主题模型

在上一篇文章中介绍了pLSA模型，接下来介绍LDA模型

2. LDA

2.1 共轭分布

首先回顾下贝叶斯公式

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)}$$

其中$P(\theta)$是参数的先验概率，$P(\theta|x)$是给定的样本后，参数的后验概率。$P(x|\theta)$是样本的似然函数。当给定了样本x后，求解系统的参数θ。

回顾极大似然估计的过程，似然函数是$P(x|\theta)$，找到一个参数θ，使得似然函数最大，即$\max_\theta P(x|\theta)$。会发现最大化的对象是分子中的一项，而$P(\theta)$被省略了。可以理解为$P(\theta)$是均匀分布的。而根据贝叶斯的思想，如果认为系统的参数也是一个随机变量，并且对参数有一个先验的认知，即有$P(\theta)$，则极大似然估计变成了极大后验概率估计MAP。具体而言，在最大化参数在给定样本后的后验概率时

$$\max_\theta P(\theta|x) = \max_\theta \frac{P(x|\theta)P(\theta)}{P(x)} \propto \max_\theta P(x|\theta)P(\theta)$$

可以发现在选择θ的过程中，$P(x)$是一个常数，因此可以省略，因此贝叶斯的思想在做极大似然估计时，是加入了对参数θ的先验信息，从而将MLE变成了MAP。到这里我们也可以看到，其实MLE就是MAP的一个特殊情况，在参数的先验概率服从均匀分布的情况下MAP即为MLE。

接着可以引出共轭分布的概念。在MAP的过程中，我们有参数的先验分布$P(\theta)$，后验分布$P(\theta|x)$，也有样本的似然函数$P(x|\theta)$，如果先验分布$P(\theta)$和后验分布$P(\theta|x)$满足同样的分布律的话，那么先验分布和后验分布被称为共轭分布，同时，先验分布$P(\theta)$被称为似然函数$P(x|\theta)$的共轭先验分布。

2.2 二项分布和Beta分布

下面以二项分布为例子来说明共轭分布

投掷一枚非均匀的硬币，使用参数为θ的伯努利模型。其中θ为硬币为正面的概率，那么结果为x的概率为

$$P(x|\theta) = \theta^x (1-\theta)^{1-x}$$

这是给定参数，样本的似然函数。那么它的共轭先验分布是什么呢？是Beta分布，具体的形式如下

$$\begin{aligned} P(\theta|\alpha,\beta) &= \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta} \\ &= \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \quad \text{将分母定义成} B(\alpha,\beta) \end{aligned}$$

它具有两个参数α和β。那我们来计算一下，根据先验和似然相乘，得到的后验概率的形式

$$\begin{aligned} P(\theta|x) &\propto P(x|\theta)P(\theta) \quad \text{这里的} P(\theta) \text{就是} P(\theta|\alpha,\beta) \\ &= \theta^x(1-\theta)^{1-x}\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta} \\ &= \frac{\theta^{x+\alpha-1}(1-\theta)^{1-x+\beta-1}}{\int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta} \\ &\propto \theta^{x+\alpha-1}(1-\theta)^{1-x+\beta-1} \quad \text{因为分母是一个常数，所以省略} \end{aligned}$$

可以看到，先验概率和后验概率都为Beta分布的形式，因此二项分布与Beta分布互为共轭分布。其中，后验概率的参数是$(x+\alpha, x+\beta)$，其中α和β在没有任何实验的前提下，先验性的给出了硬币朝上的概率分配，被称为伪计数（在实际的例子中，伪计数还出现在朴素贝叶斯的拉普拉斯平滑，分子加上1，分母加上N），这是一组超参数。当没有样本或者是小样本的时候，α和β起主要作用。而样本量大的时候，它们的作用则被削弱。

举个例子，在校门口统计一段时间内出入的男女生数据分别为N_B和N_G，估计该校男女生比例。根据大数定律，频率等于概率，则

$$\begin{cases} P_B = \frac{N_B}{N_B+N_G} \\ P_G = \frac{N_G}{N_B+N_G} \end{cases}$$

如果只观察到4个女生和1个男生，样本量很小，利用上面的结论，得到女生的比例为80%，显然过拟合了。因此可以将公式修改为

$$\begin{cases} P_B = \frac{N_B+5}{N_B+N_G+10} \\ P_G = \frac{N_G+5}{N_B+N_G+10} \end{cases} \Rightarrow \begin{cases} P_B = \frac{1+5}{1+4+10} = 40\% \\ P_G = \frac{4+5}{1+4+10} = 60\% \end{cases}$$

这样就显得更加合理了。在分子和分母上分别加的5和10，为超参数。当没有样本时，$N_B = N_G = 0$，则$P_B = P_G = 0.5$，更加合理。因此当样本量比较小时，可以尽量使得参数不要过拟合。

再举一个例子，在线性回归里，目标函数（损失函数）为预测值和真实值的MSE

$$J(\vec{\theta}) = \frac{1}{2}\sum_{i=1}^{m}\left(h_{\vec{\theta}}\left(x^{(i)}\right) - y^{(i)}\right)^2$$

为了防止过拟合，往往会在损失函数后面增加正则项，进而将损失函数改造成

$$J(\vec{\theta})=\frac{1}{2}\sum_{i=1}^{m}\left(h_{\vec{\theta}}\left(x^{(i)}\right)-y^{(i)}\right)^{2}+\lambda\sum_{j=1}^{n}\theta_{j}^{2}$$

可以发现，如果我们增加了$\lambda\sum_{j=1}^{n}\theta_j^2$，本质上是认为参数$\theta$服从高斯分布，加入了先验信息，从而达到了防止过拟合的效果。

2.3 多项分布和Dirichlet分布

参考Beta分布的形式，分子是$\theta^{\alpha-1}(1-\theta)^{\beta-1}$。如果推广成$K$项分布，则分子变成$\prod_{k=1}^{K}p_k^{\alpha_k-1}$。分母是归一化因子，具体形式为$\frac{\prod_{k-1}^{K}\Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K}\alpha_k)}$，其中$\Gamma()$为gamma函数，是阶乘在实数上的推广，具体形式为

$$\Gamma(x)=\int_{0}^{\infty}t^{x-1}e^{-t}dt$$

其中$\Gamma(n)=(n-1)!$。Gamma函数具有整数阶乘的性质，即满足$\frac{\Gamma(n+1)}{\Gamma(n)}=\frac{n!}{(n-1)!}=n$

顺便一提，在上一节介绍的Beta分布的分母，也可以写成Gamma函数的形式

$$B(\alpha,\beta)=\int_{0}^{1}\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

我们将分母定义成$\frac{1}{\Delta(\overrightarrow{\alpha})}$，那么Dirichlet分布可以写成

$$p(\vec{p}|\vec{\alpha})=\operatorname{Dir}(\vec{p}|\vec{\alpha})\triangleq\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma\left(\alpha_k\right)}\prod_{k=1}^{K}p_k^{\alpha_k-1}\triangleq\frac{1}{\Delta(\vec{\alpha})}\prod_{k=1}^{K}p_k^{\alpha_k-1}$$

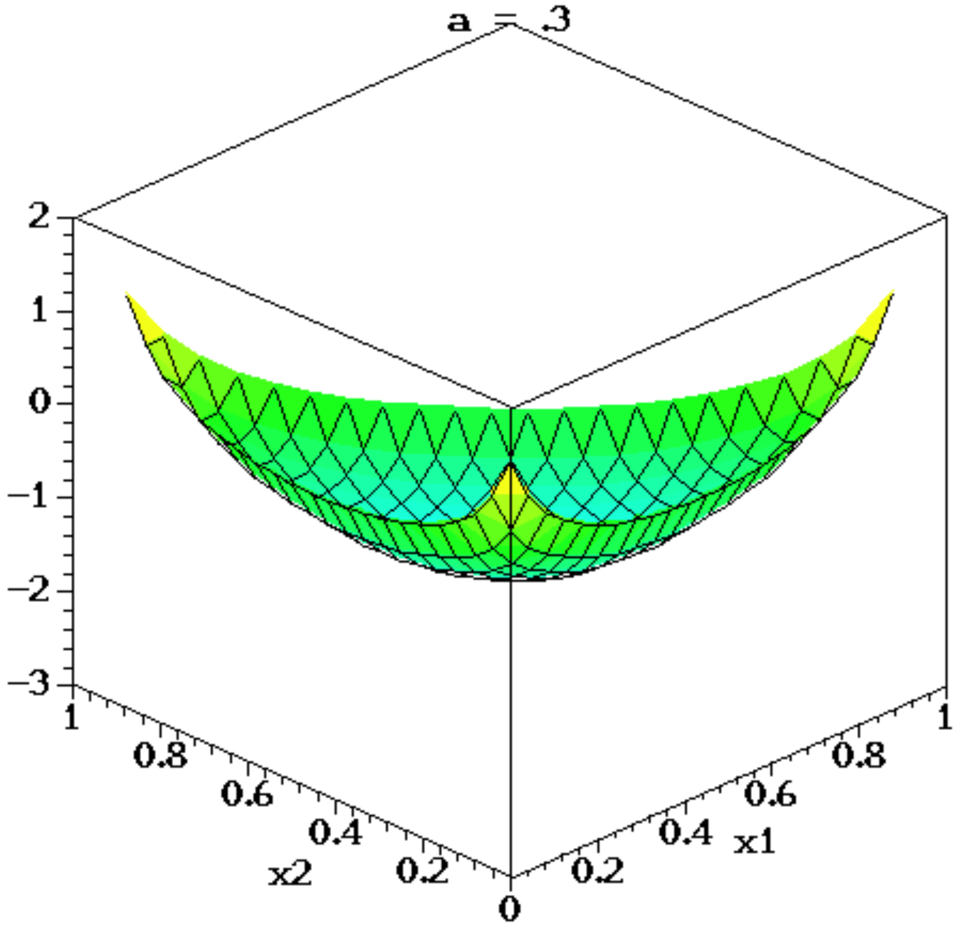
稍微对Dirichlet分布分析一下。这个分布的参数是$\overrightarrow{p}$，共有K维，并且满足$\sum_k p_k=1$，因此实际上的自由度为$K-1$，是定义在$K-1$维的单纯形(Simplex)上。在Dirichlet分布中，超参数$\overrightarrow{\alpha}$的个数是K，因此超参数的个数一定要多于参数的自由度（在Beta分布中，自由度为1，拥有两个超参数α和β）。

我们知道Dirichlet分布中，超参数$\overrightarrow{\alpha}$的个数是K，但是在实践中，我们往往会另这K个分量都等于相同的数，原因是我们不知道哪一个维度更重要，根据最大熵模型的思想，在什么都不知道的情况下，选择均匀分布。在这样的设置下，得到的分布为对称的Dirichlet分布，此时的α被称为聚集参数（concentration parameter）。此时的Dirichlet分布可以写成

$$p(\vec{p}|\alpha,K)=\operatorname{Dir}(\vec{p}|\alpha,K)\triangleq\frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K}\prod_{k=1}^{K}p_k^{\alpha-1}\triangleq\frac{1}{\Delta_K(\alpha)}\prod_{k=1}^{K}p_k^{\alpha-1}$$

$$(\vec{p}|\alpha,K)\triangleq\frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K}\prod_{k=1}^{K}p_k^{\alpha-1}\triangleq\frac{1}{\Delta_K(\alpha)}\prod_{k=1}^{K}p_k^{\alpha-1}$$

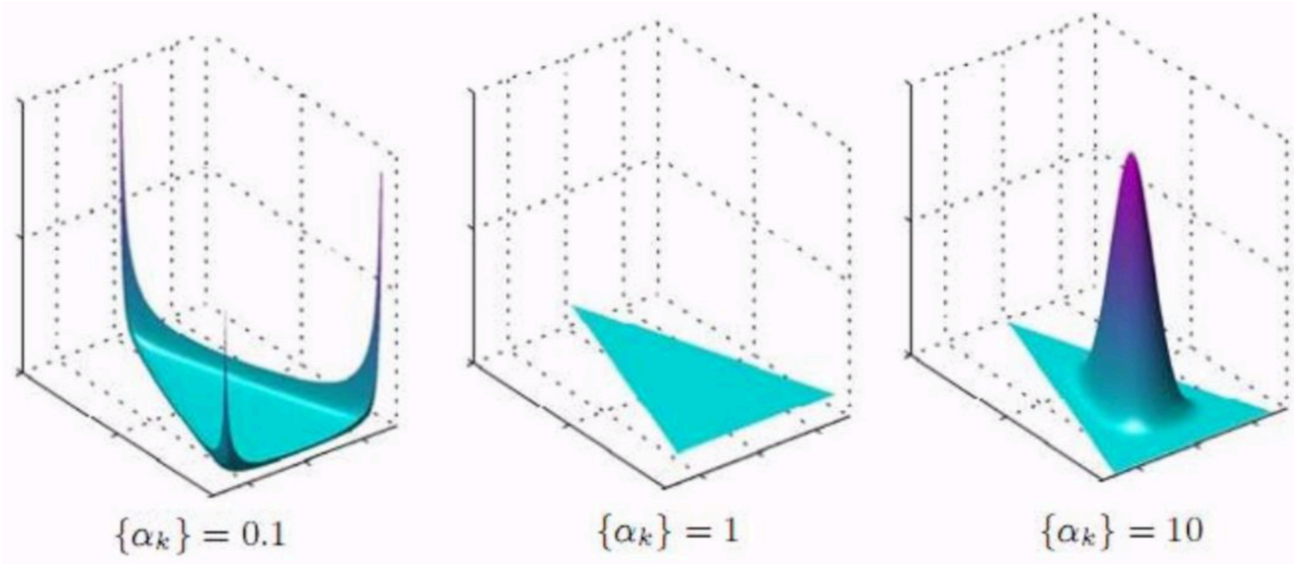
下面这幅图给出了对称的Dirichlet分布在α取不同值的情况下，概率的变化情况，其中x_1和x_2分别代表其中的两个维度，因为$x_3=1-x_1-x_2$，所以不需要展示x_3，纵轴表示$\log p$。



当$\alpha = 1$时，Dirichlet分布退化为均匀分布。

当$\alpha > 1$时，图像向上凸，表示$x_1 = x_2 = \ldots = x_K$的概率增大

当$\alpha < 1$时，图像向下凸，表示某一个维度概率$p_i = 1$，其他维度$p_{非1} = 0$的概率增大。



$\alpha = 0.1$，表示主题会集中到某一个轴上，使得文章的主题越鲜明。

$\alpha = 1$，表示每个轴取到的概率是一样的，表示每个主题都有可能，因此主题最不鲜明

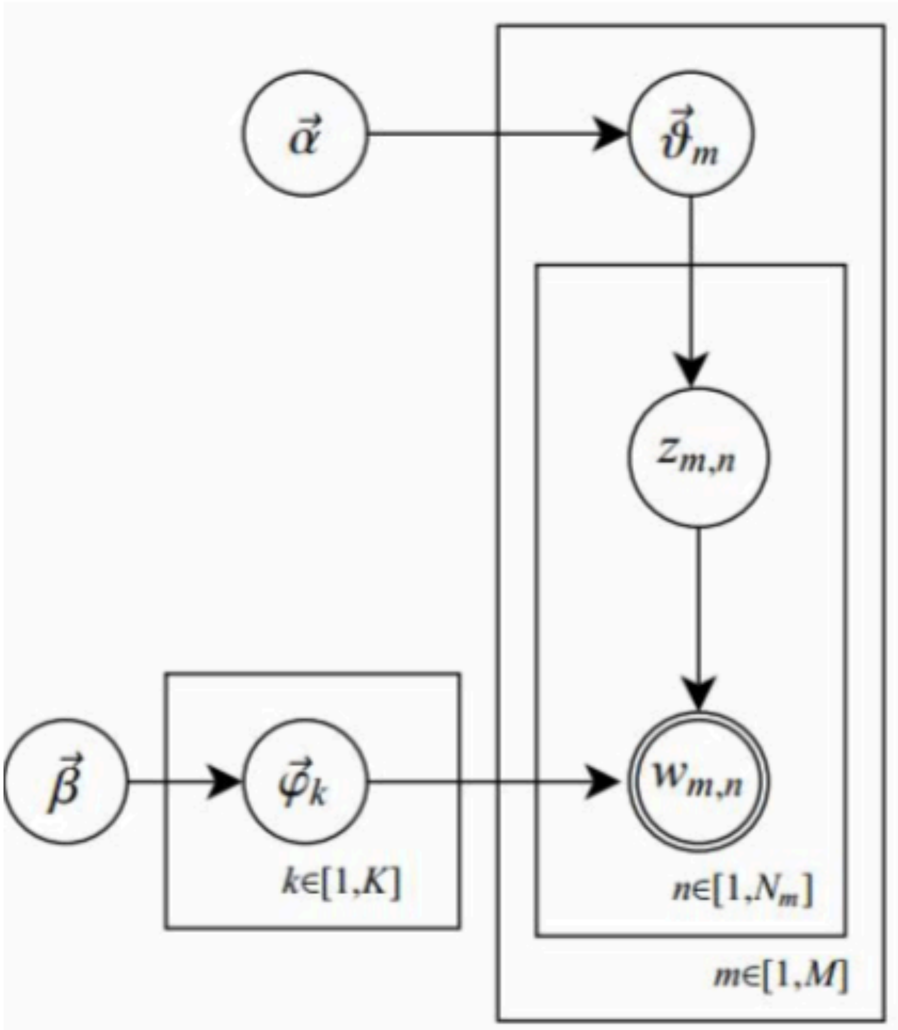
$\alpha = 10$，表示每个轴的概率相等，表示文档中的主题取到的概率相等。

2.4 LDA

给定一个语料库，包含m篇文档，其中涉及了K个主题，V个词汇(不重复)。每篇文档都有各自的主题分布，主题分布是一个多项（K点）分布。这个多项分布的参数服从Dirichlet分布，参数为α。

每个主题有各自的词分布，词分布也服从是多项（V点）分布。该多项分布的参数服从Dirichlet分布，参数为β。

对于第m篇文档的第n个词的生成过程，是首先从该文档的主题分布中，采样出一个主题来。接着选择这个主题对应的词分布。最后利用这个词分布采样出词$w_{m,n}$。不断重复这个过程，直到语料库中的M篇文档生成完毕。



上面这幅网络图，说明了主题分布与词分布生成词$w_{m,n}$的过程。首先，α和β为先验分布的参数，事先给定。ϑ_m是第m篇文档的主题分布，由α决定，$\vartheta_m = (\vartheta_{m1}, \vartheta_{m2}, \ldots, \vartheta_{mK})$，是个长度为$K$的向量，表示该篇文档属于这$K$个主题的概率。由该主题分布，采样出一个具体的主题$z_{m,n} = k, k \in [1, K]$，下标表示第$m$篇文档的第$n$个词属于什么主题。进而选择第$k$个主题的词分布$\varphi_k$。词分布$\varphi_k$由$\beta$决定，$\varphi_k = (\varphi_{k1}, \varphi_{k2}, \ldots, \varphi_{kV})$是长度为$V$的向量，表示主题$k$取到每个词的概率。

参数学习

一个词$w_{m,n} = t$的概率，可以看成每篇文档中，出现主题k的概率乘以主题k对应的词分布，该词分布出现词t的概率。然后对所有主题k积分，即可得到出现词$w_{m,n}$的似然函数。

$$p\left(w_{m,n} = t|\vec{\vartheta}_m, \underline{\Phi}\right) = \sum_{k=1}^{K} p\left(w_{m,n} = t|\vec{\varphi}_k\right) p\left(z_{m,n} = k|\vec{\vartheta}_m\right)$$

那么整个语料库出现的词的似然函数是

$$p(\mathcal{W}|\underline{\Theta}, \underline{\Phi}) = \prod_{m=1}^{M} p\left(\vec{w}_m|\vec{\vartheta}_m, \underline{\Phi}\right) = \prod_{m=1}^{M}\prod_{n=1}^{N_m} p\left(w_{m,n}|\vec{\vartheta}_m, \underline{\Phi}\right)$$

主题和词的联合分布

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

其中第一项是给定了主题z和词分布先验概率的参数β，采样词w的过程。第二项是给定了主题分布的先验概率的参数α，采样主题z的过程。下面介绍这两项的计算方法。

第一项

$$\begin{aligned}
p(\vec{w}|\vec{z},\vec{\beta}) &= \int p(\vec{w}|\vec{z},\underline{\Phi})p(\underline{\Phi}|\vec{\beta})\mathrm{d}\underline{\Phi} \quad p(\underline{\Phi}|\vec{\beta})\text{是词分布的先验概率，服从}\mathrm{Dirichlet}\text{分布} \\
&= \int p(\vec{w}|\vec{z},\underline{\Phi})\frac{1}{\Delta(\vec{\beta})}\prod_{t=1}^{T}\varphi_{z,t}^{\beta_t-1}\mathrm{d}\underline{\Phi} \quad p(\vec{w}|\vec{z},\underline{\Phi})\text{是}\mathrm{V}\text{项分布} \\
&= \int \varphi_z\frac{1}{\Delta(\vec{\beta})}\prod_{t=1}^{T}\varphi_{z,t}^{\beta_t-1}\mathrm{d}\underline{\Phi} \\
&= \int \frac{1}{\Delta(\vec{\beta})}\prod_{t=1}^{T}\varphi_{z,t}^{n_z^{(t)}+\beta_t-1}\mathrm{d}\underline{\Phi} \quad n_z^{(t)}\text{表示词}t\text{被分配给主题}z\text{的个数} \\
&= \int \prod_{z=1}^{K}\frac{1}{\Delta(\vec{\beta})}\prod_{t=1}^{V}\varphi_{z,t}^{n_z^{(t)}+\beta_t-1}\mathrm{d}\vec{\varphi}_z \quad \text{对所有的词分布做积分，因此要再加上}\prod_{z}^{K}
\end{aligned}$$

回顾Dirichlet分布的形式，

$$\mathrm{Dir}(\vec{p}|\vec{\alpha}) \triangleq \frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma\left(\alpha_k\right)}\prod_{k=1}^{K}p_k^{\alpha_k-1} \triangleq \frac{1}{\Delta(\vec{\alpha})}\prod_{k=1}^{K}p_k^{\alpha_k-1}$$

两边同时对$\vec{p}$求积分，左边等于1，右边等于$\int_{\vec{p}}\frac{1}{\Delta(\vec{\alpha})}\prod_{k=1}^{K}p_k^{\alpha_k-1}\mathrm{d}\vec{p}$，因此得到

$$\Delta(\vec{\alpha})=\int_{\vec{p}}\prod_{k=1}^{K}p_k^{\alpha_k-1}\mathrm{d}\vec{p}$$

因此

$$\int_{\vec{\varphi}_z}\prod_{t=1}^{V}\varphi_{z,t}^{n_z^{(t)}+\beta_t-1}\mathrm{d}\vec{\varphi}_z=\Delta(\vec{\beta}+\overrightarrow{n_z})$$

代入$p(\vec{w}|\vec{z},\vec{\beta})$中，得到

$$
\begin{aligned}
p(\vec{w}|\vec{z},\vec{\beta}) &= \int \prod_{z=1}^{K} \frac{1}{\Delta(\vec{\beta})} \prod_{t=1}^{V} \varphi_{z,t}^{n_z^{(t)}+\beta_t-1} \mathrm{d}\vec{\varphi}_z \\
&= \prod_{z=1}^{K} \frac{1}{\Delta(\vec{\beta})} \int \prod_{t=1}^{V} \varphi_{z,t}^{n_z^{(t)}+\beta_t-1} \mathrm{d}\vec{\varphi}_z \\
&= \prod_{z=1}^{K} \frac{\Delta(\vec{\beta}+\overrightarrow{n_z})}{\Delta(\vec{\beta})}
\end{aligned}
$$

第二项和第一项的推导过程很类似，这里就直接给出结论

$$
\begin{aligned}
p(\vec{z}|\vec{\alpha}) &= \int p(\vec{z}|\underline{\Theta})p(\underline{\Theta}|\vec{\alpha})\mathrm{d}\underline{\Theta} \\
&= \int \prod_{m=1}^{M} \frac{1}{\Delta(\vec{\alpha})} \prod_{k=1}^{K} \vartheta_{m,k}^{n_m^{(k)}+\alpha_k-1} \mathrm{d}\vec{\vartheta}_m \\
&= \prod_{m=1}^{M} \frac{\Delta\left(\vec{n}_m+\vec{\alpha}\right)}{\Delta(\vec{\alpha})}, \quad \vec{n}_m = \left\{n_m^{(k)}\right\}_{k=1}^{K}
\end{aligned}
$$