清 华 大 学

综 合 论 文 训 练

题目：以行列式点过程为先验的最大熵判别式隐狄利克雷分配模型

系　　别：物理系

专　　业：数理基础科学

姓　　名：刘　畅

指导教师：朱　军　副教授

2014 年 6 月 19 日

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：学校有权保留学位论文的复印件，允许该论文被查阅和借阅；学校可以公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存该论文。

(涉密的学位论文在解密后应遵守此规定)

签　名：刘畅　导师签名：　　日　期：2014.6.19

中文摘要

话题模型在自动分析文本信息方面取得了很大的成功。已有的话题模型，诸如隐狄利克雷分配模型（LDA），监督式隐狄利克雷分配模型（sLDA）以及最大熵判别式隐狄利克雷分配模型（MedLDA）等都可以成功地挖掘出一个文集中的话题，并能够将每篇文档对应到话题空间中的一个向量作为该文档的一个低维特征表示。然而，这些话题模型中都隐含有话题之间相互独立的先验，因而有可能会出现话题冗余，即话题之间有较大相似性的现象。这种现象会导致学习得到的话题代表性不足、不能很好地反映文档的特征等问题。为此，我们希望能为话题模型加入鼓励多样性的先验来代替独立性先验，得到更加多样的话题。行列式点过程（DPP）作为一种鼓励多样性的分布，可以作为话题的先验使用，使话题更加多样，从而提高话题的表达文档特征的能力，同时提高模型的预测能力。

本论文主要工作是提出使用 DPP 先验的 DPP-MedLDA 模型。MedLDA 模型是将 sLDA 与最大熵判别式方法结合而得到的一种更高效且具有更好预测效果的模型，是最大似然法与最大间隔方法的结合。通过使用 DPP 先验，DPP-MedLDA 模型的优化问题中增加了一个鼓励多样性的正则化项。本文给出了一种使用共轭梯度法处理这个正则化项的模型学习算法，并以此方法将 DPP-MedLDA 模型在“影评数据”（movie review data set）和“亚马逊评论数据”（Amazon review data）上进行测试。实验的定性与定量结果表明，DPP-MedLDA 模型能够显著增加话题的多样性与表达效力，且能够提高模型的预测能力。

关键词：多样性；最大熵判别式隐狄利克雷分配模型；行列式点过程

ABSTRACT

Topic models have achieved great success in automatically analyzing text information. Existing topic models, such as latent Dirichlet allocation (LDA), supervised latent Dirichlet allocation (sLDA), and maximum entropy discrimination latent Dirichlet allocation (MedLDA), can discover the latent topics in a corpus, and can assign a vector in the topic space to each of the documents as its low dimensional representation. These topic models, however, pose a latent independent prior for the topics, which may cause the problem of redundant topics, namely notable similarity among topics. This problem will lead to poor representation power of topics for documents. So we want to replace the independent prior with a prior for diversity to obtain more diverse topics. The determinantal point process (DPP) is a distribution encouraging diversity and can act as the prior for topics, so as to improve the representation power of topics and prediction power of topic models.

In this thesis, we will introduce a DPP prior for MedLDA to construct the DPP-MedLDA. By learning supervised LDA with the powerful maximum entropy discrimination principle, MedLDA is more efficient with better test results. It embodies both the maximum likelihood methods and the maximum margin principle. With a DPP prior, a regularizer for diversity appears in the optimization problem of MedLDA. We propose an effective learning method for the new model based on conjugate gradient methods to deal with this regularizer. We test DPP-MedLDA on the “movie review data set” and “Amazon review data”. The qualitative and quantitative results indicate that DPP-MedLDA can generate a set of more diverse topics with a more representative power and its predictive accuracy is improved.

Keywords: diversity; maximum entropy discrimination latent Dirichlet allocation; determinantal point process

目 录

1. 引言

1.1. 研究背景

随着当今计算机世界中电子文本总量的迅速膨胀，人们对能够自动分析大量文本数据的方法的需求与日俱增。已有的一些话题模型，如著名的 LDA（latent Dirichlet allocation，隐狄利克雷分配模型）[1]，已经能够实现自动发掘文本中隐含话题结构的功能。LDA 是一种生成式隐变量模型，是三层的层次模型。因为我们要发掘出来的话题是具有一定语义信息的，而实际的文本数据常常伴有一个概括性的响应量，这个量会很大程度上体现这个文档中话题的信息。比如一个影评的作者常常会给这个电影打分，这个分数便很大程度地体现了影评中包含的话题。为了能够利用这个文档的响应量，D. Blei 等人（2007）[2]提出了 sLDA 模型（supervised latent Dirichlet allocation, 监督式隐狄利克雷分配模型）。受 SVM（support vector machine, 支持向量机）的最大间隔（max-margin）思想的启发，J. Zhu 等人（2012）[3]提出的 MedLDA（maximum entropy discrimination latent Dirichlet allocation，最大熵判别式隐狄利克雷分配模型）将 sLDA 与最大间隔方法结合，对文集的各个文档和对应的响应量训练 SVM，得到了更加稀疏更加高效的监督式话题模型。

然而，这些话题模型都是建立在各个话题之间是相互独立的假设之上的。因而，有时为了能够得到更好的训练效果（训练时取得更大的似然），会出现话题趋同或者说话题冗余的现象。这样得到的话题虽然在训练集上能够取得较好的预测结果，但每个话题的语义并不明确清晰，而且因为话题冗余，每个文档在话题空间中的特征并不突出，在测试集上的预测结果也欠佳。我们所希望得到的是尽可能多样化的话题，使得每个话题更有概括力，真正表达出一个明确的语义，也只有这样，我们训练得到的话题才能够在分析新文档话题结构以及预测新文档响应变量时得到更好的效果。为此我们需要引入话题之间相互排斥的相关性。这对于使话题模型发掘出更有意义的话题、进而突出文档特征并提高分类或回归准确度具有很大意义。

1.2. 研究现状

当前，为模型引入相互排斥相关性的一个有效的方法是，令需要有相互排斥效应的变量服从行列式点过程（determinantal point process, DPP）的分布。行列

式点过程源自量子力学中对费米子互斥效应的描述，它能够产生更加多样的（分散的、相互排斥的）对象。近年来在统计机器学习领域中的研究表明，DPP 可以为相互排斥的随机过程建立描述贴切且便于处理的模型。DPP 的推断任务有一些很方便的性质，对于抽样、求边缘分布和条件概率都有较简洁的准确算法。A. Kulesza 和 B. Taskar 等人（2013）[4]的工作详细介绍了 DPP 的性质和一些算法，以及在一些情况下的应用。

J. Zou 等人（2012）[5]的工作给出了将 DPP 应用于生成式隐变量模型的一般方法，即将 DPP 作为需要互斥效应的隐变量的先验。此工作以 LDA 和高斯混合模型（Gaussian mixture model, GMM）两个具体的生成式隐变量模型为例，演示了加入 DPP 先验的具体方法，以及改造后模型的效果。其实验结果表明，加入 DPP 先验后的 LDA 能够自动地将停词聚集到少数几个话题之中，并且以文档的话题分配参数为特征的 SVM 分类器在测试集上的分类准确度会有提高；加入 DPP 先验的 GMM 能够提取出图片更加有信息量的特征，从而用这些特征训练出来的 SVM 分类器能够达到更高的分类准确度。

1.3. 论文研究重点和组织结构

由 J. Zou 等人（2012）[5]的工作可知，经过 DPP 先验改造后的 LDA 模型的性能得到了很大的改善。对于另一个很有效的话题模型，MedLDA 模型得到的话题仍然会出现冗余的情况。因此本工作的研究重点便是为 MedLDA 模型引入 DPP 先验来改造，并考察改造后模型的性能。

本文接下来将首先给出对已有话题模型发展过程的介绍并引出 MedLDA，之后介绍 DPP 的定义、简单性质和在机器学习领域中的发展情况，接着介绍如何将 DPP 用作 MedLDA 的先验以及求解 DPP-MedLDA 模型的具体算法，然后展示 DPP-MedLDA 模型的实验结果并与无 DPP 先验的结果对比，最后是对现有结果的总结以及对未来工作方向的展望。

2. 话题模型

在对 MedLDA 模型进行具体的改造之前，首先对 MedLDA 以及它所基于的 LDA 模型和 sLDA 模型进行介绍。由于 MedLDA 模型是由 sLDA 模型发展而来、sLDA 模型是由 LDA 模型发展而来的，因而它们遵从相同的基本量的定义。下面依照这种发展顺序分别对三个模型进行介绍，给出它们定义的基本量、模型的描述和特点以及具体的学习算法。

2.1. LDA 模型

LDA 模型由 D. Blei 等人（2003）[1]提出，它是一个三层的层次模型，其中有两层是隐变量。

2.1.1. LDA 模型中基本量的定义

LDA 模型的输入数据为文集，也就是D个文档的集合。每个文档是一个单词的集合，记文集中第d个文档的单词总数（重复出现的单词重复计数）为N_d。一个词表是一个由不重复的单词所构成的集合，词表应包含文集中出现的所有单词。记词表中不重复单词的数量为V。

LDA 模型以下述方式表示输入数据。用一个V维向量w_{dn}来表示文档d中的第n个单词，若此单词是词表中的第v个单词，则$w_{dnv}=1$且w_{dn}的其余分量都为零。用一个$V\times N_d$的矩阵w_d来表示一个文档，$w_d=(w_{d1},w_{d2},\cdots,w_{dN_d})$。各个$w_d$构成的集合为一个文档，用$w=\{w_1,w_2,\cdots,w_D\}$来表示。

LDA 模型定义“话题”为分布在词表中单词上的多项分布。若模型中给定的话题数为K，则话题的参数β是一个$K\times V$的矩阵，其中元素β_{kv}表示对于第k个话题，词表中第v个单词出现的概率。由此定义可知$\sum_{v=1}^{V}\beta_{kv}=1$。

2.1.2. LDA 模型的生成过程

LDA 模型设定，每篇文档都是由给定的K个话题构成的。这K个话题对于文档d的配比用一个K维向量θ_d来表示，其每一个分量即表示对应话题对于构建文档d的权重。θ_d是由一个以K维向量α为参数的狄利克雷分布（Dirichlet distribution）生成的。LDA 模型进一步设定，文档中的每个词都属于一个话题。用一个K维向量z_{dn}来表示文档d中第n个单词所属话题，若此单词属于话题k，则$z_{dnk}=1$且z_{dn}的其余分量都为零。z_{dn}是由一个以文档d的话题配比θ_d为参数的多项分布（multinomial distribution）生成的。最后，文档d中第n个单词w_{dn}由一个多项分

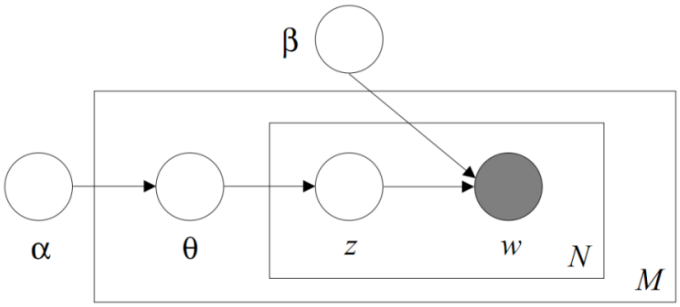

图 1：LDA 模型生成过程的图模型[1]

布生成，这个多项分布的参数是由z_{dn}所确定话题的参数。综上，LDA 设定的对于文集w中文档d的生成过程可以写为：

1) 生成文档d的话题配比$\theta_d|\alpha \sim \text{Dirichlet}(\alpha)$

2) 对于每个单词w_{dn}，

a) 生成话题所属参数$z_{dn}|\theta_d \sim \text{Multinomial}(\theta_d)$

b) 生成一个单词$w_{dn}|z_{dn},\beta \sim \text{Multinomial}(z_{dn},\beta)$

LDA 模型生成过程的图模型如图 1 所示。由上述生成过程可知，LDA 模型的参数为话题配比的先验α和话题β，并包含两个隐变量θ和z。由上述生成过程可得隐变量θ和z以及观测数据w的联合分布为：

$$p(\theta,z;w|\alpha,\beta)=\prod_{d=1}^{D}\left[p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}p(z_{dn}|\theta_d)p(w_{dn}|z_{dn},\beta)\right)\right] \tag{1}$$

其中

$$p(\theta_d|\alpha)=\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}\theta_{dk}^{\alpha_k-1} \tag{2}$$

$$p(z_{dn}|\theta_d)=\prod_{k=1}^{K}\theta_{dk}^{z_{dnk}} \tag{3}$$

$$p(w_{dn}|z_{dn},\beta)=\prod_{v=1}^{V}\prod_{k=1}^{K}\beta_{kv}^{z_{dnk}w_{dnv}} \tag{4}$$

观测数据w的边缘分布，即数据的似然为：

$$p(w|\alpha,\beta)=\int\sum_{z}p(\theta,z;w|\alpha,\beta)\,\mathrm{d}\theta$$

（5）

因此对上述似然关于模型参数α和β求最大值，得到参数的最大似然估计值（maximum likelihood estimate, MLE），即可得到由 LDA 学习得到的结果。

2.1.3. LDA 模型的变分算法

由于 LDA 模型的似然式（5）的形式复杂，直接进行最大似然估计是很不可行的。为此，D. Blei 等人（2003）[1]采用变分算法来求解模型的参数。由于边缘化（积分和求和）是对隐变量θ和z进行的，所以 D. Blei 等人（2003）[1]为隐变量引入了一个新的联合分布$q(\theta,z)$作为变分分布，且假设在此分布下θ和z相互独立，用以替代真实的分布，以使模型的求解能够处理。引入变分参数γ和ϕ，可将变分分布$q(\theta,z)$具体化为：

$$q(\theta,z|\gamma,\phi)=\prod_{d=1}^{D}\left(q(\theta_d|\gamma_d)\prod_{n=1}^{N_d}q(z_{dn}|\phi_{dn})\right)$$

（6）

其中，γ_d为K维狄利克雷分布参数，ϕ_{dn}为K维多项分布参数。ϕ满足$\sum_{k=1}^{K}\phi_{dnk}=1$。为了利用变分分布$q(\theta,z|\gamma,\phi)$，对于负对数似然，有如下性质：

$$\begin{aligned}
-\ln p(w|\alpha,\beta)&=-\ln\int\sum_{z}p(\theta,z;w|\alpha,\beta)\,\mathrm{d}\theta\\
&=-\ln\int\sum_{z}\frac{p(\theta,z;w|\alpha,\beta)}{q(\theta,z|\gamma,\phi)}q(\theta,z|\gamma,\phi)\,\mathrm{d}\theta\\
&\leq\int\sum_{z}(-)\ln\left(\frac{p(\theta,z;w|\alpha,\beta)}{q(\theta,z|\gamma,\phi)}\right)q(\theta,z|\gamma,\phi)\,\mathrm{d}\theta\\
&=-\int\sum_{z}q(\theta,z|\gamma,\phi)\ln p(\theta,z;w|\alpha,\beta)\,\mathrm{d}\theta\\
&\quad+\int\sum_{z}q(\theta,z|\gamma,\phi)\ln q(\theta,z|\gamma,\phi)\,\mathrm{d}\theta\\
&=-\mathbb{E}_{q(\theta,z|\gamma,\phi)}[\ln p(\theta,z;w|\alpha,\beta)]-\mathcal{H}\big(q(\theta,z|\gamma,\phi)\big)\\
&\triangleq\mathcal{L}^{u}(q(\theta,z|\gamma,\phi);\alpha,\beta)
\end{aligned}$$

（7）

即通过引入的隐变量的变分分布$q(\theta,z|\gamma,\phi)$找到了模型负对数似然的一个上界

$\mathcal{L}^u(q(\theta,z|\gamma,\phi);\alpha,\beta)$（上标$u$代表 unsupervised，即非监督式）。其中的不等式用到了 Jensen 不等式以及负对数函数是下凸函数的性质。为求解模型的参数，我们需要最小化负对数似然，而此时我们只需最小化这个上界。

利用式（7），可以将 LDA 模型的负对数似然上界$\mathcal{L}^u(q(\theta,z|\gamma,\phi);\alpha,\beta)$完全用模型参数$\alpha$和$\beta$以及变分参数$\gamma$和$\phi$表出。对变分参数求导并令之为零，得到变分参数的优化值满足：

$$\phi_{dnk} \propto \exp\left(\Psi(\gamma_{dk}) - \Psi\left(\sum_{j=1}^{K}\gamma_{dj}\right) + \ln\left(\sum_{v=1}^{V}\beta_{kv}w_{dnv}\right)\right)$$

$$\gamma_{dk} = \alpha_k + \sum_{n=1}^{N_d}\phi_{dnk} \tag{8}$$

其中$\Psi(\cdot)$是双伽马函数（digamma function），即伽马函数对数的导函数，$\Psi(x) = \Gamma'(x)/\Gamma(x)$。$\phi_{dnk}$的具体值可通过先计算上式中“正比于符号$\propto$”右边项的值再对$k$归一化得到。对模型参数求导并令之为零，得到模型参数的优化值满足：

$$\beta_{kv} \propto \sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi_{dnk}w_{dnv}$$

$$D\left(\Psi\left(\sum_{j=1}^{K}\alpha_j\right) - \Psi(\alpha_k)\right) + \sum_{d=1}^{D}\left(\Psi(\gamma_{dk}) - \Psi\left(\sum_{j=1}^{K}\gamma_{dj}\right)\right) = 0 \tag{9}$$

由于α的每一个分量的优化值都与其他分量相耦合，因此 D. Blei 等人（2003）[1]给出了一种使用牛顿-拉夫森（Newton-Raphson）算法的迭代算法更新α。

由于式（8）和式（9）分别依赖于模型参数和变分参数，因此 D. Blei 等人（2003）[1]使用期望最大化（expectation maximization, EM）算法来迭代求解模型参数。其期望步骤（E-step）根据当前的模型参数更新变分参数，最大化步骤（M-step）根据当前的变分参数更新模型参数，重复执行这两个步骤直到负对数似然上界收敛。其中的期望步骤（E-step），由于变分参数之间也有耦合，因此需要迭代地循环更新两个变分参数直到负对数似然上界收敛。根据上述思想，即可求解出 LDA 模型。

2.2. sLDA 模型

sLDA 模型由 D. Blei 等人（2007）[2]提出。它是对 LDA 模型的一个改进，

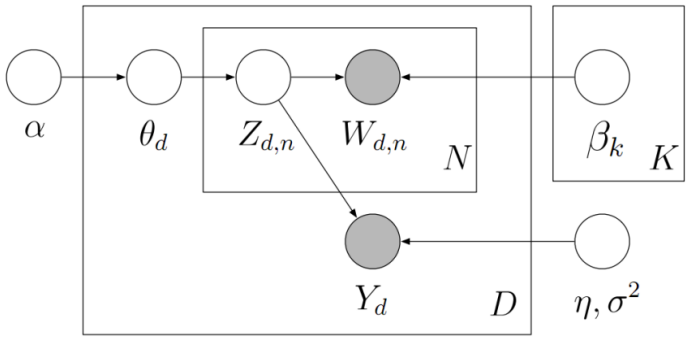

图 2：sLDA 模型生成过程的图模型[2]

使其成为监督式模型，能够利用文档的响应量来更好地发掘话题。为了将一篇文档和其响应量建立联系，D. Blei 等人（2007）[2]以文档的平均每个单词的话题所属参数作为文档的特征，并引入模型参数η和δ^2来联系文档和其响应量。

2.2.1. sLDA 模型的描述

sLDA 模型中的基本量的定义与 LDA 模型相同，只需为相应变量的加入再引入如下的量。用一个实数y_d来表示文档d的实际响应量。实际中响应量不只是实数的情况，也可以是有限分级中的一级或者是有限类别中的一类。D. Blei 等人（2007）[2]主要关注$y_d \in \mathbb{R}$的情况。作为文档的特征，定义文档d的平均每个单词的话题所属参数为

$$\bar{z}_d \triangleq \left(\frac{1}{N_d}\right)\sum_{n=1}^{N_d} z_{dn} \tag{10}$$

它是一个K维向量。为将文档的特征与其响应量相联系，我们引入模型参数η和δ^2来建立这种联系：

$$y_d|z_d, \eta, \delta \sim \mathcal{N}(\eta^T \bar{z}_d, \delta^2)$$

综上，sLDA 设定的对于文集w中文档d及其响应量y_d的生成过程可以写为：

1) 生成文档d的话题配比$\theta_d|\alpha \sim \text{Dirichlet}(\alpha)$
2) 对于每个单词w_{dn}，
   a) 生成话题所属参数$z_{dn}|\theta_d \sim \text{Multinomial}(\theta_d)$
   b) 生成一个单词$w_{dn}|z_{dn}, \beta \sim \text{Multinomial}(z_{dn}, \beta)$
3) 生成文档d的响应量$y_d|z_d, \eta, \delta^2 \sim \mathcal{N}(\eta^T \bar{z}_d, \delta^2)$

sLDA 模型生成过程的图模型如图 2 所示。由上述生成过程可知，sLDA 模型的参数为话题配比的先验α和话题β还有反映响应量特征的η和δ^2，隐变量仍为

θ和z。由上述生成过程可得隐变量θ和z以及观测数据w和y的联合分布为：

$$p(\theta, z; w, y|\alpha, \beta, \eta, \delta^2) = \prod_{d=1}^{D}\left[p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn}, \beta)\right)p(y_d|z_d, \eta, \delta^2)\right] \tag{11}$$

其中$p(\theta_d|\alpha)$、$p(z_{dn}|\theta_d)$和$p(w_{dn}|z_{dn}, \beta)$与 LDA 模型中的相同，而

$$p(y_d|z_d, \eta, \delta^2) = \frac{1}{\delta\sqrt{2\pi}} e^{-\frac{(y_d - \eta^T \bar{z}_d)^2}{2\delta^2}} \tag{12}$$

观测数据w和y的边缘分布，即数据的似然为：

$$p(w, y|\alpha, \beta, \eta, \delta^2) = \int \sum_{z} p(\theta, z; w, y|\alpha, \beta, \eta, \delta^2)\, \mathrm{d}\theta \tag{13}$$

因此求解 sLDA 模型就变为，对上述似然关于模型参数α、β、η和δ^2求最大值。

2.2.2. sLDA 模型的变分算法

与 LDA 模型相似，sLDA 模型直接进行最大似然估计也是很不可行的。为此采用同 LDA 模型类似的变分算法来求解。引入与 LDA 模型变分方法相同的变分分布和变分参数。利用变分分布$q(\theta, z|\gamma, \phi)$，使用与 LDA 相同的推导过程，可得到 sLDA 负对数似然的上界：

$$\begin{aligned} -\ln p(w, y|\alpha, \beta, \eta, \delta^2) &\leq -\mathbb{E}_{q(\theta,z|\gamma,\phi)}[\ln p(\theta, z; w, y|\alpha, \beta, \eta, \delta^2)] - \mathcal{H}\big(q(\theta, z|\gamma, \phi)\big) \\ &\triangleq \mathcal{L}^s(q(\theta, z|\gamma, \phi); \alpha, \beta, \eta, \delta^2) \end{aligned} \tag{14}$$

如此得到了 sLDA 模型负对数似然的一个上界$\mathcal{L}^s(q(\theta, z|\gamma, \phi); \alpha, \beta)$（上标$s$代表 supervised，即监督式）。将这个上界完全用模型参数α、β、η、δ^2以及变分参数γ和ϕ表出，并对变分参数和模型参数求导并令之为零便得到 sLDA 模型的参数更新方法。变分参数γ和模型参数α与β的优化值满足的方程与 LDA 模型的相同，因此这三个参数的更新方法同 LDA 模型。对于其余参数，其优化值满足：

$$\phi_{dnk} \propto \exp\left(\Psi(\gamma_{dk}) - \Psi\left(\sum_{j=1}^{K}\gamma_{dj}\right) + \ln\left(\sum_{v=1}^{V}\beta_{kv}w_{dnv}\right) + \frac{y_d}{N_d\delta^2}\eta_k - \frac{2\left(\eta^T\phi_{d,-n}\right)\eta_k + \eta_k^2}{2N_d^2\delta^2}\right) \tag{15}$$

$$\eta = (\mathbb{E}[A^TA])^{-1}\mathbb{E}[A]^Ty \tag{16}$$

$$\delta^2 = \frac{1}{D}\left(y^Ty - 2y^T\mathbb{E}[A]\eta + \mathrm{tr}\left(\mathbb{E}[A^TA]\eta\eta^T\right)\right) \tag{17}$$

其中，$A \triangleq (\bar{z}_1, \bar{z}_2, \cdots, \bar{z}_D)^T$以及$\phi_{d,-n} \triangleq \sum_{i\neq n}\phi_{di}$，且有

$$(\mathbb{E}[A])_{dk} \triangleq \left(\mathbb{E}_{q(\theta,z|\gamma,\phi)}[A]\right)_{dk} = \left(\mathbb{E}_{q(z|\phi)}[A]\right)_{dk} = \frac{1}{N_d}\sum_{n=1}^{N_d}\phi_{dnk} \tag{18}$$

$$\mathbb{E}[A^TA] \triangleq \mathbb{E}_{q(\theta,z|\gamma,\phi)}[A^TA] = \mathbb{E}_{q(z|\phi)}[A^TA] = \sum_{d=1}^{D}\frac{1}{N_d^2}\left(\sum_{n=1}^{N_d}\sum_{m\neq n}\phi_{dn}\phi_{dm}^T + \sum_{n=1}^{N_d}\mathrm{diag}\left(\phi_{dn}\right)\right) \tag{19}$$

采取与 LDA 模型相似的 EM 算法，即可求解 sLDA 模型。

2.3. MedLDA 模型

MedLDA 模型由 J. Zhu 等人（2012）[3]于 2009 年提出。这个模型是一种监督式的话题模型。它将最大间隔（maximum margin）的方法与 sLDA 模型结合，成为一种稀疏的话题模型。它能比 sLDA 提取出更加鲜明的话题特征，取得更好的准确度，也能比通常的 SVM 方法需要更少的支持向量，训练时间也会更少。

2.3.1. MedLDA 模型的描述

MedLDA 模型将 sLDA 模型中的模型参数η视作隐变量，并通过最大熵判别式方法（maximum entropy discrimination, MED）求解它所服从的分布。令$p_0(\eta)$为η的先验分布，则此时的隐变量与观测量的联合分布为：

$$p(\eta, \theta, z; y, w|\alpha, \beta, \delta^2) = p_0(\eta)p(\theta, z; y, w|\alpha, \beta, \eta, \delta^2) \tag{20}$$

其中$p(\theta,z;y,w|\alpha,\beta,\eta,\delta^2)$与 sLDA 中的相同。为能够处理似然（观测量的边缘分布），为所有隐变量引入变分分布$q(\eta,\theta,z)$，并同样令变分分布中的各量相互独立，且与η无关的部分与 sLDA 模型相同，即

$$q(\eta,\theta,z|\gamma,\phi)=q(\eta)\prod_{d=1}^{D}\left(q(\theta_d|\gamma_d)\prod_{n=1}^{N_d}q(z_{dn}|\phi_{dn})\right) \tag{21}$$

由隐变量的变分分布，使用与 LDA 相同的推导过程，可以得到 MedLDA 负对数似然的一个上界$\mathcal{L}^{bs}(q(\eta,\theta,z);\alpha,\beta,\delta^2)$（上标$bs$代表 Bayesian supervised，即贝叶斯监督式）：

$$\begin{aligned}-\ln p(w,y|\alpha,\beta,\delta^2) \leq= &-\mathbb{E}_{q(\eta,\theta,z)}[\ln p(\eta,\theta,z;w,y|\alpha,\beta,\delta^2)]-\mathcal{H}\big(q(\eta,\theta,z)\big)\\ &\triangleq \mathcal{L}^{bs}(q(\eta,\theta,z);\alpha,\beta,\delta^2)\end{aligned} \tag{22}$$

求解 MedLDA 时将似然最大化便可通过最小化这个上界来实现。这个上界与 sLDA 的负对数似然上界有如下关系：

$$\mathcal{L}^{bs}(q(\eta,\theta,z);\alpha,\beta,\delta^2)=\mathrm{KL}\big(q(\eta),p_0(\eta)\big)+\mathbb{E}_{q(\eta)}[\mathcal{L}^{s}(q(\theta,z);\alpha,\beta,\eta,\delta^2)] \tag{23}$$

LDA 和 sLDA 都是通过最小化负对数似然上界来求解模型的，也就是单纯地最大化似然。为将最大化似然与最大间隔的思想结合，我们将两个方法中的目标函数拼接起来并应用 MED 中的约束，得到 MedLDA 中待优化的目标函数：

$$\begin{aligned}&\min_{q(\eta,\theta,z),\alpha,\beta,\delta^2,\xi,\xi^*}\ \mathcal{L}^{bs}(q(\eta,\theta,z);\alpha,\beta,\delta^2)+C\sum_{d=1}^{D}(\xi_d+\xi_d^*)\\ &\forall d,s.t.:\begin{cases}y_d-\mathbb{E}_{q(\eta,z)}[\eta^T\bar{z}_d]\leq\varepsilon+\xi_d\\ -y_d+\mathbb{E}_{q(\eta,z)}[\eta^T\bar{z}_d]\leq\varepsilon+\xi_d^*\\ \xi_d,\xi_d^*\geq 0\end{cases}\end{aligned} \tag{24}$$

其中C是正则化系数，ε是 MED 中的精度参数，ξ和ξ^*是 MED 中的松弛变量。至此我们得到了 MedLDA 的优化问题。

2.3.2. MedLDA 模型的求解算法

为求解上述优化问题，为其中四个约束引入拉格朗日乘子分别为μ、μ^*、ν、ν^*并引入标准 SVM 中的 KKT 条件，则待解决的优化问题可以写为：

$$\min_{q(\eta),\gamma,\phi,\alpha,\beta,\delta^2,\xi,\xi^*,\mu,\mu^*,\nu,\nu^*} \mathcal{L}^{bs}(q(\eta,\theta,z);\alpha,\beta,\delta^2) + C\sum_{d=1}^{D}(\xi_d+\xi_d^*)$$
$$-\sum_{d=1}^{D}\mu_d\left(\mathbb{E}_{q(\eta,z)}[\eta^T\bar{z}_d]-y_d+\varepsilon+\xi_d\right)$$
$$-\sum_{d=1}^{D}\mu_d^*\left(y_d-\mathbb{E}_{q(\eta,z)}[\eta^T\bar{z}_d]+\varepsilon+\xi_d^*\right)-\sum_{d=1}^{D}(\nu_d\xi_d+\nu_d^*\xi_d^*)$$
$$\forall d, s.t.$$
$$:\begin{cases}\mu\geq 0\\ \mathbb{E}_{q(\eta,z)}[\eta^T\bar{z}_d]-y_d+\varepsilon+\xi_d\geq 0\\ \mu_d\left(\mathbb{E}_{q(\eta,z)}[\eta^T\bar{z}_d]-y_d+\varepsilon+\xi_d\right)=0\end{cases}\quad\begin{cases}\mu^*\geq 0\\ y_d-\mathbb{E}_{q(\eta,z)}[\eta^T\bar{z}_d]+\varepsilon+\xi_d^*\geq 0\\ \mu_d^*\left(y_d-\mathbb{E}_{q(\eta,z)}[\eta^T\bar{z}_d]+\varepsilon+\xi_d^*\right)=0\end{cases}$$
$$\begin{cases}\nu\geq 0\\ \xi\geq 0\\ \nu\xi=0\end{cases}\quad\begin{cases}\nu^*\geq 0\\ \xi^*\geq 0\\ \nu^*\xi^*=0\end{cases}$$

（25）

MedLDA 模型的求解即求解优化问题式（25），其中已将$q(\eta,\theta,z)$用式（21）代替为变分分布$q(\eta)$和变分分布参数γ和ϕ。由式（25）的目标函数以及式（23）可知，对于 MedLDA 模型的模型参数α、β和δ^2以及变分参数γ，其求解方法与 sLDA 的相同，只是对于δ^2需要将η用$\mathbb{E}[\eta]$来代替：

$$\delta^2=\frac{1}{D}\left(y^Ty-2y^T\mathbb{E}[A]\mathbb{E}[\eta]+\mathrm{tr}\left(\mathbb{E}[A^TA]\mathbb{E}[\eta\eta^T]\right)\right)$$

（26）

用拉格朗日乘子符号上加“ ^ ”来表示对应乘子的优化值。对于变分参量ϕ，会有两个约束条件对其有影响，对式（25）的目标函数关于ϕ求导并置为零，得到其优化值满足：

$$\phi_{dnk}\propto\exp\left(\Psi(\gamma_{dk})-\Psi\left(\sum_{j=1}^{K}\gamma_{dj}\right)+\ln\left(\sum_{v=1}^{V}\beta_{kv}w_{dnv}\right)+\frac{y_d}{N_d\delta^2}\mathbb{E}[\eta_k]\right.$$
$$\left.-\frac{2\left(\mathbb{E}\left[\eta^T\phi_{d,-n}\eta\right]\right)_k+\mathbb{E}[\eta_k^2]}{2N_d^2\delta^2}+\frac{\mathbb{E}[\eta_k]}{N_d}(\hat{\mu}_d-\hat{\mu}_d^*)\right)$$

（27）

最后一项即体现了约束带来的影响。前面诸项和 sLDA 中ϕ的更新公式相同，只需将η写作期望的形式。

式（25）中的优化函数在$\gamma,\phi,\alpha,\beta,\delta^2$都固定时，便成为一个 MED 问题。为

了求解变分分布$q(\eta)$，对优化函数关于$q(\eta)$求变分并置为零，得到$q(\eta)$的解为：

$$q(\eta) = \frac{p_0(\eta)}{Z}\exp\left(\eta^T \sum_{d=1}^{D}\left(\hat{\mu}_d - \hat{\mu}_d^* + \frac{y_d}{\delta^2}\right)\mathbb{E}[\bar{z}_d] - \frac{\eta^T \mathbb{E}[A^T A]\eta}{2\delta^2}\right) \tag{28}$$

其中A的定义同 sLDA，Z为归一化系数（配分函数）。

将优化后的$q(\eta)$代入优化问题式（25）并固定其他参数，即得到求解拉格朗日乘子最优值$\hat{\xi}, \hat{\xi}^*, \hat{\mu}, \hat{\mu}^*, \hat{v}, \hat{v}^*$的优化问题。这个优化问题可以采用标准的 SVM 的求解算法来解决。特别地，J. Zhu 等人（2012）[3]给出了当$p_0(\eta)$为K维标准正态分布时优化问题式（25）所对应的标准 SVM 优化问题。

有了上述对每个参数（函数）的优化方法，采用 EM 算法，迭代求得模型。E 步骤是对变分参数γ和ϕ的优化，M 步骤是对模型参数α、β、δ^2的优化以及求解相应的 SVM 问题以求得优化的拉格朗日乘子$\hat{\xi}, \hat{\xi}^*, \hat{\mu}, \hat{\mu}^*, \hat{v}, \hat{v}^*$。通过 EM 算法不断迭代直至负对数似然上界收敛，即可求得 MedLDA 模型。

3. 行列式点过程

行列式点过程（determinantal point process, DPP）最早由 Macchi（1975）[6]提出，当时是为了描述处于热力学平衡下的费米子系统的分布，因而当时的术语是“费米子过程”。由于费米子有泡利不相容原理的限制，即不会有两个全同的费米子占据同一个量子态，因而费米子的分布呈现出一种分散、多样的形态。费米子分布的这种排斥效应可以由行列式点过程精确描述。之后这种过程在概率与统计的领域中得到较为深入的研究。“行列式”这个术语最早是由 A. Borodin 和 G. Olshanski（2000）[7]提出的，之后在数学上的一些研究（例如，J. Hough 等人（2006）[8]、A. Borodin（2009）[9]、A. Scardicchio 等人（2009）[10]以及 F. Lavancier 等人（2012）[11]）给出了行列式点过程的抽样、边缘化、推断方法等各种概率和统计的性质。

近年来，行列式点过程在机器学习领域以其能够很好地模拟多样性的过程而受到关注。A. Kulesza 和 B. Taskar（2010）[12]提出行列式点过程的偏好性-相似性分解和结构化的行列式点过程（structured DPP）用来模拟分布在一组由结构化的对象构成的集合上具有多样性的分布。A. Kulesza 和 B. Taskar（2011a）[13]提出固定对象个数的行列式点过程（k-DPP）。A. Kulesza 和 B. Taskar（2011b）[14]提出条件化的行列式点过程（conditional DPP）并用于文档的抽取式总结。A. Kulesza 和 B. Taskar（2013）[4]给出了一个对近年来发展出来的 DPP 的性质、表示方法、学习和推断算法以及一些改进衍生版本的总结。此外，J. Gillenwater 等人（2012）[15]给出了一种基于亚模函数（submodular function）连续优化技术的行列式点过程最大后验推断的近似方法，R. Affandi 等人（2014）[16]给出了一种更加高效且可适用于大规模数据的行列式点过程参数的贝叶斯式学习方法。

下面将给出本论文需要使用的行列式点过程的一些基本概念和性质。

3.1. 行列式点过程的定义

此处我们不介绍一般性的 DPP 定义，而关注于本工作中要用到的情况。行列式点过程是在一个离散的有限的基本点集$\mathcal{Y} = \{1,2,\cdots,N\}$的幂集$2^{\mathcal{Y}}$（所有子集，包括空集和全集，构成的集合）上定义的概率分布。设$A \subset \mathcal{Y}$是一个固定的子集，Y是根据 DPP 从$\mathcal{Y}$中随机生成的一些点构成的一个子集，则

$$p(A \subset Y) = \det(K_A) \tag{29}$$

其中K是一个$N \times N$的实对称半正定方阵，K_A是K的与A中元素在$\mathcal{Y}$中的标号相对应的矩阵元构成的子方阵，且我们定义$\det K_\phi = 1$，即任意一个 DPP 生成的随机过程选中的点构成的集合都包含空集。K的特征值应处于 0 到 1 的闭区间上。K被称为“边缘核”，因为它确定了 DPP 的边缘分布：

$$p(A \subset Y) = \sum_{Y':Y' \supset A} p(Y = Y') \tag{30}$$

另一种定义是直接针对事件“从$2^{\mathcal{Y}}$中依 DPP 随机选取得到$\mathcal{Y}$的子集Y是A”定义概率，为此引入$N \times N$的实对称半正定方阵L：

$$p(Y = A) = \frac{\det(L_A)}{\det(I + L)} \tag{31}$$

其中L_A是L的与A中元素在$\mathcal{Y}$中的标号相对应的矩阵元构成的子方阵，且其中的分母是根据性质$\sum_{Y \subset \mathcal{Y}} \det(L_Y) = \det(I + L)$而得到的。$L$的特征值不必都大于等于 1。根据式（30）可导出核矩阵L和边缘核K有如下关系：

$$K = L(L + I)^{-1} = I - (L + I)^{-1}$$

$$L = K(I - K)^{-1} \tag{32}$$

作为一个说明 DPP 排斥效应的实例，考察同时包含i, j两点的子集的边缘概率。$p(i \in Y, j \in Y) = \det\begin{pmatrix} K_{ii} & K_{ij} \\ K_{ji} & K_{jj} \end{pmatrix} = K_{ii}K_{jj} - K_{ij}K_{ji} = p(i \in Y)p(j \in Y) - K_{ij}^2$。由此可知，边缘核$K$的元素$K_{ij}$描述的是元素$i$与$j$之间的相似性。对于$i = j$的情况即对角元$K_{ii}$，描述的是该点出现的边缘概率，它越大则说明点$i$在选取出的集合中出现的概率越大；对于$i \neq j$的情况即非对角元$K_{ij}$，描述的是点$i$与点$j$的相似性，并且两点越相似，两点同时出现的概率就会越低。因此我们有了这个排斥的点过程。图 3 为这种排斥性提供了一个形象的说明，展现了独立地随机产生的点过程与行列式点过程的对比。可以发现，独立地随机产生的点会出现随机聚集的现象，而行列式点过程产生的点则有一定的排斥作用，点的分布较为均匀而分散，体现了在特征空间中距离越靠近的两个点（即两点越相似）其同时出现的概率就会越低的特征。

作为 DPP 的一个限制，不论使用K还是L来表示 DPP，都不能模拟点之间有

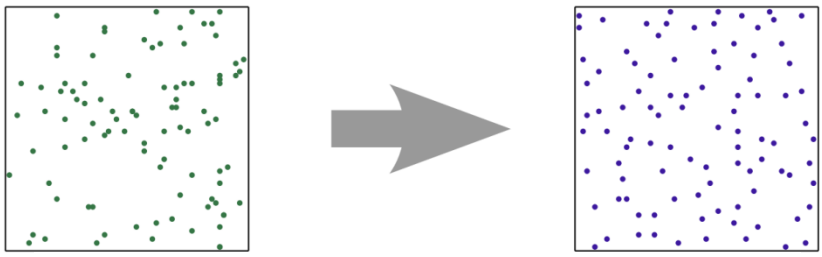

图 3：独立随机点过程（左）和行列式点过程（右）[4]

吸引相互作用的点过程。

3.2. 偏好性-相似性分解

对于核矩阵L，由于它是半正定对称的，所以可以分解为$L = B^T B$的形式，即将L视为一个 Gram 矩阵。其中B是一个$N \times N$的矩阵。我们将B拓展为$F \times N$的矩阵，其中F通常大于N，并将其逐列分解为$B_{\cdot j} = q_j \phi_j$，其中$q_j \in \mathbb{R}^+$表示对点i的偏好程度，$\phi_j \in \mathbb{R}^F$表示点i的特征，$\mathbb{R}^F$是点的特征空间。为使得q的值有意义，需要$\|\phi_j\| = 1$。此时核矩阵L的矩阵元可以写为

$$L_{ij} = q_i \phi_i^T \phi_j q_j = q_i S_{ij} q_j \tag{33}$$

其中S是一个$N \times N$的方阵，其矩阵元为

$$S_{ij} \triangleq \phi_i^T \phi_j \tag{34}$$

对于对角元，有$L_{ii} = q_i \phi_i^T \phi_i q_i = q_i^2$，因而$S$的矩阵元也可以写为

$$S_{ij} = \frac{L_{ij}}{\sqrt{L_{ii} L_{jj}}} \tag{35}$$

注意到$S_{ij} \in [-1,1]$且$S_{ii} = 1$。如前所述，矩阵元L_{ij}表示点i与点j的相似度，对于对角元$L_{ii} = q_i^2$，它单纯表示点i出现的边缘概率，因此分解后只有表示偏好程度的部分。对于非对角元$L_{ij} = q_i \phi_i^T \phi_j q_j$，它表示点$i$与点$j$之间的相似性，这种相似性包括点$i$和点$j$各自的偏好程度$q_i$和$q_j$，以及表示二者特征相似程度的二者特征向量$\phi_i$与$\phi_j$的内积$\phi_i^T \phi_j$。由于$\|\phi_j\| = 1$，因此内积$\phi_i^T \phi_j$即表示两个向量

间夹角的余弦。夹角越小，两个向量越接近，表示两个点在F维特征空间中的特征越相似，则内积越大，即S的矩阵元表示了两点的相似性。实际构造时衡量特征向量之间相似性的方法可以不限于内积。只要能够在特征向量夹角增加时减小、夹角减小时增加的$\mathbb{R}^F \times \mathbb{R}^F \to [-1,1]$映射都可以。

偏好性-相似性分解为 DPP 的构成提供了一个具有明确意义的说明，可以通过这种分解更直观地构造一个 DPP。

4. DPP-MedLDA 模型的建立

本工作中，我们对于话题模型的期望是，它能够得到更加多样的话题，即话题之间的特征差异更大。我们希望能够利用 DPP 来实现这种话题的多样性。在 MedLDA 中，话题β隐含地具有独立性先验，因而有可能会出现话题之间相似的情况。我们将独立性先验替换为具有排斥相互作用的 DPP 先验，这样就为话题引入了增加多样性的因素。下面将对加入 DPP 先验的模型进行描述，并给出它的一个具体的学习算法。

4.1. DPP 先验的构建

为了使用能够使用 DPP 作为话题的先验，我们首先要在K个话题上引入核矩阵L。依式（33）的分解，我们可以分别对话题建立表示偏好的向量q和表示相似性的矩阵S。作为先验，我们单从话题β本身所提供的信息中无法评判哪一个话题出现的概率会比其他的更大，只根据话题本身的信息我们没有理由更偏向于某一个话题；而且在对文档中的每个词进行话题归属操作时，已经有隐参数θ来体现对于实际数据，模型对不同话题的偏好了。因而表示偏好的向量我们就取为K维全一向量$(1,1,\cdots,1)^T$。另一方面，我们单从β本身提供的信息，是可以进行不同话题相似程度的度量的，也就是可以构造出一个表示相似性的矩阵S。我们采用 J. Zou 等人（2012）[5]所使用的相似性度量，即定义相似性矩阵中矩阵元S_{ij}为

$$S_{ij}(\beta)=\frac{\sum_{v=1}^{V}\left(\beta_{iv}\beta_{jv}\right)^{\rho}}{\sqrt{\sum_{v=1}^{V}\beta_{iv}^{2\rho}}\sqrt{\sum_{v=1}^{V}\beta_{jv}^{2\rho}}} \tag{36}$$

来度量话题β_i和话题β_j之间的相似性。可以发现$S_{ii}(\beta)=1$且对于完全相同的β_i和β_j，$S_{ij}(\beta)=1$，以及对于一对在一个话题中会出现的单词在另一个话题中不会出现的两个话题，$S_{ij}(\beta)=0$。

由上面对于表示偏好的向量q的论述，我们以式（36）定义的矩阵S作为核矩阵，并且取$\rho=0.5$，与 J. Zou 等人（2012）[5]使用的相同。此时我们对话题设立的先验便为：

$$p(\beta)\propto \det S(\beta) \tag{37}$$

4.2. 作为正则化项的 DPP 先验

通常的 DPP 作为一个随机过程，它衡量的是从一个基本点集中随机选取一些点构成某个子集的概率。由于 DPP 鼓励多样性，也就是子集所包含的元素越发不相似，通过 DPP 得到这个子集的概率就越高，因而我们可以用 DPP 中得到某个子集的概率来衡量这个子集中元素的不相似性，即这个子集的多样性。而在 MedLDA 中，我们需要的不是抽取出一个由尽量多样的话题构成一个话题的子集，而是度量所有话题之间的多样性。因而我们这里借助 DPP 是想通过它给出的概率来进行这种多样性的度量，而不再需要从核矩阵中抽取矩阵元构成子阵来衡量得到相应子集的概率。因而此时 DPP 得到的概率的数值可以没有绝对意义而只具有衡量多样性的相对意义，所以式（37）中定义的概率度量可以不归一化而直接采用这个数值。

与 J. Zou 等人（2012）[5]提出的问题相同，我们这里也希望能够任意调整先验的效力。参照 J. Zou 等人（2012）[5]的方法，我们可以通过将核矩阵重复Λ次并作为一个对角块矩阵的Λ个对角块，并以这个对角块矩阵为新的核矩阵计算其行列式作为先验概率。重复Λ次的操作会使话题之间的相互作用更强，原本相似的话题在重复多次后，其相似性会变得更加明显，因而会使模型对话题的相似性变得敏感，会更多地减小话题之间的相似性，达到增强 DPP 先验效力的作用。此时的 DPP 先验为新得到的对角块矩阵的行列式，即为：

$$p(\beta) = (\det S(\beta))^{\Lambda} \tag{38}$$

如同在上一段中对 DPP 先验论述的正则化作用那样，我们将$p(\beta)$视作对多样性的度量，因而不进行归一化，所以上式直接取等号。参数Λ此处的作用是控制 DPP 先验效力的参数，因而在实际使用中不限于整数。

将此先验加入到 MedLDA 模型之中，只需要在每一个以β为条件的概率上乘以该先验即可。由式（24）（或式（25））中 MedLDA 的目标函数可知，$\mathcal{L}^{bs}(q(\eta,\theta,z);\alpha,\beta,\delta^2)$这一项需要加入这个先验，进而由式（22）对$\mathcal{L}^{bs}(q(\eta,\theta,z);\alpha,\beta,\delta^2)$的定义可知，对于 DPP-MedLDA 模型，我们需要在目标函数中加入$-\Lambda \ln\det S(\beta)$这一项。由此得 DPP-MedLDA 模型的优化问题为：

$$\min_{q(\eta,\theta,z),\alpha,\beta,\delta^2,\xi,\xi^*} \mathcal{L}^{bs}(q(\eta,\theta,z);\alpha,\beta,\delta^2) - \Lambda \ln\det S(\beta) + C\sum_{d=1}^{D}(\xi_d + \xi_d^*)$$

$$\forall d, s.t.: \begin{cases} y_d - \mathbb{E}_{q(\eta,z)}[\eta^T \bar{z}_d] \le \varepsilon + \xi_d \\ -y_d + \mathbb{E}_{q(\eta,z)}[\eta^T \bar{z}_d] \le \varepsilon + \xi_d^* \\ \xi_d, \xi_d^* \ge 0 \end{cases}$$

（39）

由上式中的目标函数可知，加入 DPP 先验的作用就是在原来的目标函数中加入一个关于β的正则化项，用于鼓励β尽可能多样，这也说明了 DPP 用作正则化项的作用。从上式还可以看出，参数Λ在优化问题中的作用就是正则化系数，用以控制正则化作用的强度。它越大则 DPP 先验的效力就越强，模型就越重视话题的多样性。

4.3. DPP-MedLDA 模型的求解算法

DPP-MedLDA 模型需要求解优化问题式（39）。与 MedLDA 的优化问题式（24）对比可以发现，DPP-MedLDA 模型与它的区别只在优化β上。由式（23）（14）（6）（4）可将优化问题式（39）中与β相关的部分提取出，得到如下的最优化问题来求解β：

$$\min_{\beta} -\sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{k=1}^{K}\sum_{v=1}^{V} w_{dnv}\phi_{dnk}\ln\beta_{kv} - \Lambda\ln\det S(\beta)$$

$$\forall k, \sum_{v=1}^{V}\beta_{kv} = 1$$

$$\forall k, v, \beta_{kv} > 0$$

为体现约束条件，引入拉格朗日乘子$\rho_k, k = 1,\cdots,K$，则此时优化问题化为

$$\max_{\beta}\ L(\beta,\rho) \triangleq \sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{k=1}^{K}\sum_{v=1}^{V} w_{dnv}\phi_{dnk}\ln\beta_{kv} + \Lambda\ln\det S(\beta) + \sum_{k=1}^{K}\rho_k\left(\sum_{v=1}^{V}\beta_{kv}\right)$$

$$\forall k, v, \beta_{kv} > 0$$

（40）

由于目标函数$L(\beta)$中包含行列式，所以很难直接表达出准确解，因而需要使用数值优化的算法。我们进一步还可以发现，目标函数$L(\beta)$的参数维数是$KV + K$，通常是一个很大的数，因而不便采用牛顿法、拟牛顿法等需要计算或近似计算 Hessian 矩阵的优化方法，否则需要存储一个$(KV + K) \times (KV + K)$量级的矩阵，对空间的需求很大。因此我们采用空间消耗较小的且有一定效率的直线搜索类的

优化算法来求解。多种直线搜索优化方法（更多具体的优化方法参见 J. Nocedal 和 S. J. Wright（2006）[17]）都可以完成这个优化任务，我们接下来采用 Polak-Ribière 共轭梯度法[17,18]来求解此优化问题。

为使用共轭梯度法，首先需要计算行列式对数的梯度。我们利用如下结论：

$$\frac{\partial}{\partial \beta_{kv}} \ln \det S(\beta) = \operatorname{tr}\left(S(\beta)^{-1} \frac{\partial S(\beta)}{\partial \beta_{kv}} \right)$$

（41）

可以较为方便地求得行列式对数的梯度。由于$S(\beta)$是一个对称半正定实矩阵，因而可以采用更高效的 Cholesky 分解的方法求逆。目标函数的梯度为

$$\frac{\partial L}{\partial \beta_{kv}} = \sum_{d=1}^{D} \sum_{n=1}^{N_d} \frac{w_{dnv} \phi_{dnk}}{\beta_{kv}} + \Lambda \operatorname{tr}\left(S(\beta)^{-1} \frac{\partial S(\beta)}{\partial \beta_{kv}} \right) + \rho_k$$

$$\frac{\partial L}{\partial \rho_k} = \sum_{v=1}^{V} \beta_{kv} - 1$$

（42）

在使用共轭梯度法优化$L(\beta, \rho)$之前需要先对自变量赋初值。我们采用对应量在 MedLDA 模型中的最优解（$\Lambda = 0$时的最优解）作为初值：

$$\rho_k^{(0)} = -\sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{v=1}^{V} w_{dnv} \phi_{dnk}$$

$$\beta_{kv}^{(0)} = -\frac{1}{\rho_k^{(0)}} \sum_{d=1}^{D} \sum_{n=1}^{N_d} w_{dnv} \phi_{dnk}$$

（43）

为满足优化问题式（40）的约束条件，在实际计算时需引入一个β单个分量的最小值，记为$\beta_{\min}$。在改变β的过程中，若β的某个分量小于$\beta_{\min}$，则需将该分量设为$\beta_{\min}$（下面称此操作为“整理β”）。在求解完梯度之后，应将取得$\beta_{\min}$的β分量对应的梯度分量设为零（下面称此操作为“整理梯度”）以反映在有约束的条件下目标函数的梯度。

采用 Polak-Ribière 共轭梯度法[17,18]求解优化问题式（40）的具体步骤为：

1) 依式（43）为β, ρ赋初值为$\beta^{(0)}, \rho^{(0)}$；
2) 计算目标函数在初值处的函数值以及梯度$\mathcal{Q}^{(0)} = \nabla L\left(\beta^{(0)}, \rho^{(0)}\right)$，
   整理梯度$\mathcal{Q}^{(0)}$，令初始搜索方向为$\mathcal{P}^{(0)} = \mathcal{Q}^{(0)}$，
   设$k = 0$；
3) 只要$\left\|\mathcal{Q}^{(k)}\right\|$大于一定的阈值：
   沿着$\mathcal{P}^{(k)}$方向进行直线搜索操作来更新自变量的值成为$\beta^{(k+1)}$和$\rho^{(k+1)}$，

令$\mathcal{Q}^{(\mathcal{k}+1)}$为目标函数在更新后的自变量处的梯度，整理梯度$\mathcal{Q}^{(\mathcal{k}+1)}$，

令实数$\mathcal{X}^{(\mathcal{k}+1)}=\max\left\{0,\frac{\mathcal{Q}^{(\mathcal{k}+1)^T}(\mathcal{Q}^{(\mathcal{k}+1)}-\mathcal{Q}^{(\mathcal{k})})}{\|\mathcal{Q}^{(\mathcal{k})}\|^2}\right\}$，更新搜索方向为

$\mathcal{P}^{(\mathcal{k}+1)}=\mathcal{X}^{(\mathcal{k}+1)}\mathcal{P}^{(\mathcal{k})}+\mathcal{Q}^{(\mathcal{k}+1)}$，

更新$\mathcal{k}=\mathcal{k}+1$；

4) 停止并返回得到的β和ρ。

上述算法中用到的直线搜索操作是用来沿着给定的搜索方向$\mathcal{P}$来更新自变量使得目标函数在此方向上最大。实际中使用的直线搜索操作不一定要找到给定方向上的最大值，只需找到比出发点一定程度上大的点即可。下面给出一个简单的直线搜索算法：

1) 给定一个初始的搜索长度c以及一个扩张比例因子r（通常是 2），

计算目标函数的值$f^{(0)}$，

令$\mathcal{k}=0$；

2) 依下式更新自变量：$\left(\beta^{(\mathcal{k}+1)^T},\rho^{(\mathcal{k}+1)^T}\right)^T=\left(\beta^{(\mathcal{k})^T},\rho^{(\mathcal{k})^T}\right)^T+c(r-1)\mathcal{P}$，

整理$\beta^{(\mathcal{k}+1)}$，

计算更新后自变量处的目标函数的值$f^{(\mathcal{k}+1)}$，

$c=rc,\mathcal{k}=\mathcal{k}+1$，

只要$f^{(\mathcal{k}+1)}>f^{(\mathcal{k})}$，回到本步骤的开始部分进行循环。

3) 将$\beta^{(\mathcal{k})}$和$\rho^{(\mathcal{k})}$作为优化后的自变量返回，将$f^{(\mathcal{k})}$作为优化后的目标函数值返回，将c/r作为最终的搜索长度返回。

由于此问题中的目标函数其方向导数没有显式的表达形式不便计算（除非使用计算导数的数值方法），因此这里采用不需计算导数的上述直线搜索算法。

5. DPP-MedLDA 模型的实验结果

下面我们将通过一些在实际数据上的实验结果，从定性和定量两个角度展示 DPP-MedLDA 模型的特征和优势。从定性的角度，我们展示使用包含停词的数据和去除停词的数据这两种情况下通过 MedLDA 和 DPP-MedLDA 学习得到的诸话题的十大高频词。从定量的角度，我们使用去除停词的数据，计算在不同的正的正则化系数Λ下模型的预测准确度、话题多样性以及训练用时，考察它们随Λ的变化情况，并将它们与$\Lambda = 0$（即 MedLDA 模型）的结果对比。

我们在实验中使用的去除停词的数据是由 Pang 和 Lee（2005）所编译的“影评数据”（movie review data set）（这个数据也被 J. Zou 等人（2012）[5]使用，用来说明 DPP-LDA 处理回归问题的效果），使用的包含停词的数据是“亚马逊评论数据”（Amazon review data）。这两个数据中的文档响应量都只在 5 个数值中取值，我们对 5 个之中的每一个文档响应量的值都取 1000 个文档用于实验，其中 500 个文档用于训练，而另外 500 个文档用于测试。定性实验中固定话题数$K = 10$，对去除停词和包含停词的数据分别取$\Lambda = 0, 10^5$和$\Lambda = 0, 10^{10}$，得到在各参数下模型学习到的诸话题的十大高频词；定量实验中固定话题数$K = 5$，分别取$\Lambda = 0,\ 10^1, 10^2, 10^3, 10^4, 10^5$，对每一组参数重复进行 6 次实验，计算在这组参数下模型得到的预测准确度、话题多样性和训练用时的均值和标准差，并最终展示不同Λ下三个量的均值和标准差。实验中，$\Lambda = 0$代表不加 DPP 先验的 MedLDA 模型，是我们对比的基准。

5.1. DPP-MedLDA 模型的话题特征

此部分我们比较 MedLDA 模型（即$\Lambda = 0$的 DPP-MedLDA 模型）与加入较强 DPP 先验的 DPP-MedLDA 模型所得到的话题的特征，用得到的诸话题的十大高频词来表示。我们期望有较强 DPP 先验的 DPP-MedLDA 模型与 MedLDA 模型相比能够得到更加多样的话题，这些更多样的话题的十大高频词能够涵盖更多的实词。

首先要展示的是去除停词的“影评数据”的实验结果。取话题数$K = 5$，使用$\Lambda = 0$（即无 DPP 先验，即通常的 MedLDA 模型）与$\Lambda = 10^5$的 DPP-MedLDA 模型学习得到的话题的十大高频词列于表 1 中。

从表 1 可以看出，无 DPP 先验时，不同话题的高频词具有很大的相似性。当$\Lambda = 0$时，未被去除的次级停词“n’t”在全部 5 个话题的十大高频词中都有出

表 1：去除停词的“影评数据”在无 DPP 先验和有较强 DPP 先验的情况下得到话题的十大高频词（$K = 5$）

$\Lambda = 0$					$\Lambda = 10^5$				
T1	T2	T3	T4	T5	T1	T2	T3	T4	T5
hotel	hotel	room	room	beach	room	holiday	room	told	pool
room	good	n't	hotel	resort	hotel	euros	hotel	thought	beach
n't	pool	hotel	n't	great	n't	thing	guests	thing	food
night	food	told	stay	pool	stay	entertainment	beach	finally	resort
pool	n't	stay	area	time	night	dont	food	door	restaurant
rooms	staff	desk	nice	n't	good	fantastic	pool	wall	trip
day	day	back	breakfast	place	day	inclusive	people	move	bar
food	bar	time	pool	day	rooms	didnt	put	paid	kids
staff	room	night	rooms	stay	staff	changed	n't	reservation	times
area	restaurant	front	bed	ocean	area	apartments	door	waiting	ocean

现，而“hotel”“room”“stay”等词在其中 4 个话题中都是出现频率顶尖的词。全部话题的十大高频词包含的不同词的数量较少，话题冗余的现象很明显。而当模型有一个较强的 DPP 先验时，话题的多样性相比之下就会很明显。当$\Lambda = 10^5$时，未被去除的次级停词“dont”“didnt”只在一个话题中出现，次级停词“n’t”在两个话题中出现但在这两个话题中出现的频率有很大差别。高频词“hotel”“room”只在两个话题的十大高频词中出现。全部话题的十大高频词之间重复的词较少，涵盖了更多的词，语义表达能力更强。这些现象与我们的预期相符。

下面要展示的是包含停词的“亚马逊评论数据”的实验结果。取话题数$K = 10$，使用$\Lambda = 0$（即无 DPP 先验，即通常的 MedLDA 模型）与$\Lambda = 10^{10}$的 DPP-MedLDA 模型学习得到的话题的十大高频词列于表 2 中。

由表 2，我们可以更加明显地观察到 DPP 先验的作用。无 DPP 先验时，所有的话题的高频词几乎全部被没有具体含义的停词所占据，并且这些停词在这些话题中的频率排名都十分相似，10 个话题差异性很小，高度冗余。全部 10 个话题的十大高频词涵盖了较少的词，并且基本都是停词，只涵盖了两个具有具体含义的实词“dvd”和“movie”，而且它们出现概率还很小。可以想见这样的话题其表达效力会很差。当为模型加入一个较强的 DPP 先验后，模型学习到的话题有了一个质的飞跃。当$\Lambda = 10^{10}$时，模型可以自动将停词集中在少数几个话题之

表 2：包含停词的“亚马逊评论数据”在无 DPP 先验和有较强 DPP 先验的情况下得到话题的十大高频词（$K = 10$）

$\Lambda = 0$									
T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
the	the	the	the	the	the	the	the	the	the
a	i	i	and	a	br	a	to	and	of
br	and	it	a	of	of	and	and	br	a
of	it	a	br	to	to	of	i	of	and
and	a	and	of	and	and	to	it	a	to
is	to	to	her	is	a	in	a	is	is
dvd	this	movie	to	in	in	is	br	in	in
to	is	that	is	that	that	his	is	to	br
this	movie	of	in	as	it	he	this	this	that
it	that	was	she	he	this	for	my	it	it
$\Lambda = 10^{10}$									
T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
in	give	director	and	the	a	for	video	not	long
so	james	president	is	and	of	quot	duke	have	let
know	whose	stories	br	a	to	book	chair	be	comes
plot	among	tape	it	of	i	show	trained	one	york
through	public	starting	that	to	they	part	pants	by	ago
bit	respect	halloween	this	is	her	disney	shock	from	scott
werewolf	members	fired	was	it	all	work	chasing	at	emotional
performance	attack	narrator	with	i	who	original	bathroom	an	track
fan	hey	disturbed	as	in	she	series	sat	do	poorly
sound	drive	everytime	movie	br	there	version	costume	has	theatrical

中，从而为其他的话题收纳更多的具有具体含义的实词留出空间。具体地说，在表 2 中，T4、T5、T6 和 T9 这四个话题自动地集中了这 10 个话题中的大部分停词，T2、T3、T8、T10 这四个话题的十大高频词中没有出现停词，而 T1 和 T7 这两个话题的十大高频词中最多也只出现了两个停词。全部 10 个话题所涵盖的词包含了更多的具有具体含义的实词，并且这些实词在不同话题的十大高频词中

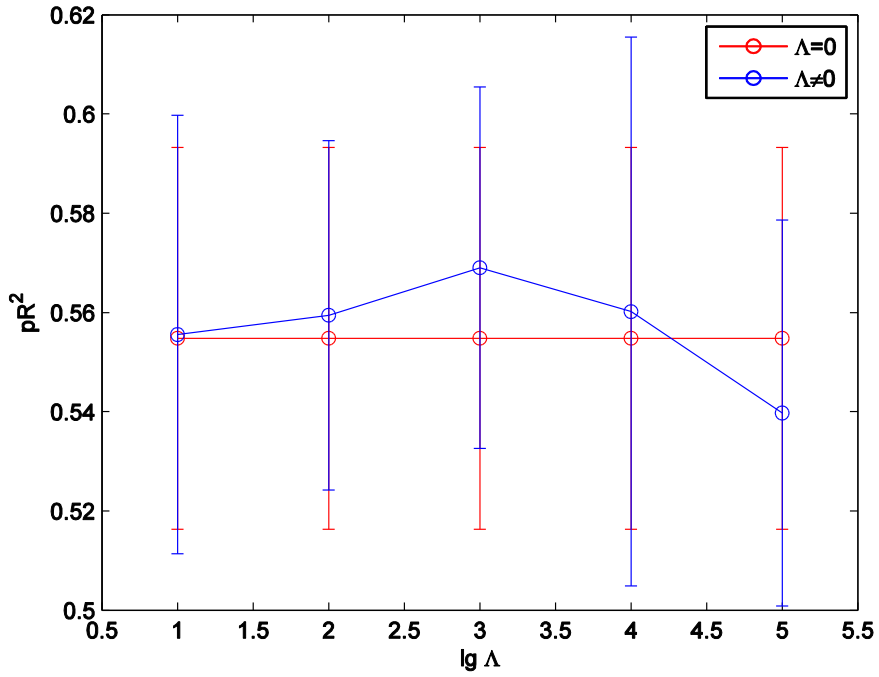

图 4：DPP-MedLDA 模型在不同Λ值下的pR^2值

也几乎没有重复出现。这些现象与无DPP先验时得到的话题形成了鲜明的对比，印证了我们的设想，达到了我们的目的。

5.2. DPP-MedLDA 模型的预测准确度

我们期望 DPP 先验的引入能够带来更好的预测准确度，因为多样性的话题能够使每个话题的语义更加明确集中，作为描述文档的特征更具有表达效力，因而依据这些话题分析文档更能反映出文档的特征，从而可以得到更准确的响应量预测值。

我们采用 D. Blei 等人（2007）[2]以及 J. Zhu 等人（2012）[3]使用的预测R^2值（predictive R^2, pR^2）来衡量响应量预测值的准确度，其定义为：

$$pR^2 \triangleq 1 - \frac{\sum_{d=1}^{D}(y_d - \hat{y}_d)^2}{\sum_{d=1}^{D}(y_d - \bar{y})^2} \tag{44}$$

其中y_d和$\hat{y}_d$分别是文档d的真实响应量和预测响应量，$\bar{y}$是整个测试集上的真实响应量的平均值。pR^2值越大则说明响应量的预测效果越好，$pR^2 = 1$说明预测值与真实值完全相同。

我们将使用“影评数据”在Λ分别取$10^1, 10^2, 10^3, 10^4, 10^5$时重复 6 次实验得到的$pR^2$值的均值和标准差对$\lg\Lambda$作图并与$\Lambda = 0$的情况对比，如图 4 所示。从中

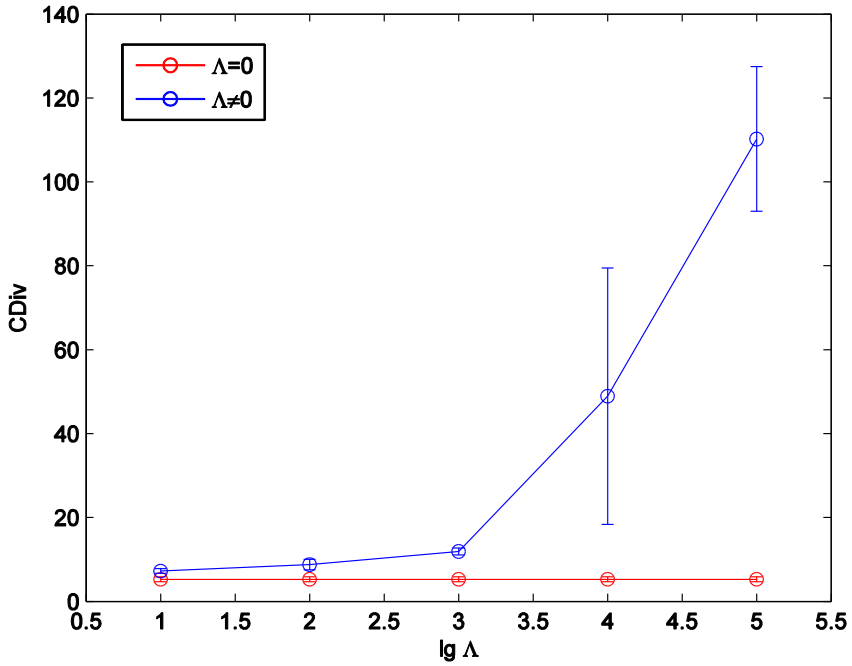

图 5：DPP-MedLDA 模型在不同Λ值下的 CDiv 值

可以看出，对于$\Lambda \neq 0$的情况，其pR^2值大体上都大于$\Lambda = 0$情况下的值，说明 DPP 先验的引入能够有效地提高模型的预测准确度。从图 4 中也可以看出，当Λ较小时pR^2值随Λ的增加而增加，说明此时通过增强 DPP 先验效力造成的话题多样性能够提高预测准确度，与预期相符。当Λ取一个适当大小的值（本实验中对应$\Lambda = 1000$）时，pR^2值取得最大，当Λ超过此值后，pR^2值随着Λ的增加反而减小。这是因为过大的Λ会使鼓励多样性的正则化项对目标函数的影响过大，使得模型一味地过分强调话题的多样性而轻视了模型参数对训练数据的拟合效果，造成预测准确度下降。

5.3. DPP-MedLDA 模型的话题多样性

我们为 MedLDA 模型引入 DPP 先验的目的就是使训练得到的话题更加多样。下面就要来验证这种多样性。在进行模型学习的过程中是用 DPP 先验来衡量话题间多样性的，所以这里我们使用另一种度量分布之间多样性的方法。KL 散度是衡量两个分布间差异的一种常见的手段，考虑到它的非对称性，我们定义如下的交叉散度 CDiv（cross KL divergence）：

$$\mathrm{CDiv} \triangleq \frac{2}{K(K-1)} \sum_{1 \le i \le j \le K} \left(\mathrm{KL}(\beta_i, \beta_j) + \mathrm{KL}(\beta_j, \beta_i) \right) \tag{45}$$

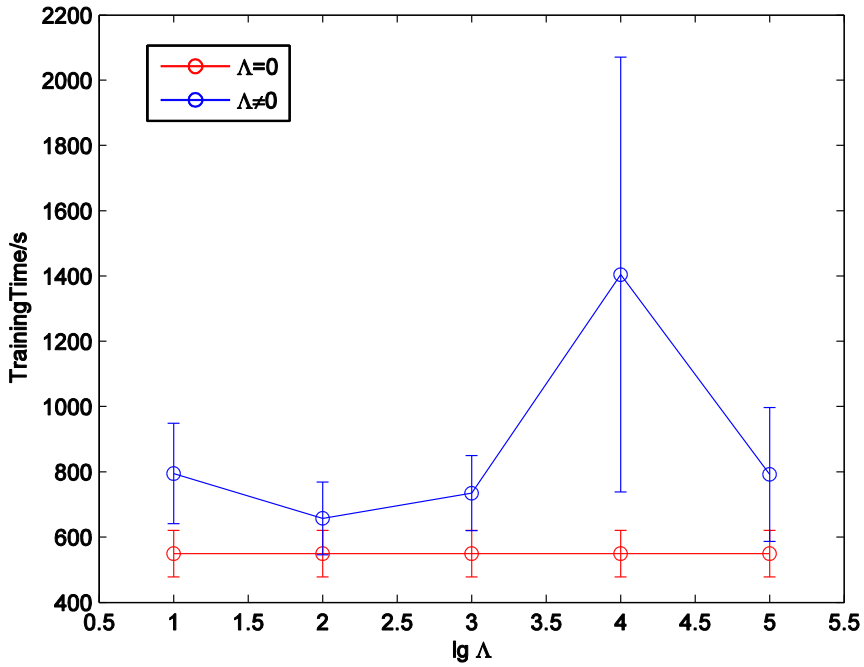

图 6：DPP-MedLDA 模型在不同Λ值下的训练时间

CDiv 值即反映K个话题之间的整体差异性，且它对于这K个话题是对称的。CDiv 值越大则说明分布之间的多样性越大。

我们将使用“影评数据”在Λ分别取$10^1, 10^2, 10^3, 10^4, 10^5$时重复 6 次实验得到的 CDiv 值的均值和标准差对$\lg \Lambda$作图并与$\Lambda = 0$的情况对比，如图 5 所示。从中可以看出，$\Lambda \neq 0$情况下的 CDiv 值总是比$\Lambda = 0$情况下的值大，并且 CDiv 值随着Λ的增加而单调地增加。这说明 DPP 先验的引入可以显著地增加话题的多样性，与我们的预期一致。

5.4. DPP-MedLDA 模型的时间效率

下面我们将考察加入 DPP 先验对于模型训练时间的影响。我们将使用“影评数据”在Λ分别取$10^1, 10^2, 10^3, 10^4, 10^5$时重复 6 次实验得到的训练时间的均值和标准差对$\lg \Lambda$作图并与$\Lambda = 0$的情况对比，如图 6 所示。从中可以看出，加入 DPP 先验后模型的训练时间基本都比不加 DPP 先验时的要长。加入 DPP 先验后，训练模型时每次最大化过程（M-step）中都需要额外增加优化话题β的步骤，因而训练时间基本都会增加。从图 4 中还可以看出，除了$\Lambda = 10^4$的情况下训练时间的均值和标准差有异样的显著增加，在其他情况下，训练时间的均值和标准差的增加量与不加 DPP 先验时的训练时间的均值和标准差都处于同一量级上，并且这些增加量对Λ的依赖关系并不明显。$\Lambda = 10^4$的情况下训练时间的均值和标准

差的显著增加可能是因为 DPP-MedLDA 的最大化过程（M-step）中的求解 SVM 的步骤需要消耗的时间对于β比较敏感，$\Lambda = 10^4$时 DPP 先验对β的修正使得之后的求解 SVM 的步骤消耗的时间很不稳定，造成训练时间的均值和标准差的显著增加。可以尝试通过使用其他方法来代替 SVM 步骤来避免这个训练时间不稳定的现象。

6. 结论与展望

本文提出了用鼓励话题多样性的先验 DPP 改造的 MedLDA 模型。文中定义了衡量话题相关性的矩阵并用于 DPP 先验，加入先验后模型在优化问题中增加了一个鼓励多样性的正则化项，采用共轭梯度法处理这个正则化项来学习话题的参数。DPP-MedLDA 模型在去除停词的“影评数据”和包含停词的“亚马逊评论数据”上的定性实验结果表明，DPP 先验的加入能够有效提高话题的多样性，使全部的话题能够展现更多方面的内容，对于有停词的数据还可以自动地将停词集中在少数几个话题中使其他话题能够表示一定的具体含义。DPP-MedLDA 模型在去除停词的“影评数据”上的定量实验结果表明，DPP 先验的加入可以使得模型在取值于一定范围内的正则化系数Λ下能够取得更好的预测准确度，话题间的“交叉散度”会变得更大，并且模型的训练时间会有一个与无 DPP 先验的训练时间相同量级的增加。这些结果与我们所期望的结果相符。

本论文涉及的工作只对用于回归问题的 MedLDA 模型加入了 DPP 先验，仍需考察用于分类问题的 MedLDA 模型加入 DPP 先验后的性能。可以想见加入 DPP 先验后能够得到更好的分类准确度以及更加多样的话题。另外，可以尝试使用 5.3 节中提出的另一种度量多样性的量 CDiv 来代替 DPP 先验用作正则化项，可以预期这样也能够达到鼓励话题多样性、提高预测准确度的效果。

插图索引

表格索引

参考文献

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993-1022, 2003.

[2] David M. Blei, Jon D. McAuliffe. Supervised topic models. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors. *Advances in Neural Information Processing Systems*, pages 121-128, Cambridge, MA, 2007. MIT Press.

[3] Jun Zhu, Amr Ahmed, and Eric P.Xing. MedLDA: Maximum Margin Supervised Topic Models. *Journal of Machine Learning Research*, 13:2237-2278, 2012.

[4] Alex Kulesza, Ben Taskar. Determinantal Point Process for Machine Learning. *arXiv preprint arXiv:1207.6083v4*, 2013.

[5] James Y. Zou, Ryan P. Adams. Priors for Diversity in Generative Latent Variable Models. *Proceedings in Neural Information Processing Systems*, 3005-3013, 2012.

[6] O. Macchi. The coincidence approach to stochastic point processes. *Advances in Applied Probability*, 7(1):83-122, 1975.

[7] Alexei Borodin, Grigori Olshanski. Distributions on partitions, point processes, and the hypergeometric kernel. *Communications in Mathematical Physics*, 211(2):335-358, 2000.

[8] J. Ben Hough, Manjunath Krishnapur, Yuval Peres, and Blint Virág. Determinantal processes and independence. *Probability Surveys*, 3:206–229, 2006.

[9] Alexei Borodin. Determinantal point processes, *arXiv preprint arXiv:0911.1153*, 2009

[10] Antonello Scardicchio, Chase E. Zachary, and Salvatore Torquato. Statistical properties of determinantal point processes in high-dimensional Euclidean spaces. *Physical Review E*, 79(4), 2009.

[11] Frédéric Lavancier, Jesper Møller, and Ege Rubak. Statistical aspects of determinantal point processes. *arXiv preprint arXiv:1205.4818,* 2012.

[12] Alex Kulesza, Ben Taskar. Structured determinantal point processes. In *Proceedings of Neural Information Processing Systems*, 2010.

[13] Alex Kulesza, Ben Taskar. k-DPPs: fixed-size determinantal point processes. In *Proceedings of the 28th International Conference on Machine Learning,* 2011a.

[14] Alex Kulesza, Ben Taskar. Learning determinantal point processes. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence,* 2011b.

[15] Jennifer Gillenwater, Alex Kulesza, and Ben Taskar. Near-Optimal MAP Inference for Determinantal Point Processes. *Advances in Neural Information Processing Systems*, 25:2744-2752, 2012.

[16] Raja Hafiz Affandi, Emily B. Fox, Ryan P. Adams, and Ben Taskar. Learning the Parameters of Determinantal Point Process Kernels. *arXiv preprint arXiv: 1402.4862v1*, 2014.

[17] Jorge Nocedal, Stephen J. Wright. Numerical Optimization: Springer series in operations research and financial engineering. *Springer-Verlag*, 2006.

[18] Elijah Polak, Gerard Ribière. Note sur la convergence de méthodes de directions conjuguées, *Revue Française d'Informatique et de Recherche Opérationnelle*, 16:35-43, 1969.

致　　谢

由衷感谢我的导师朱军老师。自加入课题组，我就受到朱老师的指导和关怀，特别是在做毕业设计时期，朱老师给了我很多帮助，包括推荐论文的参考文献和相关知识的基础读物，以及对建立模型和实现模型计算方法的指导等。朱老师与我关于之后学术上发展计划的讨论给了我很多启发和对未来工作的信心。朱老师严谨专业的学术功力让我很受鼓舞，对待学生开明负责的态度让我很感激。感谢朱军老师对去除停词的“影评数据”（movie review data set）的预处理工作以及朱军老师和汪一宁同学对包含停词的“亚马逊评论数据”（Amazon review data）的预处理工作。

感谢智能技术与系统国家重点实验室的全体同学和老师在我生活上和科研上的关心和帮助。

声　明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签　名：刘畅　　日　期：2014.6.19

附录 A　外文资料的书面翻译

生成式隐变量模型中鼓励多样性的先验

摘要

隐变量概率模型是机器学习的基石之一。这类模型为确定数据中隐藏结构的先验分布提供了一种方便明了的方法，这样这些位置的性质便可以通过后验推断来获得。这样的模型对于探索性分析和可视化、建立数据的密度模型以及为判别任务提供特征都十分有用。然而，这些模型的一个显著的局限是由先验刻画的特征通常是高度冗余的，因为内部参量的独立同分布假设。例如，混合模型的先验中没有鼓励各部分不重叠的因素，还有话题模型中没有保证只有少数几个话题中会出现同样词的因素。在本工作中，我们回顾一下隐变量概率模型的这些独立性假设，并将潜在的独立同分布先验替换为行列式点过程（DPP）。行列式点过程通过一个正定核函数使我们能够为我们的隐变量确定一种鼓励多样性的偏好。我们可以使用一个建立在不同概率分布之间的核来定义一个行列式点过程用于度量概率。我们将展示在隐式狄利克雷分配模型（LDA）还有混合模型中如何用行列式点过程先验进行最大后验（MAP）推断，并展示用此种方法获得的隐变量更加直观的表示，以及这种方法以不牺牲模型生成式特性的方式对非监督式特征提取的定量结果的改进。

1. 引言

生成式概率模型对于统计学习是个重要的工具，因为它可以实现用较简单的隐式结构来解释巨量的数据。学习得到的这些隐式结构有其特别的有用之处，诸如解释分析数据和数据的可视化，并且对提高未观测数据的预测结果也会有其作用。对于后者，我们可以把推断出的隐式结构视为一种用一种简单的方式概括出复杂的高维相互作用的特征表示。

然而，使用隐变量作为特征的核心假设，是由非监督学习得到的显著统计特征能够为判别任务提供足够的信息。这假设要求这些特征能够张成一个可以包含所有可能出现的数据的空间，且能够表现出可用于判别任务的尽可能多样的特性。而多样性则很难在生成式的框架中表达。大多数时候，模型是在特征表示是独立

的先验的情况下建立的，以期数据的一个好的拟合会需要多样的隐变量。

我们有理由认为这在实际中并不总是会发生的，以及在非监督学习中模型效力的重点通常在于增加常见情况的概率密度，而非建立新特征。例如，对于基于混合分布的生成式聚类模型中，不同的混合分量通常只在数据中的某一个“直觉组”中使用，只是因为这些分量的概率密度的形状与其他组的分布不切合。一个生成式混合模型会使用尽可能多的分量来尽可能好地拟合数据中的某一个组，这就导致了特征表示的高度冗余。类似地，诸如隐式狄利克雷分配模型（Latent Dirichlet Allocation, LDA）[1]的一个话题模型，在用于一个文本文集时，会为分布在多个话题上的相同的停词分配很大的概率密度，以期可以为常见情况尽可能好地精细调整概率。在这两种情况中，我们都希望每一个隐式分组都能够唯一地对应数据的一个特征，即第一种情况中一组数据应由一个混合分量来表示，以及第二种情况中常出现的停词应被集中在某一个组内而不是分散在多个组中。这种想法表达了一种对模型隐参数多样性的需求，而不仅仅是由独立先验所追求的后验分布最大的需求。

本文中，我们提出一种鼓励生成式概率模型中多样性的方法，即通过把隐参数的独立先验替换为行列式点过程（Determinantal Point Process, DPP）先验。行列式点过程具体地实现了在待研究空间中相似性的度量。对于本文中研究的对象，这种空间是一个可由一个正定核表示的隐变量可能的分布的空间。对于这个具体的空间，行列式点过程便可以根据格拉姆矩阵的行列式为这些可能分布的特定结构定义概率。这种构造方法自然地产生了一个偏好多样的隐参数而非冗余的隐参数的生成式隐变量模型。

行列式点过程是一个可以方便地建立可行的带有互排斥作用的点过程的统计工具，它比不能模拟相互作用的泊松过程更具有一般性（例如，参照[2]），同时也比施特劳斯[3]和吉布斯/马尔科夫[4]过程更具有可行性（不过还有只可以模拟负相互作用的代价）。霍夫等人[5]提供了行列式点过程的概率性质的一个很有用的研究，而行列式点过程的统计性质参见斯卡迪秋等人[6]和拉凡西等人[7]等文献。近来也有将行列式点过程应用于机器学习领域来为结构的集合建模[8]，以及有条件地生成一些对象的多样化的集合。我们在此所提出的方法与前面提到的工作不同。我们的方法是将行列式点过程用于层次模型并以此让隐变量拥有多样性，而非直接使观测的离散结构多样化。

2. 生成式隐变量模型中的多样性

本文中我们考虑在N个项目的数据集上生成分布族的一般的隐变量概率模型。N个项目的数据集记为$\{x_n\}_{n=1}^{N}$，其采样空间为$\mathcal{X}$，其每一个数据项都有一个隐式离散标记z_n，其值取自集合$\{1,2,\cdots,J\}$。隐式标记的取值映射到参数集合$\{\theta_j\}_{j=1}^{J}$中。由z_n确定的参数便会根据特定的分布$f(x_n|\theta_{z_n})$来生成数据。通常我们采用独立的θ_j的先验，本文中记为$\pi(\cdot)$，但是隐式变量z_n的分布可能会包含更丰富的结构。考虑以上所述，我们得到如下的一般的联合分布：

$$p\left(\{x_n,z_n\}_{n=1}^{N},\{\theta_j\}_{j=1}^{J}\right)=p(\{z_n\}_{n=1}^{N})\left[\prod_{n=1}^{N}f\left(x_n|\theta_{z_n}\right)\right]\prod_{j=1}^{J}\pi(\theta_j) \tag{A1}$$

每个分布的具体形式由具体问题确定，但这个一般性的框架会经常出现。例如，对于一个典型的混合模型（mixture model），z_n是独立地从一个多项分布中抽取的，且θ_j是确定是哪个分量的参数。对于一个混合模型（admixture model）诸如隐式狄利克雷分配模型（Latent Dirichlet Allocation, LDA）[1]，θ_j是“话题”，或者说是单词上的分布。作为混合特性，z_n会依据例如在文档的常见集合中成为单词来分享相似的结构。

这些模型通常被认为是一种依据一定原则抽取特征的方法。训练时，我们或是找出后验分布$p\left(\{\theta_j\}_{j=1}^{J}\middle|\{x_n\}_{n=1}^{N}\right)$的最大值，或是通过积分或求和消掉由数据确定的隐变量z_n来从后验分布中收集样本。之后在对测试数据x^*进行测试时，我们可以为对应的未知隐参量z^*建立条件分布。此时，z^*便是一个会概括出与x^*相关的诸多信息的“特征”。然而，对模型的这种阐述也许是值得怀疑的，我们没有为模型中的z_n赋予具体含义，当前z_n只是作为用来改善训练似然的副产品而出现的。不同的θ_j可能会为相同的数据分配实际上相同的概率，导致有歧义的特征。

2.1 作为一种度量的行列式点过程

本工作中，我们采取另一种方法来代替标准隐变量模型中的独立性假设。我们不再采用$p\left(\{\theta_j\}_{j=1}^{J}\right)=\prod_j\pi(\theta_j)$，而是在各个分量的分布$\{f(x|\theta_j)\}_{j=1}^{J}$这个集合上建立行列式点过程。利用这个行列式点过程，我们可以通过确定一个建立在各个分量的分布上的正定核函数，来定量地刻画对由重叠最小的各个分布所组成的集

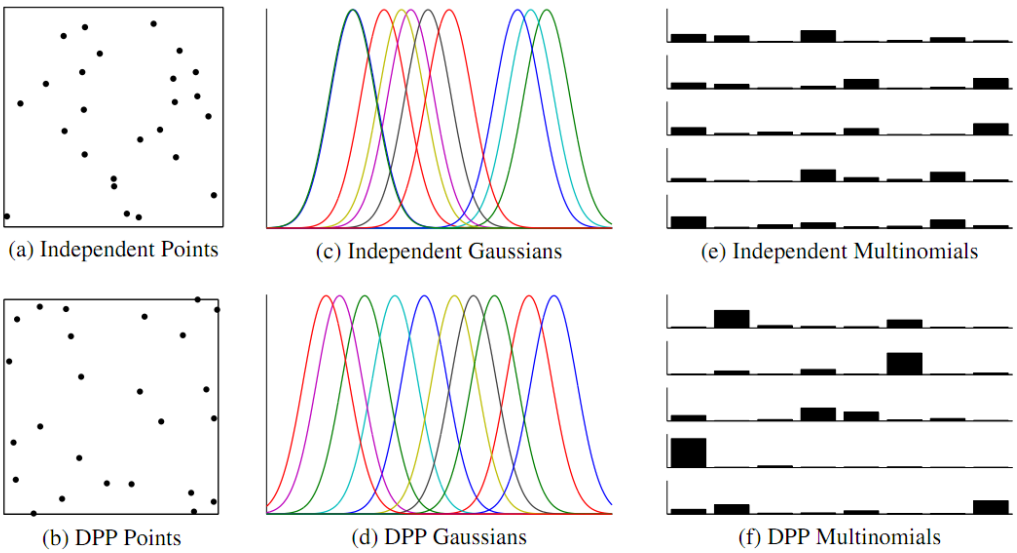

图 A1：对行列式点过程先验的解释。（a）在单位正方形内均匀地相互独立地抽取出的 25 个点；（b）由行列式点过程抽取的 25 个点；（c）十个高斯分布，它们的均值是由单位区间上的均匀分布抽取的；（d）十个高斯分布，它们的均值是由使用概率积核的行列式点过程抽取的；（e）五个随机离散分布；（f）由行列式点过程在概率积核的单形上抽取的五个随机离散分布[10]。

合的偏好。对于我们在此考虑的$f(\cdot)$的简单参数族，行列式点过程为其提供了一组“多样的”参数集$\theta = \{\theta_j\}_{j=1}^{J}$，其中“多样”的含义是针对于样本空间$\mathcal{X}$上概率的度量而言的。对于加入上述这种结构之后进行最大后验推断所得到的结果，我们期望θ_j能够描述$\mathcal{X}$中由似然恰当调整出来的、有着本质区别的各个区域，进而能够在测试时提取出得到改进的、非冗余的特征。

记Θ为θ的取值空间。则定义在Θ上的一个具体的点过程会随机地产生一个Θ的有限子集。为建立一个行列式点过程，我们首先在Θ上定义一个正定核，记为$K: \Theta \times \Theta \to \mathbb{R}$。一个特定的有限子集$\theta \subset \Theta$的概率密度便可由下式确定：

$$p(\theta \subset \Theta) \propto |K_\theta| \tag{A2}$$

其中K_θ是一个$|\theta| \times |\theta|$的正定格拉姆矩阵，它是由将核$K(\theta, \theta')$作用于θ中的元素而得到的。Θ上的核的特征谱必须在区间$[0,1]$中。本文中我们使用的核可分解为两部分：1）一个正定的相关函数$R(\theta, \theta')$，满足$R(\theta, \theta) = 1$，以及 2）一个“先验核”

$\sqrt{\pi(\theta)\pi(\theta')}$，用以表示我们对某些参数的边缘偏好（总的偏好）强于其他的。这两部分组合为我们感兴趣的核：

$$\mathrm{K}(\theta,\theta') = \mathrm{R}(\theta,\theta')\sqrt{\pi(\theta)\pi(\theta')} \tag{A3}$$

或改写为矩阵形式$\mathrm{K}_\theta = \Pi \mathrm{R}_\theta \Pi$，其中$\Pi = \mathrm{diag}(\left[\sqrt{\pi(\theta_1)}, \sqrt{\pi(\theta_2)}, \cdots\right])$。

可以发现，若对于所有的$\theta \neq \theta'$都有$\mathrm{R}(\theta,\theta') = 0$，那么这种构造方式重现了权重为$\pi(\theta)$的泊松过程，并且在这种情况下，如果$\theta$的势（其中的元素数）是提前确定的，那么它就又重现了传统的独立先验。不过更有趣的是由非对角结构的$\mathrm{R}(\theta,\theta')$所产生的集合内元素间的相互作用。这样的核总是会产生点之间的排斥效应，使得在这种先验下Θ的多样化的子集会倾向于拥有更大的概率。参看图 A1 对于独立性样本与几种不同参数下的行列式点过程之间的区别的解释。

2.2 用于概率分布的核

行列式点过程的框架使我们能够建立模拟排斥效应的生成式模型，但如同其他的基于核的先验，我们必须定义什么叫做“排斥”。现在有用于度量概率的一大类正定函数被定义了出来，在本工作中我们将使用概率积核[10]。其基本核具有如下形式：

$$K(\theta,\theta';\rho) = \int_x f(x|\theta)^\rho f(x|\theta')^\rho \mathrm{d}x \tag{A4}$$

其中$\rho > 0$。我们所需要的相关核便使用如下归一化的形式：

$$R(\theta,\theta';\rho) = K(\theta,\theta';\rho)/\sqrt{K(\theta,\theta;\rho)K(\theta',\theta';\rho)} \tag{A5}$$

这个核对于许多常见分布都具有方便的解析形式，因此它是当前模型的一个理想的构造模块。

2.3 叠加的行列式点过程

我们通常会期望先验分布具有可以任意强的性质。也就是说，按照贝叶斯先验的解释“先前观测数据的影响”，我们想要可以任意调节这种先前观测数据的量，并可以在量合适时建立一个信息量大的先验。不幸的是，标准的行列式点过程没有提供可以任意增加其强度的手段。

举例来说，取一个定义在欧几里得空间上的行列式点过程，并考虑一个点t、一个任意的单位向量w以及小的标量ϵ。利用一个参数$\delta > 1$来构造如下两对点：一对“靠近”的点$\{t, t+\epsilon w\}$，和一对“远离”的点$\{t, t+\epsilon\delta w\}$。我们希望找到一个小的$\epsilon$使得在行列式点过程下“远离”情况的可能性可以比“靠近”情况的任意地大，也就是说，我们希望如下的行列式的比值

$$r(\epsilon) = \frac{p(\{t, t+\epsilon\delta w\})}{p(\{t, t+\epsilon w\})} = \frac{1 - R(t, t+\epsilon\delta w)^2}{1 - R(t, t+\epsilon w)^2} \tag{A6}$$

在ϵ趋于零的时候无界。我们的目标是得到一个控制量级的参量使行列式点过程的作用与似然函数中的各项相比可以任意地强。如果我们在$\epsilon = 0$附近对分子分母进行泰勒展开，我们有

$$r(\epsilon) \approx \frac{1 - \left(R(t,t) + 2\delta w\epsilon\left[\frac{d}{dt}R(t,\tilde{t})\right]_{\tilde{t}=t}\right)}{1 - \left(R(t,t) + 2w\epsilon\left[\frac{d}{dt}R(t,\tilde{t})\right]_{\tilde{t}=t}\right)} = \delta \tag{A7}$$

我们可以发现，在零点附近，这个比值体现出了距离上的不同，但却不是以一种可以重新调整到更强作用的方式体现的。这意味着无论用什么样的行列式点过程先验都会有一些有限的数据集我们不能让它们凸显出来。为解决这个问题，我们引入一个参数$\lambda > 0$来推广行列式点过程，定义有限子集$\theta \subset \Theta$的概率为

$$p(\theta \subset \Theta) \propto |K_\theta|^\lambda \tag{A8}$$

对于λ为整数的情况，上式可以被看作一组有λ个完全相同的行列式点过程的“叠加实体”，这样也与我们生成式的观点相容。θ的叠加就是$\theta_\lambda = \{\lambda \text{ 个 } \theta\}$，对应的核$K_{\theta_\lambda}$就是一个$\lambda|\theta| \times \lambda|\theta|$的对角块矩阵，其中每一个对角块就是一个$K_\theta$。这与先验是赝数据的观点对应得很好：我们叠加的行列式点过程说明我们观测到λ次之前的数据集。如同其他的赝计数先验那样，λ在实际中不必为整数，并且在利用带惩罚项的对数似然进行最大后验推断的框架下，λ可以被视作一个增加行列式惩罚力度的参数。

2.4 用作正则化的行列式点过程

除了可被视作生成式框架下分布的先验，行列式点过程还可被视作进行学习时一种新的鼓励“多样性”的正则化项。我们希望求解

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}\subset\Theta}{\operatorname{argmin}}\, \mathcal{L}(\boldsymbol{\theta}; \{x_n\}_{n=1}^N) - \lambda \ln |\mathbf{K}_{\boldsymbol{\theta}}|$$

（A9）

即从Θ中选择最好的参数集$\boldsymbol{\theta}$。这里的$\mathcal{L}(\cdot)$是一个损失函数，它依赖于数据以及判别函数，并带有参数$\boldsymbol{\theta}$。由式（A3）

$$\ln |\mathbf{K}_{\boldsymbol{\theta}}| = \ln\ |\mathbf{R}_{\boldsymbol{\theta}}| + \sum_{\theta_j\in\boldsymbol{\theta}} \ln \pi(\theta_j)$$

（A10）

如果$\mathcal{L}(\cdot) = -\ln p(\{x_n\}_{n=1}^N|\boldsymbol{\theta})$，那么得到的优化结果就是最大后验估计。在此框架下，我们可以将行列式点过程惩罚与任何其他的θ的正则化项结合，比如产生稀疏性的ℓ_1正则化项。在下面的各部分中，我们会给出一些经验结果来说明这种多样性会改善模型的整体表现。

3. 最大后验推断

下面我们将固定$|\boldsymbol{\theta}|$。此时可以将核$\mathbf{K}_{\boldsymbol{\theta}}$视为是$\boldsymbol{\theta}$的一个函数，其梯度为$\frac{\partial}{\partial\theta}\log|\mathbf{K}_{\boldsymbol{\theta}}| = \operatorname{trace}(\mathbf{K}_{\boldsymbol{\theta}}^{-1}\frac{\partial\mathbf{K}_{\boldsymbol{\theta}}}{\partial\theta})$。这使我们在进行推断时可以使用一般的基于梯度的优化算法。特别地，我们可以通过使用期望最大化（Expectation Maximization, EM）算法将优化$\boldsymbol{\theta}$的过程作为一个模块部分。下面我们以两个可以直接插入行列式点过程先验的生成式隐变量模型为例来说明。

多样化的隐式狄利克雷分配模型

隐式狄利克雷分配模型（LDA）[1]是应用十分广泛的用于文本的混合模型，并渐渐应用于可以被视作“单词包”的其他类型的数据。隐式狄利克雷分配模型建立了一系列话题——在词表中单词上的分布——并认为文集中的每个单词应归属于这些话题中的一个。这种话题-单词的关系是不被观测到的，但隐式狄利克雷分配模型通过要求对任意给定的文档只有少数的话题会被呈现，来找到文档的结构。

在标准的隐式狄利克雷分配模型中，话题是K个分布在大小为V的词库上的离散分布β_k，β_{kv}表示话题k产生单词v的概率。文档一共有M个，且第m个文档有N_m个单词。文档m有一个分布在话题上的隐式多项分布，其参数记为θ_m，则这个文档中的每个词w_{mn}都有一个抽样自θ_m的话题下标z_{mn}。经典的隐式狄利克雷分配模型对β_k使用独立的狄利克雷先验，这里我们要将隐式狄利克雷分配模型“多样

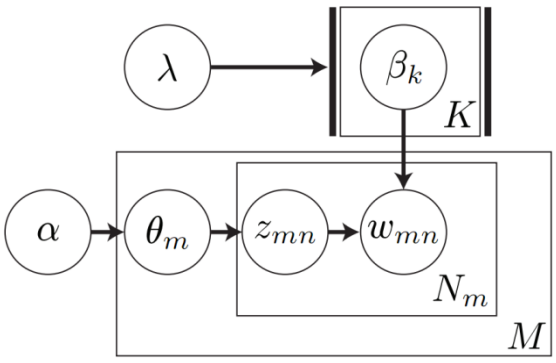

图 A2：行列式点过程-隐式狄利克雷分配模型的结构。我们将隐式狄利克雷分配模型的独立同分布的话题的标准盘记号替换为一个“双线盘”来表示一个行列式点过程。

化”，将原来的先验置换为行列式点过程。也就是说，我们引入一个相关核

$$R(\beta_k, \beta_{k'}) = \frac{\sum_{v=1}^{V}(\beta_{kv}\beta_{k'v})^{\rho}}{\sqrt{\sum_{v=1}^{V}\beta_{kv}^{2v}}\sqrt{\sum_{v=1}^{V}\beta_{k'v}^{2\rho}}} \tag{A11}$$

当β_k与$\beta_{k'}$越发相似时会趋于 1。在下面的行列式点过程-隐式狄利克雷分配模型中，我们设定$\rho = 0.5$。我们采取$\pi(\beta_k) = \text{Dirichlet}(\alpha)$，并将得到的先验写作$p(\boldsymbol{\beta}) \propto |\mathbf{K}_{\boldsymbol{\beta}}|$。我们称此模型为“行列式点过程-隐式狄利克雷分配模型”，并在图 A2 中以图模型来表示它。我们在图模型中用一个“双线盘”来表示行列式点过程，以此突出它是如何为代替独立性假设而引入的。

为了对此模型做最大后验学习，我们构造一个变分期望最大化算法的变体算法。如同隐式狄利克雷分配模型中的期望最大化算法，我们定义一个因子化的近似

$$q(\theta_m, z_m|\gamma_m, \phi_m) = q(\theta_m|\gamma_m)\prod_{n=1}^{N} q(z_{mn}|\phi_{mn}) \tag{A12}$$

在这种近似下，对于每一个文档m，话题后验分布都有一个狄利克雷分布的近似，这个分布由其参数γ_m指定。ϕ_m是一个$N \times K$的矩阵，其第n行，记为ϕ_{mn}，是一个对单词w_{mn}所属话题的多项分布。对于当前β_{kv}的估计值，γ_m和ϕ_m由迭代

法优化。具体细节参见布雷等人[1]。我们为包含行列式点过程而对变分期望最大化算法所做的拓展不会改变这些步骤。

不过，行列式点过程的引入确实影响变分期望最大化算法的最大化步骤。鼓励多样性的先验的引入会新加入一个对$\boldsymbol{\beta}$的惩罚项，所以此时最大化步骤需要求解

$$\boldsymbol{\beta}^* = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left\{ \sum_{m=1}^{M} \sum_{n=1}^{N_m} \sum_{k=1}^{K} \sum_{v=1}^{V} \phi_{mnk} w_{mn}^{(v)} \ln \beta_{kv} + \lambda \ln |\mathbf{K}_{\boldsymbol{\beta}}| \right\} \tag{A13}$$

同时满足$\boldsymbol{\beta}$的每一行求和应为 1 的约束。对于$\lambda = 0$的情况，这个优化过程便退化为原始的隐式狄利克雷分配模型中的标准更新过程，即$\beta_{kv}^* \propto \sum_{m=1}^{M} \sum_{n=1}^{N_m} \phi_{mnk} w_{mn}^{(v)}$。对于$\lambda > 0$的情况，我们采用梯度下降法来求解局部最优的$\boldsymbol{\beta}$。

多样化的高斯混合模型

这个混合模型对于生成式聚类和密度估计等任务是一个很常用的模型。给定J个分量，数据的概率为

$$p(x_n|\boldsymbol{\theta}) = \sum_{j=1}^{J} \chi_j f(x_n|\theta_j) \tag{A14}$$

θ_k通常在其先验中被视作独立的。现在我们要考察θ_k的行列式点过程先验在分量为高斯分布的情况下的性质。

行列式点过程先验对于高斯混合模型有特别的可行性。如同行列式点过程-隐式狄利克雷分配模型那样，我们采用概率积核，它在此种情况下也有一个方便的解析形式[10]。令$f_1 = \mathcal{N}(\mu_1, \Sigma_1)$，$f_2 = \mathcal{N}(\mu_2, \Sigma_2)$为两个高斯分布，则积核为：

$$K(f_1, f_2) = (2\pi)^{(1-2\rho)\frac{D}{2}} \rho^{-\frac{D}{2}} \left|\hat{\Sigma}\right|^{\frac{1}{2}} (|\Sigma_1||\Sigma_2|)^{-\frac{\rho}{2}} \exp\left(-\frac{\rho}{2}(\mu_1^T \Sigma_1^{-1} \mu_1 + \mu_2^T \Sigma_2^{-1} \mu_2 - \hat{\mu}^T \hat{\Sigma} \hat{\mu})\right)$$

其中$\hat{\Sigma} = (\Sigma_1 + \Sigma_2)^{-1}$且$\hat{\mu} = \Sigma_1^{-1}\mu_1 + \Sigma_2^{-1}\mu_2$。在方差矩阵$\sigma^2 I$是固定且各项同性的、且$\rho = 1$的特殊情况下，核可以写为

$$K\big(f(\cdot\,|\mu), f(\cdot\,|\mu')\big) = \frac{1}{(4\pi\sigma^2)^{\frac{D}{2}}} e^{-\frac{\left\|\mu-\mu'\right\|^2}{4\sigma^2}}$$

（A15）

其中D是数据的维数。

在高斯混合模型的标准的期望最大化算法中，通常引入隐式的二值变量z_{nj}，表示数据n属于分量j。期望步骤会计算出表示权值的向量$\gamma(z_{nj}) = \mathbb{E}[z_{nj}] \propto \chi_j \mathcal{N}(x_n|\mu_j, \Sigma_j)$。这个步骤同样适用于行列式点过程-高斯混合模型。更新分量权重的步骤也是一样的：$\chi_j = \frac{1}{N}\sum_{n=1}^{N}\gamma(z_{nj})$。这个算法和标准的期望最大化算法的区别是，行列式点过程-高斯混合模型的最大化步骤需要优化如下的目标函数（为使公式清晰，已用$\boldsymbol{\theta}$将$\{\mu_j, \Sigma_j\}_{j=1}^{J}$求和消去）：

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}} \left\{ \sum_{n=1}^{N}\sum_{j=1}^{J} \gamma(z_{nj})[\ln\chi_j + \ln\mathcal{N}(x_n|\mu_j, \Sigma_j)] + \lambda \ln|\mathbf{K}_{\boldsymbol{\theta}}| \right\}$$

（A16）

与行列式点过程-高斯混合模型紧密相关的模型是行列式点过程-K平均模型。它的核按照与式（A15）相同的形式作用于中心点的集合，此时σ^2只是一个控制量级的常数。令$\boldsymbol{\theta} = \{\mu_j\}$以及$z_{nj}$为强赋值的指示量，则最大化步骤为：

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}} \left\{ \sum_{\boldsymbol{n=1}}^{\boldsymbol{N}}\sum_{\boldsymbol{j=1}}^{\boldsymbol{J}} z_{nj}\left\|x_n - \mu_j\right\|^2 + \lambda \ln|\mathbf{K}_{\boldsymbol{\theta}}| \right\}$$

（A17）

根据积核，两个高斯分布之间的相似性随着它们均值间的距离的增加而指数地衰减。实际中，我们发现，当混合分量的数量$|\boldsymbol{\theta}|$大的时候，$\mathbf{K}_{\boldsymbol{\theta}}$可由一个稀疏矩阵来很好地近似。

4. 实验 1：多样化的话题模型

我们在未经筛选的 20 个新闻组的文集上不去除任何停词地测试隐式狄利克雷分配模型以及行列式点过程-隐式狄利克雷分配模型。原始的隐式狄利克雷分配模型的一个常见的令人不满的缺陷是，将隐式狄利克雷分配模型用于未经筛选的数据所得到的话题都会被停词所抢占。即使话题数从$K = 5$变化到$K = 50$，这种

表 A1：由隐式狄利克雷分配模型和行列式点过程-隐式狄利克雷分配模型学习到的表示话题中的最常出现的 10 个单词。

LDA		DPP-LDA					
typical		”stop words”		”Christianity”	”space”	”OS”	”politics”
the	the	the	and	jesus	space	file	ms
to	to	of	in	matthew	nasa	pub	myers
and	and	that	at	prophecy	astronaut	usr	god
in	it	you	from	christians	mm	available	president
of	of	by	some	church	mission	export	but
is	is	one	their	messiah	pilot	font	package
it	in	all	with	psalm	shuttle	lib	options
for	that	but	your	isaiah	military	directory	dee
that	for	do	who	prophet	candidates	format	believe
can	you	my	which	lord	ww	server	groups

令人沮丧的现象还是会发生。表 A1 的前两列列出了使用$K = 25$的隐式狄利克雷分配模型学习到的话题中的两个话题中的最常出现的 10 个单词。停词会经常出现在所有的文档中，所以它们无用地关联了各话题中包含关键信息的词。我们将 669 个最常出现的停词去除后再次进行实验，由常规的隐式狄利克雷分配模型推断出的话题还是会被没有信息量的次常见停词所占据。

行列式点过程-隐式狄利克雷分配模型可以自动地将停词归类到少数几个话题中。通过寻找停词构成的话题，剩下话题的大部分都可用来接纳更有信息量的单词。表 A1 列出了从由未经筛选的 20 个新闻组的文集中学习（$K = 25, \lambda = 10^4$）得到的行列式点过程-隐式狄利克雷分配模型中抽样出来的话题。当我们变化K或增加λ时我们可以观察到鲁棒的稳定的将停词分类到少数几个话题中的效果。在β的分布的积核的度量下，诸多话题中都十分常见的高频单词会显著地增加话题间的相似性。这种相似性会在行列式点过程中产生很大的惩罚项，所以目标函数会积极地迫使隐式狄利克雷分配模型的参数偏离会导致停词占据很多话题的大概率名额这样的后果的区域。

由行列式点过程-隐式狄利克雷分配模型学习出来的特征可以使文档分类的效果更好。我们通常使用γ_m，也就是对应每个文档的话题的后验分布，作为文档分类的特征向量。我们先得到对训练文档使用行列式点过程-隐式狄利克雷分配模型的变分期望最大化算法推断出来的$\{\gamma_{m,train}\}$，再对$\{\gamma_{m,train}\}$用 20 个新闻组中真实的话题标记训练出一个支持向量机（Support Vector Machine, SVM）分类器。对于测试文档，我们将参数α和β固定为在训练集上推断得到的值，并使用变分期望

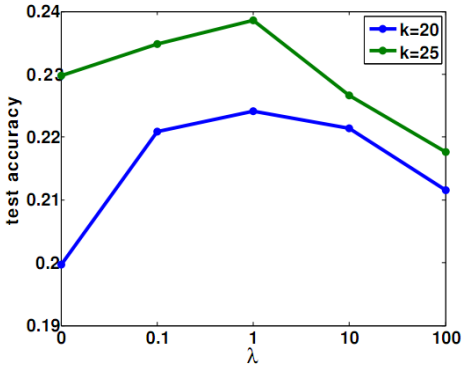

图 A3：λ对分类误差的影响

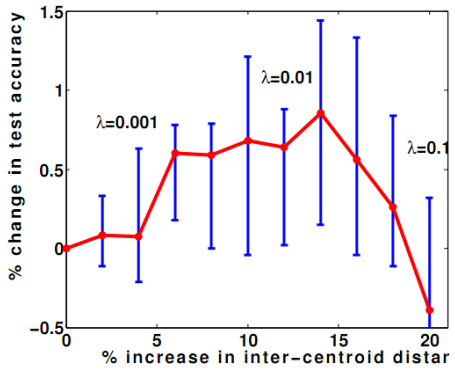

图 A4：聚类中心距离对测试误差的影响

最大化算法来找到$\{\gamma_{m,test}\}$最大后验估计。λ在一定区间上变化时其对应的平均测试分类准确度如图 A3 所示。$\lambda = 0$的情况对应原始的隐式狄利克雷分配模型。对于每次实验，我们为在一定区间上变化的λ对应的行列式点过程-隐式狄利克雷分配模型都使用同一个训练集。$\lambda = 1$的行列式点过程-隐式狄利克雷分配模型在分类测试中始终的表现始终优于隐式狄利克雷分配模型（$p < 0.001$，二项分布测试）。过大的λ值会降低分类的表现。

5. 实验 2：多样化的聚类

混合模型是一种为分类任务学习特征的常用而有效的方法。例如，寇斯等人[11]最近的工作表明，即使是简单的K-平均，作为一种为图像标记而抽取特征的方法也十分有效。在此工作中，K-平均给出了 CIFAR-10 物体识别任务中的艺术结果的状态。寇斯等人用一个分面片的过程得到的这些结果，其中的面片是从图片中随机抽样得到的，用以进行训练。每一个面片都是一个 6 乘 6 的方块，表示为一个 36 维向量空间中的点。来自训练用图片的面片用K-平均的方法组合并聚类。于是每个面片便可表示为一个K维二值特征向量：如果面片距中心k的距离比它距所有中心的平均距离小，那么这个向量的第k个元素便为 1。大体上说，这个特征向量有一半的特征元素都是零。于是来自同一个图片的面片就会被聚合起来去构造整个图片的一个特征向量。对这些图片特征训练一个支持向量机来进行分类。

行列式点过程-K-平均会产生更加有信息量的特征，因为聚类中心会相互排斥，从而在像素空间中占据差别更大的位置。我们要重现寇斯等人的实验，使用他们公布的代码进行完全相同的前期与后期处理。在此设定下，$\lambda = 0$会复现通常的K-

表 A2：CIFAR-10 数据集的测试分类准确度

training set size	K	K-means	DPP K-means	gain (%)	λ
500	30	34.81	**36.21**	1.4	0.01
1000	30	43.32	**44.27**	0.95	0.01
2000	60	52.05	**52.55**	0.50	0.01
5000	150	61.03	**61.23**	0.20	0.001
10000	300	66.36	**66.65**	0.29	0.001

平均，得到与冦斯等人[11]相同的结果。我们将行列式点过程-K-平均用于 CIFAR-10 数据集，同时会改变训练集的大小。对于每一个训练集大小，我们对于在一个范围内变化的K运行常规的K-平均算法，并由此选择出用于K-平均的能够给出最好准确度的K。然后我们将此结果与使用相同的K的行列式点过程-K-平均的结果进行对比。对于训练集中多达 10000 张图片的数据，行列式点过程-K-平均都可以产生比简单K-平均更好的测试分类准确度。这种比较基于相互匹配的设定，即我们是对同样一个给定的随机抽取的训练集和一个聚类中心的初始化，分别进行K-平均和行列式点过程-K-平均来产生聚类中心的。得到的这两组聚类中心被用来提取特征和训练分类器，然后对得到的两个分类器在相同的图片测试集上进行测试。就测试的准确度而言，行列式点过程-K-平均总是比K-平均产生更好的结果（$p < 0.001$，二项分布测试）。例如，对于大小为 1000 的训练集，取$k = 30$，我们进行 100 次实验，每次都使用一个随机抽取的训练集和随机的初始化，有 94 次的实验行列式点过程-K-平均的结果优于K-平均。正如我们期望的那样，作为正则化项，行列式点过程带来的改进会在更小的训练集上取得更显著的效果。对于全部的 CIFAR-10 数据集的 50000 张训练图片，行列式点过程-K-平均并不总是优于K-平均。

接下来我们会关心行列式点过程分离聚类中心的程度和测试集的分类准确度之间是否存在某个模式。对于 1000 张训练图片并取$k = 30$，对于每一个随机抽取的训练集和聚类中心的随机初始化，计算K-平均和行列式点过程-K-平均的聚类中心间的平均距离。我们再计算每个聚类中心集的测试准确度。图 A4 将聚类中心间距离的相对增量分为 10 组，对于每一组，我们显示出测试准确度变化的 25%、50%和 75%的分位数。测试准确度在聚类中心间距离比K-平均的聚类中心间距离增加 14%的时候最大，对应着$\lambda = 0.01$。

6. 讨论

我们介绍了一种将鼓励多样性的偏好引入生成式概率模型中的普遍的方法。我们展示了如何将行列式点过程作为能够加入到已有的学习算法中的集成模块，并讨论了行列式点过程作为鼓励多样性的正则化项的作用。我们考察了两种对多样性有需求的情况：从文档中学习话题，和图像面片的聚类。将行列式点过程加入到隐式狄利克雷分配模型中可使隐式狄利克雷分配模型能够自动地将停词聚集到少数几个类别之中，使得其他类别能够体现更加有信息量的话题。在文档和图片的分类任务中，都存在一个鼓励多样性的中间体系（由超参数λ控制），使得与标准的独立同分布的模型相比能有稳定持续的改进。计算$M \times M$的核矩阵$\mathbf{K}$的逆矩阵可能是一个计算上的瓶颈，其中M是隐式分布的数量。不过在许多情况下，比如隐式狄利克雷分配模型，M都会比数据大小小得多。我们预期，基于行列式点过程的多样性可以有效地加入到更多的生成式概率模型中，例如可以应用于隐式马尔科夫模型或更一般的时间序列的发射参数，或是应用于迁移学习的机制之中。

参考文献

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[2] J. F. C. Kingman. *Poisson Processes.* Oxford University Press, Oxford, United Kingdom, 1993.

[3] David J. Strauss. A model for clustering. *Biometrika*, 62(2):467–475, August 1975.

[4] Jesper Møller and Rasmus Plenge Waagepetersen. *Statistical Inference and Simulation for Spatial Point Processes.* Monographs on Statistics and Applied Probability. Chapman and Hall/CRC, Boca Raton, FL, 2004.

[5] J. Ben Hough, Manjunath Krishnapur, Yuval Peres, and Blint Virág. Determinantal processes and independence. *Probability Surveys*, 3:206–229, 2006.

[6] Antonello Scardicchio, Chase E. Zachary, and Salvatore Torquato. Statistical properties of determinantal point processes in high-dimensional Euclidean spaces. *Physical Review E*, 79(4), 2009.

[7] Frédéric Lavancier, Jesper Møller, and Ege Rubak. Statistical aspects of determinantal point processes. http://arxiv.org/abs/1205.4818, 2012.

[8] Alex Kulesza and Ben Taskar. Structured determinantal point processes. In *Advanced in Neural Information Processing Systems 23*, 2011.

[9] Alex Kulesza and Ben Taskar. Learning determinantal point processes. *In Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence,* 2011.

[10] Tony Jebara, Risi Kondor, and Andrew Howard. Probability product kernels. *Journal of Machine Learning Research,* 5:819–844, 2004.

[11] Adam Coates Honglak Lee and Andrew Ng. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics,* 2011.

原文索引

James Y. Zou, Ryan P. Adams. Priors for Diversity in Generative Latent Variable Models. *Neural Information Processing Systems*, 3005-3013, 2012

综合论文训练记录表

学生姓名	刘畅	学号	2010012176	班级	基应01
论文题目	以行列式点过程为先验的MedLDA模型				

主要内容以及进度安排

主要内容：MedLDA是一种监督式隐变量话题模型，并采用了最大间隔的思想来学习参数。它可以发掘出文档中的话题结构。行列式点过程(DPP)是一种可以很好地模拟多样性的模型。为了使话题模型更有效，即发掘出更加多样的话题，将DPP与话题模型结合是一个很直接的想法。目前已有DPP与LDA结合的工作，并取得了预期的效果。本工作是将DPP与MedLDA结合，将DPP作为MedLDA的先验，并通过实验验证其效果。预期能取得可以获得更加多样、更有针对性的模型。

进度安排：

1-3周：文献调研
4-6周：理论推导，完成算法设计
7-8周：程序和数据搜集
9-14周：实验实施
15-16周：实验结果分析，论文写作

指导教师签字：

考核组组长签字：

2014 年 3 月 13 日

中期考核意见

该同学工作较积极主动，较好地完成了所规定任务。

考核组组长签字：

2014 年 4 月 16 日

指导教师评语	刘畅同学的论文提出了基于行列式点过程的最大间隔话题模型，对相应优化问题进行了有效求解，并且在真实数据上验证了该方法的有效性，实验说明能显著提高话题间的多样性。刘畅同学在完成论文过程中积极主动探索，勇于克服困难，较好地完成了论文的各项工作，是一篇优秀的本科毕业论文。 指导教师签字：李亚 2014 年 6 月 12 日
评阅教师评语	该生通过行列式点过程为一种监督式隐变量话题模型 MedLDA 进行了改进，使得学到的话题多样性较好。在实验用的数据集上证明了该方法的有效性。论文写作严谨规范，符合毕业论文要求。 评阅教师签字：胡国华 2014年 6 月 12 日
答辩小组评语	论文提出了基于行列式点过程的最大间隔话题模型，并提出了有效的求解方法。真实数据上的实验验证了方法的有效性。论文写作规范，结构清晰，达到了符合训练论文的要求，是一篇优秀的毕业论文。 答辩小组组长签字：李建民 2014 年 6 月 12 日

总成绩：93

教学负责人签字：[illegible]

2014 年 6 月 19 日