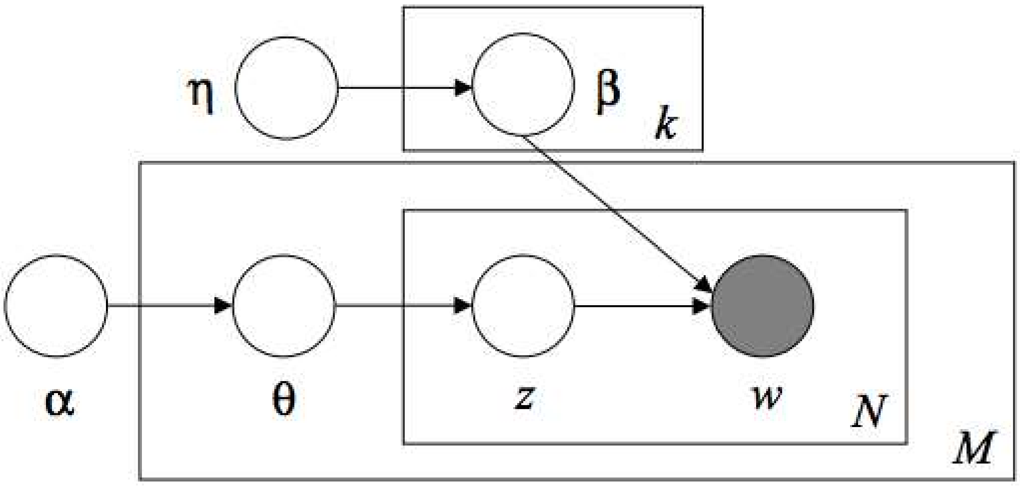
1. 联合分布



$$
\begin{aligned}
p(\theta, z, w, \beta|\alpha, \eta) &= p(\theta, z, w|\beta, \alpha)p(\beta|\eta) \\
&= p(\beta|\eta)\prod_{d=1}^{D} p(\theta_d, z_d, w_d|\beta, \alpha) \\
&= \prod_{k=1}^{K} p(\beta_k|\eta)\prod_{d=1}^{D}[p(\theta_d|\alpha)\prod_{i=1}^{L_d} p(z_{di}|\theta_d)p(w_{di}|z_{di}, \beta)] \\
&= \prod_{k=1}^{K} Dir(\beta_k|\eta)\prod_{d=1}^{D}[Dir(\theta_d|\alpha)\prod_{i=1}^{L_d} Cat(z_d|\theta_d)Cat(w_{di}|\beta_{z_{di}})]
\end{aligned}
$$

2. Gibbs Sampling

迪利克雷先验，Multinomial似然时的后验。

假设有i.i.d.的样本 $X_1, X_2, \cdots, X_N$ ，

$$p(X) \sim Cat(X|\pi)$$

π 的先验分布为 $Dir(\alpha)$

则 π 后验分布为?

$$
\begin{aligned}
p(\pi|D,\alpha) &= \frac{p(\pi|\alpha)\prod_{i=1}^{N} p(X_i|\pi)}{p(D)} \\
&\propto \prod_{k=1}^{K} \pi_k^{\alpha_k-1} \prod_{i=1}^{N}\prod_{k=1}^{K} \pi_k^{I(X_i=k)} \\
&= \prod_{k=1}^{K} \pi_k^{\alpha_k-1} \prod_{k=1}^{K} \pi_k^{N(X_i=k)} \\
&= \prod_{k=1}^{K} \pi_k^{N(X_i=k)+\alpha_k-1}
\end{aligned}
$$

所以

$$p(\pi|D,\alpha) \sim\sim Dir(\mathbf{N}+\alpha-1)$$

其中

$$N_k = N(X_i=k)$$

回到LDA

$$
\begin{aligned}
p(\theta_d|\cdot) &= Dir(\{\alpha_k + \sum_{i=1}^{L_d} I(z_{di}=k)\}) \\
p(\beta_k|\cdot) &= Dir(\{\eta_v + \sum_{d=1}^{D}\sum_{i=1}^{L_d} I(w_{di}=v, z_{di}=k)\}) \\
p(z_{di}=k|\cdot) &\propto \theta_{dk}\beta_{kw_{di}}
\end{aligned}
$$

3. Collapsed Gibbs sampling

通过积分消去 θ,β

$$
\begin{aligned}
p(z|\alpha) &= \prod_{d=1}^{D} p(z_d|\alpha) = \prod_{d=1}^{D}\int p(z_d|\theta_d)p(\theta_d|\alpha)d\theta_d \\
&= \prod_{d=1}^{D}\int p(\theta_d|\alpha)\prod_{i=1}^{L_d} p(z_{di}|\theta_d)d\theta_d \\
&= \prod_{d=1}^{D}\int Dir(\theta_d|\alpha)\prod_{i=1}^{L_d} Cat(z_{di}|\theta_d)d\theta_d
\end{aligned}
$$

由2中讨论 设

$$D = \{z_{di}, i=1,2,\cdots,L_d\}$$

则

$$
\begin{aligned}
p(\theta_d|D) &= \frac{p(D|\theta_d)p(\theta_d)}{p(D)} \\
&\sim Dir(C_d+\alpha) \\
p(D) &= \frac{B(C_d+\alpha)}{B(\alpha)} \\
&= \frac{\mathcal{T}(\sum_k \alpha_k)}{\prod_{k=1}^K \mathcal{T}(\alpha_k)} \frac{\prod_{k=1}^K \mathcal{T}(C_{dk}+\alpha_k)}{\mathcal{T}(\sum_k (C_{dk}+\alpha_k))}
\end{aligned}
$$

其中 C_{dk} 为文档d中主题为k(即 $z_{di}=k$)的单词个数。

实现中，向量 α 各维度取为一样，设为常数 α 。

则

$$
p(D) = \frac{\mathcal{T}(K\alpha)}{\mathcal{T}(\alpha)^K} \frac{\prod_{k=1}^K \mathcal{T}(C_{dk}+\alpha)}{\mathcal{T}(L_d+K\alpha))}
$$

最后

$$
p(z|\alpha) = (\frac{\mathcal{T}(K\alpha)}{\mathcal{T}(\alpha)^K})^D \prod_{d=1}^D \frac{\prod_{k=1}^K \mathcal{T}(C_{dk}+\alpha)}{\mathcal{T}(L_d+K\alpha))}
$$

同理

$$
p(w|z,\eta) = (\frac{\mathcal{T}(V\eta)}{\mathcal{T}(\eta)^V})^K \prod_{d=1}^K \frac{\prod_{v=1}^V \mathcal{T}(C_{vk}+\alpha)}{\mathcal{T}(C_k+V\eta))}
$$

其中 C_{vk} 为所有文档中单词v主题为k的个数。 C_k 为所有文档中主题为k的单词个数。

最终

$$
\begin{aligned}
p(z_{di}=k|z_{-di},w,\alpha,\eta) &= \frac{p(z,w|\alpha,\eta)}{p(z_{-di},w|\alpha,\eta)} \\
&\propto \frac{\mathcal{T}(C_{vk}+\eta)}{\mathcal{T}(C_{vk}^-+\eta)} \frac{\mathcal{T}(C_k^-+V\eta)}{\mathcal{T}(C_k+V\eta)} \frac{\mathcal{T}(C_{ik}+\eta)}{\mathcal{T}(C_{ik}^-+\eta)} \\
&= \frac{C_{vk}^-+\eta}{C_k^-+V\eta}(C_{dk}^-+\alpha)
\end{aligned}
$$

其中 $v=w_{z_{di}}$