

Hierarchical Heterogeneous Analysis of High-Dimensional Data

Yan Ren

2021103739

2023 年 1 月 12 日

目录

| | | |
|----------|--|----------|
| 1 | Model | 3 |
| 1.1 | Basic Model | 3 |
| 1.2 | Object Function | 3 |
| 1.3 | Penalty Function | 4 |
| 2 | Computation | 4 |
| 2.1 | Transformation | 4 |
| 2.2 | MCP | 6 |
| 3 | Model Preparation | 7 |
| 3.1 | Version 1 - P for β ADMM | 7 |
| 3.2 | Version 2 - P for β | 9 |
| 3.3 | Version 3 - P for both β, α | 12 |
| 3.4 | Version 4 - P for all | 12 |
| 3.5 | Notes | 15 |

1 Model

- $X = (X_1, \dots, X_n)^T \in \mathbb{R}^{n \times p}$, type I features (main G) for all observations, where $X_i \in \mathbb{R}^p$;
- $Z = (Z_1, \dots, Z_n)^T \in \mathbb{R}^{n \times q}$, type II features (main E) for all observations, where $Z_i \in \mathbb{R}^q$;
- $y = (y_1, \dots, y_n)^T \in \mathbb{R}^n$, the response vector.

1.1 Basic Model

简化后的异质性模型为：

$$y_i = X_i^T \beta'_i + Z_i^T \alpha'_i + \epsilon_i \quad (1.1)$$

其中 $\epsilon_i \sim (iid)N(0, \sigma^2)$

假设所有观测大组可根据基因（X）进行划分，大组中根据环境/图片特征变量（Z）进行进一步的划分。假设有 K 个亚组，则相同亚组中的样有相同参数。

1.2 Object Function

基于高斯混合分布，重参数化后最小化目标函数为（负对数似然+惩罚）：

$$\begin{aligned} \mathcal{L}(\beta, \alpha, \pi | X, y) &= -\frac{1}{n} \sum_{i=1}^n \log p_Y(y) + \text{pen}(\beta', \alpha') \\ &= -\frac{1}{n} \sum_{i=1}^n \log \left(\sum_{k=1}^K \pi_k f_k(y_i | X_i, Z_i) \right) + \text{pen}(\beta', \alpha') \\ &= -\frac{1}{n} \sum_{i=1}^n \log \left\{ \sum_{k=1}^K \pi_k \frac{1}{\sqrt{2\pi}\sigma_k} \exp \left(-\frac{1}{2\sigma_k^2} (y_i - X_i^T \beta'_k - Z_i^T \alpha'_k)^2 \right) \right\} + \text{pen}(\beta', \alpha') \\ &= -\frac{1}{n} \sum_{i=1}^n \log \left\{ \sum_{k=1}^K \pi_k \frac{\rho_k}{\sqrt{2\pi}} \exp \left(-\frac{1}{2} (y_i \rho_k - X_i^T \beta_k - Z_i^T \alpha_k)^2 \right) \right\} + \text{pen}(\beta, \alpha) \end{aligned} \quad (1.2)$$

其中重参数化为 $\beta_k = \beta'_k / \sigma_k, \rho_k = 1/\sigma_k$ ，惩罚函数为

$$\begin{aligned}
\text{pen}(\beta, \alpha) &= \sum_{k=1}^K \sum_{j=1}^p \text{pen}(|\beta_{kj}|, \lambda_1) + \sum_{k=1}^K \sum_{s=1}^q \text{pen}(|\alpha_{ks}|, \lambda_1) \\
&+ \sum_{k < k'} \text{pen} \left(\sqrt{\|\beta_k - \beta_{k'}\|_2^2 + \|\alpha_k - \alpha_{k'}\|_2^2}, \lambda_2 \right) \\
&+ \sum_{k < k'} \text{pen}(\|\beta_k - \beta_{k'}\|_2, \lambda_3)
\end{aligned} \tag{1.3}$$

第一项惩罚保证基因相关系数的稀疏性；第二，第三项惩罚共同作用以致模型呈现层级结构（Hierarchy）。

1.3 Penalty Function

MCP 惩罚函数

$$\text{pen}_{(\lambda, a)}(\beta) = \begin{cases} \lambda|\beta| - \frac{\beta^2}{2a}, & |\beta| \leq a\lambda \\ \frac{a}{2}\lambda^2, & |\beta| > a\lambda \end{cases} \tag{1.4}$$

MCP 惩罚函数对参数求偏导得

$$\text{pen}'_{(\lambda, a)}(\beta) = \begin{cases} \text{sgn}(\beta) \left(\lambda - \frac{|\beta|}{a} \right) & |\beta| \leq a\lambda \\ 0 & \text{otherwise} \end{cases} \tag{1.5}$$

超参数 a 常取为大于 1 的数。

2 Computation

2.1 Transformation

采用 EM 算法。E 步更新 π ，M 步更新参数 β, α 。引入变量 C_i 用于指示每个样本属于哪个类别，有后验概率如 2.1 所示

$$q_{C_i}(c_i) = \frac{\pi_{c_i} f_{c_i}(y_i | X_i, Z_i)}{\sum_{k=1}^K \pi_k f_k(y_i | X_i, Z_i)} \tag{2.1}$$

目标函数 1.2 中对数似然部分可以放缩如下

$$\begin{aligned}
\log p_Y(y) &= \log \sum_{k=1}^K p_{Y,C}(y, c) \\
&= \log E_{q_C} \left(\frac{p_{Y,C}(y, c)}{q_C(c)} \right) \\
&\geq E_{q_C} \left(\frac{p_{Y,C}(y, c)}{q_C(c)} \right) \\
&= E_{q_C} (\log p_{Y,C}(y, c)) - E_{q_C} (\log q_C(c))
\end{aligned} \tag{2.2}$$

综上，目标函数可以放缩（根据 Gaussian mixture models and the EM algorithm）为

$$\begin{aligned}
\mathcal{L}(\beta, \alpha, \pi | X, y) &= - \sum_{i=1}^n \log p_Y(y) + \text{pen}(\beta, \alpha) \\
&\leq -E_{q_C} (\log p_{Y,C}(y, c)) + E_{q_C} (\log q_C(c)) + \text{pen}(\beta, \alpha)
\end{aligned} \tag{2.3}$$

在 M 步更新参数 β, α 式 2.3 中 $E_{q_C} (\log q_C(c))$ 与参数无关。而 $E_{q_C} (\log p_{Y,C}(y, c))$ 可以进一步变形

$$\begin{aligned}
E_{q_C} \log p_{Y,C}(y, c) &= E_{q_C} \log (p_C(c) p_{Y|C}(y)) \\
&= E_{q_C} \log \left(\prod_{i=1}^n p_{C_i}(c_i) p_{Y_i|C_i}(y_i) \right) \\
&= E_{q_C} \log \left(\prod_{i=1}^n \prod_{k=1}^K \pi_{c_i} f_{Y_i|C_i}(y_i) \right)^{I(C_i=k)} \\
&= E_{q_C} \sum_{i=1}^n \sum_{k=1}^K I(C_i = k) (\log \pi_{c_i} + \log f_{Y_i|C_i}(y_i)) \\
&= \sum_{i=1}^n \sum_{k=1}^K E_{q_C} I(C_i = k) (\log \pi_{c_i} + \log f_{Y_i|C_i}(y_i)) \\
&= \sum_{i=1}^n \sum_{k=1}^K q_{C_i}(k) (\log \pi_{c_i} + \log f_{Y_i|C_i}(y_i)) \\
&= \sum_{i=1}^n \sum_{k=1}^K q_{C_i}(k) \left(\log \pi_{c_i} + \log \rho_k - \frac{1}{2} \log 2\pi - \frac{1}{2} (y_i \rho_k - X_i^T \beta_k - Z_i^T \alpha_k)^2 \right) \\
&= \sum_{i=1}^n \sum_{k=1}^K q_{C_i}(k) \left(C - \frac{1}{2} (y_i \rho_k - X_i^T \beta_k - Z_i^T \alpha_k)^2 \right)
\end{aligned} \tag{2.4}$$

其中 C 为与更新参数无关的常数项。此时最小化目标为原目标函数的上界，表达式为

$$\begin{aligned}
\mathcal{L}(\beta, \alpha | \pi, X, y) &= -\frac{1}{n} \sum_{i=1}^n \log p_Y(y) + \text{pen}(\beta, \alpha) \\
&\leq -\frac{1}{n} E_{q_C} (\log p_{Y,C}(y, c)) + \text{pen}(\beta, \alpha) \\
&= -\frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K q_{C_i}(k) \left(C - \frac{1}{2} (y_i \rho_k - X_i^T \beta_k - Z_i^T \alpha_k)^2 \right) \\
&\quad + \sum_{k=1}^K \text{pen}(\|\beta_k\|_2, \lambda_1) + \sum_{k=1}^K \text{pen}(\|\alpha_k\|_2, \lambda_1) \\
&\quad + \sum_{k < k'} \text{pen} \left(\sqrt{\|\beta_k - \beta_{k'}\|_2^2 + \|\alpha_k - \alpha_{k'}\|_2^2}, \lambda_2 \right) \\
&\quad + \sum_{k < k'} \text{pen}(\|\beta_k - \beta_{k'}\|_2, \lambda_3)
\end{aligned} \tag{2.5}$$

2.2 MCP

β, z 均为 p 维向量，现假设目标函数有

$$\mathcal{L}(x; \lambda) = \frac{\rho}{2} \|x - z\|_2^2 + \text{pen}(\|x\|_2; \lambda) \tag{2.6}$$

则 x 对应解为

$$x = \begin{cases} \frac{S(z, \frac{\lambda}{\rho})}{1 - \frac{1}{a\rho}} & \text{if } |z| \leq a\lambda \\ z & \text{if } |z| > a\lambda \end{cases} \tag{2.7}$$

其中 $S(.,.)$ 为 soft thresholding operator，定义如下

$$S(x, t) = \text{sign}(x)(|x| - t)_+ = x \left(1 - \frac{t}{|x|} \right)_+ \tag{2.8}$$

故 x 解也可写为

$$x = \begin{cases} \frac{z(1 - \frac{\lambda}{|z|\rho})_+}{1 - \frac{1}{a\rho}} & \text{if } |z| \leq a\lambda \\ z & \text{if } |z| > a\lambda \end{cases} \tag{2.9}$$

3 Model Preparation

3.1 Version 1 - P for β ADMM

为了尝试推导以及检验代码正确性，先尝试构建较为简单的模型进行实验。考虑以下目标函数

$$\begin{aligned}\mathcal{L}^{(1)}(\beta, \alpha) &= -\frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K q_{C_i}(k) \left(C - \frac{1}{2} (y_i \rho_k - X_i^T \beta_k - Z_i^T \alpha_k)^2 \right) + \sum_{k=1}^K \sum_{j=1}^p \text{pen}(|\beta_{kj}|, \lambda_1) \\ &= \frac{1}{2n} \sum_{i=1}^n \sum_{k=1}^K q_{C_i}(k) (y_i \rho_k - X_i^T \beta_k - Z_i^T \alpha_k)^2 + \sum_{k=1}^K \sum_{j=1}^p \text{pen}(|\beta_{kj}|, \lambda_1) + C\end{aligned}\quad (3.1)$$

C 表示与更新参数无关的项，后面均省略。对于类别 k 目标函数为

$$\mathcal{L}_k^{(1)}(\beta_k, \alpha_k) = \frac{1}{2n} \sum_{i=1}^n q_{C_i}(k) (y_i \rho_k - X_i^T \beta_k - Z_i^T \alpha_k)^2 + \sum_{j=1}^p \text{pen}(|\beta_{kj}|, \lambda_1) \quad (3.2)$$

对应增广拉格朗日函数为

$$\begin{aligned}\mathcal{L}_{k,\rho}^{(1)}(\beta_k, \alpha_k, \tau) &= \frac{1}{2n} \sum_{i=1}^n q_{C_i}(k) (y_i \rho_k - X_i^T \beta_k - Z_i^T \alpha_k)^2 + \sum_{j=1}^p \text{pen}(|\beta_{kj}|, \lambda_1) \\ &\quad + \sum_{j=1}^p \text{pen}(|\theta_{kj}|, \lambda_1) + \tau^T (\beta_k - \theta_k) + \frac{\rho}{2} \|\beta_k - \theta_k\|_2^2\end{aligned}\quad (3.3)$$

τ 为对应的拉格朗日乘子（向量）， $\rho > 0$ 为拉格朗日参数。给定 $(\beta_k^{(t)}, \alpha_k^{(t)}, \tau^{(t)})$ 则第 $t+1$ 步更新写为

$$\beta_k^{(t+1)} \in \arg \min_{\beta \in R^p} \mathcal{L}_{k,\rho}^{(1)}(\beta, \theta^{(t)}, \tau^{(t)}) \quad (3.4)$$

$$\theta_k^{(t+1)} \in \arg \min_{\theta \in R^p} \mathcal{L}_{k,\rho}^{(1)}(\beta^{(t+1)}, \theta, \tau^{(t)}) \quad (3.5)$$

$$\tau^{(t+1)} = \tau^{(t)} + \rho (\beta^{(t+1)} - \theta^{(t+1)}) \quad (3.6)$$

Update β_k 式 3.3 中与 β_k 相关的项为

$$\begin{aligned}\mathcal{L}_{k,\rho}^{(1)}(\beta_k) &= \frac{1}{2n} \sum_{i=1}^n q_{C_i}(k) (y_i \rho_k - X_i^T \beta_k - Z_i^T \alpha_k)^2 + \sum_{j=1}^p \text{pen}(|\beta_{kj}|, \lambda_1) \\ &\quad + \tau^T (\beta_k - \theta_k) + \frac{\rho}{2} \|\beta_k - \theta_k\|_2^2\end{aligned}\quad (3.7)$$

对 β_k 求偏导, 得

$$\beta_k = (n^{-1}X^TW_kX + \rho I_p)^{-1} (n^{-1}X^TW'_ky^{\beta_k} + \rho\theta_k - \tau) \quad (3.8)$$

其中 $y^{(\beta_k)} = y\rho_k - Z\alpha_k$, $W_k = \rho_k \text{diag}(q_{C_1}(k), \dots, q_{C_n}(k))$, I_p 为 p 维单位矩阵。

Update θ_k 式 3.3 中与 θ_k 相关的项为

$$\begin{aligned} \mathcal{L}_{k,\rho}^{(1)}(\beta_k) &= \sum_{j=1}^p \text{pen}(\theta_{kj}, \lambda_1) + \tau^T(\beta_k - \theta_k) + \frac{\rho}{2} \|\beta_k - \theta_k\|_2^2 \\ &= \frac{\rho}{2} \|\theta_k - z_k\|_2^2 + \sum_{j=1}^p \text{pen}(\theta_{kj}, \lambda_1) \end{aligned} \quad (3.9)$$

其中 $z_k = \beta_k + \frac{\tau}{\rho}$.

故 $\theta_k = (\theta_{k1}, \dots, \theta_{kp})$ 的解为

$$\theta_{kj} = \begin{cases} \frac{S(z_{kj}, \frac{\lambda}{\rho})}{1 - \frac{1}{a\rho}} & \text{if } |z_{kj}| \leq a\lambda \\ z_{kj} & \text{if } |z_{kj}| > a\lambda \end{cases} \quad (3.10)$$

Update ρ_k 惩罚项不涉及 ρ_k , 更新 ρ_k 只需要考虑拟合项。式 3.3 中未完全表示出 ρ_k 相关的项, 原待最小化的目标函数中与 ρ_k 有关的项应为 3.11. (注意此处 ρ 表示拉格朗日参数, $\rho_k = 1/\sigma_k$ 为高斯分布中的参数)

$$\mathcal{L}_{k,\rho}^{(1)}(\rho_k) = -\frac{1}{n} \sum_{i=1}^n q_{C_i}(k) \left(\log \rho_k - \frac{1}{2} (y_i \rho_k - X_i^T \beta_k - Z_i^T \alpha_k)^2 \right) \quad (3.11)$$

对 ρ_k 求导得

$$\frac{\partial \mathcal{L}_{k,\rho}^{(1)}(\rho_k)}{\partial \rho_k} = -\frac{1}{n} \sum_{i=1}^n q_{C_i}(k) (1/\rho_k - (y_i \rho_k - X_i^T \beta_k - Z_i^T \alpha_k) y_i) \quad (3.12)$$

$\frac{\partial \mathcal{L}_{k,\rho}^{(1)}(\rho_k)}{\partial \rho_k} = 0$ 时, 有下式成立。

$$\sum_{i=1}^n q_{C_i}(k) (-\rho_k^2 y_i^2 + \rho_k (X_i^T \beta_k + Z_i^T \alpha_k) y_i + 1) = 0 \quad (3.13)$$

$$-y^T W_k y \rho_k^2 + (X \beta_k + Z \alpha_k)^T W_k y \rho_k + I_n^T W_k I_n = 0$$

这是一个一元二次方程求根问题, 因为 ρ_k 应为正值, 取两个根中的正值, 解为

$$\rho_k = \frac{-B - \sqrt{B^2 - 4AC}}{2A} \quad (3.14)$$

其中 $A = -y^T W_k y$, $B = (X\beta_k + Z\alpha_k)^T W_k y$, $C = 1_n^T W_k 1_n$.

若给定每个样本类别，等价于再各个类别内部使用 ADMM 方法进行迭代求解系数；若未知样本类别，使用 EM 算法进行估计，算法流程见 1

Algorithm 1: Version 1 with sparse penalty for β (ADMM)

Input : 数据 y, X , 超参数 $n, p, K, a, \lambda, \rho$

Output: $\beta_k, k = 1, \dots, K$ 的估计值

```

1  初始化  $q_{C_i}(k) = \frac{1}{K} + N(0, 0.5^2), k = 1, \dots, K, i = 1, \dots, n$  标准化;
2  repeat
3      for  $k$  in  $1 : K$  do
4           $W_k = \text{diag}(q_{C_1}(k), \dots, q_{C_n}(k));$ 
5          更新  $\beta_k^{(iter)} = (n^{-1}X^T W_k X + \rho I_p)^{-1} (n^{-1}X^T W_k' y^{\beta_k} + \rho\theta_k - \tau);$ 
6          更新  $\theta_k^{(iter)} = \begin{cases} \frac{S(z_k, \frac{\lambda}{\rho})}{1 - \frac{1}{a\rho}} & \text{if } \|z_k\|_2 \leq a\lambda \\ z & \text{if } \|z_k\|_2 > a\lambda \end{cases};$ 
7          更新  $\tau^{(iter)} = \tau^{(iter-1)} + \rho(\beta_k^{(iter)} - \theta_k^{(iter)}).$ 
8          更新  $\rho_k^{(iter)} = \frac{-B - \sqrt{B^2 - 4AC}}{2A}.$ 
9          where  $A = -y^T W_k y; B = (X\beta_k + Z\alpha_k)^T W_k y; C = 1_n^T W_k 1_n.$ 
10     end for
11     计算基于  $\beta_k^{(iter)}$  得到的后验概率  $q_C$ 
12      $q_{C_i}^{(iter)}(k) = \frac{\pi_k^{(iter)} f_k(y_i; \beta_k^{(iter)}) / \rho_k^{(iter)}}{\sum_{k'=1}^K \pi_{k'}^{(iter)} f_{k'}(y_i; \beta_{k'}^{(iter)}) / \rho_{k'}^{(iter)}}, k = 1, \dots, K$ 
13     更新  $\pi_k^{(iter)} = \frac{1}{n} \sum_{i=1}^n q_{C_i}^{(iter)}(k)$ 
14 until 收敛 ( $\frac{\|\beta^{(iter)} - \beta^{(iter-1)}\|}{\|\beta^{(iter)}\|}$  小于临界值);
```

实验结果大体如表 14 所示。

3.2 Version 2 - P for β

依旧对目标函数

$$\mathcal{L}^{(2)}(\beta, \alpha) = \frac{1}{2n} \sum_{i=1}^n \sum_{k=1}^K q_{C_i}(k) (y_i \rho_k - X_i^T \beta_k - Z_i^T \alpha_k)^2 + \sum_{k=1}^K \sum_{j=1}^p \text{pen}(|\beta_{kj}|, \lambda_1) \quad (3.15)$$

| p | 非零个数 | epsilon | flexmix | 自己迭代 |
|----|------|---------------|---------|--------------|
| 8 | 3 | 0 | 收敛到真值 | 收敛到真值 |
| 8 | 3 | $N(0, 0.5^2)$ | 收敛到接近真值 | 收敛到接近真值 |
| 40 | 3 | 0 | 全0估计 | 距离真值有差异，详情见表 |
| 40 | 3 | $N(0, 0.5^2)$ | 全0估计 | 收敛到接近真值 |

不同于上一节，现在直接对于 β_{kj} 求偏导进行求解，而不用 ADMM 算法框架进行问题的转化。目标函数中与第 k 类中的参数相关的部分写为 $\mathcal{L}_k^{(2)}$ ，见 3.16，其中 $y'_{ik} = \rho_k y_i - X_{i(-j)}^T \beta_{k(-j)} - Z_i^T \alpha_k$ ， $y'_k = (y'_{1k}, \dots, y'_{nk})^T \in \mathbb{R}^{n \times 1}$ ， $X_{i(-j)}$ 表示 X_i 中除第 j 列外的其他特征， $X_{(-j)} = (X_{1(-j)}, \dots, X_{n(-j)})^T \in \mathbb{R}^{n \times (p-1)}$ ， $\tilde{X}_j = (X_{1j}, \dots, X_{nj})^T \in \mathbb{R}^{n \times 1}$ ， $W_k = \text{diag}(q_{C_1}(k), \dots, q_{C_n}(k)) \in \mathbb{R}^{n \times n}$

$$\begin{aligned}
\mathcal{L}_k^{(2)}(\beta_k, \alpha_k) &= \frac{1}{2n} \sum_{i=1}^n q_{C_i}(k) (y_i \rho_k - X_i^T \beta_k - Z_i^T \alpha_k)^2 + \sum_{j=1}^p \text{pen}(|\beta_{kj}|, \lambda_1) \\
&= \frac{1}{2n} \sum_{i=1}^n q_{C_i}(k) (y'_{ik} - X_{ij} \beta_{kj})^2 + \sum_{j=1}^p \text{pen}(|\beta_{kj}|, \lambda_1) \\
&= \frac{1}{2n} \sum_{i=1}^n q_{C_i}(k) (y'_{ik}^2 - 2X_{ij} y'_{ik} \beta_{kj} + X_{ij}^2 \beta_{kj}^2) + \sum_{j=1}^p \text{pen}(|\beta_{kj}|, \lambda_1) \\
&= \frac{1}{2n} \left(\tilde{X}_j^T W_k \tilde{X}_j \beta_{kj}^2 - 2\tilde{X}_j^T W_k y'_k \beta_{kj} + y_k'^T W_k y'_k \right) + \sum_{j=1}^p \text{pen}(|\beta_{kj}|, \lambda_1) \\
&= \frac{\tilde{X}_j^T W_k \tilde{X}_j}{2n} \left(\beta_{kj} - \frac{\tilde{X}_j^T W_k y'_k}{\tilde{X}_j^T W_k \tilde{X}_j} \right)^2 + \text{pen}(|\beta_{kj}|, \lambda_1) + C
\end{aligned} \tag{3.16}$$

其中 C 与 β_{kj} 无关，记 $u_{kj} = \tilde{X}_j^T W_k y'_k$ ， $l_{kj} = \tilde{X}_j^T W_k \tilde{X}_j$ ，使得上式最小的 β_{kj} 可表示为

$$\beta_{kj} = \begin{cases} \frac{S\left(\frac{u_{kj}}{l_{kj}}, \frac{n\lambda}{l_{kj}}\right)}{1 - \frac{n}{al_{kj}}} = \frac{S(u_{kj}, n\lambda)}{l_{kj} - \frac{n}{a}} & \text{if } \left| \frac{u_{kj}}{l_{kj}} \right| \leq a\lambda \\ \frac{u_{kj}}{l_{kj}} & \text{if } \left| \frac{u_{kj}}{l_{kj}} \right| > a\lambda \end{cases} \tag{3.17}$$

算法流程如 2

实验中，若 $n=400$, $p=8$ β_k 中非零项长度为 3（实验中设置 $\alpha_k = 0$ ，上面所有推导可忽略包含 α 项，同时包含 β, α 且不全为 0 的模型在第三部分实验），模型可以稳定收敛到

Algorithm 2: Version 2 with sparse penalty for β

Input : 数据 y, X , 超参数 $n, p, K, a, \lambda, \rho$

Output: $\beta_k, k = 1, \dots, K$ 的估计值

1 初始化 $q_{C_i}(k) = \frac{1}{K} + N(0, 0.5^2), k = 1, \dots, K, i = 1, \dots, n$ 标准化, 或用 flexmix 包结果初始化;

2 repeat

3 **for** k in $1 : K$ **do**

4 $W_k = \text{diag}(q_{C_1}(k), \dots, q_{C_n}(k));$

5 **for** j in $1 : p$ **do**

6 $u_{kj} = (\rho_k y - X_{(-j)} \beta_{k(-j)} - Z \alpha_k)^T W_k \tilde{X}_j$

7 $l_{kj} = \tilde{X}_j^T W_k \tilde{X}_j$

8 更新 $\beta_{kj}^{(iter)} = \begin{cases} \frac{u_{kj}}{l_{kj}}, & \frac{u_{kj}}{l_{kj}} > a\lambda \\ S(\frac{u_{kj}}{l_{kj}}, n\lambda), & \frac{u_{kj}}{l_{kj}} \leq a\lambda \end{cases}$

9 **end for**

10 更新 $\rho_k^{(iter)} = \frac{-B - \sqrt{B^2 - 4AC}}{2A}$.

11 where $A = -y^T W_k y; B = (X \beta_k + Z \alpha_k)^T W_k y; C = 1_n^T W_k 1_n$.

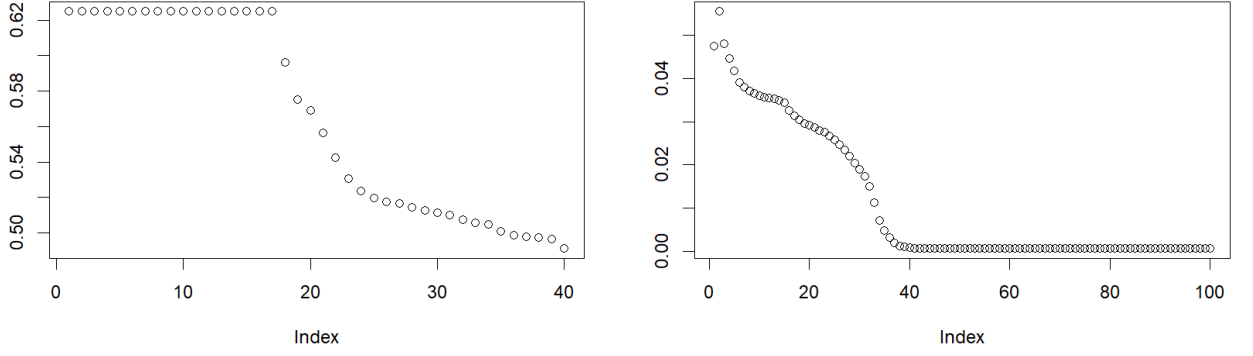
12 **end for**

13 计算基于 $\beta_k^{(iter)}$ 得到的后验概率 q_C

14 $q_{C_i}^{(iter)}(k) = \frac{\pi_k^{(iter)} f_k(y_i; \beta_k^{(iter)} / \rho_k^{(iter)})}{\sum_{k'=1}^K \pi_{k'}^{(iter)} f_{k'}(y_i; \beta_{k'}^{(iter)} / \rho_{k'}^{(iter)})}, k = 1, \dots, K$

15 更新 $\pi_k^{(iter)} = \frac{1}{n} \sum_{i=1}^n q_{C_i}^{(iter)}(k)$

16 **until** 收敛 ($\frac{\|\beta^{(iter)} - \beta^{(iter-1)}\|}{\|\beta^{(iter)}\|}$ 小于临界值);



(a) 全 0 初值

(b) flexmix 初值

图 1: 基于两种初值的参数收敛曲线（仅有 β, α 稀疏惩罚）

真值，当增加误差项或者当 p 增大时，会出现 3,-3 项估计准确而 1,-1 被压为零，或调整参数后出现部分零项估计非零，整个向量偏离真值的情况。在困难情况下使用 flexmix 包作为储值可以收敛。

3.3 Version 3 - P for both β, α

此时目标函数表示为

$$\begin{aligned} \mathcal{L}^{(3)}(\beta, \alpha) = & \frac{1}{2n} \sum_{i=1}^n \sum_{k=1}^K q_{C_i}(k) (y_i \rho_k - X_i^T \beta_k - Z_i^T \alpha_k)^2 \\ & + \sum_{k=1}^K \sum_{j=1}^p \text{pen}(|\beta_{kj}|, \lambda_1) + \sum_{k=1}^K \sum_{s=1}^q \text{pen}(|\alpha_{ks}|, \lambda_1) \end{aligned} \quad (3.18)$$

当 β_k, α_k 不全为 0 向量时，同时更新 β_k, α_k 的计算流程与上面两种情况相似，只需要注意当更新 β_{kj} 时，对应 $y'_i = \rho_k y_i - X_{i(-j)}^T \beta_{k(-j)} - Z_i^T \alpha_k$, $j = 1, 2, \dots, p$ ；而更新 α_{ks} 时，对应 $y'_i = \rho_k y_i - Z_{i(-s)}^T \alpha_{k(-s)} - X_i^T \beta_k$, $s = 1, 2, \dots, q$ ；

在 p, q 较大时更新难度增大，图 1 对比了 $p = q = 20$ 时全 0 初值及基于 flexmix 包结果继续迭代的参数距离真值距离随迭代次数变化曲线。

已经验证，该框架下 $\alpha = 0$ 时结果与 Version 2 完全相同。

3.4 Version 4 - P for all

该部分在 Version 3（仅有 β, α 的稀疏惩罚）基础上加上类别压缩惩罚函数，目标函数

补全为最终的

$$\begin{aligned}
\mathcal{L}^{(4)}(\beta, \alpha) = & \frac{1}{2n} \sum_{i=1}^n \sum_{k=1}^K q_{C_i}(k) (y_i \rho_k - X_i^T \beta_k - Z_i^T \alpha_k)^2 \\
& + \sum_{k=1}^K \sum_{j=1}^p \text{pen}(|\beta_{kj}|, \lambda_1) + \sum_{k=1}^K \sum_{s=1}^q \text{pen}(|\alpha_{ks}|, \lambda_1) \\
& + \sum_{k < k'} \text{pen} \left(\sqrt{\|\beta_k - \beta_{k'}\|_2^2 + \|\alpha_k - \alpha_{k'}\|_2^2}, \lambda_2 \right) \\
& + \sum_{k < k'} \text{pen}(\|\beta_k - \beta_{k'}\|_2, \lambda_3)
\end{aligned} \tag{3.19}$$

对应的增广拉格朗日函数为

$$\begin{aligned}
\mathcal{L}^{(4)}(\beta, \alpha, u, v, \xi, \zeta) = & \frac{1}{2n} \sum_{i=1}^n \sum_{k=1}^K q_{C_i}(k) (y_i \rho_k - X_i^T \beta_k - Z_i^T \alpha_k)^2 \\
& + \sum_{k=1}^K \sum_{j=1}^p \text{pen}(|\beta_{kj}|, \lambda_1) + \sum_{k=1}^K \sum_{s=1}^q \text{pen}(|\alpha_{ks}|, \lambda_1) \\
& + \sum_{k < k'} \text{pen} \left(\sqrt{\|v_{kk'}\|_2^2 + \|w_{kk'}\|_2^2}, \lambda_2 \right) \\
& + \sum_{k < k'} \text{pen}(\|v_{kk'}\|_2, \lambda_3) \\
& + \sum_{k < k'} \xi_{kk'}^T (\beta_k - \beta_{k'} - v_{kk'}) + \frac{\tau}{2} \sum_{k < k'} \|\beta_k - \beta_{k'} - v_{kk'}\|_2^2 \\
& + \sum_{k < k'} \zeta_{kk'}^T (\alpha_k - \alpha_{k'} - w_{kk'}) + \frac{\tau}{2} \sum_{k < k'} \|\alpha_k - \alpha_{k'} - w_{kk'}\|_2^2
\end{aligned} \tag{3.20}$$

记 $H_1 = \varepsilon \otimes I_p \in \mathbb{R}^{\frac{K(K-1)}{2} p \times Kp}$, $H_2 = \varepsilon \otimes I_q \in \mathbb{R}^{\frac{K(K-1)}{2} q \times Kq}$, $\varepsilon = \{(e_k - e_{k'}), k < k'\}^T$. $e_k \in \mathbb{R}^{K \times 1}$ 为独热向量, 除了第 k 个元素为 1 外, 其他元素都为 0; I_p, I_q 分别表示维度为 $p \times p, q \times q$ 的单位对角阵。

另外, 记

$$u_{kk'} = (v_{kk'}^T, w_{kk'}^T)^T \tag{3.21}$$

则增广拉格朗日函数可以改写为

$$\begin{aligned}
\mathcal{L}^{(4)}(\beta, \alpha, u, v, \xi, \zeta) = & \frac{1}{2n} \sum_{i=1}^n \sum_{k=1}^K q_{C_i}(k) (y_i \rho_k - X_i^T \beta_k - Z_i^T \alpha_k)^2 \\
& + \sum_{k=1}^K \sum_{j=1}^p \text{pen}(|\beta_{kj}|, \lambda_1) + \sum_{k=1}^K \sum_{s=1}^q \text{pen}(|\alpha_{ks}|, \lambda_1) \\
& + \sum_{k < k'} \text{pen}(\|u_{kk'}\|_2, \lambda_2) + \sum_{k < k'} \text{pen}(\|v_{kk'}\|_2, \lambda_3) \\
& + \frac{\tau}{2} \|H_1 \beta - v + \frac{1}{\tau} \xi\|_2^2 + \frac{\tau}{2} \|H_2 \alpha - w + \frac{1}{\tau} \zeta\|_2^2
\end{aligned} \tag{3.22}$$

注意维度 $\beta \in \mathbb{R}^{Kp \times 1}, \alpha \in \mathbb{R}^{Kq \times 1}, u, \xi \in \mathbb{R}^{\frac{K(K-1)p}{2} \times 1}, w, \zeta \in \mathbb{R}^{\frac{K(K-1)q}{2} \times 1}$ 。

Update β_{kj} 式子 3.22 中与 β_{kj} 相关的项写为

$$\begin{aligned}
\mathcal{L}^{(4)}(\beta_{kj}) = & \frac{1}{2n} \sum_{i=1}^n q_{C_i}(k) (y_i \rho_k - X_i^T \beta_k - Z_i^T \alpha_k)^2 + \text{pen}(|\beta_{kj}|, \lambda_1) + \frac{\tau}{2} \|H_1 \beta - v + \frac{1}{\tau} \xi\|_2^2 \\
= & \frac{1}{2n} (l_{kj} \beta_{kj}^2 - u_{kj} \beta_{kj}) + \frac{\tau}{2} (l'_{kj} \beta_{kj}^2 - u'_{kj} \beta_{kj}) + \text{pen}(|\beta_{kj}|, \lambda_1) + C(\beta_{kj}) \\
= & \frac{1}{2} (L_{kj} \beta_{kj}^2 - U_{kj} \beta_{kj}) + \text{pen}(|\beta_{kj}|, \lambda_1) + C(\beta_{kj}) \\
= & \frac{L_{kj}}{2} \left(\beta_{kj}^2 - \frac{U_{kj}}{L_{kj}} \right)^2 + \text{pen}(|\beta_{kj}|, \lambda_1) + C(\beta_{kj})
\end{aligned} \tag{3.23}$$

其中 $C(\beta_{kj})$ 表示与 β_{kj} 无关的项，记号补充说明如下：

$$\begin{cases}
l_{kj} = \tilde{X}_j^T W_k \tilde{X}_j \\
u_{kj} = \tilde{X}_j^T W_k (\rho_k y - X_{(-j)} \beta_{k(-j)} - Z \alpha_k) \\
l'_{kj} = (H_1^T H_1)_{kj, kj} \\
u'_{kj} = \left(H_1^T (v - \frac{1}{\tau} \xi) \right)_{kj} - \beta_{(-kj)}^T (H_1^T H_1)_{kj, (-kj)} \\
L_{kj} = \frac{1}{n} l_{kj} + \tau l'_{kj} \\
U_{kj} = \frac{1}{n} u_{kj} + \tau u'_{kj}
\end{cases} \tag{3.24}$$

$W_k = \text{diag}(q_{C_1}(k), \dots, q_{C_n}(k))$ ，解得

$$\beta_{kj} = \begin{cases} \frac{S\left(\frac{U_{kj}}{L_{kj}}, \frac{\lambda}{L_{kj}}\right)}{1 - \frac{1}{aL_{kj}}} = \frac{S(U_{kj}, \lambda)}{L_{kj} - \frac{1}{a}} & \text{if } |\frac{U_{kj}}{L_{kj}}| \leq a\lambda \\ \frac{U_{kj}}{L_{kj}} & \text{if } |\frac{U_{kj}}{L_{kj}}| > a\lambda \end{cases} \tag{3.25}$$

Update α_{ks} 更新过程完全类似 β_{kj} 的更新。

Update $v_{kk'}, w_{kk'}$ 式子 3.22 中与 $v_{kk'}, w_{kk'}$ 相关的项写为

$$\begin{aligned} \mathcal{L}^{(4)}(v_{kk'}, w_{kk'}) &= \sum_{k < k'} \text{pen}(\|u_{kk'}\|_2, \lambda_2) + \sum_{k < k'} \text{pen}(\|v_{kk'}\|_2, \lambda_3) \\ &\quad + \frac{\tau}{2} \|H_1 \beta - v + \frac{1}{\tau} \xi\|_2^2 + \frac{\tau}{2} \|H_2 \alpha - w + \frac{1}{\tau} \zeta\|_2^2 \end{aligned} \quad (3.26)$$

记

$$\begin{cases} \overline{v_{kk'}} = \beta_k - \beta_{k'} + \frac{1}{\tau} \xi_{kk'} \\ \overline{w_{kk'}} = \alpha_k - \alpha_{k'} + \frac{1}{\tau} \zeta_{kk'} \end{cases} \quad (3.27)$$

$$\begin{cases} u_{kk'} = (v_{kk'}^T, w_{kk'}^T)^T \\ \overline{u_{kk'}} = (\overline{v_{kk'}}^T, \overline{w_{kk'}}^T)^T \end{cases} \quad (3.28)$$

更新分为四种情况

1. 若 $\|\overline{u_{kk'}}\|_2 > a\lambda_2$ 且 $\|\overline{v_{kk'}}\|_2 > a\lambda_3$, 则 $w_{kk'} = \overline{w_{kk'}}$, $v_{kk'} = \overline{v_{kk'}}$;
2. 若 $\|\overline{u_{kk'}}\|_2 \leq a\lambda_2$ 且 $\frac{(1 - \lambda_2/(\tau\|\overline{u_{kk'}}\|_2))_+}{1 - 1/(a\tau)} \|\overline{v_{kk'}}\|_2 > a\lambda_3$,
 则 $w_{kk'} = \frac{(1 - \lambda_2/(\tau\|\overline{u_{kk'}}\|_2))_+}{1 - 1/(a\tau)} \overline{w_{kk'}}$, $v_{kk'} = \frac{(1 - \lambda_2/(\tau\|\overline{u_{kk'}}\|_2))_+}{1 - 1/(a\tau)} \overline{v_{kk'}}$;
3. 若 $\|\overline{w_{kk'}}\|_2^2 + \left(\frac{(1 - \lambda_3/(\tau\|\overline{v_{kk'}}\|_2))_+}{1 - 1/(a\tau)} \right)^2 \|\overline{v_{kk'}}\|_2^2 > (a\lambda_2)^2$ 且 $\|\overline{v_{kk'}}\|_2 \leq a\lambda_3$,
 则 $w_{kk'} = \overline{w_{kk'}}$, $v_{kk'} = \frac{(1 - \lambda_3/(\tau\|\overline{v_{kk'}}\|_2))_+}{1 - 1/(a\tau)} \overline{v_{kk'}}$;
4. 否则, $w_{kk'} = \frac{\overline{w_{kk'}}}{1 + \frac{\text{pen}'(u_{kk'}, \lambda_2)}{\tau\|u_{kk'}\|_2}}$, $v_{kk'} = \frac{\overline{v_{kk'}}}{1 + \frac{\text{pen}'(u_{kk'}, \lambda_2)}{\tau\|u_{kk'}\|_2} + \frac{\text{pen}'(v_{kk'}, \lambda_3)}{\tau\|v_{kk'}\|_2}}$

3.5 Notes

- $\lambda_2 = 0, \lambda_3 > 0$ 只对 β 进行压组操作, 和 α 无关。由于两个惩罚共用相同的 a , 若 $\lambda_2 > 0$ 则必有 $\lambda_2 > \lambda_3$

References

- [1] Yueyong Shi, Yuling Jiao, Yongxiu Cao, and Yan Yan Liu. An alternating direction method of multipliers for mcp-penalized regression with high-dimensional data. *Acta Mathematica Sinica, English Series*, 34:1892–1906, 2018.