LDA Gibbs sampling

贺成

2017.6.17

Target:

LDA的目标是找到每个词后潜藏（隐含）的主题，因此要计算后验概率：

$$p(\vec{z}|\vec{w}) = \frac{p(\vec{w},\vec{z})}{p(\vec{w})} = \frac{\prod_{i=1}^{W} p(w_i, z_i)}{\prod_{i=1}^{W} \sum_{k=1}^{K} p(w_i, z_i = k)}$$

文档中一个单词w_i 的概率是（W是语料中所有单词的个数）：

$$p(w_i) = \sum_{k=1}^{K} p(w_i, z_i = k)$$

因此：

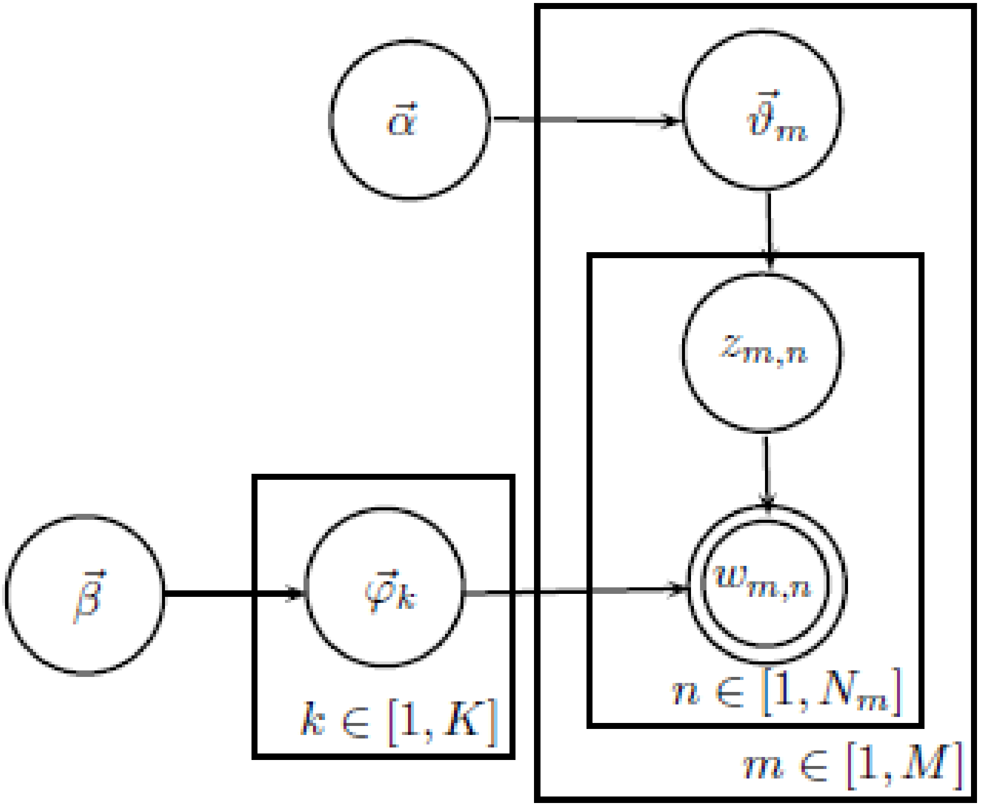
$$p(\vec{w}) = \prod_{i=1}^{W} \sum_{k=1}^{K} p(w_i, z_i = k)$$

分母时间复杂度：K^W,无法枚举

因此需要使用gibbs sampling，即用$p(z_i|\overrightarrow{z_{\neg i}},\vec{w})$来模拟$p(\vec{z}|\vec{w})$。

$$p(z_i=k|\overrightarrow{z_{\neg i}},\vec{w}) = \frac{p(\vec{w},\vec{z})}{p(\vec{w},\overrightarrow{z_{\neg i}})}$$

接下来先求$p(\vec{w},\vec{z})$：



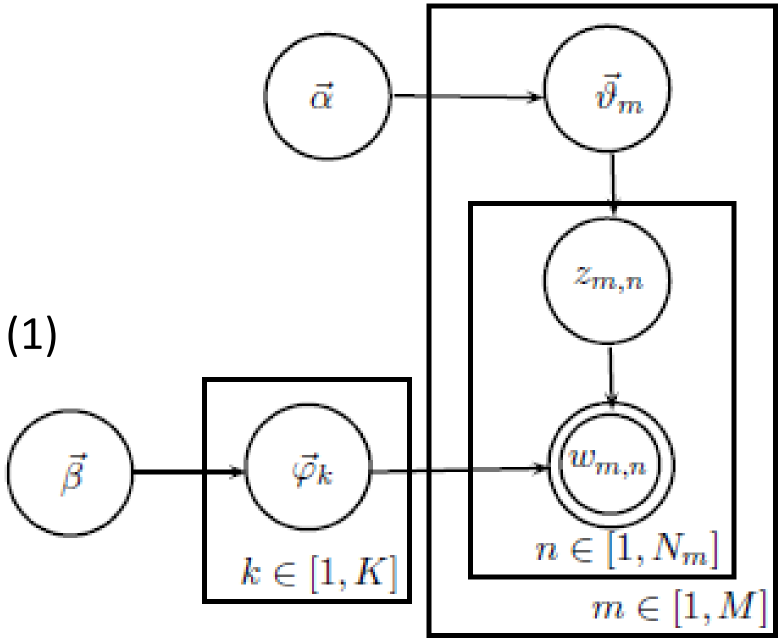
$$\begin{aligned} &p(\vec{w},\vec{z}|\vec{\alpha},\vec{\beta}) \\ &= p(\vec{w}|\vec{z},\vec{\alpha},\vec{\beta})p(\vec{z}|\vec{\alpha},\vec{\beta}) \qquad (2) \\ &= p(\vec{w}|\vec{z},\vec{\beta})p(\vec{z}|\vec{\alpha}) \qquad (3) \end{aligned}$$

注：(2)->(3)因为给定z后w和α无关，因此：
$p(\vec{w}|\vec{z},\vec{\alpha},\vec{\beta}) = p(\vec{w}|\vec{z},\vec{\beta})$
z只和α相关，因此：$p(\vec{z}|\vec{\alpha},\vec{\beta}) = p(\vec{z}|\vec{\alpha})$

求$p(\vec{z}|\vec{\alpha}\,)$:

Dirichlet分布：$\Delta(\alpha)$是归一化参数

$$p(\vec{p}|\vec{\alpha}) = Dirichlet(\vec{p}|\vec{\alpha}) = \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}p_k{}^{\alpha_k-1} = \frac{1}{\Delta(\vec{\alpha})}\prod_{k=1}^{K}p_k{}^{\alpha_k-1} \quad (1)$$



先来计算第m个文档的主题的条件分布：$p(\vec{z}_m|\vec{\alpha}\,)$

$$p(\vec{z}_m|\vec{\alpha}\,) = \int p(\vec{z}_m|\vec{\vartheta}_m)p(\vec{\vartheta}_m|\vec{\alpha})d\vec{\vartheta}_m \quad (1)$$

$$= \int \prod_{k=1}^{K}p_k{}^{n_m^{(k)}}\ Dirichlet(\vec{\alpha})\,d\vec{\vartheta}_m \quad (2)$$

$$= \int \prod_{k=1}^{K}p_k{}^{n_m^{(k)}}\frac{1}{\Delta(\vec{\alpha})}\prod_{k=1}^{K}p_k{}^{\alpha_k-1}d\vec{\vartheta}_m \quad (3)$$

$$= \frac{1}{\Delta(\vec{\alpha})}\int \prod_{k=1}^{K}p_k{}^{n_m^{(k)}+\alpha_k-1}d\vec{\vartheta}_m \quad (4)$$

$$= \frac{\Delta(\vec{n}_m+\vec{\alpha})}{\Delta(\vec{\alpha})} \quad (5)$$

注:(1)

$$p(\vec{z}_m|\vec{\alpha}\,) = \int p(\vec{z}_m,\vec{\vartheta}_m|\vec{\alpha}\,)d\vec{\vartheta}_m$$
$$= \int p(\vec{z}_m|\vec{\vartheta}_m,\vec{\alpha})p(\vec{\vartheta}_m|\vec{\alpha})d\vec{\vartheta}_m$$
$$= \int p(\vec{z}_m|\vec{\vartheta}_m)p(\vec{\vartheta}_m|\vec{\alpha})d\vec{\vartheta}_m$$

注:(4)->(5)

对(1)两端积分，得：$1 = \frac{1}{\Delta(\vec{\alpha})}\int\prod_{k=1}^{K}p_k{}^{\alpha_k-1}\,d\vec{p}$

故：$\Delta(\vec{\alpha}) = \int\prod_{k=1}^{K}p_k{}^{\alpha_k-1}\,d\vec{p}$

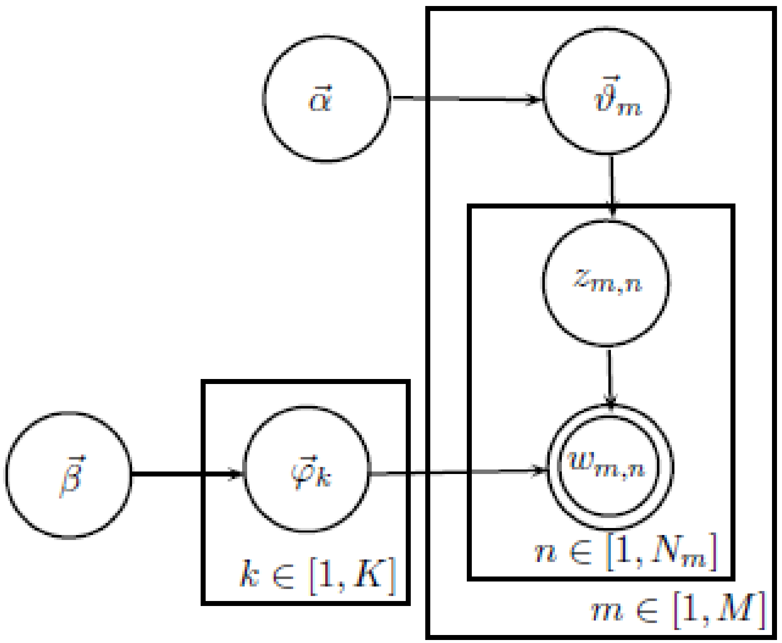
因此：$p(\vec{z}|\vec{\alpha}\,) = \prod_{m=1}^{M}\frac{\Delta(\vec{n}_m+\vec{\alpha})}{\Delta(\vec{\alpha})}$，其中在第m个文档中第k个主题的词的个数为: $n_m^{(k)}$,$\vec{n}_m = (n_m^{(1)}, n_m^{(2)}, \dots, n_m^{(K)})$

求$p(\vec{w}|\vec{z},\vec{\beta})$：

先求第k个主题对应的词的条件分布：

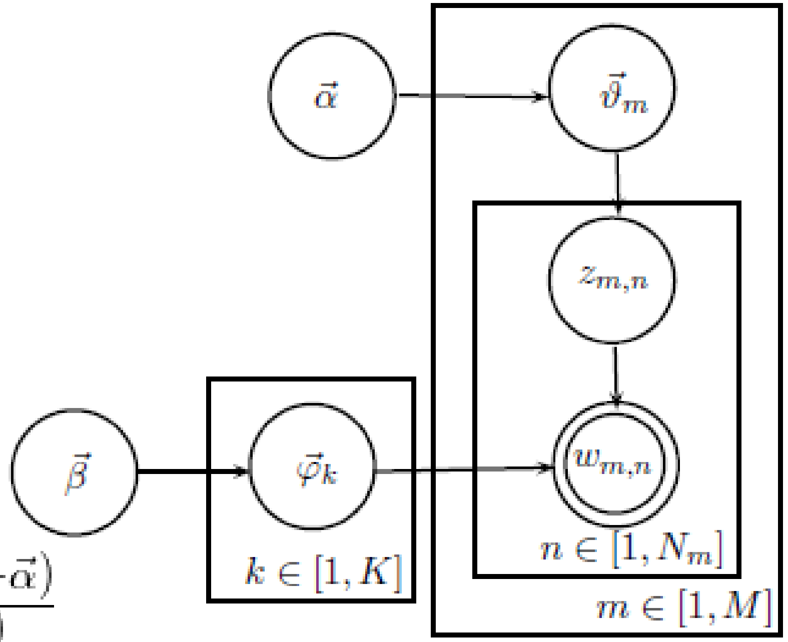
$$
\begin{aligned}
p(\vec{w}|\vec{z}_k,\vec{\beta}) &= \int p(\vec{w}|\vec{\varphi}_k,\vec{z}_k)\cdot p(\vec{\varphi}_k|\vec{\beta})d\vec{\varphi}_k \\
&= \int \prod_{v=1}^{V} p_{k,v}^{n_k^{(v)}}\cdot Dirichlet(\vec{\beta})d\vec{\varphi}_k \\
&= \int \prod_{v=1}^{V} p_{k,v}^{n_k^{(v)}}\cdot \frac{1}{\Delta(\vec{\beta})}\cdot \prod_{v=1}^{V} p_{k,v}^{\beta_v-1}d\vec{\varphi}_k \\
&= \frac{1}{\Delta(\vec{\beta})}\cdot \int \prod_{v=1}^{V} p_{k,v}^{n_k^{(v)}+\beta_v-1}d\vec{\varphi}_k \\
&= \frac{\Delta(\vec{n}_k+\vec{\beta})}{\Delta(\vec{\beta})}
\end{aligned}
$$

$$
\Rightarrow p(\vec{w}|\vec{z},\vec{\beta}) = \prod_{k=1}^{K}\frac{\Delta(\vec{n}_k+\vec{\beta})}{\Delta(\vec{\beta})}
$$



因此：$p(\vec{w},\vec{z}|\vec{\alpha},\vec{\beta}) = p(\vec{w}|\vec{z},\vec{\beta})p(\vec{z}|\vec{\alpha}\ )$

$$= \prod_{k=1}^{K}\frac{\Delta(\vec{n}_k+\vec{\beta})}{\Delta(\vec{\beta})} * \prod_{m=1}^{M}\frac{\Delta(\vec{n}_k+\vec{\alpha})}{\Delta(\vec{\alpha})}$$



i = (m,n)是二维下标，对应第m篇文档第n个词

$$p(z_i = k|\vec{z_{\neg i}},\vec{w}) = \frac{p(\vec{w},\vec{z})}{p(\vec{w},\vec{z_{\neg i}})} = \frac{\prod_{k=1}^{K}\frac{\Delta(\vec{n_k}+\vec{\beta})}{\Delta(\vec{\beta})}\cdot\prod_{m=1}^{M}\frac{\Delta(\vec{n_m}+\vec{\alpha})}{\Delta(\vec{\alpha})}}{\prod_{k=1}^{K}\frac{\Delta(\vec{n_{k,\neg i}}+\vec{\beta})}{\Delta(\vec{\beta})}\cdot\prod_{m=1}^{M}\frac{\Delta(\vec{n_{m,\neg i}}+\vec{\alpha})}{\Delta(\vec{\alpha})}}$$

$$\propto \frac{\Delta(\vec{n_k}+\vec{\beta})}{\Delta(\vec{n_{k,\neg i}}+\vec{\beta})}\cdot\frac{\Delta(\vec{n_m}+\vec{\beta})}{\Delta(\vec{n_{m,\neg i}}+\vec{\beta})}$$

计算$\dfrac{\Delta(\vec{n}_k+\vec{\beta})}{\Delta(n_{k,\neg i}+\vec{\beta})}$:

$n_k^{(t)}$第k个主题中第t个词的个数，$n_k^{(t)} = n_{k,\neg i}^{(t)} + 1$。把当前词（第i个词）排除，即第i个词的统计量减1。

$$\frac{\Delta(\vec{n}_k + \vec{\beta})}{\Delta(n_{k,\neg i} + \vec{\beta})} = \frac{\dfrac{\prod_{t=1}^{V}\Gamma(n_k^{(t)} + \beta_t)}{\Gamma(\sum_{t=1}^{V}(n_k^{(t)} + \beta_t)}}{\dfrac{\prod_{t=1}^{V}\Gamma(n_{k,\neg i}^{(t)} + \beta_t)}{\Gamma(\sum_{t=1}^{V}(n_{k,\neg i}^{(t)} + \beta_t)}}$$

$$= \frac{\prod_{t=1}^{V}\Gamma(n_k^{(t)}+\beta_t)}{\Gamma(\sum_{t=1}^{V}(n_k^{(t)}+\beta_t))}\cdot\frac{\Gamma\left(\sum_{t=1}^{V}(n_{k,\neg i}^{(t)}+\beta_t)\right)}{\prod_{t=1}^{V}\Gamma\left(n_{k,\neg i}^{(t)}+\beta_t\right)}$$

$$= \frac{\Gamma(n_k^{(t)}+\beta_t)}{\Gamma\left(n_{k,\neg i}^{(t)}+\beta_t\right)}\cdot\frac{1}{\sum_{t=1}^{V}(n_{k,\neg i}^{(t)}+\beta_t)}$$

$$= \frac{n_{k,\neg i}^{(t)}+\beta_t}{\sum_{t=1}^{V}(n_{k,\neg i}^{(t)}+\beta_t)} = \frac{n_{k,\neg i}^{(t)}+\beta_t}{\sum_{t=1}^{V} n_{k,\neg i}^{(t)}+V\beta}$$

$$\frac{\prod_{t=1}^{V}\Gamma(n_k^{(t)} + \beta_t)}{\prod_{t=1}^{V}\Gamma\left(n_{k,\neg i}^{(t)} + \beta_t\right)}$$

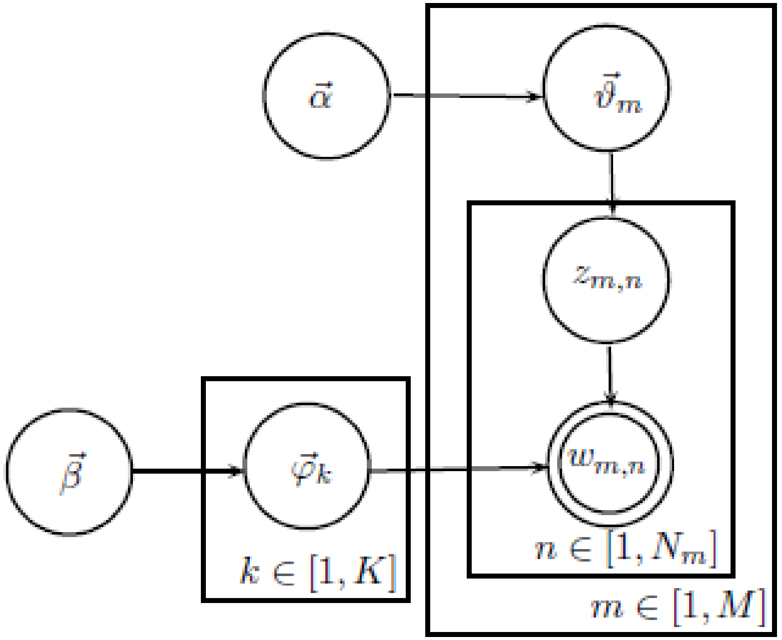
$$= \frac{\Gamma(n_k^{(1)} + \beta_1)\cdot\Gamma(n_k^{(2)} + \beta_2)\cdots\Gamma(n_k^{(i)} + \beta_i)\cdots\Gamma(n_k^{(V)} + \beta_V)}{\Gamma(n_k^{(1)} + \beta_1)\cdot\Gamma(n_k^{(2)} + \beta_2)\cdots\Gamma(n_k^{(i)} - 1 + \beta_i)\cdots\Gamma(n_k^{(V)} + \beta_V)}$$

注：通常设置所有α都相等，即$\alpha_1 = \alpha_2 = .. = \alpha$,所有β都相等，即$\beta_1 = \beta_2 = .. = \beta$

同理：$\dfrac{\Delta(\vec{n}_m+\vec{\alpha})}{\Delta(\vec{n}_{m,\neg i}+\vec{\alpha})} = \dfrac{n_{m,\neg i}^{(t)}+\alpha_k}{\sum_{k=1}^{K} n_{m,\neg i}^{(k)}+K\alpha}$

综上：

$$p(z_i = k|\vec{z}_{\neg i},\vec{w}) \propto \frac{n_{k,\neg i}^{(t)} + \beta_t}{\sum_{t=1}^{V} n_{k,\neg i}^{(t)} + V\beta} \cdot \frac{n_{m,\neg i}^{(t)} + \alpha_k}{\sum_{k=1}^{K} n_{m,\neg i}^{(k)} + K\alpha}$$



估计参数ϑ和Φ：

因为多项分布和dirichlet分布满足共轭关系，所以：

Dirichlet 先验+ 多项分布的数据$\longrightarrow$ 后验分布为Dirichlet 分布

$$Dir(\vec{p}|\vec{\alpha}) + MultCount(\vec{n}) = Dir(\vec{p}|\vec{\alpha} + \vec{n})$$

于是，在给定了参数$\vec{p}$的先验分布$Dir(\vec{p}|\vec{\alpha})$ 的时候，各个词出现频次的数据$\vec{n} \sim Mult(\vec{n}|\vec{p}, N)$ 为多项分布，所以无需计算，我们就可以推出后验分布是

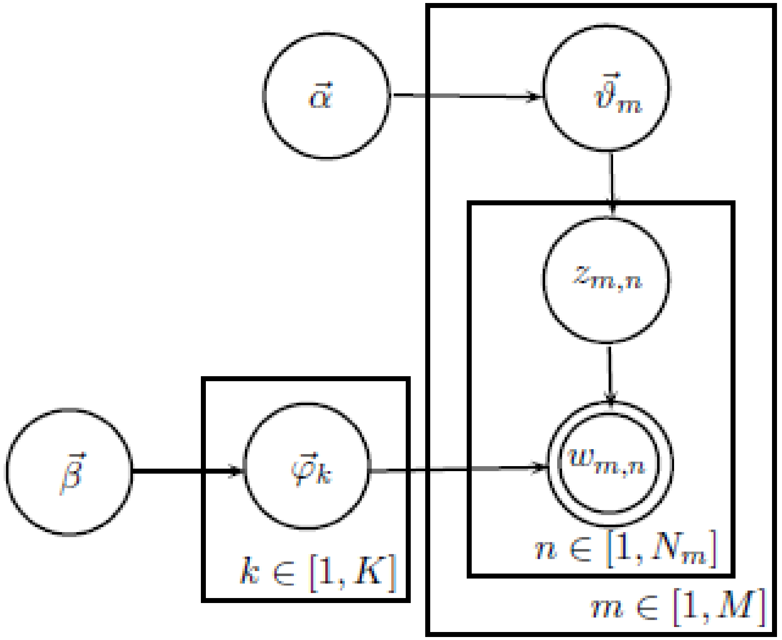
$$p(\vec{p}|\mathcal{W}, \vec{\alpha}) = Dir(\vec{p}|\vec{n} + \vec{\alpha}) = \frac{1}{\Delta(\vec{n} + \vec{\alpha})} \prod_{k=1}^{V} p_k^{n_k + \alpha_k - 1} d\vec{p}$$

在贝叶斯的框架下，参数$\vec{p}$如何估计呢？由于我们已经有了参数的后验分布，所以合理的方式是使用后验分布的极大值点，或者是参数在后验分布下的平均值。在该文档中，我们取平均值作为参数的估计值。使用上个小节中(17)式的结论，由于$\vec{p}$ 的后验分布为$Dir(\vec{p}|\vec{n} + \vec{\alpha})$，于是

$$E(\vec{p}) = \left(\frac{n_1 + \alpha_1}{\sum_{i=1}^{V}(n_i + \alpha_i)}, \frac{n_2 + \alpha_2}{\sum_{i=1}^{V}(n_i + \alpha_i)}, \cdots, \frac{n_V + \alpha_V}{\sum_{i=1}^{V}(n_i + \alpha_i)} \right)$$

也就是说对每一个p_i, 我们用下式做参数估计

$$\hat{p}_i = \frac{n_i + \alpha_i}{\sum_{i=1}^{V}(n_i + \alpha_i)}$$



所以：

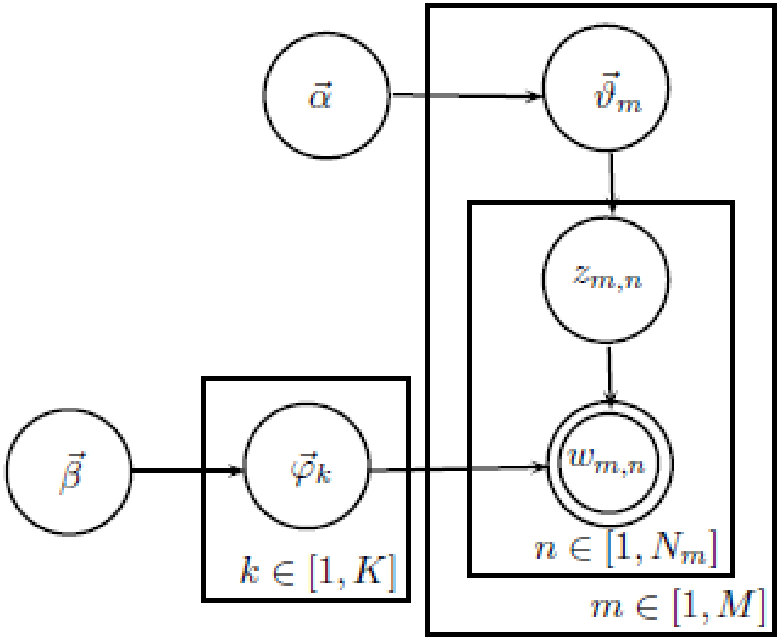
$$\varphi_{k,t} = \frac{n_k^{(t)} + \beta_t}{\sum_{t=1}^{V}(n_k^{(t)} + \beta_t)},$$

$$\vartheta_{m,k} = \frac{n_m^{(k)} + \alpha_k}{\sum_{k=1}^{K}(n_m^{(k)} + \alpha_k)}.$$

事件先验伪计数(prior pseudo-count)

第k个主题，第t个词

第m篇文章，第k个主题



参考资料:

- 《Parameter estimation for text analysis》
- 《LDA数学八卦》
- 《LDA数学漫游》
- 七月算法《主题模型》
- 徐亦达《MCMC系列》
- 吴立德《概率主题模型》
- 博客《通俗理解LDA主题模型》