

Web-based Supplementary Materials for "A Latent Model To Detect Multiple Clusters of Varying Sizes" by Minge Xie, Qiankun Sun and Joseph Naus

Web Appendix I: Gibbs Sampling Algorithms in Section 3.

A. *Gibbs sampling method in Section 3.2.* The Gibbs samples $\mathbf{b}^* = (b_1^*, b_2^*, \dots, b_{k+1}^*)'$ and $\mathbf{c}^* = (c_1^*, c_2^*, \dots, c_k^*)'$ from $f(\mathbf{b}, \mathbf{c} | \mathbf{y}, k, \boldsymbol{\theta}^{(s)})$ are generated by cycling through simulations from the (fully) conditional distributions of b_j or c_j given the rest of b 's and c 's many times until the Gibbs sampling chain is "burnt-in".

The question is how to simulate b_j or c_j from the fully conditional distributions, given the rest of b_j 's or c_j 's. By ignoring unwanted terms, it is easy to see that $f(b_j | b_l, l = 1, \dots, k+1, l \neq j, \mathbf{c}, \mathbf{y}, k) \propto f(\mathbf{b}, \mathbf{c}, \mathbf{y} | k) \propto e^{\sum_{s=1}^k Z_s(\log \alpha_s)} \psi_b(b_j) \mathbf{1}_{\{\delta=k\}}$ and $f(b_{k+1}) \propto \psi_b(b_{k+1}) \mathbf{1}_{\{\delta=k\}}$. Similarly, $f(c_j | c_l, l = 1, \dots, k, l \neq j, \mathbf{b}, \mathbf{y}, k) \propto f(\mathbf{b}, \mathbf{c}, \mathbf{y} | k) \propto e^{\sum_{s=1}^k Z_s(\log \alpha_s)} / \{T + \sum_{s=1}^k (\alpha_s - 1) c_s\}^n \psi_c(c_j) \mathbf{1}_{\{\delta=k\}}$. Thus, for a given $\boldsymbol{\theta} = (\alpha, \lambda)'$, we can use the following importance sampling method to simulate a b_j from the fully conditional distribution $f(b_j | b_l, l = 1, \dots, k+1, l \neq j, \mathbf{c}, \mathbf{y}, k)$:

STEP A. Simulate a large number, say N , random deviates e_1, e_2, \dots, e_N from a candidate distribution $\tilde{\psi}_b(b_j)$. Then, compute weight $w_l = \{\psi_b(e_l) / \tilde{\psi}_b(e_l)\} e^{\sum_{s=1}^k Z_s^{[l]}(\log \alpha_s)} \mathbf{1}_{\{\delta^{[l]}=k\}}$, for $l = 1, \dots, N$, where $Z_s^{[l]}$ is the total number of incidences in s th cluster and $\{\delta^{[l]} = k\}$ is the constraint of having k clusters but with the b_j being replaced by e_l and the rest of b 's and c 's kept the same. In the case of simulating b_{k+1} given the rest b 's and c 's, the weight can be simplified to $w_l = \{\psi_b(e_l) / \tilde{\psi}_b(e_l)\} \mathbf{1}_{\{\delta^{[l]}=k\}}$.

STEP B. Simulate b_j from one of the N values $\{e_1, e_2, \dots, e_N\}$ with respective probabilities (p_1, p_2, \dots, p_N) ; Here, $p_l = w_l / \sum_{s=1}^N w_s$.

Similarly, we can simulate a c_j from the fully conditional distribution $f(c_j | c_l, l = 1, 2, \dots, k, l \neq j, \mathbf{b}, \mathbf{y}, \delta = k)$. In this case, we simulate in Step A random deviates e_1, e_2, \dots, e_N from a candidate distribution $\tilde{\psi}_c(c_j)$ and compute weight $w_l = \{\psi_c(e_l) / \tilde{\psi}_c(e_l)\} e^{\sum_{s=1}^k Z_s^{[l]}(\log \alpha_s)} / \{T + \sum_{s \neq j} (\alpha_s - 1) c_s + (\alpha_j - 1) e_l\} \mathbf{1}_{\{\delta^{[l]}=k\}}$. Here, again, $Z_s^{[l]}$ and $\{\delta^{[l]} = k\}$ are computed with given

b and c values but now the c_j is replaced by e_l . Step B is the as in the b_j case.

In the special exponential case, $\psi_b(b_j) \sim \text{Exp}(1/\lambda_b)$ and $\psi_c(c_j) \sim \text{Exp}(1/\lambda_c)$. Since it is easy to directly simulate from a truncated exponential distribution, we suggest to pick $\tilde{\psi}_b(b_j) \propto \text{Exp}(1/\lambda_b)\mathbf{1}(\delta = k)$ and $\tilde{\psi}_c(c_j) \propto \text{Exp}(1/\lambda_c)\mathbf{1}(\delta = k)$, instead of $\tilde{\psi}_b(b_j) \propto \text{Exp}(1/\lambda_b)$ and $\tilde{\psi}_c(c_j) \propto \text{Exp}(1/\lambda_c)$, to improve computing efficiency.

B. *Gibbs sampling method in Section 3.4.* We need to use a Gibbs sampling method to simulate $\mathbf{b}^{**} = (b_1^{**}, \dots, b_{k+1}^{**})$ and $\mathbf{c}^{**} = (c_1^{**}, \dots, c_k^{**})$ from $f_\theta(\mathbf{b}, \mathbf{c}|k)$ from any set of given parameter values θ in Section 3.4. By the standard Gibbs sampling procedure, we cycle through simulating b_j or c_j from their corresponding fully conditional distributions $f_\theta(b_j|b_l, l = 1, 2, \dots, k+1, l \neq j, \mathbf{c}, \delta = k)$ or $f_\theta(c_j|c_l, l = 1, 2, \dots, k, l \neq j, \mathbf{b}, \delta = k)$ many times. We can get a set of \mathbf{b}^{**} and \mathbf{c}^{**} from $f_\theta(\mathbf{b}, \mathbf{c}|k)$. Note that, $f_\theta(b_j|b_l, l = 1, 2, \dots, k+1, l \neq j, \mathbf{c}, \delta = k) \propto \psi_b(b_j)\mathbf{1}_{(\delta=k)}$ and $f_\theta(c_j|c_l, l = 1, 2, \dots, k, l \neq j, \mathbf{b}, \delta = k) \propto \psi_c(c_j)\mathbf{1}_{(\delta=k)}$. So, these fully conditional distributions are truncated distributions. In many cases, for example $\psi_b(b_j)$ and $\psi_c(c_j)$ being exponential distributions, the truncated distribution can be directly simulated. Otherwise, we can use an importance sampling algorithm to obtain \mathbf{b}^{**} and \mathbf{c}^{**} samples, assuming we know how to simulate from $\psi_b(b_j)$ and $\psi_c(c_j)$.

Web Appendix II: Data Sets

A. *Hospital Hemoptysis Admission Data Set.* Days of hemoptysis admission at Nice University Hospital from January 1 to December 31, 1995, are: 2 8 23 29 43 48 58 60 61 63 69 71 74 74 78 80 85 86 86 87 93 105 106 108 115 117 121 126 135 140 141 156 159 179 187 188 188 191 191 198 201 214 225 225 235 235 239 249 262 271 279 279 282 292 296 302 317 323 337 342 352 354.

B. *Brucellosis counts per week from the CDC database.* The 1997-2003 weekly averages are: 0.86 1.00 1.29 0.72 1.15 1.43 0.86 1.29 1.86 1.29 2.00 1.58 1.29 1.29 1.15 2.00 0.72 1.29 2.43 2.15 3.58 1.86 1.00 2.00 3.00 2.58 3.29 2.15 2.29 3.29 3.43 2.72 2.43 1.58 2.58 2.86 3.29 3.00 2.72 1.86 1.72 2.43 3.58 2.00 1.29 2.00 1.29 2.43 3.15 2.15 3.43 7.15. The 2004 Data are: 0 0 0 0 2 3 1 1 0 5 4 1 1 0 3 2 0 1 1 4 3 2 4 6 0 9 3 2 5 4 0 10 0 0 3 1 8 5 5 4 4 5 0 27 7 12 0 5 6 6 4 2