Supporting text: Gibbs Sampling algorithm for latent factor mixed models

Let us denote by n the total sample size, L is the number of loci, and D is the dimension of the set of environmental covariates. We denote $\mathrm{N}(\mu, \Sigma)$ the multivariate Gaussian distribution of mean μ and of covariance matrix Σ, and $\Gamma^{-1}(a, b)$ is the inverse-gamma distribution of shape a and rate b (and scale $1/b$).

Prior distributions. Let us denote by $\mathrm{I}_{i,\ell}$ an indicator variable equal to 0 when the genotype data are missing at locus ℓ for individual i, and equal to 1 otherwise. The prior distributions on the LFMM parameters are defined hierarchically as follows. For all i, ℓ, the allele count is described by

$$G_{i,\ell} \mid U_i, V_\ell, \beta_\ell, \mu_\ell, \sigma^2 \sim \mathrm{N}(X_i\beta_\ell + \mu_\ell + U_i^T V_\ell, \sigma^2)^{\mathrm{I}_{i,\ell}}\,, \tag{1}$$

where μ_ℓ is a locus-specific mean, and σ^2 is a residual variance term. The factors are described as

$$U_i|\sigma_U^2 \sim \mathrm{N}(0, \sigma_U^2 \mathrm{I}_K)\,, \tag{2}$$

where I_K is the identity matrix with K dimensions, and K is the number of factors in the model. The scores are described as

$$V_\ell \sim \mathrm{N}(0, \mathrm{I}_K)\,, \tag{3}$$

where I_K is the identity matrix with K dimensions. For $1 \leq j \leq D$, the fixed effect regression coefficients are described by

$$\beta_{j\ell}|\sigma_{\beta_j}^2 \sim \mathrm{N}(0, \sigma_{\beta_j}^2)\,, \tag{4}$$

and the intercept coefficients are described by

$$\mu_\ell|\sigma_\mu^2 \sim \mathrm{N}(0, \sigma_\mu^2)\,. \tag{5}$$

The hyper-parameters, $\sigma^2_{\beta_j}$, σ^2_μ follow non-informative inverse-chi squared distributions, and σ^2_U follows an inverse-gamma distribution $\Gamma^{-1}(\eta, \eta)$ where $\eta = 10^2 - 10^3$. This parameterization encourages sparsity in factor estimates.

Conditional distributions. The LFM model is a hierarchical model with conditional distributions that can be described as follows:

$$p(\sigma_U^2|U, \eta) = \Gamma^{-1}(\eta + \frac{nK}{2}, \frac{1}{2}\sum_i U_i^T U_i + \eta) \tag{6}$$

$$p(\sigma^2_{\beta_j}|\beta) = \Gamma^{-1}(1 + \frac{L}{2}, \frac{1}{2}\sum_l \beta^2_{j\ell} + 1) \tag{7}$$

$$p(\sigma_\mu^2|\mu) = \Gamma^{-1}(1 + \frac{L}{2}, \frac{1}{2}\sum_\ell \mu_\ell^2 + 1) \tag{8}$$

$$p(U_i|G, V, \beta, \mu, \sigma_U^2, \sigma^2) = \mathrm{N}(\mu_U^i, {\Delta_U^i}^{-1})\,, \tag{9}$$

where

$$\Delta_U^i = {\sigma_U}^{-2}\mathrm{I}_K + \sigma^{-2}\sum_\ell V_\ell V_\ell^T\,, \tag{10}$$

and

$$\mu_U^i = \sigma^{-2}(\Delta_U^i)^{-1}\sum_\ell (G_{i,\ell} - X_i\beta_\ell - \mu_\ell)V_\ell\,. \tag{11}$$

In addition, we have

$$p(V_\ell|G, U, \beta, \mu, \alpha_G) = \mathrm{N}(\mu_V^\ell, {\Delta_V^\ell}^{-1})\,, \tag{12}$$

where

$$\Delta_V^\ell = {\sigma_V}^{-2}\mathrm{I}_K + \sigma^{-2}\sum_i U_i U_i^T \tag{13}$$

and

$$\mu_V^\ell = \sigma^{-2}(\Delta_V^\ell)^{-1}\sum_i (G_{i,\ell} - X_i\beta_\ell - \mu_\ell)U_i\,. \tag{14}$$

We have

$$p(\beta_\ell|G, U, V, \mu, \sigma_\beta(1)^2, \ldots, \sigma_\beta(d)^2, \sigma^2) = \mathrm{N}(\mu_\beta^\ell, {\Delta_\beta^\ell}^{-1}) \tag{15}$$

where

$$\Delta_\beta^\ell = \mathrm{diag}({\sigma_{\beta_1}}^{-2}, \cdots, {\sigma_{\beta_j}}^{-2}) + \sigma^{-2}\sum_i X_i^T X_i \tag{16}$$

and

$$\mu_\beta^\ell = \sigma^{-2}(\Delta_\beta^\ell)^{-1}\sum_i (G_{i,\ell} - U_i^T V_\ell - \mu_\ell)X_i^T\,. \tag{17}$$

Finally, we have

$$p(\mu_\ell|G, U, V, \beta, \sigma_\mu^2, \sigma^2) = \mathrm{N}(\mu_\mu^\ell, {\Delta_\mu^\ell}^{-1}) \tag{18}$$

where

$$\Delta_\mu^\ell = {\sigma_\mu}^{-2} + n\sigma^{-2} \tag{19}$$

and

$$\mu_\mu^\ell = \sigma^{-2}(\Delta_\mu^\ell)^{-1}\sum_i (G_{i,\ell} - U_i^T V_\ell - X_i\beta_\ell) \tag{20}$$

The parameter σ^2 is updated at each iteration using the current residual variance. The other parameters are updated through Gibbs sampling cycles.

Main algorithm Let nc be the number of Gibbs sampler cycles, and 'burn' be the number of cycles used for burn-in.

1. Initialize the model parameters

$$U = 0_{K,n}$$

$$V = 0_{K,L}$$

$$\beta = 0_{L,D}$$

$$\mu = 0_{L,1}$$

2. For $t = 1, \ldots,$ nc

   - Input missing values at locus ℓ for individual i,

   $$G_{i,\ell} \leftarrow {U_i^{(t-1)}}^T V_\ell^{(t-1)} + X_i^{(t-1)} \beta_\ell^{(t-1)}$$

   - Update the residual variance

   $$\sigma^{2^{(t)}} = var(G - {U^{(n-1)}}^T V^{(t-1)} - X^{(t-1)} \beta^{(t-1)})$$

   - Sample the hyper-parameters

   $${\sigma_U^2}^{(t)} \sim p(\sigma_U^2 | U^{(t-1)}, \eta)$$

   $$\sigma_\beta^{2(t)} \sim p(\sigma_\beta^2 | \beta^{(t-1)})$$

   $$\sigma_\mu^{2(t)} \sim p(\sigma_\mu^2 | \mu^{(t-1)})$$

   - For each locus ℓ, sample

   $$\mu_\ell^{(t)} \sim p(\mu_\ell | U^{(t-1)}, V^{(t-1)}, \beta^{(t-1)}, {\sigma_\mu^2}^{(t)}, \sigma^{2^{(t)}})$$

   $$\beta_\ell^{(t)} \sim p(\beta_\ell | U^{(t-1)}, V^{(t-1)}, \mu^{(t)}, {\sigma_{\beta_1}^2}^{(t)}, \ldots, {\sigma_{\beta_j}^2}^{(t)}, \sigma^{2^{(t)}})$$

   - For each individual i, sample

   $$U_i^{(t)} \sim p(U_i | \mu^{(t)}, V^{(t-1)}, \beta^{(t)}, {\sigma_U^2}^{(t)}, \sigma^{2^{(t)}})$$

   - For each locus ℓ, sample

   $$V_\ell^{(t)} \sim p(V_\ell | \mu^{(t)}, U^{(t)}, \beta^{(t)}, \sigma^{2^{(t)}})$$

3. compute the parameters

$$U = \text{mean}(U^{(\text{burn}+1)}, \ldots, U^{(\text{nc})})$$

$$V = \text{mean}(V^{(\text{burn}+1)}, \ldots, V^{(\text{nc})})$$

$$\beta = \text{mean}(\beta^{(\text{burn}+1)}, \ldots, \beta^{(\text{nc})})$$

$$\mu = \text{mean}(\mu^{(\text{burn}+1)}, \ldots, \mu^{(\text{nc})})$$

$$Z = \text{mean}(\beta^{(\text{burn}+1)}, \ldots, \beta^{(\text{nc})}) / \text{var}(\beta^{(\text{burn}+1)}, \ldots, \beta^{(\text{nc})})^{\frac{1}{2}}$$