Derivation of Supervised Indian Buffet Process

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 $\mathbf{X}$ and observed outcomes $\mathbf{Y}$.

Let $\boldsymbol{\pi}$ be a K-vector. Let $\mathbf{X}$ be $N \times D$ and $\mathbf{X}_n$ be the nth row of $\mathbf{X}$. Let $\mathbf{Z}$ be an $N \times K$ binary matrix. Let $\mathbf{Z}_i$ be the ith row of $\mathbf{Z}$ and $z_{i,k}$ be the ith element of the kth column of $\mathbf{Z}$. Let $\mathbf{A}$ be $K \times D$ and $\mathbf{A}_k$ be the kth row of $\mathbf{A}$. Let $\mathbf{Y}$ be and N-vector. Let $\boldsymbol{\beta}$ be a K-vector.

$$
\begin{aligned}
\pi_k &\sim \text{Stick-Breaking}(\alpha) \\
\mathbf{X}_i|\mathbf{Z}_i, \mathbf{A} &\sim MVN(\mathbf{Z}_i\mathbf{A}, \sigma_X^2 I_D) \\
\mathbf{A}_k &\sim MVN(\mathbf{0}, \sigma_A^2 I_D) \\
\mathbf{Y}_i|\mathbf{Z}_i, \boldsymbol{\beta} &\sim N(\mathbf{Z}_i\boldsymbol{\beta}, \tau^{-1}) \\
\tau &\sim \text{Gamma}(a, b) \\
\beta|\tau &\sim MVN(\mathbf{0}, \tau^{-1} I_K)
\end{aligned}
$$

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

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

2 Approximation

$$
\begin{aligned}
q_{\lambda_k}(\pi_k) &= \text{Beta}(\pi_k; \lambda_{k,1}, \lambda_{k,2}) \\
q_{\phi_k, \boldsymbol{\Phi}_k}(\mathbf{A}_k) &= \text{Normal}(\mathbf{A}_k; \bar{\phi}_k, \boldsymbol{\Phi}_k) \\
q_{\nu_{n,k}}(z_{n,k}) &= \text{Bernoulli}(z_{n,k}; \nu_{n,k}) \\
q_{\mathbf{m},\mathbf{S},c,d}(\beta, \tau) &= \text{MVN}(\boldsymbol{\beta}; \mathbf{m}, \mathbf{S}) \times \text{Gamma}(\tau; c, d)
\end{aligned}
$$

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

Consider the problem of computing the log posterior.

$$
\log p(\mathbf{W}|\mathbf{X}, \mathbf{Y}, \boldsymbol{\theta}) = \log p(\mathbf{W}, \mathbf{X}, \mathbf{Y}|\boldsymbol{\theta}) - \log p(\mathbf{X}, \mathbf{Y}|\boldsymbol{\theta})
$$

This is difficult because $\log p(\mathbf{X}, \mathbf{Y}|\boldsymbol{\theta}) = \log \int p(\mathbf{X}, \mathbf{Y}, \mathbf{W}|\boldsymbol{\theta}) d\mathbf{W}$ is intractable.

We therefore approximate with the distribution q,

$$
q(\mathbf{W}) = q_{\boldsymbol{\lambda}}(\boldsymbol{\pi}) q_{\boldsymbol{\phi},\boldsymbol{\Phi}}(\mathbf{A}) q_{\boldsymbol{\eta}}(\mathbf{Z}) q_{\mathbf{m},\mathbf{S},\mathbf{c},\mathbf{d}}(\boldsymbol{\beta}, \tau)
$$

$$
\underset{\tau,\phi,\nu,m,S,c,d}{D(q|p)} = \underset{\tau,\phi,\nu,m,S,c,d}{\text{argmax}} \; \mathbb{E}_q[\log(p(\mathbf{X}, \mathbf{Y}, \mathbf{W}|\boldsymbol{\theta}))] + H[q]
$$

$$p_K(\mathbf{W}, \mathbf{X}, \mathbf{Y}|\boldsymbol{\theta}) = p(\tau|a,b)\prod_{k=1}^{K}\left(p(\pi_k|\alpha)p(\mathbf{A}_k|\sigma_A^2)p(\beta_k|\tau)\prod_{i=1}^{N}p(z_{i,k}|\pi_k)\right)\prod_{i=1}^{N}p(\mathbf{X}_i|\mathbf{Z}_i,\mathbf{A},\sigma_X^2)p(Y_i|\mathbf{Z}_i,\boldsymbol{\beta},\tau)$$

3 Parameter Updates

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

3.1 Updating ϕ, Φ

$$\begin{aligned}
\mathbb{E}_{\mathbf{A}_{-k},\mathbf{Z}}[(\mathbf{X}_i - \mathbf{Z}_i\mathbf{A})(\mathbf{X}_i - \mathbf{Z}_i\mathbf{A})^T] &\propto -2\mathbf{A}_k\mathbb{E}_{\mathbf{A}_{-k},\mathbf{Z}}[z_{i,k}(\mathbf{X}_i - \mathbf{Z}_i\mathbf{A})] - \mathbf{A}_k\mathbb{E}_{\mathbf{Z}}[z_{i,k}]\mathbf{A}_k^T \\
&\propto -2\mathbf{A}_k\left[\mathbb{E}[z_{i,k}]\left(\mathbf{X}_i - \sum_{l=1}^{K}\mathbb{E}_{\mathbf{A}_{-k},\mathbf{Z}}[z_{i,k}z_{i,l}\mathbf{A}_l]\right)\right] - \mathbf{A}_k\nu_{i,k}\mathbf{A}_k^T \\
&\propto -2\mathbf{A}_k\left(\nu_{i,k}\left(\mathbf{X}_i - \sum_{l:k\neq l}\nu_{i,l}\bar{\phi}_l\right)\right) + \mathbf{A}_k\nu_{i,k}\mathbf{A}_k^T
\end{aligned}$$

$$\begin{aligned}
\log q_{\phi_k}(\mathbf{A}_k) &\propto \mathbb{E}_{\mathbf{A}_{-k},\mathbf{Z}}[\log p_K(\mathbf{W},\mathbf{X}|\boldsymbol{\theta})] \\
&\propto \mathbb{E}_{\mathbf{A}_{-k},\mathbf{Z}}[\log p_K(\mathbf{A}_k|\sigma_A^2) + \sum_{i=1}^{N}\log p_K(\mathbf{X}_i|\mathbf{Z}_i,\mathbf{A},\sigma_X^2)] \\
&\propto -\frac{1}{2\sigma_A^2}\mathbf{A}_k\mathbf{A}_k^T - \frac{1}{2\sigma_X^2}\sum_{i=1}^{N}\mathbb{E}_{\mathbf{A}_{-k},\mathbf{Z}}[(\mathbf{X}_i - \mathbf{Z}_i\mathbf{A})(\mathbf{X}_i - \mathbf{Z}_i\mathbf{A})^T] \\
&\propto -\frac{1}{2}\left[\mathbf{A}_k\left(\frac{1}{\sigma_A^2} + \frac{\sum_{i=1}^{N}\nu_{i,k}}{\sigma_n^2}\right)\mathbf{A}_k^T - 2\mathbf{A}_k\left(\frac{1}{\sigma_X^2}\sum_{i=1}^{N}\nu_{i,k}\left(\mathbf{X}_i - \left(\sum_{l:l\neq k}\nu_{i,l}\bar{\phi}_l\right)\right)\right)^T\right]
\end{aligned}$$

This is the kernel of a Normal distribution.

$$\begin{aligned}
\mathbf{A}_k &\sim MVN(\bar{\phi}_k, \boldsymbol{\Phi}_k) \\
\bar{\phi}_k &= \left[\frac{1}{\sigma_X^2}\sum_{i=1}^{N}\nu_{i,k}\left(\mathbf{X}_i - \left(\sum_{l:l\neq k}\nu_{i,l}\bar{\phi}_l\right)\right)\right]\left(\frac{1}{\sigma_A^2} + \frac{\sum_{i=1}^{N}\nu_{i,k}}{\sigma_X^2}\right)^{-1} \\
\boldsymbol{\Phi}_k &= \left(\frac{1}{\sigma_A^2} + \frac{\sum_{i=1}^{N}\nu_{i,k}}{\sigma_X^2}\right)^{-1} I
\end{aligned}$$

3.2 Updating m, S, c, and d

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

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

$$
\begin{aligned}
\boldsymbol{\beta} &\sim \text{MVN}(\mathbf{m}, \tau^{-1}\mathbf{S}) \\
\tau &\sim \text{Gamma}(c,d) \\
\mathbf{m} &= \mathbf{S}\mathbb{E}[\mathbf{Z}^T]\mathbf{Y} \\
\mathbf{S} &= (\mathbb{E}[\mathbf{Z}^T\mathbf{Z}]+I_K)^{-1} \\
c &= a + \frac{N}{2} \\
d &= b + \frac{\mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbb{E}[\mathbf{Z}]\mathbf{S}\mathbb{E}[\mathbf{Z}^T]\mathbf{Y}}{2}
\end{aligned}
$$

Updating ν

$$
\begin{aligned}
\log q_{\nu_{i,k}}(z_{i,k}) &\propto \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{A},\boldsymbol{Z}_{-ik},\boldsymbol{\beta},\tau}[\log p_K(\mathbf{W},\mathbf{X},\mathbf{Y}|\boldsymbol{\theta})] \\
&\propto \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{A},\boldsymbol{Z}_{-ik},\boldsymbol{\beta},\tau}[\log p(z_{i,k}|\pi_k) + \log p(\mathbf{X}_i|\mathbf{Z}_i,\mathbf{A},\sigma_X^2) + \log p(Y_i|\mathbf{Z}_i,\boldsymbol{\beta},\tau)]
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}_{\pi_k,\mathbf{Z}_{-ik}}[\log p(z_{i,k}|\pi_k)] &= z_{i,k}\mathbb{E}[\log(\pi_k)] + (1-z_{i,k})\mathbb{E}[\log(1-\pi_k)] \\
&= z_{i,k}[\psi(\lambda_{k,1}) - \psi(\lambda_{k,1}+\lambda_{k,2})] + (1-z_{i,k})[\psi(\lambda_{k,2}) - \psi(\lambda_{k,1}+\lambda_{k,2})] \\
&= z_{i,k}[\psi(\lambda_{k,1}) - \psi(\lambda_{k,2})] + \psi(\lambda_{k,2}) - \psi(\lambda_{k,1}+\lambda_{k,2})
\end{aligned}
$$

$$\begin{aligned}
\mathbb{E}_{\mathbf{A},\mathbf{Z}_{-nk}}[\log p(\mathbf{X}_i|\mathbf{Z}_i,\mathbf{A},\sigma_X^2)] &\propto -\frac{1}{2\sigma_X^2}\mathbb{E}_{\mathbf{A},\mathbf{Z}_{-ik}}[(\mathbf{X}_i-\mathbf{Z}_i\mathbf{A})(\mathbf{X}_i-\mathbf{Z}_i\mathbf{A})^T] \\
&\propto -\frac{1}{2\sigma_X^2}\mathbb{E}_{\mathbf{A},\mathbf{Z}_{-ik}}[-2\mathbf{Z}_i\mathbf{A}\mathbf{X}_n^T+\mathbf{Z}_i\mathbf{A}\mathbf{A}^T\mathbf{Z}_i^T] \\
&\propto -\frac{1}{2\sigma_X^2}\left[-2z_{i,k}\bar{\phi}_k\mathbf{X}_i^T+z_{i,k}(\mathrm{tr}(\mathbf{\Phi}_k)+\bar{\phi}_k\bar{\phi}_k^T)+2z_{i,k}\bar{\phi}_k\left(\sum_{l:l\neq k}\nu_{i,l}\bar{\phi}_l^T\right)\right]
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_{\mathbf{Z}_{-ik},\boldsymbol{\beta},\tau}[\log p(Y_i|\mathbf{Z}_i,\boldsymbol{\beta},\tau)] &\propto \mathbb{E}_{\mathbf{Z}_{-ik},\boldsymbol{\beta},\tau}[-\frac{\tau}{2}(Y_i-\mathbf{Z}_i\boldsymbol{\beta})(Y_i-\mathbf{Z}_i\boldsymbol{\beta})^T] \\
&\propto \mathbb{E}_{\mathbf{Z}_{-ik},\boldsymbol{\beta},\tau}[-\frac{\tau}{2}(-2\mathbf{Z}_i\boldsymbol{\beta}Y_i+\mathbf{Z}_i\boldsymbol{\beta}\boldsymbol{\beta}^T\mathbf{Z}_i^T)] \\
&\propto -\frac{c}{2d}\left(-2z_{i,k}m_kY_i+z_{i,k}\left(\frac{dS_{k,k}}{c-1}+m_k^Tm_k\right)+2z_{i,k}m_k\left(\sum_{l:l\neq k}\nu_{i,l}m_l\right)\right)
\end{aligned}$$

$$\begin{aligned}
\log\frac{\nu_{i,k}}{1-\nu_{i,k}} &= \psi(\lambda_{k,1})-\psi(\lambda_{k,2})-\frac{1}{2\sigma_X^2}\left[-2\bar{\phi}_k\mathbf{X}_i^T+(\mathrm{tr}(\mathbf{\Phi}_k)+\bar{\phi}_k\bar{\phi}_k^T)+2\bar{\phi}_k\left(\sum_{l:l\neq k}\nu_{i,l}\bar{\phi}_l^T\right)\right] \\
&\quad -\frac{c}{2d}\left(-2m_kY_i+\left(\frac{dS_{k,k}}{c-1}+m_k^Tm_k\right)+2m_k\left(\sum_{l:l\neq k}\nu_{i,l}m_l\right)\right) \\
&\equiv v_{i,k}
\end{aligned}$$

$$\begin{aligned}
\nu_{i,k} &= \frac{1}{1+\exp(-v_{i,k})} \\
\log\left(z_{i,k}^{\nu_{i,k}}(1-z_{i,k})^{1-\nu_{i,k}}\right) &= z_{i,k}\log\left(\frac{1}{1+\exp(-v_{i,k})}\right)+(1-z_{i,k})\log\left(\frac{\exp(-v_{i,k})}{1+\exp(-v_{i,k})}\right) \\
&= -z_{i,k}\log(1+\exp(-v_{i,k}))+z_{i,k}\log(1+\exp(-v_{i,k}))-z_{i,k}\log(\exp(-v))+\log\left(\frac{\exp(-v_{i,k})}{1+\exp(-v_{i,k})}\right) \\
&\propto z_{i,k}v_{i,k}
\end{aligned}$$

So this is a Bernoulli kernel.

$$\begin{aligned}
z_{i,k} &\sim \text{Bernoulli}(\nu_{i,k}) \\
\nu_{i,k} &= \frac{1}{1+\exp(-v)}
\end{aligned}$$

Updating λ

$$\begin{aligned}
\log q_{\boldsymbol{\lambda}_k}(\pi_k) &\propto \mathbb{E}_{\mathbf{A},\mathbf{Z}}[\log p_K(\mathbf{W}, \mathbf{X}, \mathbf{Y}|\boldsymbol{\theta})] \\
&\propto \mathbb{E}_{\mathbf{A},\mathbf{Z}}\left[\log p_K(\pi_k|\alpha) + \sum_{i=1}^{N} \log p_K(z_{i,k}|\pi_k)\right] \\
&= \left(\frac{\alpha}{K} - 1\right) \log \pi_k + \sum_{i=1}^{N} (\nu_{i,k} \log \pi_k + (1 - \nu_{i,k}) \log(1 - \pi_k))
\end{aligned}$$

This is a Beta kernel.

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