Bayesian Spike and Slab with Random Intercept

This work builds upon Naqvi et al. Fast Laplace Approximation for Sparse Bayesian Spike and Slab Models, by adding in a random effect to produce more robust posterior inclusion probabilities.

When the main goal is variable selection, we treat the variance components (fixed effects (σ^2), random effects (δ^2)) as hyperparameters and estimate them by empirical bayes.

The model implemented can be described as follows:

$$Y|\beta,\mu,\sigma^2 \sim N(X\beta + Z\mu, \sigma^2)$$

$$\beta|\tau_1^2,\tau_0^2 \sim \pi N(0,\tau_1^2) + (1-\pi)N(0,\tau_0^2)$$

$$\mu|\delta^2 \sim N(0,\delta^2)$$

$$P(\beta|Y,\mu,\sigma^2,\tau_1^2,\tau_0^2,\delta^2) = \frac{P(Y|\beta,\mu,\sigma^2)P(\beta|\tau_1^2,\tau_0^2)P(\mu|\delta^2)}{P(Y|\mu,\sigma^2,\tau_1^2,\tau_0^2)P(\mu|\delta^2)}$$

Where $Y \in \mathbb{R}^{N\times 1}$ is a response vector; $X \in \mathbb{R}^{N\times p}$ is a matrix of predictors; $Z \in \mathbb{R}^{N\times m}$ is a one-hot-encoded group ID matrix; μ the random effect vector; β the (sparse) regression weights; σ^2 the fixed effects variance; τ_0^2, τ_1^2 the spike and slab variances, respectively; π the spike and slab mixture proportion; δ^2 the random effects variance component.

Step 1 – Marginalize the random effects

First, we integrate out the random effects, μ

$$P(Y,\beta) = \int P(Y|\beta,\mu,\sigma^2)P(\beta|\tau_1^2,\tau_0^2)P(\mu|\delta^2)d\mu$$

$$= P(\beta|\tau_1^2,\tau_0^2)\int P(Y|\beta,\mu,\sigma^2)P(\mu|\delta^2)d\mu$$

This is a well-known integral (Gaussian convolution):

$$= P(\beta|\tau_1^2,\tau_0^2)N(Y|X\beta,\Sigma); \quad ,\Sigma = \delta^2 ZZ^T + \sigma^2 I$$

So the joint probability of Y and β is:

$$\log P(Y,\beta) = -\frac{N}{2}\log 2\pi - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(Y - X\beta)^T\Sigma^{-1}(Y - X\beta) + \sum_{j=1}^{p}\log\left(\pi N\left(\beta_j\middle|0,\tau_1^2\right) + (1-\pi)N\left(\beta_j\middle|0,\tau_0^2\right)\right)$$

Step 2 – Maximum *a posteriori* estimation

We set σ^2, δ^2 to a constant. Maximize the posterior with respect to β. To do this, we need the gradients of the posterior with respect to β:

Gradient of the likelihood:

$$\nabla_\beta \log P(Y|\beta,\sigma^2) = X^T\Sigma^{-1}Y - X^T\Sigma^{-1}X\,\beta$$

Gradient of the prior:

$$\nabla_{\beta_j}\log P\left(\beta_j\middle|\tau_1^2,\tau_0^2\right) = -\beta_j\frac{\left(\frac{\pi}{\tau_1^2}N\left(\beta_j\middle|0,\tau_1^2\right) + \frac{(1-\pi)}{\tau_0^2}N\left(\beta_j\middle|0,\tau_0^2\right)\right)}{\pi N\left(\beta_j\middle|0,\tau_1^2\right) + (1-\pi)N\left(\beta_j\middle|0,\tau_0^2\right)}$$

Step 3 – Hessian calculation

Compute the Hessian at $\hat{\beta}$

Hessian of the likelihood:

$$\nabla_\beta^2 \log P(Y|\beta,\sigma^2) = -X^T\Sigma^{-1}X$$

Hessian of the prior:

$$\nabla_\beta^2\log P(\beta|\tau_1^2,\tau_0^2) = diag\left(\frac{\nabla_\beta^2 P(\beta|\tau_1^2,\tau_0^2)}{\mathrm{P}(\beta|\tau_1^2,\tau_0^2)} - \left[\frac{\nabla_\beta P(\beta)}{P(\beta)}\right]^2\right)$$

Where we have:

$$\nabla_{\beta_j} P(\beta) = -\beta_j \left(\frac{\pi}{\tau_1^2} N(\beta_j | 0, \tau_1^2) + \frac{(1-\pi)}{\tau_0^2} N(\beta_j | 0, \tau_0^2) \right)$$

$$\nabla_{\beta_j}^2 P(\beta | \tau_1^2, \tau_0^2) = \pi \left(\frac{\beta_j^2}{\tau_1^4} - \frac{1}{\tau_1^2} \right) N_1 + (1-\pi) \left(\frac{\beta_j^2}{\tau_0^4} - \frac{1}{\tau_0^2} \right) N_0$$

$$H(\beta) = \nabla_\beta^2 \log P(Y | \beta, \sigma^2) + \nabla_\beta^2 \log P(\beta | \tau_1^2, \tau_0^2)$$

Step 4 – Estimation of variance components (Type II likelihood)

Use the Laplace approximation to integrate β out. Maximize the marginal posterior with respect to σ^2, δ^2.

The Laplace approximation is found by Taylor expansion of the log exponent of the posterior, leading to:

$$P(\beta, Y | \sigma^2, \tau_1^2, \tau_0^2, \delta^2) \approx P(\hat{\beta}, Y | \sigma^2, \tau_1^2, \tau_0^2, \delta^2) \cdot N(\beta | \hat{\beta}, -H^{-1}(\hat{\beta}))$$

$$\therefore P(Y | \sigma^2, \tau_1^2, \tau_0^2, \delta^2) = P(\hat{\beta}, Y | \sigma^2, \tau_1^2, \tau_0^2, \delta^2)(2\pi)^{\frac{p}{2}} \det(-H(\hat{\beta}))^{-\frac{1}{2}}$$

Where $\hat{\beta}$ denotes the MAP estimate.

We can use this to integrate out β, leading to the log marginal posterior:

$$\log P(Y | \sigma^2, \tau_1^2, \tau_0^2, \delta^2) = -\frac{1}{2} \log|\Sigma| - \frac{1}{2} (Y - X\beta)^T \Sigma^{-1} (Y - X\beta) - \frac{1}{2} \log|-H(\hat{\beta})| + const$$

Where const denotes terms not dependent on σ^2, δ^2

To take the derivative, we use the following matrix identities:

$$\frac{\partial}{\partial \theta} \log|\Sigma| = tr(\Sigma^{-1} \partial_\theta \Sigma)$$

$$\frac{\partial}{\partial \theta} (Y - X\beta)^T \Sigma^{-1} (Y - X\beta) = -(Y - X\beta)^T \Sigma^{-1} (\partial_\theta \Sigma) \Sigma^{-1} (Y - X\beta)$$

$$\frac{\partial}{\partial \theta} \log|-H(\beta)| = tr(-H^{-1}(\beta)(-\partial_\theta H(\beta)))$$

$$\partial_{\theta_i}\Sigma = \begin{matrix} e^{\theta_1}I; \;\; i=1 \\ e^{\theta_2}ZZ^T; i=2 \end{matrix}$$

$$\partial_\theta\big(-H(\beta)\big) = -X^T\Sigma^{-1}(\partial_\theta\Sigma)\Sigma^{-1}X$$

So we have:

$$\nabla_\theta \log P(Y|\sigma^2,\tau_1^2,\tau_0^2,\delta^2) = -\frac{1}{2}tr(\Sigma^{-1}\,\partial_\theta\Sigma) + \frac{1}{2}(Y-X\beta)^T\Sigma^{-1}(\partial_\theta\Sigma)\Sigma_{-1}(Y-X\beta)$$
$$-\frac{1}{2}tr(-H^{-1}(\beta)\big(-\partial_\theta H(\beta)\big)$$

And we maximize with respect to $\theta=(\sigma^2,\delta^2),$ parametrized as $(e^{\theta_1},e^{\theta_2})$ so we can do unconstrained optimization.

Step 5 – Model iteration

Repeat step 2-4 until convergence. We use the following convergence criterion:

$$\log P\big(\beta^{(t)}\big|X,Y,\theta^{(t)}\big) - \log P\big(\beta^{(t-1)}\big|X,Y,\theta^{(t-1)}\big) \le 0.0001$$

Step 6 – Estimation of posterior inclusion probabilities (PIPs)

Once we have estimated the MAP for β,σ^2,δ^2, we can find the PIP.

$$PIP_{\beta_j} = \int \frac{\pi N(\beta|0,\tau_1^2)}{\pi N(0,\tau_1^2)+(1-\pi)N(0,\tau_0^2)} N\left(\beta\middle|\hat{\beta},-H^{-1}(\hat{\beta})_{jj}\right)d\beta_j$$

While this integral has no closed form, it is a simple 1 dimensional integral that can be efficiently computed using Monte Carlo integration or Gaussian quadrature, as in Naqvi et al.

Appendix – derivations

Gradient of the likelihood:

$$
\begin{aligned}
\nabla_\beta\left(-\frac{N}{2}\log 2\pi-\frac{1}{2}\log|\Sigma|-\frac{1}{2}(Y-X\beta)^T\Sigma^{-1}(Y-X\beta)\right)&\\
=\nabla_\beta\left(-\frac{1}{2}(Y-X\beta)^T\Sigma^{-1}(Y-X\beta)\right)&\\
=\nabla_\beta\left(-\frac{1}{2}(Y^T-\beta^TX^T)\Sigma^{-1}(Y-X\beta)\right)&\\
=\nabla_\beta\left(-\frac{1}{2}(Y^T\Sigma^{-1}-\beta^TX^T\Sigma^{-1})(Y-X\beta)\right)&\\
=\nabla_\beta\left(-\frac{1}{2}(Y^T\Sigma^{-1}Y-Y^T\Sigma^{-1}X\beta-\beta^TX^T\Sigma^{-1}Y+\beta^TX^T\Sigma^{-1}X\beta)\right)&\\
=\nabla_\beta\left(\beta^TX^T\Sigma^{-1}Y-\frac{1}{2}\beta^TX^T\Sigma^{-1}X\beta\right)&\\
=X^T\Sigma^{-1}Y-X^T\Sigma^{-1}X\,\beta&
\end{aligned}
$$

Gradient of the prior

$$
\nabla_{\beta_j}\sum_{j=1}^{p}\log\left[\pi N\left(\beta_j\middle|0,\tau_1^2\right)+(1-\pi)N\left(\beta_j\middle|0,\tau_0^2\right)\right]
$$

$$
\begin{aligned}
&\nabla_{\beta_j}\log\left[\pi N\left(\beta_j\middle|0,\tau_1^2\right)+(1-\pi)N\left(\beta_j\middle|0,\tau_0^2\right)\right]\\
&=\frac{-\beta_j\left(\frac{\pi}{\tau_1^2}N_1+\frac{(1-\pi)}{\tau_0^2}N_0\right)}{\pi N_1+(1-\pi)N_0};\quad N_k=N(\beta_j|0,\tau_k^2)
\end{aligned}
$$

Hessian of the likelihood

$$
\begin{aligned}
\nabla_\beta^2\log P(Y,\beta|\sigma^2,\tau_1^2,\tau_0^2,\delta^2)&=\nabla_\beta(X^T\Sigma^{-1}Y-X^T\Sigma^{-1}X\,\beta)\\
&=-X^T\Sigma^{-1}X
\end{aligned}
$$

Hessian of the prior

$$
\nabla_\beta^2\log P(\beta|\tau_1^2,\tau_0^2)=\nabla_{\beta_j}\frac{-\beta_j\left(\frac{\pi}{\tau_1^2}N_1+\frac{(1-\pi)}{\tau_0^2}N_0\right)}{\pi N_1+(1-\pi)N_0}=\nabla_{\beta_j}\left(\frac{\nabla_{\beta_j}P(\beta)}{P(\beta)}\right)
$$

$$
\left[\frac{f(x)}{g(x)}\right]'=\frac{g(x)f'(x)-f(x)g'(x)}{g^2(x)}=\frac{P(\beta)\nabla_\beta^2P(\beta)}{[P(\beta)]^2}-\frac{\nabla_\beta P(\beta)\nabla_\beta P(\beta)}{[P(\beta)]^2}
$$

$$= \frac{\nabla^2_{\beta_j} P(\beta|\tau_1^2, \tau_0^2)}{P(\beta|\tau_1^2, \tau_0^2)} - \left[\frac{\nabla_{\beta_j} P(\beta|\tau_1^2, \tau_0^2)}{P(\beta|\tau_1^2, \tau_0^2)}\right]^2$$

$$\nabla^2_{\beta_j} P(\beta) = \nabla_{\beta_j}\left[\frac{-\beta_j \pi}{\tau_1^2} N_1 - \frac{\beta_j(1-\pi)}{\tau_0^2} N_0\right]$$

$$= \frac{\partial}{\partial \beta_j}\left(\frac{-\beta_j \pi}{\tau_1^2}\right) N_1 + \left(\frac{-\beta_j \pi}{\tau_1^2}\right)\frac{\partial}{\partial \beta_j} N_1 + \frac{\partial}{\partial \beta_j}\left(\frac{-\beta_j(1-\pi)}{\tau_0^2}\right) N_0 + \left(\frac{-\beta_j(1-\pi)}{\tau_0^2}\right)\frac{\partial}{\partial \beta} N_0$$