1 Setup

Consider a corpus of $N$ documents with a vocabulary of length $D$. Set $K$ to be large, where $K$ is the maximum number of topics that will be allowed (the algorithm may elect to use fewer than $K$ topics). The documents have document-term matrix $X$ and observed outcomes $Y$.

Let $\pi$ be a $K$-vector. Let $X$ be $N \times D$ and $X_n$ be the $n$th row of $X$. Let $Z$ be an $N \times K$ binary matrix. Let $Z_i$ be the $i$th row of $Z$ and $z_{i,k}$ be the $i$th element of the $k$th column of $Z$. Let $A$ be $K \times D$ and $A_k$ be the $k$th row of $A$. Let $Y$ be and $N$-vector. Let $\beta$ be a $K$-vector.

$$
\pi_k \sim \text{Stick-Breaking}(\alpha)
$$

$$
X_i | Z_i, A \sim \text{MVN}(Z_i A, \sigma_A^2 I_D)
$$

$$
A_k \sim \text{MVN}(0, \sigma_A^2 I_D)
$$

$$
Y_i | Z_i, \beta \sim N(Z_i \beta, \tau^{-1})
$$

$$
\tau \sim \text{Gamma}(a, b)
$$

$$
\beta | \tau \sim \text{MVN}(0, \tau^{-1}I_K)
$$

Stick-breaking is performed by generating $\eta_k \sim \text{Beta}(\alpha, 1)$ for $k = 1, \ldots, K$ and $\pi_k = \prod_{m=1}^{K} \eta_k$.

$Z$ can be interpreted as a topic matrix (where each topic is either present or absent, and a document can have arbitrarily many topics). $A$ maps topics onto word counts, and $\beta$ maps topics onto the response.

2 Approximation

$$
q_{\lambda_k}(\pi_k) = \text{Beta}(\pi_k; \lambda_{k,1}, \lambda_{k,2})
$$

$$
q_{\phi_k, \Phi_k}(A_k) = \text{Normal}(A_k; \bar{\phi}_k, \Phi_k)
$$

$$
q_{\nu_{n,k}}(z_{n,k}) = \text{Bernoulli}(z_{n,k}; \nu_{n,k})
$$

$$
q_{m,s,c,d}(\beta, \tau) = \text{MVN}(\beta; m, S) \times \text{Gamma}(&c, d)
$$

For notational convenience, let $W = \{\pi, Z, A, \beta, \tau\}$ and $\theta = \{\alpha, \sigma_A^2, \sigma_X^2, a, b\}$.

Consider the problem of computing the log posterior.

$$
\log p(W|X, Y, \theta) = \log p(W, X, Y|\theta) - \log p(X, Y|\theta)
$$

This is difficult because $\log p(X, Y|\theta) = \log \int p(X, Y, W|\theta) dW$ is intractable. We therefore approximate with the distribution $q$,

$$
q(W) = q_{\lambda}(\pi)q_{\phi, \Phi}(A)q_{\nu}(Z)q_{m,s,c,d}(\beta, \tau)
$$

$$
D(q||p) = \arg\max_{\tau, \phi, \nu, m, S, c, d} E_q[\log(p(X, Y, W|\theta))] + H[q]
$$
\[ p_K(W, X, Y|\theta) = p(\tau|a, b) \prod_{k=1}^{K} \left( p(\pi_k|\alpha)p(\sigma^2_\lambda)p(\beta_k|\tau) \prod_{i=1}^{N} p(z_{i,k}|\pi_k) \right) \prod_{i=1}^{N} p(X_i|Z_i, A, \sigma^2_X) p(Y_i|Z_i, \beta, \tau) \]

3 Parameter Updates

In this section, expectations are taken with respect to \( q \).

3.1 Updating \( \phi, \Phi \)

\[ E_{A_{-k}, Z}[(X_i - Z_iA)(X_i - Z_iA)^T] \propto -2A_k E_{A_{-k}, Z}[z_{i,k}(X_i - Z_iA)] - A_k E_{Z}[z_{i,k}]A_k^T \]

\[ \propto -2A_k \left[ E[z_{i,k}] \left( X_i - \sum_{l=1}^{K} E_{A_{-k}, Z}[z_{i,k}z_{i,l}A_l] \right) \right] - A_k \nu_{i,k} A_k^T \]

\[ \propto -2A_k \left( \nu_{i,k} \left( X_i - \sum_{l:k \neq l} \nu_{i,l} \bar{\phi}_l \right) \right) + A_k \nu_{i,k} A_k^T \]

\[ \log q_{\phi_k}(A_k) \propto E_{A_{-k}, Z}[\log p_K(W, X|\theta)] \]

\[ \propto E_{A_{-k}, Z}[\log p_K(A_k|\sigma^2_\lambda) + \sum_{i=1}^{N} \log p_K(X_i|Z_i, A, \sigma^2_X)] \]

\[ \propto \frac{1}{2\sigma^2_\lambda} A_k A_k^T - \frac{1}{2\sigma^2_X} \sum_{i=1}^{N} E_{A_{-k}, Z}[(X_i - Z_iA)(X_i - Z_iA)^T] \]

\[ \propto \frac{1}{2} \left[ A_k \left( \frac{1}{\sigma^2_A} + \sum_{i=1}^{N} \frac{\nu_{i,k}}{\sigma^2_n} \right) A_k^T - 2A_k \left( \frac{1}{\sigma^2_X} \sum_{i=1}^{N} \nu_{i,k} \left( X_i - \sum_{l:k \neq l} \nu_{i,l} \bar{\phi}_l \right) \right) \right] \]

This is the kernel of a Normal distribution.

\[ A_k \sim MVN(\bar{\phi}_k, \Phi_k) \]

\[ \bar{\phi}_k = \left[ \frac{1}{\sigma^2_X} \sum_{i=1}^{N} \nu_{i,k} \left( X_i - \sum_{l:k \neq l} \nu_{i,l} \bar{\phi}_l \right) \right] \left( \frac{1}{\sigma^2_A} + \sum_{i=1}^{N} \frac{\nu_{i,k}}{\sigma^2_X} \right)^{-1} \]

\[ \Phi_k = \left( \frac{1}{\sigma^2_A} + \sum_{i=1}^{N} \frac{\nu_{i,k}}{\sigma^2_X} \right)^{-1} I \]
3.2 Updating \( m, S, c, \text{ and } d \)

\[
\log q_{m, S, c, d}(\beta, \tau) \propto \log \mathbb{E}_Z[\log p_K(W, X|\theta)] \\
\propto \mathbb{E}_Z[\log p_K(Y_i|Z_i, \beta, \tau)] + \log p_K(\beta|\tau) + \log p_K(\tau) \\
\propto \mathbb{E}_Z \left[ \left( \frac{\tau}{2\pi} \right)^{N^2/2} \exp \left( -\frac{\tau \sum_{i=1}^{N} (Y_i - Z_i \beta)^2}{2} \right) \right] + \log \left( \left( \frac{\tau}{2\pi} \right)^{K/2} \exp \left( -\frac{\tau \beta^T \beta}{2} \right) \right) + \log \left( \frac{b^a \tau^{a-1} e^{-br}}{\Gamma(a)} \right) \\
\propto \frac{N}{2} \log \tau - \frac{\tau \sum_{i=1}^{N} \mathbb{E}_Z[(Y_i - Z_i \beta)^2]}{2} + \frac{K}{2} \log \tau - \frac{\tau \beta^T \beta}{2} + (a - 1) \log \tau - b \tau \\
\propto -\frac{\tau}{2} \left[ \beta^T (\mathbb{E}[Z^T Z] + I_K) \beta - 2\beta^T \mathbb{E}[Z^T] Y \right] + (a + \frac{N}{2} + K) \log \tau - \left( b + \frac{Y^T Y}{2} \right) \tau \\
\propto \frac{K}{2} \log \tau - \frac{\tau}{2} \left( \beta^T \gamma^{-1} \beta - 2\beta^T \gamma^{-1} \gamma \mathbb{E}[Z^T] Y + \mathbb{E}[Z^T] \gamma^{-1} \gamma \mathbb{E}[Z^T] Y \right) \\
+ (a + \frac{N}{2} + K) \log \tau - \left( b + \frac{Y^T Y - Y^T \mathbb{E}[Z] \mathbb{E}[Z^T] Y}{2} \right) \tau \\
\propto \frac{K}{2} \log \tau - \frac{1}{2} (\beta - \gamma \mathbb{E}[Z^T] Y)^T \gamma^{-1} (\beta - \gamma \mathbb{E}[Z^T] Y) + (a + \frac{N}{2} + K) \log \tau - \left( b + \frac{Y^T Y - Y^T \mathbb{E}[Z] \mathbb{E}[Z^T] Y}{2} \right) \tau \\
\]

Here, \( \gamma^{-1} = \mathbb{E}[Z^T Z] + I_K \).

\( \beta \sim MVN(m, \tau^{-1} S) \)

\( \tau \sim \text{Gamma}(c, d) \)

\( m = SE[Z^T] Y \)

\( S = (\mathbb{E}[Z^T Z] + I_K)^{-1} \)

\( c = a + \frac{N}{2} \)

\( d = \frac{b + \frac{Y^T Y - Y^T \mathbb{E}[Z] \mathbb{E}[Z^T] Y}{2}}{2} \)

Updating \( \nu \)

\[
\log q_{\nu, i}(z_{i,k}) \propto \mathbb{E}_{\pi_i, A_i, Z_{-i,k}, \beta, \tau}[\log p_K(W, X, Y|\theta)] \\
\propto \mathbb{E}_{\pi_i, A_i, Z_{-i,k}, \beta, \tau}[\log p(z_{i,k}|A_i) + \log p(X_i|Z_i, A, \sigma^2_A) + \log p(Y_i|Z_i, \beta, \tau)] \\
\]

\[
\mathbb{E}_{\pi_i, Z_{-i,k}}[\log p(z_{i,k}|\pi_k)] = z_{i,k} \mathbb{E}[\log(\pi_k)] + (1 - z_{i,k}) \mathbb{E}[\log(1 - \pi_k)] \\
= z_{i,k} [\psi(\lambda_{k,1}) - \psi(\lambda_{k,1} + \lambda_{k,2})] + (1 - z_{i,k})[\psi(\lambda_{k,2}) - \psi(\lambda_{k,1} + \lambda_{k,2})] \\
= z_{i,k} [\psi(\lambda_{k,1}) - \psi(\lambda_{k,2})] + \psi(\lambda_{k,2}) - \psi(\lambda_{k,1} + \lambda_{k,2}) \\
\]

3
\[\mathbb{E}_{A,Z_{-ik}}[\log p(X_i|Z_i, A, \sigma^2_X)] \propto -\frac{1}{2\sigma^2_X} \mathbb{E}_{A,Z_{-ik}}[(X_i - Z_i A)(X_i - Z_i A)^T] \]
\[\propto -\frac{1}{2\sigma^2_X} \mathbb{E}_{A,Z_{-ik}}[-2Z_i A X_i^T + Z_i A A^T Z_i^T] \]
\[\propto -\frac{1}{2\sigma^2_X} \left[ -2z_{i,k} \bar{\phi}_k X_i^T + z_{i,k} (\text{tr}(\Phi_k) + \bar{\phi}_k \bar{\phi}_k^T) + 2z_{i,k} \bar{\phi}_k \left( \sum_{l \neq k} \nu_{i,l} \bar{\phi}_l^T \right) \right] \]

\[\mathbb{E}_{Z_{-ik},\beta,T}[\log p(Y_i|Z_i, \beta, \tau)] \propto \mathbb{E}_{Z_{-ik},\beta,T}[\frac{-\tau}{2} (Y_i - Z_i \beta)(Y_i - Z_i \beta)^T] \]
\[\propto \mathbb{E}_{Z_{-ik},\beta,T}[\frac{-\tau}{2} (-2Z_i \beta Y_i + Z_i |\beta| T Z_i^T)] \]
\[\propto -\frac{c}{2d} \left( -2z_{i,k} m_k Y_i + z_{i,k} \left( \frac{dS_k}{c - 1} + \bar{m}_k m_k \right) + 2z_{i,k} m_k \left( \sum_{l \neq k} \nu_{i,l} m_l \right) \right) \]

\[\log \frac{\nu_{i,k}}{1 - \nu_{i,k}} = \psi(\lambda_{i,1}) - \psi(\lambda_{i,2}) - \frac{1}{2\sigma^2_X} \left[ -2\bar{\phi}_k X_i^T + \text{tr}(\Phi_k) + \bar{\phi}_k \bar{\phi}_k^T + 2\bar{\phi}_k \left( \sum_{l \neq k} \nu_{i,l} \bar{\phi}_l^T \right) \right] \]
\[\propto -\frac{c}{2d} \left( -2m_k Y_i + \left( \frac{dS_k}{c - 1} + \bar{m}_k m_k \right) + 2m_k \left( \sum_{l \neq k} \nu_{i,l} m_l \right) \right) \]
\[\equiv v_{i,k} \]

\[\nu_{i,k} = \frac{1}{1 + \exp(-v_{i,k})} \]

\[\log \left( z_{i,k}^{\nu_{i,k}} (1 - z_{i,k})^{1 - \nu_{i,k}} \right) = z_{i,k} \log \left( \frac{1}{1 + \exp(-v_{i,k})} \right) + (1 - z_{i,k}) \log \left( \frac{\exp(-v_{i,k})}{1 + \exp(-v_{i,k})} \right) \]
\[= -z_{i,k} \log(1 + \exp(-v_{i,k})) + z_{i,k} \log(1 + \exp(-v_{i,k})) - z_{i,k} \log(\exp(-v)) + \log \left( \frac{\exp(-v_{i,k})}{1 + \exp(-v_{i,k})} \right) \]
\[\propto z_{i,k} v_{i,k} \]

So this is a Bernoulli kernel.

\[z_{i,k} \sim \text{Bernoulli} (\nu_{i,k}) \]
\[\nu_{i,k} = \frac{1}{1 + \exp(-v)} \]

So this is a Bernoulli kernel.
Updating $\lambda$

$$
\log q_{\lambda_k}(\pi_k) \propto E_{A,Z}[\log p_K(W, X, Y|\theta)] \\
\propto E_{A,Z} \left[ \log p_K(\pi_k|\alpha) + \sum_{i=1}^{N} \log p_K(z_{i,k}|\pi_k) \right] \\
= \left( \frac{\alpha}{K} - 1 \right) \log \pi_k + \sum_{i=1}^{N} (\nu_{i,k} \log \pi_k + (1 - \nu_{i,k}) \log (1 - \pi_k))
$$

This is a Beta kernel.

$$
\pi_k \sim \text{Beta}(\lambda_{k,1}, \lambda_{k,2}) \\
\lambda_{k,1} = \frac{\alpha}{K} + \sum_{i=1}^{N} \nu_{i,k} \\
\lambda_{k,2} = 1 + \sum_{i=1}^{N} (1 - \nu_{i,k})
$$