统计建模与推断

——以混合主题模型与因子分解模型为例

预备知识
PLSA GMM -EM
LDA-VBEM CTM-VBEM LDA-MCMC1 LDA-MCMC2
PMF RMF(SGD&ALS) Soc-MF BPMF FM
组合模型 Co-TopicRegression

Z-Y LI

预备知识

为使本文内容自我包含，先讨论模型相关的一些指数族分布的性质，这将利于后续内容的展开。

一个指数族的概率分布可以写成如下的形式：

$$p(X|Y) = \exp\{\phi(Y)^T u(X) + f(X) + g(Y)\}$$

其中 $\phi(Y)$称作自然参数，$u(X)$称作充分统计量，$g(Y)$是归一化因子，或者叫作 log-partition 以确保任何的 Y 设置对 X 的积分为 1。

方便起见，重参数化上面的公式为：

$$p(X|\phi') = \exp\{\phi'^T u(X) + f(X) + \tilde{g}(\phi')\}$$

其中 $g(\phi^{-1}(t)) = \tilde{g}(t)$。两边对$\int_X p(X|\phi')=1$ 中的参数 ϕ'求微分有

$$\int_X \frac{d}{d\phi'} p(X|\phi') = 0$$

$$\int_X p(X|\phi')\,[u(X) + \frac{d}{d\phi'}\tilde{g}(\phi')] = 0$$

X 所服从的分布对应的自然统计量 $u(X)$在参数 ϕ' 下的期望是：

$$E_{p(X|\phi')} u(X) = -\frac{d}{d\phi'}\tilde{g}(\phi') = -\frac{d}{d\phi'} g\left(\phi^{-1}(\phi')\right)_{|\phi^{-1}(\phi')=Y}$$

$$= -\frac{d}{ds} g(s) \cdot \left(\frac{1}{\frac{d}{dt}\phi(t)}\right)_{|s=\phi^{-1}(\phi')=Y,t=\phi'} = -\frac{d}{ds} g(s) / \frac{d}{dt}\phi(t)_{|s=Y,t=\phi(Y)}$$

这个结论含义是，**一个指数族分布对自身充分统计量求期望就是 log-partition 对自然参数的导数。**这个结论涉及到微积分的基本知识，由于 p(X|Y)在很多情况下没有必要写成重参数化的具体形式，因而我们将利用上面的公式完成下面几个典型指数族分布的推导。

Dirichlet 分布：

$$p(\theta|\alpha) = \exp\{\phi(\alpha)^T u(\theta) + f(\theta) + g(\alpha)\} = \frac{\Gamma\left(\sum_{i=1}^{k}\alpha_i\right)}{\prod_{i=1}^{k}\Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1},$$

随机变量 θ 在（k-1）维的单纯形内取值，即$\theta_i>0, \sum_{i=1}^{k}\theta_i=1$。

$$\phi(\alpha)=\begin{pmatrix}\alpha_1 - 1\\ \dots\dots\\ \alpha_K - 1\end{pmatrix} u(\theta)=\begin{pmatrix}log\theta_1\\ \dots\dots\\ log\theta_k\end{pmatrix} g(\alpha)=\log\frac{\Gamma(\sum_{i=1}^{K}\alpha_i)}{\prod_{i=1}^{K}\Gamma(\alpha_i)} \quad f(\theta)=0,$$

$$E_{p(\theta|\alpha)} \log\theta_i = -\frac{d}{ds_i} g(s) \;/\; \frac{d}{dt_i}\phi(t)_{|s=\alpha_i,t=\alpha_i-1} = \Psi(\alpha_i) - \Psi\left(\sum_{i=1}^{K}\alpha_i\right)$$

其中 $\Psi(t) = \Gamma'(t)/\Gamma(t)$ 。

Multinomial 分布：

$p(Z_n|\theta) = \exp\{\phi(\theta)^T u(Z_n) + f(Z_n) + g(\theta)\}$，

$$\Phi(\theta) = \begin{pmatrix} log\theta_1 \\ \dots\dots \\ log\theta_k \end{pmatrix} u(Z_n) = \begin{pmatrix} Z_n^1 \\ \dots\dots \\ Z_n^K \end{pmatrix} g(\theta) = \log\left(\frac{1}{\sum_{i=1}^{k}\theta_i}\right) f(Z_n) = 0 E_{p(Z_n|\theta)} Z_n^i = \theta_i$$

Multivariate Gaussian 分布：

$$p(\eta|\mu,\Lambda^{-1}) = \sqrt{\frac{|\Lambda|}{(2\pi)^d}}\exp(-\frac{1}{2}(\eta-\mu)^T\Lambda(\eta-\mu)),$$

其中 η 是 d 维随机变量

$$\Phi(\mu,\Lambda^{-1}) = \begin{pmatrix} \Lambda\mu \\ -\frac{1}{2} vec(\Lambda) \end{pmatrix} u(\eta) = \begin{pmatrix} \eta \\ vec(\eta\eta^T) \end{pmatrix} f(\eta) = 0$$

$$g(\mu,\Lambda^{-1}) = \frac{1}{2}(\ln|\Lambda| - \mu^T\Lambda\mu - d\ln(2\pi))$$

$$E_{p(\eta|\mu,\Sigma)}\begin{pmatrix} \eta \\ vec(\eta\eta^T) \end{pmatrix} = \frac{d}{d\Phi} g(\mu,\Lambda^{-1})$$

$$= \begin{pmatrix} \mu \\ vec(\Lambda^{-1} + \mu\mu^T) \end{pmatrix}, \text{vec}(\cdot)$$将矩阵按列堆砌拉申成向量。

如果先验 θ~Dir(α), Z_n~Mult(θ)，θ 的后验形式为 $\hat{p}(\theta) \propto p(\theta|\alpha)\prod_{n=1} p(Z_n|\theta)$，由于 Dir 分布中充分统计量与 Mult 分布中自然参数有相同的形式，且在单纯形约束下 Mult 中的 g(θ)=0，因此$\widehat{p(\theta)}$服从 Dir 分布，我们称 Dir 分布是 Mult 的共轭分布。

共轭分布：

Dirichlet to Multinomial

Beta to Bernoulli

Gassian-Wishart(inverse gamma) to Multivariate Gaussian

值得说明的是，这里的 Multinomial 分布实际上指的是 Categorical 分布，2 维是 Bernoulli，也就是说在多个类别中选择一个。我们习惯上把 Bernoulli 的随机变量记作 0 与 1，表示 2 个中选择其中一个的概率，是 2 维随机变量。比如，n 个单词在文本中出现的概率。而把单词是否出现当作两个不同的事件对待的 Bernoulli 则与此不符的，这是两种不同的模型。

图模型内容参考文献

Winn J, Bishop C M, Jaakkola T. Variational Message Passing[J]. Journal of Machine Learning Research, 2005, 6(4).

Jordan M I, Ghahramani Z, Jaakkola T S, et al. An introduction to variational methods for graphical models[J]. Machine learning, 1999, 37(2): 183-233.

Andrieu C, De Freitas N, Doucet A, et al. An introduction to MCMC for machine learning[J]. Machine learning, 2003, 50(1-2): 5-43.

Bishop C M, Nasrabadi N M. Pattern recognition and machine learning[M]. New York: springer, 2006.

Wainwright M J, Jordan M I. Graphical models, exponential families, and variational inference[J]. Foundations and Trends® in Machine Learning, 2008, 1(1-2): 1-305.

Hastie T, Tibshirani R, Friedman J, et al. The elements of statistical learning: data mining, inference and prediction[J]. The Mathematical Intelligencer, 2005, 27(2): 83-85.

Neal, Radford M. "Probabilistic inference using Markov chain Monte Carlo methods." (1993).

Theory and Use of the EM Algorithm . Foundataion and Trends

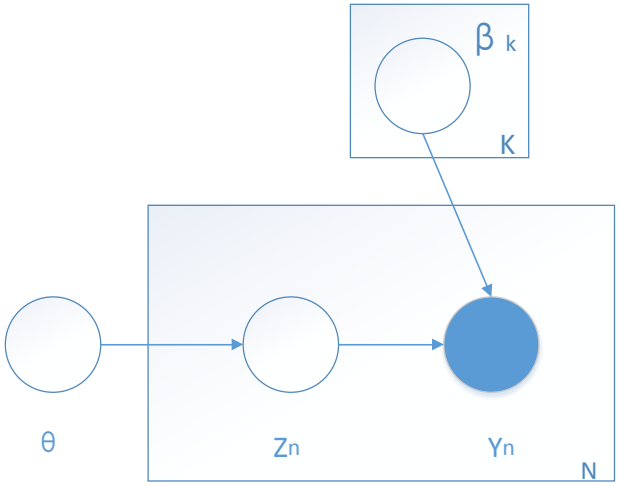
.....

概率隐语义分析 pLSA 及 EM 算法

概率隐语义分析(probabilistic Latent Semantic Analysis)模型中主题的选择由 θ 确定，话题变量 z_i 服从 θ 确定的 Multinomial 分布，记作 Z_i~Mult(θ)。文本中的单词{Y_1 ... ... Y_n}都各自对应一个主题{Z_1 ... ... Z_n}，这样文本中的每个 Y_i 服从 $\beta_{zi\cdot}$确定的 Multinomial 分布，记作 Y_i~Mult($\beta_{zi\cdot}$)。所以有以下的生成假设，对每个 Y_i ∈{Y_1 ... ... Y_n}：

(a) 从 θ 确定的 Multinomial 分布产生话题 Z_i，即 Z_i~Mult(θ)。

(b) 从 $\beta_{zi\cdot}$确定的 Multinomial 分布产生单词 Y_i，即 Y_i~Mult($\beta_{zi\cdot}$)。



在 pLSA 中，参数 θ 在（K-1）维的单纯形内取值($\theta_i>0$, $\sum_{i=1}^{K}\theta_i=1$)，而Z_n服从 θ 确定的多项式分布：$p(Z_n=k|\theta)=\theta_k$。即 Z_n 以概率 θ_k 取值为 k（k 从 1 到 K）并选择外层的第 k 个参数族来确定 Y_n 的多项式分布 $p(Y_n|Z_n=k,\boldsymbol{\beta})=\text{Mult}(\beta_k)$。$Y_n$ 代表单词对应单词表中的序号，取值从 1 到 V，因而 β_k 作为多项式分布的参数，是一个 V 维的向量。而这样得到 Y 和 Z 的联合分布如下：

$$p(\mathbf{Y},\boldsymbol{Z}|\theta,\boldsymbol{\beta})=\Pi_{n=1}^{N}\{p(Z_n|\theta)p(Y_n|Z_n,\boldsymbol{\beta})\}$$

这里直接求解参数{$\mathbf{Z}$, θ, $\boldsymbol{\beta}$}使 Y 的后验函数最大化很困难，积分掉 Z，采用近似推断的方法来得到 log 后验的一个下界：

$$\log p(\mathbf{Y}|\theta,\boldsymbol{\beta})=\log\int_{\boldsymbol{Z}}p(\mathbf{Y},\boldsymbol{Z}|\theta,\boldsymbol{\beta})\,d\mathbf{Z}$$

$$\geq\int_{\boldsymbol{Z}}q(\mathbf{Z})\log p(\mathbf{Y},\boldsymbol{Z}|\theta,\boldsymbol{\beta})d\mathbf{Z}-\int_{\boldsymbol{Z}}q(\mathbf{Z})\log q(\mathbf{Z})d\mathbf{Z}$$

其中，$\int_{\boldsymbol{Z}}q(\mathbf{Z})\log p(\mathbf{Y},\boldsymbol{Z}|\theta,\boldsymbol{\beta})d\mathbf{Z}=E_{q(\mathbf{Z})}\log p(\mathbf{Y},\boldsymbol{Z}|\theta,\boldsymbol{\beta})$，$-\int_{\boldsymbol{Z}}q(\mathbf{Z})\log q(\mathbf{Z})d\mathbf{Z}=H(q(\mathbf{Z}))$，前一项是对 log 联合分布中的变量 $\mathbf{Z}$ 在 $q(\mathbf{Z})$下求期望，后一项是$q(\mathbf{Z})$的相对熵，

$$\log p(\mathbf{Y}|\theta,\boldsymbol{\beta})\geq LB=E_{q(\mathbf{Z})}\log p(\mathbf{Y},\boldsymbol{Z}|\theta,\boldsymbol{\beta})+H(q(\mathbf{Z}))$$

很明显，$(\mathbf{Y},\mathbf{Z})$中的 Z 可以看作是不完全观察值，这样的求解方法就是 EM 算法，通常我们也称这样的 $\mathbf{Z}$ 是隐变量。$q(\mathbf{Z})$的近似分布在这里可利用后验 $q(\mathbf{Z})\propto$

$p(\mathbf{Y},\boldsymbol{Z}|\theta^{old},\boldsymbol{\beta}^{old})$，其中$\theta^{old},\boldsymbol{\beta}^{old}$从上一次 M 步中得到。由于 $q(\mathbf{Z}|\theta^{old},\boldsymbol{\beta}^{old}) \propto p(\mathbf{Y},\boldsymbol{Z}|\theta^{old},\boldsymbol{\beta}^{old}) \propto \Pi_{n=1}^{N}\{p(Z_n|\theta^{old})p(Y_n|Z_n,\boldsymbol{\beta}^{old})\}$，可见在$\{\theta^{old},\boldsymbol{\beta}^{old}\}$已知的情况下$\{Z_n\}$之间是相互独立的，所以有：

$q(\mathbf{Z}) = \Pi_{n=1}^{N} q(Z_n)$，$q(Z_n) \propto p(Z_n|\theta^{old})p(Y_n|Z_n,\boldsymbol{\beta}^{old})$，$H(q(\mathbf{Z})) = \Sigma_{n=1}^{N} H(q(Z_n))$，

$E_{q(\mathbf{z})}\log p(\mathbf{Y},\boldsymbol{Z}|\theta,\beta) = \Sigma_{n=1}^{N} E_{q(z_n)}\log p(Z_n|\theta) + \Sigma_{n=1}^{N} E_{q(z_n)}\log p(Y_n|Z_n,\boldsymbol{\beta})$，

$LB = \Sigma_{n=1}^{N}\{E_{q(z_n)}\log p(Z_n|\theta) + E_{q(z_n)}\log p(Y_n|Z_n,\boldsymbol{\beta}) + H(q(Z_n))\}$，

下面我们分别讨论模型的 E-M 步骤：在 pLSA 中$q(Z_n) \propto \theta_{Z_n}^{old}\beta_{Z_nY_n}^{old}$，

E-step：$E_{q(z_n)}\log p(Z_n|\theta) = \Sigma_{k=1}^{K} E_{q(z_n)}(\delta(Z_n = k)\log\theta_k) =$

$(\Sigma_{k=1}^{K}\theta_k^{old}\beta_{kY_n}^{old}\log\theta_k)/(\Sigma_{k=1}^{K}\theta_k^{old}\beta_{kY_n}^{old})$

$E_{q(z_n)}\log p(Y_n|Z_n,\boldsymbol{\beta}) = (\Sigma_{k=1}^{K}\theta_k^{old}\beta_{kY_n}^{old}\log\beta_{k,Y_n})/(\Sigma_{k=1}^{K}\theta_k^{old}\beta_{kY_n}^{old})$，

记$A_{kn}=\theta_k^{old}\beta_{kY_n}^{old}$，$A_n=\Sigma_{k=1}^{K}A_{kn}$，计算出 A_{kn} 与 A_n(k=1…K，n=1…N)。

其中的 $\delta(\cdot)$是指示函数，当$Z_n = k$值为 1 否则为 0。

M-step：由于 $H(q(Z_n))$是常数，所以我们只需要考虑 LB 中的$\{\theta,\beta\}$使如下目标函数最大化：

$\Sigma_{k=1}^{K}\Sigma_{n=1}^{N}\{(A_{kn}\log\theta_k)/A_n + (A_{kn}\log\beta_{k,Y_n})/A_n\}$，由于 θ 和 β_k 满足单纯形取值约束，所以添加拉格朗日项得到新的优化函数：$O_1(\theta) + O_2(\boldsymbol{\beta})$

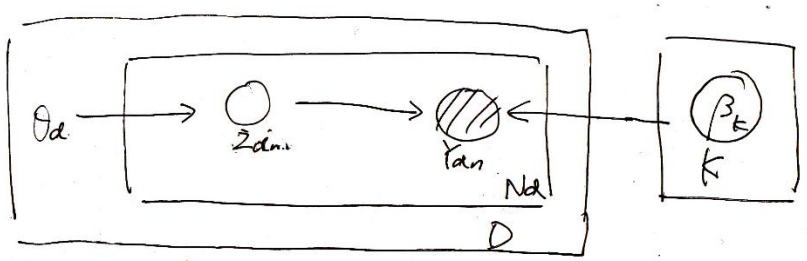
$O_1 = \Sigma_{k=1}^{K}\Sigma_{n=1}^{N}\{(A_{kn}\log\theta_k)/A_n\} + \lambda(\Sigma_{k=1}^{K}\theta_k - 1)$，

$O_2 = \Sigma_{k=1}^{K}\{\Sigma_{n=1}^{N}(A_{kn}\log\beta_{k,Y_n})/A_n + \lambda_k(\Sigma_{v=1}^{V}\beta_{kv} - 1)\}$，

最大化O_1得到：$\theta_k = \frac{1}{N}\Sigma_{n=1}^{N}\frac{A_{kn}}{A_n}$，

最大化O_2得到：$\beta_{kv} = \frac{1}{\Sigma_{v'=1}^{V}\Sigma_{n=1}^{N}\frac{A_{kn}}{A_n}\delta(Y_n=v')}\Sigma_{n=1}^{N}\frac{A_{kn}}{A_n}\delta(Y_n = v)$。

上面的 PLSA 推导可以拓展到 D 篇文档的情形，如下：



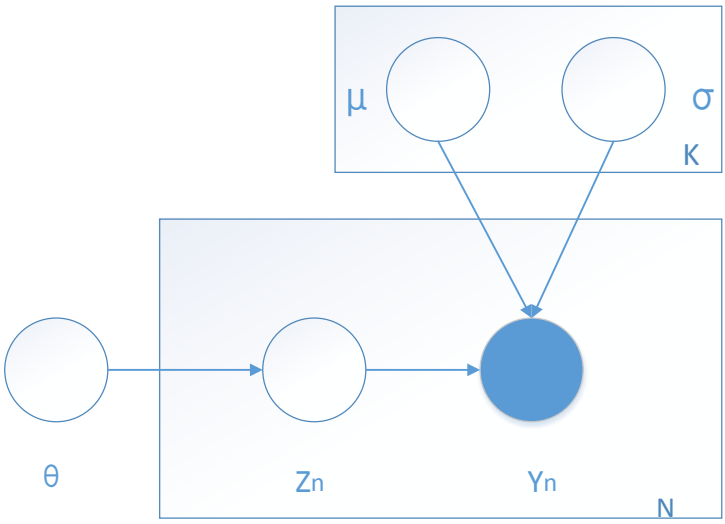
$\sum_{k=1}^{K}\sum_{d=1}^{D}\sum_{n=1}^{Nd}\{(A_{dkn}\cdot\log\theta_{dk})/A_{dkn} + (A_{dkn}\log\beta_{kY_{dn}})/A_{don}\}$

$A_{dkn} = \theta_{dk}^{old}\beta_{kY_{dn}}^{old}$　　$A_{don} = \sum_k A_{dkn}$

对θ_{dk}求O_1 obtain　$\theta_{dk} = \frac{1}{Nd}\sum_{n=1}^{Nd}\frac{A_{dkn}}{A_{don}}$

对β_{kv}求O_2 obtain　$\beta_{kv} = \frac{\sum_{d=1}^{D}\sum_{n=1}^{Nd}\frac{A_{dkn}}{A_{don}}\delta(Y_{dn}=v)}{\sum_{v'=1}^{V}\sum_{d=1}^{D}\sum_{n=1}^{Nd}\frac{A_{dkn}}{A_n}\delta(Y_{dn}=v')}$

高斯混合模型 **GMM** 及 **EM** 算法



在 GMM 中，在 Z_n 的条件下选择外层的第 k 个参数族来确定 Y_n 的一元高斯分布 $p(Y_n|Z_n=k,\boldsymbol{\mu},\boldsymbol{\sigma})=N(Y_n|\mu_k,\sigma^2{}_k)$。为方便起见让 β_k 也代表 μ_k 和 $\sigma^2{}_k$，这样两个模型

将放在一个求解框架下做讨论。在 GMM 中$q(Z_n)\propto\theta_{Z_n}^{old}\,N(Y_n|\mu_{Z_n}^{old},\sigma_{Z_n}^{old})$,

E-step： 记$A_{kn}=\theta_k^{old}\,N(Y_n|\mu_{Z_n}^{old},\sigma_{Z_n}^{old})$, $A_n=\Sigma_{k=1}^{K}A_{kn}$,

计算出 A_{kn} 与 A_n(k=1…K，n=1…N)。

$$E_{q(z_n)}\log p(Z_n|\theta)=\Sigma_{k=1}^{K}E_{q(z_n)}(\delta(Z_n=k)\log\theta_k)=(\Sigma_{k=1}^{K}A_{kn}\log\theta_k)/A_n,$$

$$E_{q(z_n)}\log p(Y_n|Z_n,\boldsymbol{\mu},\boldsymbol{\sigma})=\left(\Sigma_{k=1}^{K}A_{kn}\log N(Y_n|\mu_k,\sigma_k^2)\right)/A_n。$$

M-step： 由于 $H(q(Z_n))$是常数，所以我们只需要考虑 LB 中的$\{\theta,\beta\}$使如下目标函数最大化：$\Sigma_{k=1}^{K}\Sigma_{n=1}^{N}\{(A_{kn}\log\theta_k)/A_n+(A_{kn}\log N(Y_n|\mu_k,\sigma_k^2))/A_n\}$ 。

所以添加拉格朗日项得到新的优化函数:$O_1(\theta)+O_2(\boldsymbol{\mu},\boldsymbol{\sigma})$

$$O_1=\Sigma_{k=1}^{K}\Sigma_{n=1}^{N}\{(A_{kn}\log\theta_k)/A_n\}+\lambda(\Sigma_{k=1}^{K}\theta_k-1)$$

$$O_2=\Sigma_{k=1}^{K}\{\Sigma_{n=1}^{N}(A_{kn}\log N(Y_n|\mu_k,\sigma_k))/A_n\}$$

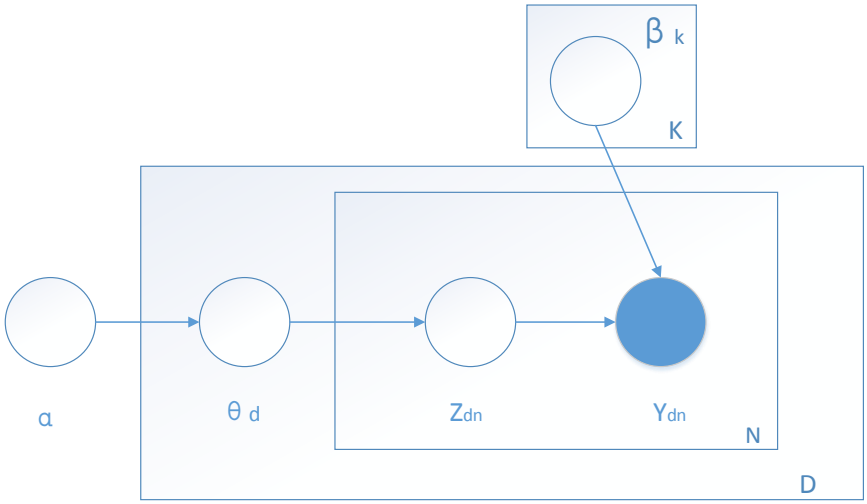
最大化 O_1 得到：$\theta_k=\frac{1}{N}\Sigma_{n=1}^{N}\frac{A_{kn}}{A_n}$，最大化 O_2 得到：

$$\mu_k=(\Sigma_{n=1}^{N}\frac{A_{kn}}{A_n}Y_n)/(\Sigma_{n=1}^{N}\frac{A_{kn}}{A_n}),\quad \sigma_k^2=(\Sigma_{n=1}^{N}\frac{A_{kn}}{A_n}(Y_n-\mu_k)^2)/(\Sigma_{n=1}^{N}\frac{A_{kn}}{A_n}),$$

可以看到，如果 GMM 中的高斯分布是多元高斯分布，仅需要考虑 O_2 最大化即可。

层级推广 **LDA&GMM-LDA** 及 **VB-EM** 算法

LDA 及 **VB-EM** 算法



LDA 的图模型表示

Latent Dirichlet Allocation 模型如图所示，它是 pLSA 的层级推广模型。pLSA 中主题的选择由θ确定，话题变量z_i服从θ确定的 Multinomial 分布，记作Z_i~Mult(θ)。文本中的单词{Y_1 …… Y_n}都各自对应一个主题{Z_1 …… Z_n}，这样文本中的每个Y_i服从$\beta_{z_i.}$确定的 Multinomial 分布，记作Y_i~Mult($\beta_{z_i.}$)。因而与此类似，LDA 的生成过程如下：

> 对每个文件 d：
>
> 从α 确定的 Dirichlet 分布产生话题分布参数θ_d，即θ_d~$Dir(\alpha)$。
>
> 对每个单词$Y_{di} \in \{Y_{d1} \ldots\ldots Y_{dn}\}$：
>
> (a) 从θ_d确定的 Multinomial 分布产生话题Z_{di}，即Z_{di}~Mult(θ_d)。
>
> (b) 从$\beta_{z_i.}$确定的 Multinomial 分布产生单词Y_{di}，即Y_{di}~Mult($\beta_{z_i.}$)。

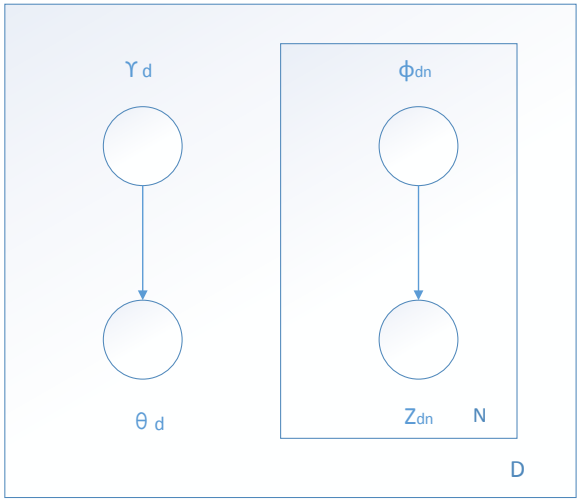
α 是k 维 Dirichlet 分布的参数，其中α_i>0。这样得到联合分布如下：

$$p(\mathbf{Y}, \boldsymbol{Z}, \boldsymbol{\theta}|\boldsymbol{\beta}, \alpha) = \Pi_{d=1}^{D} p(\theta_d|\alpha) \Pi_{n=1}^{Nd} \{p(Z_{dn}|\theta_d) p(Y_{dn}|Z_{dn}, \boldsymbol{\beta})\}$$

采用近似推断的方法来得到 log 后验的一个下界：

$$\log p(\mathbf{Y}|\alpha, \boldsymbol{\beta}) = \log \int_{(\boldsymbol{Z},\boldsymbol{\theta})} p(\mathbf{Y}, \boldsymbol{Z}, \boldsymbol{\theta}|\boldsymbol{\beta}, \alpha)\, d(\boldsymbol{Z}, \boldsymbol{\theta}) \geq LB = E_{q(\mathbf{Z},\boldsymbol{\theta})} \log p(\mathbf{Y}, \boldsymbol{Z}, \boldsymbol{\theta}|\boldsymbol{\beta}, \alpha) + H(q(\mathbf{Z}, \boldsymbol{\theta}))$$

这里$(\mathbf{Z},\boldsymbol{\theta})$是该模型的隐变量，$q(\mathbf{Z},\boldsymbol{\theta})$将选择完全可分解的变分分布，如下图所示，这样的方法叫作变分贝叶斯(Variational Bayesian)近似推断。由于 Dir 和 Mult 分布具有共轭性质，$\boldsymbol{\theta}$与$\mathbf{Z}$的所有分量在变分近似分布下相互独立，这些都将简化 E 步的计算。



VB 近似分布表

$$\theta_d \sim Dir(\gamma_d)$$

$$Z_{dn} \sim Mult(\phi_{dn})$$

$$q(\mathbf{Z},\boldsymbol{\theta}) = q(\mathbf{Z})q(\boldsymbol{\theta})$$

$$q(\mathbf{Z}) = \Pi_{d=1}^{D}\Pi_{n=1}^{Nd} q(Z_{dn}|\phi_{dn})$$

$$q(\boldsymbol{\theta}) = \Pi_{d=1}^{D} q(\theta_d|\gamma_d)$$

$$q = \Pi_{d=1}^{D} q(\theta_d)\{\Pi_{n=1}^{Nd} q(Z_{dn})\}$$

$q(\mathbf{Z},\boldsymbol{\theta})$近似分布的图模型表示和分布表

$$\log p(\mathbf{Y},\boldsymbol{Z},\boldsymbol{\theta}|\boldsymbol{\beta},\alpha) = \Sigma_{d=1}^{D}\{\log p(\theta_d|\alpha) + \Sigma_{n=1}^{Nd}\{\log p(Z_{dn}|\theta_d) + \log p(Y_{dn}|Z_{dn},\boldsymbol{\beta})\}\},$$

$$H\big(q(\mathbf{Z},\boldsymbol{\theta})\big) = \Sigma_{d=1}^{D}\{H\big(q(\theta_d)\big) + \Sigma_{n=1}^{Nd} H\big(q(Z_{dn})\big)\},$$

$$H\big(q(\boldsymbol{Z_d},\theta_d)\big) = H\big(q(\theta_d)\big) + \Sigma_{n=1}^{Nd} H\big(q(Z_{dn})\big),$$

令$LB_d = E_{q(\mathbf{Z_d},\theta_d)}\{\log p(\theta_d|\alpha) + \Sigma_{n=1}^{Nd}\{\log p(Z_{dn}|\theta_d) + \log p(Y_{dn}|Z_{dn},\boldsymbol{\beta})\}\}$

$+H(q(\boldsymbol{Z_d},\theta_d))$，则 LB=$E_{q(\mathbf{Z},\boldsymbol{\theta})}\log p(\mathbf{Y},\boldsymbol{Z},\boldsymbol{\theta}|\boldsymbol{\beta},\alpha) + H\big(q(\mathbf{Z},\boldsymbol{\theta})\big) = \Sigma_{d=1}^{D} LB_d$

由于我们引入 VB 近似分布，最大化 LB 将涉及到参数$\{\alpha,\boldsymbol{\beta}\}$与所有的$\{\boldsymbol{\phi_d},\gamma_d\}$。又由于 LB 可分解成 D 个$LB_d$下界和的形式，在参数$\{\alpha,\boldsymbol{\beta}\}$给定的情况下，最大化$LB_d$将只涉及到$\{\boldsymbol{\phi_d},\gamma_d\}$，所以我们采取分块优化的方法，即先固定$\{\alpha,\boldsymbol{\beta}\}$更新所有的$\{\boldsymbol{\phi_d},\gamma_d\}$，再固定$\{\boldsymbol{\phi},\boldsymbol{\gamma}\}$更新$\{\alpha,\boldsymbol{\beta}\}$。

LDA 变分近似推断算法

E-step: 下面在整个 E 步中，LB_d下界的讨论将会去掉 d，其中的θ、Z_n以及γ、ϕ_n均默认带有下标 d。$E_{q(Z,\theta)}\log p(\theta|\alpha) = \phi_{Dir}(\alpha)^T E_{q(\theta)}\big(u_{Dir}(\theta)\big) + g_{Dir}(\alpha)$

$$= \begin{pmatrix}\alpha_1 - 1\\ \cdots\cdots\\ \alpha_K - 1\end{pmatrix}^T \begin{pmatrix}\Psi(\gamma_1) - \Psi\big(\sum_{i=1}^{K}\gamma_i\big)\\ \cdots\cdots\\ \Psi(\gamma_K) - \Psi\big(\sum_{i=1}^{K}\gamma_i\big)\end{pmatrix} + \log\frac{\Gamma(\sum_{i=1}^{K}\alpha_i)}{\Pi_{i=1}^{K}\Gamma(\alpha_i)} \qquad \text{term①}$$

$$E_{q(\mathrm{Z},\theta)}\Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}\log \mathrm{p}(Z_n|\theta) = E_{q(\theta)}\phi_{\mathrm{Mult}}(\theta)^T\Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}E_{q(\mathrm{Z_n})}\mathrm{u}_{\mathrm{Mult}}(\mathrm{Z_n}) \quad \text{term②}$$

$$= \begin{pmatrix}\Psi(\gamma_1) - \Psi\left(\sum_{i=1}^{\mathrm{K}}\gamma_i\right)\\ \cdots\cdots \\ \Psi(\gamma_{\mathrm{K}}) - \Psi\left(\sum_{i=1}^{\mathrm{K}}\gamma_i\right)\end{pmatrix}^T \begin{pmatrix}\Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}\phi_{\mathrm{n}1}\\ \cdots\cdots \\ \Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}\phi_{\mathrm{n}\mathrm{K}}\end{pmatrix}$$

$$E_{q(\mathrm{Z},\theta)}\Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}\log \mathrm{p}(\mathrm{Y_n}|\mathrm{Z_n},\beta)=\Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}E_{q(\mathrm{Z_n})}\log\beta_{\mathrm{Z_nY_n}} \quad \text{term③}$$

$$= \Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}E_{q(\mathrm{Z_n})}\delta(Z_n = k)\log\beta_{k\mathrm{Y_n}} = \Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}\Sigma_{\mathrm{k}=1}^{K}\phi_{\mathrm{nk}}\log\beta_{k\mathrm{Y_n}}$$

$$\mathrm{H}\big(\mathrm{q}(\theta)\big) = -E_{q(\theta)}\log q(\theta) = -\phi_{\mathrm{Dir}}(\gamma)^T E_{q(\theta)}\mathrm{u}_{\mathrm{Dir}}(\theta) - \mathrm{g}_{\mathrm{Dir}}(\gamma)$$

$$= -\begin{pmatrix}\gamma_1 - 1\\ \cdots\cdots \\ \gamma_{\mathrm{K}} - 1\end{pmatrix}^{\mathrm{T}} \begin{pmatrix}\Psi(\gamma_1) - \Psi\left(\sum_{i=1}^{\mathrm{K}}\gamma_i\right)\\ \cdots\cdots \\ \Psi(\gamma_{\mathrm{K}}) - \Psi\left(\sum_{i=1}^{\mathrm{K}}\gamma_i\right)\end{pmatrix} - \log\frac{\Gamma\left(\sum_{i=1}^{\mathrm{K}}\gamma_i\right)}{\Pi_{\mathrm{i}=1}^{\mathrm{K}}\Gamma(\gamma_i)} \quad \text{term④}$$

$$\Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}\mathrm{H}\big(\mathrm{q}(\mathrm{Z_n})\big) = -\Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}E_{\mathrm{q}(\mathrm{Z_n})}\log\mathrm{q}(\mathrm{Z_n}) = -\Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}\begin{pmatrix}log\phi_{n1}\\ \cdots\cdots \\ log\phi_{nk}\end{pmatrix}^T \begin{pmatrix}\phi_{n1}\\ \cdots\cdots \\ \phi_{nk}\end{pmatrix} \quad \text{term⑤}$$

$\mathrm{LB_d} = \text{term①} + \text{term②} + \text{term③} + \text{term④} + \text{term⑤}$

这其中涉及到参数γ的有 1、2、4 涉及到参数ϕ的有 2、3、5。先考虑γ：

$$\text{term}\{1,2,4\} = \left(-\begin{pmatrix}\gamma_1 - 1\\ \cdots\cdots \\ \gamma_{\mathrm{K}} - 1\end{pmatrix} + \begin{pmatrix}\alpha_1 - 1\\ \cdots\cdots \\ \alpha_{\mathrm{K}} - 1\end{pmatrix} + \right.$$

$$\left.\begin{pmatrix}\Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}\phi_{\mathrm{n}1}\\ \cdots\cdots \\ \Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}\phi_{\mathrm{n}\mathrm{K}}\end{pmatrix}\right)^{\mathrm{T}} \begin{pmatrix}\Psi(\gamma_1) - \Psi\left(\sum_{i=1}^{\mathrm{K}}\gamma_i\right)\\ \cdots\cdots \\ \Psi(\gamma_{\mathrm{K}}) - \Psi\left(\sum_{i=1}^{\mathrm{K}}\gamma_i\right)\end{pmatrix} - \log\frac{\Gamma\left(\sum_{i=1}^{\mathrm{K}}\gamma_i\right)}{\Pi_{\mathrm{i}=1}^{\mathrm{K}}\Gamma(\gamma_i)}$$

$$\text{term}\{1,2,4\}(\gamma_{\mathrm{i}}) = (1-\gamma_i)\left(\Psi(\gamma_{\mathrm{i}}) - \Psi\left(\sum_{k=1}^{\mathrm{K}}\gamma_k\right)\right)$$

$$-\log\Gamma\left(\sum_{i=1}^{\mathrm{K}}\gamma_i\right) + \log\Gamma(\gamma_i) + \Sigma_{\mathrm{k}=1}^{\mathrm{K}}(\alpha_k - 1 + \Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}\phi_{\mathrm{nk}})(\Psi(\gamma_{\mathrm{k}}) -$$

$$\Psi(\textstyle\sum_{k=1}^{\mathrm{K}}\gamma_k))$$

上式对γ_{i}求导并置为 0，可得到一个局部最大值点：$\frac{\partial\text{term}\{1,2,4\}(\gamma_{\mathrm{i}})}{\partial\gamma_{\mathrm{i}}} =$

$$\Psi'(\gamma_{\mathrm{i}})\big(\alpha_i + \Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}\phi_{\mathrm{ni}} - \gamma_i\big) - \Psi'\left(\sum_{k=1}^{\mathrm{K}}\gamma_k\right)\Sigma_{\mathrm{k}=1}^{K}\big(\alpha_k + \Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}\phi_{\mathrm{nk}} - \gamma_k\big)$$

γ_{i}满足等式：$\gamma_{\mathrm{i}} = \alpha_i + \Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}\phi_{\mathrm{ni}}\ (\mathrm{i} = 1 \ldots \mathrm{K})$。

考虑ϕ_{n}，并加入拉格朗日约束

$$\text{term}\{2,3,5\}(\phi_{\mathrm{n}}) = \begin{pmatrix}\Psi(\gamma_1) - \Psi\left(\sum_{i=1}^{\mathrm{K}}\gamma_i\right)\\ \cdots\cdots \\ \Psi(\gamma_{\mathrm{K}}) - \Psi\left(\sum_{i=1}^{\mathrm{K}}\gamma_i\right)\end{pmatrix}^T \begin{pmatrix}\phi_{\mathrm{n}1}\\ \cdots\cdots \\ \phi_{\mathrm{n}\mathrm{K}}\end{pmatrix} -$$

$$\begin{pmatrix} log\phi_{n1} \\ \dots\dots \\ log\phi_{nk} \end{pmatrix}^T \begin{pmatrix} \phi_{n1} \\ \dots\dots \\ \phi_{nk} \end{pmatrix} + \Sigma_{\mathrm{k}=1}^{K}\phi_{\mathrm{nk}}\log\beta_{k\mathrm{Y_n}} + \lambda_{\mathrm{n}}(\Sigma_{\mathrm{k}=1}^{K}\phi_{\mathrm{nk}} - 1)$$

$$\mathrm{term}\{2,3,5\}(\phi_{\mathrm{ni}}) = \phi_{\mathrm{ni}}\left(\Psi(\gamma_{\mathrm{i}}) - \Psi\left(\sum_{i=1}^{\mathrm{K}}\gamma_i\right)\right) - \phi_{\mathrm{ni}}\log\phi_{\mathrm{ni}} + \phi_{\mathrm{ni}}\log\beta_{i\mathrm{Y_n}}$$

$$+\lambda_{\mathrm{n}}(\Sigma_{\mathrm{k}=1}^{K}\phi_{\mathrm{nk}} - 1)$$

$$\frac{\partial\mathrm{term}\{2,3,5\}(\phi_{\mathrm{ni}})}{\partial\phi_{\mathrm{ni}}} = \Psi(\gamma_{\mathrm{i}}) - \Psi\left(\sum_{i=1}^{\mathrm{K}}\gamma_i\right) + \log\beta_{i\mathrm{Y_n}} - \log\phi_{\mathrm{ni}} - 1 + \lambda_{\mathrm{n}},$$

设置导数为 0，这样事实上有：

$\phi_{\mathrm{ni}} \propto \beta_{\mathrm{iY_n}}\exp(\Psi(\gamma_{\mathrm{i}}) - \Psi\left(\sum_{i=1}^{\mathrm{K}}\gamma_i\right))$，$\Sigma_{\mathrm{k}=1}^{K}\phi_{\mathrm{nk}} = 1$。

M-step：对参数α 的更新只涉及 term①，但这里需要考虑 D 个 term①的和，目标函数如下：

$$\mathrm{Oj}(\alpha) = \begin{pmatrix} \alpha_1 - 1 \\ \dots\dots \\ \alpha_{\mathrm{K}} - 1 \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} \Sigma_{\mathrm{d}=1}^{D}\Psi(\gamma_{\mathrm{d}1}) - \Sigma_{\mathrm{d}=1}^{D}\Psi\left(\sum_{i=1}^{\mathrm{K}}\gamma_{di}\right) \\ \dots\dots \\ \Sigma_{\mathrm{d}=1}^{D}\Psi(\gamma_{\mathrm{d}\mathrm{K}}) - \Sigma_{\mathrm{d}=1}^{D}\Psi\left(\sum_{i=1}^{\mathrm{K}}\gamma_{di}\right) \end{pmatrix} + \mathrm{D}\log\frac{\Gamma(\sum_{i=1}^{\mathrm{K}}\alpha_i)}{\Pi_{\mathrm{i}=1}^{\mathrm{K}}\Gamma(\alpha_i)}$$

$$\frac{\partial\mathrm{Oj}(\alpha))}{\partial\alpha_{\mathrm{i}}} = D\left(\Psi\left(\sum_{i=1}^{\mathrm{K}}\alpha_i\right) - \Psi(\alpha_{\mathrm{i}})\right) + \Sigma_{\mathrm{d}=1}^{D}(\Psi(\gamma_{\mathrm{di}}) - \Psi\left(\sum_{k=1}^{\mathrm{K}}\gamma_{dk}\right))$$

优化方法 1：

$$\frac{\partial\mathrm{Oj}(\alpha)}{\partial\alpha_{\mathrm{i}}\partial\alpha_{\mathrm{j}}} = D(\Psi'\left(\sum_{k=1}^{\mathrm{K}}\alpha_k\right) - \delta(i,j)\Psi'(\alpha_i)),$$

则有$Hessian(\alpha) = \mathrm{diag(h)} + \mathrm{s}11^{\mathrm{T}}$,其中$\mathrm{s} = D\Psi'\left(\sum_{k=1}^{\mathrm{K}}\alpha_k\right)$，

$\mathrm{h} = \left(-\mathrm{D}\Psi'(\alpha_1), \dots\dots, -\mathrm{D}\Psi'(\alpha_K)\right)^{\mathrm{T}}$，diag 将 h 转为对角矩阵。

由于α_{i}的导数涉及到$\alpha_{\mathrm{j}}(j \neq i)$，因此我们采用 Newton-Raphson 优化来找到函数的局部稳定点：

$\alpha_{\mathrm{new}} = \alpha_{old} - Hessian(\alpha_{old})^{-1}\partial\mathrm{Oj}(\alpha^{old})$，

计算其中的黑塞矩阵和梯度向量，则有：

$$Hessian(\alpha)^{-1} = diag(h)^{-1} - \frac{diag(h)^{-1}11^T diag(h)^{-1}}{s^{-1} + \sum_{k=1}^{\mathrm{K}}h_k^{-1}}$$

$$\left(Hessian(\alpha)^{-1}\partial\mathrm{Oj}(\alpha)\right)_k = \frac{(\partial\mathrm{Oj}(\alpha))_{\mathrm{k}} - c}{h_k}，\ 其中 c = \frac{\sum_{k=1}^{\mathrm{K}}(\partial\mathrm{Oj}(\alpha))_{\mathrm{k}}/h_k}{s^{-1} + \sum_{k=1}^{\mathrm{K}}h_k^{-1}},$$

$\left(\partial\mathrm{Oj}(\alpha)\right)_{\mathrm{k}} = \frac{\partial\mathrm{Oj}(\alpha)}{\partial\alpha_{\mathrm{k}}}$如上。

循环 Newton-Raphson 迭代直到α 收敛。

优化方法 2：

令$\frac{\partial \mathrm{Oj}(\alpha)}{\partial \alpha_{\mathrm{i}}}=0$，$\Psi(\alpha_{\mathrm{i}})=\Psi\left(\sum_{i=1}^{\mathrm{K}}\alpha_i\right)+\frac{1}{D}\Sigma_{\mathrm{d}=1}^{D}(\Psi(\gamma_{\mathrm{di}})-\Psi\left(\sum_{k=1}^{\mathrm{K}}\gamma_{dk}\right)$

利用上式循环更新直到收敛。

同样β的更新只涉及 D 个 term③，但对每个β_{k}需要加入朗格朗日约束，目标函数如下：

$\Sigma_{\mathrm{d}=1}^{\mathrm{D}}\Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}\Sigma_{\mathrm{k}=1}^{K}\phi_{\mathrm{dnk}}\log\beta_{k\mathrm{Y_n}}+\Sigma_{\mathrm{k}=1}^{\mathrm{K}}\lambda_{\mathrm{k}}(\Sigma_{\mathrm{v}=1}^{\mathrm{V}}\beta_{\mathrm{kv}}-1)$

$=\Sigma_{\mathrm{d}=1}^{\mathrm{D}}\Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}\Sigma_{\mathrm{k}=1}^{K}\Sigma_{\mathrm{v}=1}^{V}\delta(\mathrm{Y_{dn}}=\mathrm{v})\phi_{\mathrm{dnk}}\log\beta_{\mathrm{kv}}+\Sigma_{\mathrm{k}=1}^{\mathrm{K}}\lambda_{\mathrm{k}}(\Sigma_{\mathrm{v}=1}^{\mathrm{V}}\beta_{\mathrm{kv}}-1)$

对上式中的某个β_{kv}求导并置 0，得：$\frac{\Sigma_{\mathrm{d}=1}^{\mathrm{D}}\Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}\delta(\mathrm{Y_{dn}}=\mathrm{v})\phi_{\mathrm{dnk}}}{\beta_{\mathrm{kv}}}+\lambda_{\mathrm{k}}=0$，

即$\beta_{\mathrm{kv}}\propto\Sigma_{\mathrm{d}=1}^{\mathrm{D}}\Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}\delta(\mathrm{Y_{dn}}=\mathrm{v})\phi_{\mathrm{dnk}}$。

输出：$\alpha,\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\gamma}$。

输入：文本单词$\mathbf{Y}$，收敛率ϵ，最大迭代次数 S，初始化$\{\alpha,\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\gamma}\}$。

For t =1…S：

For d=1…D：

$\gamma_{\mathrm{di}}=\alpha_{di}+\Sigma_{\mathrm{n}=1}^{\mathrm{N_d}}\phi_{\mathrm{dni}}\ (\mathrm{i}=1\ldots\mathrm{K})$

For n=1…$\mathrm{N_d}$：

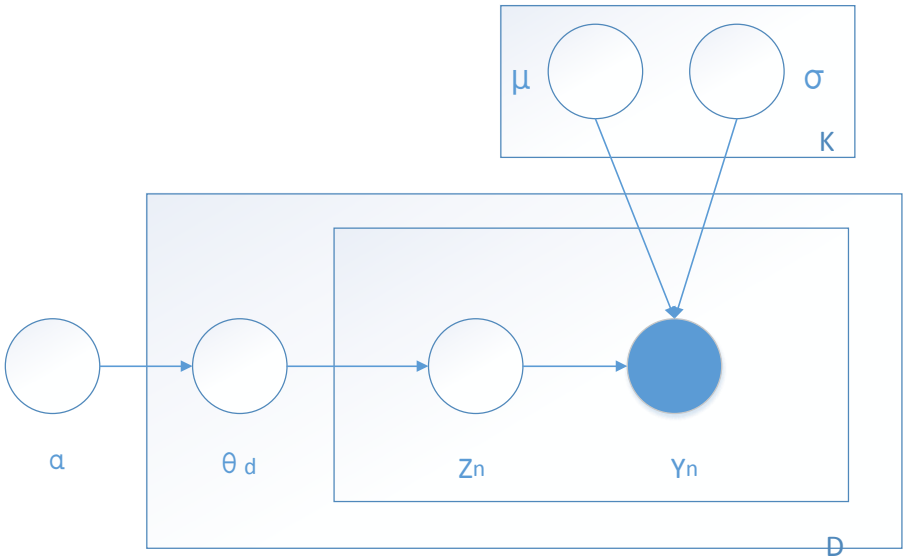
$\phi_{\mathrm{dni}}\propto\beta_{\mathrm{iY_{dn}}}\exp(\Psi(\gamma_{\mathrm{di}})-\Psi\left(\sum_{i=1}^{\mathrm{K}}\gamma_{di}\right))$，$\Sigma_{\mathrm{k}=1}^{K}\phi_{\mathrm{dnk}}=1$

Update $\beta_{\mathrm{kv}}\propto\Sigma_{\mathrm{d}=1}^{\mathrm{D}}\Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}\delta(\mathrm{Y_{dn}}=\mathrm{v})\phi_{\mathrm{dnk}}$

Update α

IF converge：break

GMM-LDA 及 VB-EM 算法



以上，我们得到了 LDA 的变分贝叶斯近似推断的 E-M 推导（VB-EM）。回顾 pLSA 和 GMM 的推断过程可以发现，两个模型的不同之处仅在涉及外层参数 β 的计算和更新上。这对 LDA 和 GMM-LDA 同样适用，下面给出 GMM-LDA 类似的过程。

E-step： 与 LDA 仅有 term③不同，

$$E_{q(Z,\theta)}\Sigma_n^{Nd}{}_{=1}\log p(Y_n|Z_n,\mu,\sigma)=\Sigma_n^{Nd}{}_{=1}E_{q(Z_n)}\log N(Y_n|\mu_{Z_n},\sigma^2{}_{Z_n})=$$

$$\Sigma_{nNd=1}E_{q(Z_n)}\delta(Z_n=k)\log N(Y_n|\mu_k,\sigma_{2k})=\Sigma_{nNd=1}\Sigma_{kK=1}\phi_{nk}\log N(Y_n|\mu_k,\sigma_{2k})$$

这样在 E 步的计算中，我们就有（省略下标 d）：

$\gamma_i=\alpha_i+\Sigma_{n=1}^{Nd}\phi_{ni}\ (i=1\dots K)$与$\phi_{ni}\propto N(Y_n|\mu_i,\sigma_i^2)(\Psi(\gamma_i)-\Psi(\sum_{i=1}^{K}\gamma_i))$，

$\Sigma_{k=1}^{K}\phi_{nk}=1$。

M-step： 对 α 的更新不涉及到{μ, σ}，同 LDA。对{μ, σ}的更新只涉及的目标函数如下：$\Sigma_{d=1}^{D}\Sigma_{n=1}^{Nd}\Sigma_{k=1}^{K}\phi_{dnk}\log N(Y_{dn}|\mu_k,\sigma_k^2)$。

不妨让所有{Y_{dn}}以下标 t 重新标号$T=\Sigma_{d=1}^{D}N_d$，

则有$\Sigma_{t=1}^{T}\Sigma_{k=1}^{K}A_{kt}\log N(Y_t|\mu_k,\sigma_k^2)$ ，最大化它则得到

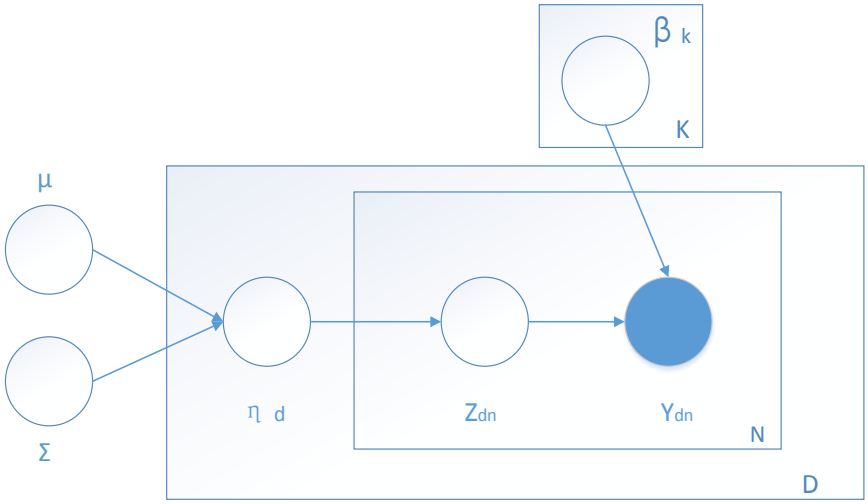
$\mu_k=(\Sigma_{t=1}^{T}A_{kt}Y_t)/(\Sigma_{t=1}^{T}A_{kt})$，$\sigma_k^2=(\Sigma_{t=1}^{T}A_{kt}(Y_t-\mu_k)^2)/(\Sigma_{t=1}^{T}A_{kt})$

这部分的内容我们详细探讨了由 pLSA、GMM 推广到层级贝叶斯的 LDA、GMM-LDA，整个模型的求解，我们都放在了一个统一的 E-M 优化框架下。共轭分布的选取简化了 E 步的推导。给出了一般求解层级贝叶斯图模型中隐变量近似分布的选取方法，一种是采用上一轮参数来求期望，这等同于将隐变量看作是不完全观察值，另一种是利用完全分解的变分分布，这里完全分解的意思是隐变量之间在变分分布下完全独立，这同样将简化 E 步的计算。

总结：利用了指数族分布的性质，如对自然统计量的期望、共轭分布的特点，这些理论基础简化了我们的计算过程，也有利于后续内容的展开。

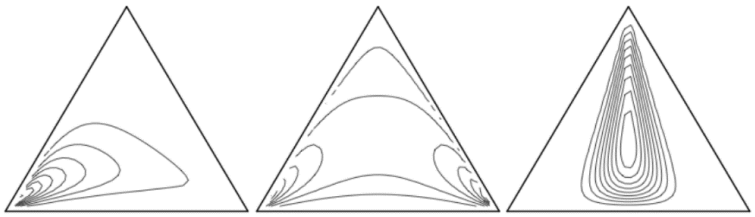
CTM 模型 及 VB-EM 算法

LDA 模型选择外层超参数分布主要基于 Dirichlet 是 Mult 的共轭先验，Dir 的选取有利于简化 VB 下界的计算，也满足主题之间的可交换性(可祥见 Dir 分布的相关性质)。但同时 Dir 不容易刻画话题之间的关联程度，为此引入罗杰斯特高斯分布(logistic normal distribution)，即$\eta \sim N(\mu, \Sigma)$，$\theta_i = \exp(\eta_i)/\Sigma_{i=1}^{K}\exp(\eta_i)$，$Z_n \sim Mult(\theta)$，这个模型如图所示是 CTM(correlated topic model)模型。



CTM(correlated topic model)的图模型表示

下图所示为 3 维罗杰斯特高斯分布密度图，左起第一个Σ为非零对角均值矩阵，



中间分量 1 和 2 具有负的相关系数，最右边的分量 1 和 2 具有正的相关系数。

罗杰斯特高斯分布图

CTM 的 VB-EM 型近似推断，近似分布$q(\mathbf{Z}, \boldsymbol{\eta})$将类似 LDA 中选择完全可分解的变分分布：

$\eta_d \sim N(\lambda_d, \gamma_d^2)$，$Z_{dn} \sim Mult(\phi_{dn})$，$q(\mathbf{Z}, \boldsymbol{\eta}) = q(\mathbf{Z})q(\boldsymbol{\eta})$，$q(\mathbf{Z}) = \Pi_{d=1}^{D}\Pi_{n=1}^{Nd}q(Z_{dn})$，$q(\boldsymbol{\eta}) = \Pi_{d=1}^{D}q(\eta_d)$，$\eta_d$的近似分布是$\lambda_d, \gamma_d^2$确定的多元高斯，$\gamma_d^2 = \mathrm{diag}(\gamma_{d1}^2 \ldots \gamma_{dK}^2)$，

即近似分布中η_d的分量相互独立，联合分布为：

$$p(\mathbf{Y},\boldsymbol{Z},\boldsymbol{\eta}|\boldsymbol{\beta},\mu,\Lambda^{-1})=\Pi_{d=1}^{D}p(\eta_d|\mu,\Lambda^{-1})\Pi_{n=1}^{Nd}\{p(Z_{dn}|\eta_d)p(Y_{dn}|Z_{dn},\boldsymbol{\beta})\}$$

CTM 变分近似推断算法推导

采用近似推断方法来得到 log 后验的一个下界：

$$\log p(\mathbf{Y}|\mu,\Lambda^{-1},\boldsymbol{\beta})=\log\int_{(\boldsymbol{Z},\boldsymbol{\eta})}p(\mathbf{Y},\boldsymbol{Z},\boldsymbol{\eta}|\boldsymbol{\beta},\alpha)\,d(\boldsymbol{Z},\boldsymbol{\eta})$$

$$\geq LB=E_{(\boldsymbol{Z},\boldsymbol{\eta})}\log p(\mathbf{Y},\boldsymbol{Z},\boldsymbol{\eta}|\boldsymbol{\beta},\alpha)+H(q(\boldsymbol{Z},\boldsymbol{\eta}))$$

$$\log p(\mathbf{Y},\boldsymbol{Z},\boldsymbol{\eta}|\boldsymbol{\beta},\mu,\Lambda^{-1})=\Sigma_{d=1}^{D}\{\log p(\eta_d|\mu,\Lambda^{-1})+\Sigma_{n=1}^{Nd}\{p(Z_{dn}|\eta_d)p(Y_{dn}|Z_{dn},\boldsymbol{\beta})\}\}$$

$$H(q(\boldsymbol{Z},\boldsymbol{\eta}))=\Sigma_{d=1}^{D}\{H(q(\eta_d))+\Sigma_{n=1}^{Nd}H(q(Z_{dn}))\},$$

$$H(q(\boldsymbol{Z_d},\eta_d))=H(q(\eta_d))+\Sigma_{n=1}^{Nd}H(q(Z_{dn})),$$

$$LB_d=$$

$$E_{q(\mathbf{Z_d},\eta_d)}\{\log p(\eta_d|\mu,\Lambda^{-1})+\Sigma_{n=1}^{Nd}\{\log p(Z_{dn}|\eta_d)+\log p(Y_{dn}|Z_{dn},\boldsymbol{\beta})\}\}+$$

$H(q(\boldsymbol{Z_d},\eta_d))$，则有：

$$LB=E_{q(\mathbf{Z},\boldsymbol{\eta})}p(\mathbf{Y},\boldsymbol{Z},\boldsymbol{\eta}|\boldsymbol{\beta},\mu,\Lambda^{-1})+H(q(\mathbf{Z},\boldsymbol{\eta}))=\Sigma_{d=1}^{D}LB_d$$

E-step：下面的讨论同样省略下标 d。

$$E_{q(Z,\eta)}\log p(\eta|\mu,\Lambda^{-1})=\phi_{gaus}(\mu,\Lambda^{-1})^{T}E_{q(\eta)}u_{gaus}(\eta)+g_{gaus}(\mu,\Lambda^{-1})$$

$$=-\frac{1}{2}tr(\Lambda(E_{q(\eta)}\eta\eta^{T}-\mu E_{q(\eta)}\eta^{T}-E_{q(\eta)}\eta\mu^{T}+\mu\mu^{T}))+\log\sqrt{\frac{|\Lambda|}{(2\pi)^{K}}}$$

$$=-\frac{1}{2}tr(\Lambda(\boldsymbol{\gamma^2}-\mu\lambda^{T}-\lambda\mu^{T}+\mu\mu^{T}+\lambda\lambda^{T}))+\log\sqrt{\frac{|\Lambda|}{(2\pi)^{K}}}\qquad \text{term①}$$

$$E_{q(Z)}\Sigma_{n=1}^{Nd}\log p(Z_n|\eta)=E_{q(\eta)}\phi_{Mult}\begin{pmatrix}\exp(\eta_1)/\Sigma_{i=1}^{K}\exp(\eta_i)\\ \cdots\cdots\\ \exp(\eta_K)/\Sigma_{i=1}^{K}\exp(\eta_i)\end{pmatrix}^{T}\Sigma_{n=1}^{Nd}E_{q(Z_n)}u_{Mult}(Z_n)$$

$$=\begin{pmatrix}\lambda_1-E_{q(\eta)}\log\Sigma_{i=1}^{K}\exp(\eta_i)\\ \cdots\cdots\\ \lambda_K-E_{q(\eta)}\log\Sigma_{i=1}^{K}\exp(\eta_i)\end{pmatrix}^{T}\begin{pmatrix}\Sigma_{n=1}^{Nd}\phi_{n1}\\ \cdots\cdots\\ \Sigma_{n=1}^{Nd}\phi_{nK}\end{pmatrix}\qquad \text{term②}$$

利用不等式：$E_{q(\eta)}\log\Sigma_{i=1}^{K}\exp(\eta_i)\leq\zeta^{-1}(\Sigma_{i=1}^{K}E_{q(\eta)}\exp(\eta_i))-1+\log\zeta$

其中在变分近似分布中$q(\eta)=\Pi_{k=1}^{K}q(\eta_k)$，有

$$E_{q(\eta)}\exp(\eta_i)=E_{q(\eta_i)}\exp(\eta_i)=\exp\{\lambda_i+\gamma_i^2/2\}$$

$$E_{q(\mathrm{Z},\theta)}\Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}\log \mathrm{p}(Z_n|\eta) \geq \begin{pmatrix} \lambda_1 - \zeta^{-1}\left(\Sigma_{\mathrm{i}=1}^{\mathrm{K}} \exp\{\lambda_{\mathrm{i}} + \gamma_{\mathrm{i}}^2/2\}\right) + 1 - \log\zeta \\ \cdots\cdots \\ \lambda_{\mathrm{K}} - \zeta^{-1}\left(\Sigma_{\mathrm{i}=1}^{\mathrm{K}} \exp\{\lambda_{\mathrm{i}} + \gamma_{\mathrm{i}}^2/2\}\right) + 1 - \log\zeta \end{pmatrix}^T \begin{pmatrix} \Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}\phi_{\mathrm{n}1} \\ \cdots\cdots \\ \Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}\phi_{\mathrm{n}\mathrm{K}} \end{pmatrix}$$

$$E_{q(\mathrm{Z},\eta)}\Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}\log \mathrm{p}(\mathrm{Y_n}|\mathrm{Z_n},\beta) = \Sigma_{\mathrm{n}=1}^{\mathrm{Nd}} E_{q(\mathrm{Z_n})} \log \beta_{\mathrm{Z_nY_n}}$$

$$= \Sigma_{\mathrm{n}=1}^{\mathrm{Nd}} E_{q(\mathrm{Z_n})} \delta(Z_n = k) \log \beta_{k\mathrm{Y_n}} = \Sigma_{\mathrm{n}=1}^{\mathrm{Nd}} \Sigma_{\mathrm{k}=1}^{K} \phi_{\mathrm{nk}} \log \beta_{k\mathrm{Y_n}} \qquad \text{term③}$$

$$\mathrm{H}\big(\mathrm{q}(\eta)\big) = -E_{q(\eta)} \log q(\eta) = -\phi_{\mathrm{gaus}}(\lambda,\gamma^2)^T E_{q(\eta)} \mathrm{u}_{\mathrm{gaus}}(\eta) - \qquad \text{term④}$$

$$\mathrm{g}_{\mathrm{gaus}}(\lambda,\gamma^2) =$$

$$-\log\sqrt{\frac{\left|\gamma^{2^{-1}}\right|}{(2\pi)^K}} + \frac{1}{2}K$$

$$\Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}\mathrm{H}\big(\mathrm{q}(\mathrm{Z_n})\big) = -\Sigma_{\mathrm{n}=1}^{\mathrm{Nd}} E_{\mathrm{q}(\mathrm{Z_n})} \log \mathrm{q}(\mathrm{Z_n}) = -\Sigma_{\mathrm{n}=1}^{\mathrm{Nd}} \begin{pmatrix} log\phi_{n1} \\ \cdots\cdots \\ log\phi_{nk} \end{pmatrix}^T \begin{pmatrix} \phi_{n1} \\ \cdots\cdots \\ \phi_{nk} \end{pmatrix} \qquad \text{term⑤}$$

由于在 term②中做了不等式放缩，$\mathrm{LB_d} \geq$ term① + term② + term③ + term④ + term⑤ 。这其中涉及到参数$\{\lambda,\gamma^2\}$的有 1、2、4 涉及到参数ϕ的有 2、3、5。先考虑参数$\{\lambda,\gamma^2\}$：

$$\mathrm{term}\{1,2,4\}(\{\lambda,\gamma^2\}) = \lambda^T\Lambda\mu - \frac{1}{2}\lambda^T\Lambda\lambda - \frac{1}{2}\mathrm{tr}(\Lambda\gamma^2)$$

$$+\frac{1}{2}\Sigma_{\mathrm{i}=1}^{\mathrm{K}}\log\boldsymbol{\gamma_i^2} + \begin{pmatrix} \lambda_1 - \zeta^{-1}\left(\Sigma_{\mathrm{i}=1}^{\mathrm{K}} \exp\{\lambda_{\mathrm{i}} + \gamma_{\mathrm{i}}^2/2\}\right) \\ \cdots\cdots \\ \lambda_{\mathrm{K}} - \zeta^{-1}\left(\Sigma_{\mathrm{i}=1}^{\mathrm{K}} \exp\{\lambda_{\mathrm{i}} + \gamma_{\mathrm{i}}^2/2\}\right) \end{pmatrix}^T \begin{pmatrix} \Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}\phi_{\mathrm{n}1} \\ \cdots\cdots \\ \Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}\phi_{\mathrm{n}\mathrm{K}} \end{pmatrix}$$

$$\frac{\partial \mathrm{term}\{1,2,4\}(\lambda)}{\partial\lambda} = -\Lambda(\lambda - \mu) + \Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}\phi_{\mathrm{n}1:\mathrm{K}} - \frac{\mathrm{N}}{\zeta}\left\{\exp\left(\lambda_{\mathrm{i}} + \frac{\gamma_i^2}{2}\right)\right\}_{\mathrm{i}=1:\mathrm{K}}$$

$$\frac{\partial \mathrm{term}\{1,2,4\}(\gamma^2)}{\partial \boldsymbol{\gamma_i^2}} = -\frac{\Lambda_{\mathrm{ii}}}{2} - \frac{\mathrm{N}}{2\zeta}\exp(\lambda_{\mathrm{i}} + \gamma_i^2/2) + 1/(2\gamma_i^2)$$

这两个式子无法得到解析解，为了得到相应的参数值使导数为 0，可使用“牛顿法”。考虑ϕ_{n}，并加入拉格朗日约束，仅有 term②与LDA稍有不同，立刻有：$\phi_{\mathrm{ni}} \propto \beta_{\mathrm{i}\mathrm{Y_n}}\exp(\lambda_{\mathrm{i}})$，$\Sigma_{\mathrm{k}=1}^{K}\phi_{\mathrm{nk}} = 1$。最后考虑对 term②中的$\zeta$更新，求导置导数为 0，立刻有：$\zeta = \Sigma_{\mathrm{i}=1}^{\mathrm{K}}\exp(\lambda_{\mathrm{i}} + \gamma_i^2/2)$

M-step：对参数μ,Λ^{-1}的更新只涉及 term①，但这里需要考虑 D 个 term①的和，目标函数如下：$\mathrm{Oj}(\mu,\Lambda^{-1}) = \Sigma_{\mathrm{d}=1}^{D}(\lambda_d^T\Lambda\mu - \frac{1}{2}\mathrm{tr}\big(\Lambda\boldsymbol{\gamma_d^2}\big) - \frac{1}{2}(\mu^{\mathrm{T}}\Lambda\mu + \lambda_d^T\Lambda\lambda_d)) + D\log\sqrt{\frac{|\Lambda|}{(2\pi)^K}}$，$\frac{\partial \mathrm{Oj}(\mu,\Lambda^{-1})}{\partial\mu} = \Sigma_{\mathrm{d}=1}^{D}(\lambda_d^T\Lambda - \Lambda\mu)$，

$$\frac{\partial \mathrm{Oj}(\mu,\Lambda^{-1})}{\partial \Lambda}=\Sigma_{\mathrm{d}=1}^{D}\left(\mu\lambda_d^T-\frac{1}{2}\boldsymbol{\gamma_d^2}-\frac{1}{2}\mu\mu^{\mathrm{T}}-\frac{1}{2}\lambda_d\lambda_d^T\right)+\frac{D}{2}\Lambda^{-1}$$

所以有：$\mu=\frac{1}{\mathrm{D}}\Sigma_{\mathrm{d}=1}^{D}\lambda_{\mathrm{d}}$，$\Lambda^{-1}=\frac{1}{\mathrm{D}}\Sigma_{\mathrm{d}=1}^{D}(\boldsymbol{\gamma_d^2}+(\lambda_d-\mu)(\lambda_d-\mu)^T)$

同样β的更新只涉及 D 个 term③，即$\beta_{\mathrm{kv}}\propto\Sigma_{\mathrm{d}=1}^{\mathrm{D}}\Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}\delta(\mathrm{Y_{dn}}=\mathrm{v})\phi_{\mathrm{dnk}}$。

输出：$\mu,\Lambda,\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\lambda},\boldsymbol{\gamma}$。

输入：文本单词$\mathbf{Y}$，收敛率ϵ，最大迭代次数 S，初始化$\{\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\lambda},\boldsymbol{\gamma}\}$。

For t =1…S：

For d=1…D：

Update λ_d,γ_d^2

For n=1…$\mathrm{N_d}$：

$\phi_{\mathrm{dni}}\propto\beta_{\mathrm{iY_{dn}}}\exp(\lambda_{\mathrm{di}})$，$\Sigma_{\mathrm{k}=1}^{K}\phi_{\mathrm{dnk}}=1$

Update $\beta_{\mathrm{kv}}\propto\Sigma_{\mathrm{d}=1}^{\mathrm{D}}\Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}\delta(\mathrm{Y_{dn}}=\mathrm{v})\phi_{\mathrm{dnk}}$

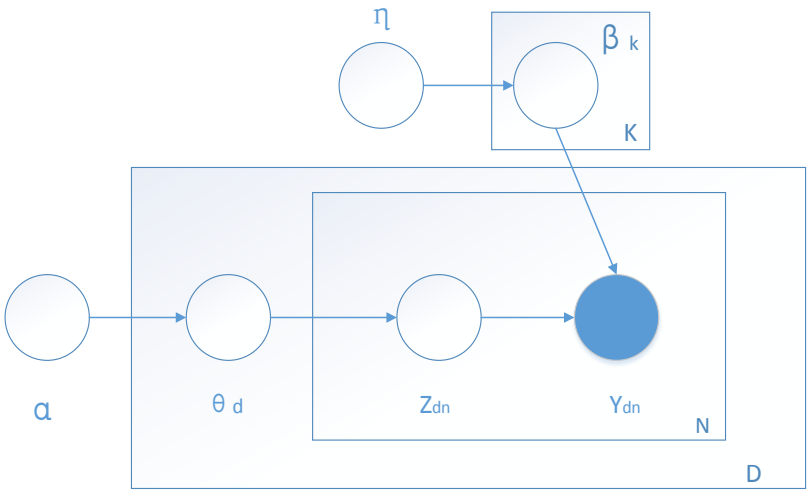
Update $\mu=\frac{1}{\mathrm{D}}\Sigma_{\mathrm{d}=1}^{D}\lambda_{\mathrm{d}}$，$\Lambda^{-1}=\frac{1}{\mathrm{D}}\Sigma_{\mathrm{d}=1}^{D}(\gamma_d^2+(\lambda_d-\mu)(\lambda_d-\mu)^T)$

IF converge：break

对称先验 LDA

LDA MCMC 算法 1

前面给出了 LDA 的 VB-EM 推断算法，这一节将介绍光滑 LDA 模型(smoothed-LDA)也称作"对称先验 LDA"，以及它的 MCMC 采样算法(Gibbs 采样算法)。



Smooth-LDA 或"对称先验 LDA"的图模型表示

由于β_k是多项式分布的参数，极大似然估计会使一些新单词（占总文件比重很少）的概率趋向于 0，这个趋势我们从 LDA 的 E-M 更新公式可以窥见：$\beta_{kv} \propto \Sigma_{d=1}^{D}\Sigma_{n=1}^{Nd}\delta(Y_{dn}=v)\phi_{dnk}$。为避免这种处理单词数量大时出现的问题，实际操作中对变量值做 laplace-smoothing 是常见的方法，但更有效的方法是增加外层超参数η，有$\beta_k \sim Dir(\eta)$，这就是 smoothed-LDA 也称作对称 LDA。考虑联合分布为：

$$p(\mathbf{Y},\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{\beta}|\alpha,\eta)=\Pi_{d=1}^{D}p(\theta_d|\alpha)\Pi_{n=1}^{Nd}\{p(Z_{dn}|\theta_d)p(Y_{dn}|Z_{dn},\boldsymbol{\beta})\}\Pi_{k=1}^{K}p(\beta_k|\eta)$$

蒙特卡洛采样算法，将联合似然看作随机域(无向图)，并利用完全条件分布构造马尔科夫链得到参数的估计值。联合分布$p(\mathbf{Y},\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{\beta}|\alpha,\eta)$写出完全条件分布有：

$$p(\theta_d|\mathbf{Z},\alpha)，\ p(Z_{dn}|\mathbf{Y},\boldsymbol{\theta},\beta)，\ p(\beta_k|\boldsymbol{Z},\boldsymbol{Y},\eta)$$

$$p(\theta_d|\mathbf{Z},\alpha)\propto p(\mathbf{Z_d},\theta_d|\alpha)=p(\mathbf{Z_d}|\theta_d)p(\theta_d|\alpha)\propto\left(\Pi_{n=1}^{N_d}\Pi_{k=1}^{K}\theta_{dk}^{\delta(Z_{dn}=k)}\right)\left(\Pi_{k=1}^{K}\theta_{dk}^{\alpha_k-1}\right)$$

$$=\Pi_{k=1}^{K}\left(\theta_{dk}^{\alpha_k-1}\Pi_{n=1}^{N_d}\theta_{dk}^{\delta(Z_{dn}=k)}\right)=\Pi_{k=1}^{K}\left(\theta_{dk}^{\Sigma_{n=1}^{N_d}\delta(Z_{dn}=k)+\alpha_k-1}\right),$$

$$\theta_d\sim Dir\left(\alpha+\left\{\sum_{n=1}^{N_d}\delta(Z_{dn}=k)\right\}_{k=1\ldots K}\right),$$

$$p(Z_{dn}=k|\mathbf{Y},\boldsymbol{\theta},\beta)\propto p(Z_{dn}=k|\theta_d)p(Y_{dn}|Z_{dn}=k,\boldsymbol{\beta})=\theta_{dk}\beta_{kY_{dn}},$$

$$Z_{dn}\sim \mathrm{Mult}\left(\{\theta_{dk}\beta_{kY_{dn}}\}_{k=1\ldots K}\right),$$

$$p(\beta_k|\boldsymbol{Z},\boldsymbol{Y},\eta)=\ p(\boldsymbol{Y}|\beta_k,\boldsymbol{Z})p(\beta_k|\eta)=\ \Pi_{d=1}^{D}\Pi_{n=1}^{N_d}\beta_{kY_{dn}}^{\delta(Z_{dn}=k)}\Pi_{v=1}^{V}\beta_{kv}^{\eta_v-1}$$

$$=\Pi_{v=1}^{V}\beta_{kv}^{\eta_v-1+\sum_{d=1}^{D}\sum_{n=1}^{N_d}\delta(Y_{dn}=v,Z_{dn}=k)},$$

$$\beta_k\sim \mathrm{Dir}\left(\eta+\left\{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\delta(Y_{dn}=v,Z_{dn}=k)\right\}_{v=1\ldots V}\right),$$

可见 Dir 和 Mult 共轭关系使得θ_d仍然为 Dir 分布，这样只需对参数α,η做更新，循环对变量做 Gibbs 采样直到消除初始值的影响，其中{}表示一个向量，而其中的每个分量对应于它右下角下标。

对称先验 LDA 的 MCMC 算法①

输出：$\boldsymbol{\beta},\mathbf{Z},\boldsymbol{\theta}$。

输入：文本单词$\mathbf{Y}$，最大迭代次数 S，初始化$\{\boldsymbol{\beta},\alpha,\eta)\}$。

For t=1…S：

$\alpha^t=\ \alpha^{t-1}+\left\{\sum_{n=1}^{N_d}\delta(Z_{dn}^{t-1}=k)\right\}_{k=1\ldots K}$

$\theta_d^t\sim \mathrm{Dir}(\alpha^t)$其中 $d=1\ldots D$

$Z_{dn}^t\sim \mathrm{Mult}\left(\{\theta_{dk}^t\beta_{kY_{dn}}^{t-1}\}_{k=1\ldots K}\right)$ 其中 $d=1\ldots D, n=1\ldots N_d$

$\eta^t=\eta^{t-1}+\{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\delta(Y_{dn}=v,Z_{dn}^t=k)\}_{v=1\ldots V}$

$\beta_k^t\sim \mathrm{Dir}(\eta^t)$其中 $k=1\ldots K$

LDA MCMC 算法 2

另一种方法利用 Dir 和 Mult 共轭关系基于给定超参数α、β对$(\boldsymbol{\theta},\boldsymbol{\beta})$做积分，这样的 LDA 求解策略叫作“坍缩贝叶斯方法”（collapsed bayesian），视$\boldsymbol{Z}$为待估计参数。联合分布为：

$$\mathrm{p}(\mathbf{Y},\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{\beta}|\alpha,\eta)=\Pi_{\mathrm{d}=1}^{\mathrm{D}}\mathrm{p}(\theta_{\mathrm{d}}|\alpha)\Pi_{\mathrm{n}=1}^{\mathrm{Nd}}\{\mathrm{p}(Z_{dn}|\theta_{\mathrm{d}})\mathrm{p}(\mathrm{Y_{dn}}|\mathrm{Z_{dn}},\boldsymbol{\beta})\}\Pi_{\mathrm{k}=1}^{\mathrm{K}}\mathrm{p}(\beta_{\mathrm{k}}|\eta)=$$

$$\left(\frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\Pi_{\mathrm{i}=1}^{\mathrm{k}}\Gamma(\alpha_i)}\right)^{D}\left(\frac{\Gamma(\sum_{i=1}^{V}\eta_i)}{\Pi_{\mathrm{i}=1}^{\mathrm{V}}\Gamma(\eta_i)}\right)^{\mathrm{K}}(\Pi_{\mathrm{k}=1}^{\mathrm{K}}\Pi_{\mathrm{i}=1}^{\mathrm{V}}\beta_{ki}^{\eta_i-1})(\Pi_{\mathrm{d}=1}^{\mathrm{D}}\Pi_{\mathrm{i}=1}^{\mathrm{K}}\theta_{di}^{\alpha_i-1})(\Pi_{\mathrm{d}=1}^{\mathrm{D}}\Pi_{\mathrm{n}=1}^{\mathrm{Nd}}\theta_{\mathrm{d}Z_{dn}}\beta_{\mathrm{Z_{dn}Y_{dn}}})=$$

$$\mathrm{const}(\alpha,\eta)(\Pi_{\mathrm{k}=1}^{\mathrm{K}}\Pi_{\mathrm{i}=1}^{\mathrm{V}}\beta_{ki}^{\eta_i-1+\Sigma_{\mathrm{d}=1}^{\mathrm{D}}\Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}\delta(Y_{dn}=i,Z_{dn}=k)})(\Pi_{\mathrm{d}=1}^{\mathrm{D}}\Pi_{\mathrm{k}=1}^{\mathrm{K}}\theta_{dk}^{\alpha_k-1+\Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}\delta(Z_{dn}=k)}$$

我们引入新的记号$\#\{\mathrm{d},\mathrm{v},\mathrm{k}\}=\Sigma_{\mathrm{n}=1}^{\mathrm{N_d}}\delta(\mathrm{Y_{dn}}=\mathrm{v},Z_{dn}=\mathrm{k})$，表示在文本 d 中单词下标为 v 且话题值为 k 的单词数量，记号$\#\{\mathrm{d},\cdot,\mathrm{k}\}=\Sigma_{\mathrm{n}=1}^{\mathrm{N_d}}\delta(Z_{dn}=k)$表示文本 d 中话题值为 k 的单词出现的次数，类似地，$\#\{\cdot,\cdot,\mathrm{k}\}=\Sigma_{\mathrm{d}=1}^{\mathrm{D}}\Sigma_{\mathrm{n}=1}^{\mathrm{N_d}}\delta(Z_{dn}=k)$，它们与$\{\mathbf{Y},\boldsymbol{Z}\}$有关。利用这个模型，可得到$\mathrm{Z_{dn}}$条件分布如下：

$$\mathrm{p}(\mathrm{Z_{dn}}=\mathrm{k}|\boldsymbol{Z}^{\neg(\mathrm{d,n})})\propto\frac{\eta_{\mathrm{Y_{dn}}}+\#\{\cdot,\mathrm{Y_{dn}},\mathrm{k}\}^{\neg(\mathrm{d,n})}}{\sum_{v=1}^{\mathrm{V}}\eta_v+\#\{\cdot,\cdot,\mathrm{k}\}^{\neg(\mathrm{d,n})}}\frac{\alpha_k+\#\{\mathrm{d},\cdot,\mathrm{k}\}^{\neg(\mathrm{d,n})}}{\sum_{k=1}^{\mathrm{K}}\alpha_k+\#\{\mathrm{d},\cdot,\cdot\}^{\neg(\mathrm{d,n})}}$$

输出：$\mathbf{Z}$。

输入：文本单词$\mathbf{Y}$，最大迭代次数 S，初始化$\{\alpha,\eta)\}$。

For t=1…S：

For d=1…D：

$$\mathrm{p}\left(\mathrm{Z_{dn}}=\mathrm{k}\middle|\boldsymbol{Z}^{\neg(\mathrm{d,n})}\right)\propto\frac{\eta_{\mathrm{Y_{dn}}}+\#\{\cdot,\mathrm{Y_{dn}},\mathrm{k}\}^{\neg(\mathrm{d,n})}}{\sum_{v=1}^{\mathrm{V}}\eta_v+\#\{\cdot,\cdot,\mathrm{k}\}^{\neg(\mathrm{d,n})}}\frac{\alpha_k+\#\{\mathrm{d},\cdot,\mathrm{k}\}^{\neg(\mathrm{d,n})}}{\sum_{k=1}^{\mathrm{K}}\alpha_k+\#\{\mathrm{d},\cdot,\cdot\}^{\neg(\mathrm{d,n})}}$$

对称先验 LDA 的 MCMC 算法②推导

$$\int_{(\boldsymbol{\theta},\boldsymbol{\beta})}\mathrm{p}(\mathbf{Y},\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{\beta}|\alpha,\eta)\,\mathrm{d}(\boldsymbol{\theta},\boldsymbol{\beta})$$

$$=\mathrm{const}(\alpha,\eta)\int_{(\boldsymbol{\theta},\boldsymbol{\beta})}(\Pi_{\mathrm{k}=1}^{\mathrm{K}}\Pi_{\mathrm{v}=1}^{\mathrm{V}}\beta_{kv}^{\eta_v-1+\#\{\cdot,\mathrm{v,k}\}})(\Pi_{\mathrm{d}=1}^{\mathrm{D}}\Pi_{\mathrm{k}=1}^{\mathrm{K}}\theta_{dk}^{\alpha_k-1+\#\{\mathrm{d},\cdot,\mathrm{k}\}})\,\mathrm{d}(\boldsymbol{\theta},\boldsymbol{\beta})$$

$$=\mathrm{const}(\alpha,\eta)(\Pi_{\mathrm{k}=1}^{\mathrm{K}}\int_{\beta_{\mathbf{k}}}\Pi_{\mathrm{v}=1}^{\mathrm{V}}\beta_{kv}^{\eta_v-1+\#\{\cdot,\mathrm{v,k}\}}\mathrm{d}\beta_{\mathbf{k}})(\Pi_{\mathrm{d}=1}^{\mathrm{D}}\int_{\theta_{\mathrm{d}}}\Pi_{\mathrm{k}=1}^{\mathrm{K}}\theta_{dk}^{\alpha_k-1+\#\{\mathrm{d},\cdot,\mathrm{k}\}}\mathrm{d}\theta_{\mathrm{d}})$$

观察上式，dirichlet 分布满足$\int_{\beta_k}\frac{\Gamma(\sum_{v=1}^{V}\eta_v+\#\{\cdot,v,k\})}{\prod_{v=1}^{V}\Gamma(\eta_v+\#\{\cdot,v,k\})}\prod_{i=1}^{V}\beta_{kv}^{\eta_v-1+\#\{\cdot,v,k\}}d\beta_k=1$，所以我们得到：

$$const(\alpha,\eta)\prod_{k=1}^{K}\frac{\prod_{v=1}^{V}\Gamma(\eta_v+\#\{\cdot,v,k\})}{\Gamma(\sum_{v=1}^{V}(\eta_v+\#\{\cdot,v,k\}))}\prod_{d=1}^{D}\frac{\prod_{k=1}^{K}\Gamma(\alpha_k+\#\{d,\cdot,k\})}{\Gamma\left(\sum_{k=1}^{K}(\alpha_k+\#\{d,\cdot,k\})\right)}$$

$$p(\mathbf{Y},\boldsymbol{Z}|\alpha,\eta)\propto\prod_{k=1}^{K}\frac{\prod_{v=1}^{V}\Gamma(\eta_v+\#\{\cdot,v,k\})}{\Gamma(\sum_{v=1}^{V}(\eta_v+\#\{\cdot,v,k\}))}\prod_{d=1}^{D}\frac{\prod_{k=1}^{K}\Gamma(\alpha_k+\#\{d,\cdot,k\})}{\Gamma(\sum_{k=1}^{K}(\alpha_k+\#\{d,\cdot,k\}))}$$

对其中的变量Z_{dn},我们有，$\neg(d,n)$的含义是除去对第 d 个文件第 n 个单词对应的变量。

$$p\left(\mathbf{Y}^{\neg(d,n)},\boldsymbol{Z}^{\neg(d,n)}\middle|\alpha,\eta\right)$$

$$\propto\prod_{k=1}^{K}\frac{\prod_{v=1}^{V}\Gamma(\eta_v+\#\{\cdot,v,k\}^{\neg(d,n)})}{\Gamma(\sum_{v=1}^{V}(\eta_v+\#\{\cdot,v,k\}^{\neg(d,n)}))}\prod_{d=1}^{D}\frac{\prod_{k=1}^{K}\Gamma(\alpha_k+\#\{d,\cdot,k\}^{\neg(d,n)})}{\Gamma(\sum_{k=1}^{K}(\alpha_k+\#\{d,\cdot,k\}^{\neg(d,n)}))}$$

$$p(Z_{dn},Y_{dn}|\mathbf{Y}^{\neg(d,n)},\boldsymbol{Z}^{\neg(d,n)},\alpha,\eta)$$

$$\propto\frac{\prod_{k=1}^{K}\frac{\prod_{v=1}^{V}\Gamma(\eta_v+\#\{\cdot,v,k\})}{\Gamma(\sum_{v=1}^{V}\eta_v+\#\{\cdot,\cdot,k\}))}\prod_{d=1}^{D}\frac{\prod_{k=1}^{K}\Gamma(\alpha_k+\#\{d,\cdot,k\})}{\Gamma(\sum_{k=1}^{K}\alpha_k+\#\{d,\cdot,\cdot\})}}{\prod_{k=1}^{K}\frac{\prod_{v=1}^{V}\Gamma(\eta_v+\#\{\cdot,v,k\}^{\neg(d,n)})}{\Gamma(\sum_{v=1}^{V}\eta_v+\#\{\cdot,\cdot,k\}^{\neg(d,n)})}\prod_{d=1}^{D}\frac{\prod_{k=1}^{K}\Gamma(\alpha_k+\#\{d,\cdot,k\}^{\neg(d,n)})}{\Gamma(\sum_{k=1}^{K}\alpha_k+\#\{d,\cdot,\cdot\}^{\neg(d,n)})}}$$

上面这个式子就是在$\boldsymbol{Z}^{\neg(d,n)},\mathbf{Y},\alpha,\eta$的条件下$Z_{dn}$的概率分布。由于$\Gamma(t+1)=t\Gamma(t)$，所以有

$$p\left(Z_{dn}=k,Y_{dn}=v\middle|\mathbf{Y}^{\neg(d,n)},\boldsymbol{Z}^{\neg(d,n)},\alpha,\eta\right)$$

$$\propto\frac{\eta_v+\#\{\cdot,v,k\}^{\neg(d,n)}}{\sum_{v=1}^{V}\eta_v+\#\{\cdot,\cdot,k\}^{\neg(d,n)}}\frac{\alpha_k+\#\{d,\cdot,k\}^{\neg(d,n)}}{\sum_{k=1}^{K}\alpha_k+\#\{d,\cdot,\cdot\}^{\neg(d,n)}}$$

我们写出两项所依赖的条件变量，事实上有如下关系：

$$p\left(Y_{dn}=v\middle|\mathbf{Y}^{\neg(d,n)},\boldsymbol{Z}^{\neg(d,n)},Z_{dn}=k,\eta\right)=\frac{\eta_v+\#\{\cdot,v,k\}^{\neg(d,n)}}{\sum_{v=1}^{V}\eta_v+\#\{\cdot,\cdot,k\}^{\neg(d,n)}}$$

$$p\left(Z_{dn}=k\middle|\boldsymbol{Y}_d^{\neg(d,n)},\mathbf{Z}_d^{\neg(d,n)},\alpha\right)=\frac{\alpha_k+\#\{d,\cdot,k\}^{\neg(d,n)}}{\sum_{k=1}^{K}\alpha_k+\#\{d,\cdot,\cdot\}^{\neg(d,n)}}$$

即前一项是在已知$\{\mathbf{Y}^{\neg(d,n)},\boldsymbol{Z}^{\neg(d,n)},\eta\}$且$Z_{dn}$选择话题 k 条件下，产生单词 v 的概率。后一项是在已知$\{\mathbf{Y}^{\neg(d,n)},\boldsymbol{Z}^{\neg(d,n)},\alpha\}$条件下产生话题 k 的概率。可见积分掉

$(\boldsymbol{\theta},\boldsymbol{\beta})$后，模型仍然保持生成含义。省略掉已知参数 **Y**，α，η，我们只需推断参数 **Z**，可利用下式做 Gibbs 采样：

$$\mathrm{p}(\mathrm{Z_{dn}}=\mathrm{k}|\boldsymbol{Z}^{\neg(\mathrm{d,n})}) \propto \frac{\eta_{\mathrm{Y_{dn}}}+\#\{\cdot,\mathrm{Y_{dn}},\mathrm{k}\}^{\neg(\mathrm{d,n})}}{\sum_{v=1}^{\mathrm{V}}\eta_v+\#\{\cdot,\cdot,\mathrm{k}\}^{\neg(\mathrm{d,n})}}\frac{\alpha_k+\#\{\mathrm{d},\cdot,\mathrm{k}\}^{\neg(\mathrm{d,n})}}{\sum_{k=1}^{\mathrm{K}}\alpha_k+\#\{\mathrm{d},\cdot,\cdot\}^{\neg(\mathrm{d,n})}}$$

此外，从推断形式上还有 message passing 算法类似 MCMC，以及 online 优化算法 online LDA(随机变分推断)。分布式、MPI 算法见参考文献。

主题模型参考文献

Hofmann T. Probabilistic latent semantic analysis[C] Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., 1999: 289-296.

Hofmann T. Unsupervised learning by probabilistic latent semantic analysis[J]. Machine learning, 2001, 42(1-2): 177-196.

Heinrich G. Parameter estimation for text analysis[J]. Technical Report. Web: http://www. arbylon. net/publications/text-est. pdf, 2005.

Blei D M. Probabilistic topic models[J]. Communications of the ACM, 2012, 55(4): 77-84.

Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. the Journal of machine Learning research, 2003, 3: 993-1022.

Lafferty J D, Blei D M. Correlated topic models[C] Advances in neural information processing systems. 2005: 147-154.

Teh Y W, Newman D, Welling M. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation[C] Advances in neural information processing systems. 2006: 1353-1360.

Darling W M. A Theoretical and Practical Implementation Tutorial on Topic Modeling and Gibbs Sampling[C] Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. 2011: 642-647.

Korsos L F, Taddy M. Gibbs Sampling for n-Gram Latent Dirichlet Allocation[R]. 2011. Technical Report Web: http://home.uchicago.edu/~lkorsos/GibbsNGram LDA.pdf

Zeng J, Cheung W K, Liu J. Learning topic models by belief propagation[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2013, 35(5): 1121-1134.

Blei D M, McAuliffe J D. Supervised topic models[J]. arXiv preprint arXiv:1003.0783, 2010.

Hoffman M, Bach F R, Blei D M. Online learning for latent dirichlet allocation[C] advances in neural information processing systems. 2010: 856-864.

Hoffman M，Stochastic Variational Inference

Porteous I, Newman D, Ihler A, et al. Fast collapsed gibbs sampling for latent dirichlet allocation[C] Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2008: 569-577.

赵鑫，李晓明．主题模型在文本挖掘中的应用[R]．北京大学技术报告

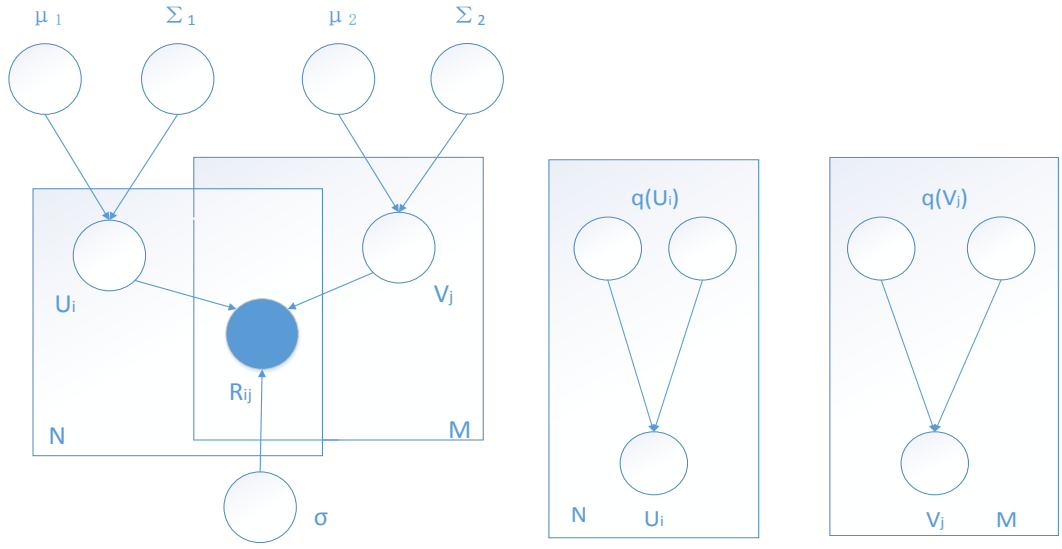
刘知远，孙茂松．基于文档主题结构的关键词抽取方法研究[D]．北京:清华大学, 2011（MPI）

丁轶群 基于概率生成模型的文本主题模型及其应用 浙江大学

Jun Zhu 正则化贝叶斯 清华大学

David Blei 主页

概率矩阵分解算法 PMF 及 VB-EM 算法



概率矩阵分解 PMF(左)变分分布假设(右)的图模型表示

概率矩阵分解(probabilistic matrix factorization)是一个基本的因子分解模型，常见于协同过滤和图像处理中，它假设特征$U_i \sim N(\mu_1, \Sigma_1)$，$V_j \sim N(\mu_2, \Sigma_2)$，其中 i=1…N，j=1…M，$U_i$和$V_j$是 d 维向量服从多元高斯分布，观察值$R_{ij} \sim N(U_i^T V_j, \sigma^2)$，$\delta(i,j)$指示观察值是否存在，记$\boldsymbol{\Theta} = \{\mu_1, \Sigma_1, \mu_2, \Sigma_2, \sigma^2\}$，该模型的联合分布写作：

$$p(\mathbf{R}, \mathbf{U}, \mathbf{V} | \boldsymbol{\Theta}) = \Pi_{i=1}^{N} p(U_i | \mu_1, \Sigma_1) \Pi_{j=1}^{M} p(V_j | \mu_2, \Sigma_2) \Pi_{i=1}^{N} \Pi_{j=1}^{M} p(R_{ij} | U_i^T V_j, \sigma^2)^{\delta(i,j)}$$

如果$\mu_1 = 0, \Sigma_1 = \sigma_1^2 I, \mu_2 = 0, \Sigma_2 = \sigma_2^2 I$，最大化上面的似然函数，等同于最小化下面的目标函数：

$$\min_{U,V} \sum_{i=1}^{N} \sum_{j=1}^{M} \delta(i,j) (R_{ij} - U_i^T V_j)^2 + \lambda_U \sum_{i=1}^{N} U_i^T U_i + \lambda_V \sum_{j=1}^{M} V_j^T V_j$$

其中，$\lambda_U = \frac{\sigma^2}{\sigma_U^2}, \lambda_V = \frac{\sigma^2}{\sigma_V^2}$，也就是说此时的 PMF 模型等价于正则化矩阵分解。若采用图模型的求解策略，外层参数作为未知参数，将所有的U_i和V_j看作隐变量，假设完全分解的变分分布$q(U_i)$与 $q(V_j)$。

$$\log \int_{(\mathbf{U},\mathbf{V})} p(\mathbf{R}, \mathbf{U}, \mathbf{V} | \boldsymbol{\Theta}) \, d(\mathbf{U}, \mathbf{V}) \geq LB$$

$$= E_{q(\mathbf{U},\mathbf{V})} \log p(\mathbf{R}, \mathbf{U}, \mathbf{V} | \boldsymbol{\Theta}) + H(q(\mathbf{U}, \mathbf{V}))$$

概率矩阵分解的 VB-EM 算法推导

其中，$H\left(q(\mathbf{U},\mathbf{V})\right)=\sum_{\mathrm{i}=1}^{\mathrm{N}}H(q(\mathrm{U_i}))+\sum_{\mathrm{j}=1}^{\mathrm{M}}H(q(\mathrm{V_j}))$，由于$\log \mathrm{p}(\mathbf{R},\mathbf{U},\mathbf{V}|\mathbf{\Theta})=$

$\sum_{\mathrm{i}=1}^{\mathrm{N}}\log \mathrm{p}(\mathrm{U_i}|\mu_1,\Sigma_1)+\sum_{\mathrm{j}=1}^{\mathrm{M}}\log \mathrm{p}\left(\mathrm{V_j}\middle|\mu_2,\Sigma_2\right)+\sum_{\mathrm{i}=1}^{\mathrm{N}}\sum_{\mathrm{j}=1}^{\mathrm{M}}\delta(\mathrm{i},\mathrm{j})\log \mathrm{p}\left(\mathrm{R_{ij}}\middle|\mathrm{U_i}^{T}\mathrm{V_j},\sigma^2\right)$

LB 中包含$\mathrm{U_i}$的项记为$LB_{\mathrm{U_i}}$，

$$LB_{\mathrm{U_i}}=E_{q(\mathrm{U_i})}\log \mathrm{p}(\mathrm{U_i}|\mu_1,\Sigma_1)+\sum_{\mathrm{j}=1}^{\mathrm{M}}\delta(\mathrm{i},\mathrm{j})E_{q(\mathrm{U_i})}\log \mathrm{p}\left(\mathrm{R_{ij}}\middle|\mathrm{U_i}^{T}\mathrm{V_j},\sigma^2\right)\quad+H\left(q(\mathrm{U_i})\right)$$

$$E_{q(\mathrm{U_i})}\log \mathrm{p}(\mathrm{U_i}|\mu_1,\Sigma_1)=\frac{1}{2}\log\frac{1}{(2\pi)^{\mathrm{d}}|\Sigma_1|}+\left(-\frac{1}{2}E_{q(\mathrm{U_i})}(\mathrm{U_i}-\mu_1)^{T}\Sigma_1^{-1}(\mathrm{U_i}-\mu_1)\right)$$

$$E_{q(\mathrm{U_i})q(\mathrm{V_j})}\log \mathrm{p}\left(\mathrm{R_{ij}}\middle|\mathrm{U_i}^{T}\mathrm{V_j},\sigma^2\right)=\frac{1}{2}\log\frac{1}{2\pi\sigma^2}-\frac{1}{2\sigma^2}E_{q(\mathrm{U_i})q(\mathrm{V_j})}\left(\mathrm{R_{ij}}-\mathrm{U_i}^{T}\mathrm{V_j}\right)^2$$

若令$q(\mathrm{U_i})=N(U_i|\Phi^{\mathrm{i}},\Sigma^{\mathrm{i}})$，$q\left(\mathrm{V_j}\right)=N(\mathrm{V_j}|\Phi^{\sim\mathrm{j}},\Sigma^{\sim\mathrm{j}})$[1]

$$H\left(q(\mathrm{U_i})\right)=-\frac{1}{2}\log\frac{1}{(2\pi)^{\mathrm{d}}|\Sigma^{\mathrm{i}}|}-\left(-\frac{1}{2}E_{q(\mathrm{U_i})}\left(\mathrm{U_i}-\Phi^{\mathrm{i}}\right)^{T}\Sigma^{\mathrm{i}^{-1}}\left(\mathrm{U_i}-\Phi^{\mathrm{i}}\right)\right)$$

$$E_{q(\mathrm{U_i})}(\mathrm{U_i}-\mu_1)^{T}\Sigma_1^{-1}(\mathrm{U_i}-\mu_1)$$

$$=\mathrm{tr}\left(\Sigma_1^{-1}(E_{q(\mathrm{U_i})}\mathrm{U_i}\mathrm{U_i}^{T}-E_{q(\mathrm{U_i})}\mathrm{U_i}\mu_1^{T}-E_{q(\mathrm{U_i})}\mu_1\mathrm{U_i}^{T}+\mu_1\mu_1^{T})\right)$$

$$=\mathrm{tr}\left(\Sigma_1^{-1}\Sigma^{\mathrm{i}}+\Sigma_1^{-1}\Phi^{\mathrm{i}}\Phi^{\mathrm{i}^{T}}-2\Sigma_1^{-1}\Phi^{\mathrm{i}}\mu_1^{T}+\Sigma_1^{-1}\mu_1\mu_1^{T}\right)$$

类似地，$E_{q(\mathrm{U_i})}(\mathrm{U_i}-\Phi)^{T}\Sigma^{\mathrm{i}^{-1}}(\mathrm{U_i}-\Phi)=\mathrm{d}$

$$E_{q(\mathrm{U_i})q(\mathrm{V_j})}\left(\mathrm{R_{ij}}-\mathrm{U_i}^{T}\mathrm{V_j}\right)^2$$

$$=\mathrm{R_{ij}^2}-2\mathrm{R_{ij}}\Phi^{\mathrm{i}^{\mathrm{T}}}\Phi^{\sim\mathrm{j}}+\mathrm{tr}(E_{q(\mathrm{U_i})}(\mathrm{U_i}\mathrm{U_i}^{T})E_{q(\mathrm{V_j})}(\mathrm{V_j}\mathrm{V_j}^{T}))$$

$$=\mathrm{R_{ij}^2}-2\mathrm{R_{ij}}\Phi^{\mathrm{i}^{\mathrm{T}}}\Phi^{\sim\mathrm{j}}+\mathrm{tr}(\Sigma^{\mathrm{i}}+\Phi^{\mathrm{i}}\Phi^{\mathrm{i}^{T}})(\Sigma^{\sim\mathrm{j}}+\Phi^{\sim\mathrm{j}}\Phi^{\sim\mathrm{j}^{T}})$$

$$LB_{\mathrm{U_i}}\left(\Phi^{\mathrm{i}},\Sigma^{\mathrm{i}}\right)=\frac{1}{2}\mathrm{tr}\left(-\Sigma_1^{-1}\Sigma^{\mathrm{i}}-\Sigma_1^{-1}\Phi^{\mathrm{i}}\Phi^{\mathrm{i}^{T}}+2\Sigma_1^{-1}\Phi^{\mathrm{i}}\mu_1^{T}\right)-\frac{1}{2}\log\frac{1}{(2\pi)^{\mathrm{d}}|\Sigma^{\mathrm{i}}|}$$

$$+\sum_{\mathrm{j}=1}^{\mathrm{M}}\delta(\mathrm{i},\mathrm{j})\left(\frac{1}{\sigma^2}\Phi^{\mathrm{i}^{\mathrm{T}}}\Phi^{\sim\mathrm{j}}\mathrm{R_{ij}}-\frac{1}{2\sigma^2}\mathrm{tr}(\Sigma^{\mathrm{i}}+\Phi^{\mathrm{i}}\Phi^{\mathrm{i}^{T}})(\Sigma^{\sim\mathrm{j}}+\Phi^{\sim\mathrm{j}}\Phi^{\sim\mathrm{j}^{T}})\right)$$

对上式中的Φ^{i}求导并令导数为 0，得到：

$$\Phi^{\mathrm{i}}=\left(\frac{1}{\sigma^2}\sum_{\mathrm{j}=1}^{\mathrm{M}}\delta(\mathrm{i},\mathrm{j})\left(\Sigma^{\sim\mathrm{j}}+\Phi^{\sim\mathrm{j}}\Phi^{\sim\mathrm{j}^{T}}\right)+\Sigma_1^{-1}\right)^{-1}\left(\Sigma_1^{-1}\mu_1+\frac{1}{\sigma^2}\sum_{\mathrm{j}=1}^{\mathrm{M}}\Phi^{\sim\mathrm{j}}\mathrm{R_{ij}}\right)$$

对协方差Σ^{i}求导，并令导数为 0 得到：

$$\Sigma^{\mathrm{i}}=\left(\frac{1}{\sigma^2}\sum_{\mathrm{j}=1}^{\mathrm{M}}\delta(\mathrm{i},\mathrm{j})\left(\Sigma^{\sim\mathrm{j}}+\Phi^{\sim\mathrm{j}}\Phi^{\sim\mathrm{j}^{T}}\right)+\Sigma_1^{-1}\right)^{-1}$$

这样我们得到 E 步，初始化所有的$\Phi^{\mathrm{i}},\Sigma^{\mathrm{i}},\Phi^{\sim\mathrm{j}},\Sigma^{\sim\mathrm{j}}$以及$\mu_1,\Sigma_1,\mu_2,\Sigma_2,\sigma^2$。

[1] 协方差和求和符号在这里都使用∑这个符号，请读者注意分辨。

E-step： 已知R_{ij}，固定$\boldsymbol{\Theta}=\{\mu_1,\Sigma_1,\mu_2,\Sigma_2,\sigma^2\}$，先更新所有的$\Sigma^{i}$与 Φ^{i}(i=1…N)，再更新所有的$\Sigma^{\sim j}$与 $\Phi^{\sim j}$(j=1…M)。

$$\Sigma^{i}=\left(\frac{1}{\sigma^2}\sum_{j=1}^{M}\delta(i,j)\left(\Sigma^{\sim j}+\Phi^{\sim j}\Phi^{\sim j^T}\right)+\Sigma_1^{-1}\right)^{-1}$$

$$\Phi^{i}=\Sigma^{i}\left(\Sigma_1^{-1}\mu_1+\frac{1}{\sigma^2}\sum_{j=1}^{M}\Phi^{\sim j}R_{ij}\right)$$

$$\Sigma^{\sim j}=\left(\frac{1}{\sigma^2}\sum_{i=1}^{N}\delta(i,j)\left(\Sigma^{i}+\Phi^{i}\Phi^{i^T}\right)+\Sigma_2^{-1}\right)^{-1}$$

$$\Phi^{\sim j}=\Sigma^{\sim j}\left(\Sigma_2^{-1}\mu_2+\frac{1}{\sigma^2}\sum_{i=1}^{M}\Phi^{i}R_{ij}\right)$$

M-step： 固定所有的Σ^{i}与 Φ^{i}(i=1…N)，$\Sigma^{\sim j}$与 $\Phi^{\sim j}$(j=1…M)。

先考虑σ^2，LB中它只包含在$\sum_{i=1}^{N}\sum_{j=1}^{M}\delta(i,j)E_{q(U_i)q(V_j)}\log p\left(R_{ij}\middle|U_i^TV_j,\sigma^2\right)$

对其中的σ^2求导，并置导数为 0 得：

$\sigma^2=\frac{1}{\sum_{i=1}^{N}\sum_{j=1}^{M}\delta(i,j)}\sum_{i=1}^{N}\sum_{j=1}^{M}\left\{R_{ij}^2-2R_{ij}\Phi^{i^T}\Phi^{\sim j}+\mathrm{tr}\left(\Sigma^{i}+\Phi^{i}\Phi^{i^T}\right)\left(\Sigma^{\sim j}+\Phi^{\sim j}\Phi^{\sim j^T}\right)\right\}$

考虑μ_1,Σ_1，LB 中它只包含在$\sum_{i=1}^{N}E_{q(U_i)}\log p(U_i|\mu_1,\Sigma_1)$

$$\frac{N}{2}log\frac{1}{(2\pi)^d|\Sigma_1|}-\frac{1}{2}\mathrm{tr}\left(\Sigma_1^{-1}\sum_{i=1}^{N}\Sigma^{i}+\Sigma_1^{-1}\sum_{i=1}^{N}\Phi^{i}\Phi^{i^T}-2\Sigma_1^{-1}\sum_{i=1}^{N}\Phi^{i}\mu_1^T+N\Sigma_1^{-1}\mu_1\mu_1^T\right)$$

对其中的μ_1求导，并置为 0 得：$\mu_1=\frac{1}{N}\sum_{i=1}^{N}\Phi^{i}$

对其中的Σ_1^{-1}求导，并置为 0 得：$\Sigma_1=\frac{1}{N}\Sigma_{i=1}^{N}(\Sigma^{i}+\left(\Phi^{i}-\mu_1\right)\left(\Phi^{i}-\mu_1\right)^T)$

同样我们有：

$$\mu_2=\frac{1}{M}\sum_{j=1}^{M}\Phi^{\sim j}$$

$$\Sigma_2=\frac{1}{M}\Sigma_{j=1}^{M}(\Sigma^{\sim j}+\left(\Phi^{\sim j}-\mu_2\right)\left(\Phi^{\sim j}-\mu_2\right)^T)$$

上面考虑的 PMF 模型以及变分分布中的协方差矩阵Σ_1,Σ_2和所有的$\Sigma^{i},\Sigma^{\sim j}$均为一般实矩阵，若考虑矩阵为对角矩阵，可在相应的更新中只取对角元素（对角近似）。但如果对角元素均相同，如$\Sigma_1=\sigma_1^2I,\Sigma_2=\sigma_2^2I$，分情况讨论如下：

①$\Sigma_1=\sigma_1^2I$，$\Sigma_2=\sigma_2^2I$，而$\Sigma^{i},\Sigma^{\sim j}$为非对角，此时变分近似分布更复杂。

E-step： Φ^{i}与$\Phi^{\sim j}$，Σ^{i}与$\Sigma^{\sim j}$的更新不变。

M-step： μ_1,μ_2,σ^2的更新不变。

$$\sigma_1^2 = \frac{1}{Nd}\Sigma_{i=1}^{N}\left(\left(\Phi^i - \mu_1\right)^T\left(\Phi^i - \mu_1\right) + \mathrm{tr}\Sigma^i\right)$$

$$\sigma_2^2 = \frac{1}{Md}\Sigma_{j=1}^{M}\left(\left(\Phi^{\sim j} - \mu_2\right)^T\left(\Phi^{\sim j} - \mu_2\right) + \mathrm{tr}\Sigma^{\sim j}\right)$$

②$\Sigma^i = diag(\gamma_1^i \ldots \gamma_d^i), \Sigma^{\sim j} = diag(\gamma_1^{\sim j} \ldots \gamma_d^{\sim j})$为对角，而$\Sigma_1, \Sigma_2$为非对角，此时变分近似分布更简单。

E-step： Φ^i与$\Phi^{\sim j}$的更新不变

$$\gamma_k^{i\,2} = \left(\frac{1}{\sigma^2}\sum_{j=1}^{M}\delta(i,j)\left(\gamma_k^{\sim j^2} + \Phi_k^{\sim j^2}\right) + \Sigma_{1,kk}^{-1}\right)^{-1}$$

$$\gamma_k^{\sim j^2} = \left(\frac{1}{\sigma^2}\sum_{i=1}^{N}\delta(i,j)\left(\gamma_k^{i\,2} + \Phi_k^{i\,2}\right) + \Sigma_{2,kk}^{-1}\right)^{-1}$$

M-step： $\sigma^2, \mu_1, \mu_2, \Sigma_1, \Sigma_2$ 更新不变。

输出：U，V。

输入：指示矩阵$Y_{ij} = \delta(i,j)$，观察值矩阵 R，最大迭代次数 S，初始化$\mu_1, \Sigma_1, \mu_2, \Sigma_2, \sigma$以及$\Sigma^i$与 Φ^i(i=1…N)，$\Sigma^{\sim j}$与 $\Phi^{\sim j}$(j=1…M)。

For t =1…S：

For i from 1 to N, j from 1 to M：

update $\Sigma^i, \Phi^i, \Sigma^{\sim j}, \Phi^{\sim j}$

update $\mu_1, \Sigma_1, \mu_2, \Sigma_2, \sigma$。

$\Delta^t = \|Y \odot (R - U^T V)\|_2^2$

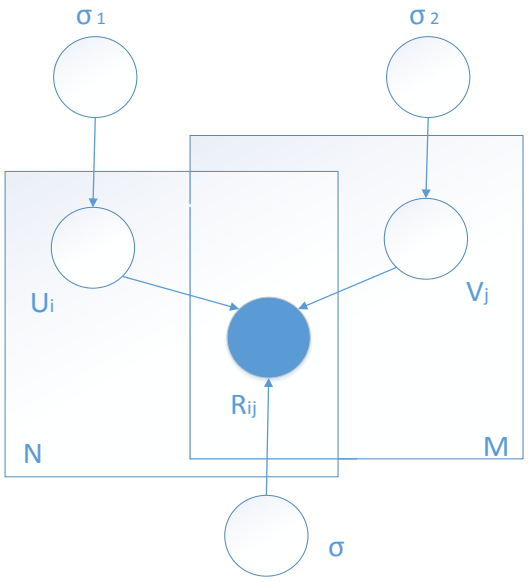
If $\sqrt{\Delta^t/\Delta^{t-1}} < \epsilon$： break

For i from 1 to N, j from 1 to M：

$U_i = \Phi^i, V_j = \Phi^{\sim j}$

吴金龙. Netflix Prize 中的协同过滤算法[D]. 北京: 北京大学, 2010.

正则矩阵分解算法 RMF 及 SGD&ALS 算法



PMF 简化情形下的图模型表示

上一小节中提到 PMF 等同于正则化矩阵分解问题的情形。这一节专门探讨正则化矩阵分解解决不完全观察矩阵近似的问题。我们定义矩阵 Y，其中的每个元素$Y_{ij} = \delta(i,j)$指示R_{ij}是否被观察，定义运算符⊙为 Hadamard 积满足$A_{n*m} \odot B_{n*m} = [a_{ij}b_{ij}]_{n*m}$，定义矩阵的$l_1$范数满足$\|A\|_1 = \sum_{ij}|a_{ij}|$，$l_2$范数满足$\|A\|_2 = (\sum_{ij} a_{ij}^2)^{1/2}$。定义$U = [U_1 \dots U_N]$，$V = [V_1 \dots V_M]$。假设特征$U_i \sim N(0, \sigma_1^2 I)$，$V_j \sim N(0, \sigma_2^2 I)$，其中 i=1…N，j=1…M，$U_i$和$V_j$是 d 维向量服从多元高斯分布。若观察值$R_{ij} \sim N(U_i^T V_j, \sigma^2)$，解决这个问题的 PMF 变分近似算法(VB-EM 算法)在附录最后讨论的情形①中已经给出，变分近似分布中$\Sigma^i, \Sigma^{\sim j}$为非对角，而所有的$\Phi^i$与$\Phi^{\sim j}$与$\mu_1, \mu_2$均设置成 0 即可。在给定外层超参数情况下，似然函数最大化将等同于下面的l_2正则最小化问题：

$\min_{U,V} \|Y \odot (R - U^T V)\|_2^2 + \lambda_U \|U\|_2^2 + \lambda_V \|V\|_2^2$，其中$\lambda_U = \frac{\sigma^2}{\sigma_U^2}, \lambda_V = \frac{\sigma^2}{\sigma_V^2}$。

给定参数λ_U和λ_V的值，利用随机梯度下降(stochastic gradient decent)或者交替最小二乘(alternative least squares)算法优化上面的目标函数，可得到所有的U_i和V_j。这个问题通常又称作“最大边际矩阵分解”(max-margin matrix factorization)M^3F，它解决大规模不完全矩阵的低秩近似问题。下面给出 SGD 和 ALS 求解l_2正则(Regularized MF)的算法流程。输入指示矩阵Y，观察值矩阵 R，正则化参数$\lambda =$

$\lambda_U = \lambda_V$，收敛率ϵ，初始化 U，V。SGD 需要学习率η。ALS 是 Ridge 回归问题。

输出：U，V。

输入：指示矩阵Y，观察值矩阵 R，正则化参数$\lambda = \lambda_U = \lambda_V$，学习率$\eta$，收敛率$\epsilon$，最大迭代次数 S，初始化 U，V。

For t =1…S：

For each (i, j) with $Y_{ij} \neq 0$：

$\Delta_{ij} = R_{ij} - U_i^T V_j$

$U_i = U_i + \eta(\Delta_{ij} V_j - \lambda U_i)$

$V_j = V_j + \eta(\Delta_{ij} U_i - \lambda V_j)$

$\Delta^t = \|Y \odot (R - U^T V)\|_2^2$

If $\sqrt{\Delta^t / \Delta^{t-1}} < \epsilon$： break

输出：U，V。

输入：指示矩阵Y，观察值矩阵 R，正则化参数$\lambda = \lambda_U = \lambda_V$，收敛率$\epsilon$，最大迭代次数 S，初始化 U，V。

For t =1…S：

For each i from 1 to N:

$U_i = \left(\sum_j Y_{ij} V_j V_j^T + \lambda I\right)^{-1} \sum_j Y_{ij} R_{ij} V_j$

For each i from 1 to M:

$V_j = \left(\sum_i Y_{ij} U_i U_i^T + \lambda I\right)^{-1} \sum_i Y_{ij} R_{ij} U_i$

$\Delta^t = \|Y \odot (R - U^T V)\|_2^2$

If $\sqrt{\Delta^t / \Delta^{t-1}} < \epsilon$： break

类似地，若假设$R_{ij} \sim L(U_i^T V_j, \sigma)$，L 代表 laplace 分布$L(x|\mu, \sigma) = \frac{1}{2\sigma} \exp(-\frac{|x-\mu|}{\sigma})$则在给定外层超参数情况下，似然函数最大化等同于下面的$\boldsymbol{l_1}$正则最小化问题[2]。

[2]我们通常又把它称作“鲁棒矩阵分解”或“稀疏低秩矩阵分解”问题。由于篇幅限制，求解这个问题的算法可详见鲁棒矩阵分解相关资料。

低秩矩阵分解参考文献：
Srebro N, Rennie J, Jaakkola T S. Maximum margin matrix factorization[C] Advances in neural information processing systems. 2004: 1329-1336.

Mazumder R, Hastie T, Tibshirani R. Spectral regularization algorithms for learning large incomplete matrices[J]. The Journal of Machine Learning Research, 2010, 99: 2287-2322.

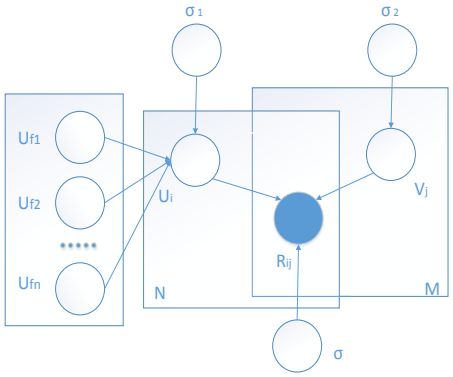
Mnih A, Salakhutdinov R. Probabilistic matrix factorization[C] Advances in neural information processing systems. 2007: 1257-1264.

Wang N, Yao T, Wang J, et al. A probabilistic approach to robust matrix factorization[M] Computer Vision–ECCV 2012. Springer Berlin Heidelberg, 2012: 126-139.

Wang, Shusen, and Zhihua Zhang. "Colorization by Matrix Completion." AAAI. 2012.

Mitra K, Sheorey S, Chellappa R. Large-scale matrix factorization with missing data under additional constraints[C] Advances in Neural Information Processing Systems. 2010: 1651-1659.

基于图约束的矩阵分解 SMF



上图所示，即为社交关系矩阵分解 Social-MF。值得注意的是，Social-MF 实际上有很多种版本，这里引述的是“基于相似度矩阵分解”，与 Purushotham S 所利用的模型是不同的。在 Social-MF 模型中，定义矩阵 S 为用户相似度矩阵(推荐系统中 side information 或者用户社交关系 构建相似度 person cosine 等)，即用户因子U_i与U_j之间的相似度为S_{ij}，Social-MF 解决下面的l_2正则最小化问题：

$$\min_{U,V} \|Y \odot (R - U^T V)\|_2^2 + \lambda_U \|U\|_2^2 + \lambda_V \|V\|_2^2 + \sum_{i>j} S_{ij} \left\|U_i - U_j\right\|_2^2$$

其中$\lambda_U = \frac{\sigma^2}{\sigma_1^2}, \lambda_V = \frac{\sigma^2}{\sigma_2^2}$，给定参数$\lambda_U$和$\lambda_V$的值。可见 Social-MF 使得相似程度更大的用户因子之间的欧式距离更小，这将有利于将能够定义用户距离的辅助信息融入到协同过滤矩阵分解中。上式等价于最小化下面的目标函数：

$$\min_{U,V} \|Y \odot (R - U^T V)\|_2^2 + \lambda_U U^T U + \lambda_V V^T V + \mathrm{tr}(ULU^T)$$

其中$L = D - S$，D是对角矩阵满足$D_{ii} = \sum_j S_{ij}$，称L为 Laplacian 矩阵。因而若令$\lambda_U = \lambda_V = \alpha$，添加控制参数β，重写上面的目标函数为：

$$f = \frac{1}{2}\|Y \odot (R - U^T V)\|_2^2 + \frac{1}{2}\alpha \mathrm{tr}(VV^T) + \frac{1}{2}\mathrm{tr}[U(\alpha I + \beta L)U^T]$$

上式中α，β，L均已知，对 f 中的V_j求导得到：

$$\frac{\partial f}{\partial V_j} = \left(\alpha I + \sum_{i=1}^{N} Y_{ij} U_i U_i^T\right) V_j - \sum_{i=1}^{N} Y_{ij} R_{ij} U_i$$

设置梯度下降的步长η，对所有的项目因子V_j完成更新，即$V_j = V_j - \eta \frac{\partial f}{\partial V_j}$。

对 f 的第一项$g = \frac{1}{2}\|Y \odot (R - U^T V)\|_2^2$中的$U_{di}$求导得到：

$$\frac{\partial g}{\partial U_{id}} = \left(\sum_{j=1}^{M} Y_{ij} V_{jd}^2\right) U_{id} - \sum_{j=1}^{M} Y_{ij} V_{jd} (R_{ij} - U_i^T V_j + U_{id} V_{jd})$$

则有$\frac{\partial g}{\partial U_i} = W U_{*d} - x$，其中 W 是 N 维对角矩阵有$W_{ii} = \sum_{j=1}^{M} Y_{ij} V_{jd}^2$，$x$是 N 维向量$x_i = \sum_{j=1}^{M} Y_{ij} V_{jd} (R_{ij} - U_i^T V_j + U_{id} V_{jd})$。所以对 f 中的$U_i$求导得到：

$$\frac{\partial f}{\partial U_{*d}} = (W + \alpha I + \beta L) U_{*d} - x$$

设置梯度下降的步长η，对所有的维度 d 完成更新，即$U_{*d} = U_{*d} - \eta \frac{\partial f}{\partial U_{*d}}$。

注：如果相似度不是事先给定的，这种在 U 的维度之间增加约束的方法需要利用图模型推断来学习，则参考下面论文 kernelized probabilistic matrix factorization

http://people.ee.duke.edu/~lcarin/kpmf_sdm_final.pdf

Li W J, Yeung D Y. TagiCoFi: tag informed collaborative filtering[C] Proceedings of the third ACM conference on Recommender systems. ACM, 2009: 69-

贝叶斯概率矩阵分解 BPMF 及 MCMC 算法

在这一小节将要介绍的贝叶斯概率矩阵分解 Bayesian PMF 利用 Gibbs 采样求解模型。回顾 LDA 模型的 Gibbs 采样所利用的完全条件分布，Dir 与 Mult 的共轭关系使得对参数θ的采样可以在 Dir 分布下完成，这将简化采样的过程。

PMF 中，U_i的条件分布[3]同比例于$p(U_i|\mu_1,\Sigma_1)\Pi_{j=1}^{M}p\left(R_{ij}\middle|U_i^T V_j,\sigma^2\right)^{\delta(i,j)}$，这个分布仍然是多元高斯分布，即有：

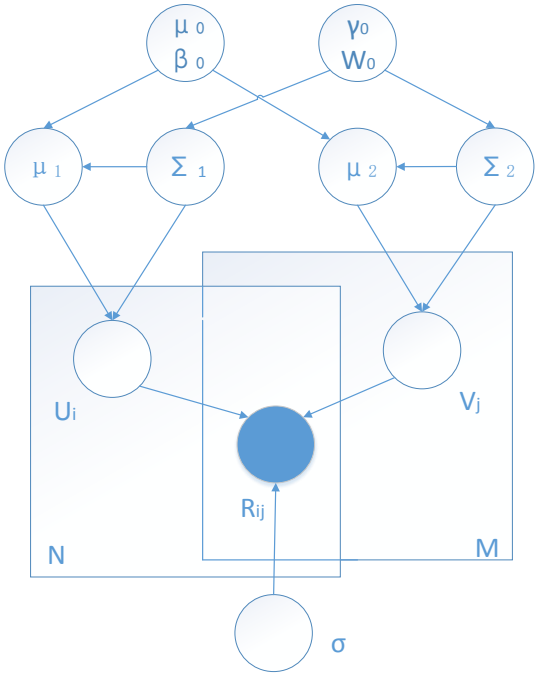
$\widehat{p(U_\iota)} \propto p(U_i|\mu_1,\Sigma_1)\Pi_{j=1}^{M}p\left(R_{ij}\middle|U_i^T V_j,\sigma^2\right)^{\delta(i,j)} \propto N(U_i|\mu_i^*,\Sigma_i^*)$，其中：

$$\Sigma_i^* = \left(\Sigma_1^{-1} + \frac{1}{\sigma^2}\sum_{j=1}^{M}\delta(i,j)V_j V_j^T\right)^{-1}$$

$$\mu_i^* = \Sigma_i^*\left(\frac{1}{\sigma^2}\sum_{j=1}^{M}\delta(i,j)V_j R_{ij} + \Sigma_1^{-1}\mu_1\right)$$

因此U_i和V_j的采样可在多元高斯分布下完成。如果在参数$\boldsymbol{\Theta} = \{\mu_1,\Sigma_1,\mu_2,\Sigma_2,\sigma^2\}$给定情况下，利用$U_i$和$V_j$的条件分布，通过采样(构造马尔科夫链)得到稳定分布时所有U_i和V_j的值，这样事实上得到了 MAP (maximum a posterior) 估计，把这一步看作 E 步，而在所有U_i和V_j的值给定情况下对外层参数$\boldsymbol{\Theta}$求导做更新看作 M 步，我们把这样的方法叫作“MCMC-EM”算法。下面给出的模型基于贝叶斯的求解策略，为外层参数增加超参数先验，这样外层参数可通过采样完成更新，这个模型叫作“贝叶斯概率矩阵分解模型 BPMF”。

[3] 本文中“后验分布”与“条件分布”或者“联合分布”与“似然函数”通常等于或同比于相同的目标函数，只不过在对待隐变量、已知变量、参数时的说法不同。



贝叶斯概率矩阵分解的图模型表示

该模型的图表示，参数$\mu_0, \beta_0, \gamma_0, W_0$是最外层的超参数，此时$\mu_1, \Sigma_1$的先验分布叫作“高斯-逆威沙特分布”(Gaussian inverse-Wishart)，具体地：

$$p(\mu_1|\mu_0, \beta_0, \Sigma_1) = N(\mu_1|\mu_0, \beta_0^{-1}\Sigma_1)$$，其中β_0是实值变量

$$p(\Sigma_1|\gamma_0, W_0) = W(\Sigma_1^{-1}|W_0, \gamma_0)$$

其中 W 代表威沙特分布$W(\Sigma_1^{-1}|W_0, \gamma_0) \propto |\Sigma_1^{-1}|^{\frac{\gamma_0-d-1}{2}} \exp(-\frac{1}{2}tr(W_0^{-1}\Sigma_1^{-1}))$，其中$\gamma_0$是自由度，$W_0$是 d 维 scale 矩阵。

$$p(\mu_1, \Sigma_1|\mu_0, \beta_0, \gamma_0, W_0) \propto N(\mu_1|\mu_0, \beta_0^{-1}\Sigma_1)W(\Sigma_1^{-1}|W_0, \gamma_0) \propto$$

$$\sqrt{|\beta_0|}|\Sigma_1^{-1}|^{\frac{\gamma_0-d}{2}} \exp\left(-\frac{1}{2}tr(W_0^{-1}\Sigma_1^{-1}+\beta_0(\mu_1-\mu_0)(\mu_1-\mu_0)^T\Sigma_1^{-1})\right)$$

此时，联合分布写为：

$$p(\mathbf{R}, \mathbf{U}, \mathbf{V}|\mathbf{\Theta}, \mu_0, \beta_0, \gamma_0, W_0)$$

$$= p(\mu_1, \Sigma_1|\mu_0, \beta_0, \gamma_0, W_0)$$

$$\Pi_{i=1}^{N} p(U_i|\mu_1, \Sigma_1)\Pi_{j=1}^{M} p(V_j|\mu_2, \Sigma_2)\Pi_{i=1}^{N}\Pi_{j=1}^{M} p(R_{ij}|U_i^T V_j, \sigma^2)^{\delta(i,j)}$$

Salakhutdinov R, Mnih A. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo[C] Proceedings of the 25th international conference on Machine learning. ACM, 2008: 880-887

贝叶斯概率矩阵分解 BPMF 的 MCMC 算法推导

μ_1,Σ_1的后验分布同比于其中包含μ_1,Σ_1的部分，即

$$p(\mu_1,\Sigma_1|\mu_0,\beta_0,\gamma_0,W_0)\Pi_{i=1}^{N}p(U_i|\mu_1,\Sigma_1)\propto$$

$$\sqrt{|\beta_0|}|\Sigma_1^{-1}|^{\frac{\gamma_0-d+N}{2}}\exp\left(-\frac{1}{2}\mathrm{tr}\left(W_0^{-1}\Sigma_1^{-1}+\beta_0(\mu_1-\mu_0)(\mu_1-\mu_0)^T\Sigma_1^{-1}\right.\right.$$

$$\left.\left.+\sum_{i=1}^{N}(U_i-\mu_1)(U_i-\mu_1)^T\Sigma_1^{-1}\right)\right)$$

我们考虑其中的$\beta_0(\mu_1-\mu_0)(\mu_1-\mu_0)^T+\sum_{i=1}^{N}(U_i-\mu_1)(U_i-\mu_1)^T$，它等于下式：

$$\sum_{i=1}^{N}\left(U_i-\frac{\sum_{j=1}^{N}U_j}{N}\right)\left(U_i-\frac{\sum_{j=1}^{N}U_j}{N}\right)^T+\frac{N\beta_0}{N+\beta_0}\left(\mu_0-\frac{\sum_{j=1}^{N}U_j}{N}\right)\left(\mu_0-\frac{\sum_{j=1}^{N}U_j}{N}\right)^T+(N$$

$$+\beta_0)\left(\mu_1-\frac{\beta_0\mu_0+\sum_{i=1}^{N}U_i}{N+\beta_0}\right)\left(\mu_1-\frac{\beta_0\mu_0+\sum_{i=1}^{N}U_i}{N+\beta_0}\right)^T$$

可见μ_1,Σ_1的后验分布仍然是“高斯-逆威沙特分布”，记其中的变量：

$$U_{aver}=\frac{\sum_{j=1}^{N}U_j}{N}$$

$$S_{aver}=\frac{1}{N}\sum_{i=1}^{N}(U_i-U_{aver})(U_i-U_{aver})^T$$

后验参数如下：

$$\beta_0^*=N+\beta_0,\qquad \gamma_0^*=\gamma_0+N,\qquad \mu_0^*=\frac{\beta_0\mu_0+NU_{aver}}{N+\beta_0}$$

$$[W_0^*]^{-1}=W_0^{-1}+\hat{N}S_{aver}+\frac{N\beta_0}{N+\beta_0}(\mu_0-U_{aver})(\mu_0-U_{aver})^T$$

$$p(\widehat{\mu_1,\Sigma_1})\propto N(\mu_1|\mu_0^*,{\beta_0^*}^{-1}\Sigma_1)W(\Sigma_1^{-1}|W_0^*,\gamma_0^*)$$

对μ_2,Σ_2的后验分布推导则与此类似。而对U_i和V_j的采样概率分布所服从的参数则与“MCMC-EM”算法中的 E 步相同。

输出：U，V。

输入：指示矩阵Y，观察值矩阵 R，收敛率ϵ，最大迭代次数 S，固定值σ，初始化$\mu_0, \beta_0, \gamma_0, W_0, U, V$。

For t =1…S：

compute $\beta_0^*, \gamma_0^*, \mu_0^*, W_0^*$ with U

sample $\Sigma_1^{-1} \sim \text{Wishart}(W_0^*, \gamma_0^*)$ $\mu_1 \sim N\left(\mu_0^*, {\beta_0^*}^{-1}\Sigma_1\right)$

compute $\beta_0^*, \gamma_0^*, \mu_0^*, W_0^*$ with V

sample $\Sigma_2^{-1} \sim \text{Wishart}(W_0^*, \gamma_0^*)$ $\mu_2 \sim N\left(\mu_0^*, {\beta_0^*}^{-1}\Sigma_2\right)$

For each i from 1 to N：

compute μ_i^*, Σ_i^* sample $U_i \sim N(\mu_i^*, \Sigma_i^*)$

For each i from 1 to M：

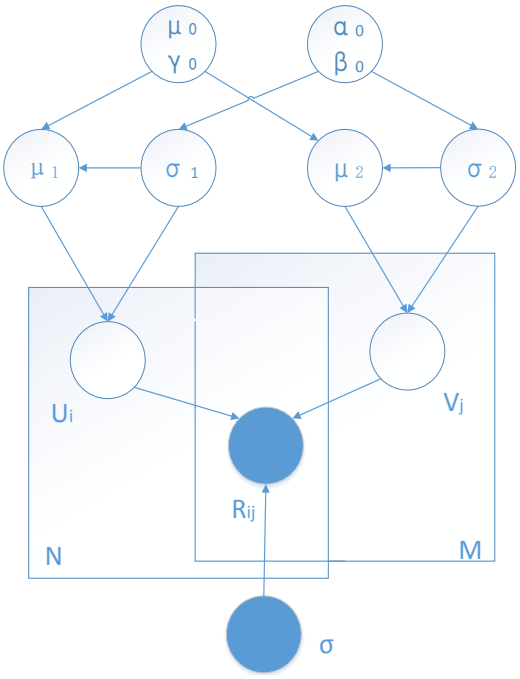
compute $\mu_{\sim j}^*, \Sigma_{\sim j}^*$ sample $V_j \sim N(\mu_{\sim j}^*, \Sigma_{\sim j}^*)$

$\Delta^t = \|Y \odot (R - U^T V)\|_2^2$

If $\sqrt{\Delta^t / \Delta^{t-1}} < \epsilon$： break

贝叶斯概率因子分解器 FM 及 MCMC 算法

BPMF 利用高斯-逆威沙特先验与多变元高斯分布共轭的特性完成采样，高斯分布的共轭分布设置还有很多种，一种简化版本是利用“逆伽马分布”，这就是下面介绍的“因子分解机模型”(factorization machine)[4]。该模型的图表示，



叶斯概率因子分解机 FM 的图模型表示

参数$\mu_0, \gamma_0, \alpha_0, \beta_0$是最外层的超参数，此时$\mu_1, \sigma_1$的先验分布叫作“高斯-逆伽马分布”(gaussian inverse-gamma)，具体地：

$p(\mu_1|\mu_0, \gamma_0, \sigma_1) = N(\mu_1|\mu_0, \gamma_0^{-1}\sigma_1^2 I)$，其中$\gamma_0$是实值变量

$p(\sigma_1^2|\alpha_0, \beta_0) = \Gamma(\sigma_1^{2^{-1}}|\alpha_0, \beta_0)$

其中Γ代表伽马分布$\Gamma(x|\alpha_0, \beta_0) = \frac{x^{\alpha_0-1}}{\Gamma(\alpha_0)\beta_0^{\alpha_0}}\exp(-\frac{x}{\beta_0})$，这样有：

$$p(\mu_1|\mu_0, \gamma_0, \sigma_1)p(\sigma_1|\alpha_0, \beta_0) = N(\mu_1|\mu_0, \gamma_0^{-1}\sigma_1^2 I)\Gamma(\sigma_1^{-1}|\alpha_0, \beta_0) \propto$$

$$(\sigma_1^{-1})^{\alpha_0-1+d}\exp\left(\left(-\frac{1}{\beta_0}-\frac{\gamma_0}{2}(\mu_1-\mu_0)^T(\mu_1-\mu_0)\right)\left(\sigma_1^{2^{-1}}\right)\right)$$

我们考虑“高斯-逆伽马分布”作为先验时μ_1, σ_1^2的后验分布，同比例于下式：

$p(\mu_1|\mu_0, \gamma_0, \sigma_1)p(\sigma_1^2|\alpha_0, \beta_0)\Pi_{i=1}^{N}p(U_i|\mu_1, \sigma_1)$

其中$\gamma_0(\mu_0-\mu_1)^T(\mu_0-\mu_1) + \sum_{i=1}^{N}(U_i-\mu_1)^T(U_i-\mu_1)$可以写为下式：

[4] 这里给出的 FM 模型，参考 Rendle. Factorization Machine 中的参数设置。

$$\sum_{i=1}^{N}\left(U_i-\frac{\sum_{j=1}^{N}U_j}{N}\right)^T\left(U_i-\frac{\sum_{j=1}^{N}U_j}{N}\right)+\frac{N\gamma_0}{N+\gamma_0}\left(\mu_0-\frac{\sum_{j=1}^{N}U_j}{N}\right)^T\left(\mu_0-\frac{\sum_{j=1}^{N}U_j}{N}\right)+(N+\gamma_0)\left(\mu_1-\frac{\gamma_0\mu_0+\sum_{i=1}^{N}U_i}{N+\gamma_0}\right)^T\left(\mu_1-\frac{\gamma_0\mu_0+\sum_{i=1}^{N}U_i}{N+\gamma_0}\right)$$

可见，该后验分布仍然是“高斯-逆伽马分布”

$$p(\mu_1|\mu_0,\gamma_0,\sigma_1)p(\sigma_1^2|\alpha_0,\beta_0)\Pi_{i=1}^{N}p(U_i|\mu_1,\sigma_1)\propto N\left(\mu_1\middle|\mu_0^*,\gamma_0^{*-1}\sigma_1^2I\right)\Gamma\left(\sigma_1^{2-1}\middle|\alpha_0^*,\beta_0^*\right)$$

$$\gamma_0^*=N+\gamma_0,\ \mu_0^*=\frac{\gamma_0\mu_0+\sum_{i=1}^{N}U_i}{N+\gamma_0},\ \alpha_0^*=\alpha_0+dN,\ \beta_0^*=\left(\frac{1}{\beta_0}+\frac{1}{2}\left(\sum_{i=1}^{N}\left(U_i-\frac{\sum_{j=1}^{N}U_j}{N}\right)^T\left(U_i-\frac{\sum_{j=1}^{N}U_j}{N}\right)+\frac{N\gamma_0}{N+\gamma_0}\left(\mu_0-\frac{\sum_{j=1}^{N}U_j}{N}\right)^T\left(\mu_0-\frac{\sum_{j=1}^{N}U_j}{N}\right)\right)\right)^{-1}$$

由于在$\mu_1,\Sigma_1=\sigma_1^2I$给定情况下，$U_i$的条件分布于 BPMF 中一致所以采样分布参考 BPMF，而μ_2,Σ_2则与此类似。

输出：U，V。

输入：指示矩阵Y，观察值矩阵 R，收敛率ϵ，最大迭代次数 S，固定值σ，初始化$\mu_0,\gamma_0,\alpha_0,\beta_0$,U,V。

For t =1…S：

compute $\mu_0^*,\gamma_0^*,\alpha_0^*,\beta_0^*$ with U

sample $\sigma_1^{-1}\sim\Gamma(\alpha_0^*,\beta_0^*)$ $\mu_1\sim N\left(\mu_0^*,\gamma_0^{*-1}\sigma_1^2I\right)$

compute $\mu_0^*,\gamma_0^*,\alpha_0^*,\beta_0^*$ with V

sample $\sigma_2^{-1}\sim\Gamma(\alpha_0^*,\beta_0^*)$ $\mu_2\sim N\left(\mu_0^*,\gamma_0^{*-1}\sigma_2^2I\right)$

For each i from 1 to N：

compute μ_i^*,Σ_i^* sample $U_i\sim N(\mu_i^*,\Sigma_i^*)$

For each i from 1 to M：

compute $\mu_{\sim j}^*,\Sigma_{\sim j}^*$ sample $V_j\sim N(\mu_{\sim j}^*,\Sigma_{\sim j}^*)$

$\Delta^t=\|Y\odot(R-U^TV)\|_2^2$

If $\sqrt{\Delta^t/\Delta^{t-1}}<\epsilon$： break

LIBFM 开源工具参考文献

Rendle S. Factorization machines[C]. Data Mining (ICDM), 2010 IEEE 10th International Conference on. IEEE, 2010: 995-1000.

Rendle S. Factorization machines with libfm[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2012, 3(3): 57-72.

协同过滤问题的参考文献：

Shapira B. Recommender systems handbook[M]. Springer, 2011.

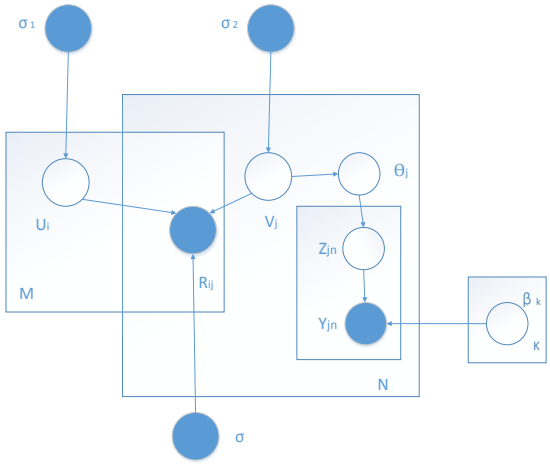
Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender systems[J]. Computer, 2009, 42(8): 30-37.

Su X, Khoshgoftaar T M. A survey of collaborative filtering techniques[J]. Advances in artificial intelligence, 2009, 2009: 4.

Ma H, Yang H, Lyu M R, et al. Sorec: social recommendation using probabilistic matrix factorization[C] Proceedings of the 17th ACM conference on Information and knowledge management. ACM, 2008: 931-940.

Hu L, Cao J, Xu G, et al. Personalized recommendation via cross-domain triadic factorization[C]. Proceedings of the 22nd international conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2013: 595-606.

协同主题回归模型 **Collaborative Topic Regression**



CTR 图模型表示

如图 CTR 模型组合了概率矩阵分解 PMF 和主题模型 LDA，将文本信息融入到因子分解特征中，利用假设$V_j \sim N(\theta_j, \sigma_2^2 I)$，$V_j$服从主题分布参数$\theta_j$为均值，$\sigma_2^2 I$为协方差的多元高斯分布。其中$\sigma, \sigma_1, \sigma_2, \alpha$均为给定的参数，该模型的联合似然函数写为：

$$p(U, V, \mathbf{Z}, \boldsymbol{\theta} | \boldsymbol{\beta}, \mathbf{Y})$$

$$= \Pi_{i=1}^{N} p(U_i | 0, \sigma_1^2 I) \Pi_{j=1}^{M} \{ p(V_j | \theta_j, \sigma_2^2 I) \Pi_{n=1}^{N_j} (\theta_{j Z_{jn}} \beta_{Z_{jn} Y_{jn}}) \} \Pi_{i=1}^{N} \Pi_{j=1}^{M} p(R_{ij} | U_i^T V_j, \sigma^2)^{\delta(i,j)}$$

积分掉其中的$\mathbf{Z}$后，令$\frac{1}{\lambda_u} = \sigma_1^2, \frac{1}{\lambda_v} = \sigma_2^2, \sigma^2 = 1$，则有$\log p(U, V, \boldsymbol{\theta} | \boldsymbol{\beta}, \mathbf{Y})$等于下式：

$$L(\boldsymbol{\theta}, \boldsymbol{\beta}, U, V) = -\frac{\lambda_u}{2} \sum_{i=1}^{N} U_i^T U_i - \frac{\lambda_v}{2} \sum_{j=1}^{M} (V_j - \theta_j)^T (V_j - \theta_j) - \sum_{ij} \frac{Y_{ij}}{2} (R_{ij} - U_i^T V_j)^2 +$$

$$\sum_{j=1}^{M} \sum_{n=1}^{N_j} \log \sum_k \theta_{jk} \beta_{k Y_{jn}}$$

优化这个目标函数，对U_i, V_j求导并置导数为 0，类似于 RMF(ALS)有：

$$U_i = (V \mathrm{diag}(Y_{i*}) V^T + \lambda_U I)^{-1} V \mathrm{diag}(Y_{i*}) R_{i*}^T$$

$$V_j = \left(U \mathrm{diag}(Y_{*j}) U^T + \lambda_V I\right)^{-1} \left(U \mathrm{diag}(Y_{*j}) R_{*j} + \lambda_V \theta_j\right)$$

其中Y_{*j}取 Y 矩阵的第 j 列，Y_{i*}则去第 i 行，diag 将向量拉伸成对角矩阵。对参数$\boldsymbol{\theta}, \boldsymbol{\beta}$的更新需要对目标函数 L 应用 E-M 型近似推断，

$$\log \sum_k \frac{\theta_{jk} \beta_{k Y_{jn}}}{\phi_{jnk}} \phi_{jnk} \geq \sum_k (\phi_{jnk} \log \theta_{jk} \beta_{k Y_{jn}} - \phi_{jnk} \log \phi_{jnk})$$

其中$\sum_k \phi_{jnk} = 1$，对 L 中包含θ_j的项：

$$L(\theta_j, \phi_j) \geq -\frac{\lambda_v}{2} (V_j - \theta_j)^T (V_j - \theta_j) + \sum_{n=1}^{N_j} \sum_k (\phi_{jnk} \log \theta_{jk} \beta_{k Y_{jn}} - \phi_{jnk} \log \phi_{jnk})$$

由于加入θ_j的拉格朗日约束无法得到解析解，计算梯度如下：

$$\frac{\partial L(\theta_j,\phi_j)}{\partial \theta_{jk}} = \frac{\sum_{n=1}^{N_j}\phi_{jnk}}{\theta_{jk}} + \lambda_V(V_{jk} - \theta_{jk})$$

得到$\theta_j = \theta_j + \eta\frac{\partial L(\theta_j,\phi_j)}{\partial \theta_j}$，$\eta$是步长，但$\theta_j$满足单纯形$\sum_k \theta_{jk} = 1, \theta_{jk} > 0$约束，故投影，即求解下式:

$$\min_v \frac{1}{2}\left\|\theta_j - v\right\|_2^2$$，其中 v 满足$\sum_k v_k = 1, v_k > 0$。

Duchi J, Shalev-Shwartz S, Singer Y, et al. Efficient projections onto the L1-ball for learning in high dimensions[C] Proceedings of the 25th ICML. ACM, 2008: 272-279.

然后更新$\theta_j = v$。最后，最优ϕ_{jnk}满足：$\phi_{jnk} \propto \theta_{jk}\beta_{kY_{jn}}$，类似 LDA 中 EM 步骤，最优$\beta_{kv}$满足：$\beta_{kv} \propto \sum_{j,n}\phi_{jnk}\,\delta(Y_{jn} = v)$。

此外，组合它们的方法还有很多，可见参考文献。

参考文献

Wang C, Blei D M. Collaborative topic modeling for recommending scientific articles[C] Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2011: 448-456.

Adams R P, Dahl G E, Murray I. Incorporating side information in probabilistic matrix factorization with gaussian processes[J]. arXiv preprint arXiv:1003.4944, 2010.

Agarwal D, Chen B C. fLDA: matrix factorization through latent dirichlet allocation[C] Proceedings of the third ACM international conference on Web search and data mining. ACM, 2010: 91-100.

Purushotham S, Liu Y, Kuo C C J. Collaborative Topic Regression with Social Matrix Factorization for Recommendation Systems[J]. arXiv preprint arXiv:1206.4684, 2012.

Mackey L W, Weiss D, Jordan M I. Mixed membership matrix factorization[C] Proceedings of the 27th International Conference on Machine Learning (ICML-10). 2010: 711-718.

Shan H, Banerjee A. Generalized probabilistic matrix factorizations for collaborative filtering[C] Data Mining (ICDM), 2010 IEEE 10th International Conference on. IEEE, 2010: 1025-1030.