

Appendix to “Variational approximation for  
mixtures of linear mixed models” published in the  
Journal of Computational and Graphical Statistics

Siew Li Tan and David J. Nott

Dec 2012

## A Derivation of variational lower bound for Algorithm 1

The variational lower bound can be written as  $E_q\{\log p(y, \theta)\} - E_q\{\log q(\theta)\}$ , where  $E_q(\cdot)$  denotes the expectation with respect to  $q$ . Consider the first term,  $E_q\{\log p(y, \theta)\}$ . Let  $\zeta_{ij} = I(z_i = j)$  where  $I(\cdot)$  denotes the indicator function. We have

$$\begin{aligned} \log p(y, \theta) = & \sum_{i=1}^n \sum_{j=1}^k \zeta_{ij} \left\{ \log p(y_i | z_i = j, \beta_j, a_i, b_j, \Sigma_{ij}) + \log p(a_i | \sigma_{a_j}^2) + \log p_{ij} \right\} + \log p(\delta) \\ & + \sum_{j=1}^k \left\{ \log p(\beta_j) + \log p(b_j | \sigma_{b_j}^2) + \log p(\sigma_{a_j}^2) + \log p(\sigma_{b_j}^2) + \sum_{l=1}^g \log p(\sigma_{jl}^2) \right\}. \end{aligned}$$

Taking expectations with respect to  $q$ , we obtain

$$\begin{aligned} E_q\{\log p(y, \theta)\} = & -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^k q_{ij} \left[ \sum_{l=1}^g \kappa_{il} \left\{ \log \lambda_{jl}^q - \psi(\alpha_{jl}^q) \right\} + \xi_{ij}^T \Sigma_{ij}^{q-1} \xi_{ij} + \text{tr}(\Sigma_{ij}^{q-1} \Lambda_{ij}) \right. \\ & \left. + s_1 \left\{ \log \lambda_{a_j}^q - \psi(\alpha_{a_j}^q) \right\} - \frac{\alpha_{a_j}^q}{\lambda_{a_j}^q} \left\{ \text{tr}(\Sigma_{a_i}^q) + \mu_{a_i}^{qT} \mu_{a_i}^q \right\} \right] - \frac{N+ns_1+k(p+s_2)}{2} \log(2\pi) \\ & + \sum_{i=1}^n \sum_{j=1}^k q_{ij} \log p_{ij} - \frac{1}{2} \sum_{j=1}^k \left[ \log |\Sigma_{\beta_j}| + \text{tr}(\Sigma_{\beta_j}^{-1} \Sigma_{\beta_j}^q) + \mu_{\beta_j}^{qT} \Sigma_{\beta_j}^{-1} \mu_{\beta_j}^q \right] \\ & - \frac{1}{2} \sum_{j=1}^k \left[ s_2 \left\{ \log \lambda_{b_j}^q - \psi(\alpha_{b_j}^q) \right\} + \frac{\alpha_{b_j}^q}{\lambda_{b_j}^q} \left\{ \text{tr}(\Sigma_{b_j}^q) + \mu_{b_j}^{qT} \mu_{b_j}^q \right\} \right] + \log p(\mu_\delta^q) \\ & + \sum_{j=1}^k \left[ \alpha_{a_j} \log \lambda_{a_j} - \log \Gamma(\alpha_{a_j}) - (\alpha_{a_j} + 1) \left\{ \log \lambda_{a_j}^q - \psi(\alpha_{a_j}^q) \right\} - \frac{\lambda_{a_j} \alpha_{a_j}^q}{\lambda_{a_j}^q} \right. \\ & \left. + \alpha_{b_j} \log \lambda_{b_j} - \log \Gamma(\alpha_{b_j}) - (\alpha_{b_j} + 1) \left\{ \log \lambda_{b_j}^q - \psi(\alpha_{b_j}^q) \right\} - \frac{\lambda_{b_j} \alpha_{b_j}^q}{\lambda_{b_j}^q} \right] \\ & + \sum_{j=1}^k \sum_{l=1}^g \left[ \alpha_{jl} \log \lambda_{jl} - \log \Gamma(\alpha_{jl}) - (\alpha_{jl} + 1) \left\{ \log \lambda_{jl}^q - \psi(\alpha_{jl}^q) \right\} - \frac{\lambda_{jl} \alpha_{jl}^q}{\lambda_{jl}^q} \right] \end{aligned} \quad (1)$$

where  $\Gamma(\cdot)$  and  $\psi(\cdot)$  denote the gamma and digamma functions respectively,  $p_{ij}$  is evaluated by setting  $\delta = \mu_\delta^q$ ,  $p(\mu_\delta^q)$  denotes the prior distribution for  $\delta$  evaluated at  $\mu_\delta^q$ ,  $\xi_{ij} = y_i - X_i \mu_{\beta_j}^q - W_i \mu_{a_i}^q - V_i \mu_{b_j}^q$ ,  $\Sigma_{ij}^{q-1} = \text{blockdiag} \left( \frac{\alpha_{j1}^q}{\lambda_{j1}^q} I_{\kappa_{i1}}, \dots, \frac{\alpha_{jg}^q}{\lambda_{jg}^q} I_{\kappa_{ig}} \right)$  and  $\Lambda_{ij} = X_i \Sigma_{\beta_j}^q X_i^T + W_i \Sigma_{a_i}^q W_i^T + V_i \Sigma_{b_j}^q V_i^T$ .

Turning to the second term,  $E_q\{\log q(\theta)\}$ , we have

$$\begin{aligned}
E_q\{\log q(\theta)\} &= \sum_{j=1}^k \left[ E_q\{\log q(\beta_j)\} + E_q\{\log q(b_j)\} + E_q\{\log q(\sigma_{b_j}^2)\} + E_q\log q(\sigma_{a_j}^2) \right] \\
&\quad + \sum_{j=1}^k \sum_{l=1}^g E_q\{\log q(\sigma_{jl}^2)\} + \sum_{i=1}^n E_q\{\log q(a_i)\} + \sum_{i=1}^n \sum_{j=1}^k q_{ij} \log q_{ij} \\
&= \sum_{j=1}^k \left[ -\frac{1}{2} \log |\Sigma_{\beta_j}^q| - \frac{1}{2} \log |\Sigma_{b_j}^q| + (\alpha_{b_j}^q + 1) \psi(\alpha_{b_j}^q) - \log \lambda_{b_j}^q - \log \Gamma(\alpha_{b_j}^q) \right. \\
&\quad \left. - \alpha_{b_j}^q + (\alpha_{a_j}^q + 1) \psi(\alpha_{a_j}^q) - \log \lambda_{a_j}^q - \log \Gamma(\alpha_{a_j}^q) - \alpha_{a_j}^q \right] + \sum_{i=1}^n \sum_{j=1}^k q_{ij} \log q_{ij} \\
&\quad + \sum_{j=1}^k \sum_{l=1}^g \left[ (\alpha_{jl}^q + 1) \psi(\alpha_{jl}^q) - \log \lambda_{jl}^q - \log \Gamma(\alpha_{jl}^q) - \alpha_{jl}^q \right] - \frac{1}{2} \sum_{i=1}^n \log |\Sigma_{a_i}^q| \\
&\quad - \frac{k(p+s_2)+ns_1}{2} \{\log(2\pi) + 1\}
\end{aligned} \tag{2}$$

and putting (1) and (2) together gives the lower bound for Algorithm 1.

## B Algorithm 2 (partial centering when $X_i = W_i$ )

(Updates in steps 4 and 7 remain the same as in Algorithm 1 with  $s_1 = p$ .) Initialize:  $q_{ij}$  for  $i = 1, \dots, n$ ,  $j = 1, \dots, k$ ,  $\mu_{b_j}^q$ ,  $\mu_{\beta_j}^q$ ,  $\frac{\alpha_{a_j}^q}{\lambda_{a_j}^q}$  and  $\frac{\alpha_{b_j}^q}{\lambda_{b_j}^q}$  for  $j = 1, \dots, k$ ,  $\frac{\alpha_{jl}^q}{\lambda_{jl}^q}$  for  $j = 1, \dots, k$ ,  $l = 1, \dots, g$ . Do until the change in the lower bound between iterations is less than a tolerance:

1. For  $i = 1, \dots, n$ ,

$$\begin{aligned}
\Sigma_{\eta_i}^q &\leftarrow \left\{ \sum_{j=1}^k q_{ij} \frac{\alpha_{a_j}^q}{\lambda_{a_j}^q} I_p + X_i^T (\sum_{j=1}^k q_{ij} \Sigma_{ij}^{q-1}) X_i \right\}^{-1}, \\
\mu_{\eta_i}^q &\leftarrow \Sigma_{\eta_i}^q \sum_{j=1}^k q_{ij} \left\{ \frac{\alpha_{a_j}^q}{\lambda_{a_j}^q} \mu_{\beta_j}^q + X_i^T \Sigma_{ij}^{q-1} (y_i - V_i \mu_{b_j}^q) \right\}.
\end{aligned}$$

2. For  $j = 1, \dots, k$ ,

$$\begin{aligned}
\Sigma_{\beta_j}^q &\leftarrow \left( \Sigma_{\beta_j}^{-1} + \frac{\alpha_{a_j}^q}{\lambda_{a_j}^q} \sum_{i=1}^n q_{ij} I_p \right)^{-1}, \\
\mu_{\beta_j}^q &\leftarrow \Sigma_{\beta_j}^q \frac{\alpha_{a_j}^q}{\lambda_{a_j}^q} \sum_{i=1}^n q_{ij} \mu_{\eta_i}^q.
\end{aligned}$$

3. For  $j = 1, \dots, k$ ,

$$\begin{aligned}
\Sigma_{b_j}^q &\leftarrow \left( \frac{\alpha_{b_j}^q}{\lambda_{b_j}^q} I_{s_2} + \sum_{i=1}^n q_{ij} V_i^T \Sigma_{ij}^{q-1} V_i \right)^{-1}, \\
\mu_{b_j}^q &\leftarrow \Sigma_{b_j}^q \sum_{i=1}^n q_{ij} V_i^T \Sigma_{ij}^{q-1} (y_i - X_i \mu_{\eta_i}^q).
\end{aligned}$$

5. For  $i = 1, \dots, n$ ,  $j = 1, \dots, k$ ,  $q_{ij} \leftarrow \frac{p_{ij} \exp(c_{ij})}{\sum_{l=1}^k p_{il} \exp(c_{il})}$ , where

$$c_{ij} = \frac{p}{2}(\psi(\alpha_{a_j}^q) - \log \lambda_{a_j}^q) - \frac{1}{2} \left[ \omega_{ij}^T \Sigma_{ij}^{q-1} \omega_{ij} + \text{tr} \left\{ \Sigma_{ij}^{q-1} (X_i \Sigma_{\beta_j}^q X_i^T + V_i \Sigma_{b_j}^q V_i^T) \right\} \right] \\ + \sum_{l=1}^g \frac{\kappa_{il}}{2} \left\{ \psi(\alpha_{jl}^q) - \log \lambda_{jl}^q \right\} - \frac{\alpha_{a_j}^q}{2\lambda_{a_j}^q} \left\{ (\mu_{\eta_i}^q - \mu_{\beta_j}^q)^T (\mu_{\eta_i}^q - \mu_{\beta_j}^q) + \text{tr}(\Sigma_{\eta_i}^q + \Sigma_{\beta_j}^q) \right\}.$$

6. For  $j = 1, \dots, k$ ,

$$\alpha_{a_j}^q \leftarrow \alpha_{a_j} + \frac{p}{2} \sum_{i=1}^n q_{ij}, \\ \lambda_{a_j}^q \leftarrow \lambda_{a_j} + \frac{1}{2} \sum_{i=1}^n q_{ij} \{ (\mu_{\eta_i}^q - \mu_{\beta_j}^q)^T (\mu_{\eta_i}^q - \mu_{\beta_j}^q) + \text{tr}(\Sigma_{\eta_i}^q + \Sigma_{\beta_j}^q) \}.$$

8. For  $j = 1, \dots, k$ ,  $l = 1, \dots, g$ ,

$$\alpha_{jl}^q \leftarrow \alpha_{jl} + \frac{1}{2} \sum_{i=1}^n q_{ij} \kappa_{il}, \\ \lambda_{jl}^q \leftarrow \lambda_{jl} + \frac{1}{2} \sum_{i=1}^n q_{ij} \left\{ (\omega_{ij})_{\kappa_{il}}^T (\omega_{ij})_{\kappa_{il}} + \text{tr}(X_i \Sigma_{\eta_i}^q X_i^T + V_i \Sigma_{b_j}^q V_i^T)_{\kappa_{il}} \right\}, \\ \text{where } \omega_{ij} = y_i - X_i \mu_{\eta_i}^q - V_i \mu_{b_j}^q.$$

The variational lower bound is given by

$$\frac{1}{2} \sum_{j=1}^k \left[ \log |\Sigma_{\beta_j}^{-1} \Sigma_{\beta_j}^q| - \text{tr}(\Sigma_{\beta_j}^{-1} \Sigma_{\beta_j}^q) - \mu_{\beta_j}^{qT} \Sigma_{\beta_j}^{-1} \mu_{\beta_j}^q + \log |\Sigma_{b_j}^q| - \frac{\alpha_{b_j}^q}{\lambda_{b_j}^q} \left\{ \mu_{b_j}^{qT} \mu_{b_j}^q + \text{tr}(\Sigma_{b_j}^q) \right\} \right] \\ + \sum_{j=1}^k \left[ \alpha_{a_j} \log \frac{\lambda_{a_j}}{\lambda_{a_j}^q} + \log \frac{\Gamma(\alpha_{a_j}^q)}{\Gamma(\alpha_{a_j})} + \frac{p \sum_{i=1}^n q_{ij}}{2} \left\{ \psi(\alpha_{a_j}^q) - \log \lambda_{a_j}^q \right\} + \psi(\alpha_{a_j}^q)(\alpha_{a_j} - \alpha_{a_j}^q) + \alpha_{a_j}^q \right. \\ \left. - \frac{\lambda_{a_j} \alpha_{a_j}^q}{\lambda_{a_j}^q} + \alpha_{b_j} \log \frac{\lambda_{b_j}}{\lambda_{b_j}^q} + \log \frac{\Gamma(\alpha_{b_j}^q)}{\Gamma(\alpha_{b_j})} - \frac{\lambda_{b_j} \alpha_{b_j}^q}{\lambda_{b_j}^q} - \frac{s_2}{2} \log \lambda_{b_j}^q + \alpha_{b_j}^q \right] + \sum_{j=1}^k \sum_{l=1}^g \left[ \alpha_{jl} \log \frac{\lambda_{jl}}{\lambda_{jl}^q} + \alpha_{jl}^q \right. \\ \left. + \log \frac{\Gamma(\alpha_{jl}^q)}{\Gamma(\alpha_{jl})} + \frac{\sum_{i=1}^n \kappa_{il} q_{ij}}{2} \left\{ \psi(\alpha_{jl}^q) - \log \lambda_{jl}^q \right\} + \psi(\alpha_{jl}^q)(\alpha_{jl} - \alpha_{jl}^q) - \frac{\lambda_{jl} \alpha_{jl}^q}{\lambda_{jl}^q} \right] + \frac{1}{2} \sum_{i=1}^n \log |\Sigma_{\eta_i}^q| \\ - \sum_{i=1}^n \sum_{j=1}^k \frac{q_{ij}}{2} \left[ \omega_{ij}^T \Sigma_{ij}^{q-1} \omega_{ij} + \text{tr} \left\{ \Sigma_{ij}^{q-1} (X_i \Sigma_{\eta_i}^q X_i^T + V_i \Sigma_{b_j}^q V_i^T) \right\} + \frac{\alpha_{a_j}^q}{\lambda_{a_j}^q} \left\{ \text{tr}(\Sigma_{\eta_i}^q + \Sigma_{\beta_j}^q) + \right. \right. \\ \left. \left. (\mu_{\eta_i}^q - \mu_{\beta_j}^q)^T (\mu_{\eta_i}^q - \mu_{\beta_j}^q) \right\} \right] + \sum_{i=1}^n \sum_{j=1}^k q_{ij} \log \frac{p_{ij}}{q_{ij}} + \frac{k(p+s_2)+np-N \log(2\pi)}{2} + \log p(\mu_{\delta}^q)$$

where  $\omega_{ij} = y_i - X_i \mu_{\eta_i}^q - V_i \mu_{b_j}^q$ . In the examples, when Algorithm 2 is used in conjunction with the VGA to fit a 1-component mixture ( $j = 1$ ), we set  $q_{ij} = 1$  for  $i = 1, \dots, n$ ,  $\frac{\alpha_{jl}^q}{\lambda_{jl}^q} = 1$  for  $l = 1, \dots, g$ ,  $\frac{\alpha_{b_j}^q}{\lambda_{b_j}^q} = 1$ ,  $\frac{\alpha_{a_j}^q}{\lambda_{a_j}^q} = 0.1$ ,  $\mu_{b_j}^q = 0$  and  $\mu_{\beta_j}^q = 0$  for initialization.

### C Algorithm 3 (full centering when $X_i = W_i = V_i$ )

(Updates in step 4 remains the same as in algorithm 1.) Initialize:  $q_{ij}$  for  $i = 1, \dots, n$ ,  $j = 1, \dots, k$ ,  $\mu_{\nu_j}^q$ ,  $\mu_{\beta_j}^q$ ,  $\frac{\alpha_{a_j}^q}{\lambda_{a_j}^q}$  and  $\frac{\alpha_{b_j}^q}{\lambda_{b_j}^q}$  for  $j = 1, \dots, k$ ,  $\frac{\alpha_{jl}^q}{\lambda_{jl}^q}$  for  $j = 1, \dots, k$ ,  $l = 1, \dots, g$ . Do until the change in the lower bound between iterations is less than a tolerance:

1. For  $i = 1, \dots, n$ ,

$$\Sigma_{\rho_i}^q \leftarrow \left\{ \sum_{j=1}^k q_{ij} \frac{\alpha_{a_j}^q}{\lambda_{a_j}^q} I_p + X_i^T (\sum_{j=1}^k q_{ij} \Sigma_{ij}^{q-1}) X_i \right\}^{-1},$$

$$\mu_{\rho_i}^q \leftarrow \Sigma_{\rho_i}^q \sum_{j=1}^k q_{ij} \left( \frac{\alpha_{a_j}^q}{\lambda_{a_j}^q} \mu_{\nu_j}^q + X_i^T \Sigma_{ij}^{q-1} y_i \right).$$

2. For  $j = 1, \dots, k$ ,

$$\Sigma_{\nu_j}^q \leftarrow \left\{ \left( \frac{\alpha_{b_j}^q}{\lambda_{b_j}^q} + \frac{\alpha_{a_j}^q}{\lambda_{a_j}^q} \sum_{i=1}^n q_{ij} \right) I_p \right\}^{-1},$$

$$\mu_{\nu_j}^q \leftarrow \Sigma_{\nu_j}^q \left( \frac{\alpha_{b_j}^q}{\lambda_{b_j}^q} \mu_{\beta_j}^q + \frac{\alpha_{a_j}^q}{\lambda_{a_j}^q} \sum_{i=1}^n q_{ij} \mu_{\rho_i}^q \right).$$

3. For  $j = 1, \dots, k$ ,

$$\Sigma_{\beta_j}^q \leftarrow \left( \Sigma_{\beta_j}^{-1} + \frac{\alpha_{b_j}^q}{\lambda_{b_j}^q} I_p \right)^{-1},$$

$$\mu_{\beta_j}^q \leftarrow \Sigma_{\beta_j}^q \frac{\alpha_{b_j}^q}{\lambda_{b_j}^q} \mu_{\nu_j}^q.$$

5. For  $i = 1, \dots, n$ ,  $j = 1, \dots, k$ ,  $q_{ij} \leftarrow \frac{p_{ij} \exp(c_{ij})}{\sum_{l=1}^k p_{il} \exp(c_{il})}$ , where

$$\begin{aligned} c_{ij} = & \frac{p}{2} \left\{ \psi(\alpha_{a_j}^q) - \log \lambda_{a_j}^q \right\} - \frac{1}{2} \left\{ (y_i - X_i \mu_{\rho_i}^q)^T \Sigma_{ij}^{q-1} (y_i - X_i \mu_{\rho_i}^q) + \text{tr}(\Sigma_{ij}^{q-1} X_i \Sigma_{\rho_i}^q X_i^T) \right\} \\ & - \frac{\alpha_{a_j}^q}{2\lambda_{a_j}^q} \left\{ (\mu_{\rho_i}^q - \mu_{\nu_j}^q)^T (\mu_{\rho_i}^q - \mu_{\nu_j}^q) + \text{tr}(\Sigma_{\rho_i}^q + \Sigma_{\nu_j}^q) \right\} + \sum_{l=1}^g \frac{\kappa_{il}}{2} \left\{ \psi(\alpha_{jl}^q) - \log \lambda_{jl}^q \right\}. \end{aligned}$$

6. For  $j = 1, \dots, k$ ,

$$\alpha_{a_j}^q \leftarrow \alpha_{a_j} + \frac{p}{2} \sum_{i=1}^n q_{ij},$$

$$\lambda_{a_j}^q \leftarrow \lambda_{a_j} + \frac{1}{2} \sum_{i=1}^n q_{ij} \left\{ (\mu_{\rho_i}^q - \mu_{\nu_j}^q)^T (\mu_{\rho_i}^q - \mu_{\nu_j}^q) + \text{tr}(\Sigma_{\rho_i}^q + \Sigma_{\nu_j}^q) \right\}.$$

7. For  $j = 1, \dots, k$ ,

$$\alpha_{b_j}^q \leftarrow \alpha_{b_j} + \frac{p}{2},$$

$$\lambda_{b_j}^q \leftarrow \lambda_{b_j} + \frac{1}{2} \left\{ (\mu_{\nu_j}^q - \mu_{\beta_j}^q)^T (\mu_{\nu_j}^q - \mu_{\beta_j}^q) + \text{tr}(\Sigma_{\nu_j}^q + \Sigma_{\beta_j}^q) \right\}.$$

8. For  $j = 1, \dots, k$ ,  $l = 1, \dots, g$ ,

$$\alpha_{jl}^q \leftarrow \alpha_{jl} + \frac{1}{2} \sum_{i=1}^n q_{ij} \kappa_{il},$$

$$\lambda_{jl}^q \leftarrow \lambda_{jl} + \frac{1}{2} \sum_{i=1}^n q_{ij} \left\{ (y_i - X_i \mu_{\rho_i}^q)^T_{\kappa_{il}} (y_i - X_i \mu_{\rho_i}^q)_{\kappa_{il}} + \text{tr}(X_i \Sigma_{\rho_i}^q X_i^T)_{\kappa_{il}} \right\}.$$

The variational lower bound is given by

$$\begin{aligned}
& \sum_{j=1}^k \left[ \frac{1}{2} \log |\Sigma_{\nu_j}^q| - \frac{1}{2} \log |\Sigma_{\beta_j}^{-1} \Sigma_{\beta_j}^q| - \frac{1}{2} \text{tr}(\Sigma_{\beta_j}^{-1} \Sigma_{\beta_j}^q) - \frac{1}{2} \mu_{\beta_j}^{qT} \Sigma_{\beta_j}^{-1} \mu_{\beta_j}^q + \alpha_{a_j} \log \frac{\lambda_{a_j}}{\lambda_{a_j}^q} + \log \frac{\Gamma(\alpha_{a_j}^q)}{\Gamma(\alpha_{a_j})} \right. \\
& + \alpha_{a_j}^q + \frac{p \sum_{i=1}^n q_{ij}}{2} \left\{ \psi(\alpha_{a_j}^q) - \log \lambda_{a_j}^q \right\} + \psi(\alpha_{a_j}^q)(\alpha_{a_j} - \alpha_{a_j}^q) - \frac{\lambda_{a_j} \alpha_{a_j}^q}{\lambda_{a_j}^q} + \alpha_{b_j} \log \frac{\lambda_{b_j}}{\lambda_{b_j}^q} + \log \frac{\Gamma(\alpha_{b_j}^q)}{\Gamma(\alpha_{b_j})} \\
& - \frac{\lambda_{b_j} \alpha_{b_j}^q}{\lambda_{b_j}^q} - \frac{p}{2} \log \lambda_{b_j}^q + \alpha_{b_j}^q - \frac{\alpha_{b_j}^q}{2 \lambda_{b_j}^q} \left\{ (\mu_{\nu_j}^q - \mu_{\beta_j}^q)^T (\mu_{\nu_j}^q - \mu_{\beta_j}^q) + \text{tr}(\Sigma_{\nu_j}^q + \Sigma_{\beta_j}^q) \right\} \left. \right] + \frac{1}{2} \sum_{i=1}^n \log |\Sigma_{\rho_i}^q| \\
& + \sum_{j=1}^k \sum_{l=1}^g \left[ \alpha_{jl} \log \frac{\lambda_{jl}}{\lambda_{jl}^q} + \psi(\alpha_{jl}^q)(\alpha_{jl} - \alpha_{jl}^q) + \log \frac{\Gamma(\alpha_{jl}^q)}{\Gamma(\alpha_{jl})} + \frac{\sum_{i=1}^n \kappa_{il} q_{ij}}{2} \left\{ \psi(\alpha_{jl}^q) - \log \lambda_{jl}^q \right\} - \frac{\lambda_{jl} \alpha_{jl}^q}{\lambda_{jl}^q} \right. \\
& + \alpha_{jl}^q \left. \right] - \sum_{i=1}^n \sum_{j=1}^k \frac{q_{ij}}{2} \left[ \frac{\alpha_{a_j}^q}{\lambda_{a_j}^q} \left\{ \text{tr}(\Sigma_{\rho_i}^q + \Sigma_{\nu_j}^q) + (\mu_{\rho_i}^q - \mu_{\nu_j}^q)^T (\mu_{\rho_i}^q - \mu_{\nu_j}^q) \right\} + \text{tr} \left\{ \Sigma_{ij}^{q-1} (X_i \Sigma_{\rho_i}^q X_i^T) \right\} \right. \\
& + (y_i - X_i \mu_{\rho_i}^q)^T \Sigma_{ij}^{q-1} (y_i - X_i \mu_{\rho_i}^q) \left. \right] + \log p(\mu_{\delta}^q) + \sum_{i=1}^n \sum_{j=1}^k q_{ij} \log \frac{p_{ij}}{q_{ij}} + \frac{p(2k+n) - N \log(2\pi)}{2}.
\end{aligned}$$

In the examples, when Algorithm 3 is used in conjunction with the VGA to fit a 1-component mixture ( $j = 1$ ), we set  $q_{ij} = 1$  for  $i = 1, \dots, n$ ,  $\frac{\alpha_{jl}^q}{\lambda_{jl}^q} = 10$  for  $l = 1, \dots, g$ ,  $\frac{\alpha_{a_j}^q}{\lambda_{a_j}^q} = 0.1$ ,  $\frac{\alpha_{b_j}^q}{\lambda_{b_j}^q} = 0.01$ ,  $\mu_{\beta_j}^q = 0$  and  $\mu_{\nu_j}^q = 0$  for initialization. We note that the rate of convergence of Algorithm 3 can be sensitive to the initialization of  $\frac{\alpha_{jl}^q}{\lambda_{jl}^q}$ ,  $\frac{\alpha_{a_j}^q}{\lambda_{a_j}^q}$  and  $\frac{\alpha_{b_j}^q}{\lambda_{b_j}^q}$  and observed that an initialization satisfying  $\frac{\alpha_{b_j}^q}{\lambda_{b_j}^q} < \frac{\alpha_{a_j}^q}{\lambda_{a_j}^q} < \frac{\alpha_{jl}^q}{\lambda_{jl}^q}$  works better.

## D Example on clustering of yeast galactose data

The yeast galactose data of Ideker *et al.* (2001) has four replicate hybridizations for each of 20 cDNA array experiments. We consider a subset of 205 genes previously analyzed by Yeung *et al.* (2003) and Ng *et al.* (2006) whose expression patterns reflect four functional categories in the Gene Ontology (GO) listings (Ashburner *et al.*, 2000). Approximately 8% of the data are missing and Yeung *et al.* (2003) used a  $k$ -nearest neighbour method to impute the missing data values. Yeung *et al.* (2003) and Ng *et al.* (2006) evaluated the performance of their clustering algorithms by how closely the clusters compared with the four categories in the GO listings. They used the adjusted Rand index (Hubert and Arabie, 1985) to assess the degree of agreement between their partitions and the four functional categories.

We use this example to illustrate the way that our model can make use of covariates in the

mixing weights, unlike previous analyses of this data set. In particular, we use the GO listings as covariates in the mixture weights. Let  $u_i$  be a vector of length  $d = 4$  where the  $l$ th element is 1 if the functional category of gene  $i$  is  $l$  and 0 otherwise. Instead of looking at the data with missing values imputed by the  $k$ -nearest neighbour method, we consider the original data containing 8% missing values, since our model has the capability to handle missing data. This data set can be accessed from <http://expression.washington.edu/publications/kayee/yeunggb2003/gal205.txt>. Taking  $n = 205$  genes, let  $y_{itr}$  denote the  $r$ th repetition of the expression profile for gene  $i$  at experiment  $t$ ,  $0 \leq r \leq 4$ , and  $R_{it}$  denote the number of replicate hybridizations data available for gene  $i$  in experiment  $t$ ,  $i = 1, \dots, 205$ ,  $t = 1, \dots, 20$ . For each  $i = 1, \dots, n$ ,  $y_i$  is a vector of  $n_i$  observations where  $n_i = \sum_{t=1}^{20} R_{it}$  and  $y_i = (y_{i11}, \dots, y_{i14}, \dots, y_{i,20,1}, \dots, y_{i,20,4})^T$ , with missing observations omitted.  $V_i$  is a  $n_i \times 80$  matrix obtained from  $I_{80}$  by removing the  $(tr)$ th row if the observation for experiment  $t$  at the  $r$ th repetition is not available.  $X_i$  is a  $n_i \times 20$  matrix,

$$X_i = \begin{bmatrix} 1_{R_{i1}} & 0_{R_{i1}} & \dots & 0_{R_{i1}} \\ 0_{R_{i2}} & 1_{R_{i2}} & \dots & 0_{R_{i2}} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{R_{i20}} & 0_{R_{i20}} & \dots & 1_{R_{i20}} \end{bmatrix}$$

and  $W_i = X_i$ . For the error terms, we set  $g = 20$  with  $\kappa_{il} = R_{il}$ ,  $i = 1, \dots, n$ ,  $l = 1, \dots, g$ , so that the error variance of each mixture component is allowed to vary between different experiments. We used the following priors,  $\delta \sim N(0, 1000I)$ ,  $\beta_j \sim N(0, 1000I)$  for  $j = 1, \dots, k$ , and  $IG(2, 0.12)$  for  $\sigma_{a_j}^2$ ,  $\sigma_{b_j}^2$ ,  $j = 1, \dots, k$  and  $\sigma_{jl}^2$ ,  $j = 1, \dots, k$ ,  $l = 1, \dots, g$ .

Applying VGA using Algorithm 2 (with partial centering) for five times, we obtained a 7-component mixture on all five trials with similar results. The clustering of a 7-component mixture with the highest estimated log marginal likelihood among the five trials is shown in Figure 1. Some merge moves such as merging cluster 1 with 2, cluster 4 with 7 or cluster 4 with 6 were considered but these did not result in a higher estimated log marginal likelihood. The same holds for the other 7-component mixtures. The number of optimal clusters obtained using VGA is the same as that reported in Ng *et al.* (2006) although there are slight differences in the clusterings. In particular, instead of having one cluster containing all the genes from Category 4, we observed that two or three of the genes in Category 4 were consistently separated from the cluster containing the remaining genes from Category 4. Fitted probabilities from the gating function are shown in Figure 2. These were obtained by substituting  $\delta$  with  $\mu_\delta^q$  from the variational posterior into  $P(z_i = j) = p_{ij} = \frac{\exp(u_i^T \delta_j)}{\sum_{l=1}^k \exp(u_i^T \delta_l)}$

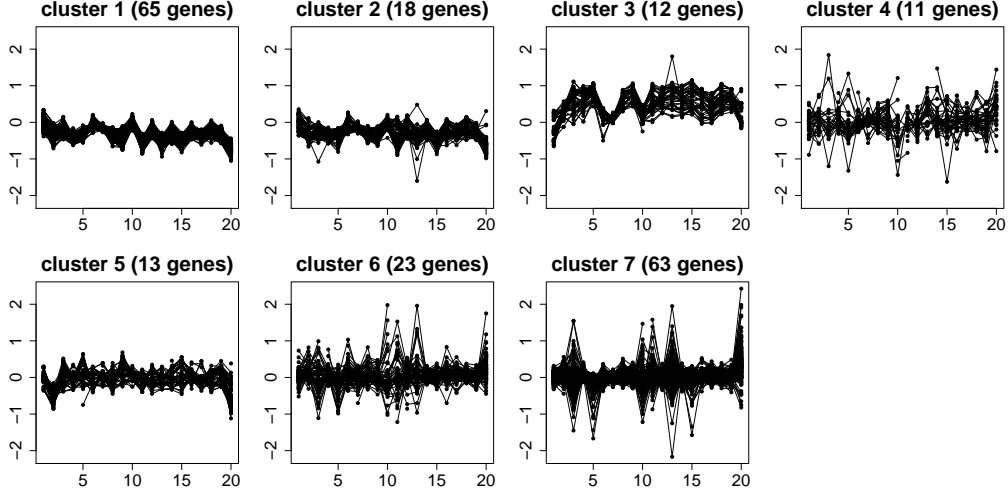


Figure 1: Clustering results for yeast galactose data obtained from VGA using Algorithm 2. The  $x$ -axis are the experiments and  $y$ -axis are the gene expression profiles. GO listings were used as covariates in the mixture weights.

which represents the probability that observation  $i$  belongs to component  $j$  of the mixture conditional on the category that observation  $i$  belongs to in the GO listings.

To investigate the impact of reparametrizing the model using hierarchical centering, we applied VGA using Algorithm 1 five times. This time, we obtained a 6-component mixture twice and a 7-component mixture thrice. The average estimated log marginal log likelihood attained by Algorithm 1 was 7901 which is lower than the average of 8201 attained by Algorithm 2. For fitting a 7-component model, VGA with Algorithm 1 took an average of 3418 seconds, while Algorithm 2 took an average of 1758 seconds. While these results may

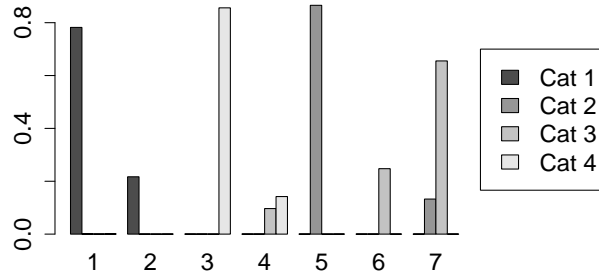


Figure 2: Fitted probabilities from gating function. The  $x$ -axis are the clusters and  $y$ -axis are the probabilities.



not be conclusive, the gain in efficiency in using Algorithm 2 over Algorithm 1 is clear. By using hierarchical centering, the computation time was reduced by nearly half in this example.

## References

- Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. *Nature Genetics*, 25, 25–29.
- Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2, 193–218.
- Ideker, T., Thorsson, V., Ranish, J.A., Christmas, R., Buhler, J., Eng, J.K., Bumgarner, R., Goodlett, D.R., Aebersold, R. and Hood, L. (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, 292, 929–934.
- Ng, S.K., McLachlan, G.J., Wang, K., Ben-Tovim Jones, L. and Ng, S.-W. (2006). A mixture model with random-effects components for clustering correlated gene-expression profiles. *Bioinformatics*, 22, 1745–1752.
- Yeung, K.Y., Medvedovic, M. and Bumgarner, R.E. (2003). Clustering gene-expression data with repeated measurements. *Genome Biology*, 4, Article R34.