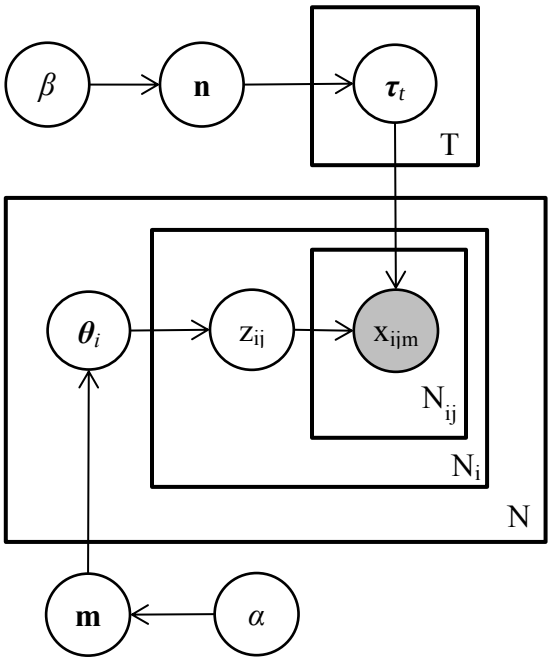
Derivation of Posterior Probabilities for TpT-LDA



Definitions:

1. Topic – A multinomial probability distribution over all words in the vocabulary. The total number of topics in an LDA model needs to be either pre-determined or an appropriate number needs to be selected by cross-validation.
2. Document – Documents are represented as bag-of-words. In context of Twitter, each ***tweet*** treated as a document.
3. User – This is a unique person sending out messages. Typically, each person will have a particular distribution over topics based on personal preferences. In Twitter, a person is identified by the unique Twitter *handle*.

Notation:

N = Number of users.
N_i = Number of documents for *i*-th user.
N_{ij} = Number of words within the *j*-th document of the *i*-th user.
T = Number of topics.
V = Number of unique words in vocabulary.
W = Size of the entire corpus, i.e. the total number of words across all document and all users.
X = The entire corpus – the complete set of words across all document and all users.
$\boldsymbol{\Theta} = \{\boldsymbol{\theta}_i\}$ = Distribution over topics for *i*-th user; $i = 1,\ldots,N$
$\mathbf{T} = \{\boldsymbol{\tau}_t\}$ = Distribution over words for *t*-th topic; $t = 1,\ldots,T$
$\mathbf{Z} = \{z_{ij}\}$ = Topic for *j*-th document of *i*-th user.
$\mathbf{x}_{ij}$ = *j*-th document of *i*-th user; $j = 1,\ldots,N_i$
x_{ijm} – *m*-th word in the *j*-th document of *i*-th user.
β, α – Dirichlet concentration parameters

n, **m** – uniform Dirichlet hyper-priors

Unlike the regular LDA model, each document in our setup belongs to only one topic. This is reasonable in our context because we expect the documents to be short and therefore be generated from only one topic.

The generative process is:

1. For each user u_i:
    a. Draw the topic distribution $\boldsymbol{\theta}_i \sim Dir(\alpha \boldsymbol{m})$
    b. For each document:
        i. Draw topic $z_{ij} \sim Multi(\boldsymbol{\theta}_i)$
        ii. Draw the words within the document $\mathbf{x}_{ij} \sim Multi\left(\boldsymbol{\tau}_{z_{ij}}\right)$

The complete data likelihood is:

$$p(\mathbf{X}, \boldsymbol{Z}, \boldsymbol{\Theta}, \boldsymbol{T} | \beta \boldsymbol{n}, \alpha \boldsymbol{m}) = p(\boldsymbol{T}|\beta \boldsymbol{n}) p(\boldsymbol{\Theta}|\alpha \boldsymbol{m}) p(\boldsymbol{Z}|\boldsymbol{\Theta}) p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{T})$$

Marginalizing out $\boldsymbol{Z}$ to get the observed data likelihood leads to the following for each document:

$$p(\boldsymbol{x}_{ij}|\boldsymbol{\theta}_i, \boldsymbol{T}) = \sum_{t=1}^{T} \left[\prod_{m}^{N_{ij}} p(x_{ijm}|\boldsymbol{\tau}_t) \right] p(z_{ij} = t|\boldsymbol{\theta}_i)$$

Now, the likelihood for the entire corpus is:

$$p(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{T}) = \prod_{i=1}^{N} \prod_{j=1}^{N_i} p(\boldsymbol{x}_{ij}|\boldsymbol{\theta}_i, \boldsymbol{T}) \qquad (1)$$

where:

$$P(\boldsymbol{T}|\beta \boldsymbol{n}) = Dir(\boldsymbol{T}|\beta \boldsymbol{n})$$
$$P(\boldsymbol{\theta}_i|\alpha \boldsymbol{m}) = Dir(\boldsymbol{\theta}_i|\alpha \boldsymbol{m})$$
$$P(\mathrm{z}_{ij}|\boldsymbol{\theta}_i) = Multi(\mathrm{z}_{ij}|\boldsymbol{\theta}_i)$$
$$P\left(x_{ijm}|\mathrm{z}_{ij}, \boldsymbol{\tau}_{z_{ij}}\right) = Multi\left(x_{ijm}|\mathrm{z}_{ij}, \boldsymbol{\tau}_{z_{ij}}\right)$$

Exact inference of the parameters is hard for the LDA model. Therefore, we use Gibbs sampling. Most of the derivations below are similar to that of the original LDA model in (Heinrich, 2008).

The Collapsed Gibbs sampler is often more efficient than the regular Gibbs sampler where we try to avoid inference on some parameters. In our case, these parameters are $\boldsymbol{\Theta}$ and $\boldsymbol{T}$. These parameters can be interpreted as statistics of the associations between the observed x_{ijm} and the corresponding z_{ij}. The target inference is the distribution $p(\boldsymbol{Z}|\boldsymbol{X})$:

$$p(\boldsymbol{Z}|\boldsymbol{X}) = \frac{p(\boldsymbol{Z}, \boldsymbol{X})}{p(\boldsymbol{X})} = \frac{\prod_{i=1}^{W} p(z_i, x_i)}{\prod_{i=1}^{W} \sum_{t=1}^{T} p(z_i = t, x_i)}$$

The Gibbs sampler runs a Markov chain that uses the full conditional $p(z_i|\boldsymbol{Z}_{\neg i}, \boldsymbol{X})$ in order to simulate $p(\boldsymbol{Z}|\boldsymbol{X})$. i refers to the tuple (n, m) (i.e., n-th user's m-th document). '$\neg i$' indicates that

document *i* is excluded. The joint distribution $p(\boldsymbol{Z}, \boldsymbol{X})$ can be factored as:

$$p(\boldsymbol{Z}, \boldsymbol{X}|\alpha, \beta) = p(\boldsymbol{X}|\boldsymbol{Z}, \beta)p(\boldsymbol{Z}|\alpha) \quad (2)$$

$p(\boldsymbol{X}|\boldsymbol{Z}, \beta)$ is a multinomial over the word counts given the associated topics.

$$p(\boldsymbol{X}|\boldsymbol{Z}, \beta) = \prod_{i=1}^{W} p(x_i|z_i) = \prod_{i=1}^{W} \tau_{z_i, x_i} \quad (3)$$

The above equation (3) can also be written as:

$$p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{T}) = \prod_{t=1}^{T} \prod_{m=1}^{V} \tau_{t,m}^{n_t^{(m)}} \quad (4)$$

In equation (4), $n_t^{(m)}$ refers to the number of times that the word *m* occurs in the corpus across all documents within topic *t*. The distribution $p(\boldsymbol{X}|\boldsymbol{Z}, \beta)$ can be computed by integrating over $\boldsymbol{T}$.

$$\begin{aligned} p(\boldsymbol{X}|\boldsymbol{Z}, \beta) &= \int p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{T})p(\boldsymbol{T}|\beta)d\boldsymbol{T} \\ &= \int \prod_{t=1}^{T} \frac{1}{\Delta(\boldsymbol{\beta})} \prod_{m=1}^{V} \tau_{t,m}^{n_t^{(m)}+\beta_m-1} d\boldsymbol{\tau}_t \\ &= \prod_{t=1}^{T} \frac{\Delta(\boldsymbol{n}_t + \boldsymbol{\beta})}{\Delta(\boldsymbol{\beta})}, \boldsymbol{n}_t = \left\{n_t^{(m)}\right\}_{m=1}^{V}, \Delta(\boldsymbol{\beta}) = \frac{\prod_{m=1}^{V} \Gamma(\beta_m)}{\Gamma(\sum_{m=1}^{V} \beta_m)} \end{aligned}$$

Similar to the above, we will now derive $p(\boldsymbol{Z}|\alpha)$. We first start with $p(\boldsymbol{Z}|\boldsymbol{\Theta})$:

$$p(\boldsymbol{Z}|\boldsymbol{\Theta}) = \prod_{i=1}^{W} p(z_i|\theta_i) = \prod_{n=1}^{N} \prod_{t=1}^{T} p(z_i = t|u_i, \theta_i) = \prod_{n=1}^{N} \prod_{t=1}^{T} \theta_{n,t}^{n_n^{(t)}} \quad (5)$$

In equation (5), u_i is the user to which z_i belongs; θ_i refers to u_i's topic distribution. $n_n^{(t)}$ refers to the number of times that topic *t* has been observed with a document of user *n*. We will now integrate out $\boldsymbol{\Theta}$.

$$\begin{aligned} p(\boldsymbol{Z}|\alpha) &= \int p(\boldsymbol{Z}|\boldsymbol{\Theta})p(\boldsymbol{\Theta}|\alpha)d\boldsymbol{\Theta} \\ &= \int \prod_{n=1}^{N} \frac{1}{\Delta(\boldsymbol{\alpha})} \prod_{t=1}^{T} \theta_{n,t}^{n_n^{(t)}+\alpha_t-1} d\boldsymbol{\theta}_n \\ &= \prod_{n=1}^{N} \frac{\Delta(\boldsymbol{n}_n + \boldsymbol{\alpha})}{\Delta(\boldsymbol{\alpha})}, \boldsymbol{n}_n = \left\{n_n^{(t)}\right\}_{t=1}^{T}, \Delta(\boldsymbol{\alpha}) = \frac{\prod_{t=1}^{T} \Gamma(\alpha_t)}{\Gamma(\sum_{t=1}^{T} \alpha_t)} \end{aligned}$$

The joint distribution (2) now becomes:

$$p(\boldsymbol{Z}, \boldsymbol{X}|\alpha, \beta) = \prod_{t=1}^{T} \frac{\Delta(\boldsymbol{n}_t + \boldsymbol{\beta})}{\Delta(\boldsymbol{\beta})} \cdot \prod_{n=1}^{N} \frac{\Delta(\boldsymbol{n}_n + \boldsymbol{\alpha})}{\Delta(\boldsymbol{\alpha})}$$

For Gibbs sampling, we need to sample $p(z_i = t|\boldsymbol{Z}_{\neg i}, \boldsymbol{X})$.

$$p(z_i = t|\boldsymbol{Z}_{\neg i}, \boldsymbol{X}) = \frac{p(\boldsymbol{X}, \boldsymbol{Z})}{p(\boldsymbol{X}, \boldsymbol{Z}_{\neg i})} = \frac{p(\boldsymbol{X}|\boldsymbol{Z})}{p(\boldsymbol{X}_{\neg i}|\boldsymbol{Z}_{\neg i})p(w_i)} \cdot \frac{p(\boldsymbol{Z})}{p(\boldsymbol{Z}_{\neg i})}$$

$$\propto \frac{\Delta(\boldsymbol{n}_t + \boldsymbol{\beta})}{\Delta(\boldsymbol{n}_{t,\neg i} + \boldsymbol{\beta})} \cdot \frac{\Delta(\boldsymbol{n}_n + \boldsymbol{\alpha})}{\Delta(\boldsymbol{n}_{n,\neg i} + \boldsymbol{\alpha})}$$

$$= \frac{\prod_{m \in \boldsymbol{X}_i} \Gamma\left(n_t^{(m)} + \beta_m\right) \Gamma\left(\sum_{m=1}^{V} n_{t,\neg i}^{(m)} + \beta_m\right)}{\prod_{m \in \boldsymbol{X}_i} \Gamma\left(n_{t,\neg i}^{(m)} + \beta_m\right) \Gamma\left(\sum_{m=1}^{V} n_t^{(m)} + \beta_m\right)} \cdot \frac{\Gamma\left(n_n^{(t)} + \alpha_t\right) \Gamma\left(\sum_{t'=1}^{T} n_{n,\neg i}^{(t')} + \alpha_{t'}\right)}{\Gamma\left(n_{n,\neg i}^{(t)} + \alpha_t\right) \Gamma\left(\sum_{t'=1}^{T} n_n^{(t')} + \alpha_{t'}\right)}$$

$$= \frac{\prod_{m \in \boldsymbol{X}_i} \prod_{j=0}^{f_m} \left(n_{t,\neg i}^{(m)} + \beta_m + j\right)}{\prod_{j=0}^{length(\boldsymbol{X}_i)} \left(\sum_{m=1}^{V} n_{t,\neg i}^{(m)} + \beta_m + j\right)} \cdot \frac{n_{n,\neg i}^{(t)} + \alpha_t}{\left[\sum_{t'=1}^{T} n_n^{(t')} + \alpha_{t'}\right] - 1}$$

$$\propto \frac{\prod_{m \in \boldsymbol{X}_i} \prod_{j=0}^{f_m} \left(n_{t,\neg i}^{(m)} + \beta_m + j\right)}{\prod_{j=0}^{length(\boldsymbol{X}_i)} \left(\sum_{m=1}^{V} n_{t,\neg i}^{(m)} + \beta_m + j\right)} \left(n_{n,\neg i}^{(t)} + \alpha_t\right) \quad (6)$$

In the above, counts '$\boldsymbol{n}_{\cdot,\neg i}^{(\cdot)}$' indicate that the document i was excluded. f_m is the count of occurrences of word m in document i. $length(\boldsymbol{X}_i)$ is the total count of words in document i. $\{m \in \boldsymbol{X}_i\}$ is the set of unique words in document i.

The multinomial parameters ($\boldsymbol{\Theta}$,$\boldsymbol{T}$) can be computed as:

$$p(\boldsymbol{\theta}_n|\boldsymbol{z}_n, \boldsymbol{\alpha}) = \frac{1}{Z_{\theta_n}} \prod_{m=1}^{N_n} p\left(z_{n,m}|\boldsymbol{\theta}_n\right) \cdot p(\boldsymbol{\theta}_n|\boldsymbol{\alpha}) = Dir(\boldsymbol{n}_n + \boldsymbol{\alpha})$$

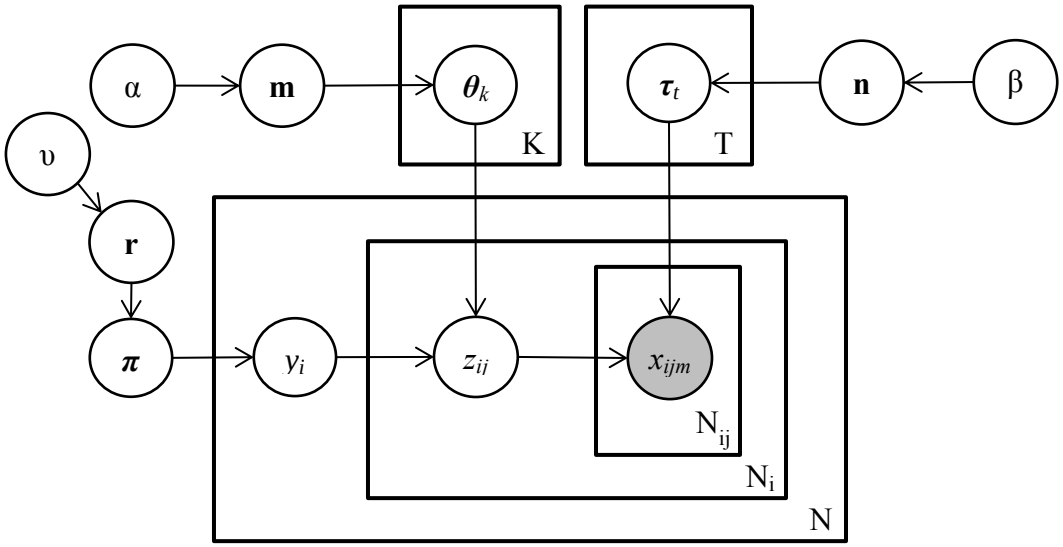
$$p(\boldsymbol{\tau}_t|\boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{\beta}) = \frac{1}{Z_{\tau_t}} \prod_{\{i:z_i=t\}} p(w_i|\boldsymbol{\tau}_t) \cdot p(\boldsymbol{\tau}_t|\boldsymbol{\beta}) = Dir(\boldsymbol{n}_t + \boldsymbol{\beta})$$

Since both the above are Dirichlet distributions, we can compute the expectations easily:

$$\tau_{t,m} = \frac{n_t^{(m)} + \beta_m}{\sum_{m=1}^{V} n_t^{(m)} + \beta_m} \quad (7)$$

$$\theta_{n,t} = \frac{n_n^{(t)} + \alpha_t}{\sum_{t'=1}^{T} n_n^{(t')} + \alpha_{t'}} \qquad (8)$$

Derivation of Posterior Probabilities for GTpT-LDA



Definitions:
In this version of the model, we continue with the same three definitions as in the previous model and add one more:

4. Group – We define a group as an entity whose users share the same topic distribution. In our model, we allow one user to belong to only one group.

The major change here is that we now define the topic distributions for groups of users so that we can impose a clustering over those users.

Notation:
N = Number of users.
N_i = Number of documents for *i*-th user.
N_{ij} = Number of words within the *j*-th document of the *i*-th user.
T = Number of topics.
K = Number of user groups.
V = Number of unique words in vocabulary.
W = Size of the entire corpus, i.e. the total number of words across all documents and all users.
X = The entire corpus – the complete set of words across all documents and all users.
Θ = $\{\boldsymbol{\theta}_k\}$ = Topic proportions for *k*-th user-group; $k = 1,\ldots,K$. This is a T-by-K matrix.
T = $\{\boldsymbol{\tau}_t\}$ = Distribution over words for *t*-th topic; $t = 1,\ldots,T$
Y = $\{y_i\}$ = Group to which the *i*-th user belongs; $i=1,\ldots N$
Z = $\{z_{ij}\}$ = Topic for *j*-th document of *i*-th user.
$\mathbf{x}_{ij}$ = *j*-th document of *i*-th user; $j = 1,\ldots,N_i$

x_{ijm} – *m*-th word in the *j*-th document of *i*-th user.
α, β, υ – Dirichlet concentration parameters
n, **m, r** – uniform Dirichlet hyper-priors

The generative process is:

1. For each user u_i:
    a. Draw the group membership $y_i \sim Multi(\boldsymbol{\pi})$
    b. Draw a topic distribution for the group $\boldsymbol{\theta}_{y_i} \sim Dir(\alpha \boldsymbol{m})$
    c. For each document of user u_i:
        i. Draw a topic $z_{ij} \sim Multi(\boldsymbol{\theta}_{y_i})$
        ii. For each words within the document
            Draw $x_{ijm} \sim Multi\left(\boldsymbol{\tau}_{z_{ij}}\right)$

The complete data likelihood is:

$$p(\mathbf{X}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\Theta}, \boldsymbol{T}, \boldsymbol{\pi} | \beta \boldsymbol{n}, \alpha \boldsymbol{m}) = p(\boldsymbol{T}|\beta \boldsymbol{n}) p(\boldsymbol{\pi}|\upsilon \boldsymbol{r}) p(\boldsymbol{Y}|\boldsymbol{\pi}) p(\boldsymbol{\Theta}|\alpha \boldsymbol{m}) p(\boldsymbol{Z}|\boldsymbol{Y}, \boldsymbol{\Theta}) p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{T})$$

Marginalizing out $\boldsymbol{Z}$, $\boldsymbol{Y}$ to get the observed data likelihood leads to the following for the *i*-th user where $\boldsymbol{x}_i$ is the complete set of words for the *i*-th user across all documents of that user:

$$p(\boldsymbol{x}_i|\boldsymbol{\Theta}, \boldsymbol{T}, \boldsymbol{\pi}) = \sum_{k=1}^{K} p(y_i = \pi_k) \left\{ \prod_{j=1}^{N_i} \left[\sum_{t=1}^{T} p(z_{ij} = t|\boldsymbol{\theta}_k) \left(\prod_{m=1}^{N_{ij}} p(x_{ijm}|\boldsymbol{\tau}_t) \right) \right] \right\}$$

Now, the likelihood for the entire corpus is (where *N* is the total number of users):

$$p(\boldsymbol{X}|\boldsymbol{\Theta}, \boldsymbol{T}, \boldsymbol{\pi}) = \prod_{i=1}^{N} p(\boldsymbol{x}_i|\boldsymbol{\Theta}, \boldsymbol{T}, \boldsymbol{\pi}) \qquad (9)$$

where:

$$P(\boldsymbol{\pi}|\upsilon \boldsymbol{r}) = Dir(\boldsymbol{\pi}|\upsilon \boldsymbol{r})$$
$$P(\boldsymbol{T}|\beta \boldsymbol{n}) = Dir(\boldsymbol{T}|\beta \boldsymbol{n})$$
$$P(\boldsymbol{\theta}_i|\alpha \boldsymbol{m}) = Dir(\boldsymbol{\theta}_i|\alpha \boldsymbol{m})$$
$$P(y_i|\boldsymbol{\pi}) = Multi(y_i|\boldsymbol{\pi})$$
$$P(\mathrm{z}_{ij}|y_i, \boldsymbol{\theta}_{y_i}) = Multi(\mathrm{z}_{ij}|y_i, \boldsymbol{\theta}_{y_i})$$
$$P\left(x_{ijm}|\mathrm{z}_{ij}, \boldsymbol{\tau}_{z_{ij}}\right) = Multi\left(x_{ijm}|\mathrm{z}_{ij}, \boldsymbol{\tau}_{z_{ij}}\right)$$

Here again, we use collapsed Gibbs sampling instead of exact inference. In this case, these parameters are $\boldsymbol{\Theta}$, $\boldsymbol{T}$, and $\boldsymbol{\pi}$. These parameters can be interpreted as statistics of the associations between the observed x_{ijm} and the corresponding y_i and z_{ij}. The target inference is the distribution $p(\boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{X})$:

$$p(\boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{X}) = \frac{p(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{X})}{p(\boldsymbol{X})}$$

The Gibbs sampler runs a Markov chain that uses the full conditional $p(y_i, \boldsymbol{z}_i|\boldsymbol{Y}_{\neg i}, \boldsymbol{Z}_{\neg i}, \boldsymbol{X})$ in

order to simulate $p(\boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{X})$. Here, '$i$' refers to all documents of the i-th user. '$\neg i$' indicates that all documents of the i-th are excluded. The joint distribution $p(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{X})$ can be factored as:

$$p(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{X}|\alpha, \beta, \upsilon) = p(\boldsymbol{X}|\boldsymbol{Z}, \beta)p(\boldsymbol{Z}|\boldsymbol{Y}, \alpha)p(\boldsymbol{Y}|\upsilon) \quad (10)$$

$p(\boldsymbol{X}|\boldsymbol{Z}, \beta)$ is a multinomial over the word counts given the associated topics.

$$p(\boldsymbol{X}|\boldsymbol{Z}, \beta) = \prod_{i=1}^{W} p(x_i|z_i) = \prod_{i=1}^{W} \tau_{z_i, x_i} \quad (11)$$

The above equation (11) can also be written as (once $\boldsymbol{Z}$ is known, $\boldsymbol{X}$ is independent of $\boldsymbol{Y}$):

$$p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\Theta}, \boldsymbol{T}, \boldsymbol{\pi}) = \prod_{t=1}^{T} \prod_{m=1}^{V} \tau_{t,m}^{n_t^{(m)}} \quad (12)$$

In equation (12), $n_t^{(m)}$ refers to the number of times that the word m occurs in the corpus across all documents belonging to topic t. The distribution $p(\boldsymbol{X}|\boldsymbol{Z}, \beta)$ can be computed by integrating over $\boldsymbol{T}$.

$$\begin{aligned} p(\boldsymbol{X}|\boldsymbol{Z}, \beta) &= \int p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{T})p(\boldsymbol{T}|\beta)d\boldsymbol{T} \\ &= \int \prod_{t=1}^{T} \frac{1}{\Delta(\boldsymbol{\beta})} \prod_{m=1}^{V} \tau_{t,m}^{n_t^{(m)} + \beta_m - 1} d\boldsymbol{\tau}_t \\ &= \prod_{t=1}^{T} \frac{\Delta(\boldsymbol{n}_t + \boldsymbol{\beta})}{\Delta(\boldsymbol{\beta})}, \boldsymbol{n}_t = \left\{ n_t^{(m)} \right\}_{m=1}^{V}, \qquad \Delta(\boldsymbol{\beta}) = \frac{\prod_{m=1}^{V} \Gamma(\beta_m)}{\Gamma(\sum_{m=1}^{V} \beta_m)} \end{aligned}$$

Similar to the above, we can derive $p(\boldsymbol{Z}|\boldsymbol{Y}, \alpha)$ as:

$$p(\boldsymbol{Z}|\boldsymbol{Y}, \boldsymbol{\Theta}, \alpha) = \prod_{i=1}^{W} p(z_i|y_i) = \prod_{n=1}^{N} \prod_{k=1}^{K} \prod_{t=1}^{T} p(z_i = t|y_i = k) = \prod_{n=1}^{N} \prod_{k=1}^{K} \prod_{t=1}^{T} \theta_{k,t}^{m_{k,n}^{(t)}} \quad (13)$$

In equation (13), $i \equiv (n, k, t)$; and y_i refers to the group for the i-th instance. $m_{k,n}^{(t)}$ refers to the number of times the topic t has been observed within the documents of user n who belongs to group k. Note that each user can belong to only one group. We will now integrate out $\boldsymbol{\Theta}$.

$$\begin{aligned} p(\boldsymbol{Z}|\boldsymbol{Y}, \alpha) &= \int p(\boldsymbol{Z}|\boldsymbol{Y}, \boldsymbol{\Theta})p(\boldsymbol{\Theta}|\alpha)d\boldsymbol{\Theta} \\ &= \int \prod_{n=1}^{N} \prod_{k=1}^{K} \frac{1}{\Delta(\boldsymbol{\alpha})} \prod_{t=1}^{T} \theta_{k,t}^{m_{k,n}^{(t)} + \alpha_t - 1} d\boldsymbol{\theta}_k \\ &= \int \prod_{k=1}^{K} \frac{1}{\Delta(\boldsymbol{\alpha})} \prod_{t=1}^{T} \theta_{k,t}^{m_k^{(t)} + \alpha_t - 1} d\boldsymbol{\theta}_k \\ &= \prod_{k=1}^{K} \frac{\Delta(\boldsymbol{m}_k + \boldsymbol{\alpha})}{\Delta(\boldsymbol{\alpha})}, \boldsymbol{m}_k = \left\{ m_k^{(t)} \right\}_{t=1}^{T}, \qquad \Delta(\boldsymbol{\alpha}) = \frac{\prod_{t=1}^{T} \Gamma(\alpha_t)}{\Gamma(\sum_{t=1}^{T} \alpha_t)} \end{aligned}$$

Here, $m_k^{(t)}$ is the number of times that topic t has been observed in group k across all users belonging to the group k.

Now for evaluating $\mathrm{p}(\boldsymbol{Y}|\upsilon)$ we first start with:

$$\mathrm{p}(\boldsymbol{Y}|\boldsymbol{\pi},\upsilon)=\prod_{k=1}^{K}\pi_k^{n^{(k)}}$$

Next, we get $\mathrm{p}(\boldsymbol{Y}|\upsilon)$ by integrating out $\boldsymbol{\pi}$:

$$p(\boldsymbol{Y}|\upsilon)=\int p(\boldsymbol{Y}|\boldsymbol{\pi})p(\boldsymbol{\pi}|\upsilon)d\boldsymbol{\pi}=\frac{\Delta(\boldsymbol{n}+\boldsymbol{\upsilon})}{\Delta(\boldsymbol{\upsilon})},\boldsymbol{n}=\left\{n^{(k)}\right\}_{k=1}^{K},\Delta(\boldsymbol{\upsilon})=\frac{\prod_{k=1}^{K}\Gamma(\upsilon_k)}{\Gamma(\sum_{k=1}^{K}\upsilon_k)}$$

Here, $n^{(k)}$ is the number of users in *k*-th group.

The joint distribution (10) now becomes:

$$p(\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{X}|\alpha,\beta)=\left(\prod_{t=1}^{T}\frac{\Delta(\boldsymbol{n}_t+\boldsymbol{\beta})}{\Delta(\boldsymbol{\beta})}\right)\cdot\left(\prod_{k=1}^{K}\frac{\Delta(\boldsymbol{m}_k+\boldsymbol{\alpha})}{\Delta(\boldsymbol{\alpha})}\right)\cdot\left(\frac{\Delta(\boldsymbol{n}_k+\boldsymbol{\upsilon})}{\Delta(\boldsymbol{\upsilon})}\right) \qquad (14)$$

For Gibbs sampling, we need to sample $p(y_i|\boldsymbol{Y}_{\neg i},\boldsymbol{z}_i,\boldsymbol{Z}_{\neg i},\boldsymbol{X})$, followed by $p(z_i|y_i,\boldsymbol{Y}_{\neg i},\boldsymbol{Z}_{\neg i},\boldsymbol{X})$.

While computing $p(y_i|\boldsymbol{Y}_{\neg i},\boldsymbol{z}_i,\boldsymbol{Z}_{\neg i},\boldsymbol{X})$, '$i$' refers to all documents of the *i*-th user. '$\neg i$' means that all documents of user 'i' are excluded.

$$\begin{aligned}p(y_i=k|\boldsymbol{Y}_{\neg i},\boldsymbol{z}_i,\boldsymbol{Z}_{\neg i},\boldsymbol{X})&=\frac{p(\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{X})}{p(\boldsymbol{Y}_{\neg i},\boldsymbol{z}_i,\boldsymbol{Z}_{\neg i},\boldsymbol{X})}\\&=\frac{p(\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{Z})}{p(\boldsymbol{X}_{\neg i}|\boldsymbol{Y}_{\neg i},\boldsymbol{Z}_{\neg i})p(\boldsymbol{z}_i,\boldsymbol{x}_i)}\cdot\frac{p(\boldsymbol{Z}|\boldsymbol{Y})}{p(\boldsymbol{Z}_{\neg i}|\boldsymbol{Y}_{\neg i})}\cdot\frac{p(\boldsymbol{Y})}{p(\boldsymbol{Y}_{\neg i})}\end{aligned}$$

Now, since $p(\boldsymbol{z}_i,\boldsymbol{x}_i)$ does not depend on y_i,

$$p(y_i=k|\boldsymbol{Y}_{\neg i},\boldsymbol{z}_i,\boldsymbol{Z}_{\neg i},\boldsymbol{X})\propto\frac{p(\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{Z})}{p(\boldsymbol{X}_{\neg i}|\boldsymbol{Y}_{\neg i},\boldsymbol{Z}_{\neg i})}\cdot\frac{p(\boldsymbol{Z}|\boldsymbol{Y})}{p(\boldsymbol{Z}_{\neg i}|\boldsymbol{Y}_{\neg i})}\cdot\frac{p(\boldsymbol{Y})}{p(\boldsymbol{Y}_{\neg i})} \qquad (15)$$

$$=\left(\prod_{t=1}^{T}\frac{\Delta(\boldsymbol{n}_t+\boldsymbol{\beta})}{\Delta(\boldsymbol{n}_{t,\neg i}+\boldsymbol{\beta})}\right)\cdot\left(\frac{\Delta(\boldsymbol{m}_k+\boldsymbol{\alpha})}{\Delta(\boldsymbol{m}_{k,\neg i}+\boldsymbol{\alpha})}\right)\cdot\left(\frac{\Delta(\boldsymbol{n}+\boldsymbol{\upsilon})}{\Delta(\boldsymbol{n}_{\neg i}+\boldsymbol{\upsilon})}\right) \qquad (16)$$

Equation (16) follows from substituting (14) into (15). Also, note that $\Delta(\boldsymbol{m}_k+\boldsymbol{\alpha})$ is obtained by removing all documents of user *i* from group y_i and adding them to group *k*. All product terms other than those for group *k* cancel out from the middle term of equation (14).

$$\begin{aligned}=&\left(\prod_{t=1}^{T}\frac{\prod_{m=1}^{V}\Gamma\left(n_t^{(m)}+\beta_m\right)\Gamma\left(\sum_{m=1}^{V}n_{t,\neg i}^{(m)}+\beta_m\right)}{\prod_{m=1}^{V}\Gamma\left(n_{t,\neg i}^{(m)}+\beta_m\right)\Gamma\left(\sum_{m=1}^{V}n_t^{(m)}+\beta_m\right)}\right)\\&\times\left(\frac{\prod_{t=1}^{T}\Gamma\left(m_k^{(t)}+\alpha_t\right)\Gamma\left(\sum_{t=1}^{T}m_{k,\neg i}^{(t)}+\alpha_t\right)}{\prod_{t=1}^{T}\Gamma\left(m_{k,\neg i}^{(t)}+\alpha_t\right)\Gamma\left(\sum_{t=1}^{T}m_k^{(t)}+\alpha_t\right)}\right)\\&\times\left(\frac{\prod_{k=1}^{K}\Gamma\left(n^{(k)}+\upsilon_k\right)\Gamma\left(\sum_{l=1}^{K}n_{\neg i}^{(l)}+\upsilon_l\right)}{\prod_{k=1}^{K}\Gamma\left(n_{\neg i}^{(k)}+\upsilon_k\right)\Gamma\left(\sum_{l=1}^{K}n^{(l)}+\upsilon_l\right)}\right)\end{aligned}$$

$$= \left(\prod_{t=1}^{T} \frac{\prod_{m=1}^{V} \prod_{j=0}^{n_m - 1} \left(n_{t,\neg i}^{(m)} + \beta_m + j \right)}{\prod_{j=0}^{\#actions_i^t - 1} \left(\sum_{m=1}^{V} n_{t,\neg i}^{(m)} + \beta_m + j \right)} \right) \times \left(\frac{\prod_{t=1}^{T} \prod_{j=0}^{m_{y_i,i}^{(t)} - 1} \left(m_{k,\neg i}^{(t)} + \alpha_t + j \right)}{\prod_{j=0}^{\left(\sum_{t=1}^{T} m_{y_i,i}^{(t)} \right) - 1} \left(\sum_{t=1}^{T} m_{k,\neg i}^{(t)} + \alpha_t + j \right)} \right) \times \left(\frac{n_{\neg i}^{(k)} + \upsilon_k}{\left[\sum_{l=1}^{K} n^{(l)} + \upsilon_l \right] - 1} \right)$$

In the above expression, $\sum_{t=1}^{T} m_{y_i,i}^{(t)}$ is the total number of documents for user *i* (who belongs to group y_i) and $m_{y_i,i}^{(t)}$ is the total number of documents the user *i* has in topic *t*. After removing all the terms that do not depend on '*k*', we have:

$$p(y_i = k | \boldsymbol{Y}_{\neg i}, \boldsymbol{z}_i, \boldsymbol{Z}_{\neg i}, \boldsymbol{X}) \propto \left(\frac{\prod_{t=1}^{T} \prod_{j=0}^{m_{y_i,i}^{(t)} - 1} \left(m_{k,\neg i}^{(t)} + \alpha_t + j \right)}{\prod_{j=0}^{\left(\sum_{t=1}^{T} m_{y_i,i}^{(t)} \right) - 1} \left(\sum_{t=1}^{T} m_{k,\neg i}^{(t)} + \alpha_t + j \right)} \right) \times \left(n_{\neg i}^{(k)} + \upsilon_k \right) \quad (17)$$

We now compute $p(z_i = t | y_i, \boldsymbol{Z}_{\neg i}, \boldsymbol{X})$. In this case, $i \equiv (p,q)$ refers to the *p*-th document of the *q*-th user. All documents of a user belong to the same group. For ease of notation, now z_i and y_i now refer to the topic and user-group associated the *i*-th document respectively. We observe that once y_i is given, we know the group to which z_i belongs and it is independent of the topic distributions of other groups. z_i is also independent of all documents in other groups. Therefore, $p(z_i = t | y_i, \boldsymbol{Z}_{\neg i}, \boldsymbol{X}) = p\left(z_i = t \middle| y_i, \boldsymbol{Z}_{\neg i}^{y_i}, \boldsymbol{X}^{y_i} \right)$ where $\left(\boldsymbol{Z}_{\neg i}^{y_i}, \boldsymbol{X}^{y_i} \right)$ refers to the data in the group y_i only. The following derivation is analogous to (6) with only difference being the superscripts y_i.

$$p\left(z_i = t \middle| y_i, \boldsymbol{Z}_{\neg i}^{y_i}, \boldsymbol{X}^{y_i} \right) = \frac{p(y_i, \boldsymbol{Z}^{y_i}, \boldsymbol{X}^{y_i})}{p\left(y_i, \boldsymbol{Z}_{\neg i}^{y_i}, \boldsymbol{X}^{y_i} \right)} = \frac{p(\boldsymbol{X}^{y_i} | \boldsymbol{Z}^{y_i})}{p\left(\boldsymbol{X}_{\neg i}^{y_i} | \boldsymbol{Z}_{\neg i}^{y_i} \right) p\left(\boldsymbol{w}_i^{y_i} \right)} \cdot \frac{p(\boldsymbol{Z}^{y_i})}{p\left(\boldsymbol{Z}_{\neg i}^{y_i} \right)}$$

$$\propto \frac{\Delta\left(\boldsymbol{n}_t^{y_i} + \boldsymbol{\beta} \right)}{\Delta\left(\boldsymbol{n}_{t,\neg i}^{y_i} + \boldsymbol{\beta} \right)} \cdot \frac{\Delta\left(\boldsymbol{m}_{y_i} + \boldsymbol{\alpha} \right)}{\Delta\left(\boldsymbol{m}_{y_i,\neg i} + \boldsymbol{\alpha} \right)}$$

$$= \frac{\prod_{m \in \boldsymbol{X}_i} \Gamma\left(n_t^{y_i,m} + \beta_m \right) \Gamma\left(\sum_{m=1}^{V} n_{t,\neg i}^{y_i,m} + \beta_m \right)}{\prod_{m \in \boldsymbol{X}_i} \Gamma\left(n_{t,\neg i}^{y_i,m} + \beta_m \right) \Gamma\left(\sum_{m=1}^{V} n_t^{y_i,m} + \beta_m \right)} \times \frac{\Gamma\left(m_{y_i}^{(t)} + \alpha_t \right) \Gamma\left(\sum_{t=1}^{T} m_{y_i,\neg i}^{(t)} + \alpha_t \right)}{\Gamma\left(m_{y_i,\neg i}^{(t)} + \alpha_t \right) \Gamma\left(\sum_{t=1}^{T} m_{y_i}^{(t)} + \alpha_t \right)}$$

$$= \frac{\prod_{m \in \boldsymbol{X}_i} \prod_{j=0}^{f_m} \left(n_{t,\neg i}^{y_i,m} + \beta_m + j \right)}{\prod_{j=0}^{length(\boldsymbol{X}_i)} \left(\sum_{m=1}^{V} n_{t,\neg i}^{y_i,m} + \beta_m + j \right)} \times \frac{m_{y_i,\neg i}^{(t)} + \alpha_t}{\left[\sum_{t=1}^{T} m_{y_i}^{(t)} + \alpha_t \right] - 1}$$

$$\propto \frac{\prod_{m \in X_i} \prod_{j=0}^{f_m} \left(n_{t,\neg i}^{y_i,m} + \beta_m + j\right)}{\prod_{j=0}^{length(X_i)} \left(\sum_{m=1}^{V} n_{t,\neg i}^{y_i,m} + \beta_m + j\right)} \times \left(m_{y_i,\neg i}^{(t)} + \alpha_t\right) \quad (18)$$

In the above, $n_t^{y_i,m}$ is the count of occurrences of word *m* in topic *t* in group y_i. $m_{y_i}^{(t)}$ is the total number of documents of topic *t* in group y_i. '$m_{y_i,\neg i}^{(t)}$' indicates that the document *i* was excluded. f_m is the count of occurrences of word *m* in document *i*. $length(\boldsymbol{X}_i)$ is the total count of words in document *i*. $\{m \in \boldsymbol{X}_i\}$ is the set of unique words in document *i*.

The multinomial parameters $\boldsymbol{\theta}_k$ for the group *k* can be computed as:

$$p\left(\boldsymbol{\theta}_k | \boldsymbol{Z}^{(k)}, \boldsymbol{\alpha}\right) = \frac{1}{Z_{\boldsymbol{\theta}_k}} \prod_{i=1}^{N^{(k)}} p\left(z_i^{(k)} | \boldsymbol{\theta}_k\right) \cdot p(\boldsymbol{\theta}_k | \boldsymbol{\alpha}) = Dir(\boldsymbol{m}_k + \boldsymbol{\alpha}),$$
$$\boldsymbol{m}_k = \left\{m_k^{(t)}\right\}_{t=1}^{T}$$

Here, $N^{(k)}$ is the number of users in group *k*, and $\boldsymbol{Z}^{(k)}$ are the membership assignments for all users in group *k*. As already mentioned earlier, $m_k^{(t)}$ is the number of times topic *t* has been observed in group *k* across all users belonging to the group *k*.

Now, we compute the parameters $\boldsymbol{T}$. The word distribution within each topic is dependent on the entire dataset since they are shared by all groups.

$$p(\boldsymbol{\tau}_t | \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{X}, \alpha, \beta, \upsilon) = \frac{1}{Z_{\boldsymbol{\tau}_t}} \prod_{\{i: z_i = t\}} p(x_i | \boldsymbol{\tau}_t) \cdot p(\boldsymbol{\tau}_t | \boldsymbol{\beta}) = Dir(\boldsymbol{n}_t + \boldsymbol{\beta}),$$
$$\boldsymbol{n}_t = \left\{n_t^{(m)}\right\}_{m=1}^{V}$$

The group distribution can now be computed as:

$$p(\boldsymbol{\pi} | \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{X}, \alpha, \beta, \upsilon) = p(\boldsymbol{\pi} | \boldsymbol{Y}, \upsilon) = \frac{1}{Z_\pi} \prod_{\{i=1..N\}} p(y_i | \boldsymbol{\pi}) \cdot p(\boldsymbol{\pi} | \boldsymbol{\upsilon}) = Dir(\boldsymbol{n} + \boldsymbol{\upsilon}),$$
$$\boldsymbol{n} = \left\{n^{(k)}\right\}_{k=1}^{K}$$

Since both the above are Dirichlet distributions, we can compute the expectations easily:

$$\theta_k^t = \frac{m_k^t + \alpha_t}{\sum_{t'=1}^{T} m_k^{t'} + \alpha_{t'}} \quad (19)$$

$$\tau_{t,m} = \frac{n_t^{(m)} + \beta_m}{\sum_{m'=1}^{V} n_t^{(m')} + \beta_{m'}} \quad (20)$$

$$\pi_k = \frac{n^{(k)} + \upsilon_k}{\sum_{k'=1}^{K} n^{(k')} + \upsilon_{k'}} \quad (21)$$

References

[1] Heinrich, Gregor. Parameter Estimation for Text Analysis, 2008