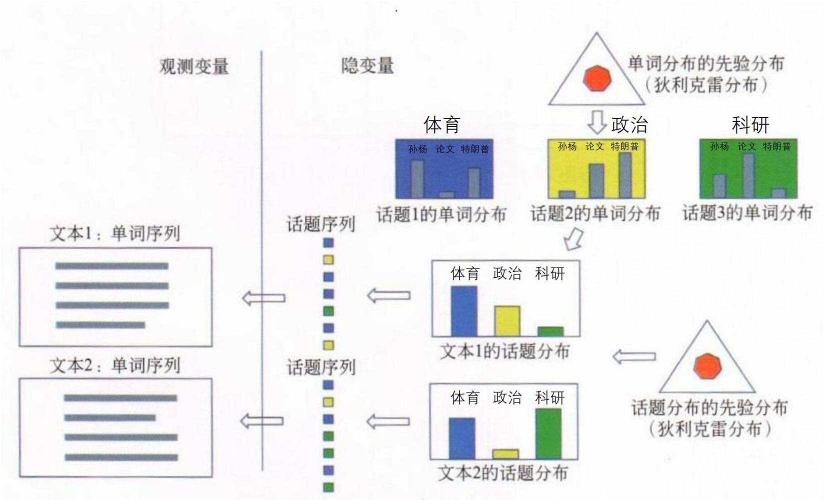
马尔可夫链蒙特卡罗法
在潜在狄利克雷分配中的应用
MCMC for LDA

西安交通大学管理学院
信息管理与电子商务系
智能决策与机器学习研究中心
刘佳鹏

LDA回顾

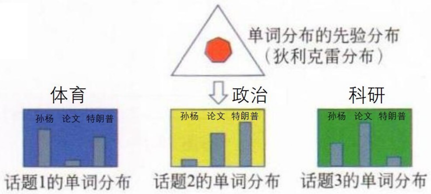

观测变量
隐变量
单词分布的先验分布
（狄利克雷分布）
体育
政治
科研
孙杨 论文 特朗普
孙杨 论文 特朗普
孙杨 论文 特朗普
话题1的单词分布
话题2的单词分布
话题3的单词分布
文本1：单词序列
话题序列
体育 政治 科研
文本1的话题分布
文本2：单词序列
话题序列
体育 政治 科研
文本2的话题分布
话题分布的先验分布
（狄利克雷分布）

LDA回顾

- 潜在狄利克雷分配（LDA）使用三个集合:
- (1) 单词集合 $W=\{w_1,\cdots,w_v,\cdots,w_V\}$, 其中 w_v 是第 v 个单词, $v=1,2,\cdots,V$, V 是单词的个数
- (2) 文本集合 $\boldsymbol{w}=\{\boldsymbol{w}_1,\cdots,\boldsymbol{w}_m,\cdots,\boldsymbol{w}_M\}$, 其中 $\boldsymbol{w}_m$ 是第 m 个文本, $m=1,2,\cdots,M$, M 是文本的个数
  - 文本 $\boldsymbol{w}_m$ 是一个单词序列 $\boldsymbol{w}_m=(w_{m1},\cdots,w_{mn},\cdots,w_{mN_m})$, 其中 w_{mn} 是文本 $\boldsymbol{w}_m$ 的第 n 个单词, $n=1,2,\cdots,N_m$, N_m 是文本 $\boldsymbol{w}_m$ 中单词的个数
- (3) 话题集合 $Z=\{z_1,\cdots,z_k,\cdots,z_K\}$, 其中 z_k 是第 k 个话题, $k=1,2,\cdots,K$, K 是话题的个数

LDA回顾

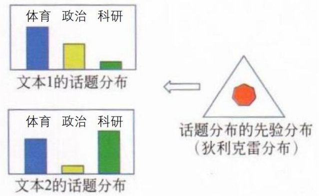
- **话题的单词分布及其先验分布：**



- 每一个话题 z_k 由一个“单词的条件概率分布 $p(w \mid z_k)$” 决定，$w \in W$
- 分布 $p(w \mid z_k)$ 服从多项分布(严格意义上类别分布)，其参数为 φ_k
  - 参数 φ_k 是一个 V 维向量 $\varphi_k = (\varphi_{k1}, \varphi_{k2}, \cdots, \varphi_{kV})$，其中 φ_{kv} 表示话题 z_k 生成单词 w_v 的概率
  - 所有话题的参数向量构成一个 $K \times V$ 矩阵 $\boldsymbol{\varphi} = \{\varphi_k\}_{k=1}^K$
  - 参数 φ_k 服从狄利克雷分布(先验分布)，其超参数为 β
  - 超参数 β 也是一个 V 维向量 $\beta = (\beta_1, \beta_2, \cdots, \beta_V)$

LDA回顾

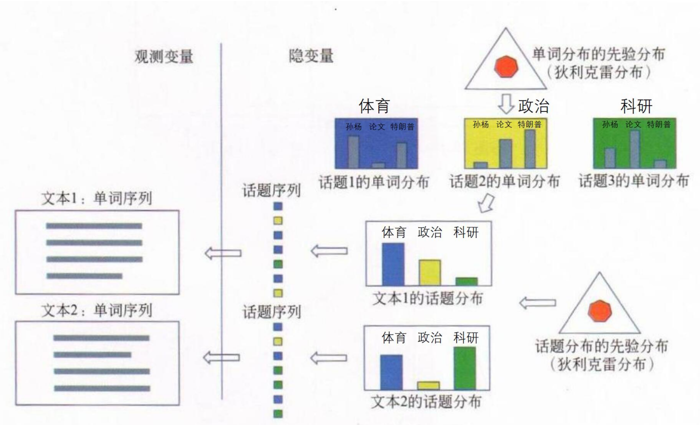
- **文本的话题分布及其先验分布：**



- 每一个文本 $\mathbf{w}_m$ 由一个“话题的条件概率分布 $p(z \mid \mathbf{w}_m)$” 决定，$z \in Z$
- 分布 $p(z \mid \mathbf{w}_m)$ 服从多项分布(严格意义上类别分布)，其参数为 θ_m
  - 参数 θ_m 是一个 K 维向量 $\theta_m = (\theta_{m1}, \theta_{m2}, \cdots, \theta_{mK})$，其中 θ_{mk} 表示文本 $\mathbf{w}_m$ 生成话题 z_k 的概率
  - 所有文本的参数向量构成一个 $M \times K$ 矩阵 $\boldsymbol{\theta} = \{\theta_m\}_{m=1}^{M}$
  - 参数 θ_m 服从狄利克雷分布(先验分布)，其超参数为 α
  - 超参数 α 也是一个 K 维向量 $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_K)$

LDA回顾

- 每一个文本 $\boldsymbol{w}_m$ 中的每一个单词 w_{mn} 由该文本的话题分布 $p(z \mid \boldsymbol{w}_m)$ 以及所有话题的单词分布 $p(w \mid z_k)$ 决定



LDA回顾

- LDA文本集合的生成过程如下:
- 给定单词集合 W, 文本集合 $\boldsymbol{w}$, 话题集合 Z, 狄利克雷分布的超参数 α 和 β
- (1) **生成话题的单词分布:**
  随机生成 K 个话题的单词分布。具体过程如下, 按照狄利克雷分布 $\mathrm{Dir}(\beta)$ 随机生成一个参数向量 φ_k, $\varphi_k \sim \mathrm{Dir}(\beta)$, 作为话题 z_k 的单词分布 $p(w \mid z_k), w \in W, k = 1, 2, \cdots, K$

LDA回顾

- (2) **生成文本的话题分布:**
  随机生成 M 个文本的话题分布。具体过程如下：按照狄利克雷分布 $\mathrm{Dir}(\alpha)$ 随机生成一个参数向量 $\theta_m, \theta_m \sim \mathrm{Dir}(\alpha)$, 作为文本 $\boldsymbol{w}_m$ 的话题分布 $p\left(z \mid \boldsymbol{w}_m\right), m=1,2,\cdots,M$

LDA回顾

- (3) **生成文本的单词序列:**
  随机生成 M 个文本的 N_m 个单词。文本 $\boldsymbol{w}_m(m=1,2,\cdots,M)$ 的单词 $w_{mn}(n=1,2,\cdots,N_m)$ 的生成过程如下:
  (3-1) 首先按照多项分布 $\operatorname{Mult}(\theta_m)$ 随机生成一个话题 $z_{mn}, z_{mn}\sim\operatorname{Mult}(\theta_m)$
  (3-2) 然后按照多项分布 $\operatorname{Mult}(\varphi_{z_{mn}})$ 随机生成一个单词 $w_{mn}, w_{mn}\sim\operatorname{Mult}(\varphi_{z_{mn}})$
- 注: 文本 $\boldsymbol{w}_m$ 本身是单词序列 $\boldsymbol{w}_m=(w_{m1},w_{m2},\cdots,w_{mN_m})$, 对应着隐式的话题序列$\boldsymbol{z}_m=(z_{m1},z_{m2},\cdots,z_{mN_m})$

LDA回顾

（LDA 的文本生成算法）

（1）对于话题 z_k（$k = 1, 2, \cdots, K$）：

生成多项分布参数 $\varphi_k \sim \mathrm{Dir}(\beta)$，作为话题的单词分布 $p(w|z_k)$；

（2）对于文本 $\mathbf{w}_m$（$m = 1, 2, \cdots, M$）：

生成多项分布参数 $\theta_m \sim \mathrm{Dir}(\alpha)$，作为文本的话题分布 $p(z|\mathbf{w}_m)$；

（3）对于文本 $\mathbf{w}_m$ 的单词 w_{mn}（$m = 1, 2, \cdots, M$，$n = 1, 2, \cdots, N_m$）：

（a）生成话题 $z_{mn} \sim \mathrm{Mult}(\theta_m)$，作为单词对应的话题；

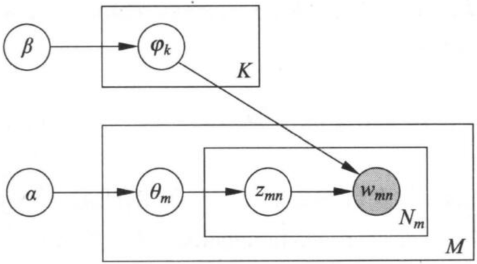
（b）生成单词 $w_{mn} \sim \mathrm{Mult}(\varphi_{z_{mn}})$。

LDA回顾

- LDA 的文本生成过程中，假定话题个数 K 给定，实际通常通过实验选定
- 狄利克雷分布的超参数 α 和 β 通常也是事先给定的
  - 在没有其他先验知识的情况下，可以假设向量 α 和 β 的所有分量均为 1，这时的文本的话题分布 θ_m 是对称的，话题的单词分布 φ_k 也是对称的

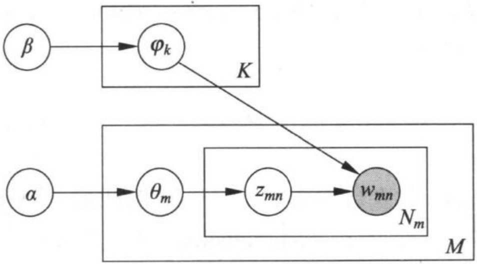
LDA回顾

- LDA模型本质是一种概率图模型（probabilistic graphical model）
- 下图为LDA作为概率图模型的板块表示（plate notation），亦称为盘式记法



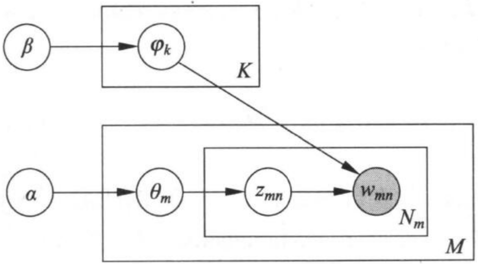
- 结点表示随机变量，实心结点是观测变量，空心结点是隐变量
- 有向边表示概率依存关系
- 矩形（板块）表示重复，板块内数字表示重复的次数

LDA回顾



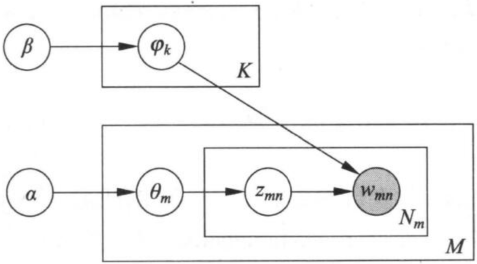
- 结点 α 和 β 是模型的超参数，结点 φ_k 表示话题的单词分布的参数，结点 θ_m 表示文本的话题分布的参数，结点 z_{mn} 表示话题，结点 w_{mn} 表示单词

LDA回顾



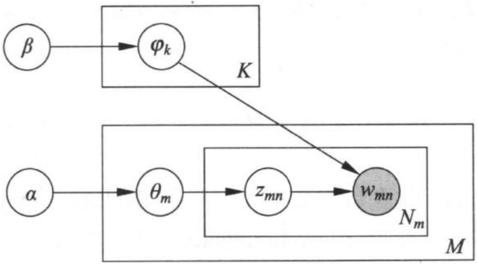
- 结点 β 指向结点 φ_k, 重复 K 次, 表示根据超参数 β 生成 K 个话题的单词分布的参数 φ_k

LDA回顾



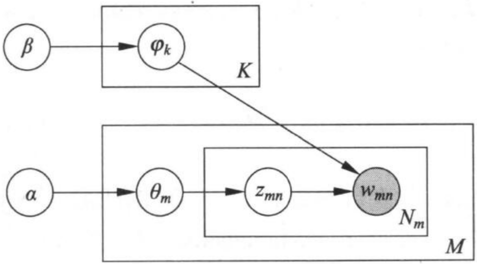
- 结点 α 指向结点 θ_m，重复 M 次，表示根据超参数 α 生成 M 个文本的话题分布的参数 θ_m

LDA回顾



- 结点 θ_m 指向结点 z_{mn}，重复 N_m 次，表示根据文本的话题分布 θ_m 生成 N_m 个话题 z_{mn}

LDA回顾

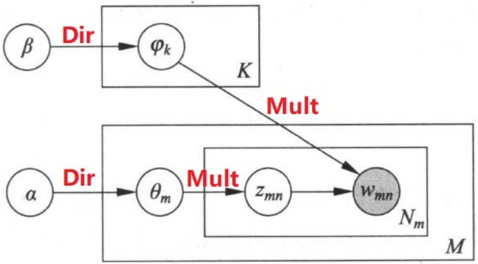


- 结点 z_{mn} 指向结点 w_{mn}，同时 K 个结点 φ_k 也指向结点 w_{mn}，表示根据话题 z_{mn} 以及 K 个话题的单词分布 φ_k 生成单词 w_{mn}

MCMC for LDA

- LDA模型中文本的单词序列是观测变量，文本的话题序列是隐变量，文本的话题分布和话题的单词分布也是隐变量
- 利用LDA进行话题分析，就是对给定文本集合，学习到每个文本的话题分布，以及每个话题的单词分布
- 这就是LDA模型的学习目标：给定文本集合，通过后验概率分布的估计，推断模型的所有参数

MCMC for LDA



- 未知参数：（1）话题的单词分布 $\varphi_{1:K}$ （2）文本的话题分布 $\theta_{1:M}$ （3）话题变量 z_{mn}，$m = 1, \cdots, M$，$n = 1, \cdots, N_m$
- 已知部分：（1）观测数据：文本的单词序列 $\boldsymbol{w} = \{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_m, \cdots, \boldsymbol{w}_M\}$ （2）超参数 α、β
- 目标：后验分布

$$p(\varphi_{1:K}, \theta_{1:M}, \{z_{mn}\} \mid \boldsymbol{w}, \alpha, \beta)$$

MCMC for LDA

- 目标：后验分布 $p(\varphi_{1:K}, \theta_{1:M}, \{z_{mn}\} \mid \boldsymbol{w}, \alpha, \beta)$
- 困难：该后验分布没有闭式解/解析解（closed-form/analytical solution）/ 直接计算该后验分布是不可行的（intractable）
- 用 $\Phi = \{\varphi_{1:K}, \theta_{1:M}, \{z_{mn}\}\}$ 表示参数集合

$$
\begin{aligned}
& p(\varphi_{1:K}, \theta_{1:M}, \{z_{mn}\} \mid \boldsymbol{w}, \alpha, \beta) \\
&= p(\Phi \mid \boldsymbol{w}, \alpha, \beta) \\
&= \frac{p(\Phi \mid \alpha, \beta) p(\boldsymbol{w} \mid \Phi)}{p(\boldsymbol{w} \mid \alpha, \beta)} \\
&= \frac{p(\Phi \mid \alpha, \beta) p(\boldsymbol{w} \mid \Phi)}{\int_\Phi p(\boldsymbol{w}, \Phi \mid \alpha, \beta) \mathrm{d}\Phi} \\
&= \frac{p(\Phi \mid \alpha, \beta) p(\boldsymbol{w} \mid \Phi)}{\int_\Phi p(\Phi \mid \alpha, \beta) p(\boldsymbol{w} \mid \Phi) \mathrm{d}\Phi}
\end{aligned}
$$

归一化因子 $\int_\Phi p(\Phi \mid \alpha, \beta) p(\boldsymbol{w} \mid \Phi) \mathrm{d}\Phi$ 没有解析解

MCMC for LDA

- 目标：后验分布 $p(\varphi_{1:K}, \theta_{1:M}, \{z_{mn}\} \mid \boldsymbol{w}, \alpha, \beta)$
- 采样方法：采样一组满足后验分布 $p(\varphi_{1:K}, \theta_{1:M}, \{z_{mn}\} \mid \boldsymbol{w}, \alpha, \beta)$ 的样本模拟该后验分布
- 同时采样所有参数的样本是困难的

$$(\varphi_{1:K}, \theta_{1:M}, \{z_{mn}\}) \sim p(\varphi_{1:K}, \theta_{1:M}, \{z_{mn}\} \mid \boldsymbol{w}, \alpha, \beta)$$

  因为难以考虑变量之间的相关关系

MCMC for LDA

- 采用分块采样的方法

$$
\begin{aligned}
&\text{for } k = 1, \cdots, K, \\
&\varphi_k \sim p(\varphi_k \mid \boldsymbol{w}, \alpha, \beta, \Phi \setminus \{\varphi_k\}) \\
&\quad = p(\varphi_k \mid \boldsymbol{w}, \alpha, \beta, \varphi_{k'=1,\cdots,k-1,k+1,\cdots,K}, \theta_{1:M}, \{z_{mn}\}_{m=1,\cdots,M,n=1,\cdots,N_m}),
\end{aligned}
$$

$$
\begin{aligned}
&\text{for } m = 1, \cdots, M, \\
&\theta_m \sim p(\theta_m \mid \boldsymbol{w}, \alpha, \beta, \Phi \setminus \{\theta_m\}) \\
&\quad = p(\theta_m \mid \boldsymbol{w}, \alpha, \beta, \varphi_{1:K}, \theta_{m'=1,\cdots,m-1,m+1,\cdots,M}, \{z_{m'n'}\}_{m'=1,\cdots,M,n'=1,\cdots,N_{m'}}),
\end{aligned}
$$

$$
\begin{aligned}
&\text{for } m = 1, \cdots, M, \quad n = 1, \cdots, N_m, \\
&z_{mn} \sim p(z_{mn} \mid \boldsymbol{w}, \alpha, \beta, \Phi \setminus \{z_{mn}\}) \\
&\quad = p(z_{mn} \mid \boldsymbol{w}, \alpha, \beta, \varphi_{1:K}, \theta_{m'=1,\cdots,M}, \{z_{m'n'}\}_{m'=1,\cdots,M,n'=1,\cdots,N_{m'}} \setminus \{z_{mn}\}),
\end{aligned}
$$

MCMC for LDA

- 设定采样次数 T (e.g., 10000)
- 随机指定初始样本 $\varphi_{1:K}^{(0)}, \theta_{1:M}^{(0)}, \{z_{mn}^{(0)}\}$
- 采样过程如下：

$$
\begin{aligned}
&\text{for } t = 1, \cdots, T, \\
&\quad \text{for } k = 1, \cdots, K, \\
&\qquad \varphi_k^{(t)} \sim p(\varphi_k \mid \boldsymbol{w}, \alpha, \beta, \varphi_{k'=1,\cdots,k-1}^{(t)}, \varphi_{k'=k+1,\cdots,K}^{(t-1)}, \theta_{1:M}^{(t-1)}, \{z_{mn}^{(t-1)}\}_{m=1,\cdots,M,n=1,\cdots,N_m}), \\
&\quad \text{for } m = 1, \cdots, M, \\
&\qquad \theta_m^{(t)} \sim p(\theta_m \mid \boldsymbol{w}, \alpha, \beta, \varphi_{1:K}^{(t)}, \theta_{m'=1,\cdots,m-1}^{(t)}, \theta_{m'=m+1,\cdots,M}^{(t-1)}, \{z_{m'n'}^{(t-1)}\}_{m'=1,\cdots,M,n'=1,\cdots,N_{m'}}), \\
&\quad \text{for } m = 1, \cdots, M, \quad n = 1, \cdots, N_m, \\
&\qquad z_{mn}^{(t)} \sim p(z_{mn} \mid \boldsymbol{w}, \alpha, \beta, \varphi_{1:K}^{(t)}, \theta_{m'=1,\cdots,M}^{(t)}, \{z_{m'n'}^{(t)}\}_{m'=1,\cdots,m-1,n'=1,\cdots,N_{m'}}, \{z_{mn'}^{(t)}\}_{n'=1,\cdots,n-1}, \\
&\qquad\qquad \{z_{mn'}^{(t-1)}\}_{n'=n+1,\cdots,N_m}, \{z_{m'n'}^{(t-1)}\}_{m'=m+1,\cdots,M,n'=1,\cdots,N_{m'}})
\end{aligned}
$$

- 输出：样本 $\{\varphi_k^{(t)}\}$, $\{\theta_m^{(t)}\}$, $z_{mn}^{(t)}$, $\quad t = 1, \cdots, T$

MCMC for LDA

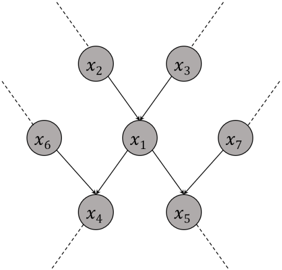
- 根据采样得到的满足后验分布 $p(\varphi_{1:K}, \theta_{1:M}, \{z_{mn}\} \mid \boldsymbol{w}, \alpha, \beta)$ 的样本

$$\{\varphi_k^{(t)}\}, \{\theta_m^{(t)}\}, z_{mn}^{(t)}, \ t = 1, \cdots, T,$$

可以计算模型参数（1）话题的单词分布 $\varphi_{1:K}$ （2）文本的话题分布 $\theta_{1:M}$ （3）话题变量 z_{mn}，$m = 1, \cdots, M$, $n = 1, \cdots, N_m$ 的统计量（如样本均值、样本方差等）

- 注：丢弃燃烧期(burn-in period)采集的样本，例如只使用$t = \frac{T}{2} + 1, \cdots, T$周期内的样本

MCMC for LDA



- 在上面的有向图中，结点 x_1 的双亲结点是 x_2 和 x_3，它的孩子结点是 x_4 和 x_5，孩子结点的其他双亲结点是 x_6 和 x_7
- 理论上可以证明，任意结点 x_i 的满条件分布仅与其双亲结点、孩子结点以及孩子结点的其他双亲结点相关（条件相关），而与其他结点无关（条件独立）

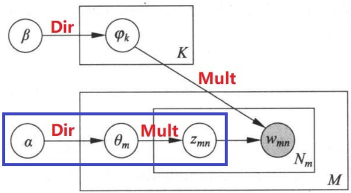
$$p(x_i \mid x_{-i}) = p\left(x_i \mid x_{\text{parent}(x_i)}, x_{\text{child}(x_i)}, x_{\text{parent}(\text{child}(x_i))}\right)$$

- 在上图的例子中

$$p(x_1 \mid x_{-1}) = p(x_1 \mid x_2, x_3, x_4, x_5, x_6, x_7)$$

- 有向图中与某结点 x_i 相关的部分被称为**马尔可夫毯**(Markov blanket)
- 应用上述结论可以简化LDA模型中的采样过程

MCMC for LDA

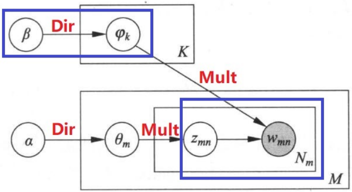


- **文本的话题分布 θ_m 的采样**

$$
\begin{aligned}
\theta_m &\sim p(\theta_m \mid \boldsymbol{w}, \alpha, \beta, \Phi \setminus \{\theta_m\}) \\
&= p(\theta_m \mid \boldsymbol{w}, \alpha, \beta, \varphi_{1:K}, \theta_{m'=1,\cdots,m-1,m+1,\cdots,M}, \{z_{m'n'}\}_{m'=1,\cdots,M, n'=1,\cdots,N_{m'}}) \\
&= p(\theta_m \mid \alpha, \{z_{mn}\}_{n=1,\cdots,N_m}) \propto p(\theta_m \mid \alpha) p(\{z_{mn}\}_{n=1,\cdots,N_m} \mid \theta_m) \\
&\propto p(\theta_m \mid \alpha) \prod_{n=1}^{N_m} p(z_{mn} \mid \theta_m) \propto \mathrm{Dir}(\theta_m \mid \alpha) \prod_{n=1}^{N_m} \mathrm{Mult}(z_{mn} \mid \theta_m) \\
&\propto \frac{1}{\mathrm{B}(\alpha)} \prod_{k=1}^{K} \theta_{mk}^{\alpha_k - 1} \cdot \prod_{n=1}^{N_m} \prod_{k=1}^{K} \theta_{mk}^{\mathbb{I}(z_{mn}=k)} \propto \frac{1}{\mathrm{B}(\alpha)} \prod_{k=1}^{K} \theta_{mk}^{\alpha_k - 1} \cdot \prod_{k=1}^{K} \theta_{mk}^{\sum_{n=1}^{N_m} \mathbb{I}(z_{mn}=k)} \\
&\propto \frac{1}{\mathrm{B}(\alpha)} \prod_{k=1}^{K} \theta_{mk}^{\alpha_k + \sum_{n=1}^{N_m} \mathbb{I}(z_{mn}=k) - 1} \qquad = \mathrm{Dir}(\theta_m \mid \alpha'_m)
\end{aligned}
$$

其中 $\alpha'_m = (\alpha'_{m1}, \cdots, \alpha'_{mK})$，$\alpha'_{mk} = \alpha_k + \sum_{n=1}^{N_m} \mathbb{I}(z_{mn} = k)$，$k = 1, \cdots, K$

MCMC for LDA



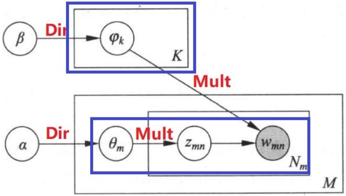
▶ **话题的单词分布 φ_k 的采样**

$$\varphi_k \sim p(\varphi_k \mid \boldsymbol{w}, \alpha, \beta, \Phi \setminus \{\varphi_k\})$$
$$= p(\varphi_k \mid \boldsymbol{w}, \alpha, \beta, \varphi_{k'=1,\cdots,k-1,k+1,\cdots,K}, \theta_{1:M}, \{z_{mn}\}_{m=1,\cdots,M,n=1,\cdots,N_m})$$
$$= p(\varphi_k \mid \beta, \boldsymbol{w}, \varphi_{k'=1,\cdots,k-1,k+1,\cdots,K}, \{z_{mn}\}_{m=1,\cdots,M,n=1,\cdots,N_m})$$
$$= p(\varphi_k \mid \beta) p(\boldsymbol{w} \mid \varphi_{1:K}, \{z_{mn}\}_{m=1,\cdots,M,n=1,\cdots,N_m}) = \mathrm{Dir}(\varphi_k \mid \beta) \prod_{m=1}^{M} \prod_{n=1}^{N_m} \varphi_{z_{mn}, i(w_{mn})}$$
$$= \mathrm{Dir}(\varphi_k \mid \beta) \prod_{k'=1}^{K} \prod_{v=1}^{V} \varphi_{k'v}^{\sum_{m=1}^{M} \sum_{n=1}^{N_m} \mathbb{I}(z_{mn}=k', w_{mn}=w_v)}$$
$$\propto \mathrm{Dir}(\varphi_k \mid \beta) \prod_{v=1}^{V} \varphi_{kv}^{\sum_{m=1}^{M} \sum_{n=1}^{N_m} \mathbb{I}(z_{mn}=k, w_{mn}=w_v)} = \mathrm{Dir}(\varphi_k \mid \beta'_k)$$

其中 $i(w_{mn}) \in \{1, \cdots, V\}$ 表示单词 w_{mn} 的索引，$\beta'_k = (\beta'_{k1}, \cdots, \beta'_{kV})$，

$$\beta'_{kv} = \beta_v + \sum_{m=1}^{M} \sum_{n=1}^{N_m} \mathbb{I}(z_{mn} = k, w_{mn} = w_v), \quad v = 1, \cdots, V$$

MCMC for LDA



- **话题变量 z_{mn} 的采样**

$$
\begin{aligned}
z_{mn} &\sim p(z_{mn} \mid \mathbf{w}, \alpha, \beta, \Phi \setminus \{z_{mn}\}) \\
&= p(z_{mn} \mid \mathbf{w}, \alpha, \beta, \varphi_{1:K}, \theta_{m'=1,\cdots,M}, \{z_{m'n'}\}_{m'=1,\cdots,M,n'=1,\cdots,N_{m'}} \setminus \{z_{mn}\}) \\
&= p(z_{mn} \mid \theta_m, w_{mn}, \varphi_{1:K}) \\
&= p(z_{mn} \mid \theta_m) p(w_{mn} \mid z_{mn}, \varphi_{1:K})
\end{aligned}
$$

从而有

$$
z_{mn} = k \propto \theta_{mk} \cdot \varphi_{k,i(w_{mn})}, \quad k = 1, \cdots, K
$$

其中 $i(w_{mn}) \in \{1, \cdots, V\}$ 表示单词 w_{mn} 的索引，进一步有

$$
p(z_{mn} = k) = \frac{\theta_{mk}\varphi_{k,i(w_{mn})}}{\sum_{k'=1}^{K} \theta_{mk'}\varphi_{k',i(w_{mn})}}, \quad k = 1, \cdots, K
$$

MCMC for LDA

- **LDA的Gibbs抽样算法：**
- 输入：文本的单词序列 $\boldsymbol{w} = \{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_m, \cdots, \boldsymbol{w}_M\}$，$\boldsymbol{w}_m = (w_{m1}, \cdots, w_{mn}, \cdots, w_{mN_m})$
- 参数：超参数 α 和 β，话题个数 K，采样次数 T
- 输出：满足文本的话题分布 θ_m 的后验分布的样本 $\{\theta_m^{(t)}\}_{t=1,\cdots,T}$，满足话题的单词分布 φ_k 的后验分布的样本 $\{\varphi_k^{(t)}\}_{t=1,\cdots,T}$，满足话题变量 z_{mn} 的后验分布的样本 $\{z_{mn}^{(t)}\}_{t=1,\cdots,T}$
- （1）为每个文本 $\boldsymbol{w}_m$ 引入计数变量 $r_m = (r_{m1}, \cdots, r_{mK})$，其中 r_{mk} 表示文本 $\boldsymbol{w}_m$ 中的话题 k 的计数，初值设为0；为每个主题 k 引入计数变量 $s_k = (s_{k1}, \cdots, s_{kV})$，其中 s_{kv} 表示话题 k 中的单词 v 的计数，初值设为0
- （2）对所有文本 $\boldsymbol{w}_m$，$m = 1, \cdots, M$ 中的所有单词 w_{mn}，$n = 1, \cdots, N_m$，
- 抽取话题变量 $z_{mn} = z_k \sim \mathrm{Mult}(\frac{1}{K})$
- 增加文本-话题计数 $r_{mk} = r_{mk} + 1$
- 增加话题-单词计数 $s_{k,i(w_{mn})} = s_{k,i(w_{mn})} + 1$
- （3）循环下列采样过程 T 次，$t = 1, \cdots, T$，
- （a）对所有文本 $\boldsymbol{w}_m$，$m = 1, \cdots, M$，抽取文本的话题分布 $\theta_m^{(t)} \sim \mathrm{Dir}(\alpha'_m)$，其中 $\alpha'_m = (\alpha'_{m1}, \cdots, \alpha'_{mK})$，$\alpha'_{mk} = \alpha_k + r_{mk}$，$k = 1, \cdots, K$
- （b）对所有话题 $k = 1, \cdots, K$，抽取话题的单词分布 $\varphi_k^{(t)} \sim \mathrm{Dir}(\beta'_k)$，其中 $\beta'_k = (\beta'_{k1}, \cdots, \beta'_{kV})$，$\beta'_{kv} = \beta_v + s_{kv}$，$v = 1, \cdots, V$
- （c）所有的 r_{mk} 和 r_{kv}，$m = 1, \cdots, M$，$k = 1, \cdots, K$，$v = 1, \cdots, V$ 归零
- （d）对所有文本 $\boldsymbol{w}_m$，$m = 1, \cdots, M$ 中的所有单词 w_{mn}，$n = 1, \cdots, N_m$，按照以下概率分布抽取话题变量 $z_{mn}^{(t)} = z_k$，并增加文本-话题计数 $r_{mk} = r_{mk} + 1$ 和话题-单词计数 $s_{k,i(w_{mn})} = s_{k,i(w_{mn})} + 1$

$$p(z_{mn}^{(t)} = k) = \frac{\theta_{mk}^{(t)} \varphi_{k,i(w_{mn})}^{(t)}}{\sum_{k'=1}^{K} \theta_{mk'}^{(t)} \varphi_{k',i(w_{mn})}^{(t)}}, \quad k = 1, \cdots, K$$