Robust Subgroup Analysis of High Dimensional Data

Chao Cheng
Author: Chao Cheng, Xingdong Feng

School of Statistics and Management
Shanghai University of Finance and Economics

December 11, 2019

Outline

Introduction

- Heterogeneity is widely presented in practice, precision medicine, the Iris Dataset, etc.
- Subgroup analysis can be done as a means of investigating heterogeneous results.
  - Based on observed attributes: clinical analysis
  - Explore the latent subgroups:
    - mixture model
    - penalty method

Model Setting

Consider the following model

$$y_i = \mu_i + \boldsymbol{x}_i^T \beta + \varepsilon_i, \qquad i = 1, 2, \cdots, n, \tag{1}$$

where β is a p-dimensional vector.

- Heterogeneous treatment effect:
  There is a partition $\boldsymbol{G}_0 = \{G_1, \cdots, G_{K_0}\}$ of $\{1, \cdots, n\}$, such that

  $$\mu_i = \alpha_k, \text{ for all } i \in G_k.$$

  Therefore α_k is the common value for subgroup G_k.
- Sparse covariate effect in high dimension:
  Only the first q entries of β are non-zero, where $q \ll n$.

Proposed M-Estimator

We propose a penalized M-estimator for subgroup analysis and variable selection simulteneously.

$$\left(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\beta}}\right) = \underset{\boldsymbol{\mu},\boldsymbol{\beta}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \rho\left(y_i - \mu_i - \boldsymbol{x}_i^T \boldsymbol{\beta}\right) + \sum_{1 \le i < j \le n} P_{\lambda_1}\left(\mu_i - \mu_j\right) + \sum_{j=1}^{p} P_{\lambda_2}\left(\beta_j\right), \quad (2)$$

where ρ is a loss function and P_{λ_1} and P_{λ_2} are two penalties with parameters λ_1 and λ_2 respectively.

- $\mathrm{E}\psi(\varepsilon) = 0$ where ψ is the derivative of ρ.
- ρ is differentiable except at finite number of points.
- L_1: $\rho(x) = |x|$.
- L_2: $\rho(x) = x^2$.
- *Huber*:

$$\rho(x; c) = \begin{cases} \frac{1}{2}x^2 & |x| \leq c \\ c|x| - \frac{1}{2}c^2 & |x| > c, \end{cases}$$

  where c is a positive constant.
- Quantile regression: $\rho(x; \tau) = x(\tau - 1_{x<0})$, where $\tau \in (0, 1)$.

- LASSO: $P_\lambda(x) = \lambda|x|$.
- SCAD:

$$P'_{\lambda,\gamma}(x) = \lambda\left\{I(x \le \lambda) + \frac{(\gamma\lambda - x)_+}{(\gamma-1)\lambda} I(x > \lambda)\right\}, \quad x > 0, \quad \gamma > 2.$$

- MCP:

$$P'_{\lambda,\gamma}(x) = \lambda\left(1 - \frac{x}{\lambda\gamma}\right)_+, \quad x > 0, \quad \gamma > 1.$$

Alternating Direction Method of Multipliers (ADMM)

We can rewrite the objective function in (2) to a optimization form:

$$\begin{aligned} &\min \frac{1}{n}\sum_{i=1}^{n}\rho\left(z_i\right)+\sum_{1\leq i<j\leq n}P_{\lambda_1}\left(s_{ij}\right)+\sum_{j=1}^{p}P_{\lambda_2}\left(w_j\right) \\ &\text{s.t.}\begin{cases}\boldsymbol{z}=\boldsymbol{y}-\boldsymbol{\mu}-\boldsymbol{X}\boldsymbol{\beta}\\ \boldsymbol{s}=\boldsymbol{D}\boldsymbol{\mu}\\ \boldsymbol{w}=\boldsymbol{\beta}\end{cases}, \end{aligned} \tag{3}$$

where $\boldsymbol{D}$ is the $\frac{n(n-1)}{2}\times n$ pairwise difference matrix, hence $s_{ij}=\mu_i-\mu_j$.

Alternating Direction Method of Multipliers (ADMM)

The augmented lagrangian form of (3) is

$$
\begin{aligned}
& L\left(\beta, \mu, \boldsymbol{z}, \boldsymbol{s}, \boldsymbol{w}, \boldsymbol{q}_1, \boldsymbol{q}_2, \boldsymbol{q}_3\right) \\
=& \frac{1}{n} \sum_{i=1}^{n} \rho\left(z_i\right)+\sum_{1 \leq i<j \leq n} P_{\lambda_1}\left(s_{ij}\right)+\sum_{j=1}^{p} P_{\lambda_2}\left(w_j\right) \\
&+\frac{r_1}{2}\|\boldsymbol{y}-\mu-\boldsymbol{X} \beta-\boldsymbol{z}\|_2^2+\frac{r_2}{2}\|\boldsymbol{D} \mu-\boldsymbol{s}\|_2^2+\frac{r_3}{2}\|\beta-\boldsymbol{w}\|_2^2 \\
&+\left\langle\boldsymbol{y}-\mu-\boldsymbol{X} \beta-\boldsymbol{z}, \boldsymbol{q}_1\right\rangle+\left\langle\boldsymbol{D} \mu-\boldsymbol{s}, \boldsymbol{q}_2\right\rangle+\left\langle\beta-\boldsymbol{w}, \boldsymbol{q}_3\right\rangle,
\end{aligned}
\tag{4}
$$

where r_1, r_2 and r_3 are positive constants, $\boldsymbol{q}_1$, $\boldsymbol{q}_2$ and $\boldsymbol{q}_3$ are multiplier vectors. $\|\boldsymbol{a}\|_2=\sqrt{\boldsymbol{a}^T \boldsymbol{a}}$ and $\langle\boldsymbol{a}, \boldsymbol{b}\rangle=\boldsymbol{a}^T \boldsymbol{b}$.

Alternating Direction Method of Multipliers (ADMM)

The augmented lagrangian form

$$
\begin{aligned}
& L\left(\beta, \mu, \boldsymbol{z}, \boldsymbol{s}, \boldsymbol{w}, \boldsymbol{q}_1, \boldsymbol{q}_2, \boldsymbol{q}_3\right) \\
=& \frac{1}{n} \sum_{i=1}^{n} \rho\left(z_i\right)+\sum_{1 \leq i<j \leq n} P_{\lambda_1}\left(s_{ij}\right)+\sum_{j=1}^{p} P_{\lambda_2}\left(w_j\right) \\
+& \frac{r_1}{2}\|\boldsymbol{y}-\mu-\boldsymbol{X} \beta-\boldsymbol{z}\|_2^2+\frac{r_2}{2}\|\boldsymbol{D} \mu-\boldsymbol{s}\|_2^2+\frac{r_3}{2}\|\beta-\boldsymbol{w}\|_2^2 \\
+& \left\langle\boldsymbol{y}-\mu-\boldsymbol{X} \beta-\boldsymbol{z}, \boldsymbol{q}_1\right\rangle+\left\langle\boldsymbol{D} \mu-\boldsymbol{s}, \boldsymbol{q}_2\right\rangle+\left\langle\beta-\boldsymbol{w}, \boldsymbol{q}_3\right\rangle .
\end{aligned}
$$

- L is quadratic in $\left(\mu^T, \beta^T\right)^T$.
- L is convex in $\boldsymbol{z}$, $\boldsymbol{s}$ and $\boldsymbol{w}$ when r_1, r_2, r_3 are properly chosen.
- L has the form of independent summation in $\boldsymbol{z}$, $\boldsymbol{s}$ and $\boldsymbol{w}$ when others are given.

Alternating Direction Method of Multipliers(ADMM)

We solve (4) in a coordinate descent fashion. For a given $\left(\beta^{(k)}, \mu^{(k)}, \boldsymbol{z}^{(k)}, \boldsymbol{s}^{(k)}, \boldsymbol{w}^{(k)}, \boldsymbol{q}_1^{(k)}, \boldsymbol{q}_2^{(k)}, \boldsymbol{q}_3^{(k)}\right)$ at step k. The update at step $k+1$ is given by:

- $\beta^{(k+1)} = \underset{\beta}{\operatorname{argmin}}\, L\left(\beta, \mu^{(k)}, \boldsymbol{z}^{(k)}, \boldsymbol{s}^{(k)}, \boldsymbol{w}^{(k)}, \boldsymbol{q}_1^{(k)}, \boldsymbol{q}_2^{(k)}, \boldsymbol{q}_3^{(k)}\right)$:

  If $p \leq n$, then

$$\beta^{(k+1)} = \left(r_1 \boldsymbol{X}^T \boldsymbol{X} + r_3 \boldsymbol{I}_p\right)^{-1} \boldsymbol{d}_1^{(k)}.$$

  If $p > n$, then

$$\beta^{(k+1)} = \frac{1}{r_3}\left\{\boldsymbol{I}_p - r_1 \boldsymbol{X}^T \left(r_1 \boldsymbol{X}\boldsymbol{X}^T + r_3 \boldsymbol{I}_n\right)^{-1} \boldsymbol{X}\right\} \boldsymbol{d}_1^{(k)},$$

  where $\boldsymbol{d}_1^{(k)} = r_1 \boldsymbol{X}^T \left(\boldsymbol{y} - \mu^{(k)} - \boldsymbol{z}^{(k)}\right) + r_3 \boldsymbol{w}^{(k)} + \boldsymbol{X}^T \boldsymbol{q}_1^{(k)} - \boldsymbol{q}_3^{(k)}$.

Alternating Direction Method of Multipliers(ADMM)

- $\boldsymbol{\mu}^{(k+1)} = \underset{\boldsymbol{\mu}}{\operatorname{argmin}}\, L\left(\boldsymbol{\beta}^{(k+1)}, \boldsymbol{\mu}, \boldsymbol{z}^{(k)}, \boldsymbol{s}^{(k)}, \boldsymbol{w}^{(k)}, \boldsymbol{q}_1^{(k)}, \boldsymbol{q}_2^{(k)}, \boldsymbol{q}_3^{(k)}\right)$:

$$\boldsymbol{\mu} = \left(r_1 \boldsymbol{I}_n + r_2 \boldsymbol{D}^T \boldsymbol{D}\right)^{-1} \boldsymbol{d}_2^{(k)},$$

where $\boldsymbol{d}_2^{(k)} = r_1\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{(k+1)} - \boldsymbol{z}^{(k)}\right) + r_2 \boldsymbol{D}^T \boldsymbol{s}^{(k)} + \boldsymbol{q}_1^{(k)} - \boldsymbol{D}^T \boldsymbol{q}_2^{(k)}$.

Alternating Direction Method of Multipliers(ADMM)

- $\boldsymbol{z}^{(k+1)} = \underset{\boldsymbol{z}}{\operatorname{argmin}}\, L\left(\beta^{(k+1)}, \boldsymbol{\mu}^{(k+1)}, \boldsymbol{z}, \boldsymbol{s}^{(k)}, \boldsymbol{w}^{(k)}, \boldsymbol{q}_1^{(k)}, \boldsymbol{q}_2^{(k)}, \boldsymbol{q}_3^{(k)}\right)$.
  The update of $\boldsymbol{z}$ depends on the choice of the loss ρ, and it can be computed elementwisely.
  - L_1:
    $$z_i^{(k+1)} = S\left(y_i - \mu_i^{(k+1)} - \boldsymbol{x}_i^T \beta^{(k+1)} + \frac{q_{1,i}^{(k)}}{r_1}, \frac{1}{nr_1}\right),$$
    where $S(x, \lambda)$ is the soft-thresholding function.
  - L_2:
    $$z_i^{(k+1)} = \frac{y_i - \mu_i^{(k+1)} - \boldsymbol{x}_i^T \beta^{(k+1)} + \frac{q_{1,i}^{(k)}}{r_1}}{1 + \frac{2}{nr_1}}.$$
    For L_2, $\boldsymbol{z}$ can be omitted in the algorithm.

Alternating Direction Method of Multipliers(ADMM)

- $\boldsymbol{z}^{(k+1)} = \underset{\boldsymbol{z}}{\operatorname{argmin}}\, L\left(\beta^{(k+1)}, \boldsymbol{\mu}^{(k+1)}, \boldsymbol{z}, \boldsymbol{s}^{(k)}, \boldsymbol{w}^{(k)}, \boldsymbol{q}_1^{(k)}, \boldsymbol{q}_2^{(k)}, \boldsymbol{q}_3^{(k)}\right)$.
  The update of $\boldsymbol{z}$ depends on the choice of the loss ρ, and it can be computed elementwisely. Let $d_{z,i}^{(k)} = y_i - \mu_i^{(k+1)} - \boldsymbol{x}_i^T \beta^{(k+1)} + \frac{q_{1,i}^{(k)}}{r_1}$.
  - *Huber*:

$$z_i^{(k+1)} = \begin{cases} S\left(d_{z,i}^{(k)}, \dfrac{c}{nr_1}\right) & \left(\dfrac{1}{nr_1}+1\right)c < \left|d_{z,i}^{(k)}\right| \\ \dfrac{d_{z,i}^{(k)}}{1+\frac{1}{nr_1}} & \left(\dfrac{1}{nr_1}+1\right)c \geq \left|d_{z,i}^{(k)}\right| \end{cases}.$$

  - Quantile regression:

$$z_i^{(k+1)} = \begin{cases} d_{z,i}^{(k)} - \dfrac{\tau}{nr_1} & \dfrac{\tau}{nr_1} < d_{z,i}^{(k)} \\ 0 & \dfrac{\tau-1}{nr_1} \leq d_{z,i}^{(k)} \leq \dfrac{\tau}{nr_1} \\ d_{z,i}^{(k)} + \dfrac{1-\tau}{nr_1} & d_{z,i}^{(k)} \leq \dfrac{\tau-1}{nr_1} \end{cases}.$$

Alternating Direction Method of Multipliers(ADMM)

- $\boldsymbol{s}^{(k+1)} = \underset{\boldsymbol{s}}{\operatorname{argmin}}\, L\left(\beta^{(k+1)}, \boldsymbol{\mu}^{(k+1)}, \boldsymbol{z}^{(k+1)}, \boldsymbol{s}, \boldsymbol{w}^{(k)}, \boldsymbol{q}_1^{(k)}, \boldsymbol{q}_2^{(k)}, \boldsymbol{q}_3^{(k)}\right)$.
  The update of $\boldsymbol{s}$ depends on the choice of the loss P_{λ_1}, and it can be computed elementwisely. Let $d_{s,ij}^{(k)} = \mu_i^{(k+1)} - \mu_j^{(k+1)} + \frac{q_{2,ij}^{(k)}}{r_2}$.
  - LASSO:
  $$s_{ij}^{(k+1)} = S\left(d_{s,ij}^{(k)}, \frac{\lambda_1}{r_2}\right).$$
  - MCP with $r_2\gamma_1 > 1$:
  $$s_{ij}^{(k+1)} = \begin{cases} \dfrac{S\left(d_{s,ij}^{(k)}, \frac{\lambda_1}{r_2}\right)}{1 - \frac{1}{r_2\gamma_1}} & \left|d_{s,ij}^{(k)}\right| \le \gamma_1\lambda_1 \\ d_{s,ij}^{(k)} & \left|d_{s,ij}^{(k)}\right| > \gamma_1\lambda_1 \end{cases}.$$

Alternating Direction Method of Multipliers(ADMM)

- $\boldsymbol{s}^{(k+1)} = \underset{\boldsymbol{s}}{\operatorname{argmin}}\, L\left(\beta^{(k+1)}, \boldsymbol{\mu}^{(k+1)}, \boldsymbol{z}^{(k+1)}, \boldsymbol{s}, \boldsymbol{w}^{(k)}, \boldsymbol{q}_1^{(k)}, \boldsymbol{q}_2^{(k)}, \boldsymbol{q}_3^{(k)}\right)$.
  The update of $\boldsymbol{s}$ depends on the choice of the loss P_{λ_1}, and it can be computed elementwisely. Let $d_{s,ij}^{(k)} = \mu_i^{(k+1)} - \mu_j^{(k+1)} + \frac{q_{2,ij}^{(k)}}{r_2}$.
  - SCAD with $r_2\,(\gamma_1 - 1) > 1$:

$$s_{ij}^{(k+1)} = \begin{cases} S\left(d_{s,ij}^{(k)}, \frac{\lambda_1}{r_2}\right) & \left|d_{s,ij}^{(k)}\right| \le \left(1 + \frac{1}{r_2}\right)\lambda_1 \\ \frac{S\left(d_{s,ij}^{(k)}, \frac{\gamma_1\lambda_1}{r_2(\gamma_1-1)}\right)}{1 - \frac{1}{r_2(\gamma_1-1)}} & \left(1 + \frac{1}{r_2}\right)\lambda_1 < \left|d_{s,ij}^{(k)}\right| \le \gamma_1\lambda_1 \\ d_{s,ij}^{(k)} & \left|d_{s,ij}^{(k)}\right| > \gamma_1\lambda_1 \end{cases}.$$

Alternating Direction Method of Multipliers(ADMM)

- $\boldsymbol{w}^{(k+1)} = \underset{\boldsymbol{w}}{\operatorname{argmin}}\, L\left(\beta^{(k+1)}, \boldsymbol{\mu}^{(k+1)}, \boldsymbol{z}^{(k+1)}, \boldsymbol{s}^{(k+1)}, \boldsymbol{w}, \boldsymbol{q}_1^{(k)}, \boldsymbol{q}_2^{(k)}, \boldsymbol{q}_3^{(k)}\right)$.
  The update of $\boldsymbol{w}$ depends on the choice of the loss P_{λ_2}, and it can be computed elementwisely. Let $d_{w,j}^{(k)} = \beta_j^{(k+1)} + \frac{q_{3,j}^{(k)}}{r_3}$.
  - LASSO:
  $$w_j^{(k+1)} = S\left(d_{w,j}^{(k)}, \frac{\lambda_2}{r_3}\right).$$
  - MCP with $r_3\gamma_2 > 1$:
  $$w_j^{(k+1)} = \begin{cases} \dfrac{S\left(d_{w,j}^{(k)}, \frac{\lambda_2}{r_3}\right)}{1 - \frac{1}{r_3\gamma_2}} & \left|d_{w,j}^{(k)}\right| \le \gamma_2\lambda_2 \\ d_{w,j}^{(k)} & \left|d_{w,j}^{(k)}\right| > \gamma_2\lambda_2 \end{cases}.$$

Alternating Direction Method of Multipliers(ADMM)

- $\boldsymbol{w}^{(k+1)} = \underset{\boldsymbol{w}}{\operatorname{argmin}}\, L\left(\beta^{(k+1)}, \boldsymbol{\mu}^{(k+1)}, \boldsymbol{z}^{(k+1)}, \boldsymbol{s}^{(k+1)}, \boldsymbol{w}, \boldsymbol{q}_1^{(k)}, \boldsymbol{q}_2^{(k)}, \boldsymbol{q}_3^{(k)}\right)$.
  The update of $\boldsymbol{w}$ depends on the choice of the loss P_{λ_2}, and it can be computed elementwisely. Let $d_{w,j}^{(k)} = \beta_j^{(k+1)} + \frac{q_{3,j}^{(k)}}{r_3}$.
  - SCAD with $r_3\,(\gamma_2 - 1) > 1$:

$$w_j^{(k+1)} = \begin{cases} S\left(d_{w,j}^{(k)}, \dfrac{\lambda_2}{r_3}\right) & \left|d_{w,j}^{(k)}\right| \le \left(1 + \dfrac{1}{r_3}\right)\lambda_2 \\ \dfrac{S\left(d_{w,j}^{(k)}, \frac{\gamma_2\lambda_2}{r_3(\gamma_2-1)}\right)}{1 - \frac{1}{r_3(\gamma_2-1)}} & \left(1 + \dfrac{1}{r_3}\right)\lambda_2 < \left|d_{w,j}^{(k)}\right| \le \gamma_2\lambda_2 \\ d_{w,j}^{(k)} & \left|d_{w,j}^{(k)}\right| > \gamma_2\lambda_2 \end{cases}.$$

Alternating Direction Method of Multipliers(ADMM)

- Update the multipliers by

$$\begin{cases} \boldsymbol{q}_1^{(k+1)} = \boldsymbol{q}_1^{(k)} + r_1 \left(\boldsymbol{y}^{(k+1)} - \boldsymbol{\mu}^{(k+1)} - \boldsymbol{X}\boldsymbol{\beta}^{(k+1)} - \boldsymbol{z}^{(k+1)} \right) \\ \boldsymbol{q}_2^{(k+1)} = \boldsymbol{q}_2^{(k)} + r_2 \left(\boldsymbol{D}\boldsymbol{\mu}^{(k+1)} - \boldsymbol{s}^{(k+1)} \right) \\ \boldsymbol{q}_3^{(k+1)} = \boldsymbol{q}_3^{(k)} + r_3 \left(\boldsymbol{\beta}^{(k+1)} - \boldsymbol{w}^{(k+1)} \right) \end{cases},$$

Convergence of the Algorithm

Conditions on the loss function

- C1 ρ is continuous on $\mathcal{R}$, and differentiable except finite many points.
- C2 ρ is convex.
- C3 ρ has unique minimal point at 0 and $\rho(0) = 0$.

Convergence of the Algorithm

Theorem

Denote the primal residual at step m by

$$\boldsymbol{r}^{(m)} = \begin{pmatrix} \boldsymbol{y} - \boldsymbol{u}^{(m)} - \boldsymbol{X}^T \beta^{(m)} - \boldsymbol{z}^{(m)} \\ \boldsymbol{D}\mu^{(m)} - \boldsymbol{s}^{(m)} \\ \beta^{(m)} - \boldsymbol{w}^{(m)} \end{pmatrix}, \tag{5}$$

and the dual residual by

$$\eta^{(m+1)} = \begin{pmatrix} r_1 \left(\boldsymbol{z}^{(m+1)} - \boldsymbol{z}^{(m)}\right) - r_2 \boldsymbol{D}^T \left(\boldsymbol{s}^{(m+1)} - \boldsymbol{s}^{(m)}\right) \\ r_1 \boldsymbol{X}^T \left(\boldsymbol{z}^{(m+1)} - \boldsymbol{z}^{(m)}\right) - r_3 \left(\boldsymbol{w}^{(m+1)} - \boldsymbol{w}^{(m)}\right) \end{pmatrix}. \tag{6}$$

These residuals of the proposed algorithm satisfy that

$$\lim_{m\to\infty} \left\| \boldsymbol{r}^{(m)} \right\|_2^2 = 0, \qquad \lim_{m\to\infty} \left\| \eta^{(m)} \right\|_2^2 = 0$$

for the Lasso, SCAD and MCP penalties if Conditions (C1)–(C3) hold.

Searching Intervals for λ

We search the grid points of a rectange plane, from $\left(\lambda_{1,max}, \lambda_{2,max}\right)$ to $\left(\lambda_{1,min}, \lambda_{2,min}\right)$, for the solution pathes.
We choose $\left(\lambda_{1,max}, \lambda_{2,max}\right)$ by

$$\lambda_{1,max} = \frac{1}{n} \left\| \boldsymbol{D} \left(\boldsymbol{D}^T \boldsymbol{D} \right)^{-1} \begin{pmatrix} \psi\left(y_1 - c\right) \\ \psi\left(y_2 - c\right) \\ \vdots \\ \psi\left(y_n - c\right) \end{pmatrix} \right\|_\infty ,$$

and

$$\lambda_{2,max} = \frac{1}{n} \left\| \sum_{i=1}^{n} \left(\psi\left(y_i - c\right) \right) \boldsymbol{x}_i \right\|_\infty ,$$

where $c = \underset{\mu}{\operatorname{argmin}} \sum_{i=1}^{n} \rho\left(y_i - \mu\right)$.

Searching Intervals for λ

A toy example

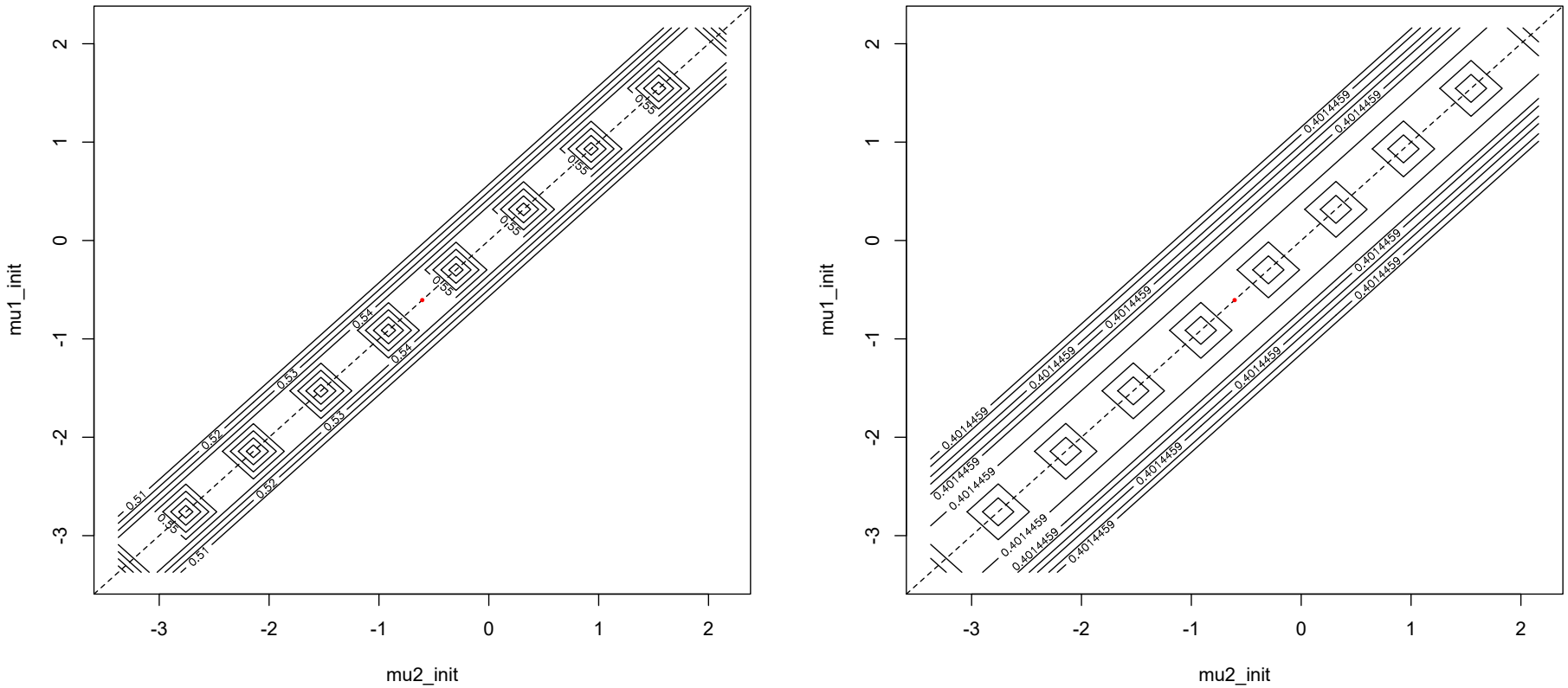

Figure: Contour figure about loss value using SCAD(left) or MCP(right) penalty of a toy example. $\rho(x) = |x|$, $n = 2$, $p = q = 0$, $\varepsilon \sim \mathrm{N}\left(0, 0.5^2\right)$, $\mu_1 = 1$, $\mu_2 = -1$. The red dot is $\mathrm{median}(\boldsymbol{y})$.

mBIC

We use the modified BIC to select the tuning parameters.

$$\mathrm{BIC}\left(\hat{\boldsymbol{\mu}}\left(\boldsymbol{\lambda}\right),\hat{\boldsymbol{\beta}}\left(\boldsymbol{\lambda}\right)\right)=\log\left(\frac{1}{n}\sum_{i=1}^{n}\rho\left(y_i-\boldsymbol{z}_i^T\hat{\mu}_i-\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}\right)\right)+\left|\hat{\mathrm{S}}_{\boldsymbol{\lambda}}\right|\phi_n, \quad (7)$$

where $\left|\hat{\mathrm{S}}_{\boldsymbol{\lambda}}\right|=\hat{K}+\hat{q}$, $\phi_n=c\frac{\log n}{n}\mathrm{loglog}\left(n+p\right)$ for some constant c.

Practical Computing Techniques

- Warm start strategy.
- A clustering method as a post-processor.
  - Rounding or hard thresholding.
  - Find the grouping result from $\{\hat{s}_{ij}\}$.
  - A simple clustering method, like k-means.
- Post-selection estimation
- Parallel/Distributed computation. We use OpenMP with C++.

Given the real subgroup structure $\boldsymbol{G}_0 = \left\{G_1, \cdots, G_{K_0}\right\}$, the original moder can be seen as

$$y_i = \boldsymbol{g}_i^T \boldsymbol{\alpha} + \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i,$$

where $\boldsymbol{g}_i$s are K_0-dimensional indicator vector, $g_{ij} = 1$ if $i \in G_j$. $\boldsymbol{\alpha}$ is the grouping effect, for all $i \in G_k$, $\mu_i = \alpha_k$.

The oracle estimator is

$$\begin{aligned}\left(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}\right) &= \left(\tilde{\boldsymbol{\alpha}}, \left(\tilde{\boldsymbol{\beta}}_A^T, \mathbf{0}_{p-q}^T\right)^T\right) \\ &= \underset{\boldsymbol{\alpha}, \boldsymbol{\beta}_A}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \rho\left(y_i - \boldsymbol{g}_i^T \boldsymbol{\alpha} - \boldsymbol{x}_{A,i}^T \boldsymbol{\beta}_A\right),\end{aligned} \tag{8}$$

and $\tilde{\mu}_i = \boldsymbol{g}_i^T \tilde{\boldsymbol{\alpha}}$.

Oracle Property

Assumptions

(A1) There exist constants M_1, C_1 and C_2 such that $\left|x_{ij}\right| \leq M_1$ for all $1 \leq i \leq n$, $1 \leq j \leq p$ and

$$C_1 \leq \lambda_{\min}\left(\frac{1}{n}\left(\boldsymbol{G} \quad \boldsymbol{X}_A\right)^T\left(\boldsymbol{G} \quad \boldsymbol{X}_A\right)\right)$$
$$\leq \lambda_{\max}\left(\frac{1}{n}\left(\boldsymbol{G} \quad \boldsymbol{X}_A\right)^T\left(\boldsymbol{G} \quad \boldsymbol{X}_A\right)\right) \leq C_2.$$

(A2) $\max\{K_0, q\} = O(n^{c_1})$ for some $0 \leq c_1 < \frac{1}{3}$.

(A3) Let $b_n = \min\left(\min_{i \neq j}\left|\alpha_{0,i} - \alpha_{0,j}\right|, \min_{1 \leq j \leq q}\left|\beta_{0,j}\right|\right)$. Then there exist constans c_2 and M_3 such that

$$2c_1 < c_2 \leq 1 \quad \text{and} \quad n^{(1-c_2)/2} b_n \geq M_3.$$

(A4) $\psi(\varepsilon_i)$ is uniformly subgaussian in a neighbourhood around 0. That means there exists a positive constant c_3 such that for all constant $c \in [-c_3, c_3]$,

$$P(|\psi(\varepsilon_i + c)| > x) \leq 2\exp\left(-c_4 x^2\right),$$

where c_4 is some positive constant.

(A5) $\mathrm{E}\psi(\varepsilon_i)$ and $\mathrm{Var}(\psi(\varepsilon_i))$ are uniformly continuous, which means $\mathrm{E}\psi(\varepsilon_i + \Delta)$ and $\mathrm{Var}\psi(\varepsilon_i + \Delta)$ is close to $\mathrm{E}\psi(\varepsilon_i)$ and $\mathrm{Var}\psi(\varepsilon_i)$.

Oracle Property

Theorem

Suppose that Assumptions (A1)–(A5) hold. Also if $\max(\lambda_1, \lambda_2) = o\left(n^{-(1-c_2)/2}\right)$, $\sqrt{q(K_0+q)} = o\left(\sqrt{n}\lambda_2\right)$, $(K_0+q)\log n = o(n\lambda_2)$, $\log(p) = o\left(n\lambda_2^2\right)$, $n\lambda_2^2 \to \infty$ *and* $\sqrt{\log n} = o\left(n\lambda_1 |G_k|_{\min}\right)$. *Then with probability approaching to 1, the oracle estimater is among the local minimizer of the penalized regression function (2) with penalties SCAD or MCP.*

Selection Consistency

Theorem

Assume the Assumptions (A1)–(A5), and (A6),(A7) in Appendix hold. Then for any sequence $\phi_n \to 0$ *satisfying* $\log(n+p)/n = o(\phi_n)$*, we have*

$$P\left(\inf_{S \neq S_0, |S| \leq K_u + q_U} \mathrm{BIC}\left(\tilde{\delta}(S)\right) > \mathrm{BIC}\left(\tilde{\delta}(S_0)\right)\right) \to 1$$

where $K_u \in (K_0, \infty)$ *and* $q_U \in (q, \infty)$ *are the upper bounds for the number of subgroups and active coefficients respectively.* $\tilde{\delta}(S)$ *denotes the unpenalized M-estimator given the model structure S. And* S_0 *represents the true model structure.*

Simulation

Basic Settings

Consider $y_i = \mu_i + \boldsymbol{x}_i^T \beta + \varepsilon_i$, $i = 1, 2, \cdots, n$, under various scenarios.

- $n \in \{200, 1000\}$, $p \in \{5, 50, 100, 500\}$.
- $q = 5$, $\beta = \left(\mathbf{1}_5^T, \mathbf{0}_{p-5}^T\right)^T$.
- $K = 2$ with $\alpha = (-1, 1)^T$ or $K = 3$ with $\alpha = (-2, 0, 2)^T$, or $K = 5$ with $\alpha = (-4, -2, 0, 2, 4)^T$.
- μ_is follow the independent multinormial distribution over α with equal probability.
- $\boldsymbol{x}_i$ follows standard normal distribution.
- $\varepsilon_i = 0.5\epsilon_i$ independently, where
  1. $\epsilon_i \sim \mathrm{N}(0, 1)$
  2. $\epsilon_i \sim t(5)$.
  3. $\epsilon_i \sim 0.95 \times \mathrm{N}(0, 1) + 0.05 \times \mathrm{N}\left(0, 10^2\right)$.

Simulation

Basic Settings

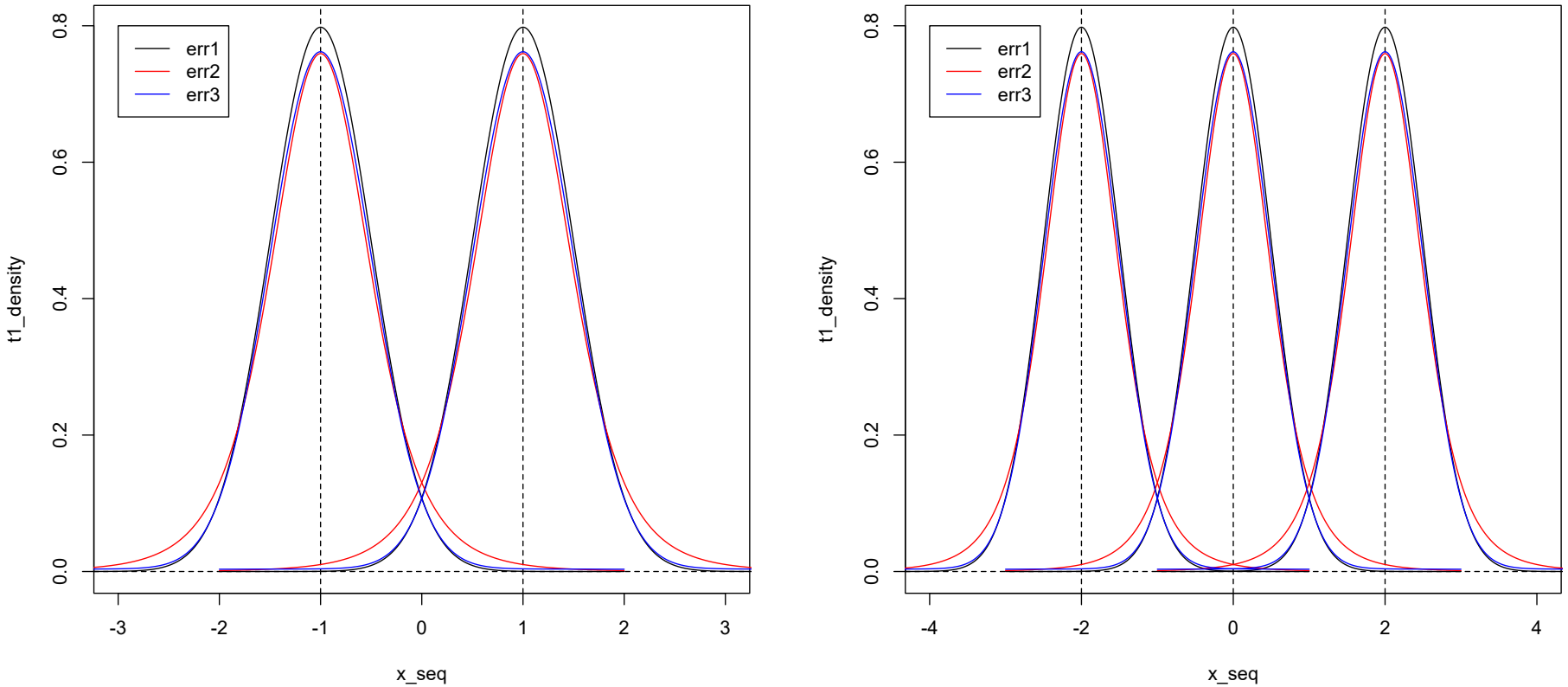

Figure: A demonstration about data points overlapping under our simulation settings.

Simulation

Basic Settings

Consider $y_i = \mu_i + \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i$, $i = 1, 2, \cdots, n$, under various scenarios.

- The proposed method with L_1, L_2 and *Huber* loss.
- Another candidate method RSI (Zhang et al., 2019, Robust subgroup identification).
- $\text{MAE}_{\boldsymbol{\mu}}$ and $\text{MAE}_{\boldsymbol{\beta}}$: the mean absolute error of $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$.
- $\bar{K}$, $\tilde{K}$, $\bar{q}$ and $\tilde{q}$: the average and median value of the estimated number of subgroups and active covariates respectively.
- *RI* (Rand, 1971, Rand Index): this index describes how close two grouping results are and is computed by

$$RI(\boldsymbol{G}_1, \boldsymbol{G}_2) = \frac{2}{n(n-1)} \times (TP + TN).$$

Benchmark for Subgroup Analysis

$n = 200$, $p = q = 5$, $K = 2$

Method	MAE_μ	MAE_β	$\bar{K}$	$\tilde{K}$	$\bar{RI}$	$\tilde{RI}$
L_1	0.116	0.040	2.000	2	0.941	0.942
sd	0.040	0.014	0.000		0.024	0.000
L_2	0.111	0.038	2.090	2	0.940	0.951
sd	0.059	0.016	0.514		0.042	0.000
Huber	0.105	0.037	2.000	2	0.944	0.951
sd	0.039	0.015	0.000		0.024	0.000
RSI	0.135	0.071	2.020	2	0.939	0.951
sd	0.154	0.082	0.141		0.057	0.000
L_1^{1000}	0.123	0.042	2.050	2	0.936	0.942
sd	0.051	0.018	0.261		0.034	0.000

Table: Benchmark for subgroup analysis. $n = 200$, $p = q = 5$, $K = 2$. Error type: Case 1.

Benchmark for Subgroup Analysis

$n = 200$, $p = q = 5$, $K = 3$

Method	MAE_μ	MAE_β	$\bar{K}$	$\tilde{K}$	$\bar{RI}$	$\tilde{RI}$
L_1	0.529	0.081	2.370	2	0.819	0.743
sd	0.307	0.042	0.485		0.103	0.000
L_2	0.302	0.063	3.830	3	0.878	0.917
sd	0.301	0.032	1.583		0.139	0.000
Huber	0.224	0.060	3.120	3	0.916	0.932
sd	0.162	0.032	0.573		0.055	0.000
RSI	0.252	0.089	2.870	3	0.916	0.944
sd	0.224	0.035	0.367		0.076	0.000
L_1^{1000}	0.238	0.068	3.310	3	0.906	0.919
sd	0.126	0.037	0.787		0.051	0.000

Table: Benchmark for subgroup analysis. $n = 200$, $p = q = 5$, $K = 3$. Error type: Case 1.

Benchmark for Subgroup Analysis

$n = 200$, $p = q = 5$, $K = 2$

Method	MAE_μ	MAE_β	$\bar{K}$	$\tilde{K}$	$\bar{RI}$	$\tilde{RI}$
L_1	0.172	0.045	2.000	2	0.892	0.896
sd	0.043	0.018	0.000		0.030	0.000
L_2	0.461	0.057	2.070	2	0.753	0.861
sd	0.375	0.026	1.380		0.183	0.000
Huber	0.167	0.044	2.000	2	0.893	0.896
sd	0.045	0.018	0.000		0.031	0.000
RSI	0.177	0.067	2.000	2	0.892	0.896
sd	0.053	0.037	0.000		0.034	0.000
L_1^{1000}	0.186	0.047	2.130	2	0.882	0.896
sd	0.070	0.022	0.506		0.049	0.000

Table: Benchmark for subgroup analysis. $n = 200$, $p = q = 5$, $K = 2$. Error type: Case 2.

Benchmark for Subgroup Analysis

$n = 200$, $p = q = 5$, $K = 3$

Method	MAE_μ	MAE_β	$\bar{K}$	$\tilde{K}$	$\bar{RI}$	$\tilde{RI}$
L_1	0.718	0.098	2.090	2	0.750	0.735
sd	0.170	0.040	0.288		0.055	0.000
L_2	0.746	0.084	3.330	3	0.671	0.814
sd	0.505	0.035	2.216		0.248	0.000
Huber	0.298	0.069	3.060	3	0.880	0.890
sd	0.168	0.032	0.509		0.051	0.000
RSI	0.355	0.102	2.820	3	0.869	0.899
sd	0.269	0.086	0.411		0.074	0.000
L_1^{1000}	0.299	0.073	3.330	3	0.877	0.886
sd	0.113	0.035	0.792		0.041	0.000

Table: Benchmark for subgroup analysis. $n = 200$, $p = q = 5$, $K = 3$. Error type: Case 2.

Subgroup Analysis and Variable Selection

$n = 200$, $p = 50$, $q = 5$, $K = 2$

Method	MAE_μ	MAE_β	$\bar{K}$	$\tilde{K}$	$\bar{q}$	$\tilde{q}$	$\bar{RI}$	$\tilde{RI}$
L_1	0.255	0.008	2.040	2	4.980	5	0.829	0.835
sd	0.077	0.006	0.197		0.141		0.051	0.000
L_2	1.017	0.034	1.370	1	3.630	5	0.501	0.501
sd	0.040	0.041	0.544		2.205		0.001	0.000
Huber	0.906	0.051	2.380	2	2.870	5	0.700	0.612
sd	0.682	0.043	0.582		2.295		0.151	0.000
RSI	0.674	0.098	5.240	5.5	50	50	0.633	0.604
sd	0.184	0.024	2.920		0.000		0.080	0.000

Table: Subgroup analysis and variable selection. $n = 200$, $p = 50$, $q = 5$, $K = 2$. Error type: Case 3.

Subgroup Analysis and Variable Selection

$n = 200$, $p = 50$, $q = 50$, $K = 3$

Method	MAE_μ	MAE_β	$\bar{K}$	$\tilde{K}$	$\bar{q}$	$\tilde{q}$	$\bar{RI}$	$\tilde{RI}$
L_1	0.828	0.019	2.000	2	4.780	5	0.712	0.718
sd	0.086	0.015	0.000		0.645		0.025	0.000
L_2	1.425	0.100	1.000	1	0.000	0	0.333	0.333
sd	0.031	0.000	0.000		0.000		0.000	0.000
Huber	0.813	0.023	2.200	2	5.250	5	0.709	0.709
sd	0.250	0.016	0.512		1.403		0.104	0.000
RSI	1.055	0.153	4.440	3	50	50	0.662	0.668
sd	0.141	0.038	2.702		0.000		0.033	0.000

Table: Subgroup analysis and variable selection. $n = 200$, $p = 50$, $q = 5$, $K = 3$. Error type: Case 3.

Subgroup Analysis and Variable Selection

$n = 200$, $p = 100$, $q = 5$, $K = 2$

Method	MAE_μ	MAE_β	$\bar{K}$	$\tilde{K}$	$\bar{q}$	$\tilde{q}$	$\bar{RI}$	$\tilde{RI}$
L_1	0.236	0.003	2.000	2	5.000	5	0.838	0.844
sd	0.069	0.002	0.000		0.000		0.047	0.000
L_2	0.993	0.050	1.000	1	0.000	0	0.501	0.501
sd	0.014	0.000	0.000		0.000		0.000	0.000
Huber	0.384	0.006	1.940	2	5.430	5	0.773	0.840
sd	0.298	0.006	0.489		1.130		0.139	0.000
RSI	0.891	0.121	6.630	7	100.000	100	0.537	0.532
sd	0.239	0.028	2.820		0.000		0.029	0.000

Table: Subgroup analysis and variable selection. $n = 200$, $p = 100$, $q = 5$, $K = 2$. Error type: Case 2.

Subgroup Analysis and Variable Selection

$n = 200$, $p = 100$, $q = 5$, $K = 3$

Method	MAE_μ	MAE_β	$\bar{K}$	$\tilde{K}$	$\bar{q}$	$\tilde{q}$	$\bar{RI}$	$\tilde{RI}$
L_1	0.818	0.009	2.000	2	4.740	5	0.719	0.726
sd	0.082	0.007	0.000		0.705		0.025	0.000
L_2	1.424	0.048	1.020	1	0.250	0	0.333	0.333
sd	0.041	0.004	0.141		0.575		0.005	0.000
Huber	0.736	0.011	2.300	2	4.900	5	0.733	0.724
sd	0.273	0.008	0.522		1.078		0.101	0.000
RSI	1.357	0.192	6.560	7	100.000	100	0.608	0.613
sd	0.305	0.050	2.607		0.000		0.040	0.000

Table: Subgroup analysis and variable selection. $n = 200$, $p = 100$, $q = 5$, $K = 3$. Error type: Case 2.

High Dimensional Dataset

$n = 200$, $p = 500$, $q = 5$

K	Method	MAE_μ	MAE_β	$\bar{K}$	$\tilde{K}$	$\bar{q}$	$\tilde{q}$	$\bar{RI}$	$\tilde{RI}$
2	L_1	0.389	0.002	2.000	2	4.360	5	0.782	0.819
	sd	0.304	0.003	0.000		1.375		0.105	0.000
	L_2	0.995	0.010	1.010	1	0.000	0	0.501	0.501
	sd	0.031	0.000	0.100		0.000		0.002	0.000
	Huber	0.585	0.003	2.140	2	4.290	5	0.732	0.773
	sd	0.462	0.003	0.450		1.924		0.131	0.000
3	L_1	0.938	0.004	2.000	2	3.610	4	0.677	0.680
	sd	0.140	0.003	0.000		1.456		0.049	0.000
	L_2	1.425	0.010	1.000	1	0.000	0	0.333	0.333
	sd	0.036	0.000	0.000		0.000		0.000	0.000
	Huber	0.925	0.005	2.300	2	3.410	4	0.687	0.686
	sd	0.236	0.003	0.461		1.718		0.065	0.000

Table: Subgroup analysis and variable selection in high dimension. $n = 200$, $p = 500$, $q = 5$. Error type: Case 2.

High Dimensional Dataset

$n = 200$, $p = 500$, $q = 5$

K	Method	MAE_μ	MAE_β	$\bar{K}$	$\tilde{K}$	$\bar{q}$	$\tilde{q}$	$\bar{RI}$	$\tilde{RI}$
2	L_1	0.458	0.003	2.000	2	4.170	5	0.747	0.792
	sd	0.316	0.003	0.000		1.393		0.113	0.000
	L_2	0.993	0.010	1.000	1	0.000	0	0.501	0.501
	sd	0.015	0.000	0.000		0.000		0.000	0.000
	Huber	0.932	0.005	1.990	2	3.110	3	0.623	0.593
	sd	0.398	0.003	0.502		1.769		0.103	0.000
3	L_1	0.941	0.004	2.000	2	3.560	4	0.672	0.675
	sd	0.131	0.002	0.000		1.351		0.045	0.000
	L_2	1.427	0.010	1.000	1	0.000	0	0.333	0.333
	sd	0.035	0.000	0.000		0.000		0.000	0.000
	Huber	1.000	0.006	2.160	2	2.870	3	0.657	0.645
	sd	0.164	0.003	0.368		1.548		0.047	0.000

Table: Subgroup analysis and variable selection in high dimension. $n = 200$, $p = 500$, $q = 5$. Error type: Case 3.

High Dimensional Dataset

$n = 1000$, $p = q = 5$, $K = 2$

Method	MAE_μ	MAE_β	$\bar{K}$	$\tilde{K}$	$\bar{RI}$	$\tilde{RI}$
L_1	0.072	0.018	2.000	2	0.951	0.953
sd	0.015	0.006	0.000		0.010	0.000
L_2	0.284	0.025	6.690	7	0.739	0.738
sd	0.037	0.008	0.506		0.037	0.000
Huber	0.070	0.015	2.000	2	0.951	0.951
sd	0.015	0.005	0.000		0.010	0.000
RSI_{DAC}	0.718	0.310	6.520	7	0.684	0.676
sd	0.387	0.173	2.772		0.096	0.000

Table: Subgroup analysis and variable selection in high dimension, $n = 1000$, $p = q = 5$, $K = 2$. Error type: Case 1.

High Dimensional Dataset

$n = 1000$, $p = q = 5$, $K = 5$

Method	MAE_μ	MAE_β	$\bar{K}$	$\tilde{K}$	$\bar{RI}$	$\tilde{RI}$
L_1	0.542	0.058	3.940	5	0.857	0.958
sd	0.548	0.046	1.469		0.143	0.000
L_2	0.185	0.041	5.920	6	0.952	0.952
sd	0.062	0.017	0.929		0.014	0.000
Huber	0.473	0.049	4.100	5	0.875	0.960
sd	0.531	0.035	1.382		0.136	0.000
RSI_{DAC}	1.229	0.249	6.200	7	0.739	0.736
sd	0.287	0.162	1.820		0.031	0.000

Table: Subgroup analysis and variable selection in high dimension, $n = 1000$, $p = q = 5$, $K = 5$. Error type: Case 1.

Remarks

- It is conceptually straightforward to extend to other models, such as logistic regression.
- Speed up ADMM.
- Theoretical results for selection consistency.
- It requires careful design and study of the algorithm for extra large datasets.

Reference

Rand, W. M. (1971), “Objective Criteria for the Evaluation of Clustering Methods,” *Journal of the American Statistical Association*, 66, 846–850.

Zhang, Y., Wang, H. J., and Zhu, Z. (2019), “ROBUST SUBGROUP IDENTIFICATION,” *Statistica Sinica*.