密级：_____

浙江大學

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 2014 年 01 月 06 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: January 06, 2014

摘要

随着现代存储和通信技术的发展，存储不断增长的大规模数据已经不再是难事，人们迫切地想从这些数据中获得有用的信息，因而如何有效地处理复杂数据并做有效的建模分析已成为计算机科学亟待解决的问题。例如，电子商务网站的数据库中既包含某件商品的文本描述、图片又包含用户评论和购买信息，对这些多关系数据建模分析是机器学习和数据挖掘领域的热点。

概率模型方法被广泛用于文本分类、信息检索、文本和图像标注、协同过滤、链接预测等问题。这些方法通常假设一个概率模型来刻画数据的生成过程，通过最大化似然概率得到模型参数的估计值，进而利用模型参数对新的数据做预测。本文的主要工作包括：

系统地总结了两种基本模型，混合主题模型和因子分解模型的相关解法。混合主题模型是一种聚类模型，它可以分析隐主题以及对单词和图像做标注。因子分解模型的经典例子是概率矩阵分解，通过假设矩阵元素由交互的因子内积，以达到对不完全观察矩阵填充的目的，常用于评分预测协同过滤、图像恢复、背景提取等问题。

提出基于混合主题模型和因子分解模型的多关系模型，以处理同时具有链接信息与辅助信息的数据。本文有效地利用上述基本模型的解法，给出了三种建模方法和对应求解算法。在公开数据集上的实验表明融入辅助信息的协同过滤模型有更好的预测精度。

关键词： 概率图模型 混合主题模型 因子分解 协同过滤 关系数据挖掘

Abstract

As the modern storage and communication technologies evolve, it has never been a hard commission to store and manage large growing dataset. But, people urge to know the important information from the large dataset, therefore building effective models to process complex dataset is becoming a critical problem in computer science. For example, a database of an electronic website frequently contains an item's text description、images、users' comments and purchase records.

Probabilistic modeling methods are prevailing used to handle these tasks such as text classification、information retrieval、text and image annotation、collaborative filtering、link prediction etc. These methods commonly assume a probabilistic model to describe the producing procedure of the dataset, the estimate values of the parameters are finally obtained by maximizing the likelihood function, so as to apply this results to make predictions for the new coming data. We summary our work as follows:

We at first introduce two basic models, topic models and factor decomposition models. Topic model as a kind of clustering model, is used to analysis latent topics and make annotations to word and image's codeword. The classic example of factor decomposition model is probabilistic matrix factorization, to assume the element of a matrix is produced by two mutual factor's inner product, the final recovered matrix could be used in collaborative filtering and image background extraction.

we propose the multi-relational models based on these basic models. For example, combine topic model and factor decomposition model to deal with a dataset containing multiple information. The experiments on some public dataset reveal that our collaborative filtering models exploiting auxiliary information achieve better performance.

Keywords： probabilistic graphical models, mixture topic model, factor decomposition, collaborative filtering, relational data-mining

目录

摘要 i

Abstract ii

第 1 章 绪论 8

1.1 研究背景 8

1.2 研究现状 9

1.3 本文工作与组织架构 10

第 2 章 混合主题模型与因子分解模型 11

2.1 混合主题模型 11

2.1.1 LDA 模型 12

2.1.2 CTM 模型 14

2.1.3 对称先验 LDA 17

2.2 因子分解模型 21

2.2.1 概率矩阵分解 PMF 及 VB-EM 算法 21

2.2.2 $l1$ 与$l2$ 正则矩阵分解 RMF 及 SGD&ALS 算法 25

2.2.3 贝叶斯概率矩阵分解 BPMF 及 MCMC 算法 28

2.2.4 贝叶斯概率因子分解机 FM 及 MCMC 算法 32

2.3 本章小结 34

第 3 章 多关系模型 35

3.1 主要模型 35

3.1.1 融入辅助信息协同主题回归模型 CTR 及其优化算法 35

3.1.2 融入辅助信息概率矩阵分解 PMF-CTM 及 VB-EM 算法 37

3.1.3 融入辅助信息概率矩阵分解 PMF-LDA 及 MCMC 算法 41

3.2 本章小结 43

第 4 章 实验结果与比较 44

4.1 混合主题模型实验 44
4.1.1 数据集 44
4.1.2 评价指标与实验结果 44
4.2 因子分解与多关系模型实验比较 50
4.2.1 数据集 50
4.2.2 评价指标与实验结果 50
4.3 本章小结 51
第 5 章 总结和展望 52
5.1 总结 52
5.2 展望 52
附录 54
参考文献 62
攻读硕士学位期间主要的研究成果 67
致谢 68

图示

图示 1-1 本文算法衍生关系图 10
图示 2-1 几个基本主题模型衍生关系图 11
图示 2-2 LDA 的图模型表示 12
图示 2-3 $q(\mathbf{Z}, \boldsymbol{\theta})$近似分布的图模型表示和分布表 13
图示 2-4 CTM(correlated topic model)的图模型表示 15
图示 2-5 罗杰斯特高斯分布图 15
图示 2-6 Smooth-LDA 或“对称先验 LDA”的图模型表示 17
图示 2-7 概率矩阵分解 PMF(左)变分分布假设(右)的图模型表示 21
图示 2-8 PMF 简化情形下的图模型表示 25
图示 2-9 贝叶斯概率矩阵分解的图模型表示 29
图示 2-10 贝叶斯概率因子分解机 FM 的图模型表示 32
图示 3-1 融入辅助信息协同话题回归模型 CTR 的图模型表示 35
图示 3-2 融入辅助信息概率矩阵分解 PMF-CTM 的图模型表示 37
图示 3-3 融入辅助信息概率矩阵分解 PMF-LDA 的图模型表示 41

算法

算法 2-1 LDA 的 VB-EM 算法 14
算法 2-2 CTM 的 VB-EM 算法 16
算法 2-3 对称先验 LDA 的 VB-EM 算法 18
算法 2-4 对称先验 LDA 的 MCMC 算法① 19
算法 2-5 对称先验 LDA 的 MCMC 算法② 20
算法 2-6 PMF(VB-EM 算法) 24
算法 2-7 RMF(SGD 算法) 27
算法 2-8 RMF(ALS 算法) 27
算法 2-9 BPMF(MCMC 算法) 31
算法 2-10 MF(MCMC 算法) 34
算法 3-1 PMF-CTM(VB-EM 算法) 36
算法 3-2 PMF(VB-EM 算法) 40
算法 3-3 PMF-LDA(MCMC 算法) 42

表格

表格 4-1 AP 与 Nips 数据集 44
表格 4-2 hetrec2011-2k 数据集 44
表格 4-3 AP 数据集上 K=5 时概率最大的前 20 个单词 45
表格 4-4 Nips 数据集前 1000 文本上 K=5 时概率最大的前 20 个单词 ... 45
表格 4-5 主题模型复杂度随话题参数变化图 47
表格 4-6 Lasfm 数据集上标签推荐的平均召回率 48
表格 4-7 Delicious 数据集上标签推荐的平均召回率 48
表格 4-8 Lastfm 数据集上项目推荐的平均召回率 49
表格 4-9 Delicious 数据集上项目推荐的平均召回率 49
表格 4-10 Netflix 数据集上 RMSE 结果 50
表格 4-11 Movielens 数据集上 RMSE 结果 51

第1章 绪论

1.1 研究背景

当今的互联网已经进入 Web2.0 时代，一个网站可能同时包含文字、图片、链接、音乐等信息，并为用户提供社交、购物、通讯等多种服务，在互联网公司的服务器中存储大量的复杂数据和历史信息。以人工智能的观点，我们寄希望于计算机可以有效地提取出文本、图片的有效特征，同时依据已有特征的标注信息完成模型训练，进而对新的内容做预测以达到自动分析文本和图片的目的，提高搜索引擎的效率。我们也希望计算机可以依据用户的历史信息，为用户推荐新的朋友、商品、音乐、新闻、链接等。这迫切地需要有效的建模方法来处理日益增长的多关系型数据。因此，对这些多关系数据做有效的建模分析和用户行为预测已成为计算机科学与统计学、管理科学最为重要的研究热点，这一学科就是机器学习与数据挖掘。统计学习是机器学习理论发展过程中最为主流的方法，概率图模型与贝叶斯学习是其中的一种，它通常假设一个概率模型来刻画数据的相互关系，通过最大化似然函数得到模型参数的估计值。概率模型建模方法已经在人工智能诸多领域得到了广泛的应用。

例如，混合主题模型，更为广义的是多项式混合模型，因其有效的无监督学习特性，是文本分析和高维数据聚类最为常用的建模方法。对单词做聚类、抽取隐主题、计算文档相似度、文档分类、自动关键词标注，这些自然语言处理中的常见问题都可以利用主题模型的建模方法加以解决。

例如，因子分解模型，通常又具体到矩阵分解、立方分解、张量分解等模型，对动态背景图像抽取、图片恢复、图像着色，这些计算机视觉中的常见问题都可以利用因子分解模型的建模方法加以解决。另外，对协同过滤、社交链接预测等推荐系统问题，也常利用因子分解模型做建模。社会网络建模工作中的谱图理论，常利用矩阵计算原理，典型的例子是利用社交网络做谱聚类，因此这一类模型具有较高的应用建模价值。

1.2 研究现状

本文主要研究两种概率模型，它们分别是混合主题模型与因子分解模型。概率模型广泛应用在数据挖掘等相关领域中，是机器学习领域重要的建模方法[4,5,6]。对结构化数据的概率建模方法，通常用层次贝叶斯图模型表示。推断这种图模型的一般性方法是：变分近似推理[2,5]与蒙特卡洛采样[3,7]。

混合主题模型的典型例子是 PLSA[8,9]，它的变形模型广泛用于文本分析中[10]。LDA 模型[11,12]和 CTM 模型[13]为 PLSA 增加了外层先验，是它的层级贝叶斯推广。本文中 LDA 模型引述了两种版本，求解它的方法有变分近似推断算法[12,14]和 Gibbs 采样算法[15,16,20]以及置信传播算法[17]。除此以外，LDA 的监督学习算法[18]和在线学习算法[19]都扩展了主题模型的用途和训练速率。国内外的自然语言处理研究者，也从模型功能和算法训练速率两个方面对主题模型做了深入的研究，例如文本挖掘模型与文档关键词抽取[21,22]，以及结合大边界学习提高主题模型监督学习能力[23,24]，还有非参数贝叶斯模型[25,26,27,28]、社会化标签推荐[29]、图像标注与分类[30]，这些理论和方法，为主题混合模型构建了一套完整的机器学习体系，扩展了它的用途并提高了求解效率。

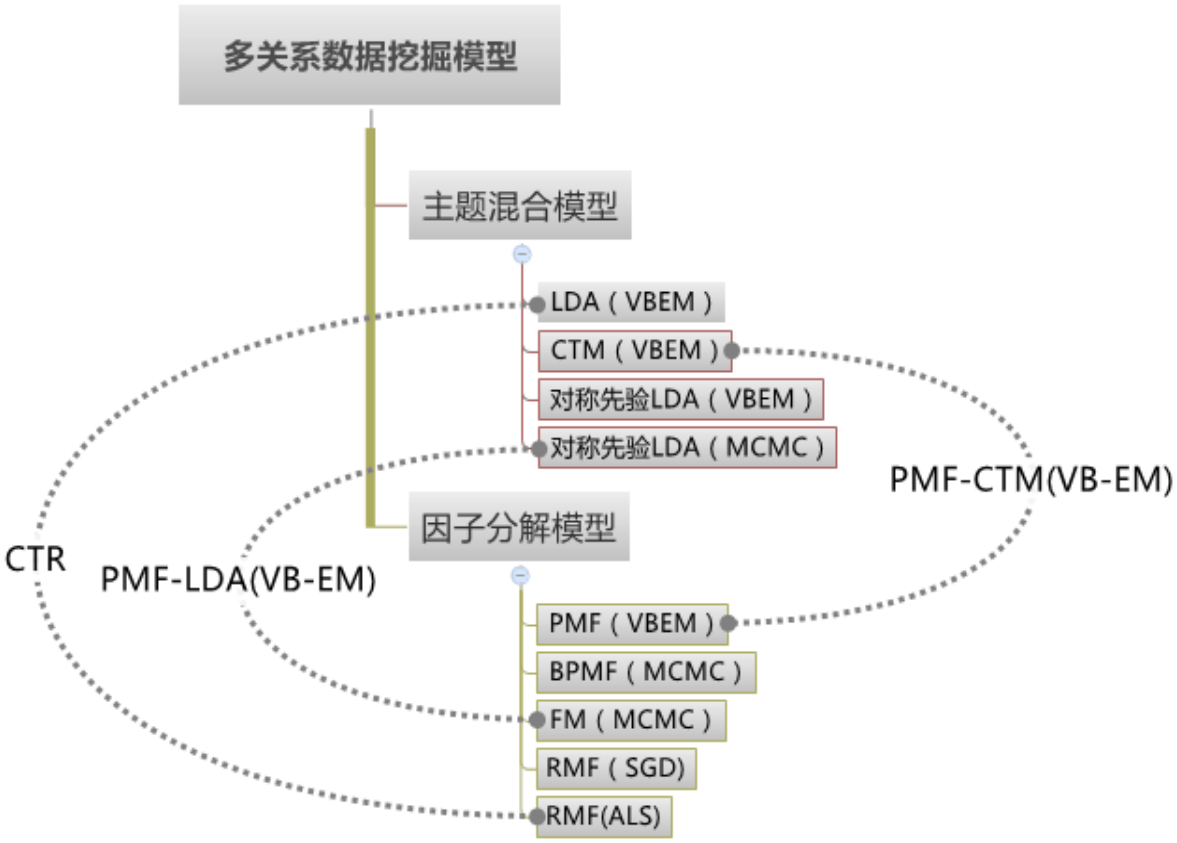
因子分解模型的典型例子是低秩矩阵分解[31,32]，理论上它们等价于概率矩阵分解模型[33,34,35]，都可以用来解决大规模不完全观察矩阵的填充问题。矩阵分解用于关系数据挖掘，典型的如社交网络挖掘[37,40,41]和协同过滤建模[36,38,39]。协同过滤方法是推荐系统建模的主要方法[42,43,44,45]，在最新的研究工作中，结合社交网络和文本辅助信息的协同过滤方法[46,47,48]相较于传统方法在预测精确度上有了显著的提高。很多研究工作者，为推荐系统开发了因子分解模型的工具集[49,50,51]，使得因子分解模型成为一套成熟的建模工具。另外，矩阵分解在计算机视觉中被用于背景图像抽取[52]、图片恢复[38]、图像着色[53]等。

最新的学术工作中，研究者们为因子分解模型融入辅助信息来提高协同过滤的效果。例如协同过滤问题中，利用高斯过程向用户因子融入社交信息[54]，利用主题模型向物品因子融入文本和属性信息[55,56,57,58]。而本文所提出的多关系模型，

正是受到这些研究工作的启发，做基于上述两种模型的组合方法，并给出一套完整的建模求解框架。

1.3 本文工作与组织架构

本文的多关系数据挖掘模型主要基于两种基本模型，第二章中详细论述主题混合模型与因子分解模型，在第三章中详细论述基于这两种基本模型的多关系模型。本文的组织架构与算法衍生关系如下图所示。

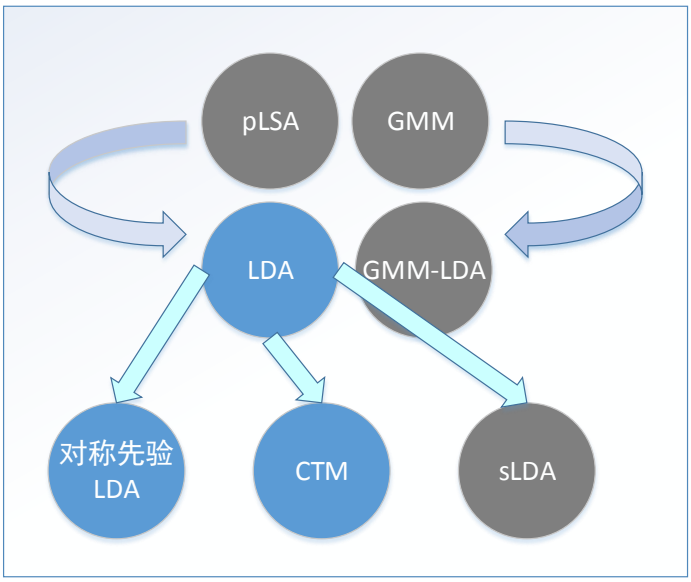


图示 1-1 本文算法衍生关系图

第四章中给出实验结果与比较。首先，在两个公开数据集上完成主题模型关键词抽取，在两个公开数据集上完成基于标签的推荐。其次在两个公开数据集上完成了因子分解模型和多关系模型的协同过滤实验。这些实验都取得了较为良好的效果。

第2章 混合主题模型与因子分解模型

2.1 混合主题模型

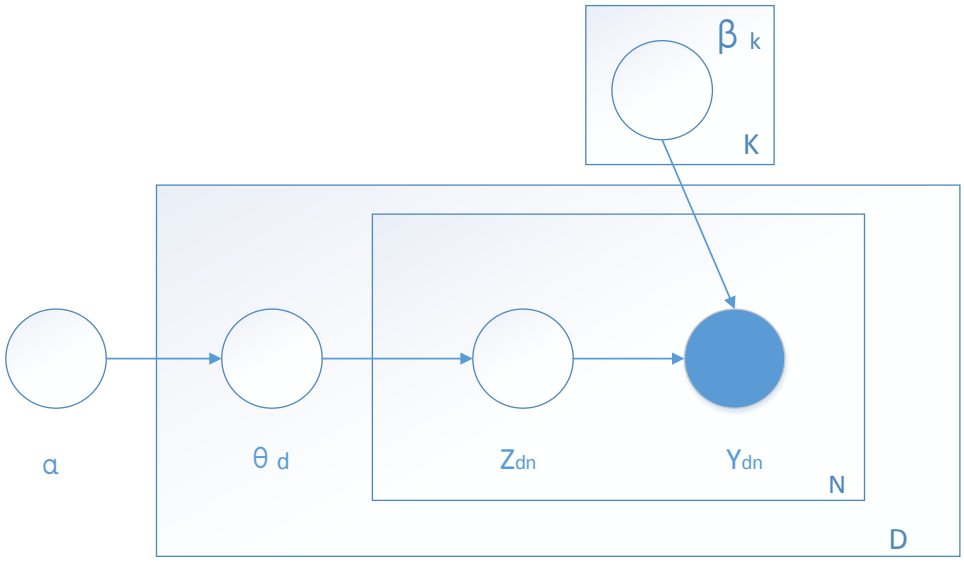


图示 2-1　几个基本主题模型衍生关系图

基本的混合主题模型，如概率隐语义分析(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)，即模型生成过程中选择外部参数族由多项式分布确定。前者多用于对文本中单词做隐含主题聚类，后者多用于对图像特征以及其它高维数据聚类。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。“LDA”是两层有向图模型，是 pLSA 的层级推广模型，GMM-LDA 与此类似，这些模型之间的衍生关系见图示 2-1。

由于篇幅所限，本章第一部分将主要介绍“LDA”与“对称先验 LDA”以及“CTM”模型。这部分内容将阐述上述模型的一般求解方法，即利用“变分近似 VBEM 推断法”和“蒙特卡洛 MCMC 采样法”最大化似然，前者在 E-M 算法框架下利用变分近似分布最大化下界，后者利用完全条件分布构造马尔科夫链进行循环采样。这些算法将在后续章节中被引用。

2.1.1 LDA 模型



图示 2-2　LDA 的图模型表示

Latent Dirichlet Allocation 模型如上图所示，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$\{Y_1 \dots\dots Y_n\}$都各自对应一个主题$\{Z_1 \dots\dots Z_n\}$，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~Mult($\beta_{z_i\cdot}$)。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$Y_{di} \in \{Y_{d1} \dots\dots Y_{dn}\}$：

(a) 从θ_d确定的 Multinomial 分布产生话题Z_{di}，即Z_{di}~Mult(θ_d)。

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

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：

$$\log p(\mathbf{Y}|\alpha,\boldsymbol{\beta}) = \log \int_{(\boldsymbol{Z},\boldsymbol{\theta})} p(\mathbf{Y},\boldsymbol{Z},\boldsymbol{\theta}|\boldsymbol{\beta},\alpha)\, d(\boldsymbol{Z},\boldsymbol{\theta}) \geq LB = E_{q(\mathbf{Z},\boldsymbol{\theta})}\log p(\mathbf{Y},\boldsymbol{Z},\boldsymbol{\theta}|\boldsymbol{\beta},\alpha) + H(q(\mathbf{Z},\boldsymbol{\theta})) \quad (2.1)$$

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)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。

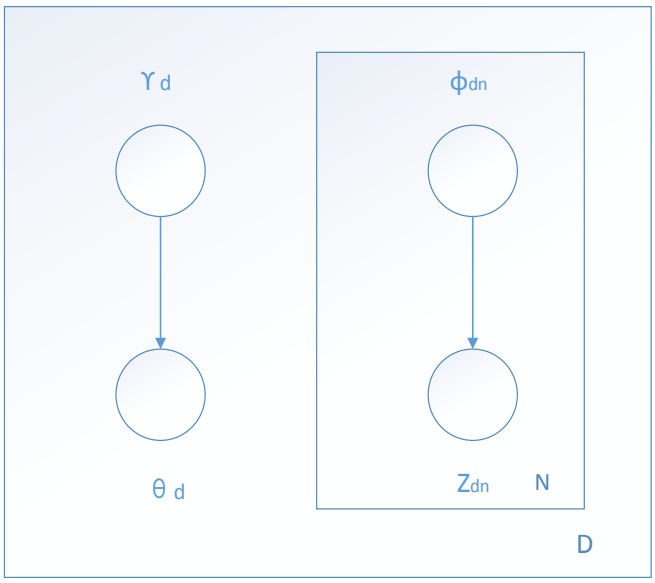
$$\log \mathrm{p}(\mathbf{Y},\boldsymbol{Z},\boldsymbol{\theta}|\boldsymbol{\beta},\alpha)=\Sigma_{\mathrm{d}=1}^{\mathrm{D}}\{\log \mathrm{p}(\theta_{\mathrm{d}}|\alpha)+\Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}\{\log \mathrm{p}(Z_{dn}|\theta_{\mathrm{d}})+\log \mathrm{p}(\mathrm{Y_{dn}}|\mathrm{Z_{dn}},\boldsymbol{\beta})\}\}\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

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

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

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

$+H(q(\boldsymbol{Z_d},\theta_d))$，则 LB=$E_{q(\mathbf{Z},\boldsymbol{\theta})}\log p(\mathbf{Y},\boldsymbol{Z},\boldsymbol{\theta}|\boldsymbol{\beta},\alpha)+\ H\big(q(\mathbf{Z},\boldsymbol{\theta})\big)=\Sigma_{\mathrm{d}=1}^{\mathrm{D}}LB_d$　（2.2）



VB 近似分布表

$\theta_{\mathrm{d}}\sim Dir(\gamma_d)$

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

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

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

$\mathrm{q}(\boldsymbol{\theta})=\Pi_{\mathrm{d}=1}^{\mathrm{D}}\mathrm{q}(\theta_{\mathrm{d}}|\gamma_{\mathrm{d}})$

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

图示 2-3　$q(\mathbf{Z},\boldsymbol{\theta})$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$\{\alpha,\boldsymbol{\beta}\}$。具体的 E-M 步骤推导可详见附录。[1]

[1] 文中加粗变量表示变量集合。

算法 2-1　LDA 的 VB-EM 算法[2]

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：

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

　　For n=1…N_d：

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

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

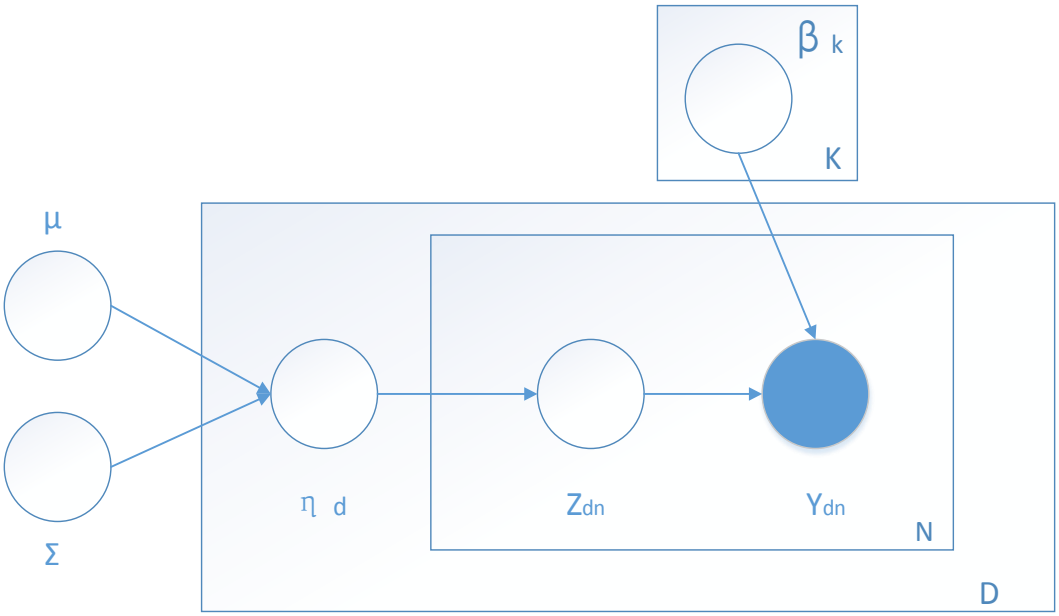
　Update α

　IF converge：break

2.1.2 CTM 模型

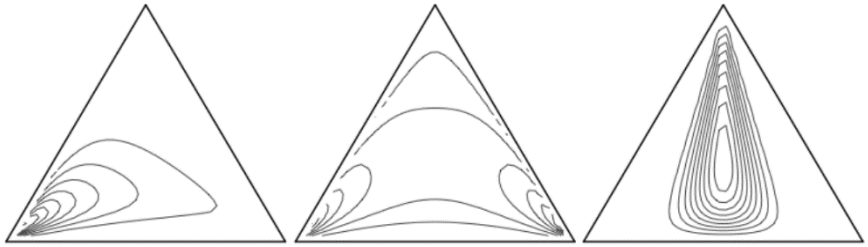
LDA 模型选择外层超参数分布主要基于 Dirichlet 是 Mult 的共轭先验，Dir 的选取有利于简化 VB 下界的计算，也满足主题之间的可交换性(可祥见 Dir 分布的相关性质)。但同时 Dir 不容易刻画话题之间的关联程度，为此引入罗杰斯特高斯分布(logistic normal distribution)，即 $\eta \sim N(\mu, \Sigma)$，$\theta_i = \exp(\eta_i)/\Sigma_{i=1}^{K}\exp(\eta_i)$，$Z_n \sim Mult(\theta)$，这个模型如下图所示是 CTM(correlated topic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

[2] α 的更新详见附录。



图示 2-4　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$\boldsymbol{\Sigma}$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 2-5　罗杰斯特高斯分布图

CTM 的 VB-EM 型近似推断，近似分布$q(\mathbf{Z},\boldsymbol{\eta})$将类似图示-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：

$$p(\mathbf{Y},\boldsymbol{Z},\boldsymbol{\eta}|\boldsymbol{\beta},\mu,\Lambda^{-1}) = \Pi_{d=1}^{D}p(\eta_d|\mu,\Lambda^{-1})\Pi_{n=1}^{Nd}\{p(Z_{dn}|\eta_d)p(Y_{dn}|Z_{dn},\boldsymbol{\beta})\} \quad (2.3)$$

采用近似推断方法来得到 log 后验的一个下界：

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

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

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

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

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

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

$H(q(\boldsymbol{Z_d},\eta_d))$，则有：

$$\mathrm{LB}=E_{q(\mathbf{Z},\boldsymbol{\eta})}\mathrm{p}(\mathbf{Y},\boldsymbol{Z},\boldsymbol{\eta}|\boldsymbol{\beta},\mu,\Lambda^{-1}) + H\big(q(\mathbf{Z},\boldsymbol{\eta})\big) = \Sigma_{\mathrm{d}=1}^{\mathrm{D}}LB_d \quad (2.4)$$

具体的 E-M 步骤推导可详见附录。

算法 2-2　CTM 的 VB-EM 算法[3]

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

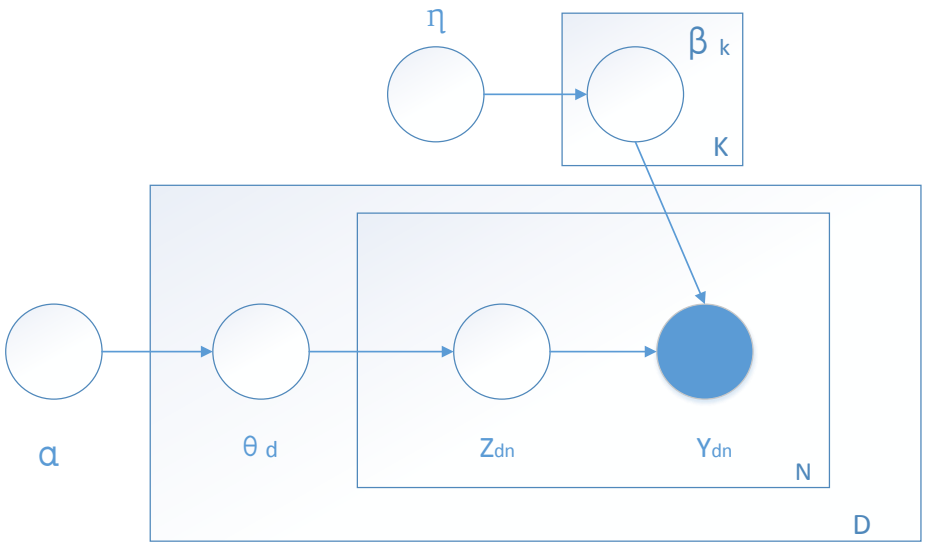
Update $\mu = \frac{1}{\mathrm{D}}\Sigma_{\mathrm{d}=1}^{D}\lambda_{\mathrm{d}}$，$\Lambda^{-1} = \frac{1}{\mathrm{D}}\Sigma_{\mathrm{d}=1}^{D}(\gamma_d^2 + (\lambda_d - \mu)(\lambda_d - \mu)^T)$

IF converge：break

[3] λ_d,γ_d^2的更新详见附录

2.1.3 对称先验 LDA

第 1 节的讨论中给出了 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 VB-EM 推断、MCMC 采样算法(Gibbs 采样算法)。



图示 2-6　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“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η，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：

$$p(\mathbf{Y},\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{\beta}|\alpha,\eta)=\Pi_{d=1}^{D}p(\theta_d|\alpha)\Pi_{n=1}^{Nd}\{p(Z_{dn}|\theta_d)p(Y_{dn}|Z_{dn},\boldsymbol{\beta})\}\Pi_{k=1}^{K}p(\beta_k|\eta) \quad (2.5)$$

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，$\log \mathrm{p}(\mathbf{Y},\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{\beta}|\alpha,\eta) = \Sigma_{\mathrm{d}=1}^{\mathrm{D}}\{\log \mathrm{p}(\theta_{\mathrm{d}}|\alpha) + \Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}\{\log \mathrm{p}(Z_{dn}|\theta_{\mathrm{d}}) + \log \mathrm{p}(\mathrm{Y_{dn}}|\mathrm{Z_{dn}},\boldsymbol{\beta})\}\} + \Sigma_{\mathrm{k}=1}^{\mathrm{K}} \log \mathrm{p}(\beta_{\mathrm{k}}|\eta)$ ，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$\mathrm{LB_d}$下界将会由 6 个 term 来表示。具体的 E-M 步骤详见附录。

算法 2-3　对称先验 LDA 的 VB-EM 算法[4]

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

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

For t =1…S：

For d=1…D：

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

For n=1…$\mathrm{N_d}$：

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

For k=1…K：　$\lambda_{\mathrm{ki}} = \eta_{\mathrm{i}} + \Sigma_{\mathrm{d}=1}^{\mathrm{D}}\Sigma_{\mathrm{n}=1}^{\mathrm{Nd}}\phi_{\mathrm{dnk}}\delta(\mathrm{Y_{dn}} = \mathrm{i})$

Update α，η

IF converge：break

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

$$\mathrm{p}(\theta_{\mathrm{d}}|\mathbf{Z},\alpha)，\ \mathrm{p}(\mathrm{Z_{dn}}|\mathbf{Y},\boldsymbol{\theta},\beta)，\ \mathrm{p}(\beta_{\mathrm{k}}|\boldsymbol{Z},\boldsymbol{Y},\eta)$$

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

[4] α，η的更新详见附录。

$$= \Pi_{k=1}^{K}\left(\theta_{dk}^{\alpha_k-1}\Pi_{n=1}^{N_d}\theta_{dk}^{\delta(Z_{dn}=k)}\right) = \Pi_{k=1}^{K}\left(\theta_{dk}^{\sum_{n=1}^{N_d}\delta(Z_{dn}=k)+\alpha_k-1}\right), \quad (2.6)$$

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

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

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

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

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

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

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。

算法 2-4　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：

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

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

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：

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

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

$$\mathrm{const}(\alpha,\eta)(\Pi_{\mathrm{k}=1}^{\mathrm{K}}\Pi_{\mathrm{i}=1}^{\mathrm{V}}\beta_{ki}^{\eta_i-1+\Sigma_{\mathrm{d}=1}^{\mathrm{D}}\Sigma_{\mathrm{n}=1}^{\mathrm{N_d}}\delta(Y_{dn}=i,Z_{dn}=k)})(\Pi_{\mathrm{d}=1}^{\mathrm{D}}\Pi_{\mathrm{k}=1}^{\mathrm{K}}\theta_{dk}^{\alpha_k-1+\Sigma_{\mathrm{n}=1}^{\mathrm{N_d}}\delta(Z_{dn}=k)}) \quad (2.7)$$

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，具体的推导可详见附录。

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

算法 2-5　对称先验 LDA 的 MCMC 算法②

输出：$\mathbf{Z}$。

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

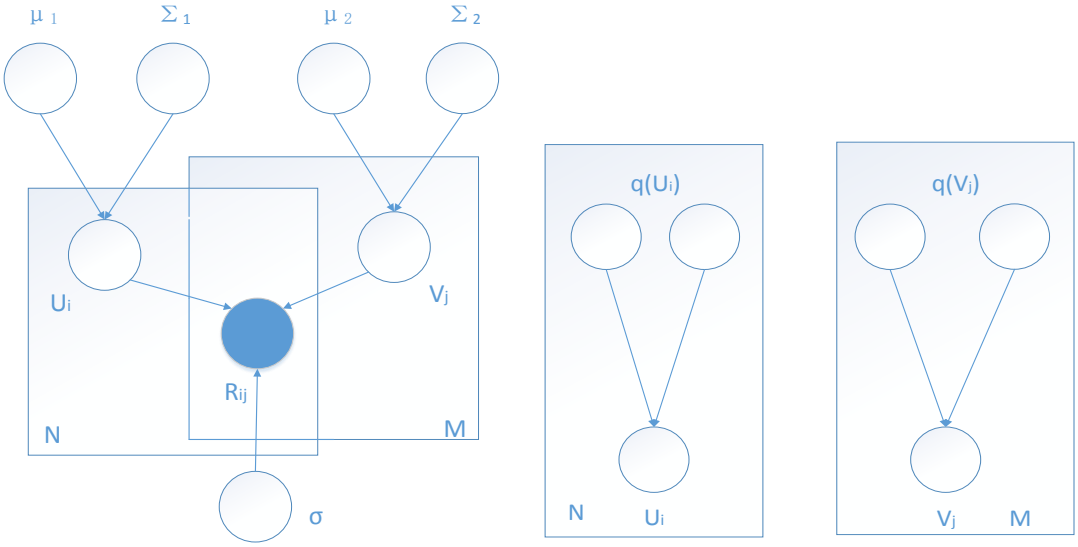
For t=1…S：

For d=1…D：

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

2.2 因子分解模型

2.2.1 概率矩阵分解 PMF 及 VB-EM 算法



图示 2-7　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：

$$p(\mathbf{R}, \mathbf{U}, \mathbf{V} | \boldsymbol{\Theta}) = \Pi_{i=1}^{N} p(U_i | \mu_1, \Sigma_1) \Pi_{j=1}^{M} p(V_j | \mu_2, \Sigma_2) \Pi_{i=1}^{N} \Pi_{j=1}^{M} p(R_{ij} | U_i^T V_j, \sigma^2)^{\delta(i,j)} \quad (2.8)$$

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 $q(V_j)$。

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

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

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

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

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

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

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

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

若令$q(\mathrm{U_i}) = N(U_i|\Phi^{\mathrm{i}},\Sigma^{\mathrm{i}})$，$q\big(\mathrm{V_j}\big) = N(\mathrm{V_j}|\Phi^{\sim\mathrm{j}},\Sigma^{\sim\mathrm{j}})$[5]

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

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

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

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

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

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

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

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

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

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

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：

[5] 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。

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

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

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

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

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

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

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

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

对其中的σ^2求导，并置导数为 0 得：

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

考虑μ_1,Σ_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μ_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Σ_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Σ_1,Σ_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$\Sigma_1=\sigma_1^2I,\Sigma_2=\sigma_2^2I$，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：　Φ^{i}与$\Phi^{\sim \mathrm{j}}$，Σ^{i}与$\Sigma^{\sim \mathrm{j}}$的更新不变。

M-step：　μ_1, μ_2, σ^2的更新不变。

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

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

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：　Φ^{i}与$\Phi^{\sim \mathrm{j}}$的更新不变

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

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

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

算法 2-6　PMF(VB-EM 算法)

输出：U，V。

输入：指示矩阵$\mathrm{Y_{ij}} = \delta(\mathrm{i}, \mathrm{j})$，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$\Sigma^{\mathrm{i}}$与 Φ^{i}(i=1…N)，$\Sigma^{\sim \mathrm{j}}$与 $\Phi^{\sim \mathrm{j}}$(j=1…M)。

For t =1…S：

For i from 1 to N, j from 1 to M：

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

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

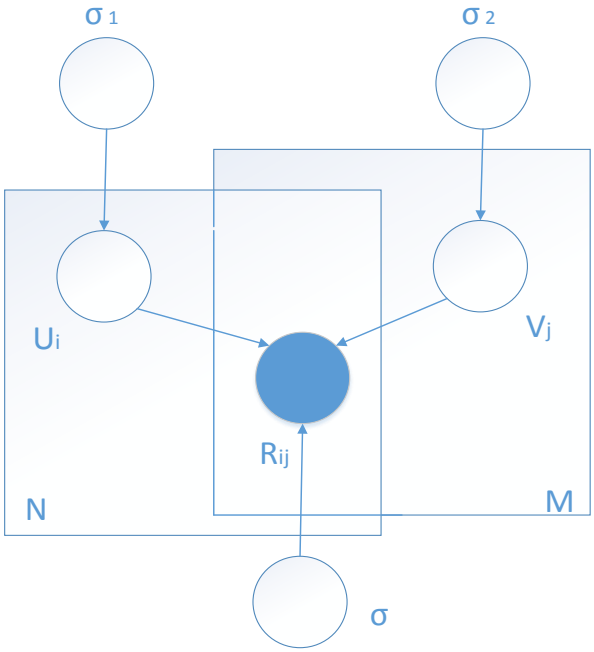
$\Delta^{\mathrm{t}} = \| \mathrm{Y} \odot (\mathrm{R} - \mathrm{U}^{\mathrm{T}} \mathrm{V}) \|_2^2$

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

For i from 1 to N, j from 1 to M：

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

2.2.2 l_1与l_2正则矩阵分解 RMF 及 SGD&ALS 算法



图示 2-8　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$Y_{ij} = \delta(i,j)$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$\|A\|_2 = (\sum_{ij} a_{ij}^2)^{1/2}$，定义矩阵的原子模(nuclear-norm)是它的奇异值(singular-value)的和即$\|A\|_* = \sum_i \sigma_i(A)$，其中$\sigma_i(A)$表示 A 的第 i 大的奇异值。定义$U = [U_1 \ldots U_N]$，$V = [V_1 \ldots V_M]$。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。

若假设观察值$R_{ij} \sim N(U_i^T V_j, \sigma^2)$，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$\Sigma^i, \Sigma^{\sim j}$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。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：

$\min_{U,V} \|Y \odot (R - U^T V)\|_2^2 + \lambda_U \|U\|_2^2 + \lambda_V \|V\|_2^2$，其中$\lambda_U = \frac{\sigma^2}{\sigma_U^2}, \lambda_V = \frac{\sigma^2}{\sigma_V^2}$。 （2.9）

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，如协同过滤和图像恢复。

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：[6]

$\min_{U,V} \|Y \odot (R - U^T V)\|_1 + \lambda_U \|U\|_2^2 + \lambda_V \|V\|_2^2$，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 需要学习率$\eta$。ALS 是 Ridge 回归问题。

[6]由于篇幅限制，求解这个问题的相关算法可详见相关资料。

算法 2-7　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$\eta$，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 $Y_{ij} \neq 0$：

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

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

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

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

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

算法 2-8　RMF(ALS 算法)

输出：U，V。

输入：指示矩阵Y，观察值矩阵 R，正则化参数$\lambda = \lambda_U = \lambda_V$，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

2.2.3 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[7]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：

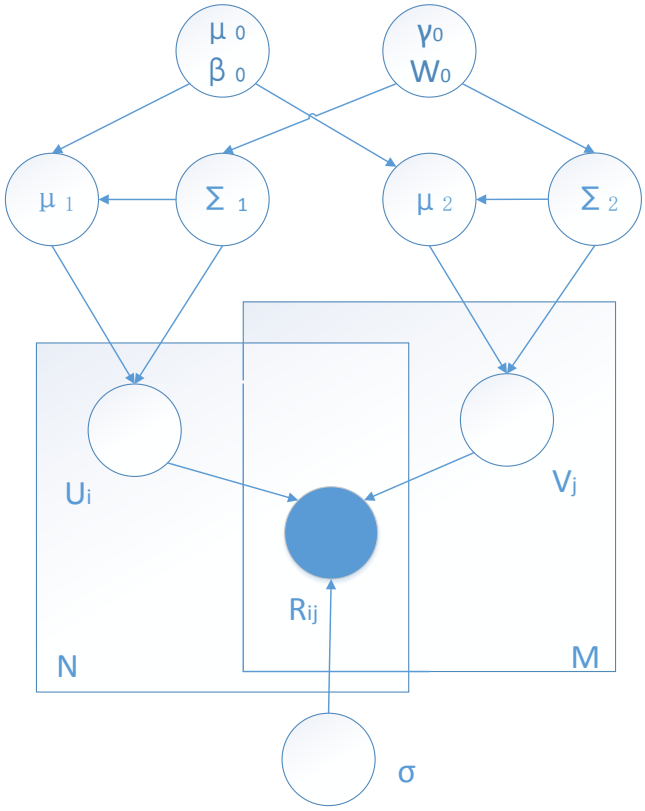
$\widehat{p(U_\iota)} \propto p(U_i|\mu_1,\Sigma_1)\Pi_{j=1}^{M}p\left(R_{ij}\middle|U_i^T V_j,\sigma^2\right)^{\delta(i,j)} \propto N(U_i|\mu_i^*,\Sigma_i^*)$，其中：

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

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

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”。

[7] 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 2-9　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：

$$p(\mathbf{R}, \mathbf{U}, \mathbf{V}|\mathbf{\Theta}, \mu_0, \beta_0, \gamma_0, W_0) = p(\mu_1, \Sigma_1|\mu_0, \beta_0, \gamma_0, W_0)$$

$$\Pi_{i=1}^{N} p(U_i|\mu_1, \Sigma_1)\Pi_{j=1}^{M} p(V_j|\mu_2, \Sigma_2)\Pi_{i=1}^{N}\Pi_{j=1}^{M} p(R_{ij}|U_i^T V_j, \sigma^2)^{\delta(i,j)} \quad (2.10)$$

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的采样则与“MCMC-EM”算法中的 E 步相同。

算法 2-9　BPMF(MCMC 算法)

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

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

For each i from 1 to N：

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

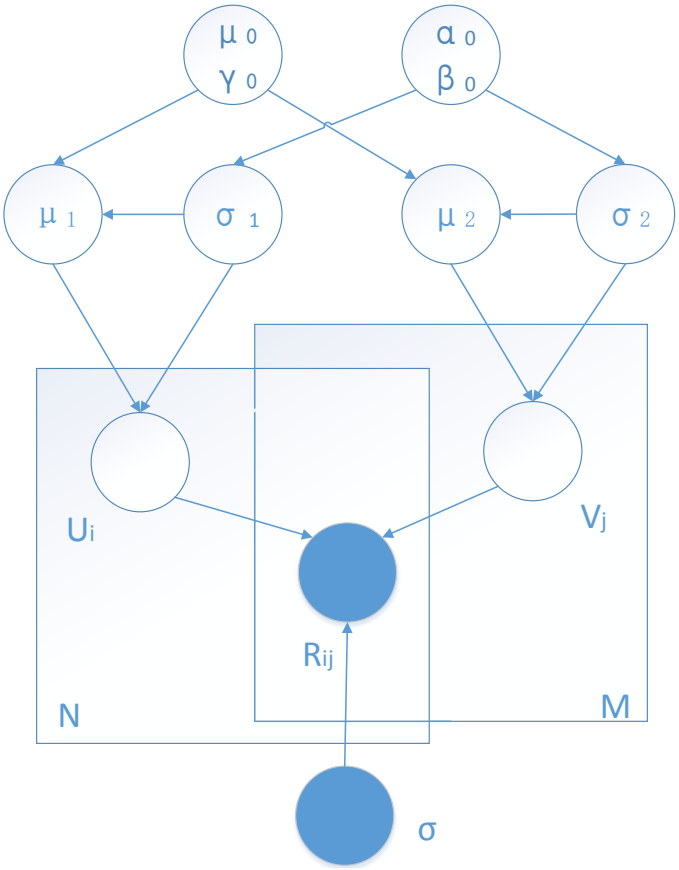
sample $V_j \sim N(\mu_{\sim j}^*, \Sigma_{\sim j}^*)$

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

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。

2.2.4 贝叶斯概率因子分解机 FM 及 MCMC 算法



图示 2-10　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γ_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：

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

可见，该后验分布仍然是“高斯-逆伽马分布”

$$p(\mu_1|\mu_0,\gamma_0,\sigma_1)p(\sigma_1^2|\alpha_0,\beta_0)\Pi_{i=1}^{N}p(U_i|\mu_1,\sigma_1)\propto N\left(\mu_1\middle|\mu_0^*,{\gamma_0^*}^{-1}\sigma_1^2 I\right)\Gamma\left({\sigma_1^2}^{-1}\middle|\alpha_0^*,\beta_0^*\right) \tag{2.11}$$

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

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

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

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

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μ_2,Σ_2则与此类似，给出算法流程如下：

算法 2-10 MF(MCMC 算法)

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 $\sigma_1^{-1} \sim \Gamma(\alpha_0^*, \beta_0^*)$

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

For each i from 1 to N：

compute μ_i^*, Σ_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

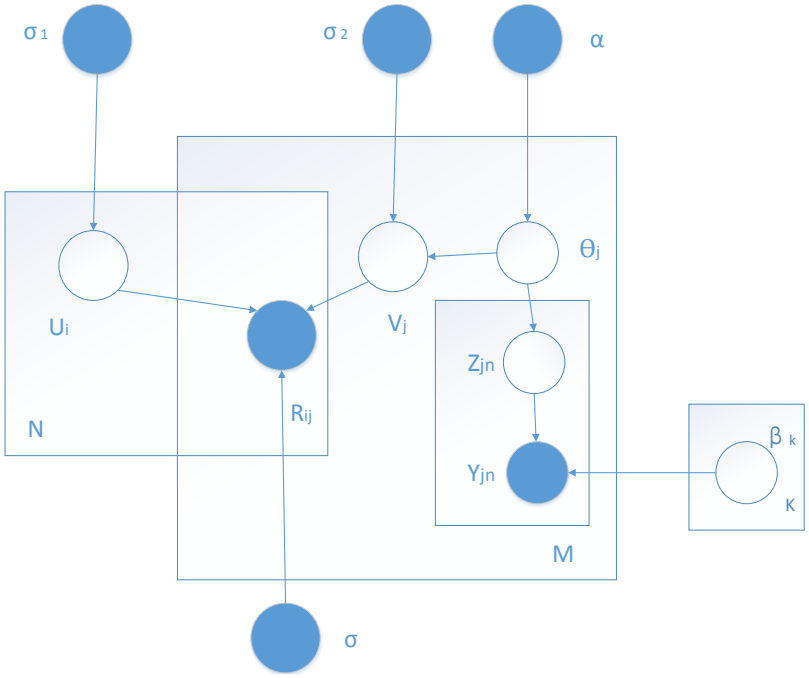
2.3 本章小结

本章主要介绍了混合主题模型的经典模型 LDA 的三种形式和两种图模型推断解法，因子分解模型的经典模型矩阵分解的四种先验假设及推断解法。总结和给出了一般性的求解框架，后续章节内容是这些模型的应用和延伸。

第3章 多关系模型

3.1 主要模型

3.1.1 融入辅助信息协同主题回归模型 CTR 及其优化算法



图示 3-1 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：

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

$$\sum_{j=1}^{M} \sum_{n=1}^{N_j} \log \sum_k \theta_{jk} \beta_{k Y_{jn}} \qquad (3.1)$$

$=\delta(i,j)$。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：

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

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

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：

$$\mathrm{L}(\theta_\mathrm{j},\phi_j)\geq-\frac{\lambda_v}{2}\left(\mathrm{V_j}-\theta_\mathrm{j}\right)^\mathrm{T}\left(\mathrm{V_j}-\theta_\mathrm{j}\right)+\sum_{\mathrm{n}=1}^{\mathrm{N_j}}\sum_\mathrm{k}(\phi_{jnk}\log\theta_{jk}\beta_{kY_{jn}}-\phi_{jnk}\log\phi_{jnk})\quad(3.2)$$

由于加入θ_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[59]:

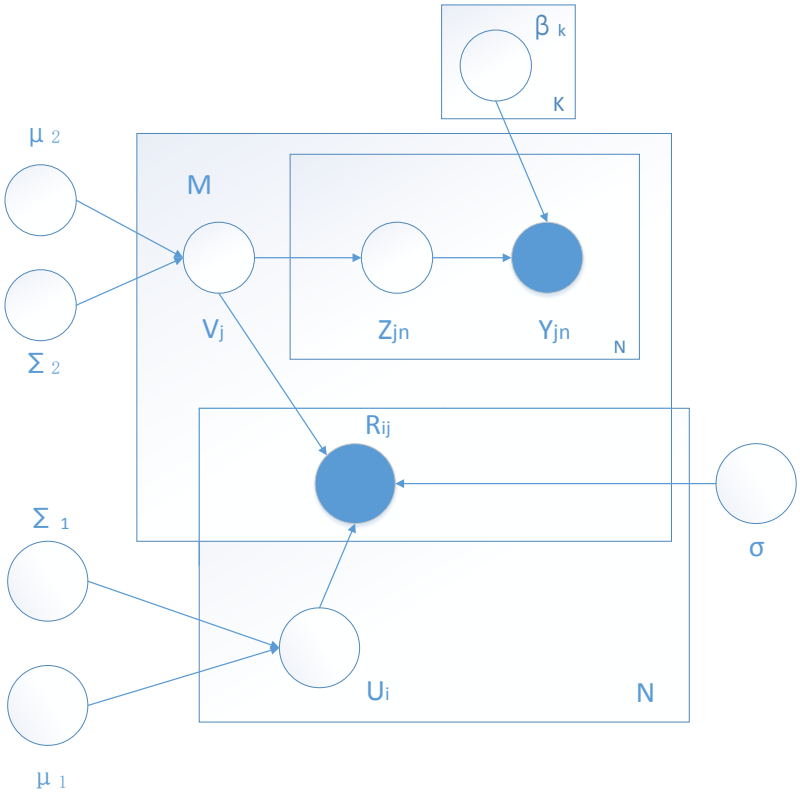
$$\min_\mathrm{v}\frac{1}{2}\left\|\theta_\mathrm{j}-\mathrm{v}\right\|_2^2,\ \text{其中 v 满足}\sum_\mathrm{k}\mathrm{v_k}=1,\mathrm{v_k}>0。$$

然后更新θ_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β_{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。把这个算法编号为“算法 3-1 PMF-CTM(VBEM 算法)”，算法流程略。

3.1.2 融入辅助信息概率矩阵分解 PMF-CTM 及 VB-EM 算法



图示 3-2 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：

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

$$p(\mathbf{Y},\boldsymbol{Z},\mathbf{V}|\boldsymbol{\beta},\mu,\sigma) = \Pi_{j=1}^{M} p(V_j|\mu_2,\Sigma_2)\Pi_{n=1}^{Nd}\{p(Z_{jn}|V_j)p(Y_{jn}|Z_{jn},\boldsymbol{\beta})\} \quad \text{(CTM)} \quad (3.3)$$

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μ_1,Σ_1,σ^2和所有Φ^i,Σ^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$\phi_{dn}, \boldsymbol{\beta}, \mu_2, \Sigma_2, \Phi^{\sim j}, \Sigma^{\sim j}$的更新主要考虑联合似然中的：

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

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$V_j$公共部分的推断，涉及到对$\Phi^{\sim j}, \Sigma^{\sim j}$的更新。这部分的 log 似然为：$\sum_{i=1}^{N}\delta(i,j)\log\ p\left(R_{ij}\middle|U_i^T V_j, \sigma^2\right) + \log p\left(V_j\middle|\mu_2, \Sigma_2\right) + \sum_{n=1}^{Nd}\log p\left(Z_{jn}\middle|V_j\right)$，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，

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

$$E_{q(Z,V_j)}\log p\left(V_j\middle|\mu_2, \Sigma_2\right) = \phi_{gaus}(\mu_2, \Sigma_2)^T E_{q(V_j)}u_{gaus}\left(V_j\right) + g_{gaus}(\mu_2, \Sigma_2)$$

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

$$E_{q(Z,V_j)}\Sigma_{n=1}^{Nd}\log p\left(Z_{jn}\middle|V_j\right) = E_{q(V_j)}\phi_{Mult}\left(\begin{pmatrix}\exp(V_{j_1})/\Sigma_{i=1}^{K}\exp(V_{j_i})\\ \cdots\cdots\\ \exp(V_{j_K})/\Sigma_{i=1}^{K}\exp(V_{j_i})\end{pmatrix}\right)^T \Sigma_{n=1}^{Nd}E_{q(Z_n)}u_{Mult}\left(Z_{jn}\right)$$

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

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

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

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

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

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

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

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

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

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

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

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

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

算法 3-2 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 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 $\mu_1, \Sigma_1, \mu_2, \Sigma_2, \sigma, \phi_{jns}, \zeta, \beta_{kv}$。

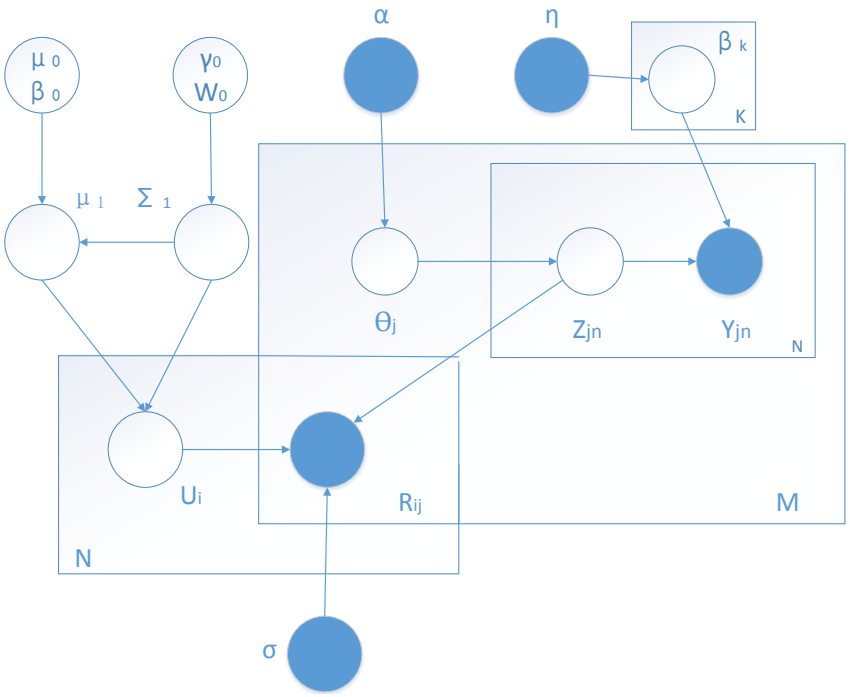
$\Delta^t = \|Y \odot (R - U^T V)\|_2^2$

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

For i from 1 to N, j from 1 to M：

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

3.1.3 融入辅助信息概率矩阵分解 PMF-LDA 及 MCMC 算法



图示 3-3 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$R_{ij}\sim N(U_i^T\left(\Sigma_{n=1}^{N_j}Z_{jn}/N_j\right),\sigma^2)$，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,μ_1,Σ_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：

$$p\left(Z_{jn}=k\middle|\boldsymbol{Z}^{\neg(j,n)}\right)\propto \qquad (3.4)$$

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

利用这个式子可完成对所有Z_{jn}的采样。

算法 3-3 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ϵ，最大迭代次数 S，固定值σ，初始化$\mu_0, \beta_0, \gamma_0, W_0, U, \mathbf{Z}, \alpha, \eta$。

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

For each i from 1 to N：

compute μ_i^*, Σ_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 $V_j = \Sigma_{n=1}^{N_j} Z_{jn} \Big/ N_j$

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

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

3.2 本章小结

本章的多关系模型从形式上来说是混合主题模型与因子模型的组合。给出了三种建模方法，涉及到如何将文本话题分布与矩阵分解因子相结合，这三种方法总结如下：

1. CTR 模型利用高斯回归假设来约束文本话题参数与矩阵分解特征。
2. PMF-CTM 模型将多元高斯分布产生的变量，既解释为矩阵分解特征，又同时当作罗杰斯特变换后的文本话题参数。
3. PMF-LDA 模型将文本的主题均值作为矩阵分解特征，这样稍作改变 LDA 的 Gibbs 采样公式即可得到新的更新公式。

这三种方法都是将文本话题参数与矩阵分解特征联系起来，构建一个完整的图模型再优化图模型的联合似然函数完成推断。

第4章 实验结果与比较

4.1 混合主题模型实验

4.1.1 数据集

这一部分，将选用2个数据集来测试LDA模型算法并展示结果。AP 数据集来自Associated Press，Nips数据集来自国际神经信息处理会议，统计信息见下表：

dataset	AP	Nips
Documents	2246	10474
Words	1000	15276

表格 4-1 AP 与 Nips 数据集

再选用hetrec2011-2k数据集来测试主题模型在推荐应用中的效果，它包含两个数据集Lastfm与Delicious，统计信息见下表：

Dataset	Lastfm	Delicious
Users	1892	1867
Items	18022	69226
tags	9749	42925
user-tags-items	186479	437593
user-items-relations	92834	104799
user-items-replay	92834	
item-tags-weight		487131

表格 4-2 hetrec2011-2k 数据集

4.1.2 评价指标与实验结果

AP 与 Nips 数据集中，实现对称先验 LDA 的 VB-EM 算法，并设定话题数量 K=10。训练后得到参数$\boldsymbol{\beta}$，并展示相应话题的前 20 个最大概率单词如下两张表格所示。然后利用，LDA 的 Gibbs 采样算法与 VB-EM 算法完成基于标签的推荐，

并在两个公开数据集 Delicious 与 Lastfm 上做对比。

Court	bush	percent	i	government
federal	president	million	police	military
million	i	year	people	two
new	soviet	new	two	people
case	new	market	years	officials
company	party	billion	new	soviet
state	dukakis	prices	city	united
judge	house	stock	school	police
last	political	last	time	states
department	campaign	sales	like	south
attorney	states	dollar	day	war
drug	congress	rose	just	official
law	government	oil	children	army
years	democratic	higher	home	troops
year	people	cents	three	today
government	reagan	price	yearold	killed
officials	committee	trading	first	president
office	united	rate	students	last
trial	gorbachev	york	dont	force
two	national	rates	man	news

表格 4-3 AP 数据集上 K=5 时概率最大的前 20 个单词

Unit	single	figure	learning	neural
set	neural	model	model	function
data	time	state	algorithm	hidden
rate	paper	networks	input	recognition
neural	general	training	training	visual
input	method	output	data	output
time	error	input	results	parameters
learning	sets	neurons	set	noise
structure	learning	algorithm	information	references
step	maximum	learning	number	networks
node	net	system	networks	context
error	classification	layer	performance	problem
number	features	order	error	time
dimensional	components	vector	based	analog
similar	problem	models	time	pattern
algorithms	data	weights	state	defined
properties	presented	units	system	expected
estimation	input	data	figure	model
type	information	small	linear	computation
long	present	functions	distribution	representations

表格 4-4 Nips 数据集前 1000 文本上 K=5 时概率最大的前 20 个单词

用蓝色字体标出了 5 个话题互不交叉的单词。从 AP 数据集上可以得到，话题 1 的主要单词是 court federal state attorney law government officials trial，我们可以认为这个话题主要与“法律、审判、法庭、法院”等内容相关。话题 2 的主要单词是 bush president party political campaign congress democratic committee gorbachev，这样可以认为该话题主要与“政治，总统，国会，政党，民主”等内容相关。话题 3 的主要单词是 market prices stock sales dollar oil trading rates，可以认为这个话题主要与“市场，定价，期货，美元，石油，贸易”有关。话题 4 的主要单词 people city school children home students man，可以认为这个话题主要与“人群，孩子，学校，家庭，城市”有关。话题 5 的主要单词是 military soviet police war army troops killed force，可以认为这个话题主要与“军队，武力”相关。很明显，Nips 是国际著名的机器学习学术会议，完整数据集包含 1987 到 1999 年的数据，由于选取的文档是前 1000 份，这些内容多来自最早先的会议。话题 1 中 rate structure step node dimensional similar properties estimation type，我们可以认为该话题与“结构，结点，维度，相似”这些结构学习的话题相关。话题 2 中 general maximum classification features components problem，可以认为该话题与“分类，主成，特征”等基于特征的学习问题相关。话题 5 中 hidden recognition visual noise references context analog pattern computation representations 则与“隐层与显层，模式与表示”等神经网络推理与表示问题有关。

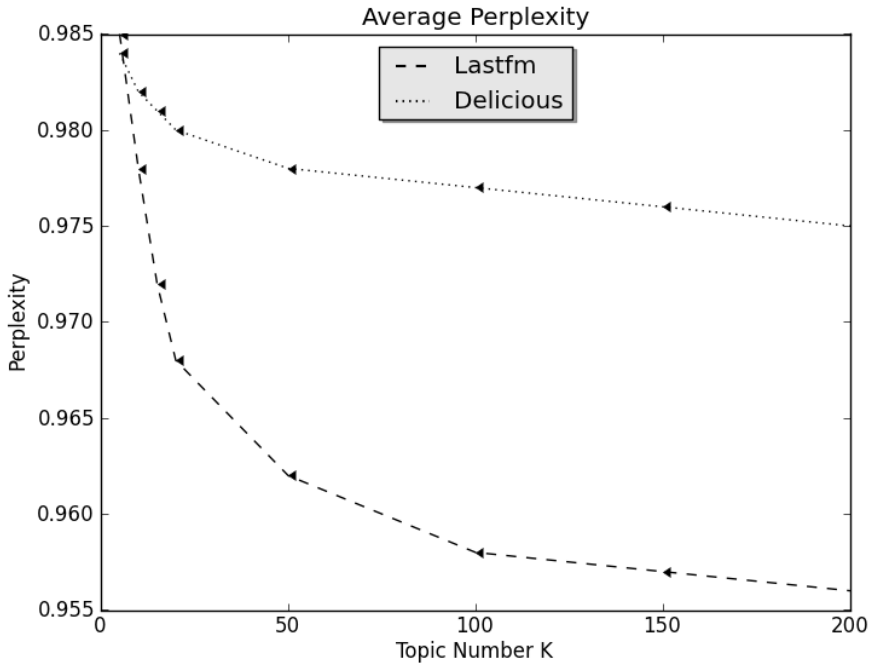
由上面的分析可见，主题模型 LDA 算法可以有效地处理 AP 数据集和 Nips 数据集并对单词做有效的聚类。下面我们应用主题模型算法来完成基于标签的推荐。Lastfm、Delicious 数据集是标签系统数据，它们允许用户为自己喜欢的商品打上相应的标签（单词），我们的任务是利用这些已知数据，完成对商品和标签的推荐。由于主题模型恰好可以处理标签（单词）数据，基于物品全局的标签信息和用户的历史标签数据训练模型，就可以依据话题为新的物品打上适当的标签，进而完成标签推荐。

复杂性(perplexity)，是概率主题模型训练程度的一个重要指标，更多的话题数 K 将会得到更低的复杂性和高似然率，它的评价公式如下：

$$p(Y_{dn}) = \frac{\sum_k \gamma_{dk}\beta_{kY_{dn}}}{\sum_k \gamma_{dk}}$$

$$perplexity(U,I) = \exp\{-\sum_{d,n} p(Y_{dn}) / N_{U,I}\} \quad (4.1)$$

对两个公开数据集 Lastfm 与 Delicious 抽取其中 80%的数据作为训练数据集，而剩余的 20%的数据作为测试数据集，这样得到的实验结果如下：

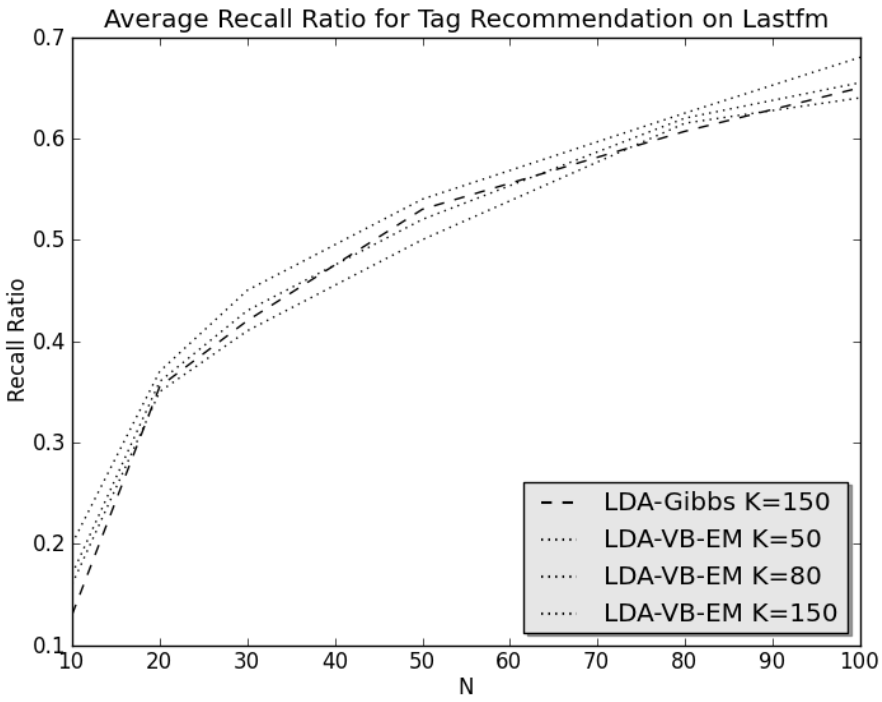


表格 4-5 主题模型复杂度随话题参数变化图

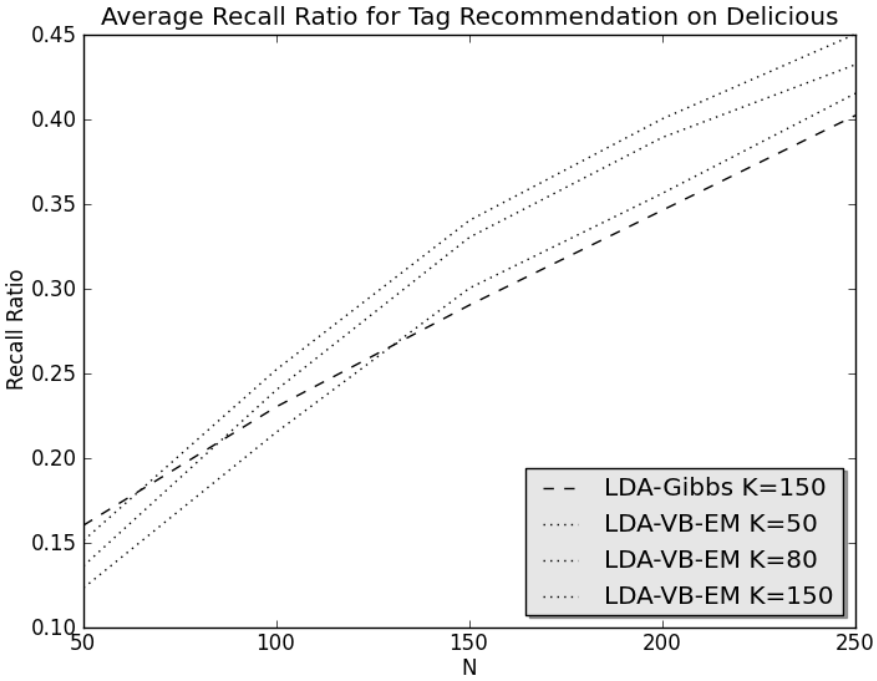
从上图中可以看到，话题数量 K 值越大所得到的似然率越高，同时 perplexity 越低，因而 K 选择 50、80、150 这三个值完成后面的实验。Top-N 推荐是指为用户推荐 N 个待选项，刻画推荐算法准确率的指标召回率(Recall Ratio)为：

$$Recall@N = \frac{number\ of\ items\ the\ user\ likes\ in\ Top\ N}{total\ number\ of\ items\ the\ user\ accept} \quad (4.2)$$

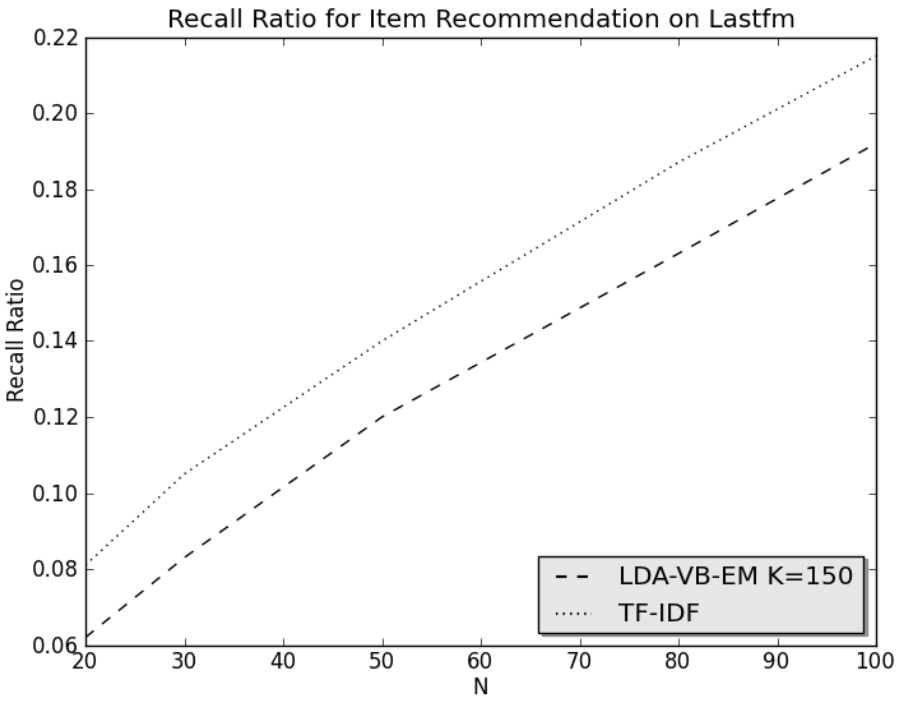
对应到主题模型，则将用户看作一个文件(a corpus of words)，项目和标签都作为单词(word)。对最终的训练结果计算如概率值，用户 u 选择话题 k 的概率为$\theta_{uk} \propto \sum_n \phi_{unk}$，$p(i|u) = \sum_k \theta_{uk}\beta_{ki}$。取p(i|u)值最大的前 N 个 i 构成待推荐列表。实验结果如下图：



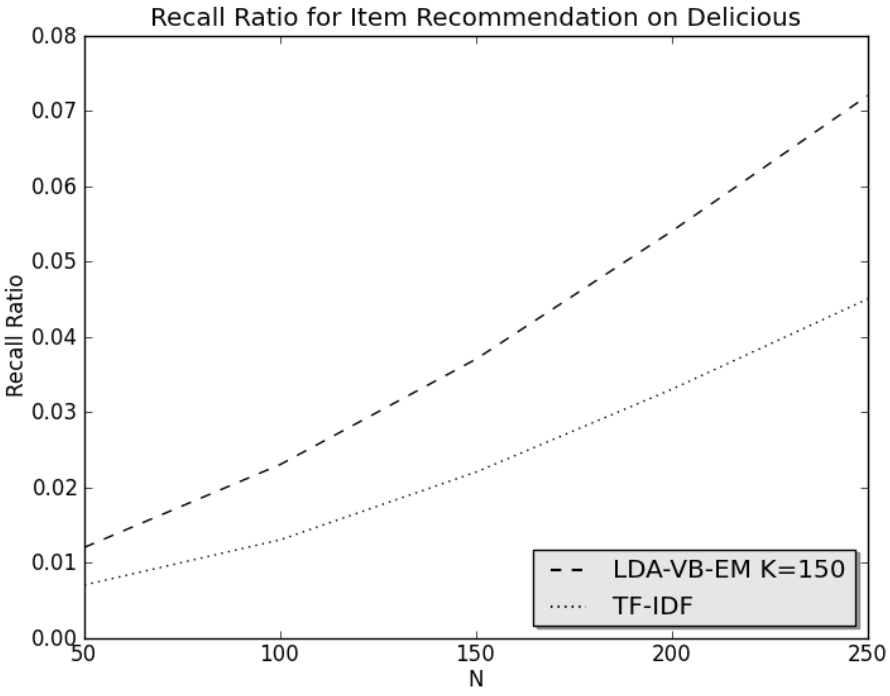
表格 4-6 Lasfm 数据集上标签推荐的平均召回率



表格 4-7 Delicious 数据集上标签推荐的平均召回率



表格 4-8 Lastfm 数据集上项目推荐的平均召回率



表格 4-9 Delicious 数据集上项目推荐的平均召回率

4.2 因子分解与多关系模型实验比较

4.2.1 数据集

选用两个公开数据集 Movielens 与 Netflix 的一部分子集。Movielens 电影评分数据由 GroupLens Research 项目小组提供，它包含三个大小分别为 100k，1M 与 10M 的数据集。电影评分数值范围是 1 到 5，间隔 0.5 分。此外，Movielens 数据提供了电影属性，这将作为辅助信息使用。Netflix 数据集来自 Netflix Prize 比赛，评分跨度为 1 到 5 的整数。在本文的实验中，我们选择的 Movielens 数据集包含 100,000 条评分有 943 个用户 ID 与 1682 个电影 ID，而 Netflix 数据集则收集其中的 900,000 条评分有 6040 个用户 ID 与 3950 个电影 ID。

4.2.2 评价指标与实验结果

采用最小方均根误差 RMSE 刻画评分预测算法的准确程度，计算公式如下：

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{n=1}^{N}(r_n-\hat{r}_n)^2}{N}} \tag{4.3}$$

在这里，所有的预测评分$\hat{r}_n$都需要经过后处理再得到相应的 RMSE 值。每次实验，随机划分其中的 80%作为训练集，20%作为测试集。为了方便起见在 Movielens 数据集上 SGD 学习率这里采用固定值$\eta=0.001$而 Netflix 数据集上$\eta=0.01$。每次实验将 80%数据集化分成四份，做四次交叉验证确定好最适当的正则化参数$\lambda=\lambda_V=\lambda_U\in\{0.1，0.01，0.001\}$，经验迭代次数则为在验证集上取得最好 RMSE 的四次迭代次数平均值，再利用数据集的 80%完整部分完成训练。

方法 \ D	5	10	15	20	25	30
SGD	0.956	0.943	0.938	0.929	0.922	0.917
ALS	0.942	0.936	0.923	0.915	0.912	0.905
PMF	0.985	0.974	0.964	0.957	0.953	0.950
BPMF	0.933	0.925	0.917	0.912	0.904	0.893
FM	0.923	0.912	0.905	0.903	0.896	0.891

表格 4-10 Netflix 数据集上 RMSE 结果

方法 \ D	5	10	15	20	25	30
PMF	0.845	0.852	0.868	0.879	0.891	0.903
PMF-CTM	0.839	0.843	0.862	0.877	0.885	0.892
BPMF	0.821	0.835	0.847	0.859	0.864	0.879
PMF-LDA	0.817	0.824	0.838	0.852	0.859	0.871
ALS	0.802	0.815	0.824	0.825	0.829	0.833
CTR	0.796	0.808	0.812	0.818	0.821	0.829

表格 4-11 Movielens 数据集上 RMSE 结果

4.3 本章小结

本章实验结果有两个部分，前一部分主要是主题模型的应用：文本关键词提取、基于标签的推荐，后一部分主要是协同过滤实验。

从表格 4-6、4-7 中可见 K 值越大相应的召回率越高，而 Gibbs 采样算法的精度略低于相同 K 值的 VBEM 算法。表格 8、9 中可见对项目推荐的召回率 TF-IDF 与 VBEM 算法在两个数据集上表现各异，表明 LDA 与 TF-IDF 具有等效的作用。

从表格 4-10 中可以看到在 Netflix 数据集上，5 种 baseline 算法随不同特征维数 D 变化的 RMSE 预测值。可见 D 值越大所得到的 RMSE 值越小，PMF、BPMF 等算法的效果都要好于 SGD 算法，而 ALS、PMF 与 BPMF、FM 之间的差别不大。这里需要说明的是，由于单一数据集和实验初始值设置的影响，这不能说明各算法之间的优劣。

从表格 4-11 中可以看到在 Movielens 数据集上，多关系模型算法与 baseline 算法随不同特征维数 D 变化的 RMSE 预测值。可见 D 值变大所得到的 RMSE 变大，这说明特征维数并不是越大越好。PMF-CTM 优于 PMF，PMF-LDA 优于 BPMF 算法，CTR 与 ALS 相同参数设置一致时 CTR 优于 ALS 算法。依据图 1-1 所示的模型关系，融入辅助信息的协同过滤算法更为有效。

第5章 总结和展望

5.1 总结

本文对多关系数据挖掘模型的阐述，主要介绍了两种基本的概率模型建模方法即混合主题模型和因子分解模型，并给出基于变分近似推断和蒙特卡洛采样两种求解策略的对应算法。基于这两种模型的特点，引出了融合辅助信息的多关系模型，有效地结合了两种模型的功能，并为此提出一套完整的模型求解策略。具体研究内容如下：

1. 对现有主题模型算法的研究和应用。研究多项式混合模型从 PLSA、GMM 到层级贝叶斯推广模型 LDA，以及改进模型 CTM。总结几种重要的主题模型方法并推导其中关键步骤。应用 LDA 完成基于标签的推荐系统算法，比较了不同话题数量下，VBEM 与 MCMC 采样算法的效果。
2. 对现有因子分解模型算法的研究和应用。研究概率矩阵分解模型，如 PMF、BPMF、FM，以及它们与正则矩阵分解的联系。总结几种主要的矩阵分解建模方法并给出了 VBEM、MCMC、SGD、ALS 几种相应的解法，在公开数据集上比较了协同过滤效果
3. 提出基于混合主题模型和因子分解模型的多关系模型。给出三种建模方法，有效地利用了基本模型求解策略，扩展了因子模型和主题模型功能，给出相应的三种优化求解算法，有效地利用辅助信息提高了基本协同过滤方法。

5.2 展望

多关系数据挖掘是丰富多彩的研究领域，本文所论述的融合辅助信息的概率矩阵分解模型只是这类模型中非常具有特点的一个。就这类模型的应用背景而言，也是更广阔的研究话题，把辅助信息融入协、同过滤模型中，就是数据挖掘和推

荐系统领域研究的热点。

例如，提高模型的准确率和求解速度以提高模型适应大数据的能力，这需要考虑主题模型和因子模型采用不同策略以更有效地结合，还需要利用贝叶斯在线学习方法来求解图模型。例如，考虑文本和辅助信息建模的其它模型，寻找引入辅助信息的新方法，以提高基本协同过滤系统的准确率。还有，将用户的社会化网络关系融入到矩阵分解中，将用户浏览历史的时序信息融入到矩阵分解中等。因而，无论是从混合主题模型与因子分解模型更有效结合的观点，还是从应用背景的角度，这方面的研究都是值得进一步深入和探索。

附录

LDA 变分近似推断算法

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

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

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

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

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

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

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

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

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

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

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：

$$\mathrm{term}\{1,2,4\} = \left(-\begin{pmatrix} \gamma_1 - 1 \\ \cdots\cdots \\ \gamma_\mathrm{K} - 1 \end{pmatrix} + \begin{pmatrix} \alpha_1 - 1 \\ \cdots\cdots \\ \alpha_\mathrm{K} - 1 \end{pmatrix} + \begin{pmatrix} \Sigma_{\mathrm{n}=1}^{\mathrm{Nd}} \phi_{\mathrm{n}1} \\ \cdots\cdots \\ \Sigma_{\mathrm{n}=1}^{\mathrm{Nd}} \phi_{\mathrm{n}\mathrm{K}} \end{pmatrix} \right)^{\mathrm{T}} \begin{pmatrix} \Psi(\gamma_1) - \Psi\left(\sum_{i=1}^{\mathrm{K}} \gamma_i\right) \\ \cdots\cdots \\ \Psi(\gamma_\mathrm{K}) - \Psi\left(\sum_{i=1}^{\mathrm{K}} \gamma_i\right) \end{pmatrix} - \log \frac{\Gamma(\sum_{i=1}^{\mathrm{K}} \gamma_i)}{\prod_{\mathrm{i}=1}^{\mathrm{K}} \Gamma(\gamma_i)}$$

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

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

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

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

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

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

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

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

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

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

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：

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

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

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：$\frac{\Sigma_{d=1}^{D}\Sigma_{n=1}^{Nd}\delta(Y_{dn}=v)\phi_{dnk}}{\beta_{kv}}+\lambda_k=0$，

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

CTM 变分近似推断算法

E-step：下面的讨论同样省略下标 d。

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

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

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

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

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

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

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

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

$$E_{q(Z,\theta)}\Sigma_{n=1}^{Nd}\log p(Z_n|\eta)\ge\begin{pmatrix}\lambda_1-\zeta^{-1}(\Sigma_{i=1}^{K}\exp\{\lambda_i+\gamma_i^2/2\})+1-\log\zeta\\ \cdots\cdots\\ \lambda_K-\zeta^{-1}(\Sigma_{i=1}^{K}\exp\{\lambda_i+\gamma_i^2/2\})+1-\log\zeta\end{pmatrix}^T\begin{pmatrix}\Sigma_{n=1}^{Nd}\phi_{n1}\\ \cdots\cdots\\ \Sigma_{n=1}^{Nd}\phi_{nK}\end{pmatrix}$$

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

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

$$H(q(\eta))=-E_{q(\eta)}\log q(\eta)=-\phi_{gaus}(\lambda,\gamma^2)^T E_{q(\eta)}u_{gaus}(\eta)-g_{gaus}(\lambda,\gamma^2)=$$

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

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

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$\{\lambda, \gamma^2\}$：

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

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

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

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：$\phi_{\mathrm{ni}} \propto \beta_{\mathrm{i}Y_\mathrm{n}}\exp(\lambda_\mathrm{i})$，$\Sigma_{\mathrm{k=1}}^{K}\phi_{\mathrm{nk}} = 1$。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：$\mathrm{Oj}(\mu,\Lambda^{-1}) = \Sigma_{\mathrm{d=1}}^{D}(\lambda_d^T\Lambda\mu - \frac{1}{2}\mathrm{tr}(\Lambda\boldsymbol{\gamma_d^2}) - \frac{1}{2}(\mu^\mathrm{T}\Lambda\mu + \lambda_d^T\Lambda\lambda_d)) + D\log\sqrt{\frac{|\Lambda|}{(2\pi)^K}}$，$\frac{\partial\mathrm{Oj}(\mu,\Lambda^{-1})}{\partial\mu} = \Sigma_{\mathrm{d=1}}^{D}(\lambda_d^T\Lambda - \Lambda\mu)$，

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

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

对称先验 LDA 变分近似推断算法

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γ的有 1、2、4 涉及到参数ϕ的有 2、3、5，实际计算可

以发现 E 步与非对称 LDA 相同：

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

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

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

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β_{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λ_{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λ_{ki}求导，则有

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

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

设上式为 0，有：

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

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。

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

对称先验 LDA 的 MCMC 算法②

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

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

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

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

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

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

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

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

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

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

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

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 $\mathbf{Y}$，α，η，我们只需推断参数 $\mathbf{Z}$，可利用下式做 Gibbs 采样：

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

参考文献

[1] Winn J M, Bishop C M. Variational message passing[C]. Journal of Maching Learning Research. 2005: 661-694.

[2] 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.

[3] 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.

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

[5] 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.

[6] 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.

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

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

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

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

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

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

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

[14] 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.

[15] 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.

[16] Korsos L F, Taddy M. Gibbs Sampling for n-Gram Latent Dirichlet Allocation[J]. 2011.

[17] Zeng J, Cheung W, Liu J. Learning topic models by belief propagation[J]. 2011.

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

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

[20] 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.

[21] 赵鑫, 李晓明. 主题模型在文本挖掘中的应用. 北京大学技术报告

[22] 刘知远, 孙茂松. 基于文档主题结构的关键词抽取方法研究. 清华大学博士学位论文

[23] Zhu J, Ahmed A, Xing E P. MedLDA: maximum margin s。upervised topic models for regression and classification[C] Proceedings of the 26th Annual International Conference on Machine Learning. ACM, 2009: 1257-1264.

[24] Jiang Q, Zhu J, Sun M, et al. Monte carlo methods for maximum margin supervised topic models[C] Advances in Neural Information Processing Systems. 2012: 1601-1609.

[25] Teh Y W, Jordan M I, Beal M J, et al. Hierarchical dirichlet processes[J]. Journal of the american statistical association, 2006, 101(476).

[26] Blei D M, Jordan M I. Variational inference for Dirichlet process mixtures[J]. Bayesian analysis, 2006, 1(1): 121-143.

[27] Blei D M, Jordan M I. Variational methods for the Dirichlet process[C] Proceedings of the twenty-first international conference on Machine learning. ACM, 2004: 12.

[28] 周建英，王飞跃，曾大军．分层 Dirichlet 过程及其应用综述,自动化学报,Vol.37, No.4, pp.390-407, 2011. 4.

[29] Krestel R, Fankhauser P, Nejdl W. Latent dirichlet allocation for tag recommendation[C] Proceedings of the third ACM conference on Recommender systems. ACM, 2009: 61-68.

[30] Wang C, Blei D, Li F F. Simultaneous image classification and annotation[C] Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009: 1903-1910.

[31] Srebro N, Rennie J, Jaakkola T S. Maximum margin matrix factorization[C] Advances in neural information processing systems. 2004: 1329-1336.

[32] 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.

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

[34] 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.

[35] 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.

[36] Introduction to statistical relational learning[M]. The MIT press, 2007.

[37] Li W J. Latent factor models for statistical relational learning[D]. The Hong Kong University of Science and Technology, 2010.

[38] Shan H. Probabilistic Models for Multi-relational Data Analysis[D]. university of minesota, 2012.

[39] Agarwal D, Chen B C. Regression-based latent factor models[C] Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009: 19-28.

[40] Singh A P, Gordon G J. Relational learning via collective matrix factorization[C] Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2008: 650-658.

[41] Gao S, Denoyer L, Gallinari P, et al. Latent factor blockmodel for modelling relational data[M] Advances in Information Retrieval. Springer Berlin Heidelberg, 2013: 447-458.

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

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

[44] 吴金龙, 鄂维南. Netflix Prize 中的协同过滤算法. 北京大学博士学位论文

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

[46] 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.

[47] 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.

[48] Loni B, Shi Y, Larson M, et al. Cross-Domain Collaborative Filtering with Factorization Machines[J].

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

[50] Rendle S. libFM 1.3-Manual[J]. 2012.

[51] Chen T, Zhang W, Lu Q, et al. SVDFeature: A Toolkit for Feature-based Collaborative Filtering[J]. Journal of Machine Learning Research, 2012, 13: 3619-3622.

[52] 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.

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

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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 international conference on Machine learning. ACM, 2008: 272-279.

攻读硕士学位期间主要的研究成果

一、参与的科研项目

1、 作为技术骨干参与 NSFC 项目（No. 61272303）“融入社会化信息情景相关推荐系统关键技术研究”，2013. 1-2016. 12，项目来源：国家自然科学基金委员会。

2、 作为技术骨干参与 973 项目子课题（No. 2010CB327903）：“面向多义性对象的学习理论和方法”，2010. 1-2011. 12，项目来源：国家重点基础研究项目。

3、 KDD-Cup 2012. Track 1 “Recommending users based on the Social network of Tencent Weibo”，在全球 658 个参赛对中排名第 9 名。

二、发表的学术论文

[1] Li, Zhengyang, and Congfu Xu. "Tag-based top-N recommendation using a pairwise topic model." Information Reuse and Integration (IRI), 2013 IEEE 14th International Conference on. IEEE, 2013.

致谢

浙江大学的研究生生活就要结束了，回首这三年的时光，需要感谢在学习和生活上给我帮助的很多人。

感谢我的导师徐从富老师，在这里我们有一个氛围融洽的实验室，老师和我们一起读书、读论文、探讨问题的学习环境，在这里学到的科研知识开阔了我的视野，老师严谨治学的态度、对研究工作充满激情和真诚的待人处事方式，这些言传身教时刻激励着我们，也会在以后的工作和生活中受用。感谢老师在本文的选题、写作等工作中的耐心指导。

感谢张志华老师、钱徽老师对我的关心和指导。

感谢实验室的大家，吴小琼、梁晨、钟豪、王鑫、尹志老师，以及已经毕业的师兄、师姐。感谢你们给我学习和生活上的帮助。

最后，我的家人、所有的朋友，感谢你们一直以来的支持和鼓励。

李正洋

2013-01-05