1 Within group variable selection

Let's focus on only one group for now. Our model is as follows:

$$P(Y_i = k|X_i, \beta, \gamma) = \frac{\exp\Big(X_i \cdot \beta_k \cdot 1_{k \neq K}\Big)}{1 + \sum_{\ell=1}^{K-1} \exp\Big(X_i \cdot \beta_\ell\Big)}, \quad k = 1, \ldots, K. \tag{1}$$

We also have

$$\begin{aligned} \gamma_j \sim & Bernoulli(1 - \rho_j) \\ \beta_{jk}|\gamma_j \sim & Normal(0, \tau_j^2), \end{aligned}$$

where $\tau_j^2 = \sigma_{0j}^{2(1-\gamma_j)} \sigma_{1j}^{2\gamma_j}$. We can sample from the posterior $P(\gamma, \beta|Y_i = k)$:

Step 1: Sample from $P(\gamma|\beta, Y_1 = k_1, \ldots, Y_n = k_n) \propto P(Y_1 = k_1, \ldots, Y_n = k_n|\beta, \gamma) \cdot P(\beta|\gamma) \cdot P(\gamma)$. We can directly sample from

$$P(\gamma|\beta, Y_1 = k_1, \ldots, Y_n = k_n) \propto \exp\Big(\sum_{\ell=1}^{K-1} \sum_{j=1}^{p} -\frac{\beta_{j\ell}^2}{2\tau_j^2}\Big) \cdot \prod_{j=1}^{p} (1 - \rho_j)^{\gamma_j} \rho_j^{(1-\gamma_j)} \prod_{\ell=1}^{K-1} \tau_j^{-1}$$

to produce B samples $\hat{\gamma}^{(1)}, \ldots, \hat{\gamma}^{(B)}$. This expression can be written as

$$\log P(\gamma|\beta, Y_1 = k_1, \ldots, Y_n = k_n) = (1 - K) \sum_{j=1}^{p} \log \tau_j - \frac{1}{2} \sum_{\ell=1}^{K-1} \sum_{j=1}^{p} \frac{\beta_{j\ell}^2}{\tau_j^2} + \sum_{j=1}^{p} \Big(\gamma_j \log(1 - \rho_j) + (1 - \gamma_j) \log \rho_j\Big).$$

To sample γ, we can sample one γ_j at a time while holding the remaining γ_{-j} fixed; the log likelihood needed here is

$$\log P(\gamma_j|\beta_j, Y_1 = k_1, \ldots, Y_n = k_n) = (1 - K) \log \tau_j - \frac{1}{2\tau_j^2} \sum_{\ell=1}^{K-1} \beta_{j\ell}^2 + \gamma_j \log(1 - \rho_j) + (1 - \gamma_j) \log \rho_j.$$

Step 2: Sample from $P(\beta|\gamma, Y_1 = k_1, \ldots, Y_n = k_n) \propto P(Y_1 = k_1, \ldots, Y_n = k_n|\beta, \gamma) \cdot P(\beta|\gamma)$. We can use MCMC to sample from

$$P(\beta|\gamma, Y_1 = k_1, \ldots, Y_n = k_n) \propto \exp\Big(\sum_{\ell=1}^{K-1} \sum_{j=1}^{p} -\frac{\beta_{j\ell}^2}{2\tau_j^2}\Big) \prod_{i=1}^{n} \frac{\exp\Big(X_i \cdot \beta_{k_i} \cdot 1_{k_i \neq K}\Big)}{1 + \sum_{\ell=1}^{K-1} \exp\Big(X_i \cdot \beta_\ell\Big)}$$

to produce B samples $\hat{\beta}^{(1)}, \ldots, \hat{\beta}^{(B)}$. Use $\hat{\beta}^{(B)}$ in the next step. This expression can be written as

$$\log P(\beta|\gamma, Y_1 = k_1, \ldots, Y_n = k_n) = -\frac{1}{2}\sum_{\ell=1}^{K-1}\sum_{j=1}^{p}\frac{\beta_{j\ell}^2}{\tau_j^2} + \sum_{i=1}^{n} X_i \cdot \beta_{k_i} \cdot 1_{k_i \neq K} - \sum_{i=1}^{n}\log\Big(1 + \sum_{\ell=1}^{K-1}\exp(X_i \cdot \beta_\ell)\Big).$$

To sample β, we can sample one β_j vector at a time while holding the remaining β_{-j} fixed; define $U(\beta_j)$ as

$$\begin{aligned} U(\beta_j) &= -\log P(\beta_j|\gamma_j, Y_1 = k_1, \ldots, Y_n = k_n) \\ &= \frac{1}{2\tau_j^2}\sum_{\ell=1}^{K-1}\beta_{j\ell}^2 - \sum_{i=1}^{n} X_i \cdot \beta_{k_i} \cdot 1_{k_i \neq K} + \sum_{i=1}^{n}\log\Big(1 + \sum_{\ell=1}^{K-1}\exp(X_i \cdot \beta_\ell)\Big). \end{aligned}$$

Then

$$\frac{\partial}{\partial \beta_j}U(\beta_j) = \frac{1}{\tau_j^2}\beta_j + \sum_{i=1}^{n}\left(\frac{X_{ij}\exp(X_i \cdot \beta)}{1 + \sum_{\ell=1}^{K-1}\exp(X_i\beta_\ell)}\right) - X_{Y,j},$$

where

$$\begin{aligned} \exp(X_i \cdot \beta) &= [\exp(X_i \cdot \beta_1), \ldots, \exp(X_i \cdot \beta_{K-1})] \\ X_{Y,j} &= [\sum_{i \in A_1} X_{ij}, \ldots, \sum_{i \in A_{K-1}} X_{ij}] \end{aligned}$$

for

$$A_\ell = \{i; Y_i = \ell, i = 1, \ldots, n\}, \quad \ell = 1, \ldots, K-1.$$

We can use Langevin dynamics to sample β_j:

$$\beta_j^{(t+1)} = \beta_j^{(t)} - \frac{\epsilon^2}{2}\frac{\partial}{\partial \beta_j}U(\beta_j^{(t)}) + \epsilon \cdot \boldsymbol{Z}^{(t)},$$

where $\epsilon > 0$ is the step size and $\boldsymbol{Z}^{(t)} \sim N(0, I_{K-1})$.

Repeat steps 1 and 2 a large number of times T until we get a large number of samples

$(\hat{\beta}^{(1)}, \hat{\gamma}^{(1)}), \ldots, (\hat{\beta}^{(T)}, \hat{\gamma}^{(T)})$. Use some criterion like the median probability criterion to determine which γ_j are active.