200000016.

Ke, Z. T., Fan, J. and Wu, Y. (2015) Homogeneity pursuit. *Journal of the American Statistical Association*, **110**, 175–194. URL http://dx.doi.org/10.1080/01621459.2014.892882.

Liu, D., Liu, R. Y. and Xie, M. (2015) Multivariate meta-analysis of heterogeneous studies using only summary statistics: Efficiency and robustness. *Journal of the American Statistical Association*, **110**, 326–340. URL http://dx.doi.org/10.1080/01621459.2014.899235.

Ma, S. and Huang, J. (2016) A concave pairwise fusion approach to subgroup analysis. *Journal of the American Statistical Association*. URL http://dx.doi.org/10.1080/01621459.2016.1148039. To appear.

Wang, H., Li, B. and Leng, C. (2009) Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71**, 671–683. URL http://dx.doi.org/10.1111/j.1467-9868.2008.00693.x.

Xie, M. and Singh, K. (2013) Confidence distribution, the frequentist distribution estimator of a parameter: A review. *International Statistical Review*, **81**, 3–39. URL http://dx.doi.org/10.1111/insr.12000.