1 MCMC Algorithm

In this section we describe the steps of the MCMC algorithm. We implement a Gibbs sampling, which requires a Metropolis update for some parameters.

1. The update of β, the vector of regression coefficients, is the standard conjugate update from a Normal model, but conditional on the Spike and Slab selection. Note that in our application the observation y_i are also indexed by the ethnicity indicator g. For ease of notation we drop the subscript g as we assume the regression coefficients to be the same across ethnicities and we simply assume to have a total of n observations. We introduce the latent variable indicator vector $\omega = \left(\omega_1, \ldots, \omega_{p-1}\right)$, where the element ω_j is equal to 1 if the j_{th} covariate is included in the model and 0 otherwise. If $\omega_j = 0$ then the corresponding β_j is equal to zero. Let β_ω denote the sub-vector of β including elements for which the corresponding ω_j is equal to 1 (slab component of the model) and let $\mathbf{X}_\omega$ be the design matrix consisting only of those columns of $\mathbf{X}$ corresponding to non-zero effects. Then the conditional distribution of β_ω

$$
\begin{aligned}
p(\beta_\omega \mid rest) &\propto \prod_{i=1}^{n} \mathrm{N}\left(y_i \mid \beta_{0i} + x_{\omega i}\beta_\omega, \tau_i^2\right) \times \prod_{j:\omega_j=1} \mathrm{N}\left(\beta_j \mid \mu_\beta, \tau_\beta^2\right) \\
&= \mathrm{N}\left(\beta_\omega \mid \tilde{\mu}_\beta, \tilde{C}_\beta\right)
\end{aligned}
$$

where

$$
\begin{aligned}
\tilde{C}_\beta &= \tau_\beta^2 I + \mathbf{X}_\omega' V \mathbf{X}_\omega \\
V &= \begin{bmatrix} \tau_1^2 & 0 & \cdots & 0 \\ 0 & \tau_2^2 & \ddots & 0 \\ \vdots & 0 & \ddots & \cdots \\ 0 & \cdots & 0 & \tau_n^2 \end{bmatrix} \\
\tilde{\mu}_\beta &= \tilde{C}_\beta^{-1} \left(\tau_\beta^2 \mu_\beta + \mathbf{X}_\omega' V y\right)
\end{aligned}
$$

and $y = (y_1, \ldots, y_n)$. Here μ_β is the vector of appropriate dimension whose elements are all equal to μ_β.

2. The update of ω is performed evaluating the model marginal likelihood individually for each covariate (with the intercept β_{0i} always included) as

$$p(\omega_j = 1 \mid \omega_{\setminus j}, rest) = \left[1 + \frac{1-\pi}{\pi} \frac{p(y \mid \omega_j = 0, \omega_{\setminus j}, \tau_1^2, \ldots, \tau_n^2)}{p(y \mid \omega_j = 1, \omega_{\setminus j}, \tau_1^2, \ldots, \tau_n^2)}\right]^{-1}$$

where $\omega_{\setminus j}$ denotes the vector ω excluding ω_j. $p(y \mid \omega_j = 1, \omega_{\setminus j}, \tau_1^2, \ldots, \tau_n^2)$ represents the marginal likelihood of the model obtained marginalising with respect to β:

$$\begin{aligned}
&p(y \mid \omega_j = 1, \omega_{\setminus j}, \tau_1^2, \ldots, \tau_n^2) = \\
&= \int_\beta p(y, \beta \mid \tau_1^2, \ldots, \tau_n^2, \omega_j = 1, \omega_{\setminus j})\, d\beta \\
&= \int_\beta p(y \mid \beta, \tau_1^2, \ldots, \tau_n^2, \omega_j = 1, \omega_{\setminus j}) p(\beta)\, d\beta \\
&= -\frac{1}{2} n \log(2\pi) + \frac{1}{2} \log\left(|\tau_\beta^2 I|\right) - \frac{1}{2} \log\left(|\tilde{C}_\beta|\right) \\
&= -\frac{1}{2}\left(\tilde{y}'\tilde{y} - \tilde{\mu}_\beta' \tilde{C}_\beta \tilde{\mu}_\beta\right)
\end{aligned}$$

where $|A|$ is the determinant of the matrix A and $\tilde{y} = (\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_n)'$, where $\tilde{y}_i = (y_i - \beta_{0i})\tau_i$.

3. The update of π is a straightforward conjugate update from a Beta-Bernoulli model

$$p(\pi \mid rest) = \text{Beta}\left(\pi_a + \sum \omega_j, \pi_b + (p-1) - \sum \omega_j\right)$$

4. To update the DGDP we adopt a truncated stick-breaking approach, i.e. we approximate the infinite mixture with a finite mixture with L components where L is large. A discussion on the truncation level can be found in Ishwaran and James (2001) and Barcella, De Iorio, Favaro, and Rosner (2017). We perform the following steps in order to update the parameters of the DGDP.
   (a) *Resampling the cluster allocation vector, given the rest.* Conditionally on the remaining parameters in the model, the allocation vectors, s_g, are independent. Note that we have an allocation vector for each ethnicity

g. Let s_{ig} be the cluster indicator for observation i in group g, with $s_{ig} \in 1,\dots,L$, for $i=1,\dots,n$. We draw s_{ig} from

$$p(s_{ig}=k \mid rest) \propto \psi_{kg}\mathrm{N}(y_{ig} \mid \beta_{0i} + \sum_{j:\omega_j=1} \beta_j x_{ij}, \tau_i^2)$$

for $k=1,\dots,L$.

(b) *Resampling the mixture weights, ψ_{kg}, given the rest.* Conditionally on g and the remaining parameters in the model, the mixture weights for each group are independent. This is a straightforward update due to the conjugacy between the Generalised Dirichlet distribution on $\psi_{1g},\dots,\psi_{Lg}$ and the Multinomial distribution on s:

$$\phi_{kg} \mid rest \sim \mathrm{Beta}\left(\mu_g\upsilon + \sum_{i=1}^{n_g}\mathrm{I}(s_{ig}=k), (1-\mu_g)\upsilon + \sum_{i=1}^{n_g}\mathrm{I}(s_{ig}>k)\right)$$

where n_g is the number of observations in group g and $\mathrm{I}(\cdot)$ represents the indicator function, assuming value 1 if the inner condition is satisfied and 0 otherwise. Then the weights ψ_{kg} can be obtained using a stick-breaking procedure.

(c) *Resampling $\mu=(\mu_1,\dots,\mu_G)$ given the rest.* Conditionally on the weights ψ_{kg} and υ we update μ with a Metropolis-Hastings step, using a Multivariate Normal proposal. In this case the data corresponds to the set of sticks ϕ_{kg} and the likelihood is given by the product of Beta distributions. See Barcella et al. (2017) for details.

(d) *Resampling υ given the rest.* Conditionally on the weights ψ_{kg} and μ we update υ with a Metropolis-Hastings step, with a Gamma proposal and data given by the set sticks ϕ_{kg} for $g=1,\dots,G$.

(e) *Resampling of the locations $\theta_k=(\beta_{0k},\tau_k^2)$ given the rest.* The locations of the DGDP are a priori *iid* realisations of the base measure $G_0 = \mathrm{N}\left(m_0,\kappa_0^2\right) \times \mathrm{Gamma}(\tau_a,\tau_b)$. Given the clustering structure defined by the allocation vector s, the update of θ_k is performed separately for each cluster and it is a straightforward conjugate update:

$$p(\beta_{0k},\tau_k^2 \mid rest) \propto G_0(\beta_{0k},\tau_k^2)\prod_{i,g:s_{ig}=k}\mathrm{N}(y_{ig} \mid \beta_{0k} + \sum_{j:\omega_j=1}\beta_j x_{ig}, \tau_k^2)$$

for $k=1,\dots,L$

2 Tables

Here we provide the list of metabolites, anthropometric and clinical covariates included in the analysis.

References

Barcella, W., M. De Iorio, S. Favaro, and G. L. Rosner (2017): "Dependent generalized dirichlet process priors for the analysis of acute lymphoblastic leukemia," *Biostatistics*, 19, 342–358.

Ishwaran, H. and L. F. James (2001): "Gibbs sampling methods for stick-breaking priors," *Journal of the American Statistical Association*, 96, 161–173.

Table 1: List of metabolites included in the analysis. Each listed lipoprotein is further fractioned according to the content of triglycerides, phospholipids, cholesterol esters and free cholesterol.

Abbreviation	Full name
acace	Acetoacetate
ace	Acetate
ala	Alanine
alb	Albumin
apoa1	Apolipoprotein A-I
apob	Apolipoprotein B
bohbut	3-hydroxybutyrate
cit	Citrate
crea	Creatinine
dha	22:6, docosahexaenoic acid
faw3	Omega-3 fatty acids
faw6	Omega-6 fatty acids
gln	Glutamine
glol	Glycerol
gly	Glycine
gp	Glycoprotein acetyls, mainly a1-acid glycoprotein
his	Histidine
ile	Isoleucine
la	18:2, linoleic acid
lac	Lactate
leu	Leucine
mufa	Monounsaturated fatty acids; 16:1, 18:1
pc	Phosphatidylcholine and other cholines
phe	Phenylalanine
pufa	Polyunsaturated fatty acids
pyr	Pyruvate
sfa	Saturated fatty acids
sm	Sphingomyelins
tyr	Tyrosine
unsatdeg	Estimated degree of unsaturation
val	Valine

Abbreviation	Full name
lipids_s_hdl	Small HDL lipids compounds
lipids_m_hdl	Medium HDL lipids compounds
lipids_l_hdl	Large HDL lipids compounds
lipids_xl_hdl	Extra large HDL lipids compounds
lipids_s_ldl	Small LDL lipids compounds
lipids_m_ldl	Medium LDL lipids compounds
lipids_l_ldl	Large LDL lipids compounds
lipids_idl	IDL lipids compounds
lipids_xs_vldl	Extra small VLDL lipids compounds
lipids_s_vldl	Small VLDL lipids compounds
lipids_m_vldl	Medium VLDL lipids compounds
lipids_l_vldl	Large VLDL lipids compounds
lipids_xl_vldl	Extra large VLDL lipids compounds
lipids_xxl_vldl	Extra extra large VLDL lipids compounds

Table 2: List of clinical and anthropometric covariates

Abbreviation	**Full name**
Age	Age at the first visit (baseline)
WHR	Waist to Hip Ratio
Thigh skinfold	Thigh skinfold
Sagittal diam	Sagittal diameter
Subscap skinfold	Subscapular skinfold
Supiliac skinfold	Suprailiac skinfold
Thigh circumf	Thigh circumference
Triceps skinfold	Triceps skinfold
bp_avdias	Blood pressure diastolic
bp_avsys	Blood pressure systolic
AST	Aspartate aminotransferase
GGT	Gamma glutamyltransferase
Sex female	Dummy variable for female sex
Smoke_Ex	Dummy variable ex smoker
Smoke_Current	Dummy variable current smoker