S3 Appendix

A tangent based method.

The integrations required to obtain the lower bound of the marginal log-likelihood are non-trivial and no closed form equations for \( q(z, \alpha, \beta) \) exists. Additional variational parameters \( \mathbf{v} = (\mathbf{a}^T, \mathbf{b}^T)^T \) are now introduced such that a lower bound of the marginal log-likelihood is found using

\[
\ln p(y) \geq \sum_z \int q_v(z) q_v(\alpha) q_v(\beta) \ln \left( \frac{p_v(y, z, \alpha, \beta)}{q_v(z) q_v(\alpha) q_v(\beta)} \right) d\alpha d\beta. \tag{1}
\]

\( \mathbf{a} = (\mathbf{a}_1^T, \ldots, \mathbf{a}_n^T)^T \) where each of the \( \mathbf{a}_i \) vectors are of length \( K_i \) while \( \mathbf{b} \) is of length \( n \). \( p_v(y, z, \alpha, \beta) \) should be viewed as a lower bound for \( p(y, z, \alpha, \beta) \) while \( q_v(z) \), \( q_v(\alpha) \) and \( q_v(\beta) \) indicates that the distributions of \( z, \alpha \) and \( \beta \) are functions of \( \mathbf{v} \) which could have different functional forms.

The function \( f(x) = -\ln(1 + e^x) \) can be presented as the maximum of a family of parabolas \(^1\) where

\[
-\ln(1 + e^x) = \max_{\xi \in \mathbb{R}} \left( A(\xi) x^2 - \frac{1}{2} x + C(\xi) \right) \text{ for all } x \in \mathbb{R},
\]

\[
A(\xi) = -\tanh \left( \frac{1}{2} \xi \right) / (4\xi),
\]

\[
C(\xi) = \frac{1}{2} \xi - \ln(1 + e^\xi) + \xi \tanh \left( \frac{1}{2} \xi \right) / 4.
\]

A lower bound for the quantities \( b(W\alpha) \) and \( b(X\beta) \) are obtained by using the above result such
that

\[
\ln p \geq y^T \text{diag}(\tilde{z})W\alpha + z^T X\beta + \ln \pi(\alpha, \beta) + \tilde{z}^T \left( A(a) \odot (W\alpha)^2 - \frac{1}{2}W\alpha + C(a) \right) + 1^T_n \left( A(b) \odot (X\beta)^2 - \frac{1}{2}X\beta + C(b) \right) \\
\geq -\frac{1}{2} \alpha^T B_1 \alpha + B_2 \alpha - \frac{1}{2} \beta^T D_1 \beta + D_2 \beta + \tilde{z}^T C(a) + 1^T_n C(b) + E. \tag{2}
\]

From equation (2) it is apparent that both \( q_{\nu}(\alpha) \) and \( q_{\nu}(\beta) \) are multivariate Gaussian distributions; specifically \( q_{\nu}(\alpha) \sim N(B_1^{-1}B_2^T, B_1^{-1}) \) and \( q_{\nu}(\beta) \sim N(D_1^{-1}D_2^T, D_1^{-1}) \). Using various known matrix
identities regarding multivariate Gaussian distributions it can be shown that

\[ c_i^{(T)} = x_i \mu_\beta + 1_k^T C(a_i) - \frac{1}{2} 1_k^T w_i \mu_\alpha + \text{tr} (d_i) \]  

(3)

where \( d_i = w_i^T \text{diag}(A(a_i)) w_i (\Sigma_\alpha + \mu_\alpha \mu_\alpha^T) \). One way of estimating the variational parameters is to numerically maximise the right hand side of equation (1) with respect to \( K \). This approach might be feasible although it is not attempted here. Similar to [1] we however devise an Tangent algorithm in order to estimate \( v \).

**Estimation of the variational parameters - The E-step**

Denote the ‘new variational parameters’ as \( a_{(N)} \) and \( b_{(N)} \) respectively. Further denote \( v^{(new)} \) as \( N \) and \( v^{(old)} \) as \( O \) for notational convenience below. Treating \( y, z, \alpha, \beta \) as the ‘complete data’, the E-step of the Tangent algorithm is found by calculating the conditional expectation of the right hand side of equation (2) which equals \( Q(N|O) = T_{\alpha,N} + T_{\beta,N} + E \) where

\[ T_{\alpha,N} = \text{tr} \left( -\frac{1}{2} B_{1,N} (\Sigma_{\alpha,O} + \mu_{\alpha,O} \mu_{\alpha,O}^T) \right) + B_{2} \mu_{\alpha,O} + B_{3,N} \]
\[ T_{\beta,N} = \text{tr} \left( -\frac{1}{2} D_{1,N} (\Sigma_{\beta,O} + \mu_{\beta,O} \mu_{\beta,O}^T) \right) + D_{2} \mu_{\beta,O} + D_{3,N}. \]

where

\[ B_{3,N} = \tilde{p}^T C(a_N) - \frac{1}{2} (\mu_\alpha^0)^T (\Sigma_\alpha^0)^{-1} (\mu_\alpha^0) \]
\[ D_{3,N} = 1_n^T C(b_N) - \frac{1}{2} (\mu_\beta^0)^T (\Sigma_\beta^0)^{-1} (\mu_\beta^0). \]
Note that here the additional subscript $N$ indicates the dependence on the ‘new variational parameters’ which are estimated in the M-step, while subscript $O$ indicates the dependence on the ‘old variational parameters’.

**Estimation of the variational parameters - The M-step**

Since $Q(N|O)$ separates in two functions of which one only depends on $a_{(N)}$ and the second only depends on $b_{(N)}$, it can be shown that

\[
(a_{(N)})^2 = \text{diagonal} \left( W \left( \Sigma_{a,O} + \mu_{a,O} \mu_{a,O}^T \right) W^T \right) \quad (4)
\]

\[
(b_{(N)})^2 = \text{diagonal} \left( X \left( \Sigma_{\beta,O} + \mu_{\beta,O} \mu_{\beta,O}^T \right) X^T \right) \quad (5)
\]

**References**