1 Bayesian bi-level variable selection with overlapping groups

The setup of this problem is as follows. Let $X \in \mathbb{R}^{n \times p}$ be our predictor variables, and Y be our categorical response: $Y_i \in \{1, \ldots, K\}$ for $i = 1, \ldots, n$. We have G groups of possibly overlapping variables, with $p_1, \ldots p_G$ variables,where $\sum_{\ell=1}^{G} p_\ell \geq p$. Our goal is to select the important groups, as well as the important variables within those groups. To solve this problem, we define a set of indicator variables at both the *group* and *within group* levels (hence the name *bi-level selection*), and sample from the posterior of those indicators.

2 Within group variable selection

Let's focus on only one group for now, where p is the number of variables in the group. Our model is as follows:

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

We also have for $j = 1, \ldots, p$ and $k = 1, \ldots, K-1$,

$$\begin{aligned}
\rho_j &\sim Beta(\alpha_1, \alpha_2) \\
\sigma_{0j}^2 &\sim IG(\lambda_0, \nu_0) \\
\sigma_{1j}^2 &\sim IG(\lambda_1, \nu_1) \\
\gamma_j | \rho_j &\sim Bernoulli(1 - \rho_j) \\
\beta_{jk} | \gamma_j, \sigma_{0j}, \sigma_{1j} &\sim Normal(0, \tau_j^2),
\end{aligned}$$

where $\tau_j^2 = (1 - \gamma_j)\sigma_{0j}^2 + \gamma_j \sigma_{1j}^2$. Whenever $\gamma_j = 1$ and variable j is deemed important, we have $\beta_{jk} \sim N(0, \sigma_{1j}^2)$. Whenever $\gamma_j = 0$ and thus variable j is deemed *not* important, we have $\beta_{jk} \sim N(0, \sigma_{0j}^2)$. Thus, we want $\sigma_{0j}^2 << \sigma_{1j}^2$ with σ_{0j}^2 very small.

We can sample from the posterior $P(\gamma, \beta, \sigma_0^2, \sigma_1^2, \rho | Y_i = k, X)$:

Step 1: Sample from

$$\begin{aligned}
P(\gamma | \beta, \rho, \sigma_0^2, \sigma_1^2, Y, X) &\propto P(Y | \beta, \sigma_0^2, \sigma_1^2, X) \cdot P(\beta | \gamma, \sigma_0^2, \sigma_1^2) \cdot P(\gamma | \rho) \\
&\propto P(\beta | \gamma, \sigma_0^2, \sigma_1^2) \cdot P(\gamma | \rho).
\end{aligned}$$

Since the γ_j are binary, we can directly sample from

$$P(\gamma | \beta, \rho, \sigma_0^2, \sigma_1^2, Y, X) \propto \exp\Big(\sum_{\ell=1}^{K-1} \sum_{j=1}^{p} -\frac{\beta_{j\ell}^2}{2\tau_j^2} \Big) \cdot \prod_{j=1}^{p} (1 - \rho_j)^{\gamma_j} \rho_j^{(1-\gamma_j)} \prod_{\ell=1}^{K-1} \tau_j^{-1}$$

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

$$\log P(\gamma|\beta,\rho,\sigma_0^2,\sigma_1^2,Y,X) = (1-K)\sum_{j=1}^{p}\log\tau_j - \frac{1}{2}\sum_{\ell=1}^{K-1}\sum_{j=1}^{p}\frac{\beta_{j\ell}^2}{\tau_j^2} + \sum_{j=1}^{p}\Big(\gamma_j\log(1-\rho_j) + (1-\gamma_j)\log\rho_j\Big).$$

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

$$\log P(\gamma_j|\beta_j,\rho_j,\sigma_{0j}^2,\sigma_{1j}^2,Y,X) = (1-K)\log\tau_j - \frac{1}{2\tau_j^2}\sum_{\ell=1}^{K-1}\beta_{j\ell}^2 + \gamma_j\log(1-\rho_j) + (1-\gamma_j)\log\rho_j.$$

Step 2: Sample from

$$P(\beta|\gamma,\sigma_0^2,\sigma_1^2,\rho,Y_1=k_1,\ldots,Y_n=k_n,X) \propto P(Y_1=k_1,\ldots,Y_n=k_n|\beta,X)\cdot P(\beta|\gamma,\sigma_0^2,\sigma_1^2).$$

We can use MCMC to sample from

$$P(\beta|\gamma,\sigma_0^2,\sigma_1^2,\rho,Y_1=k_1,\ldots,Y_n=k_n,X) \propto \exp\Big(\sum_{\ell=1}^{K-1}\sum_{j=1}^{p}-\frac{\beta_{j\ell}^2}{2\tau_j^2}\Big)\prod_{i=1}^{n}\frac{\exp\Big(X_i\cdot\beta_{k_i}\cdot 1_{k_i\neq K}\Big)}{1+\sum_{\ell=1}^{K-1}\exp\Big(X_i\cdot\beta_\ell\Big)}$$

to produce B samples $\hat\beta^{(1)},\ldots,\hat\beta^{(B)}$. The log of this expression can be written as

$$\log P(\beta|\gamma,\sigma_0^2,\sigma_1^2,\rho,Y_1=k_1,\ldots,Y_n=k_n,X) = -\frac{1}{2}\sum_{\ell=1}^{K-1}\sum_{j=1}^{p}\frac{\beta_{j\ell}^2}{\tau_j^2} + \sum_{i=1}^{n}X_i\cdot\beta_{k_i}\cdot 1_{k_i\neq K} - \sum_{i=1}^{n}\log\Big(1+\sum_{\ell=1}^{K-1}\exp(X_i\cdot\beta_\ell)\Big).$$

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

$$\begin{aligned}U(\beta_j) &= -\log P(\beta_j|\gamma_j,\sigma_{0j}^2,\sigma_{1j}^2,\rho_j,Y_1=k_1,\ldots,Y_n=k_n,X)\\ &= \frac{1}{2\tau_j^2}\sum_{\ell=1}^{K-1}\beta_{j\ell}^2 - \sum_{i=1}^{n}X_i\cdot\beta_{k_i}\cdot 1_{k_i\neq K} + \sum_{i=1}^{n}\log\Big(1+\sum_{\ell=1}^{K-1}\exp(X_i\cdot\beta_\ell)\Big).\end{aligned}$$

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

$$\exp(X_i \cdot \beta) = [\exp(X_i \cdot \beta_1), \ldots, \exp(X_i \cdot \beta_{K-1})]$$

$$X_{Y,j} = [\sum_{i \in A_1} X_{ij}, \ldots, \sum_{i \in A_{K-1}} X_{ij}]$$

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 $\boldsymbol{Z}^{(t)} \sim N(0, I_{K-1})$.

Step 3: For $j = 1, \ldots, p$ sample $P(\sigma_{0j}^2, |\gamma, \beta, \sigma_1^2, \rho, Y, X)$ and $P(\sigma_1^2, |\gamma, \beta, \sigma_{0j}^2, \rho, Y, X)$:

$$\sigma_{0j}^2 \sim IG\left(\lambda_0 + (1-\gamma_j)\frac{K}{2},\ \nu_0 + \frac{(1-\gamma_j)}{2}\sum_{k=1}^{K} \beta_{jk}^2\right)$$

$$\sigma_{1j}^2 \sim IG\left(\lambda_1 + \gamma_j \frac{K}{2},\ \nu_1 + \frac{\gamma_j}{2}\sum_{k=1}^{K} \beta_{jk}^2\right).$$

Step 4: For $j = 1, \ldots, p$, sample $P(\rho_j | \gamma, \beta, \sigma_{0j}^2, \sigma_{1j}^2, Y, X)$:

$$\rho_j \sim Beta\Big(\alpha_1 + 1 - \gamma_j,\ \alpha_2 + \gamma_j\Big).$$

Repeat steps $1-4$ 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.