密级：_____

浙江大学

# 硕 士 学 位 论 文


论文题目 **多关系数据挖掘中的概率模型研究**

作者姓名 李正洋

指导教师 徐从富 副教授

学科(专业) 计算机应用技术

所在学院 计算机学院

提交日期 2013 年 12 月 22 日

A Dissertation Submitted to Zhejiang University for the Degree of Master of Engineering


# TITLE: Study on Probabilistic Models in Multi-relational Data Mining

Author: Zhengyang Li

Supervisor: Asso. Prof. Congfu Xu

Subject: Computer Application Technology

College: Computer Science & Technology

Submitted Date: December 22, 2014

# 摘要

**关键词： 概率图模型 贝叶斯学习 主题混合模型 因子分解 协同过滤 关系数据挖掘**

# Abstract

**Keywords：**

# 目录

# 图示

# 第1章 概率图模型的推断

## 1.1 基本推断方法

### 1.1.1 变分近似推断

### 1.1.2 蒙特卡洛采样

## 1.2 指数族分布

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，这将有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，$u(X)$称作自然统计量，$g(Y)$是归一化因子(log-partition)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$\phi$求微分有$\int_X \frac{d}{d\phi'} p(X|\phi')$=0, 也即$\int_X p(X|\phi')\,[u(X) + \frac{d}{d\phi'}\tilde{g}(\phi')]$=0。X 的分布 p 对应的自然统计量$u(X)$在参数$\phi$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自然统计量求期望就是 log-partition 对自然参数的导数。

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

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

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

$$\Psi(t)=\Gamma'(t)/\Gamma(t)$$

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$\eta$是 d 维向量

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

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

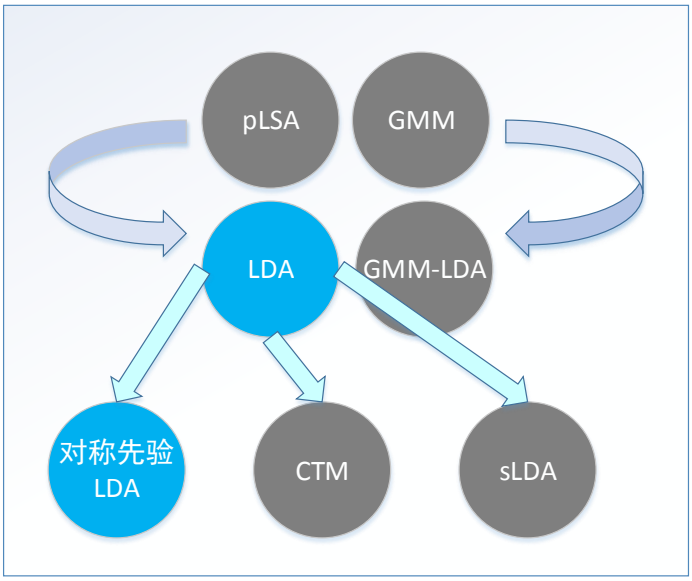
$$E_{p(\eta|\mu,\Sigma)}\begin{pmatrix}\eta\\ vec(\eta\eta^T)\end{pmatrix}=\frac{d}{d\Phi}g(\mu,\Lambda^{-1})$$

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

如果先验$\theta\sim Dir(\alpha), Z_n\sim Mult(\theta)$，$\theta$的后验形式为 $\widehat{p(\theta)}\propto p(\theta|\alpha)\prod_{n=1}p(Z_n|\theta)$，由于 Dir 分布中自然统计量与 Mult 分布中自然参数有相同的形式，且在 Dir 的条件下 Mult 中的 $g(\theta)=0$，因此$\widehat{p(\theta)}$服从 Dir 分布，所以我们称 Dir 分布是 Mult 的共轭分布。

# 第2章 混合主题模型

## 2.1 主要模型和算法




图示 1 几个基本主题模型衍生关系图

在这一章中我们将介绍主题模型的几个基本模型及其相应的推断算法，这些模型之间的衍生关系见图示-1。概率隐语义分析(probabilistic Latent Semantic Analysis)和高斯混合模型(Gaussian Mixture Model)都是多项式混合模型(Multinomial Mixture Model)，即模型生成过程中所选择的外部参数族由多项式分布确定。这两种模型因其有效的无监督聚类特性，被广泛应用在文本挖掘和计算机视觉的相关领域。本章将以 pLSA 与 GMM 这两个模型为基础，阐述图模型的一般建模方法，以及基于这两个基本模型的层次贝叶斯推广模型如 LDA 与 GMM-LDA，以及改进版本的 CTM 与“对称先验 LDA”以及 sLDA。本章内容将阐述上述模型的一般求解方法，即“近似推断法”和“采样法”。后续章节也将以此为基础。在本章节的最后，我们将实验比较“对称先验 LDA”的主题抽取效果。

这一节首先介绍两个基本模型：概率隐语义分析 pLSA 与高斯混合模型 GMM，并利用“EM 算法”框架推导两个模型的求解过程。接着介绍层级贝叶斯推广模

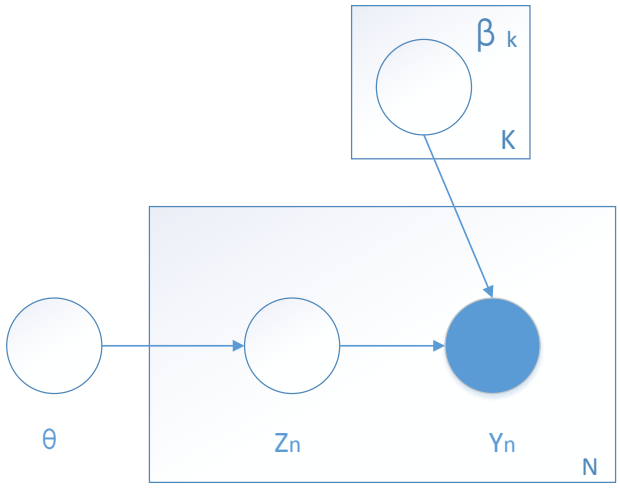
型 LDA 和 GMM-LDA，并阐述图模型中“近似推断”的一般方法。

### 2.1.1 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$\theta$确定，话题变量$z_i$服从$\theta$确定的 Multinomial 分布，记作$Z_i \sim Mult(\theta)$。文本中的单词$\{Y_1 \ldots\ldots Y_n\}$都各自对应一个主题$\{Z_1 \ldots\ldots Z_n\}$，这样文本中的每个$Y_i$服从$\beta_{z_i\cdot}$确定的 Multinomial 分布，记作$Y_i \sim Mult(\beta_{z_i\cdot})$。所以有以下的生成假设，对每个$Y_i \in \{Y_1 \ldots\ldots Y_n\}$：

(a) 从$\theta$确定的 Multinomial 分布产生话题$Z_i$，即$Z_i \sim Mult(\theta)$。

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




图示 2　概率隐语义分析 pLSA 模型的图表示

在 pLSA 中，参数$\theta$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$\theta$确定的多项式分布：$p(Z_n = k|\theta) = \theta_k$。即$Z_n$以概率$\theta_k$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$p(Y_n|Z_n = k, \boldsymbol{\beta}) = Mult(\beta_k)$。$Y_n$代表单词对应单词表中的序号，取值从 1 到 V，因而$\beta_k$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$\{\mathbf{Z}, \theta, \boldsymbol{\beta}\}$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：

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

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

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

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

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 $\boldsymbol{Z}$ 是隐变量。$\mathrm{q}(\mathbf{Z})$的近似分布在这里可利用后验$\mathrm{q}(\mathbf{Z}) \propto \mathrm{p}(\mathbf{Y},\boldsymbol{Z}|\theta^{old},\boldsymbol{\beta}^{old})$，其中$\theta^{old},\boldsymbol{\beta}^{old}$从上一次 M 步中得到。由于$\mathrm{q}(\mathbf{Z}|\theta^{old},\boldsymbol{\beta}^{old}) \propto \mathrm{p}(\mathbf{Y},\boldsymbol{Z}|\theta^{old},\boldsymbol{\beta}^{old}) \propto \Pi_{\mathrm{n}=1}^{\mathrm{N}}\{\mathrm{p}(Z_n|\theta^{\mathrm{old}})\mathrm{p}(\mathrm{Y_n}|\mathrm{Z_n},\boldsymbol{\beta}^{\mathrm{old}})\}$，可见在$\{\theta^{old},\boldsymbol{\beta}^{old}\}$已知的情况下$\{\mathrm{Z_n}\}$之间是相互独立的，所以有：

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

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

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

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

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

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

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

记$\mathrm{A_{kn}}$=$\theta_{\mathrm{k}}^{\mathrm{old}}\beta_{\mathrm{k}Y_n}^{\mathrm{old}}$，$\mathrm{A_n}$=$\Sigma_{\mathrm{k}=1}^{\mathrm{K}}\mathrm{A_{kn}}$，计算出$\mathrm{A_{kn}}$与$\mathrm{A_n}$(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$\mathrm{H}(\mathrm{q}(\mathrm{Z_n}))$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：

$\Sigma_{\mathrm{k}=1}^{\mathrm{K}}\Sigma_{\mathrm{n}=1}^{\mathrm{N}}\{(\mathrm{A_{kn}}\log\theta_k)/\mathrm{A_n} + (\mathrm{A_{kn}}\log\beta_{k,Y_n})/\mathrm{A_n}\}$，由于$\theta$ 和$\beta_{\mathrm{k}}$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：$\mathrm{O}_1(\theta) + \mathrm{O}_2(\boldsymbol{\beta})$

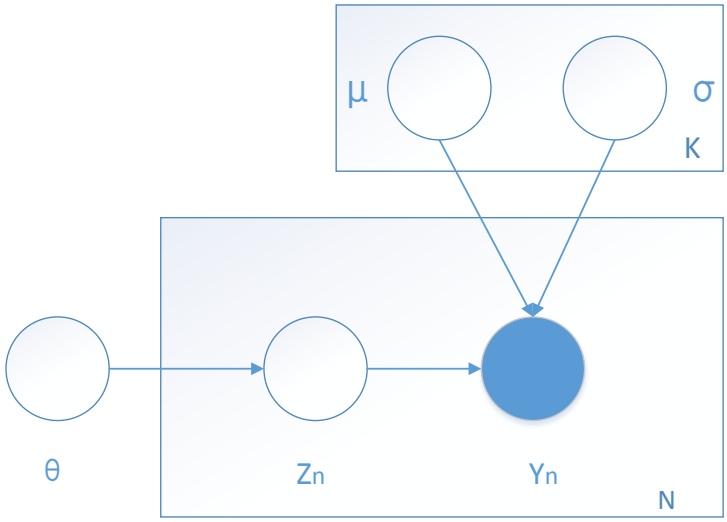
$O_1 = \Sigma_{k=1}^{K}\Sigma_{n=1}^{N}\{(A_{kn}\log\theta_k)/A_n\} + \lambda(\Sigma_{k=1}^{K}\theta_k - 1)$,

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

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

### 2.1.2 高斯混合模型 GMM 及 EM 算法




图示 3 高斯混合模型 GMM 的图表示

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$p(Y_n|Z_n = k, \boldsymbol{\mu}, \boldsymbol{\sigma}) = N(Y_n|\mu_k, \sigma_k^2)$。为方便起见让$\beta_k$也代表$\mu_k$和$\sigma_k^2$，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：

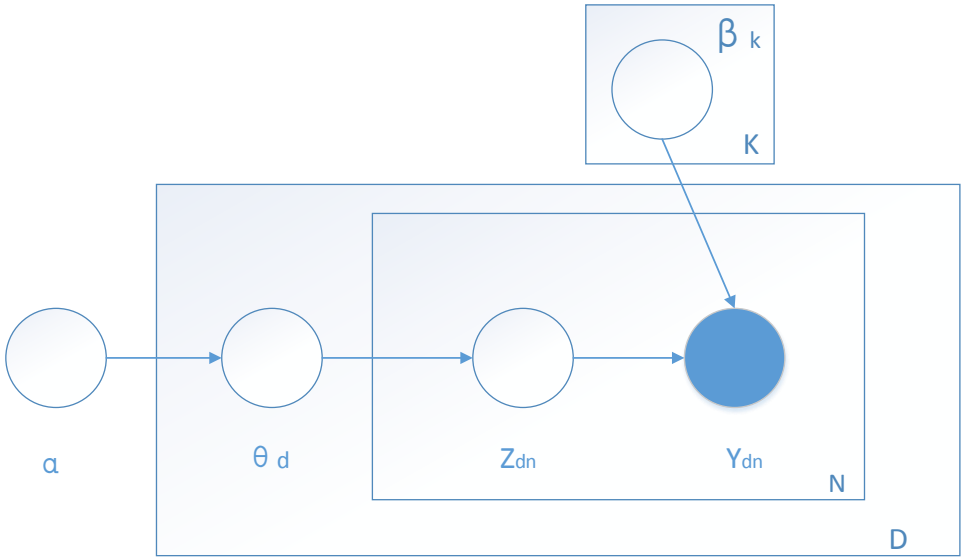
$\mu_k = (\Sigma_{n=1}^{N}\frac{A_{kn}}{A_n}Y_n)/(\Sigma_{n=1}^{N}\frac{A_{kn}}{A_n})$，$\sigma_k^2 = (\Sigma_{n=1}^{N}\frac{A_{kn}}{A_n}(Y_n-\mu_k)^2)/(\Sigma_{n=1}^{N}\frac{A_{kn}}{A_n})$，

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。

### 2.1.3 层级推广 LDA&GMM-LDA 及 VB-EM 算法

pLSA 与 GMM 作为基本的混合模型，广泛应用于文本分析和数据聚类。pLSA 在自然语言处理中的变种模型还有很多，比如我们希望它能够处理一个文本族，这时候图表-1 中的θ变量将会有 D 个，图表示也将会多出一层。同样地，我们也希望 GMM 混合模型能够应用到图像库，这都不可避免地需要增加图模型的层数。这样，整个推断过程仍然可以利用上述 EM 求解框架完成，只需要将 M 步中对外层参数β的更新步骤放在所有文件参数更新之后即可，而在β已知的情况下文件之间是相互独立。层级贝叶斯的求解策略，希望有一个外层参数α来刻画θ这个“主题分布参数”，这样做的好处是不需要初始化θ，在α 给定的情况下每个文件的θ参数相互独立，满足了文件之间的可交换性。由于 Dirichlet 是 Multinomial 的共轭分布，选择 Dirichlet 分布即θ~Dir(α)将有利于模型的推断(对 Mult 分布中自然参数在 Dir 下求期望将变得容易），这个模型被称作 LDA(latent dirichlet allocation)，同样有 GMM-LDA。

#### 2.1.3.1 LDA 模型




图示 4 LDA 的图模型表示

α是 k 维 Dirichlet 分布的参数，其中$\alpha_i>0$。为方便起见让$\beta_k$也代表$\mu_k$和$\sigma_k^2$，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，如图表-3 所示，这样的方法叫作变分贝叶斯(Variational Bayesian)VB 近似推断。在 VB 近似分布中，$\boldsymbol{\theta}$与$\mathbf{Z}$的所有分量都相互独立，这将简化 E 步的计算。

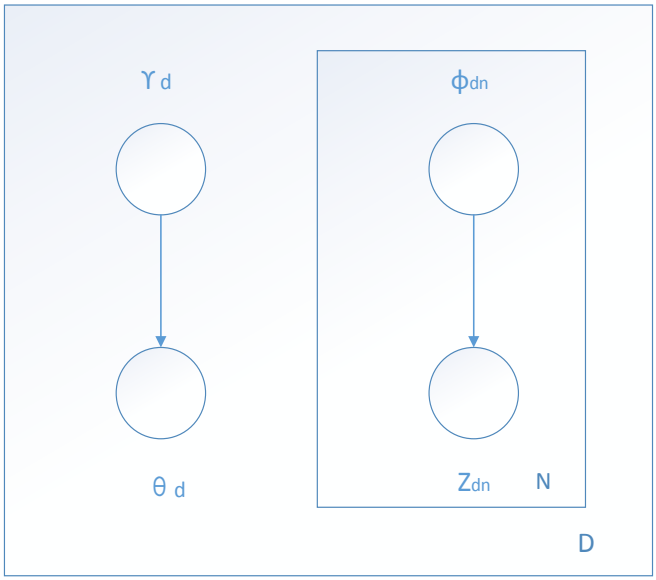
$$\log p(\mathbf{Y},\boldsymbol{Z},\boldsymbol{\theta}|\boldsymbol{\beta},\alpha)=\Sigma_{d=1}^{D}\{\log p(\theta_d|\alpha)+\Sigma_{n=1}^{Nd}\{\log p(Z_{dn}|\theta_d)+\log p(Y_{dn}|Z_{dn},\boldsymbol{\beta})\}\},$$

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

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

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




VB 近似分布表

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

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

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

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

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

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

图示 5 $q(\mathbf{Z},\boldsymbol{\theta})$近似分布的图模型表示和分布表

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$\{\alpha,\boldsymbol{\beta}\}$。有了上面的结论，我们将展开对 LDA 与 GMM-LDA 的 E-M 步骤，

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$\theta$、$Z_n$以及$\gamma$、$\phi_n$均默认带有下标 d。$E_{q(Z,\theta)}\log p(\theta|\alpha) = \phi_{Dir}(\alpha)^T E_{q(\theta)}(u_{Dir}(\theta)) + g_{Dir}(\alpha)$

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

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

$$= \begin{pmatrix}\Psi(\gamma_1) - \Psi(\sum_{i=1}^{K}\gamma_i)\\ \cdots\\ \Psi(\gamma_K) - \Psi(\sum_{i=1}^{K}\gamma_i)\end{pmatrix}^T \begin{pmatrix}\Sigma_{n=1}^{Nd}\phi_{n1}\\ \cdots\\ \Sigma_{n=1}^{Nd}\phi_{nK}\end{pmatrix}$$

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

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

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

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

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

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

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：

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

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

$$\mathrm{term}\{1,2,4\}(\gamma_{\mathrm{i}}) = (1 - \gamma_i)\Big( \Psi(\gamma_{\mathrm{i}}) - \Psi\big(\textstyle\sum_{k=1}^{\mathrm{K}} \gamma_k\big)\Big)$$

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

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

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

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

考虑$\phi_{\mathrm{n}}$，并加入拉格朗日约束

$$\mathrm{term}\{2,3,5\}(\phi_{\mathrm{n}}) = \begin{pmatrix} \Psi(\gamma_1) - \Psi\big(\sum_{i=1}^{\mathrm{K}} \gamma_i\big) \\ \dots \dots \\ \Psi(\gamma_{\mathrm{K}}) - \Psi\big(\sum_{i=1}^{\mathrm{K}} \gamma_i\big) \end{pmatrix}^T \begin{pmatrix} \phi_{\mathrm{n}1} \\ \dots \dots \\ \phi_{\mathrm{n}\mathrm{K}} \end{pmatrix} - \begin{pmatrix} log\phi_{n1} \\ \dots \dots \\ log\phi_{nk} \end{pmatrix}^T \begin{pmatrix} \phi_{n1} \\ \dots \dots \\ \phi_{nk} \end{pmatrix} +$$

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

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

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

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

设置导数为 0，这样事实上有：

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

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：

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

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

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$\alpha_{\mathrm{i}}$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：

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

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

$(\partial \mathrm{Oj}(\alpha))_{\mathrm{k}} = \frac{\partial \mathrm{Oj}(\alpha)}{\partial \alpha_{\mathrm{k}}}$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$\beta_{\mathrm{k}}$需要加入朗格朗日约束，目标函数如下：

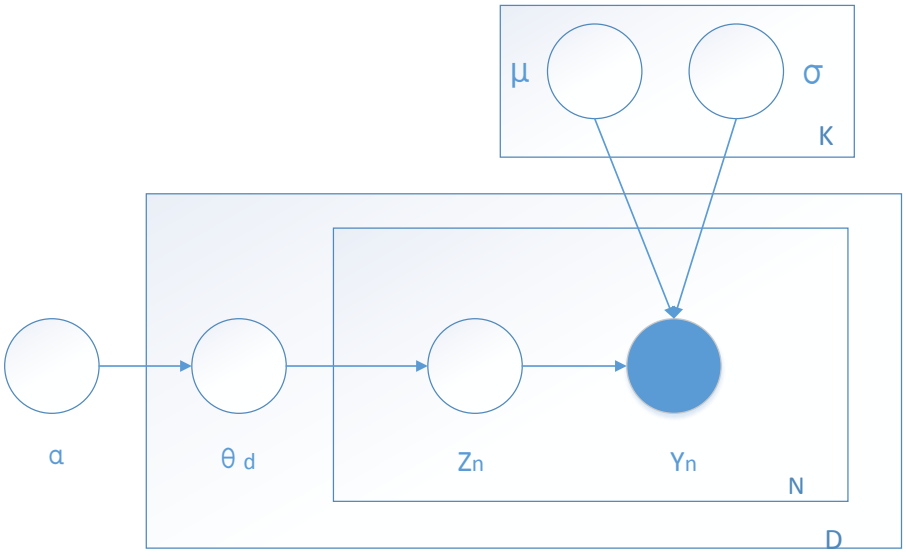
$$\Sigma_{\mathrm{d}=1}^{\mathrm{D}}\Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}\Sigma_{\mathrm{k}=1}^{K}\phi_{\mathrm{dnk}}\log\beta_{k\mathrm{Y_n}} + \Sigma_{\mathrm{k}=1}^{\mathrm{K}}\lambda_{\mathrm{k}}(\Sigma_{\mathrm{v}=1}^{\mathrm{V}}\beta_{\mathrm{kv}} - 1)$$

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

对上式中的某个$\beta_{\mathrm{kv}}$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$\beta_{\mathrm{kv}} \propto \Sigma_{\mathrm{d}=1}^{\mathrm{D}}\Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}\delta(\mathrm{Y_{dn}} = \mathrm{v})\phi_{\mathrm{dnk}}$。

#### 2.1.3.2 GMM-LDA 模型




图示 6　GMM-LDA 模型的图表示

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，

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

$=\Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}E_{q(\mathrm{Z_n})}\delta(Z_n=k)\log \mathrm{N}(\mathrm{Y_n}|\mu_{\mathrm{k}},\sigma_{\mathrm{k}}^2)=\Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}\Sigma_{\mathrm{k}=1}^{K}\phi_{\mathrm{nk}}\log \mathrm{N}(\mathrm{Y_n}|\mu_{\mathrm{k}},\sigma_{\mathrm{k}}^2)$

这样在 E 步的计算中，我们就有（省略下标 d）：

$\gamma_{\mathrm{i}}=\alpha_i+\Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}\phi_{\mathrm{ni}}\ (\mathrm{i}=1\ldots\mathrm{K})$与$\phi_{\mathrm{ni}}\propto \mathrm{N}(\mathrm{Y_n}|\mu_{\mathrm{i}},\sigma_{\mathrm{i}}^2)(\Psi(\gamma_{\mathrm{i}})-\Psi(\sum_{i=1}^{\mathrm{K}}\gamma_i))$，

$\Sigma_{\mathrm{k}=1}^{K}\phi_{\mathrm{nk}}=1$。

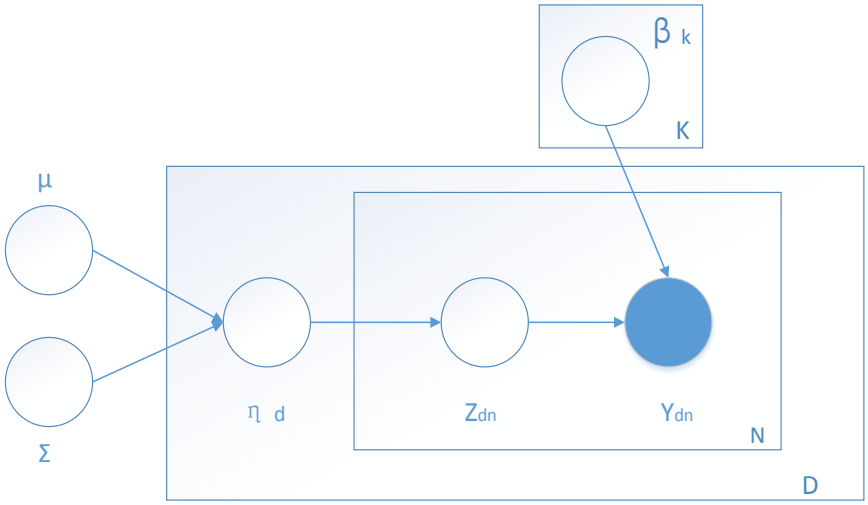
M-step： 对α的更新不涉及到$\{\mu,\sigma\}$，同 LDA。对$\{\mu,\sigma\}$的更新只涉及的目标函数如下：$\Sigma_{\mathrm{d}=1}^{\mathrm{D}}\Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}\Sigma_{\mathrm{k}=1}^{K}\phi_{\mathrm{dnk}}\log \mathrm{N}(\mathrm{Y_{dn}}|\mu_{\mathrm{k}},\sigma_{\mathrm{k}}^2)$。不妨让所有的$\{\mathrm{Y_{dn}}\}$以下标 t 重新标号$\mathrm{T}=\Sigma_{\mathrm{d}=1}^{\mathrm{D}}\mathrm{N_d}$，则有$\Sigma_{\mathrm{t}=1}^{\mathrm{T}}\Sigma_{\mathrm{k}=1}^{\mathrm{K}}\mathrm{A_{kt}}\log N(Y_t|\mu_k,\sigma_k^2)$，最大化它则得到

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

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。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，这将大大简化 E 步的计算。我们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。

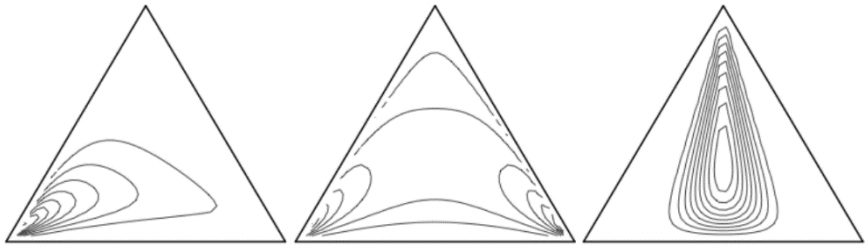
### 2.1.4 改进模型 CTM 及 VB-EM 算法

以上的 LDA 和 GMM-LDA 模型，都是基于 Dirichlet 分布为先验的，从前面的叙述中我们了解到，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$\eta\sim \mathrm{N}(\mu,\Sigma)$，$\theta_{\mathrm{i}}=\exp(\eta_i)/\Sigma_{\mathrm{i}=1}^{\mathrm{K}}\exp(\eta_i)$，$\mathrm{Z_n}\sim\mathrm{Mult}(\theta)$，这个模型如图表-5 所示是 CTM(correlated topic model)模型。我们把这个模型放在本节的最后阐述，作为 pLSA 层级贝叶斯推广的另一个范例。




图示 7 CTM(correlated topic model)的图模型表示

图-6 中所示为 3 维罗杰斯特高斯分布密度图，左起第一个$\boldsymbol{\Sigma}$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。



图示 8 罗杰斯特高斯分布图

CTM 的 VB-EM 型近似推断，近似分布$q(\mathbf{Z},\boldsymbol{\eta})$将类似图表-3 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$\eta_d$的分量相互独立，联合分布为：

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

采用近似推断方法来得到 log 后验的一个下界：

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

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

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

$H(q(\boldsymbol{Z}, \boldsymbol{\eta})) = \Sigma_{\mathrm{d}=1}^{\mathrm{D}} \{\mathrm{H}\big(\mathrm{q}(\eta_{\mathrm{d}})\big) + \Sigma_{\mathrm{n}=1}^{\mathrm{Nd}} \mathrm{H}\big(\mathrm{q}(\mathrm{Z}_{\mathrm{dn}})\big)\}$,

$H\big(q(\boldsymbol{Z_d}, \eta_d)\big) = \mathrm{H}\big(\mathrm{q}(\eta_{\mathrm{d}})\big) + \Sigma_{\mathrm{n}=1}^{\mathrm{Nd}} \mathrm{H}\big(\mathrm{q}(\mathrm{Z}_{\mathrm{dn}})\big)$,

$\mathrm{LB_d} = E_{q(\mathbf{Z_d}, \eta_{\mathrm{d}})} \{\log \mathrm{p}(\eta_{\mathrm{d}}|\mu, \Lambda^{-1}) + \Sigma_{\mathrm{n}=1}^{\mathrm{Nd}} \{\log \mathrm{p}(Z_{dn}|\eta_{\mathrm{d}}) + \log \mathrm{p}(\mathrm{Y}_{\mathrm{dn}}|\mathrm{Z}_{\mathrm{dn}}, \boldsymbol{\beta})\}\} +$

$H(q(\boldsymbol{Z_d}, \eta_d))$，则有：

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

E-step：下面的讨论同样省略下标 d。

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

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

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

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

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

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

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

$E_{q(\eta)} \exp(\eta_i) = E_{q(\eta_{\mathrm{i}})} \exp(\eta_i) = \exp\{\lambda_{\mathrm{i}} + \gamma_{\mathrm{i}}^2/2\}$

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

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

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

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

$$-\log \sqrt{\frac{|\gamma^{2^{-1}}|}{(2\pi)^K}} + \frac{1}{2} K \qquad \text{term④}$$

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

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$\phi$的有 2、3、5。先考虑参数$\{\lambda,\gamma^2\}$：

$$\mathrm{term}\{1,2,4\}(\{\lambda,\gamma^2\})=\lambda^T\Lambda\mu-\frac{1}{2}\lambda^T\Lambda\lambda-\frac{1}{2}\mathrm{tr}(\Lambda\gamma^2)$$
$$+\frac{1}{2}\Sigma_{\mathrm{i=1}}^{\mathrm{K}}\log\boldsymbol{\gamma_i^2}+\begin{pmatrix}\lambda_1-\zeta^{-1}\left(\Sigma_{\mathrm{i=1}}^{\mathrm{K}}\exp\{\lambda_\mathrm{i}+\gamma_\mathrm{i}^2/2\}\right)\\ \cdots\cdots\\ \lambda_\mathrm{K}-\zeta^{-1}\left(\Sigma_{\mathrm{i=1}}^{\mathrm{K}}\exp\{\lambda_\mathrm{i}+\gamma_\mathrm{i}^2/2\}\right)\end{pmatrix}^{T}\begin{pmatrix}\Sigma_{\mathrm{n=1}}^{\mathrm{Nd}}\phi_{\mathrm{n1}}\\ \cdots\cdots\\ \Sigma_{\mathrm{n=1}}^{\mathrm{Nd}}\phi_{\mathrm{nK}}\end{pmatrix}$$

$$\frac{\partial\mathrm{term}\{1,2,4\}(\lambda)}{\partial\lambda}=-\Lambda(\lambda-\mu)+\Sigma_{\mathrm{n=1}}^{\mathrm{Nd}}\phi_{\mathrm{n1:K}}-\frac{\mathrm{N}}{\zeta}\left\{\exp\left(\lambda_\mathrm{i}+\frac{\gamma_i^2}{2}\right)\right\}_{\mathrm{i=1:K}}$$

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

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$\phi_\mathrm{n}$，并加入拉格朗日约束，仅有 term②与LDA稍有不同，立刻有：$\phi_{\mathrm{ni}}\propto\beta_{\mathrm{iY_n}}\exp(\lambda_\mathrm{i})$，$\Sigma_{\mathrm{k=1}}^{K}\phi_{\mathrm{nk}}=1$。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$\mu,\Lambda^{-1}$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：$\mathrm{Oj}(\mu,\Lambda^{-1})=\Sigma_{\mathrm{d=1}}^{D}(\lambda_d^T\Lambda\mu-\frac{1}{2}\mathrm{tr}(\Lambda\boldsymbol{\gamma_d^2})-\frac{1}{2}(\mu^{\mathrm{T}}\Lambda\mu+\lambda_d^T\Lambda\lambda_d))+D\log\sqrt{\frac{|\Lambda|}{(2\pi)^K}}$，$\frac{\partial\mathrm{Oj}(\mu,\Lambda^{-1})}{\partial\mu}=\Sigma_{\mathrm{d=1}}^{D}(\lambda_d^T\Lambda-\Lambda\mu)$，

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

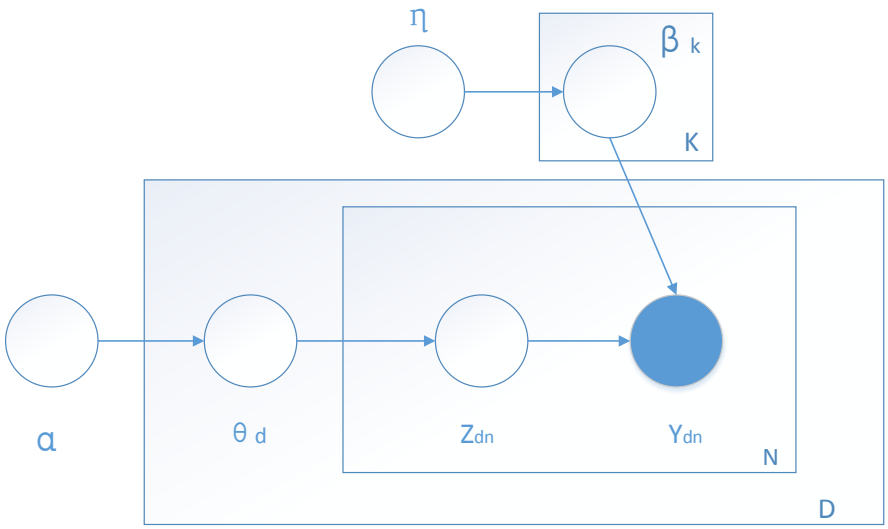
所以有：$\mu=\frac{1}{\mathrm{D}}\Sigma_{\mathrm{d=1}}^{D}\lambda_\mathrm{d}$，$\Lambda^{-1}=\frac{1}{\mathrm{D}}\Sigma_{\mathrm{d=1}}^{D}(\boldsymbol{\gamma_d^2}+(\lambda_d-\mu)(\lambda_d-\mu)^T)$

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

### 2.1.5 对称先验 LDA 及 VB-EM 算法

第 1 节的讨论中给出了 LDA 的 VB-EM 推断算法，这一节将首先介绍光滑 LDA 模型(smoothed-LDA)也称作“对称先验 LDA”，以及它的 VB-EM 推断、采

样算法。




图示 9 Smooth-LDA 或“对称先验 LDA”的图模型表示

由于$\beta_k$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 在第 1 节中的 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“过拟合”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$\eta$，有$\beta_k \sim Dir(\eta)$，这就是 smoothed-LDA 也称作对称 LDA。对称先验 LDA 的变分推断，可以让$q(\beta_k)$为变分分布$\beta_k \sim Dir(\lambda_k)$。下面首先给出 VB-EM 推导，考虑联合分布为：

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

采用近似推断的方法来得到 log 后验的一个下界：

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

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

这里$(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{\beta})$是该模型的隐变量，$q((\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{\beta}))$将选择完全可分解的变分分布，$\log p(\mathbf{Y},\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{\beta}|\alpha,\eta) = \Sigma_{d=1}^{D}\{\log p(\theta_d|\alpha) + \Sigma_{n=1}^{Nd}\{\log p(Z_{dn}|\theta_d) + \log p(Y_{dn}|Z_{dn},\boldsymbol{\beta})\}\} + \Sigma_{k=1}^{K}\log p(\beta_k|\eta)$ ，类似于 LDA 的推断(非对称先验)，这里稍有不同就是对含有$\boldsymbol{\beta}$变量的下界求期望。这样$LB_d$下界将会由 6 个 term 来表示。

E-step: term①与 term②、term④、term⑤都与 LDA 的推断(非对称先验)相同。

$$E_{q(\mathrm{Z},\theta,\boldsymbol{\beta})}\Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}\log \mathrm{p}(\mathrm{Y_n}|\mathrm{Z_n},\beta)=\Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}E_{q(\mathrm{Z_n})q(\boldsymbol{\beta})}\log \beta_{\mathrm{Z_nY_n}}=$$
$$\Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}\Sigma_{\mathrm{k}=1}^{K}\phi_{\mathrm{nk}}(\Psi(\lambda_{\mathrm{kY_{dn}}})-\Psi(\Sigma_{\mathrm{v}=1}^{\mathrm{V}}\lambda_{\mathrm{kv}}))$$

这其中涉及到参数$\gamma$的有 1、2、4 涉及到参数$\phi$的有 2、3、5，实际计算可以发现 E 步与非对称 LDA 相同：

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

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

$$E_{q(\boldsymbol{\beta})}\log \mathrm{p}(\beta_{\mathrm{k}}|\eta)=\begin{pmatrix}\eta_1-1\\ \ldots\ldots\\ \eta_{\mathrm{V}}-1\end{pmatrix}^{\mathrm{T}}\begin{pmatrix}\Psi(\lambda_{\mathrm{k1}})-\Psi\left(\sum_{i=1}^{\mathrm{V}}\lambda_{\mathrm{ki}}\right)\\ \ldots\ldots\\ \Psi(\lambda_{\mathrm{kV}})-\Psi\left(\sum_{i=1}^{\mathrm{V}}\lambda_{\mathrm{ki}}\right)\end{pmatrix}+\log\frac{\Gamma(\sum_{i=1}^{\mathrm{V}}\eta_i)}{\Pi_{\mathrm{i}=1}^{\mathrm{V}}\Gamma(\eta_i)}$$

这里不同的是涉及对变分参数$\boldsymbol{\lambda}$的更新，对其中的一个参数$\beta_{\mathrm{k}}$有：

$$\mathrm{H}(\mathbf{q}(\beta))=-E_{\mathbf{q}(\beta)}\log\mathbf{q}(\beta)=-\phi_{\mathrm{Dir}}(\lambda)^{T}E_{\mathbf{q}(\beta)}u_{\mathrm{Dir}}(\beta)-g_{\mathrm{Dir}}(\lambda)$$

$$=-\begin{pmatrix}\lambda_1-1\\ \ldots\ldots\\ \lambda_V-1\end{pmatrix}^{\mathrm{T}}\begin{pmatrix}\Psi(\lambda_1)-\Psi\left(\sum_{i=1}^{\mathrm{V}}\lambda_i\right)\\ \ldots\ldots\\ \Psi(\lambda_{\mathrm{V}})-\Psi\left(\sum_{i=1}^{\mathrm{V}}\lambda_i\right)\end{pmatrix}-\log\frac{\Gamma(\sum_{i=1}^{\mathrm{V}}\lambda_i)}{\Pi_{\mathrm{i}=1}^{\mathrm{V}}\Gamma(\lambda_i)}$$

D 个 term③中包含$\lambda_{\mathrm{k}}$的项为$\Sigma_{\mathrm{d}=1}^{\mathrm{D}}\Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}\phi_{\mathrm{nk}}(\Psi(\lambda_{\mathrm{kY_n}})-\Psi(\Sigma_{\mathrm{v}=1}^{\mathrm{V}}\lambda_{\mathrm{kv}}))$，这样 LB 下界中包含$\lambda_{\mathrm{k}}$的项，可以写为：

$$\begin{pmatrix}\eta_1-\lambda_{\mathrm{k1}}\\ \ldots\ldots\\ \eta_{\mathrm{V}}-\lambda_{kV}\end{pmatrix}^{\mathrm{T}}\begin{pmatrix}\Psi(\lambda_{\mathrm{k1}})-\Psi\left(\sum_{i=1}^{\mathrm{V}}\lambda_{ki}\right)\\ \ldots\ldots\\ \Psi(\lambda_{\mathrm{kV}})-\Psi\left(\sum_{i=1}^{\mathrm{V}}\lambda_{ki}\right)\end{pmatrix}+\Sigma_{\mathrm{d}=1}^{\mathrm{D}}\Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}\phi_{\mathrm{dnk}}(\Psi(\lambda_{\mathrm{kY_{dn}}})$$
$$-\Psi(\Sigma_{\mathrm{v}=1}^{\mathrm{V}}\lambda_{\mathrm{kv}}))-\log\frac{\Gamma(\sum_{i=1}^{\mathrm{V}}\lambda_{ki})}{\Pi_{\mathrm{i}=1}^{\mathrm{V}}\Gamma(\lambda_{ki})}$$

对其中包含的$\lambda_{ki}$求导，则有

$$(\eta_{\mathrm{i}}-\lambda_{\mathrm{ki}})(\Psi'(\lambda_{\mathrm{ki}})-\Psi'(\Sigma_{\mathrm{v}=1}^{\mathrm{V}}\lambda_{\mathrm{kv}}))+\Sigma_{\mathrm{d}=1}^{\mathrm{D}}\Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}\phi_{\mathrm{dnk}}\delta(\mathrm{Y_{dn}}=\mathrm{i})(\Psi'(\lambda_{\mathrm{kY_{dn}}})$$

$$-\Psi'\left(\Sigma_{v=1}^{V}\lambda_{kv}\right))$$

设上式为 0，有：

$$\lambda_{ki}=\eta_i+\Sigma_{d=1}^{D}\Sigma_{n=1}^{Nd}\phi_{dnk}\delta(Y_{dn}=i)$$

以上我们给出了$\{\boldsymbol{\gamma},\boldsymbol{\phi},\boldsymbol{\lambda}\}$这三种变分分布的 E 步更新。

M-step：对参数α 的更新只涉及 term①，因此与非对称 LDA 的解法一样。而包含参数η的 LB 下界目标函数与α的同型，如下式，因此可以参照非对称先验 LDA 中α的更新策略。

$$Ob(\eta)=\begin{pmatrix}\eta_1-1\\ \cdots\cdots\\ \eta_V-1\end{pmatrix}^{T}\begin{pmatrix}\Sigma_{k=1}^{K}\Psi(\lambda_{k1})-\Sigma_{k=1}^{K}\Psi\left(\sum_{i=1}^{V}\lambda_{ki}\right)\\ \cdots\cdots\\ \Sigma_{k=1}^{K}\Psi(\lambda_{kV})-\Sigma_{k=1}^{K}\Psi\left(\sum_{i=1}^{V}\lambda_{ki}\right)\end{pmatrix}+\log\frac{\Gamma(\sum_{i=1}^{V}\eta_i)}{\Pi_{i=1}^{V}\Gamma(\eta_i)}$$

### 2.1.6 对称先验 LDA 的 MCMC 与 BP 算法

联合分布$p(\mathbf{Y},\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{\beta}|\alpha,\eta)$写出完全条件分布有：

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

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

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

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

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

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

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

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

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

可见 Dir 和 Mult 共轭关系使得$\theta_d$仍然为 Dir 分布，这样只需对参数$\alpha,\eta$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。

For t=1…T：(需要初始化$\boldsymbol{\beta},\alpha,\eta$)

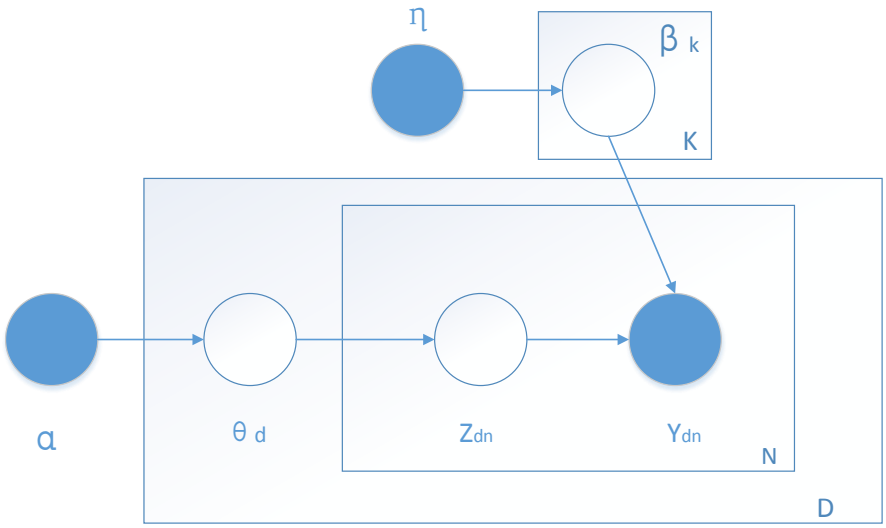
$$\theta_d^t \sim \mathrm{Dir}\left(\alpha^{t-1}+\left\{\sum_{n=1}^{N_d}\delta(Z_{dn}^{t-1}=k)\right\}_{k=1\ldots K}\right)，其中\ d=1\ldots D。$$

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

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

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

另一种 Gibbs 采样方法利用 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，见下图。




图示 10　坍缩 Gibbs 采样情形下的 LDA 图模型表示

联合分布为：

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

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

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

我们引入新的记号$\#\{d,v,k\}=\Sigma_{n=1}^{N_d}\delta(Y_{dn}=v,Z_{dn}=k)$，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$\#\{d,\cdot,k\}=\Sigma_{n=1}^{N_d}\delta(Z_{dn}=k)$表示文本 d 中话题值为 k 的单词出现的次数，类似地，$\#\{\cdot,\cdot,k\}=\Sigma_{d=1}^{D}\Sigma_{n=1}^{N_d}\delta(Z_{dn}=k)$，它们与$\{\mathbf{Y},\boldsymbol{Z}\}$有关。

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

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

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

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

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

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

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。

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

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

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

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

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：

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

以上，我们推导了“对称先验 LDA”的 VB-EM 算法和 Gibbs 采样算法，下面给出该模型的在置信传播算法。

图模型的“置信传播”(Belief Propagation)算法基于“因子分解图”(factor graph)做信息传递(massage passing)，它将变量之间的关系看作是马尔科夫随机域(markov random field)，即无向图模型。由于“因子分解图”对条件分布和联合分布同样适用，因此“置信传播”(Belief Propagation)算法同样适用于有向图模型。

仍然利用对称先验 LDA 模型的相应假设，基于 Dir 和 Mult 共轭关系积分掉$(\boldsymbol{\theta},\boldsymbol{\beta})$之后有：

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

因此构建因子函数如下：

$$f_{\theta_d}(\boldsymbol{Y_d},\boldsymbol{Z_d},\alpha)=\frac{\Pi_{k=1}^{K}\Gamma(\alpha_k+\#\{d,\cdot,k\})}{\Gamma(\sum_{k=1}^{K}(\alpha_k+\#\{d,\cdot,k\}))}$$

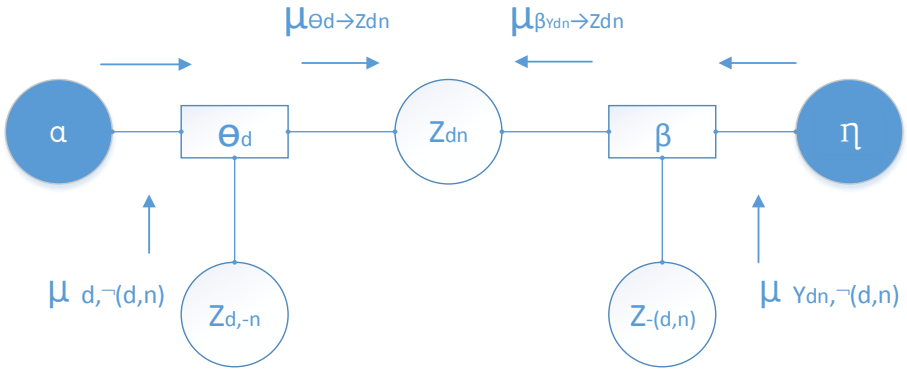
$$f_{\beta_{\cdot v}}(\boldsymbol{Y},\boldsymbol{Z},\eta)=\Pi_{k=1}^{K}\frac{\Gamma(\eta_v+\#\{\cdot,v,k\})}{\Gamma(\sum_{v=1}^{V}(\eta_v+\#\{\cdot,v,k\}))}$$

$$p(\mathbf{Y},\boldsymbol{Z}|\alpha,\eta)\propto\Pi_{d=1}^{D}f_{\theta_d}(\boldsymbol{Y_d},\boldsymbol{Z_d},\alpha)\Pi_{v=1}^{V}f_{\beta_{\cdot v}}(\boldsymbol{Y},\boldsymbol{Z},\eta)$$

上式是成立的，这涉及到分母部分不变化的讨论。因子函数中包含$Z_{dn}$的部分作为我们的传递信息$\mu_{dn}(k)$，由于$p(\mathbf{Y},\boldsymbol{Z}|\alpha,\eta)$同比例于以下条件概率：

$$\begin{aligned}&p\left(Z_{dn}=k,Y_{dn}=v,\mathbf{Y}^{\neg(d,n)},\boldsymbol{Z}^{\neg(d,n)}\middle|\alpha,\eta\right)\\&\propto f_{\theta_d}\left(Z_{dn}=k,\boldsymbol{Y_d}^{\neg(n)},\boldsymbol{Z_d}^{\neg(n)},\alpha\right)f_{\beta_{\cdot v}}\left(\mathbf{Y}^{\neg(d,n)},\boldsymbol{Z}^{\neg(d,n)},Y_{dn}=v,Z_{dn}=k,\eta\right)\\&\propto\frac{\alpha_k+\#\{d,\cdot,k\}^{\neg(d,n)}}{\sum_{k=1}^{K}\alpha_k+\#\{d,\cdot,\cdot\}^{\neg(d,n)}}\frac{\eta_v+\#\{\cdot,v,k\}^{\neg(d,n)}}{\sum_{v=1}^{V}\eta_v+\#\{\cdot,\cdot,k\}^{\neg(d,n)}}\end{aligned}$$

这与前面所得到的结论一致。由 LDA 模型的马尔科夫性构建信息传递的图模型表示如下。




图示 11　BP 情形下的 LDA 信息传递因子图

用以下分支信息向量$\mu_d$与$\mu_v$代替条件分布中的相应值可确定传递信息$\mu_{dn}(k)$。

$$\mu_{d,\neg(d,n)}(k) = \#\{d,\cdot,k\}^{\neg(d,n)},$$

$$\mu_{v,\neg(d,n)}(k) = \#\{\cdot,v,k\}^{\neg(d,n)},$$

“置信传播”算法需要已知参数$\alpha,\eta,\mathbf{Y},V,K$并初始归一化向量$\mu_d(k)$与$\mu_v(k)$。循环 T 次或更多次(d=1…D, n=1…Nd)：

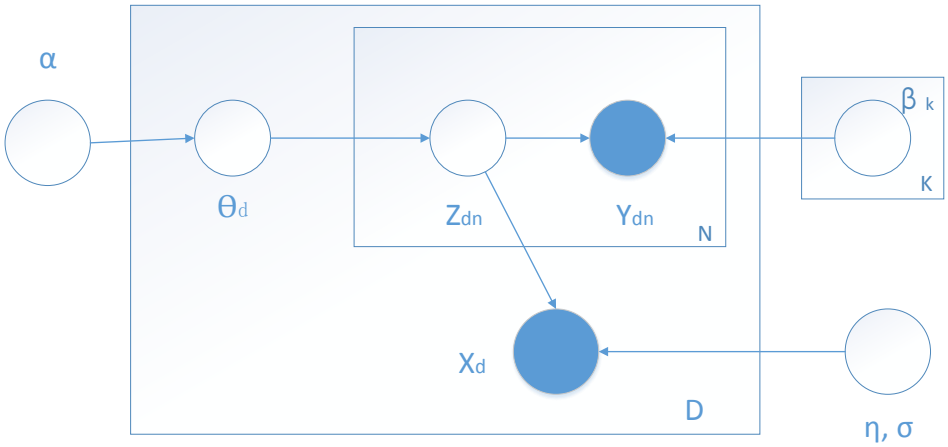
$$\mu_{dn}(k) \propto \frac{\eta_{Y_{dn}} + \mu_{Y_{dn},\neg(d,n)}(k)}{\sum_{v=1}^{V}\{\eta_v + \mu_{v,\neg(d,n)}(k)\}} \frac{\alpha_k + \mu_{d,\neg(d,n)}(k)}{\sum_{t=1}^{K}\{\alpha_k + \mu_{d,\neg(d,n)}(t)\}}$$

其中$\mu_{v,\neg(d,n)}(k) = \mu_v(k) - \delta(Y_{dn} = v)$，$\mu_{d,\neg(d,n)}(k)$则近似取$\mu_d(k)$。最后计算出参数$(\boldsymbol{\theta},\boldsymbol{\beta})$：

$$\theta_d(k) = \frac{\alpha_k + \mu_d(k)}{\sum_{k=1}^{K}\{\alpha_k + \mu_d(k)\}}, \qquad \beta_{kv} = \frac{\eta_v + \mu_v(k)}{\sum_{v=1}^{V}\{\eta_v + \mu_v(k)\}}$$

可见 Collapsed BP 算法与 Collapsed Gibbs 算法的不同是利用信息传递向量循环迭代而不是循环采样。

### 2.1.7 监督学习主题模型 sLDA 及 VB-EM 算法



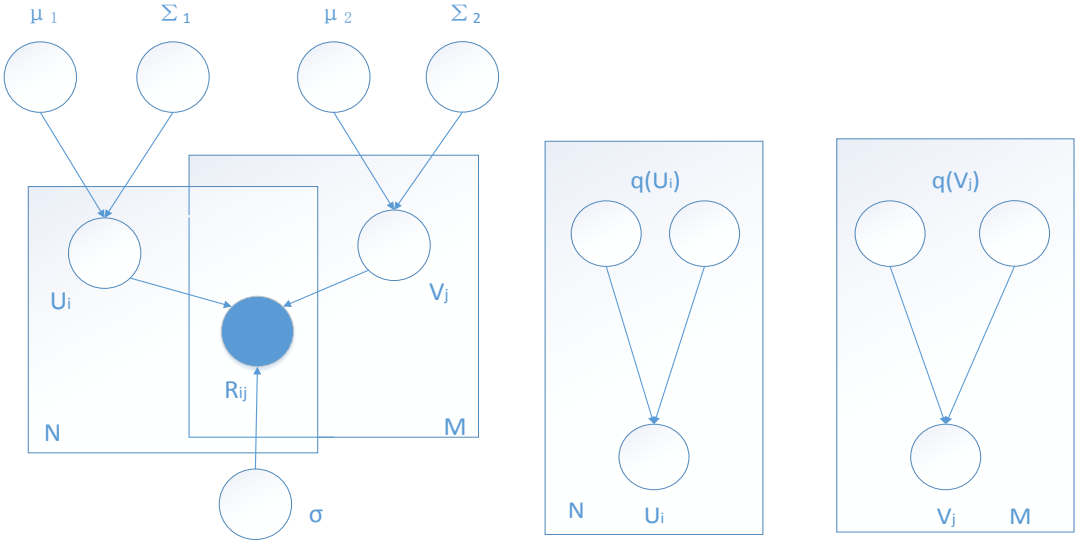

图示 12　监督主题模型 sLDA 的图模型表示

监督主题模型 sLDA(supervised LDA or topic-model)假设文本的分类标签值$X_d$满足如下关系：$X_d \sim N\left(\eta^T\left(\frac{\sum_{n=1}^{N_d}\{Z_{dn}\}}{N_d}\right),\sigma^2\right)$，这里$Z_{dn}$表示一个第 k 分量为 1 其余分量为 0 的向量(若$Z_{dn} = k$)，$X_d$可以是多分类标签。这个模型的求解仍然采用 VB-EM 算法，

# 第3章 因子分解模型

## 3.1 主要模型和算法

### 3.1.1 概率矩阵分解 PMF 及 VB-EM 算法




图示 13 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$q(\mathrm{U_i})$与 $q(\mathrm{V_j})$。

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

对上式中的$\Phi^i$求导并令导数为 0，得到：

[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。

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

对协方差$\Sigma^{\mathrm{i}}$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$\mu_1,\Sigma_1,\mu_2,\Sigma_2,\sigma^2$。

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

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

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

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

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

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

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

对其中的$\sigma^2$求导，并置导数为 0 得：

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

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

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

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

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

同样我们有：

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

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

上面考虑的 PMF 模型以及变分分布中的协方差矩阵$\Sigma_1,\Sigma_2$和所有的$\Sigma^{\mathrm{i}},\Sigma^{\sim\mathrm{j}}$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$\Sigma_1=\sigma_1^2 I,\Sigma_2=\sigma_2^2 I$，分情况讨论如下：

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

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

M-step： $\mu_1,\mu_2,\sigma^2$的更新不变。

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

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

②$\Sigma^{\mathrm{i}}=diag(\gamma_1^{\mathrm{i}}\ldots\gamma_{\mathrm{d}}^{\mathrm{i}}),\Sigma^{\sim\mathrm{j}}=diag(\gamma_1^{\sim\mathrm{j}}\ldots\gamma_{\mathrm{d}}^{\sim\mathrm{j}})$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： $\Phi^{\mathrm{i}}$与$\Phi^{\sim\mathrm{j}}$的更新不变

$$\gamma_{\mathrm{k}}^{\mathrm{i}\,2}=\left(\frac{1}{\sigma^2}\Sigma_{\mathrm{j}=1}^{\mathrm{M}}\delta(\mathrm{i},\mathrm{j})\left(\gamma_{\mathrm{k}}^{\sim\mathrm{j}\,2}+\Phi_{\mathrm{k}}^{\sim\mathrm{j}\,2}\right)+\Sigma_{1,\mathrm{kk}}^{-1}\right)^{-1}$$

$$\gamma_{\mathrm{k}}^{\sim\mathrm{j}\,2}=\left(\frac{1}{\sigma^2}\Sigma_{\mathrm{i}=1}^{\mathrm{N}}\delta(\mathrm{i},\mathrm{j})\left(\gamma_{\mathrm{k}}^{\mathrm{i}\,2}+\Phi_{\mathrm{k}}^{\mathrm{i}\,2}\right)+\Sigma_{2,\mathrm{kk}}^{-1}\right)^{-1}$$

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

算法 1 PMF(VB-EM 算法)

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 $\Phi^i$(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 $\mu_1, \Sigma_1, \mu_2, \Sigma_2, \sigma$。

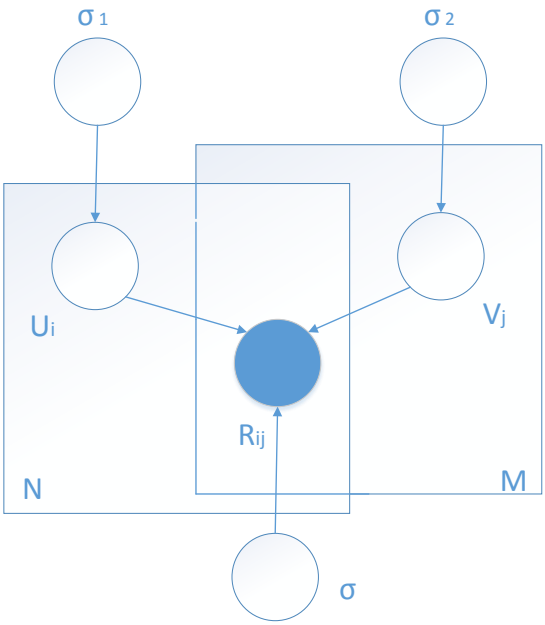
$\Delta^t = \|Y \odot (R - U^T V)\|_2^2$

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

For i from 1 to N, j from 1 to M：

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

### 3.1.2 $l_1$与$l_2$正则矩阵分解 RMF 及 SGD&ALS 算法




图示 14 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，以及它们与 PMF 建模的联系。我们定义矩阵 Y，其中的每个元素$\mathrm{Y_{ij}} = \delta(\mathrm{i},\mathrm{j})$指示$\mathrm{R_{ij}}$是否被观察，定义运算符⊙为 Hadamard 积满足$\mathrm{A_{n*m}} \odot \mathrm{B}_{n*m} = \left[\mathrm{a_{ij}b_{ij}}\right]_{n*m}$，定义矩阵的$l_1$范数满足$\|\mathrm{A}\|_1 = \sum_{ij}|a_{ij}|$，$l_2$范数满足$\|\mathrm{A}\|_2 = (\sum_{ij} a_{ij}^2)^{1/2}$，定义矩阵的原子模(nuclear-norm)是它的奇异值(singular-value)的和即$\|\mathrm{A}\|_* = \sum_{\mathrm{i}} \sigma_i(A)$，其中$\sigma_i(A)$表示 A 的第 i 大的奇异值。定义$\mathrm{U} = [\mathrm{U}_1 \ldots \mathrm{U_N}]$，$\mathrm{V} = [\mathrm{V}_1 \ldots \mathrm{V_M}]$。假设特征$\mathrm{U_i} \sim \mathrm{N}(0, \sigma_1^2 I)$，$\mathrm{V_j} \sim \mathrm{N}(0, \sigma_2^2 I)$，其中 i=1…N，j=1…M，$\mathrm{U_i}$和$\mathrm{V_j}$是 d 维向量服从多元高斯分布。

若假设观察值$\mathrm{R_{ij}} \sim N(\mathrm{U_i^T V_j}, \sigma^2)$，解决这个问题的变分近似算法(VB-EM 算法)在上一小节最后讨论的情形①中已经给出，变分近似分布中$\Sigma^{\mathrm{i}}, \Sigma^{\sim \mathrm{j}}$为非对角，而所有的$\Phi^{\mathrm{i}}$与$\Phi^{\sim \mathrm{j}}$与$\mu_1, \mu_2$均设置成 0 即可。而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：

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

给定参数$\lambda_U$和$\lambda_V$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$\mathrm{U_i}$和$\mathrm{V_j}$。这个问题通常又称作“最大边际矩阵分解”(max-margin matrix factorization)$\mathrm{M^3F}$，它解决大规模不完全矩阵的低秩近似问题，如协同过滤和图像恢复。

类似地，若假设$\mathrm{R_{ij}} \sim L(\mathrm{U_i^T V_j}, \sigma)$，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]

$$\min_{\mathrm{U,V}} \|\mathrm{Y} \odot (\mathrm{R} - \mathrm{U^T V})\|_1 + \lambda_U \|\mathrm{U}\|_2^2 + \lambda_V \|\mathrm{V}\|_2^2$$，其中$\lambda_U = \frac{2\sigma}{\sigma_U^2}, \lambda_V = \frac{2\sigma}{\sigma_V^2}$

我们通常又把它称作“鲁棒矩阵分解”(robust matrix factorization)或“稀疏低秩矩阵分解”(sparse low-rank matrix factorization)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$\epsilon$，初始化 U，V。SGD 需

[2]由于篇幅限制，求解这个问题的相关算法可详见相关资料。

要学习率η。ALS 是 Ridge 回归问题。

算法 2 RMF(SGD 算法)

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η，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

Else： $\eta = 0.1 * \eta$

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

算法 3 RMF(ALS 算法)

### 3.1.3 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$p(U_i|\mu_1,\Sigma_1)\Pi_{j=1}^{M}p\left(R_{ij}\middle|U_i^TV_j,\sigma^2\right)^{\delta(i,j)}$，这个分布仍然是多元高斯分布，即有：

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

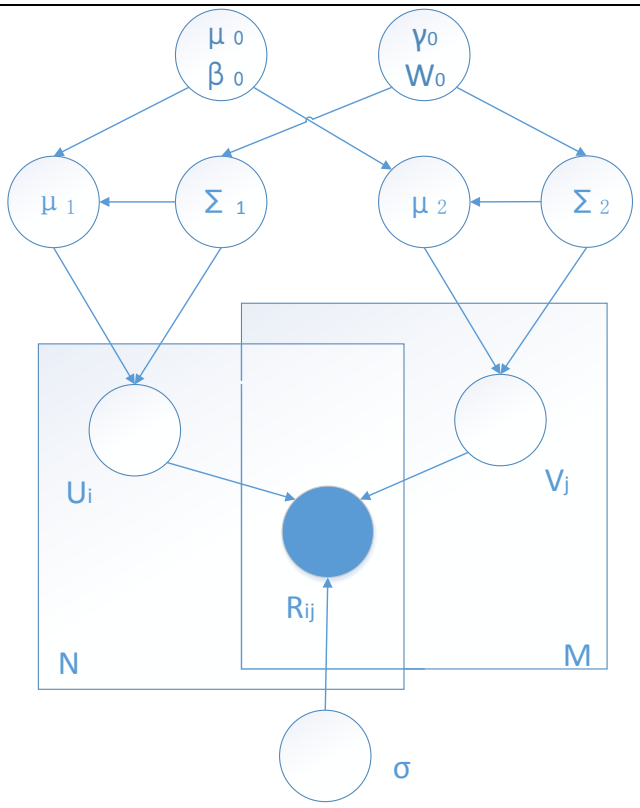
$$\Sigma_i^* = \left(\Sigma_1^{-1} + \frac{1}{\sigma^2}\sum_{j=1}^{M}\delta(i,j)V_jV_j^T\right)^{-1}$$

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

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。




图示 15　贝叶斯概率矩阵分解的图模型表示

上图是该模型的图表示，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$\beta_0$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：

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

$$= p(\mu_1, \Sigma_1|\mu_0, \beta_0, \gamma_0, W_0)\Pi_{i=1}^{N} p(U_i|\mu_1, \Sigma_1)\Pi_{j=1}^{M} p(V_j|\mu_2, \Sigma_2)\Pi_{i=1}^{N}\Pi_{j=1}^{M} p(R_{ij}|U_i^T V_j, \sigma^2)^{\delta(i,j)}$$

$\mu_1, \Sigma_1$的后验分布同比于其中包含$\mu_1, \Sigma_1$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$\mu_1, \Sigma_1$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：

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

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

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

对$\mu_2, \Sigma_2$的后验分布推导则与此类似。而对$U_i$和$V_j$的采样则与“MCMC-EM”算法中的 E 步相同。

算法 4 BPMF(MCMC 算法)

输出：U，V。

输入：指示矩阵Y，观察值矩阵 R，收敛率$\epsilon$，最大迭代次数 S，固定值$\sigma$，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 $\Sigma_1^{-1} \sim \text{Wishart}(W_0^*, \gamma_0^*)$

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

For each i from 1 to N：

compute $\mu_i^*, \Sigma_i^*$

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

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

sample $\Sigma_2^{-1} \sim \text{Wishart}(W_0^*, \gamma_0^*)$

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

For each i from 1 to M：

compute $\mu_{\sim j}^*, \Sigma_{\sim j}^*$

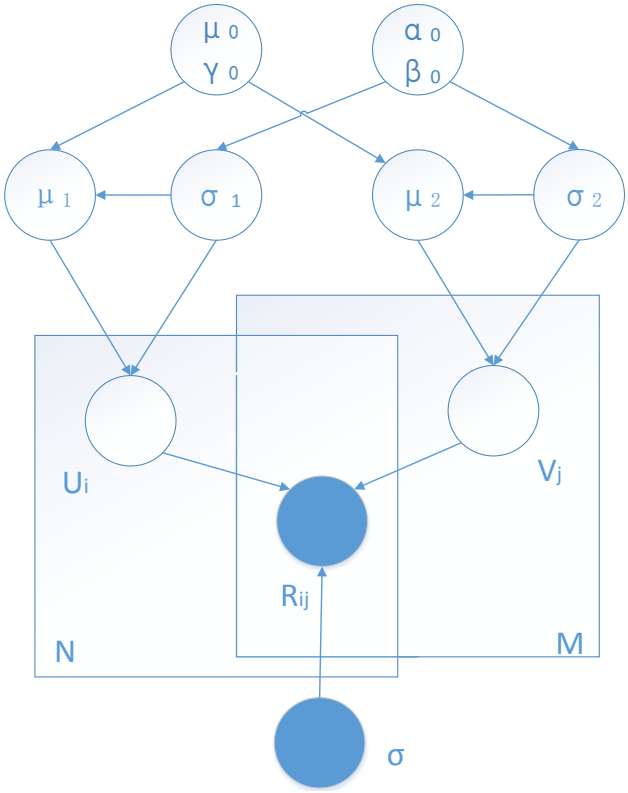
sample $V_j \sim N(\mu_{\sim j}^*, \Sigma_{\sim j}^*)$

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

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

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)，这里给出一个简化版本。

### 3.1.4 贝叶斯概率因子分解机 FM 及 MCMC 算法




图示 16　贝叶斯概率因子分解机 FM 的图模型表示

上图是该模型的图表示，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$\Gamma$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$\mu_1, \sigma_1^2$的后验分布，同比例于下式：

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

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：

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

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

可见，该后验分布仍然是“高斯-逆伽马分布”

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

$$\gamma_0^*=N+\gamma_0$$

$$\mu_0^*=\frac{\gamma_0\mu_0+\sum_{i=1}^{N}U_i}{N+\gamma_0}$$

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

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

$$\alpha_0^*=\alpha_0+dN$$

由于在$\mu_1,\Sigma_1=\sigma_1^2 I$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$\mu_2,\Sigma_2$则与此类似，给出算法流程如下：

算法 5 MF(MCMC 算法)

输出：U，V。

输入：指示矩阵Y，观察值矩阵 R，收敛率$\epsilon$，最大迭代次数 S，固定值$\sigma$，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 $\sigma_1^{-1} \sim \Gamma(\alpha_0^*, \beta_0^*)$

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

For each i from 1 to N：

compute $\mu_i^*, \Sigma_i^*$

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

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

sample $\sigma_2^{-1} \sim \Gamma(\alpha_0^*, \beta_0^*)$

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

For each i from 1 to M：

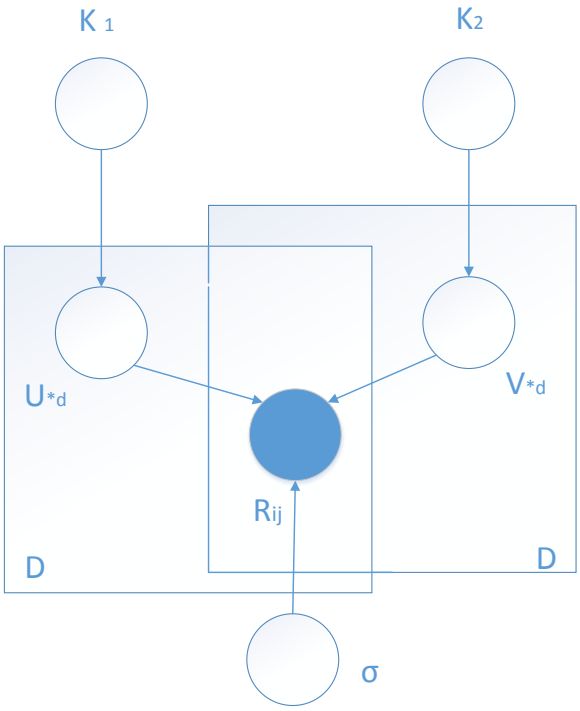
compute $\mu_{\sim j}^*, \Sigma_{\sim j}^*$

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

### 3.1.5 核概率矩阵分解 KPMF&社会化约束矩阵分解 SMF 相关算法




图示 17　核概率矩阵分解的图模型表示

在解决协同过滤问题中，通常利用用户之间的相似度量或社会关系图来提高 MF 的效果，这样的模型叫作社会化矩阵分解(Social MF)。这一小节将首先探讨核概率矩阵分解(kernel PMF)，我们将在稍后看到 SMF 是 KPMF 的一个特例。

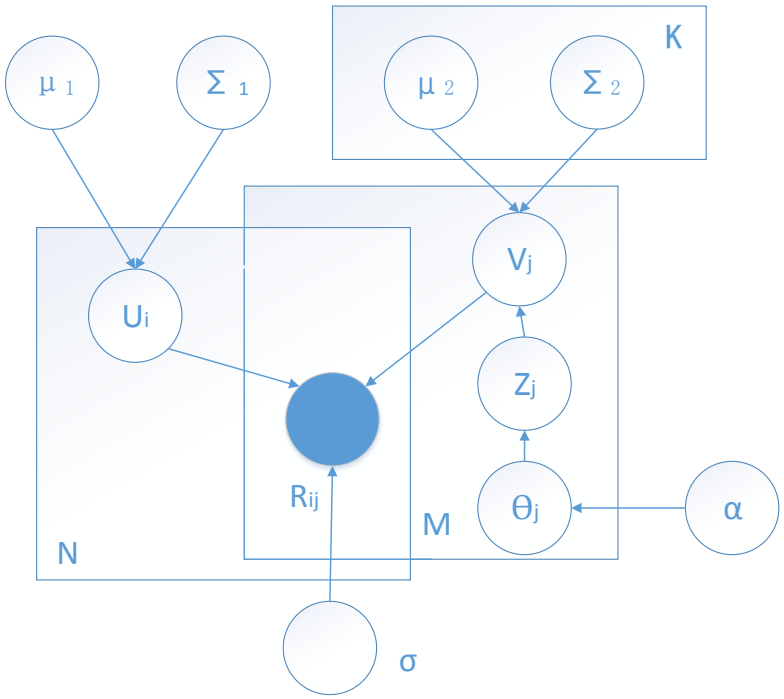
KPMF 的图模型表示如上图，记$U = [U_1 \dots U_N]$，$V = [V_1 \dots V_M]$，$U_i$和$V_j$是 d 维向量服从多元高斯分布。与 PMF 不同之处在于它假设 U 和 V 矩阵的每行从 0 均值的高斯过程(guassian process)GP 中产生，记为：

$$U_{d*} \sim GP(0, K_1), d = 1 \dots D,$$
$$V_{d*} \sim GP(0, K_2), d = 1 \dots D,$$

其中$K_1$是 N 维矩阵，$K_2$是 M 维矩阵。

高斯过程 GP 作为U 和 V的先验，

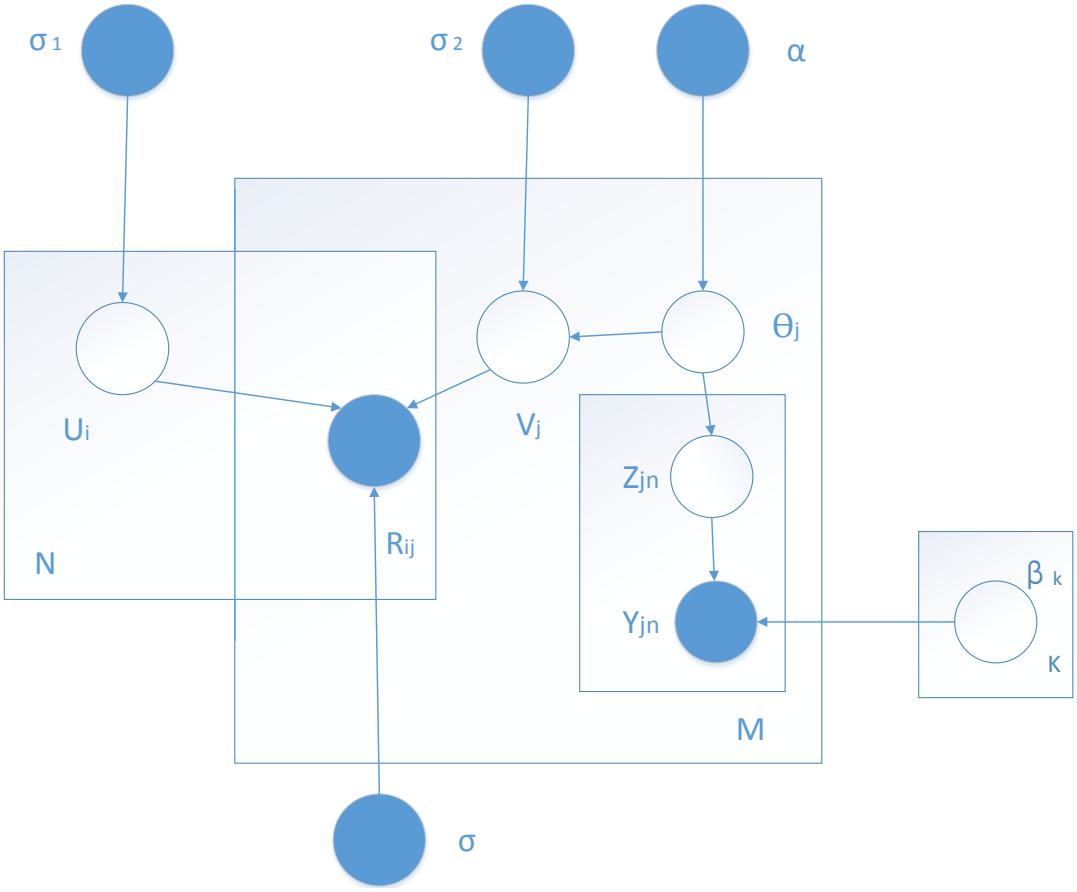
### 3.1.6 混合参数概率矩阵分解 MPMF 及 VB-EM&MCMC 算法




图示 18　混合参数概率矩阵分解 MPMF 的图模型表示

# 第4章 多关系模型

## 4. 1. 1 融入辅助信息协同主题回归模型 CTR 及其优化算法




图示 19 融入辅助信息协同话题回归模型 CTR 的图模型表示

融入辅助信息协同主题回归模型(collaborative topic regression)如上图所示，模型组合了概率矩阵分解 PMF 和主题模型 LDA，这种将文本信息融入到特征中的做法，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：

$\mathrm{L}(\boldsymbol{\theta},\boldsymbol{\beta},\mathrm{U},\mathrm{V}) = -\frac{\lambda_u}{2}\sum_{\mathrm{i}=1}^{\mathrm{N}}\mathrm{U}_\mathrm{i}^T\mathrm{U}_\mathrm{i} - \frac{\lambda_v}{2}\sum_{\mathrm{j}=1}^{\mathrm{M}}(\mathrm{V}_\mathrm{j}-\theta_\mathrm{j})^{\mathrm{T}}(\mathrm{V}_\mathrm{j}-\theta_\mathrm{j}) - \sum_{ij}\frac{Y_{ij}}{2}(\mathrm{R}_{\mathrm{ij}}-\mathrm{U}_\mathrm{i}^T\mathrm{V}_\mathrm{j})^2 + \sum_{\mathrm{j}=1}^{\mathrm{M}}\sum_{\mathrm{n}=1}^{\mathrm{N_j}}\log\sum_\mathrm{k}\theta_{jk}\beta_{kY_{jn}}$ ，其中$\mathrm{Y}_{\mathrm{ij}} = \delta(i,j)$。优化这个目标函数，对$\mathrm{U_i},\mathrm{V_j}$求导并置导数为 0，类似于 RMF(ALS)有：

$$\mathrm{U_i} = (\mathrm{Vdiag}(\mathrm{Y_{i*}})\mathrm{V}^{\mathrm{T}} + \lambda_\mathrm{U}\mathrm{I})^{-1}\mathrm{Vdiag}(\mathrm{Y_{i*}})\mathrm{R}_{\mathrm{i*}}^T$$

$$\mathrm{V_j} = \left(\mathrm{Udiag}(\mathrm{Y_{*j}})\mathrm{U}^{\mathrm{T}} + \lambda_\mathrm{V}\mathrm{I}\right)^{-1}\left(\mathrm{Udiag}(\mathrm{Y_{*j}})\mathrm{R_{*j}} + \lambda_\mathrm{V}\theta_\mathrm{j}\right)$$

其中$\mathrm{Y_{*j}}$取 Y 矩阵的第 j 列，$\mathrm{Y_{i*}}$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，

$$\log\sum_\mathrm{k}\frac{\theta_{jk}\beta_{kY_{jn}}}{\phi_{jnk}}\phi_{jnk} \geq \sum_\mathrm{k}(\phi_{jnk}\log\theta_{jk}\beta_{kY_{jn}} - \phi_{jnk}\log\phi_{jnk})$$

其中$\sum_\mathrm{k}\phi_{jnk} = 1$，对 L 中包含$\theta_\mathrm{j}$的项：

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

由于加入$\theta_\mathrm{j}$的拉格朗日约束无法得到解析解，故计算梯度如下：

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

得到$\theta_\mathrm{j} = \theta_\mathrm{j} + \eta\frac{\partial\mathrm{L}(\theta_\mathrm{j},\phi_j)}{\partial\theta_j}$，$\eta$是步长，但$\theta_\mathrm{j}$满足单纯形$\sum_\mathrm{k}\theta_\mathrm{jk} = 1, \theta_\mathrm{jk} > 0$约束，故做投影，即求解下式[4]：

$$\min_\mathrm{v}\frac{1}{2}\left\|\theta_\mathrm{j}-\mathrm{v}\right\|_2^2, \text{ 其中 v 满足}\sum_\mathrm{k}\mathrm{v_k} = 1, \mathrm{v_k} > 0。$$

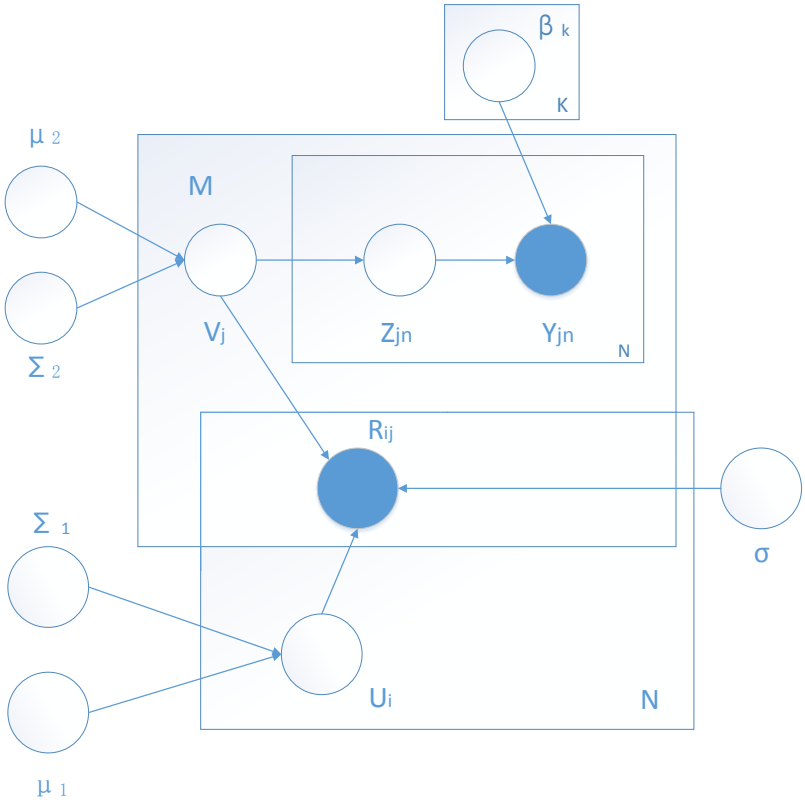
然后更新$\theta_\mathrm{j}$= v。最后，最优$\phi_{jnk}$满足：$\phi_{jnk} \propto \theta_{jk}\beta_{kY_{jn}}$，

类似 LDA 中 EM 步骤，最优$\beta_{kv}$满足：$\beta_{kv} \propto \sum_{\mathrm{j,n}}\phi_{jnk}\,\delta(\mathrm{Y_{jn}} = \mathrm{v})$。

组合 LDA 与 PMF 模型，利用$\mathrm{V_j}\sim\mathrm{N}(\theta_\mathrm{j},\sigma_2^2\mathrm{I})$并置外层超参数给定是 CTR 模型的主要思想，考虑到 CTM 模型中文本主题分布参数来自高斯罗杰斯特分布，这就很自然地将多元高斯分布产生的变量既当做 LDA 中文本主题参数又作为 PMF 中的特征因子，这个模型就是下面介绍的 PMF-CTM 模型。

[4] 约束投影优化问题将在附录中给出

### 4.1.2 融入辅助信息概率矩阵分解 PMF-CTM 及 VB-EM 算法




图示 20　融入辅助信息概率矩阵分解 PMF-CTM 的图模型表示

PMF 与 CTM 的联合分布分别为：

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

$$p(\mathbf{Y},\boldsymbol{Z},\mathbf{V}|\boldsymbol{\beta},\mu,\sigma) = \Pi_{j=1}^{M} p(V_j|\mu_2,\Sigma_2)\Pi_{n=1}^{Nd}\{p(Z_{jn}|V_j)p(Y_{jn}|Z_{jn},\boldsymbol{\beta})\} \quad (CTM)$$

PMF-CTM 的联合分布如下，是 PTM 与 CTM 的乘积并去掉重复的$\Pi_{j=1}^{M} p(V_j|\mu_2,\Sigma_2)$

$$\Pi_{i=1}^{N} p(U_i|\mu_1,\Sigma_1)\Pi_{i=1}^{N}\Pi_{j=1}^{M} p(R_{ij}|U_i^T V_j,\sigma^2)^{\delta(i,j)}\Pi_{j=1}^{M}\left\{p(V_j|\mu_2,\Sigma_2)\Pi_{n=1}^{Nd}\{p(Z_{jn}|V_j)p(Y_{jn}|Z_{jn},\boldsymbol{\beta})\}\right\}$$

其中，$p(Z_{jn}=k|V_j) = \exp(V_{j_k})/\Sigma_{i=1}^{K}\exp(V_{j_k})$，PMF-CTM 模型的 VB-EM 求解需要假设变分近似分布，采用完全可分解的分布$q(U_i) = N(U_i|\Phi^i,\Sigma^i)$，$q(V_j) = N(V_j|\Phi^{\sim j},\Sigma^{\sim j})$，$Z_{dn}\sim Mult(\phi_{dn})$，其中协方差矩阵$\Sigma^i = diag(\gamma_1^i \ldots \gamma_d^i), \Sigma^{\sim j} = diag(\gamma_1^{\sim j} \ldots \gamma_d^{\sim j})$均为对角矩阵。

对$\mu_1, \Sigma_1, \sigma^2$和所有$\Phi^{\mathrm{i}}, \Sigma^{\mathrm{i}}$的更新与 PMF 中情形②相同，这是因为在其它参数固定情况下，包含这部分量的联合似然与 PMF 相同。而对$\phi_{dn}, \boldsymbol{\beta}, \mu_2, \Sigma_2, \Phi^{\sim \mathrm{j}}, \Sigma^{\sim \mathrm{j}}$的更新主要考虑联合似然中的：

$$\Pi_{\mathrm{i}=1}^{\mathrm{N}} \Pi_{\mathrm{j}=1}^{\mathrm{M}} \mathrm{p}\left(\mathrm{R}_{\mathrm{ij}} \middle| \mathrm{U}_{\mathrm{i}}^{T} \mathrm{V}_{\mathrm{j}}, \sigma^2\right)^{\delta(\mathrm{i},\mathrm{j})} \Pi_{\mathrm{j}=1}^{\mathrm{M}} \left\{ \mathrm{p}\left(\mathrm{V}_{\mathrm{j}} \middle| \mu_2, \Sigma_2\right) \Pi_{\mathrm{n}=1}^{\mathrm{Nd}} \left\{ \mathrm{p}\left(\mathrm{Z}_{\mathrm{jn}} \middle| \mathrm{V}_{\mathrm{j}}\right) \mathrm{p}\left(\mathrm{Y}_{\mathrm{jn}} \middle| \mathrm{Z}_{\mathrm{jn}}, \boldsymbol{\beta}\right) \right\} \right\}$$

这一项的后一部分是 CTM 联合似然，因此对$\mu_2, \Sigma_2, \phi_{dn}, \boldsymbol{\beta}$的更新与 CTM 一致。这样只需要考虑包含$\mathrm{V}_{\mathrm{j}}$公共部分的推断，涉及到对$\Phi^{\sim \mathrm{j}}, \Sigma^{\sim \mathrm{j}}$的更新。这部分的 log 似然为：$\sum_{\mathrm{i}=1}^{\mathrm{N}} \delta(\mathrm{i},\mathrm{j}) \log\ \mathrm{p}\left(\mathrm{R}_{\mathrm{ij}} \middle| \mathrm{U}_{\mathrm{i}}^{T} \mathrm{V}_{\mathrm{j}}, \sigma^2\right) + \log \mathrm{p}\left(\mathrm{V}_{\mathrm{j}} \middle| \mu_2, \Sigma_2\right) + \sum_{\mathrm{n}=1}^{\mathrm{Nd}} \log \mathrm{p}\left(\mathrm{Z}_{\mathrm{jn}} \middle| \mathrm{V}_{\mathrm{j}}\right)$，为了得到它的 LB 下界，需要对上式在变分近似分布下求期望，

$$E_{q(\mathrm{U}_{\mathrm{i}}) q(\mathrm{V}_{\mathrm{j}})} \log \mathrm{p}\left(\mathrm{R}_{\mathrm{ij}} \middle| \mathrm{U}_{\mathrm{i}}^{T} \mathrm{V}_{\mathrm{j}}, \sigma^2\right) = \frac{1}{2} \log \frac{1}{2\pi\sigma^2} - \frac{1}{2\sigma^2} \left(\mathrm{R}_{\mathrm{ij}}^2 - 2\mathrm{R}_{\mathrm{ij}} \Phi^{\mathrm{i}^{\mathrm{T}}} \Phi^{\sim \mathrm{j}} + \mathrm{tr}(\Sigma^{\mathrm{i}} + \Phi^{\mathrm{i}} \Phi^{\mathrm{i}^{T}})(\Sigma^{\sim \mathrm{j}} + \Phi^{\sim \mathrm{j}} \Phi^{\sim \mathrm{j}^{T}})\right)$$

$$E_{q(\mathrm{Z},\mathrm{V}_{\mathrm{j}})} \log \mathrm{p}\left(\mathrm{V}_{\mathrm{j}} \middle| \mu_2, \Sigma_2\right) = \phi_{\mathrm{gaus}}(\mu_2, \Sigma_2)^{T} E_{q(\mathrm{V}_{\mathrm{j}})} \mathrm{u}_{\mathrm{gaus}}\left(\mathrm{V}_{\mathrm{j}}\right) + \mathrm{g}_{\mathrm{gaus}}(\mu_2, \Sigma_2)$$

$$= -\frac{1}{2} \mathrm{tr}\left(\Sigma_2^{-1} \left(\boldsymbol{\gamma}^{\sim j^2} - \mu_2 \Phi^{\sim \mathrm{j}^{T}} - \Phi^{\sim \mathrm{j}} \mu_2^{\mathrm{T}} + \mu_2 \mu_2^{\mathrm{T}} + \Phi^{\sim \mathrm{j}} \Phi^{\sim \mathrm{j}^{T}}\right)\right) + \log \sqrt{\frac{|\Sigma_2^{-1}|}{(2\pi)^K}}$$

$$E_{q(\mathrm{Z},\mathrm{V}_{\mathrm{j}})} \Sigma_{\mathrm{n}=1}^{\mathrm{Nd}} \log \mathrm{p}\left(\mathrm{Z}_{\mathrm{jn}} \middle| \mathrm{V}_{\mathrm{j}}\right) = E_{q(\mathrm{V}_{\mathrm{j}})} \phi_{\mathrm{Mult}} \left( \begin{pmatrix} \exp(\mathrm{V}_{\mathrm{j}_1}) / \Sigma_{\mathrm{i}=1}^{\mathrm{K}} \exp(\mathrm{V}_{\mathrm{j}_{\mathrm{i}}}) \\ \cdots \cdots \\ \exp(\mathrm{V}_{\mathrm{j}_K}) / \Sigma_{\mathrm{i}=1}^{\mathrm{K}} \exp(\mathrm{V}_{\mathrm{j}_{\mathrm{i}}}) \end{pmatrix} \right)^{T} \Sigma_{\mathrm{n}=1}^{\mathrm{Nd}} E_{q(\mathrm{Z}_{\mathrm{n}})} \mathrm{u}_{\mathrm{Mult}}\left(\mathrm{Z}_{\mathrm{jn}}\right)$$

$$\geq \begin{pmatrix} \Phi^{\sim \mathrm{j}}{}_1 - \zeta^{-1} \left(\Sigma_{\mathrm{i}=1}^{\mathrm{K}} \exp\{\Phi^{\sim \mathrm{j}}{}_{\mathrm{i}} + \gamma_{\mathrm{i}}^{\sim \mathrm{j}^2} / 2\}\right) + 1 - \log \zeta \\ \cdots \cdots \\ \Phi^{\sim \mathrm{j}}{}_{\mathrm{K}} - \zeta^{-1} \left(\Sigma_{\mathrm{i}=1}^{\mathrm{K}} \exp\{\Phi^{\sim \mathrm{j}}{}_{\mathrm{i}} + \gamma_{\mathrm{i}}^{\sim \mathrm{j}^2} / 2\}\right) + 1 - \log \zeta \end{pmatrix}^{T} \begin{pmatrix} \Sigma_{\mathrm{n}=1}^{\mathrm{Nd}} \phi_{\mathrm{n}1} \\ \cdots \cdots \\ \Sigma_{\mathrm{n}=1}^{\mathrm{Nd}} \phi_{\mathrm{nK}} \end{pmatrix}$$

$$H\left(q\left(\mathrm{V}_{\mathrm{j}}\right)\right) = -\frac{1}{2} \log \frac{1}{(2\pi)^{\mathrm{K}} |\Sigma^{\sim \mathrm{j}}|} - \frac{K}{2}$$

所以，我们得到了包含$\mathrm{V}_{\mathrm{j}}$部分(涉及到$\Phi^{\sim \mathrm{j}}, \Sigma^{\sim \mathrm{j}}$)的下界：$\mathrm{LB}(\mathrm{V}_{\mathrm{j}})$

$$\geq -\frac{1}{2\sigma^2} \Sigma_{\mathrm{i}=1}^{\mathrm{N}} \delta(\mathrm{i},\mathrm{j}) \left(\mathrm{R}_{\mathrm{ij}}^2 - 2\mathrm{R}_{\mathrm{ij}} \Phi^{\mathrm{i}^{\mathrm{T}}} \Phi^{\sim \mathrm{j}} + \mathrm{tr}\left(\Sigma^{\mathrm{i}} + \Phi^{\mathrm{i}} \Phi^{\mathrm{i}^{T}}\right)\left(\boldsymbol{\gamma}^{\sim j^2} + \Phi^{\sim \mathrm{j}} \Phi^{\sim \mathrm{j}^{T}}\right)\right)$$

$$-\frac{1}{2} \mathrm{tr}\left(\Sigma_2^{-1} \left(\boldsymbol{\gamma}^{\sim j^2} - \mu_2 \Phi^{\sim \mathrm{j}^{T}} - \Phi^{\sim \mathrm{j}} \mu_2^{\mathrm{T}} + \mu_2 \mu_2^{\mathrm{T}} + \Phi^{\sim \mathrm{j}} \Phi^{\sim \mathrm{j}^{T}}\right)\right) +$$

$$\begin{pmatrix} \Phi^{\sim j}{}_1 - \zeta^{-1}\left(\Sigma_{i=1}^{K}\exp\left\{\Phi^{\sim j}{}_i + \frac{\gamma_i^{\sim j^2}}{2}\right\}\right) + 1 - \log\zeta \\ \cdots\cdots \\ \Phi^{\sim j}{}_K - \zeta^{-1}\left(\Sigma_{i=1}^{K}\exp\left\{\Phi^{\sim j}{}_i + \frac{\gamma_i^{\sim j^2}}{2}\right\}\right) + 1 - \log\zeta \end{pmatrix}^{T} \begin{pmatrix} \Sigma_{n=1}^{Nd}\phi_{n1} \\ \cdots\cdots \\ \Sigma_{n=1}^{Nd}\phi_{nK} \end{pmatrix}$$

$$-\frac{1}{2}\log\frac{1}{(2\pi)^K\left|\boldsymbol{\gamma}^{\sim j^2}\right|}$$

E-step：
$$\gamma_s^{i\,2} = \left(\frac{1}{\sigma^2}\sum_{j=1}^{M}\delta(i,j)\left(\gamma_s^{\sim j^2} + \Phi_s^{\sim j^2}\right) + \Sigma_{1,ss}^{-1}\right)^{-1}$$

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

$$\frac{\partial LB(V_j)}{\partial\Phi^{\sim j}} = \frac{1}{\sigma^2}\sum_{i=1}^{N}\delta(i,j)R_{ij}\Phi^i - (\Sigma_2^{-1} - \frac{1}{\sigma^2}\sum_{i=1}^{N}\delta(i,j)(\Sigma^i + \Phi^i\Phi^{iT}))\Phi^{\sim j} + \Sigma_2^{-1}\mu_2$$

$$+ \Sigma_{n=1}^{Nd}\phi_{n1:K} - \frac{N}{\zeta}\left\{\exp\left(\Phi^{\sim j}{}_s + \gamma_s^{\sim j^2}/2\right)\right\}_{s=1:K}$$

$$\frac{\partial LB(V_j)}{\partial\gamma_s^{\sim j^2}} = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}\delta(I,j)\left(\Sigma^i + \Phi^i\Phi^{iT}\right)_{ss} - \frac{\Sigma_2^{-1}{}_{ss}}{2} - \frac{N}{2\zeta}\exp(\Phi^{\sim j}{}_s + \frac{\gamma_s^{\sim j^2}}{2})$$

$$+ \frac{1}{\left(2\gamma_s^{\sim j^2}\right)}$$

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”。

$\phi_{jns} \propto \beta_{sY_{jn}}\exp(\Phi^{\sim j}{}_s)$，$\Sigma_{s=1}^{K}\phi_{jns} = 1$。

$$\zeta = \Sigma_{s=1}^{K}\exp(\Phi^{\sim j}{}_s + \frac{\gamma_s^{\sim j^2}}{2})$$

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

$$\mu_1 = \frac{1}{N}\sum_{i=1}^{N}\Phi^i$$

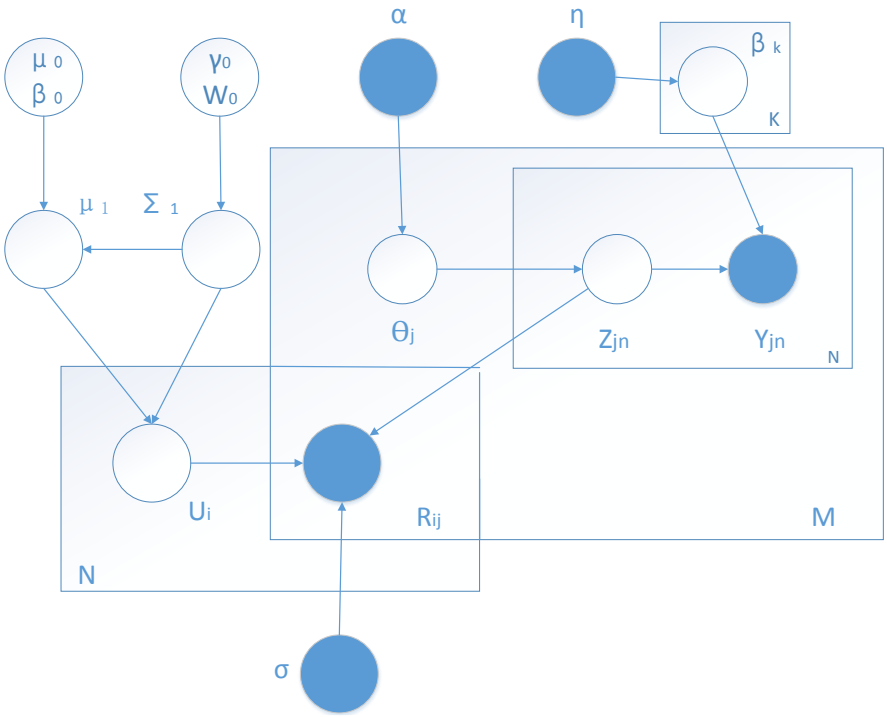
$$\Sigma_1 = \frac{1}{N}\Sigma_{i=1}^{N}(\Sigma^i + \left(\Phi^i - \mu_1\right)\left(\Phi^i - \mu_1\right)^T)$$

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

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

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

### 4.1.3 融入辅助信息概率矩阵分解 PMF-LDA 及 MCMC 算法




图示 21　融入辅助信息概率矩阵分解 PMF-LDA 的图模型表示

CTR 利用高斯噪声(最小二乘)假设将文本主题分布参数和矩阵分解特征结合在一起，PMF-CTM 中利用逻杰斯特变换将矩阵分解特征转变为主题分布参数。而 PMF-LDA 模型，直接将第 j 个文本的主题均值作为矩阵分解特征。即假设观察值$R_{ij}$满足$R_{ij} \sim N(U_i^T\left(\Sigma_{n=1}^{N_j} Z_{jn} \Big/ N_j\right), \sigma^2)$，这里若$Z_{jn} = k$则把它看作仅有第 k 个分量为 1 其它全为 0 的向量。这个模型利用 Gibbs 采样求解，需要写出完全条件分布。

考虑在 **Z** 已知情况下，对$U_i, \mu_1, \Sigma_1$的采样以及参数$\mu_0, \beta_0, \gamma_0, W_0$的更新则类似于 BPMF(MCMC)，而在$U_i$已知情况下，$Z_{jn}$的条件分布同比例于联合分布中包含$Z_{jn}$的部分，若采用 collapsed-LDA 相同的处理方法，可得到$Z_{jn}$的条件分布如下：

$$p(Z_{jn}=k|\boldsymbol{Z}^{\neg(j,n)})$$

$$\propto \frac{\eta_{Y_{jn}}+\#\{\cdot,Y_{jn},k\}^{\neg(j,n)}}{\sum_{v=1}^{V}\eta_v+\#\{\cdot,\cdot,k\}^{\neg(j,n)}}\frac{\alpha_k+\#\{j,\cdot,k\}^{\neg(j,n)}}{\sum_{k=1}^{K}\alpha_k+\#\{j,\cdot,\cdot\}^{\neg(j,n)}}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\delta(i,j)\left(R_{ij}-\frac{U_i^T\sum_{t\neq n}^{N_j}Z_{jt}}{N_j}-\frac{U_i^T Z_{jn}}{N_j}\right)^2\right)$$

利用这个式子可完成对所有$Z_{jn}$的采样，

算法 6 PMF-LDA(MCMC 算法)

输出：U，V。

输入：指示矩阵Y，观察值矩阵 R，收敛率$\epsilon$，最大迭代次数 S，固定值$\sigma$，初始化$\mu_0,\beta_0,\gamma_0,W_0,U,\boldsymbol{Z},\alpha,\eta$。

For t =1…S：

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

sample $\Sigma_1^{-1}\sim Wishart(W_0^*,\gamma_0^*)$

sample $\mu_1\sim N(\mu_0^*,\beta_0^{*-1}\Sigma_1)$

For each i from 1 to N：

compute $\mu_i^*,\Sigma_i^*$

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

For each j from 1 to M：

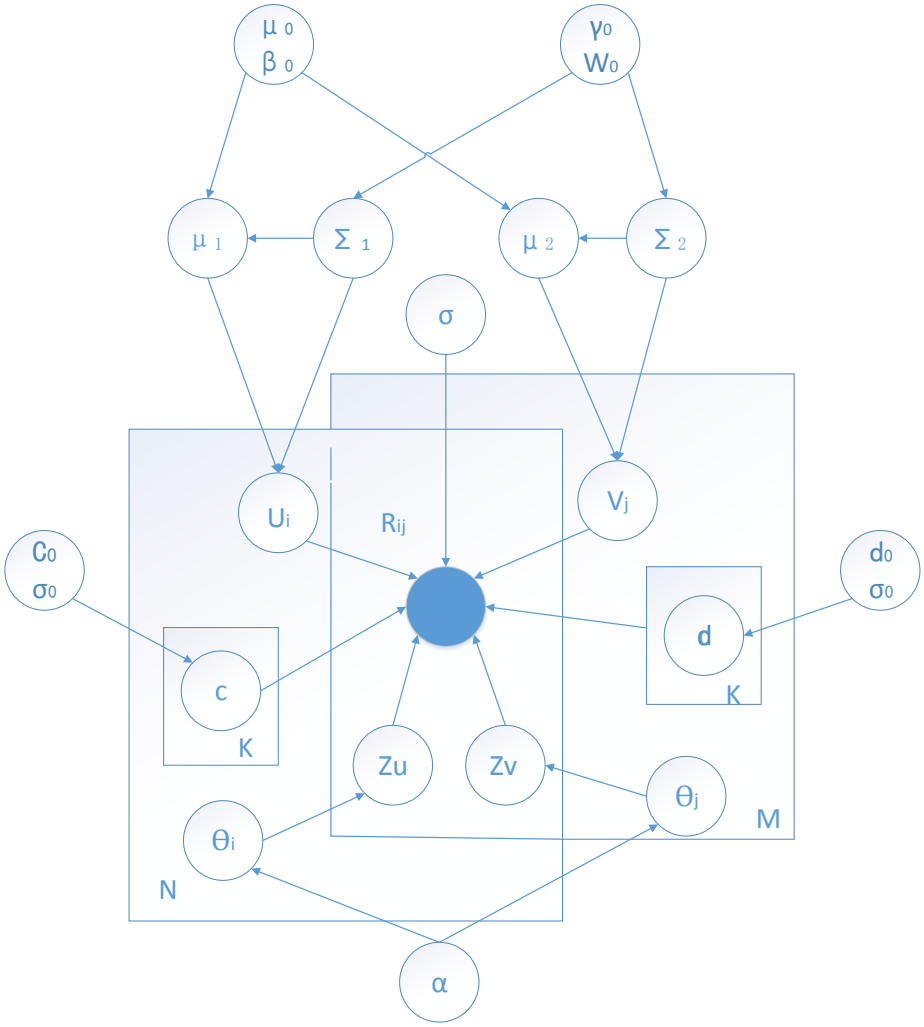
sample each $Z_{jn}$ for n from 1 to $N_j$

let $V_j=\sum_{n=1}^{N_j}Z_{jn}/N_j$

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

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

### 4. 1. 4 混合成员关系概率矩阵分解 MMMF 及 MCMC 算法




图示 22　混合成员关系概率矩阵分解 MMMF 的图模型表示