

Student Name: Gurpreet Singh
Roll Number: 150259
Date: April 20, 2018

Question 1

Firstly, we need to write the conditional posteriors, and show that the distribution is conditionally conjugate, which is the required condition for Gibbs Sampling.

Using Bayes Rule, we can write,

$$\begin{aligned}\mathbb{P}[\mu | \mathbf{x}, \tau] &= \frac{\mathbb{P}[\mathbf{x} | \mu, \tau] \mathbb{P}[\mu]}{\int \mathbb{P}[\mathbf{x} | \mu, \tau] \mathbb{P}[\mu] d\mu} \\ &\propto \mathbb{P}[\mathbf{x} | \mu, \tau] \mathbb{P}[\mu] \\ &= \mathcal{N}(\mathbf{x} | \mu, \tau^{-1}) \cdot \mathcal{N}(\mu | \mu_o, s_o)\end{aligned}$$

Using Gaussian properties (from <http://tinyurl.com/on38cle>), we can directly write the posterior as

$$\mathbb{P}[\mu | \mathbf{x}, \tau] = \mathcal{N}(\mu | \mu_N(\tau), s_N(\tau)) \quad (1)$$

where

$$\begin{aligned}\mu_N(\tau) &= \frac{\mu_o + s_o \tau \sum_{n=1}^N x_n}{1 + s_o N \tau} \\ s_N(\tau) &= \frac{1}{s_o} + N \tau\end{aligned}$$

Similarly, we can find the posterior for the variance paramter

$$\mathbb{P}[\tau | a, b] \propto \mathcal{N}(\mathbf{x} | \mu, \tau) \cdot \text{Gamma}(\tau | a, b)$$

Since this is again a conjugate pair, we can write the form of the posterior (from <http://tinyurl.com/on38cle>) as follows

$$\mathbb{P}[\tau | \mathbf{x}, \mu] = \text{Gamma}(a_N(\mu), b_N(\mu)) \quad (2)$$

where

$$\begin{aligned}a_N(\mu) &= a + \frac{N}{2} \\ b_N(\mu) &= b + \frac{\sum_{n=1}^N (x_n - \mu)^2}{2}\end{aligned}$$

Note. Sampling from both Gaussian and Beta distributions is easy, and therefore there need not be any intermediate steps.

We can now write the Gibbs Sampling algorithm as given in Algorithm 1.

Algorithm 1: Gibbs Sampler for Univariate Gaussian

Input: Conditional Posteriors for parameters (μ, τ) and set of data points $\mathbf{x} = [x_1 \dots x_N]$

Output: Set of samples $\left\{(\mu^{(t)}, \tau^{(t)})\right\}_{t=T_o}^T$

Steps:

1. Initialize an estimate $(\mu^{(0)}, \tau^{(0)})$

2. For $t = 1, 2 \dots T$,

Sample $\mu^{(t+1)} \sim \mathbb{P}\left[\mu \mid \mu_N\left(\tau^{(t)}\right), s_N\left(\tau^{(t)}\right)\right]$ ([Equation 1](#))

Sample $\tau^{(t+1)} \sim \mathbb{P}\left[\tau \mid a_N\left(\mu^{(t+1)}\right), b_N\left(\mu^{(t+1)}\right)\right]$ ([Equation 2](#))

Question 2

Part A

In order to generate a Gibbs sampler, we need to find the conditional posteriors for all the parameters $(\beta_k, \theta_d, z_{d,n})$.

Note. I have represented $\mathbf{z}_{d,n}$ and $z_{d,n}$ as the one-hot representation and the cluster index respectively. Similarly for $\mathbf{w}_{d,n}$ and $w_{d,n}$

Note. Θ represents the set $\{\theta_d\}_{d=1}^D$ and \mathbf{B} represents $\{\beta_k\}_{k=1}^K$

1. Conditional Posterior for β_k

Using Bayes Rule, we have

$$\mathbb{P}[\beta_k | \mathbf{W}, \beta_{-k}, \mathbf{Z}, \Theta] \propto \mathbb{P}[\mathbf{W} | \beta_k, \beta_{-k}, \mathbf{Z}] \mathbb{P}[\beta_k]$$

Since only the words where $z_{d,n} = k$ will depend on β_k , we can write this as

$$\mathbb{P}[\beta_k | \mathbf{W}, \beta_{-k}, \mathbf{Z}, \Theta] \propto \prod_{\substack{d,n \\ z_{d,n}=k}} \mathbb{P}[w_{d,n} | \beta_k] \mathbb{P}[\beta_k]$$

Since Multinoulli and Dirichlet are conjugate priors, from the standard results, we can write the conditional posterior as

$$\mathbb{P}[\beta_k | \mathbf{W}, \beta_{-k}, \mathbf{Z}, \Theta] = \text{Dirichlet}(\eta_{k,1} \dots \eta_{k,V}) \quad (3)$$

where

$$\begin{aligned} \eta_{k,v} &= \eta + \sum_{d,n} \mathbb{I}[z_{d,n} = k] \mathbb{I}[w_{d,n} = v] \\ &= \eta + \sum_{d,n} \mathbf{z}_{d,n}^k \mathbf{w}_{d,n}^v \end{aligned}$$

2. Conditional Posterior for θ_d

Using Bayes Rule, we have

$$\mathbb{P}[\theta_d | \mathbf{W}, \theta_{-d}, \mathbf{Z}, \mathbf{B}] \propto \mathbb{P}[\mathbf{W} | \theta_d, \theta_{-d}, \mathbf{Z}, \mathbf{B}] \mathbb{P}[\mathbf{Z} | \theta_d, \theta_{-d}] \mathbb{P}[\theta_d]$$

The likelihood, given \mathbf{Z} and \mathbf{B} is independent of Θ . Therefore, we have

$$\begin{aligned} \mathbb{P}[\theta_d | \mathbf{W}, \theta_{-d}, \mathbf{Z}, \mathbf{B}] &\propto \mathbb{P}[\mathbf{Z} | \theta_d, \theta_{-d}] \mathbb{P}[\theta_d] \\ &\propto \prod_n \mathbb{P}[z_{d,n} | \theta_d] \mathbb{P}[\theta_d] \\ &= \prod_n \text{Multinoulli}(z_{d,n} | \theta_d) \text{Dirichlet}(\theta_d | \alpha \dots \alpha) \end{aligned}$$

This is again a conjugate prior, and therefore we can directly write the conditional posterior as

$$\mathbb{P}[\theta_d | \mathbf{W}, \theta_{-d}, \mathbf{Z}, \mathbf{B}] = \text{Dirichlet}(\theta_d | \alpha_{d,1} \dots \alpha_{d,K}) \quad (4)$$

where

$$\alpha_{d,k} = \alpha + \sum_n \mathbf{z}_{d,n}^k$$

3. Conditional Posterior for $z_{d,n}$

Since only the n^{th} word of the d^{th} document depends on $z_{d,n}$, using Bayes Rule, we can write

$$\begin{aligned}\mathbb{P}\left[z_{d,n} \mid \mathbf{W}, z_{-d,n}, \boldsymbol{\Theta}, \mathbf{B}\right] &\propto \mathbb{P}\left[w_{d,n} \mid z_{d,n}, \mathbf{B}\right] \mathbb{P}\left[z_{d,n} \mid \boldsymbol{\theta}_d\right] \\ &= \text{Multinomial}\left(w_{d,n} \mid \boldsymbol{\beta}_{z_{d,n}}\right) \text{Multinoulli}\left(\boldsymbol{\theta}_d\right) \\ &= \prod_k \left(\beta_{k,w_{d,n}}\right)^{z_{d,n}^k} \prod_k \left(\theta_{d,k}\right)^{z_{d,n}^k}\end{aligned}$$

Therefore, we have the conditional posterior distribution of $z_{d,n}$ as

$$\mathbb{P}\left[z_{d,n} \mid \mathbf{W}, z_{-d,n}, \boldsymbol{\Theta}, \mathbf{B}\right] = \text{Multinoulli}\left(\gamma_{d,n}^1 \dots \gamma_{d,n}^K\right) \quad (5)$$

where

$$\gamma_{d,n}^k = \frac{\beta_{k,w_{d,n}} \cdot \theta_{d,k}}{\sum_{k'=1}^K \left(\beta_{k',w_{d,n}} \cdot \theta_{d,k'}\right)^{z_{d,n}^{k'}}}$$

We can now give the complete Gibbs Sampler as in Algorithm 2

Part B

1. Mean of conditional posterior of β_k

We can use properties of the Dirichlet distribution to write the mean of β_k using the conditional posterior as

$$\begin{aligned}\mathbb{E}[\beta_k] &= \frac{1}{\sum_{v=1}^V \eta_{k,v}} \boldsymbol{\eta}_k \\ &= \frac{1}{V\eta + \sum_{d,n} \mathbf{z}_{d,n}^k} \boldsymbol{\eta}_k\end{aligned}$$

The intuition behind this is that all other parameters fixed, we compute the posterior such that the probabilities are best representative of the words we expect to be generated for some topic, and therefore expected value of each topic probability is directly proportional to the number of times a word has occurred in that topic, along with some inclination to the prior. The expectation is similar to what we see in a standard Mixture Model case.

2. Mean of conditional posterior of θ_d

Similar to previous case

$$\begin{aligned}\mathbb{E}[\boldsymbol{\theta}_d] &= \frac{1}{\sum_{k=1}^K \alpha_{d,k}} \boldsymbol{\alpha}_d \\ &= \frac{1}{N_d + K\alpha} \boldsymbol{\alpha}_d\end{aligned}$$

The intuition behind this is that all other parameters fixed, we compute the posterior such that the probabilities are best representative of the topics, and therefore expected value of each topic probability is directly proportional to the population of that topic, along with some inclination to the prior.

Algorithm 2: Gibbs Sampler for LDA Model

Input: Conditional Posteriors for parameters $(\beta_k, \theta_d, z_{d,n})$ and set of words $\{w_{d,n}\}_{d,n}$

Output: Set of samples $\left\{ \left\{ (\beta_k^{(t)}, \theta_d^{(t)}, z_{n,d}^{(t)}) \right\}_{d,n,k} \right\}_{t=T_o}^T$

Steps:

1. For all $k \in [K]$, $d \in [D]$, $n \in [N]$, initialize the estimates as $(\beta_k^{(0)}, \theta_d^{(0)}, z_{n,d}^{(0)})$

2. For $t = 1, 2, \dots, T$

a. For all $k \in [K]$, sample β_k as

$$\{\beta_k\}^{(t+1)} \sim \text{Dirichlet} \left(\{\eta_{k,1}\}^{(t+1)} \dots \{\eta_{k,V}\}^{(t+1)} \right)$$

where

$$\forall v \in [V], \quad \{\eta_{k,v}\}^{(t+1)} = \eta + \sum_{d,n} \{\mathbf{z}_{d,n}^k\}^{(t)} \mathbf{w}_{d,n}^k$$

b. For all $d \in [D]$, sample θ_d as

$$\{\theta_d\}^{(t+1)} \sim \text{Dirichlet} \left(\{\alpha_{d,1}\}^{(t+1)} \dots \{\alpha_{d,K}\}^{(t+1)} \right)$$

where

$$\forall k \in [K], \quad \{\alpha_{d,k}\}^{(t+1)} = \alpha + \sum_n \{\mathbf{z}_{d,n}^k\}^{(t)}$$

c. For all $d \in [D]$, $n \in [N]$, sample $z_{d,n}$ as

$$\{z_{d,n}\}^{(t+1)} \sim \text{Multinoulli} \left(\{\gamma_{d,n}^1\}^{(t+1)} \dots \{\gamma_{d,n}^K\}^{(t+1)} \right)$$

where

$$\forall k \in [K], \quad \{\gamma_{d,n}^k\}^{(t+1)} = \frac{\{\beta_{k,w_{d,n}}\}^{(t+1)} \cdot \{\theta_{d,k}\}^{(t+1)}}{\sum_{k'=1}^K \left(\{\beta_{k',w_{d,n}}\}^{(t+1)} \cdot \{\theta_{d,k'}\}^{(t+1)} \right) \{\mathbf{z}_{d,n}^{k'}\}^{(t+1)}}$$

3. Return $\left\{ \left\{ (\beta_k^{(t)}, \theta_d^{(t)}, z_{n,d}^{(t)}) \right\}_{d,n,k} \right\}_{t=T_o}^T$

Question 3

Part A

We can write the marginal distribution as

$$\begin{aligned}
\mathbb{P}[\mathbf{Z} | \alpha] &= \int_{\boldsymbol{\pi}} \mathbb{P}[\mathbf{Z}, \boldsymbol{\pi} | \alpha] d\boldsymbol{\pi} \\
&= \int_{\boldsymbol{\pi}} \mathbb{P}[\mathbf{Z} | \boldsymbol{\pi}] \mathbb{P}[\boldsymbol{\pi} | \alpha] d\boldsymbol{\pi} \\
&= \prod_{k=1}^K \int_{\pi_k} \mathbb{P}[\mathbf{Z}_{:,k} | \pi_k] \mathbb{P}[\pi_k | \alpha] d\pi_k \\
&= \prod_{k=1}^K \int_{\pi_k} \left\{ \prod_{n=1}^N \mathbb{P}[Z_{n,k} | \pi_k] \right\} \mathbb{P}[\pi_k | \alpha] d\pi_k \\
&= \prod_{k=1}^K \frac{1}{B(\alpha/K, 1)} \int_{\pi_k} \left\{ \prod_{n=1}^N \pi_k^{Z_{n,k}} (1 - \pi_k)^{1 - Z_{n,k}} \right\} \pi_k^{\alpha/K - 1} d\pi_k \\
&= \prod_{k=1}^K \frac{1}{B(\alpha/K, 1)} \int_{\pi_k} (\pi_k)^{\sum_{n=1}^N Z_{n,k} + \alpha/K - 1} (1 - \pi_k)^{N - \sum_{n=1}^N Z_{n,k}} d\pi_k \\
&= \prod_{k=1}^K \frac{1}{B(\alpha/K, 1)} B\left(\sum_{n=1}^N Z_{n,k} + \alpha/K, N + 1 - \sum_{n=1}^N Z_{n,k}\right) \times \int_{\pi_k} \text{Beta}\left(\sum_{n=1}^N Z_{n,k} + \alpha/K, N + 1 - \sum_{n=1}^N Z_{n,k}\right) d\pi_k
\end{aligned}$$

Since the integrals will evaluate to 1, we have

$$\mathbb{P}[\mathbf{Z} | \alpha] = \prod_{k=1}^K \frac{1}{B(\alpha/K, 1)} B\left(\sum_{n=1}^N Z_{n,k} + \alpha/K, N + 1 - \sum_{n=1}^N Z_{n,k}\right) \quad (6)$$

Part B

We can write the desired quantity as

$$\begin{aligned}
\mathbb{P}[Z_{n,k} | Z_{-n,k}, \alpha] &= \frac{\mathbb{P}[\mathbf{Z} | \alpha]}{\mathbb{P}[Z_{-n,k} | \alpha]} \\
&= \frac{B\left(\sum_{n'=1}^N Z_{n',k} + \alpha/K, N + 1 - \sum_{n'=1}^N Z_{n',k}\right)}{\sum_{Z_{n',k} \in \{0,1\}} B\left(\sum_{n'=1}^N Z_{n',k} + \alpha/K, N + 1 - \sum_{n'=1}^N Z_{n',k}\right)} \\
&= \frac{\Gamma\left(\sum_{n'=1}^N Z_{n',k} + \alpha/K\right) \Gamma\left(N + 1 - \sum_{n'=1}^N Z_{n',k}\right)}{\sum_{Z_{n',k} \in \{0,1\}} \Gamma\left(\sum_{n'=1}^N Z_{n',k} + \alpha/K\right) \Gamma\left(N + 1 - \sum_{n'=1}^N Z_{n',k}\right)}
\end{aligned}$$

Suppose $S_{n,k} = \sum_{\substack{n=1 \\ n \neq n'}}^N Z_{n',k}$. Using the property that $\Gamma(a) = (a-1)\Gamma(a-1)$, then we can write

$$\begin{aligned} \mathbb{P} \left[Z_{n,k} = 0 \mid Z_{-n,k}, \alpha \right] &= \frac{1}{1 + \frac{N - S_{n,k}}{S_{n,k} + \alpha/K}} \\ &= \frac{S_{n,k} + \alpha/K}{N + \alpha/K} \end{aligned}$$

Therefore, we can write the distribution as a Bernoulli, given as follows.

$$\mathbb{P} \left[Z_{n,k} \mid Z_{-n,k}, \alpha \right] = \text{Bernoulli} \left(\frac{S_{n,k} + \alpha/K}{N + \alpha/K} \right) \quad (7)$$

Clearly, as $K \rightarrow \infty$,

$$\mathbb{P} \left[Z_{n,k} \mid Z_{-n,k}, \alpha \right] = \text{Bernoulli} \left(\frac{S_{n,k}}{N} \right)$$

Part C

We can write the number of ones in a column as the sum of the entries in that column (since each entry is a Bernoulli Random Variable). Suppose S_k denotes the number of ones in column k , i.e. $S_k = \sum_{n=1}^N Z_{n,k}$, we can write

$$\mathbb{E} \left[S_k \mid \alpha \right] = \mathbb{E} \left[\sum_{n=1}^N Z_{n,k} \mid \alpha \right]$$

From the law of iterated expectations, we have

$$\begin{aligned} \mathbb{E} \left[S_k \mid \alpha \right] &= \mathbb{E} \left[\mathbb{E} \left[\sum_{n=1}^N Z_{n,k} \mid \pi_k \right] \mid \alpha \right] \\ &= \mathbb{E} \left[\sum_{n=1}^N \mathbb{E} \left[Z_{n,k} \mid \pi_k \right] \mid \alpha \right] \\ &= \mathbb{E} \left[\sum_{n=1}^N \pi_k \mid \alpha \right] \\ &= \frac{\alpha \cdot N}{\alpha + K} \end{aligned}$$

Therefore the expected value of number of ones for column k is given as

$$\mathbb{E} [S_k] = \frac{\alpha \cdot N}{\alpha + K}$$

Total expected number of ones, denoted by S , can simply be calculated as the sum of expected number of ones for each column, since using equation 6, we can say that the entries are independent across columns. Therefore,

$$\mathbb{E}[S] = K \frac{\alpha \cdot N}{\alpha + K}$$

Question 4

With the assumptions that A is symmetric, we can formulate the generative story using a combination of the Mixed Membership Stochastic Blockmodel (MMSB) with the Latent Dirichlet Allocation (LDA) Model as given in Algorithm 3

Algorithm 3: Generative Story

1. For $k = 1 \dots K$, sample

$$\beta_k \sim \text{Dirichlet}(\eta \dots \eta)$$
2. For $d = 1 \dots D$, sample

$$\theta_d \sim \text{Dirichlet}(\alpha \dots \alpha)$$
3. For $d = 1 \dots D$,
 For $d' = 1 \dots d$, sample

$$r_{d \rightarrow d'} \sim \text{Multinoulli}(\theta_d)$$

$$r_{d' \rightarrow d} \sim \text{Multinoulli}(\theta_{d'})$$

$$A_{d,d'} \sim \text{Bernoulli}(\gamma_{r_{d \rightarrow d'}, r_{d' \rightarrow d}})$$

$$A_{d',d} = A_{d,d'}$$
4. For $d = 1 \dots D$,
 For $n = 1 \dots N_d$, sample

$$z_{d,n} \sim \text{Multinoulli}(\theta_d)$$

$$w_{d,n} \sim \text{Bernoulli}(\beta_{z_{d,n}})$$

Another alternative approach to modelling this could be to consider the matrix γ to be computed as a function of the words in the documents. The alternative algorithm is given in Algorithm 4

The second model allows the link between documents to be based on the number of common words the two documents have. We can also propose more complex models, for example have embeddings for the words and compute similarity with more complex functions.

Algorithm 4: Generative Story (Alternate)

1. For $k = 1 \dots K$, sample

$$\beta_k \sim \text{Dirichlet}(\eta \dots \eta)$$

2. For $d = 1 \dots D$, sample

$$\theta_d \sim \text{Dirichlet}(\alpha \dots \alpha)$$

3. For $d = 1 \dots D$,

For $n = 1 \dots N_d$, sample

$$z_{d,n} \sim \text{Multinoulli}(\theta_d)$$

$$w_{d,n} \sim \text{Bernoulli}(\beta_{z_{d,n}})$$

4. For $d = 1 \dots D$,

For $d' = 1 \dots d$, sample

$$\gamma_{d,d'} = \frac{|\mathbf{w}_d \cap \mathbf{w}_{d'}|}{|\mathbf{w}_d \cup \mathbf{w}_{d'}|}$$

$$A_{d,d'} \sim \text{Bernoulli}(\gamma_{d,d'})$$

$$A_{d',d} = A_{d,d'}$$