EFRI 模型推导

April 18, 2020

1 模型概述

EFRI 模型涉及大量标量、超参数、隐变量和可观测变量，所用符号及其含义如表 1所示。

各个超参数、隐变量和可观测变量之间的概率依赖关系如图 1所示。为了简化表示，图中变量没有加索引。需要注意的是，$\boldsymbol{\xi}$ 和 $\boldsymbol{\pi}$ 都服从狄利克雷分布，二者的超参数不同，分别为 $(\boldsymbol{\beta}_\xi)$ 和 $(\boldsymbol{\beta}_\pi)$，在这里都简写为 $\boldsymbol{\beta}$。$\boldsymbol{\lambda}$、$\boldsymbol{\psi}$、$\boldsymbol{c}$ 和 h 的方差变量 $\boldsymbol{\sigma}^2$ 也是同理。

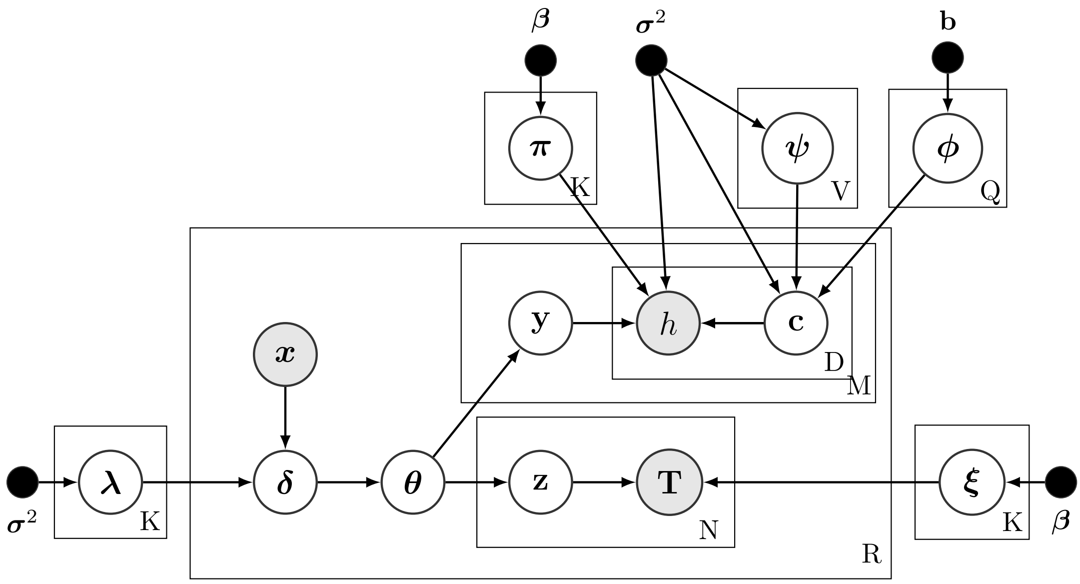

图 1: EFRI 的贝叶斯网络图

表 1: EFRI 符号表

变量种类	符号	含义
标量	R	区域的总个数。
	K	城市功能个数（聚类个数）。
	F	POI 兴趣点类型的总数。
	M	区域中评论的个数。
	N	区域中时间区间的个数。
	D	评论中包含的城市特征个数。
	V	城市特征的总个数。
	Q	城市活动的总个数。
	G	嵌入向量的维度。
	I	移动起止点统计量的取值区间个数。
超参数	$\boldsymbol{\sigma}^2$	高斯分布的方差变量。
	$\boldsymbol{\beta}$	狄利克雷分布的超参数。
	$\boldsymbol{b}$	拉普拉斯分布的方差变量。
隐变量	$\boldsymbol{\theta}_r$	区域 r 的城市功能分布。
	$\boldsymbol{\xi}_k$	主题 k 下，移动起止点向量的数值在 I 个区间的分布。
	$\boldsymbol{\pi}_k$	第 k 个城市功能下各个城市活动的分布。
	$\boldsymbol{\psi}_v$	城市特征 v 的嵌入向量。
	$\boldsymbol{\phi}_q$	城市活动 q 的嵌入向量。
	$\boldsymbol{\lambda}_k$	城市功能 k 中各个不同兴趣点 POI 的权重。
	$\boldsymbol{\delta}_r$	原 POI 特征向量 $\mathbf{x}_r$ 经过加权后的 POI 特征向量。
	$\boldsymbol{c}_{r,m,d}$	区域 r 内评论 m 中的第 d 个城市特征与各城市活动的相关性。
	$\boldsymbol{z}_{r,n}$	区域 r 内第 n 个时间区间所对应的功能分布。
	$\boldsymbol{y}_{r,m}$	区域 r 内评论 m 所对应的功能分布。
可观测变量	$\mathbf{T}_r$	区域 r 的移动起止点矩阵。
	$\mathbf{x}_r$	区域 r 的 POI 特征向量。
	$h(r,m,d)$	区域 r 内评论 m 中第 d 个词位置对应的函数值。

2 变分推断

在公式 1中，L_p 表示 EFRI 模型的对数似然函数，L_1 表示评论文本数据的似然函数，L_2 表示移动起止点数据的似然函数。求解模型的目标是最大化如公式 1所示的对数似然函数。

$$\begin{aligned} L_p =& L_1 + L_2, \\ L_1 =& \log[p(h|\boldsymbol{\lambda}, \boldsymbol{\delta}, \boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\psi}, \mathbf{y}, \mathbf{c}) \cdot p(\boldsymbol{\lambda}, \boldsymbol{\delta}, \boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\psi}, \mathbf{y}, \mathbf{c}|\boldsymbol{\sigma}^2, \boldsymbol{\beta}, \mathbf{b})], \\ L_2 =& \log[p(\mathbf{T}|\boldsymbol{\lambda}, \boldsymbol{\delta}, \boldsymbol{\theta}, \boldsymbol{\xi}, \mathbf{z}) p(\boldsymbol{\lambda}, \boldsymbol{\delta}, \boldsymbol{\theta}, \boldsymbol{\xi}, \mathbf{z}|\boldsymbol{\sigma}^2, \boldsymbol{\beta}, \mathbf{b})]. \end{aligned} \tag{1}$$

如图 1所示，变量间的依赖关系如下所示：

- $p(\boldsymbol{\lambda}_k) = \mathcal{N}(0, \sigma_\lambda^2 I)$，$\boldsymbol{\lambda}_k$ 表示功能 k 中各个不同兴趣点的权重，由高斯分布生成。需要注意的是，这里的 σ_λ 的下标 λ 不是索引，是为了跟其它隐变量 $\boldsymbol{\psi}$、c 等的参数有区分。
- $\delta_{r,k} = \frac{\exp\{\delta_{r,k}\}}{\sum_{k'} \exp\{\delta_{r,k'}\}} \frac{\exp\{\boldsymbol{x}_r^T \boldsymbol{\lambda}_k\}}{\sum_{k'} \exp\{\boldsymbol{x}_r^T \boldsymbol{\lambda}_{k'}\}}$，$\boldsymbol{\delta}_r$ 表示原 POI 特征向量 $\boldsymbol{x}_r$ 经过 $\boldsymbol{\lambda}_k$ 加权并规范化后的 POI 特征向量。
- $p(\boldsymbol{\theta}_r) = Dir(\boldsymbol{\delta}_r)$，$\boldsymbol{\theta}_r$ 表示区域 r 的主题分布，服从参数为 $\boldsymbol{\delta}_r$ 的狄利克雷分布。
- $p(\boldsymbol{z}_{r,n}) = Mult(\boldsymbol{\theta}_r)$，$\boldsymbol{z}_{r,n}$ 表示区域 r 内第 n 个时间区间所对应的主题分布，服从参数为 $\boldsymbol{\theta}_r$ 的多项式分布。
- $p(\boldsymbol{\xi}_k) = Dir(\boldsymbol{\beta}_\xi)$，$\boldsymbol{\xi}_k$ 表示主题 k 下，移动轨迹向量的数值在 I 个区间的分布。服从参数为 $\boldsymbol{\beta}_\xi$ 的狄利克雷分布。需要注意的是，这里的 $\boldsymbol{\beta}_\xi$ 是一个向量，下标 ξ 不是索引，是为了跟另一个隐变量 $\boldsymbol{\pi}$ 的参数有区分。
- $p(\boldsymbol{t}_{r,n}) = Mult(\boldsymbol{\xi}_{z_{r,n}})$，$\boldsymbol{t}_{r,n}$ 表示输入的移动起止点特征矩阵 $\boldsymbol{T}_r$ 的第 n 个向量。
- $p(\boldsymbol{y}_{r,m}) = Mult(\boldsymbol{\theta}_r)$，$\boldsymbol{y}_{r,m}$ 表示区域 r 内评论 m 所对应的主题分布，服从参数为 $\boldsymbol{\theta}_r$ 的多项式分布。
- $p(\boldsymbol{\pi}_k) = Dir(\boldsymbol{\beta})$，$\boldsymbol{\pi}_k$ 表示第 k 个城市主题下各个城市功能的分布。

- $p(\boldsymbol{\psi}_v) = \mathcal{N}(0, \sigma_\psi^2 I)$，$\boldsymbol{\psi}_v$ 表示第 v 个城市特征的词嵌入向量，服从方差为 σ_ψ^2 的高斯分布。
- $p(\boldsymbol{\phi}_q) = \mathcal{L}aplace(0, \boldsymbol{b})$，$\boldsymbol{\phi}_q$ 表示第 q 个城市功能的词嵌入向量，服从位置参数为 0、尺度参数为 $\boldsymbol{b}$ 的拉普拉斯分布。
- $p(\boldsymbol{c}_{r,m,d}) = \mathcal{N}(\boldsymbol{\Phi}\boldsymbol{\psi}_{v(d)}^T, \sigma_c^2)$，$\boldsymbol{c}_{r,m,d}$ 表示在区域 r 内评论 m 中城市特征 d 与各个城市功能的相关性。
- $p(h) = \mathcal{N}(\boldsymbol{\pi}_{\boldsymbol{y}_{r,m}}^T \boldsymbol{c}_{r,m,d}, \sigma_w^2)$，其中 h 表示区域 r 内评论 m 中第 d 个位置对应的可观测函数值 $h(r, m, d)$。

公式 1涉及到隐变量的耦合，难以直接求解。因此使用变分推断（Variational Inference）EM 算法 [1, 2] 来求解参数。变分推断基于平均场理论（Mean Field Assumption），假设所有的隐变量都是通过各自的独立分布形成的。在 EFRI 中，假设每个隐变量都各自服从对应的变分分布，如图 2所示。与 1一样，图 2省略了变量索引。若一个隐变量只有一个变分分布的参数，如 y，则 $\mathbf{y}$ 对应的变分参数用 $\mathbf{y}'$ 的形式来表示。若一个隐变量服从的变分分布中有两个不同参数，则用增加上标的形式来表示变分参数，如 $\boldsymbol{\lambda}$ 的期望参数和方差参数分别表示为 $\boldsymbol{\lambda}^{(\mu)}$ 和 $\boldsymbol{\lambda}^{(\sigma^2)}$。

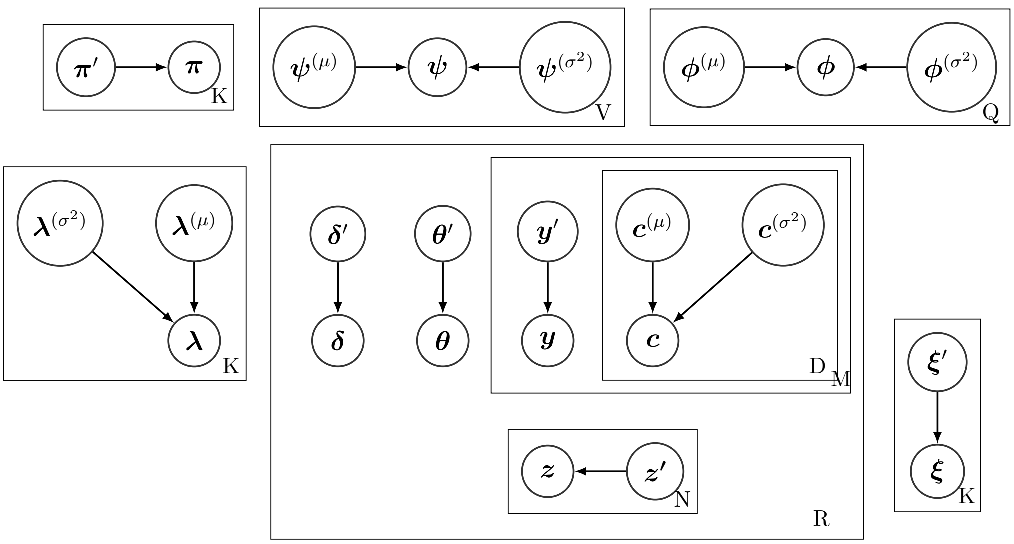

图 2: EFRI 的独立变分分布图

由图 2可以得到隐变量的变分分布

$$
\begin{aligned}
&q(\boldsymbol{\lambda},\boldsymbol{\delta},\boldsymbol{\theta},\boldsymbol{z},\boldsymbol{\xi},\boldsymbol{y},\boldsymbol{\pi},\boldsymbol{\psi},\boldsymbol{\phi},\boldsymbol{c}) \qquad (2)\\
=&\prod_{k=1}^{K} q(\boldsymbol{\lambda}_k)q(\boldsymbol{\pi}_k)q(\boldsymbol{\xi}_k)\prod_{v=1}^{V} q(\boldsymbol{\psi}_v)\prod_{q=1}^{Q} q(\boldsymbol{\phi}_q)\prod_{r=1}^{R}\left[q(\boldsymbol{\delta}_r)q(\boldsymbol{\theta}_r)\prod_{n=1}^{N} q(\boldsymbol{z}_{r,n})\prod_{m=1}^{M}\left(q(\boldsymbol{y}_{r,m})\prod_{d=1}^{D} q(\boldsymbol{c}_{r,m,d})\right)\right].
\end{aligned}
$$

变分推断的目标是用 $q(\boldsymbol{\lambda},\boldsymbol{\theta},\boldsymbol{z},\boldsymbol{\xi},\boldsymbol{y},\boldsymbol{\pi},\boldsymbol{\psi},\boldsymbol{\phi},\boldsymbol{c})$ 来近似地估计对数似然函数 L_p，希望这两个分布尽可能的相似，也就是希望这两个概率分布之间的 KL 距离 $KL(q||p)$ 尽可能小。模型的求解目标变为计算合适的隐变量的值, 使对应的 $q(\boldsymbol{\lambda},\boldsymbol{\theta},\boldsymbol{z},\boldsymbol{\xi},\boldsymbol{y},\boldsymbol{\pi},\boldsymbol{\psi},\boldsymbol{\phi},\boldsymbol{c})$ 能尽可能近似对数似然函数 L_p，进而使用 EM 算法迭代。

根据 Bishop [3] 对变分推断的介绍，模型的对数似然函数可以表示为：

$$
L_p = \mathrm{ELBO} + KL(q||p), \tag{3}
$$

其中 ELBO(**E**vidence **L**ower **BO**und) 是模型的最大化证据下界，表示为：

$$
\mathrm{ELBO} = \mathbb{E}_q[L_p] - \mathbb{E}_q[q(\boldsymbol{\lambda},\boldsymbol{\theta},\boldsymbol{z},\boldsymbol{\xi},\boldsymbol{y},\boldsymbol{\pi},\boldsymbol{\psi},\boldsymbol{\phi},\boldsymbol{c})]. \tag{4}
$$

由于对数似然部分 L_p 与 KL 散度无关，可以看做常量，因此模型的求解目标从最小化 KL 散度等价转换为求 ELBO 最大值。

3 算法步骤

Algorithm Steps 根据变分推断求解 EFRI 的步骤如下：

1. 初始化所有的变分参数和超参数。

2. 进行 EFRI 的 E 步骤循环迭代。

   (a) 对于 $\forall k \in \{1,2,\ldots,K\}, \forall i \in \{1,2,\ldots,I\}$，更新 $\xi'_{k,i}$：

$$
\xi'_{k,i} = (\beta_\xi)_i + \sum_{r=1}^{R}\sum_{n=1}^{N} z'_{r,n,k} t_{r,n,i}.
$$

(b) 对于 $\forall k \in \{1,2,\ldots,K\}, \forall q \in \{1,2,\ldots,Q\}$，更新 $\pi'_{k,q}$：

$$\pi'_{k,q} = (\beta_\pi)_q + \sum_{r=1}^{R}\sum_{m=1}^{M} y'_{r,m,k} \sum_{d=1}^{D}\sum_{v=1}^{V} h(r,m,d)^v.$$

其中 $h(r,m,d)^v = h(r,m,d)$ 当且仅当区域 r 内评论 m 中第 d 个城市特征是 v 时成立，否则 $h(r,m,d)^v = 0$。

(c) 对于 $\forall q \in \{1,2,\ldots,Q\}, \forall g \in \{1,2,\ldots,G\}$，更新 $\phi^{(\mu)}_{q,g}$：

$$\phi^{(\mu)}_{q,g} = \frac{\sum_{r=1}^{R}\sum_{m=1}^{M}\sum_{d=1}^{D}\sum_{k=1}^{K}\sum_{v=1}^{V} y'_{r,m,k}\frac{\pi'_{k,q}}{\sum_{q'}\pi'_{k,q'}}\psi^{(\mu)}_{v,g} h(r,m,d)^v - \frac{\sigma_w^2}{b_g}}{\sum_{r=1}^{R}\sum_{m=1}^{M}\sum_{d=1}^{D}\sum_{k=1}^{K} y'_{r,m,k}(\frac{\pi'_{k,q}}{\sum_{q'}\pi'_{k,q'}}\cdot\frac{\pi'_{k,q}+1}{\sum_{q'}\pi'_{k,q'}+1})[(\psi^{(\mu)}_{v,g})^2 + \psi^{(\sigma^2)}_{v,g}]}.$$

(d) 对于 $\forall q \in \{1,2,\ldots,Q\}, \forall g \in \{1,2,\ldots,Q\}$，更新 $\phi^{(\sigma^2)}_{q,g}$：

$$\phi^{(\sigma^2)}_{q,g} = \frac{\sigma_w^2}{\sum_{r=1}^{R}\sum_{m=1}^{M}\sum_{d=1}^{D}\sum_{k=1}^{K} y'_{r,m,k}(\frac{\pi'_{k,q}}{\sum_{q'}\pi'_{k,q'}}\cdot\frac{\pi'_{k,q}+1}{\sum_{q'}\pi'_{k,q'}+1})[(\psi^{(\mu)}_{v,g})^2 + \psi^{(\sigma^2)}_{v,g}]}.$$

(e) 对于 $\forall v \in \{1,2,\ldots,V\}, \forall g \in \{1,2,\ldots,G\}$，更新 $\psi^{(\mu)}_{v,g}$：

$$\psi^{(\mu)}_{v,g} = \frac{\sum_{r=1}^{R}\sum_{m=1}^{M}\sum_{d=1}^{D}\sum_{k=1}^{K}\sum_{q=1}^{Q} y'_{r,m,k}\frac{\pi'_{k,q}}{\sum_{q'}\pi'_{k,q'}}\phi^{(\mu)}_{q,g} h(r,m,d)^v}{\sum_{r=1}^{R}\sum_{m=1}^{M}\sum_{d=1}^{D}\sum_{k=1}^{K}\sum_{q=1}^{Q} y'_{r,m,k}(\frac{\pi'_{k,q}}{\sum_{q'}\pi'_{k,q'}}\cdot\frac{\pi'_{k,q}+1}{\sum_{q'}\pi'_{k,q'}+1})[(\phi^{(\mu)}_{q,g})^2 + \phi^{(\sigma^2)}_{q,g}] + \frac{\sigma_w^2}{(\sigma_\psi)_g^2}}.$$

(f) 对于 $\forall v \in \{1,2,\ldots,V\}, \forall g \in \{1,2,\ldots,G\}$，更新 $\psi^{(\sigma^2)}_{v,g}$：

$$\psi^{(\sigma^2)}_{v,g} = \frac{\sigma_w^2}{\sum_{r=1}^{R}\sum_{m=1}^{M}\sum_{d=1}^{D}\sum_{k=1}^{K}\sum_{q=1}^{Q} y'_{r,m,k}(\frac{\pi'_{k,q}}{\sum_{q'}\pi'_{k,q'}}\cdot\frac{\pi'_{k,q}+1}{\sum_{q'}\pi'_{k,q'}+1})[(\phi^{(\mu)}_{q,g})^2 + \phi^{(\sigma^2)}_{q,g}] + \frac{\sigma_w^2}{(\sigma_\psi)_g^2}}.$$

(g) 对于 $\forall k \in \{1,2,\ldots,K\}, \forall f \in \{1,2,\ldots,F\}$，更新 $\lambda^{(\mu)}_{k,f}$：

$$\lambda^{(\mu)}_{k,f} = (\sigma_\lambda^2)_f \sum_{r=1}^{R} \delta'_{r,k} x_{r,f}.$$

(h) 对于 $\forall k \in \{1,2,\ldots,K\}, \forall f \in \{1,2,\ldots,F\}$，更新 $\lambda^{(\sigma^2)}_{k,f}$：

$$\lambda^{(\sigma^2)}_{k,f} = (\sigma_\lambda^2)_f.$$

(i) 对于 $\forall r \in \{1, 2, \ldots, R\}, \forall k \in \{1, 2, \ldots, K\}$，更新 $\delta'_{r,k}$：

$$\delta'_{r,k} = \frac{1}{\Psi(\sum_{j=1}^{K} \theta'_{r,j}) - \Psi(\theta'_{r,k}) - \sum_{f=1}^{F} x_{r,f} \lambda_{v,g}^{(\mu)}}.$$

其中 $\Psi(\cdot)$ 函数表示 $\Gamma(\cdot)$ 函数的二阶导数。

(j) 对于 $\forall r \in \{1, 2, \ldots, R\}, \forall k \in \{1, 2, \ldots, K\}$，更新 $\theta'_{r,k}$：

$$\theta'_{r,k} = \delta'{}_{r,k} + \sum_{n=1}^{N} z'{}_{r,n,k} + \sum_{m=1}^{M} y'{}_{r,m,k}.$$

(k) 对于 $\forall r \in \{1, 2, \ldots, R\}, \forall m \in \{1, 2, \ldots, M\}, \forall k \in \{1, 2, \ldots, K\}$，更新 $y'_{r,m,k}$：

$$y'_{r,m,k} \propto \exp\{-\sum_{d=1}^{D} \sum_{q=1}^{Q} \sum_{g=1}^{G} \sum_{v=1}^{V} \frac{1}{2\sigma_w^2} \rho + \Psi(\theta'_{r,k} - \Psi(\sum_{j=1}^{K} \theta'_{r,j}))\}.$$

其中 $\boldsymbol{y}'_{r,m}$ 是一个加和为 1 的向量，所以其中的每个元素 $y'_{r,m,k}$ 不能直接求得，需要进行等比例缩放。ρ 表示为

$$\begin{aligned}\rho = & [h(r,m,d)^v]^2 - 2 \times \frac{\pi'_{k,q}}{\sum_{q'} \pi'_{k,q'}} \phi_{q,g}^{(\mu)} \psi_{v,g}^{(\mu)} h(r,m,d)^v \\ & + (\frac{\pi'_{k,q}}{\sum_{q'} \pi'_{k,q'}} \cdot \frac{\pi'_{k,q} + 1}{\sum_{q'} \pi'_{k,q'} + 1})[(\phi_{q,g}^{(\mu)})^2 + \phi_{q,g}^{(\sigma^2)}][(\psi_{v,g}^{(\mu)})^2 + \psi_{v,g}^{(\sigma^2)}].\end{aligned} \tag{5}$$

(l) 对于 $\forall r \in \{1, 2, \ldots, R\}, \forall n \in \{1, 2, \ldots, N\}, \forall k \in \{1, 2, \ldots, K\}$，更新 $z'_{r,n,k}$：

$$z'_{r,n,k} \propto \exp\{\sum_{i=1}^{I} \left\{ t_{r,n,i} \left[\Psi(\xi'_{k,i}) - \Psi\left(\sum_{j=1}^{K} \xi'_{k,j} \right) \right] \right\} + \Psi(\theta'_{r,k}) - \Psi(\sum_{j=1}^{K} \theta'_{r,j})\}.$$

与 $y'_{r,m,k}$ 相似，$z'_{r,n,k}$ 也需要等比例缩放，使 $\boldsymbol{z}'_{r,n}$ 成为一个加和为 1 的向量。

(m) 如果所有变分参数均已收敛，则结束本次循环，否则回到步骤 2a

3. 进行 EFRI 的 M 步骤循环迭代。

(a) 对于 $\forall g \in \{1, 2, \ldots, G\}$，更新 b_g：

$$b_g = \frac{\sum_{q=1}^{Q} \phi_{q,g}^{(\mu)}}{Q}.$$

(b) 对于 $\forall f \in \{1, 2, \ldots, F\}$，更新 $(\sigma_\lambda^2)_f$：

$$(\sigma_\lambda^2)_f = \frac{\sum_{k=1}^{K}((\lambda_{k,f}^{(\mu)})^2 + \lambda_{k,f}^{(\sigma^2)})}{K}.$$

(c) 对于 $\forall v \in \{1, 2, \ldots, V\}$，更新 $(\sigma_\psi^2)_v$：

$$(\sigma_\psi^2)_v = \frac{\sum_{g=1}^{G}((\psi_{v,g}^{(\mu)})^2 + \psi_{v,g}^{(\sigma^2)})}{G}.$$

(d) 更新 σ_w^2：

$$\sigma_w^2 = \frac{\sum_{r=1}^{R}\sum_{m=1}^{M}\sum_{d=1}^{D}\sum_{k=1}^{K}\sum_{q=1}^{Q}\sum_{g=1}^{G}\sum_{v=1}^{V} y'_{r,m,k}\rho}{\sum_{r=1}^{R}\sum_{m=1}^{M}\sum_{d=1}^{D}\sum_{k=1}^{K} y'_{r,m,k}}.$$

其中 ρ 的定义如公式 5所示。

(e) 对于 $\forall q \in \{1, 2, \ldots, Q\}$，更新 $(\hat{\beta_\pi})_q$：

$$(\hat{\beta_\pi})_q = (\beta_\pi)_q + \sum_{s=1}^{Q} \frac{K\Psi(\sum_{q'=1}^{Q}(\beta_\pi)_{q'}) - K\Psi((\beta_\pi)_q) + \sum_{k=1}^{K}\left(\Psi(\pi'_{k,q}) - \Psi(\sum_{j=1}^{Q}\xi'_{k,j})\right)}{K\Psi'(\sum_{q'=1}^{Q}(\beta_\pi)_{q'}) - \delta(q,s)K\Psi'((\beta_\pi)_q)}.$$

其中 $\Psi(\cdot)$ 函数表示 $\Gamma(\cdot)$ 函数的二阶导数，$\Psi'(\cdot)$ 函数表示 $\Gamma(\cdot)$ 函数的三阶导数。当且仅当 $q = s$ 时 $\delta(q, s) = 1$ 成立，否则 $\delta(q, s) = 0$。

(f) 对于 $\forall i \in \{1, 2, \ldots, I\}$，更新 $(\hat{\beta_\xi})_i$：

$$(\hat{\beta_\xi})_i = (\beta_\xi)_i + \sum_{s=1}^{I} \frac{K\Psi(\sum_{i'=1}^{Q}(\beta_\xi)_{i'}) - K\Psi((\beta_\xi)_i) + \sum_{k=1}^{K}\left(\Psi(\xi'_{k,i}) - \Psi(\sum_{j=1}^{I}\xi'_{k,j})\right)}{K\Psi'(\sum_{i'=1}^{Q}(\beta_\xi)_{i'}) - \delta(q,s)K\Psi'((\beta_\xi)_i)}.$$

(g) 如果所有超参数均已收敛，则结束本次循环，否则回到步骤 3a

4. 如果所有参数均收敛，则算法结束。否则回到步骤 2。

References

[1] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. The Journal of Machine Learning Research, 3:993 C1022, 2003.

[2] Blei D M , Kucukelbir A , Mcauliffe J D . Variational Inference: A Review for Statisticians[J]. Journal of the American Statistical Association, 2017:0-0.

[3] Christopher M Bishop. Pattern Recognition and Machine Learning (Information Science and Statistics)[M]. Springer-Verlag New York, Inc. 2006.