Modeling population structure under hierarchical Dirichlet processes: Appendix

Lloyd T. Elliott[*], Maria De Iorio[†], Stefano Favaro[‡,§], Kaustubh Adhikari[¶] and Yee Whye Teh[*,‖]

Appendix

We develop a Gibbs sampler algorithm, in which each variable is updated given the remaining variables fixed at their current value.

Update for G_0**.** See section 2.4 of main manuscript.

Update for G_i**.** See section 2.4 of main manuscript.

Update for n_{ik} ***and*** m_{ik}**.** Assume that at iteration w there are K populations in the model. We need to resample the seating arrangement of the Chinese restaurant of G_i. Updating n_{ik} is straightforward as the new sample is simply the number of linked segments with $z_{il} = k$. That is,

$$n_{ik} = \sum_{l=1}^{L} \mathbf{1}(s_{il} = 0)\mathbf{1}(z_{il} = k).$$

Then, m_{ik} can be sampled from the distribution of the number of tables in Chinese Restaurant Process with $n_i k$ customers and mass parameter αq_{0k}. That is,

$$\begin{aligned} m_{ik} &= \sum_{j=1}^{n_{ik}} b_j \\ b_j &\sim \text{Bernoulli}\left(\frac{\alpha q_{ik}}{\alpha q_{ik} + j - 1}\right) \end{aligned}$$

where $b_j = 1$ if customer j joins a new table.

Update for θ_{kl}**.** Assume that at iteration w there K populations in the model. The posterior $p(\theta_{kl}) \mid \text{rest})$ is the same as in a simple parametric Bayesian model using as observations all the markers for which $z_{il} = k$. In the case of SNP data the conditional posterior of θ_{kl} is a Beta distribution.

[*]Department of Statistics, University of Oxford, Oxford OX1 3TG, U.K.
[†]Department of Statistical Science, University College London, London WC1E 6BT, U.K. m.deiorio@ucl.ac.uk
[‡]Department of Economics and Statistics, 10134 Torino, Italy.
[§]Collegio Carlo Alberto, 10024 Moncalieri, Italy.
[¶]Department of Genetics, Evolution and Environment, University College London, London WC1E 6BT, U.K.
[‖]Google DeepMind, London EC4A 3TW, U.K.

Update for α**.** The concentration parameter α_0 governs the distribution of the number of of θ_{kl}'s in each mixture. We follow Teh et al. (2012). Assume that at iteration w there K populations in the model. Let $m_{\bullet\bullet} = \sum_{i,k} m_{ik}$ and $n_{i\bullet} = \sum_k n_{ik}$. We introduce latent variable $w_i \in [0,1]$ and $t_i \in \{0,1\}$, $i = 1, \ldots, N$, with

$$\begin{aligned} w_i \mid \alpha_0 &\sim \text{Beta}(\alpha_0 + 1, n_{i\bullet}) \\ t_i \mid \alpha_0 &\propto \left(\frac{n_{i\bullet}}{\alpha_0 + n_{i\bullet}}\right)^{t_i}. \end{aligned}$$

If α is given a $\Gamma(a,b)$ hyperprior, then given w_i and t_i, α has a Gamma distribution with parameters $a + m_{\bullet\bullet} - \sum_{i=1}^N t_i$ and $b - \sum_{i=1}^N \log w_i$.

Update for α_0**.** Given the total number $m_{\bullet\bullet} = \sum_{i,k} m_{ik}$ of the θ_{kl}'s, the concentration parameter α_0 governs the distribution of the number of population K. We use the auxiliary method of Escobar and West (1995). If α_0 is given a $\Gamma(a,b)$ hyperprior, it can be resampled by introducing a latent variable $\gamma \sim \text{Beta}(\alpha_0 + 1, m_{\bullet\bullet})$ and

$$p(\alpha_0 \mid K, \gamma) = \pi\Gamma(a + K, b - \log(\gamma)) + (1 - \pi)\Gamma(a + K - 1, b - \log(\gamma))$$

where $\pi/(1-\pi) = (a + K - 1)/m_{\bullet\bullet}(b - \log(\gamma))$.

Update for r**.** The rate r of the Poisson process is given a Uniform prior on some interval $[r_L, r_U]$. We use a random walk Metropolis step to update r with proposal distribution centred around the current value.

Update for hyperparameters in the base measure H**.** The proportion of the model involving the hyperparameters in H is a convential parametric model. Hence, conditioning on all the other variables usually leaves us with a standard Bayesian model, often in conjugate form. In the case of SNP data, we have taken $\theta_k \sim H = \prod_{l=1}^L \text{Beta}(c\mu_l, c(1-\mu_l))$. We specify independent $\text{Beta}(a_l, b_l)$ for each μ_l, $l = 1, \ldots, L$, and update μ_l using a random walk Metropolis step with proposal distribution centred around the current value.

References

Escobar, M. D. and West, M. (1995). "Bayesian density estimation and inference using mixtures." *Journal of the American Statistical Association*, 90(430): 577–588. 2

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2012). "Hierarchical Dirichlet processes." *Journal of the American Statistical Association*, 101(476): 1566–1581. 2