# LDA - Collapsed Gibbs Sampling

document ← topic $\xleftarrow[\theta]{\text{distributed}}$ words

first step, second step

document { topic 1 $p_1$ { $w_1$ ⋮ $w_2$ ⋮ $w_n$

topic 2 $p_2$

⋮

topic n $p_3$

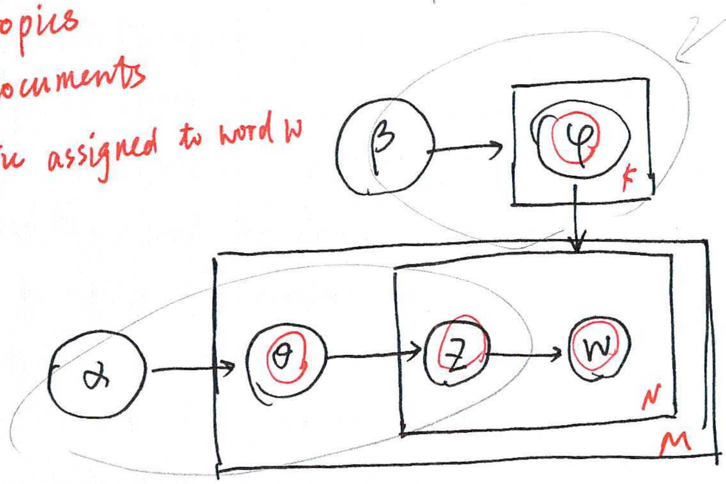
for topic ?

N: words

K: topics

M: documents

z : topic assigned to word w




for document ?

$\alpha, \beta$: parameters of two seperate Dirichelet distributions

Multinomial($\theta$) : distribution over topics for a given document

Multinomial($\varphi$) : distribution over words for a given topic

⇒ $\theta$ : $M \times K$ ← topic
↑ document

joint : $p(\theta, \phi, z, W \mid \alpha, \beta)$

$$= \prod_{j=1}^{K} P(\phi_j \mid \beta) \prod_{i=1}^{M} P(\theta_i \mid \alpha) \prod_{n=1}^{N} P(z_{j,n} \mid \theta_i) P(W_{i,n} \mid z_{i,n}, \phi_{z_{i,n}})$$

$\phi$ : $K \times N$ ← words
↑ topic

The generative process :

1) sample all $\theta$ from $Dir(\alpha)$

2) sample all $\varphi$ from $\cancel{\phi}$ $Dir(\beta)$

Here, $z_{n,i}$ only indicates a index of topic ?

3) For each word index $n$ in document $i$ where $i \in \{1, \ldots, M\}$

i) assign topic $z_{n,i}$ for $n$ from $Multi(\theta_i)$

ii) sample word from ~~Muti~~ $Muti(\phi_{z_{n,i}})$

Gibbs: We want to know $\theta, \varphi, z$ from joint distribution

$$P(\theta, \varphi, z, W \mid \alpha, \beta) = \prod_{j=1}^{K} P(\varphi_j \mid \beta) \prod_{i=1}^{M} P(\theta_i \mid \alpha) \prod_{n=1}^{N} P(z_{i,n} \mid \theta_i) P(W_{i,n} \mid z_{i,n}, \varphi_{z_{i,n}})$$

Posterior:

$\theta_{i,k}$ (the proportion of topic $k$ in document $i$) is $\theta_{i,k} = \dfrac{n_i^k + \alpha_k}{\sum_{k=1}^{K} n_i^k + \alpha_k}$

the number of words $\boxed{\text{in document}}$ $i$ that have been assigned to topic $k$

we can get it from $z_i$

$\varphi_{k,w}$ (the proportion of word $w$ in topic $k$) $\varphi_{k,w} = \dfrac{n_w^k + \beta_w}{\sum_{w=1}^{W} n_w^k + \beta_w}$

the number of times word $w$ is assigned to topic $k$ (over all documents in the corpus)

$Z$

$$P(z_i = j \mid z_{-i}, W) \propto \overset{①}{P}(W_i \mid z_i = j, z_{-i}, W_{-i}) \overset{②}{P}(z_i = j \mid z_{-i})$$

① $P(W_i \mid z_i = j, z_{-i}, W_{-i}) = \int \underbrace{P(W_i \mid z_i = j, \varphi_j)}_{= \varphi_{W_i, j}} P(\varphi_j \mid z_{-i}, W_{-i}) \, d\varphi_j$

$$P(W_i \mid z_i = j, z_{-i}, W_{-i}) = \frac{n_{-i,j}^{W_i} + \beta}{n_{-i,j} + W\beta}$$

$n_{-i,j}^{W_i}$: total # $w_i$ instances assigned to topic $j$, not including the current $w_i$

$n_{-i,j}$: total # words of words assigned to topic $j$ not including the current word

② $P(z_i = j \mid z_{-i}) = \int \underbrace{P(z_i = j \mid \theta_{d_i})}_{\theta_{d_i, j}} P(\theta_{d_i} \mid z_{-i}) \, d\theta_{d_i}$

$n_{-i,j}^{d_i}$: # words assigned to topic $j$ in doc $d_i$ not counting the current one.

$n_{-i}^{d_i}$: # words in doc $d_i$ not counting the current one

$$P(z_i = j \mid z_{-i}) = \frac{n_{-i,j}^{d_i} + \alpha}{n_{-i}^{d_i} + K\alpha}$$

so $$P(z_i = j \mid z_{-i}, W) \propto \frac{n_{-i,j}^{W_i} + \beta}{n_{-i,j} + W\beta} \cdot \frac{n_{-i,j}^{d_i} + \alpha}{n_{-i}^{d_i} + K\alpha}$$