学校代码	10699
分 类 号	TP391.4
密　　级	
学　　号	2011201822

题目　狄利克雷过程在主题分割和音素分割中的应用

作者　　杨超

学 科 、专 业　　计算机应用技术

指 导 教 师　　谢磊

申请学位日期　　2014 年 2 月

西 北 工 业 大 学

硕 士 学 位 论 文

（学位研究生）

题目：狄利克雷过程在主题分割和音素分割中的应用

作　　者：杨超

学科专业：计算机应用技术

指导教师：谢磊

2014 年　2　月

The Application of Dirichlet Processes in Topic Segmentation and Phoneme Segmentation

A Dissertation Submitted for the Degree of Master

on Computer Application and Technology

by Yang Chao

Under the Supervision of

Xie Lei

School of Computer Science

Northwestern Polytechnical University, Xi'an, China

February, 2014

摘要

在处理序列数据时，经常需要通过建立分割模型来描述由子现象单元构成的动态现象。一般而言，这类任务大致分为两类，第一类主要考虑建模单元内聚性以及相邻单元的差异性，如对一篇文章进行自动分段，常用的方法是在全局上寻找使某种准则最优的分割。另一类任务的单元间则具有较强的复现性，如基因序列分割和说话人检测，前者在观察值上有严格的重复子序列，后者虽然观察值并没有严格的重复，但是却存在某种相似性。对这类任务常用的方法是为子单元建立简单的模型，然后利用马尔科夫过程来描述它们之间的转化。不过这些方法都依赖于预先给定的参数个数—全局最优分割需要给定分割的个数，马尔科夫模型需要给定状态个数。本文通过引入狄利克雷过程先验，建立起非参数贝叶斯方法，使得模型可以随着数据的变化自动调整参数个数。

本文分别选取广播新闻故事分割 (story segmentation) 和类因素分割 (phoneme segmentation) 任务作为这两类分割任务的代表，将的基于狄利克雷过程的非参数模型应用于其中。

故事分割是主题分割的一种特殊形式，需要将一段文本流分割为故事流，本文主要研究新闻广播中的故事分割，即将完整的广播新闻节目分割为多段独立的新闻故事。本文从基于全局最大化似然的无监督贝叶斯方法出发，通过引入依赖于距离的中国餐馆过程 (distance dependent Chinese restaurant process, dd-CRP)，为其建立了非参数模型。和几种常用的方法对比，本文提出的新方法可以自动发现合适的参数个数，并在 F1 值上得到了最优的结果。

声学单元分割也称类音素分割，是语音研究中的一项基础任务。这一任务的目的是将音频流分割为类似音素的最小单元，以用于更高层的任务。本文用不同的高斯分布描述不同的声学单元，用马尔科夫过程描述单元间的转化，并增加一个分层狄利克雷过程作为模型参数的先验，建立起含有无限状态的隐马尔科夫模型，称之为分层狄利克雷过程隐马尔科夫模型 (hierarchical Dirichlet process hidden Markov model, HDP-HMM)。实验结果表明，这一模型具有可观的分割效果，并可以有效的用于对序列中子单元的聚类分析。

主题词： 狄利克雷过程，非参数贝叶斯模型，依赖于距离的中国餐馆过程，隐马尔科夫模型，主题分割，音素分割

Abstract

Segmentation model is used to describe dynamical phenomena formed of sub units when handling sequence data. In general，there are mainly two kinds of these tasks . The first is the task such as paragraph segmentation for documents which focuses on the cohesion of each unit and the diversity between adjacent units. The common approach is to find the best segmentation according some global optimization rules. For another kind task, the dynamical phenomena are formed of a set of reduplicate units. These include modeling the gene sequence which forms of a set of sub-sequences and the speaker diarization task, in which the data switches between different but reduplicate speaker voices. For these cases, Markov process with switches between a set of simpler models, are employed to describe the observed data. But these approaches typically rely on pre-specified number of parameters. The global optimization segmentation needs the number of segmentations and the Markov model needs the number of states. In this thesis a Bayesian nonparametric approach that allows for auto-adaptation of parameter number is developed by introducing a Dirichlet process prior on the model parameters.

This thesis studies two basic and essential tasks as specialization of the two kinds of task mentioned before.

The first is story segmentation, a special form of topic segmentation, which concerning converting a text stream into a story stream. This thesis focuses on the story segmentation task of broadcast news so that the goal is to cut the entire news into multiple independent news stories. This thesis develops a novelty nonparametric approach by plugging a distance dependent Chinese restaurant process into an unsupervised Bayesian segmentation approach.Experiments show that our approach outperforms both supervised and unsupervised baseline approaches and the segmentation number can be automatically learned from data.

The kind of Markov switch model is studied via modelling unsupervised acoustic unit segmentation task(also called phoneme segmentation), which is a basic task in speech research area. This task needs to cut the speech signal stream into segments of acoustic unit which can be viewed as the smallest unit forming human speech. Then the acoustic unit can be used for higher level task. In this thesis, each kind acoustic unit is modeled as a Gaussian model and the switch between these units is modelled as a Markov process. A HDP prior is put on this model to get a hidden Markov model with unbounded state number. The experiment results show the segmentation performance of our model is comparable with other state-of-art methods. And our model also shows a good property of cluster analysis for acoustic unit.

Key Words: Dirichlet process, Bayesian nonparametric, distance dependent Chinese restaurant process, hidden Markov model,story segmentation,acoustic unit segmentation

目录

1 绪论

1.1 研究背景

本论文受到国家自然科学基金面上项目《基于 DBN 协同建模的中文及跨语种语音结构事件检测研究》（编号 61175018）的资助，旨在研究利用非参数贝叶斯模型对广播新闻故事分割任务和音素分割任务进行建模和分析。

1.1.1 非参数贝叶斯模型

目前，概率模型已经成为了机器学习领域中的重要方法。由于许多不同领域的问题在利用概率模型进行抽象描述时，其本质上往往是一致的，只是在特征表示上有一些区别，因此可以容易的将某领域建立的概率模型应用在到其他领域的任务中，并获得可观的效果。所以，对于概率模型的一般性质和算法的研究非常重要。20 世纪以来，概率论、数理统计和最优化理论等领域的发展为此提供了充分的基础，目前许多的概率模型理论都是将这些经典在机器学习这个主题下进行重新的解释。近年来，Koller 和 Jordan 等人更是从更加宏观的角度对利用概率建模进行了研究，发展出的图模型理论，为不同领域的现象建模提供了一套统一的概率建模方法和推断的框架 [1, 2, 3]。

利用概率模型进行推断的结果是一个概率分布，这给模型带来了一定的鲁棒性，也更符合人们解释现象时的行为。然而，由于这个推断的结果是和模型的参数相关的，必须考虑模型的参数学习问题。对于模型的参数，一般有两种处理方法，一是利用最大似然法得到参数的值，然后固定该参数值进行推断。另一种则是为参数增加先验分布，得到参数的一组分布，当进行推断时，对参数所有取值下的推断结果进行加权，这一方法也叫做贝叶斯方法。无论采用哪一种方法，都需要固定参数的个数进行参数学习。而然，对于许多问题，其模型对于参数的个数是敏感的，如果参数的个数设置不当，会使得模型的复杂度和数据不一致，是的模型不能有效的建模数据。比如对于聚类问题，当模型的聚类个数和数据的真实聚类个数不等时，得到的结果就无法令人满意。对于如何选择合适参数个数，通常称为模型选择 [4, 5, 6] 问题。对于规模较小的问题，可以用人工实验来进行，但是也要耗费大量的人力物力，而对于大多数实际任务，比如对于海量互联网数据的分析任务，单次实验即需要调动成百上千台服务器运行数个月才有结果。所以通过人工实验选择参数个数是不可行的, 必须寻找合适的方法来解决模型选择

的问题。

近年来，高斯过程 (Gaussian process,GP)[7]、狄利克雷过程 (Dirichlet process,DP)[8, 9] 等随机过程，由于其具备的特殊性质，可以用于建立能够根据数据分布自适应调整参数个数的模型，这类模型被称为非参数贝叶斯模型 (Bayesian nonparametric models)[10]。其中的狄利克雷过程具有良好的聚类性质，其聚类成分个数可以随着数据的变化而改变，从而被广泛的应用于建模含有混合成分的非参数模型，并在许多实际应用中都得到了可观的结果。本文将对这一过程的相关模型和算法进行详细深入的研究，并应用在广播新闻故事分割和类音素这两个语言和语音处理任务中，取得了良好的效果。

1.1.2 广播新闻故事分割

广播新闻故事分割是指在新闻广播的音视频流上切分为出多个新闻故事的任务 [11, 12]，它是主题分割任务的一种特殊形式。主题分割泛指一系列需要对序列数据按照其主题语义进行切分的任务。对于广播新闻的检索系统，从整段新闻中分割出具有语义相关性质的小段落是非常重要的，目前，面对日益增长的海量数据，如果使用人工的方法来标注工作量巨大，一般是不可能完成的。因此，寻找到合适的自动分割方法十分关键。除了用于新闻检索中，这一任务也可以进一步推广到一般的视频 (此处的视频泛指包含音频的视频) 检索任务中。如今网络视频数据爆炸性增长，单个视频文件可能长达几十分钟或者数小时，此时必须对其进行有效地切分以提供更加细粒度的检索能力。另外，通过对视频进行有效地分割，可以帮助对视频进行进一步的分析和理解，这是其他许多任务不可缺少的环节。

对于这个任务，许多经典的方法都是基于词汇黏合（lexical cohesion）关系，即假设每个故事内的不同句子其用词是近似的，而不同故事之间的用词是有差异性的，这样将文档划分为黏合度较高的片段即可得到故事的边界。其中的典型代表是 Hearst 提出的 TextTiling 方法 [13, 14, 15]，其将句子转化为词频向量表示，然后利用一些相似度测度函数来考察相邻句子间的相似度，由于边界两侧的句子分属不同故事，其用词差距会导致相似度较低，因此，可以用相似度曲线上的谷点来检测边界。不过，通过寻找局部最小值来检测边界的结果只是局部最优的，Malioutov 等人提出的最小割算法 (Mincut)[16]，利用一个动态规划方法来寻找全局最优的故事边界。

另外，近几年基于主题模型的方法被应用在这一任务中，这类方法利用一系列主题模型如潜在语义分析 (latent semantic analysis，LSA)、概率潜在语义分析 (PLSA(Probability Latent Semantic Analysis，PLSA)、隐狄利克雷分配 (latent dirichlet allocation，LDA) 等 [17, 18]，先从含有边界信息的语料中学习出一组主题，然后将待分割语料的词频表示映射的这组主题表示上，从而得到一个更低维

度的稀疏主题表示，进而利用 textiling 或者 Mincut 等方法在主题表示上进行进一步分割 [19, 20, 21, 22]。Zheng 和 Lu 等人从利用流形学习的方法，可以从语料中学习到更加适合分割的表示 [23, 24]，这些方法获得了相比于直接利用词频表示的方法获得了更好的分割结果，但是这类方法需要标注好的语料作为获取主题模型的训练集，是一种有监督的方法。

1.1.3 类音素分割

类音素分割任务是指将一段音频流分割为类音素单元流，这里的类音素是指一些类似于语音学中音素 (phoneme)，用于构成语音的最小单元，所以也称声学单元分割 (acoustic unit segmentation) 或者音素分割 (phoneme segmentation)。对于许多传统的语音处理任务，如语音识别与合成，都是基于已经标注好的语料进行的，因此并不需要进行类音素分割。然而目前互联网数据库拥有海量的未标注语音资源，如果利用诸如深度神经网络的模型，可以通过这些数据自动学习到效果好的模型 [25]，但是，对这些语音的标注需要大量的人力，人工进行标注是不现实的。因此，利用无监督的方法处理海量语音数据的研究非常必要。另外，许多语言，如一些非洲部落的语言，甚至未被语言学家研究过，如果可以找到一种无监督的框架，能够方便的推广到对各种语言的建模，就可以极大地加快对这些语言的研究。目前这一类问题被称为低资源语音研究 (low resource)[26, 27]，意指利用最低限度的资源对语音进行无监督的研究，是目前语音领域研究的热点之一。而对低资源语音研究中，从特征层面往上，第一个任务的便是子词级相关研究，即从原始的声音特征流中发现不同的音素，作为语音的最小组成单元，因此，音素分割具有重要的研究意义。

在音素分割的研究中，常用声学变化的峰值作为可能的边界，许多方法利用不同的声学变化准则进行分割 [28, 29, 30]。Qiao 引入了失真率的概念并以此为基础定义了代价函数，通过层次聚类算法不断地将相邻的语音帧合称为一个片段，得到最终的分割 [31]。Scharenborg 分析了这一自下而上的方法的局限性，利用自上而下的信息进行富足分割，来弥补单一使用声学变化为线索的不足 [32]。本文以文献 [33, 34] 为参考，从概率的角度进行建模，将音素的分割和聚类统一起来，并针对音素个数未知的问题，建立非参数模型进行研究。

1.2 本文主要工作和创新点

本文旨在利用概率模型建模序列数据解决分割问题。为了解决传统概率模型面临的模型选择问题，本文引入狄利克雷过程先验，针对故事分割和类音素分割两类序列分割问题建立了不同的非参数贝叶斯模型，并进行了实验验证与分析。

本文的主要工作概括如下:

1. 关于概率图模型理论的研究 概率图模型作为描述现象的一种方法，广泛的用于各个领域。本文研究了基本的概率图建模方法，包括常见的概率分布，概率图模型的表示方法，以及相关的推断算法。深入探讨了如何用概率图模型描述的数据不同性质，以及如何从简单的模型扩展到复杂的图模型。

2. 关于非参数贝叶斯模型理论的研究 传统的参数模型面临着模型选择问题，如果参数选择不当，会出现过拟合或者欠拟合的问题。非参数模型则通过数据来调整参数过程以解决模型选择的问题。本文对一类常用的非参数模型 -狄利克雷过程进行了深入的研究。研究包括基本的狄利克雷过程，狄利克雷过程混合以及分层狄利克雷过程模型。并深入讨论了这类模型的基于 stick-breaking 构造和中国餐馆过程 (Chinese restaurant process，CRP) 的构造以及其在不同构造下的推断算法。对于中国餐馆过程，本文还研究了一种其对应的更一般的形式，称为依赖于距离的的中国餐馆过程。另外，本文还研究了将分层狄利克雷过程用于构造无线状态隐马尔科夫序列模型。

3. 故事分割任务的非参数方法建模 传统的基于 Mincut 的故事分割方法使用余弦相似度和交叉熵作为词汇黏合度测量，然后在根据经验设计出函数来度量故事内聚性，对于不同问题也需要设计不同的函数。本文从概率的观点出发，利用贝叶斯模型对广播新闻故事分割任务进行建模，得到一种基于联合似然的内聚性度量。另外，基于 MinCut 框架的算法需要给定切分个数，而新闻的故事个数实际上是未知，这是 MinCut 相关算法的一个弊端。本文提出一种新的方法，通过增加一个与依赖于距离的的中国餐馆过程 (distance dependent CRP,dd-CRP)，为广播新闻故事分割任务建立了一个非参数贝叶斯模型，从而解除了这一限制，并通过食盐验证了这种方法的优势。

4. 音素分割任务的非参数方法建模 类音素分割和故事分割相比，除了在任务领域和数据特征上的不同，其最主要的区别在于，类音素分割任务中许多类音素单元是重复出现的，而故事分割却不具备这一性质。因此，本文为类音素分割任务建立了不同的分割模型。对于类音素分割的建模，要考虑语音数据的时序性，即同一个语音单元对应的连续多帧是相似的，还需要考虑到音素之间的转换的统计性。本文用隐马尔科夫模型建模帧间的时序关系，用自跳转描述音素的持续性，用互跳转描述音素之间的转换。然而，隐马尔科模型需要给定模型中的状态个数，由于实际类音素单元的个数未知，本文将狄利克雷过程作为隐马尔科夫模型的先验，通过建立非参数模型，自适应的发现合适的状态个数。另外，本文还引入一个粘滞参数来建模音素的持续性，使得模型具有一定的自跳转偏执。这种方法称

为基于 sticky-HDPHMM 模型的方法。另外，类音素分割任务往往是作为类音素发现任务的一个子任务，即需要对分割得到的结果再进行聚类。本文的方法将这两个过程统一起来，其分割和聚类的过程被蕴含在模型的迭代推断过程之中。实验结果验证了本文方法的有效性。

本文主要的创新点如下:

1. 本文通过建立概率生成模型来描述新闻广播故事的生成过程，得到的内聚性度量拥有很好的物理解释，并且具有更好的切分效果。进一步，针对广播新闻故事分割的故事个数未知的问题，本文提出了一种基于依赖于距离的中国餐馆过程的方法，有效的解决了这一问题。

2. 通过将分层狄利克雷过程 (XXX,HDP) 做为隐马尔科夫模型的先验，为音素分割任务建立了一个含有无限状态的隐马尔科夫模型，称之为 sticky-HDPHMM 模型，解决了音素个数未知的问题，并取得了良好的音素分割效果。

1.3 本文的组织结构

第一章 综述了论文的研究背景和内容，并给出了本文的主要工作，创新点以及组织结构。

第二章 回顾了概率模型的基础知识，包括基本的概率分布，有向图和无向图模型的表示，置信传播算法。基于 Gibbs 采样推断。并研究了常见的概率模型之间的关系。

第三章 研究了狄利克雷过程的相关模型，对基于不同构造的推断和参数学习方法进行了研究。

第四章 首先回顾了故事分割任务的非概率模型方法，然后提出了利用概率模型建模的新模型，并给出了实验的结果和分析。

第五章 描述了类音素分割任务，以及相关的无限状态隐马尔科夫模型，包括对模型进行了自跳转偏置变量以后的改进算法，给出了实验结果和分析。

第六章 总结了全文并给出了一些下一步需要研究和解决的问题.

2 概率模型基础知识

本文主要使用概率模型对分割问题进行建模和研究 [35]，所以本章给出文中需要用到的基本概率知识，包括指数函数族的性质，图模型的表示与消息传播算法，gibbs 采样算法以及一些常用的概率模型。

2.1 指数函数族

2.1.1 指数函数族分布

在使用概率方法对问题进行建模时，经常使用的一类称为指数族的分布：

$$p(x|\eta) = h(x)g(\eta)\exp\{\eta^T u(x)\} \tag{2.1}$$

其中 η 是分布的参数，也称为自然参数，$\boldsymbol{u}(x)$ 是关于 x 的函数，$g(\eta)$ 是归一化系数，该函数需要满足概率的归一化条件：

$$g(\eta)\int h(x)\exp\{\eta^T u(x)\}dx = 1 \tag{2.2}$$

许多常用的分布如伯努利分布，高斯分布，泊松分布等都是属于指数族的分布。例如伯努利分布：

$$p(x|\mu) = \mu^x(1-\mu)^{1-x} \tag{2.3}$$

可以将其变换为指数族分布的一般形式:

$$\begin{aligned} p(x|\mu) &= \exp\{x\ln\mu + (1-x)\ln(1-\mu)\} \\ &= (1-\mu)\exp\left\{\ln(\frac{\mu}{1-\mu})x\right\} \end{aligned} \tag{2.4}$$

和指数族分布的一般形式（式(2.1)）比较，有:

$$\begin{aligned} \eta &= \ln(\frac{\mu}{1-\mu}) \\ u(x) &= x \\ h(x) &= 1 \\ g(\eta) &= \sigma(-\eta) \end{aligned} \tag{2.5}$$

同样，对于高斯分布:

$$p(x|\mu,\sigma^2)=\frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}\exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\} \tag{2.6}$$

其指数族一般形式为:

$$\begin{aligned}\boldsymbol{\eta}&=\begin{pmatrix}\mu/\sigma^2\\-1/2\sigma^2\end{pmatrix}\\ \boldsymbol{u}(x)&=\begin{pmatrix}x\\x^2\end{pmatrix}\\ h(x)&=(2\pi)^{-1/2}\\ g(\eta)&=(-2\eta_2)^{1/2}\exp\left(\frac{{\eta_1}^2}{4\eta_2}\right)\end{aligned} \tag{2.7}$$

2.1.2 充分统计量

假设有一组独立同分布于某一指数分布的样本 $\boldsymbol{X}=\{x_1,...,x_N\}$，其联合似然函数为:

$$p(\boldsymbol{X}|\eta)=(\prod_{n=1}^{N}h(x_n))g(\eta)^N\exp\{\eta^T\sum_{n=1}^{N}u(x_n)\} \tag{2.8}$$

如果用最大似然方法进行参数估计，取 $\eta=\eta_{ML}$ 使得上式的梯度为零，有:

$$-\nabla\ln g(\eta_{ML})=\frac{1}{N}\sum_{n=1}^{N}u(x_n) \tag{2.9}$$

可以看出，参数的最大似然估计解只和 $\sum_n u(x_n)$ 有关，所以 $\sum_n u(x_n)$ 又称为分布的充分统计量。这样，当需要进行参数估计时，不必保存所有的数据，而只需要记录充分统计量即可。比如对于伯努利分布，只要记录 x_n，对于高斯分布，只要记录 x_n 和 x_n^2。‘

2.1.3 共轭先验

贝叶斯方法将参数看作是一个随机变量而不是一个值，经常需要估计它在观察值上的后验分布 $p(\eta|X)$，或者希望进行贝叶斯预测:

$$p(x_{new}|X)=\int p(x_{new}|\eta)p(\eta|X)d\eta \tag{2.10}$$

对于这些任务，可以利用贝叶斯条件概率公式得到其后验概率：

$$p(\eta|X) = \frac{p(\eta)p(X|\eta)}{P(X)} \tag{2.11}$$

然而为了使用该公式，必须为参数增加一个先验分布。这里引入共轭分布的概念，如果对于 $p(X|\eta)$，关于 η 的某一分布 $p(\eta|\lambda)$ 和其对应的后验分布 $p(\eta|X)$ 是同一族的分布，则称 $p(\eta|\lambda)$ 是 $p(X|\eta)$ 的共轭分布。共轭分布是贝叶斯方法中非常重要的知识，后面会看到，利用共轭分布，可以使得模型的许多推断任务可以解析的进行。

注意，考虑下面的分解：

$$p(\eta|X, x_{n+1}) \propto p(x_{n+1}|\eta)p(\eta|X) \tag{2.12}$$

其中 $p(\eta|X)$ 可看做 η 的先验，$p(\eta|X, x_{n+1})$ 可看做 η 的后验，$p(x_{n+1}|\eta$ 可看做似然，若先验不是关于似然共轭的，则 $p(\eta|X, x_{n+1})$ 和 $p(\eta|X)$ 具有不同的形式，即后验分布的函数族随着数据的变化而变化，那么对于这样的分布难以进行研究。另外，上面这种分解也展现了贝叶斯方法具有天然的在线 (on line) 性质。

考虑下面的一种先验分布：

$$p(\eta|\chi, \upsilon) = f(\chi, \upsilon)g(\eta)^{\upsilon} \exp\left\{\upsilon\eta^T\chi\right\} \tag{2.13}$$

与式(2.8)结合，可得到 η 的后验分布：

$$p(\eta|X, \chi, \upsilon) \propto g(\eta)^{\upsilon+N} \exp\left\{\eta^T\left(\sum_{n=1}^{N} u(x_n) + \upsilon\chi\right)\right\} \tag{2.14}$$

可以看出，这个后验分布和先验分布具有相同的函数形式，从而说明指数族存在一个通用的共轭先验。这个先验具有一定的物理意义，可认为其假设存在 υ 个伪观察样本，且每个观察的统计量是 χ。

多项分布和高斯分布是两种最常见的概率分布，分别用于处理离散和连续数据，关于他们的共轭性质分析，可以参考 [36, 37]。

2.2 图模型

概率图模型提供了一个一致的框架来表示一组随机变量之间的条件依赖关系。对于概率图模型的综述以及相关的推断算法，请参考文献 [1, 2]。目前这一框架已经发展出了一些通用的有效的推断算法，如置信传播 (Belief Propagation，

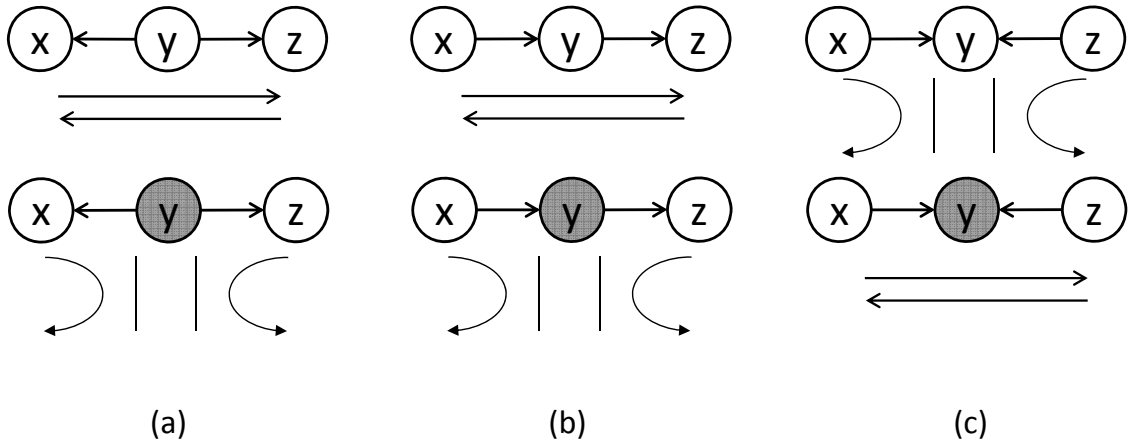

图 2.1: 不同情况下的独立性质示意图

BP) 和变分推断 (Variarional Inference)。许多经典的模型如隐马尔科夫模型 (hidden Markov models,HMM)，状态空间模型 (state space models) 都可以归到图模型的框架下，相应的推断算法如前向后向算法和卡尔曼滤波等都可以看做是置信传播算法的特殊形式。

一般而言，用集合 $G = (V, E)$ 表示一个图模型，其包含了用来表示随机变量的节点集合 V，以及一组用来表示随机变量见的概率关系的边集合 E。E 中的元素 (i, j) 表示节点 i,$j \in V$ 之间的依赖关系。在可视化的图形表示里，通常每个节点用圆圈表示，无色的圆圈代表未观察到的变量，有色的圆圈代表观察到的变量。有向图的边用箭头表示，即对于 $(i, j) \in E$，在图中对应一个从 i 指向 j 的箭头，其中 i 称为父节点，j 称为子节点。无向图的边则用无箭头的连接线段表示。关于有向图和无向图，下面会详细讨论。另外，模型中的参数用圆角矩形表示。

2.2.1 有向图模型

在有向图中，对于节点 i，可以有零或多个父节点，也可以有零个或多个子节点，本文用集合 $\Gamma(j) = \{i \in V|(i, j) \in E\}$ 表示节点 j 的父节点集合。对于没有子节点的节点，称其为根节点。对于没有父节点的节点，称其为叶节点。:

在概率图表示下，对于一组随机变量，其联合概率可以分解为一系列条件概率的乘积:

$$p(x_V) = \prod_{i \in V} p(x_i|\boldsymbol{x}_{\Gamma(i)}) \tag{2.15}$$

其中 $\boldsymbol{x_A}$ 表示集合 $x_i|i \in \boldsymbol{A}$. 很容易证明，对于一个无环图，上式是一个满足概率条件的密度函数。

独立性质 通过有向图模型可以很容易写出联合分布，但是由其定义的节点之间的独立关系却不显然的。这里仅讨论三种简单的拓扑，其他任何复杂的拓扑

形式的独立性质都可以从这几种基础形式根据一定的规则得到。

考虑图2.1(a) 中尾对尾情况，在未观察到 y 时，x 和 z 不是独立的：

$$p(x,z)=\int p(x,y,z)dy=\int p(x)p(y,z|x)dy=p(x)p(z|x)\neq p(x)p(z) \tag{2.16}$$

如图2.1(a) 上图所示，可以认为未观察到 y 时，没有阻止 x 到 z 的联系，所以 x 和 z 是不独立的。

而在观察到 y 时，x 和 z 则是条件独立的：

$$\begin{aligned} p(x,z|y)p(y)&=p(x)p(y|x)p(z|y)\\ p(x,z|y)&=p(x|y)p(z|y)\end{aligned} \tag{2.17}$$

如图2.1(a) 下图，可以认为此时由于 y 被观察到，阻止了 x 到 z 的联系。

对于另外两种情况也是类似的。考虑图2.1(b) 头对尾的情况，易得：

$$\begin{aligned} p(x,z)&\neq p(x)p(z)\\ p(x,z|y)&=p(x|y)p(z|y)\end{aligned} \tag{2.18}$$

考虑图2.1(c) 头对头的情况，易得：

$$\begin{aligned} p(x,z)&=p(x)p(z)\\ p(x,z|y)&\neq p(x|y)p(z|y)\end{aligned} \tag{2.19}$$

注意，头对头时的情况和上面两种不一样，在未观察到 y 的时候，x 和 z 是独立的，然而在观察到 y 的时候，x 和 z 却不是条件独立的，这一现象称为 explaining away。更加详细的讨论可参考 [35]。

另外，对于任意两个节点集，对于他们的（条件）独立性质有一个称为 d-划分的定理，利用这个定理，对于任何拓扑，都能够得到其节点间的独立关系。

D-分割 对于 A、B 和 C 三个互不相交的节点集，考察 A 中任意节点到 B 中任意节点之间的路径。如果某一条路径满足下面的情况之一，则称该路径是被阻塞的：

- 路径中存在尾对尾的节点或者头对尾的节点，且这些节点在 C 集合中。
- 路径中存在的所有头对头节点以及它们的子孙节点都不在在 C 集合中。

如果 A 到 B 的所有路径都是阻塞的，则称 A 集合和 B 集合是被 C 集合 d-分割的。

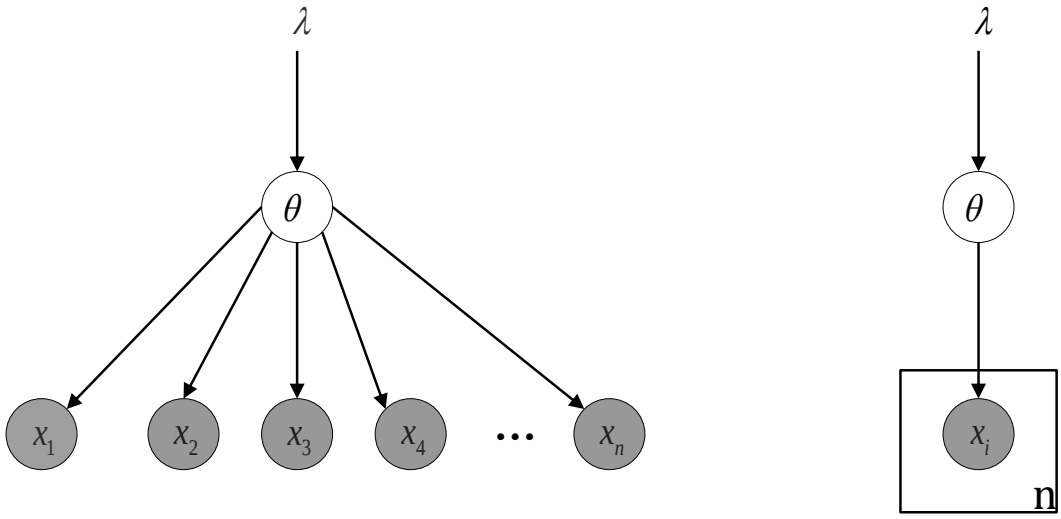

图 2.2: 左图：多个独立同分布的节点。右图：使用堆叠记法进行简化

马尔科夫 blanket 对于有向图模型中某一节点，由该节点的父节点、子节点以及同子父节点（和该节点拥有相同子节点的节点）构成的集合称为的该节点马尔科夫 blanket。根据上面介绍的定理，可以很容易推导出，当观察到某节点的马尔科夫 blanket 时，该节点与其他所有节点是条件独立的。

堆叠记法 对于多个独立同分布变量，如图左所示，如果一一表示出来，会非常繁琐，所以一般将其简化为用单个节点和带有个数标记的方框来表示, 如图右所示。

2.2.2 无向图模型

本文主要使用有向图模型来进行建模，不过有向图的一些推断算法需要先将其转换为无向图的形式，这一转换非常简单，只要将边上的箭头去除，并将所有的同子父节点相连即可。

在有向图中，随机变量的联合分布可以依照节点的拓扑进行因子化。无向图模型也表征了一些条件独立的性质，可以用来进行因子化。不过，其表征条件独立的方式和有向图有所不同。对于无向图，假设 V_i、V_j 和 V_k 是三个互不相交的节点集，如果从 V_i 中的某一节点到 V_k 中的某一节点必然要经过 V_j，则称 V_j 是一个分割集。对于无向图模型，其中任意两个节点集在其分割集上是条件独立的，即:

$$p(\boldsymbol{x}_{V_i}, \boldsymbol{x}_{V_k}|\boldsymbol{x}_{V_j}) = p(\boldsymbol{x}_{V_i}|\boldsymbol{x}_{V_j})p(\boldsymbol{x}_{V_k}|\boldsymbol{x}_{V_j}) \tag{2.20}$$

这个性质称为无向图模型的全局马尔科夫性，因此无向图模型又称马尔科夫随机

场。

注意，当一个有向图转化为无向图表示后，此时的无向图形式仍能表征出的独立性是原来的有向图中也有的。但是反之则不然，一些有向图中的独立性，在无向图里无法表现。这说明无向图和有向图并不等价。但是，这不意味着无向图的表示空间就是有向图的表示空间的子集，因为也存在一些无向图是无法用有向图表示的。考虑一个最简单的三个节点全连接的无向图，根据之前的讨论易知，只有 9 种有向图对应的无向图是这种形式，而这 9 种有向图必然存在某两个节点在观察到或者未观察到另一个节点时是条件独立的。而这个无向图却可以表示出另一种模型: 这三个节点两两之间，无论在什么情况都是不独立的。故无向图的表示空间不是有向图的表示空间的子集。

对于无向图模型，其联合概率的分解形式并不像有向图那么直接。令 C 表示一个无向图模型 G 中的所有全联通子图组成的集合，这里定义一组全联通子图上的函数 $\psi_c(.)$，如果关于随机变量 V 的联合分布可以分解为如下形式:

$$p(\boldsymbol{x}_V) \propto \prod_{c \in C} \psi_c(\boldsymbol{x}_c) \tag{2.21}$$

则这组随机变量的联合分布对应的无向图表示 G 是具有全局马尔科夫性质的。反之，如果对于所有的 x 满足其概率 p(x) 是严格正的，则其联合分布对应的马尔科夫随机场一定可以按上式分解。这个理论称为 Hammersley-Clifford 定理。具体的证明以及进一步讨论可参考 [38, 39]。

pairwise 马尔科夫随机场 考虑一种不存在环的无向图，或称为树状的无向图模型。如果一个有向图其对应的无向图是树状的，则其有向图形式和无向图形式表征的独立性是一致。对于树状的模型，任何一个单一的节点 v 都是一个分割集，它将随机变量分为两个集合，这两个集合中的节点在观察到 v 时条件独立的。这种独立性对于其对应的有向图也是一样的情况。

对于 pairwise 马尔科夫随机场，其所有的全联通子图集合即由其所有的节点和所有的边组成的集合。故其分解形式为:

$$p(\boldsymbol{x}_V) \propto \prod_{(i,j) \in E} \psi_{ij}(x_i, x_j) \prod_{i \in V} \psi_i(x_i) \tag{2.22}$$

许多重要的序列模型，如隐马尔科夫模型、状态空间模型等，其对应的无向图模型都是 pairwise 马尔科夫随机场。

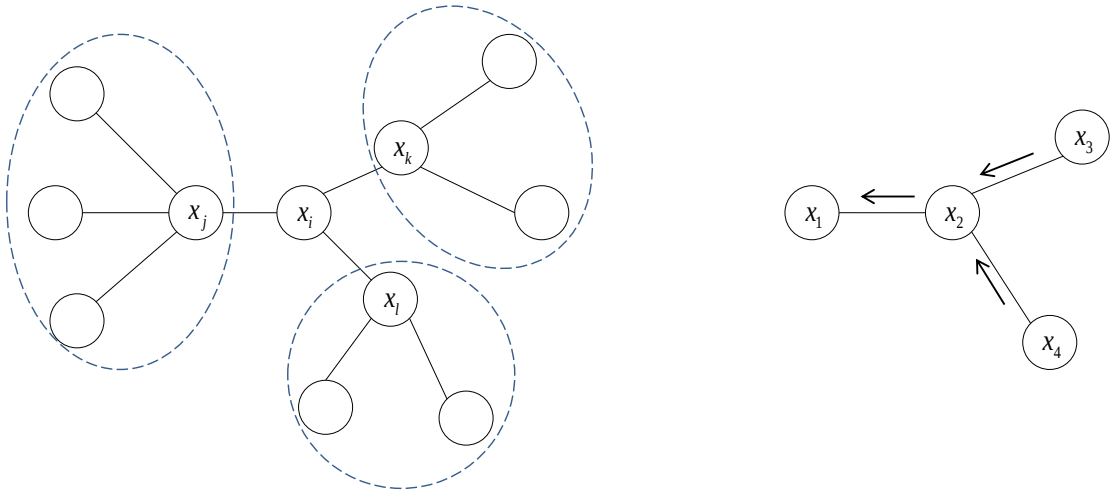

图 2.3: 左图：右图：简单的消息传播示意图

2.2.3 置信传播算法

在大部分应用中，随机变量的联合空间都会很大，以至于推断任务变得很难。比如，考虑一个含有 N 个节点的图模型，每个节点的取值空间是一个含有 K 个值的离散空间，则联合状态空间的大小为 K^N, 如果需要计算某个变量在观察值下的条件边缘概率，对于含有 K 个值的图，则需要求 K_{N-1} 项的和。

对于树形的图模型，可以通过递归的进行局部计算，从而有效的降低算法的复杂度。由于树状模型中任何一个节点都是分割集，他将原模型划分为几个树形的子图模型。下面介绍一种叫做置信传播的算法 [40]，可以通过将子图计算的结果联合起来得到全图的计算结果。

考虑图2.3右的情况，关于 $\boldsymbol{x}$ 的联合概率分布可以分解成:

$$p(\boldsymbol{x}) \propto \psi_{12}(x_1, x_2)\psi_{23}(x_2, x_3)\psi_{24}(x_2, x_4)\psi_1(x_1)\psi_2(x_2)\psi_3(x_3)\psi_4(x_4) \tag{2.23}$$

求边缘概率分布 $p(x_1)$ 时，可将不同的积分项分离:

$$p(x_1) \propto \psi_1(x_1) \int_{x_2} \psi_{12}(x_1, x_2)\psi_2(x_2) \underbrace{\left[\int_{x_3} \psi_{23}(x_2, x_3)\psi_3(x_3)dx_3\right]}_{m_{32}(x_2)} \underbrace{\left[\int_{x_4} \psi_{24}(x_2, x_4)\psi_4(x_4)dx_4\right]}_{m_{42}(x_2)} dx_2 \tag{2.24}$$

这里，定义一个称为消息的变量 $m_{ji}(x_i)$ 来表示上式中对 x_j 进行积分且结果是关于 x_i 的函数的项, 这样上式就变为了:

$$m_{21}(x_1) \propto \int_{x_2} \psi_{12}(x_1, x_2)\psi_2(x_2)m_{32}(x_2)m_{42}(x_2)dx_2 \tag{2.25}$$

如果考虑更一般的情况，假设图中存在一组观察节点 $\boldsymbol{y}$，不失一般性，假设 y_i 和 x_i 相连，则可以将式(2.25)推广到一般的情况:

$$m_{ji}(x_i) = \int_{x_j} \left(\psi_{ij}(x_i, x_j)\psi_i(x_i, y_i) \prod_{k \in \Gamma(j)i} m_{kj}(x_j) \right) dx_j \tag{2.26}$$

可以将这个公式看做是一个消息在图模型的节点间传递的过程。对于节点 i，收集其所有子节点传来的消息，然后将其加工再传给其父节点。

另外，叶子节点上的初始消息为:

$$m_{.i}(x_i) = 1 \tag{2.27}$$

利用上面的公式，从叶子节点递推的计算出所有节点之间传播的消息后，就可以得到任意需要的边缘概率分布:

$$p(x_i|\boldsymbol{y}) = \frac{1}{Z}\psi_i(x_i, y_i) \prod_{j \in \Gamma(i)} m_{ji} \tag{2.28}$$

其中:

$$Z = \int_{x_i} \psi_i(x_i, y_i) \prod_{j \in \Gamma(i)} m_{ji} dx_i \tag{2.29}$$

一些常用的序列模型中的经典算法如隐马尔科夫模型中的前向后向算法 [41]，状态空间模型中的卡尔曼滤波算法等都是置信传播算法的特殊形式。对于非树形的图模型，上面的算法不再成立，需要使用 loopy 置信传播算法，其体细节可参考 [42]。

2.2.4 有限混合模型

本文在2.1节介绍了指数函数族，用于表征观察数据的分布情况。然而有时候数据并不是简单的单一分布，而是服从多重模态分布。此时，经常用有限混合分布来建模数据。一个含有 K 个混合成分的模型的如下:

$$p(x|\pi, \theta_1, ..., \theta_K) = \sum_{k=1}^{K} \pi_k f(x|\theta_k) \tag{2.30}$$

其中每个混合成分对应着一个概率密度函数 $f(x|\theta_k)$，这里统一记为 $F(\theta_k)$。

对于混合模型，每个观察 x_i 的生成过程是：首先根据一个参数为 π 的 K 维的

多项分布选择一个成分 k，然后从第 k 个成分对应的概率密度函数中生成 x_i:

$$\begin{aligned} z_i &\sim \pi \\ x_i &\sim F(\theta_{z_i}) \end{aligned} \tag{2.31}$$

z_i 是用来表示 x_i 对应隐成分的指示变量，即 $z_i = k$ 表示 x_i 是从第 k 个成分的概率密度中采样的。

在大部分应用中，$f(x|\theta_k)$ 都是指数函数族，比如当其是高斯分布时，对应的混合模型就是常用的混合高斯模型 (GMM)。

对于这个模型，可以用最大化观察值联合似然的方法来估计模型的参数值，也可以用贝叶斯方法计算参数的后验概率分布。不过当使用贝叶斯方法时，需要为参数加上一个共轭先验分布:

$$\begin{aligned} &\theta_k \sim G_0(\lambda), \quad k = 1, ..., K \\ &\pi \sim Dir(\frac{\alpha}{K}, ..., \frac{\alpha}{K}) \end{aligned} \tag{2.32}$$

这里还可以用另一种方法来表示上述的贝叶斯混合模型[1]。考虑一个在聚类参数空间 Θ(即 θ 所在的空间，而不是观察值的空间) 上的离散分布 G:

$$\begin{aligned} &\theta_k \sim G_0(\lambda), k = 1, ..., K \\ &\pi \sim Dir(\frac{\alpha}{K}, ..., \frac{\alpha}{K}) \\ &G(\theta) = \sum_{k=1}^{K} \pi_k \delta_{\theta_k} \end{aligned} \tag{2.33}$$

其中 δ_{θ_k} 表示一个关于 θ 的函数, 其在 θ_k 处取 1 而在其他点处为 0 的函数。

然后，按照如下过程生成 x_i:

$$\begin{aligned} \phi_i &\sim G \\ x_i &\sim F(\phi_i) \end{aligned} \tag{2.34}$$

这个生成过程和上面使用只是变量 z 的生成过程是一致的。图2.4给出了这两种不同的表示法对应的图模型。

需要注意的是，有限混合模型中的混合数是确定的。然而，在实际问题中，往往不知道成分的个数，从而需要用到一些模型选择的方法来选择合适的 K。后文将讨论一种称为 Dirichlet 过程混合的模型，可以有效地解决这个问题。

[1]理解这种表示方法对于理解 Dirichlet 过程非常重要。

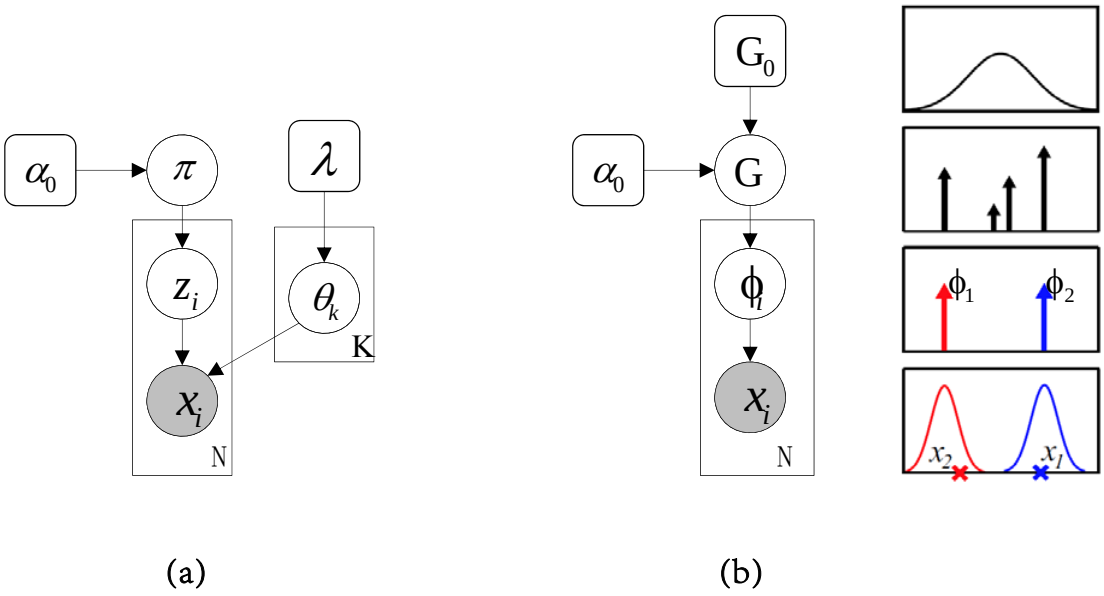

图 2.4: 有限混合模型的图模型表示

2.2.5 隐马尔科夫模型

混合模型假设样本是可交换顺序的，但是在一些问题中，样本却是有顺序相关性的。比如在对语音序列建模中，由于语音信号是连续变化的，相邻时间片的样本是相关的，故不能直接使用混合模型进行建模。这里考虑对混合模型进行时间上的扩展。假设对于 t 时刻的样本，其成分变量 z_t 并不是服从从参数为 π 的同一个离散分布，而是服从一个和前 n 个时刻相关的离散分布，即:

$$z_t \sim \pi_{z_{t-n},...z_{t-1}} \tag{2.35}$$

注意，在这个模型里，π 的个数是 K^n 个，其中 $z_t \in \{1, ..., K\}$.

这一模型称为 n 阶隐马尔科夫模型 (hidden Markov models,HMM)。其中经常使用的是 1 阶 HMM，即 z_t 服从一个和 z_{t-1} 相关的离散分布。隐马尔科夫模型在对各种序列建模的任务中应用广泛，如生物信息挖掘中对基因序列的建模任务，自然语言处理中的分词、标注、实体名提取等任务，语音中对音素的建模任务。但是，隐马尔科夫模型中的 z_t 服从一个 K 维的离散分布，这使得在实际应用中，需要用模型选择的方法来确定一个最优的 K，后文将讨论对其的改进方法。

2.2.6 隐狄利克雷分配

上面考虑的都是单组数据的问题，然而，许多任务需要处理多组数据，这些数据的生成过程往往是相关的，但是并不完全相同。如果单独为每组数据建模，就无法利用到他们之间的关联性，而如果将其看做是单一的一组可交换的数据，

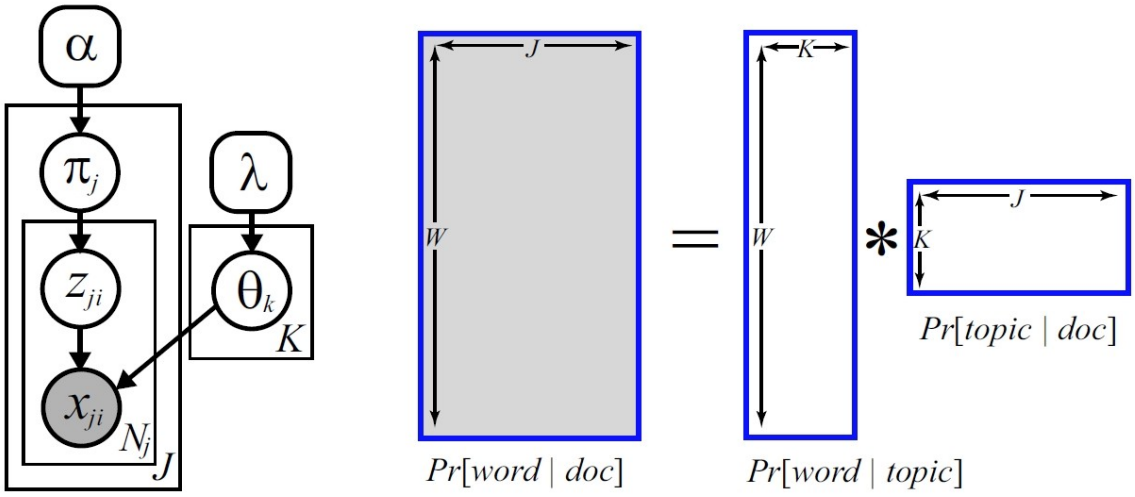

图 2.5: LDA 的图模型表示

又会丢失掉他们天生的判别信息。比如在文本语料中，有多个文档，如果为每个文档建立单独的混合模型，则没有建模文档之间的相关性，如果把所有文档看做是一个可交换性的文档建立一个混合模型，则丢失了文档之间的判别信息。所以可以考虑建立一种层次模型，通过在不同的组之间共享参数来建模相关性，而用另一些参数来控制这组共享参数的权重从而表示他们的区分性。

隐狄利克雷分配 (Latent Dirichlet allocation)[18] 就是利用这种思路，通过对混合模型进行层次扩展得到的一种用来建模多组相关数据的层次贝叶斯模型。考虑一个数据集，含有 J 组数据 $\boldsymbol{x} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_J)$，其中第 j 组数据集中有 N_j 个样本 $\boldsymbol{x}_j = (x_{j1}, ..., x_{jN_j})$。整个数据集表示为

在利用 LDA 建模时，假设每组数据内的样本是可交换的，并且独立同分布于一个含有 K 个成分的混合模型，而各组的 K 个成分是共享的。用 $\{\theta_k\}_{k=1}^K$ 表示混合成分的参数，用 $\boldsymbol{\pi}_j$ 表示第 j 组数据的混合权重，则对于第 j 组数据有:

$$p(x_{ji}|\boldsymbol{\pi}_j, \theta_1, ..., \theta_K) = \sum_{k=1}^{K} \pi_{jk} f(x_{ji}|\theta_k), \pi_j \in \Pi_{K-1} \tag{2.36}$$

可以看出，LDA 使用组间共享参数 θ 用来建模样本间的相关性，使用组内自有的 $\boldsymbol{\pi}_j$ 用来建模各组的差异。和贝叶斯混合模型一样，需要为参数 $\boldsymbol{\pi}$ 加上一个先验[2]。LDA 假设不同组之间是具有交换性的，根据 De Finetti 定理 [43]，这组权重参数可以看做是从同一个先验分布中独立采样的，所以这里加入一个离散分布的共轭先验，即 Dirichlet 分布:

$$\boldsymbol{\pi}_j \sim Dir(\alpha) j = 1, ...J \tag{2.37}$$

另外，也需要为成分参数 θ 加上一个先验 $G_0(\lambda)$，此时得到的 LDA 模型对应的图

[2]如果模型不加这个先验，就是概率隐语义分析模型 (PLSA)

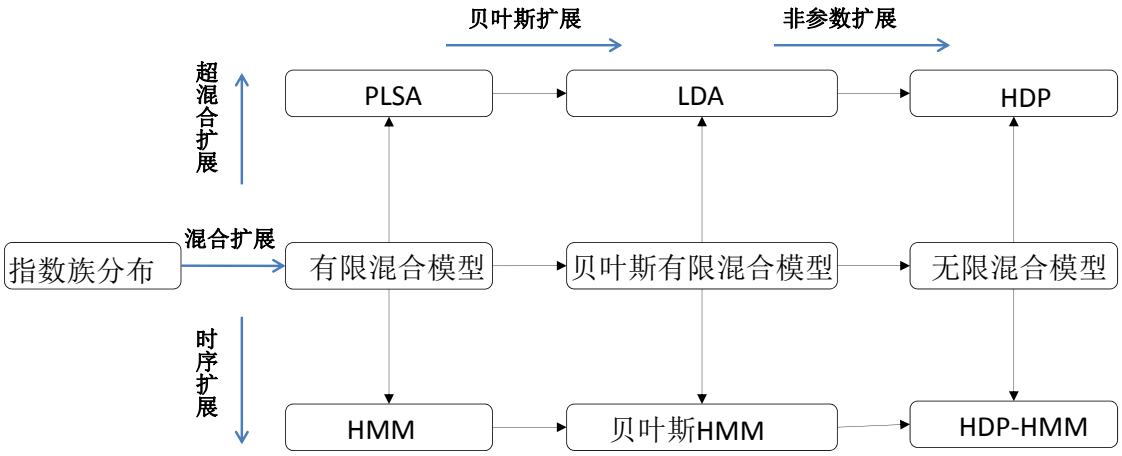

图 2.6: 常用的概率模型之间关系

模型表示见图2.5.

如果用混合模型的另一种表示法（式(2.33)和式(2.34)）来表示每组数据的混合模型，则可以的到 LDA 的另一种等价生成过程:

$$\begin{aligned} G_j(\theta) &= \sum_{k=1}^{K} \pi_{jk} \delta_{\theta_k} \\ \pi_j &\sim Dir(\frac{\alpha}{K}, ..., \frac{\alpha}{K}) \\ \theta_k &\sim G_0(\lambda), k = 1, ..., K \\ \phi_{ji} &\sim G \\ x_{ji} &\sim F(\phi_{ji}) \end{aligned} \tag{2.38}$$

LDA 最早是被用于文本分析，它可以从文档中无监督的学习到多个主题 (即单词的分布)，以及得到文档在主题维度上的表示。对于离散数据，LDA 可以得到一个低秩的矩阵分解近似解。不过，只要选择合适的 $F(\theta)$，LDA 也可以用来建模连续数据。其相关的推断算法可以参考文献 [18, 36]。

和有限混合模型和隐马尔科夫模型类似，LDA 同样需要手动设定混合成分个数，后文中将介绍一种 LDA 的非参数模型扩展，可以自动的从数据中推断出混合成分的个数。

2.2.7 概率模型间关系

本章介绍的几种概率模型，通过对基本的指数族分布进行混合、超混合和时序扩展，得到了一系列具有不同性质的模型。其中，混合模型用于建模多成分混合的数据，超混合模型用于建模共享多个成分的多组数据，时序模型用来建模在时间上相关的数据。进一步，通过贝叶斯扩展，可以得到这些模型对应的贝叶斯

模型。这些模型的参数个数是固定的，统称为参数模型。对于某个模型，其概率图拓扑和概率函数已经确定，此时其模型的复杂度由参数的个数决定。然而，如果复杂度与问题不一致，会导致过拟合或者欠拟合的问题，这就需要通过手工调整参数个数，来选择复杂度合适的模型，这一过程也称为模型选择。在下一章将研究的 Dirichlet 过程，可以用来对这些模型进行进一步扩展，使得其变为非参数贝叶斯模型。这类非参数模型的参数个数会根据数据的分布而自适应的变化，从而可以为研究模型选择的问题提供了新的方法。图2.6给出了这些模型之间的关系。

2.3 Gibbs 采样

对于一个概率分布 $p(x)$，$x \in \chi$, 如果能设计一个概率分布 $q(.|.)$，使得从初始状态 $x^{(0)} \in \chi$ 出发，在 $t > 0$ 时按照该分布不断采样，最终得到的 x 的经验分布和 $p(x)$ 是一致的:

$$x^{(t)} \sim q(x|x^{(t-1)})t = 1, 2, ... \tag{2.39}$$

那么只需要迭代足够多次，即可以得到真实分布的估计。

这个方法称为马尔科夫链蒙特卡洛方法 (Markov chain Monte Carlo,MCMC)[44]，其关键在于找到一个使上述假设成立的 q。Metropolis-Hastings 算法提供了一种通用的框架来构造 q，具体细节可参考 [45]。

这里介绍一种特殊的 MCMC，称为 Gibbs 采样。假设样本空间是 N 维的，对于待采样的 N 个变量 $x = \{x_1, ..., x_N\}$，如果每个变量在给定其他 N-1 个变量时的条件概率都是可以计算的，那么在第 t 次迭代时进行如下的采样:

$$\begin{aligned} x_i^{(t)} &\sim p(x_i|x_j^{(t-1)}, j \neq i) \quad i = i(t) \\ x_j^{(t)} &\sim x_j^{(t-1)} \qquad\qquad\qquad\quad j \neq i(t) \end{aligned} \tag{2.40}$$

即对第 $i(t)$ 维变量进行重新采样，而保持其他维的变量不变。

当这个过程迭代进行无限次时，$x_{(t)}$ 可以收敛到 $p(x)$ 的真实采样。

2.3.1 有限混合模型的 Gibbs 采样

这里用有限混合模型的推断作为例子来展示 Gibbs 采样的过程。在有限模型 (图2.4) 中，观察值为 $x = \{x_i\}_{i=1}^N$，需要推断的变量为成分指示变量 $z = \{z_i\}_{i=1}^N$、混合参数 π、成分参数 $\theta = \{\theta_k\}_{k=1}^K$。这里主要是展示 Gibbs 采样，不考虑模型的参数的学习问题，假定其为固定值。

根据 Gibbs 采样算法，需要求得每个变量在其他变量已知时的条件概率。首

先考虑 z，对于 z_i，根据概率模型中的概率依赖关系，可以得到:

$$\begin{aligned} p(z_i = k|z^{-i}, x, \pi, \theta_i, ..., \theta_K) &\sim p(z_i = k|\pi)p(x_i|z_i, \theta_1, ...\theta_K) \\ &= \pi_k f(x_i|\theta_k) \end{aligned} \tag{2.41}$$

其中 z^{-i} 表示 z 中除去 z_i 的其他变量。

下面给出详细的解释。首先，有

$$\begin{aligned} p(z_i = k|z^{-i}, x, \pi, \theta_i, ..., \theta_K) &= \frac{p(z_i = k, x_i|z^{-i}, x^{-i}, \pi, \theta_i, ..., \theta_K)}{\sum_{k=1}^{K} p(z_i = k, x_i|z^{-i}, x^{-i}, \pi, \theta_i, ..., \theta_K)} \\ &\sim p(z_i = k, x_i|z^{-i}, x^{-i}, \pi, \theta_i, ..., \theta_K) \\ &\sim p(z_i = k|z^{-i}, x^{-i}, \pi, \theta_i, ..., \theta_K)p(x_i|z_i, z^{-i}, x^{-i}, \pi, \theta_i, ..., \theta_K) \end{aligned} \tag{2.42}$$

对于右手边第一项，当观察到 π 时，z_i 和 z^{-i}，θ^{-z_i}，x^{-i} 条件独立。当未观察到 x_i 时，z_i 和 θ_{z_i} 独立，从而第一项可以写为 $p(z_i = k|\pi)$。对于右手边第二项，当观察到 z_i 和 $theta_{z_i}$ 时，x_i 和其余变量条件独立，故可以写为 $p(x_i|z_i, \theta_{z_i})$。

对于 π 和 $\theta = \{\theta_k\}_{k=1}^{K}$，可知他们在观察到 z 的情况下是条件独立的，结合模型中其他的独立性质，有:

$$p(\pi, \theta_i, ..., \theta_K|z, x) = p(\pi|z)\prod_{k=1}^{K} p(\theta_k|\{x_i|z_i = k\}) \tag{2.43}$$

其中每一项利用贝叶斯条件法则即可求得。

2.3.2 Collapsed Gibbs 采样

Gibbs 采样需要迭代较多次才能收敛，根据 Rao–Blackwellized 理论，如果可以用解析的方法积分掉某些变量以减小采样空间，则采样的过程可以更快的收敛。这种方法通常称为 Rao–Blackwellized Gibbs 采样或者 collapsed Gibbs 采样。

仍然考虑有限混合模型的采样问题，由于为 π 增加的是共轭先验，故考虑将 π 解析的积分掉。则此时待采样变量减少为 $z = \{z_i\}_{i=1}^{N}$ 和 $\theta = \{\theta_k\}_{k=1}^{K}$。根据概率图模型中的独立关系，有:

$$p(z_i = k|z^{-i}, x, \theta_i, ..., \theta_K) \sim p(z_i = k|z^{-i})p(x_i|z_i, \theta_1, ...\theta_K) \tag{2.44}$$

如果 θ 的先验 $G_0(\lambda)$ 是共轭先验，则可以将 θ 也解析的积分掉。此时的采样空间进一步变小，待采样的变量仅为 $z = \{z_i\}_{i=1}^{N}$，根据概率图模型中的独立关系，

有:

$$p(z_i = k|z^{-i}, x, \theta_i, ..., \theta_K) \sim p(z_i = k|z^{-i})p(x_i|\{x_j|z_j = k, j \neq i\}) \tag{2.45}$$

3 Dirichlet 过程

如今，使用概率模型来建模观察数据的分布已经成为了机器学习领域的主流方法。然而，当数据量和模型复杂度 (通常指参数个数) 不一致时，传统的参数模型会导致过拟合或者欠拟合的问题，而模型的复杂度选择需要依靠人工选择，比如做多组实验进行交叉验证。非参数模型则通过在一个无限维的参数空间里寻找合适模型，使得参数的个数可以根据数据自动的调整来解决这一问题。近几年 Guassian 过程、Dirichlet 过程等方法逐渐进入研究者的视野，许多相关的模型和算法被提出，并在某些应用上取得了良好的效果。本章主要介绍其中与 Dirichlet 过程相关的模型和算法。

3.1 Dirichlet 过程

随机过程是定义在函数空间上的分布，它的每个采样都是一个概率测度。Dirichlet 过程则是一种特殊的随机过程：如果对服从于某个随机过程的概率分布 G，G 在任何划分下都是服从 Dirichlet 分布的，则称该随机过程为 Dirichlet 过程。其严格定义如下：

假设 G_0 是测度空间 Θ 上的随机概率分布，α_0 是正实数, 对于测度空间 Θ 上的概率分布 G，如果其满足以下条件：

对测度空间 Θ 的任意一个有限可测划分 $(A_1, ..., A_r)$，均有以下关系存在：

$$(G(A_1), ..., G(A_r)) \sim Dir(\alpha_0 G_0(A_1), ..., \alpha_0 G_0(A_r)) \tag{3.1}$$

则称 G 服从由基分布 G_0 和参数 α_0 组成的 Dirichlet 过程 (Dirichlet process,DP)，记作 $G \sim DP(\alpha_0, G_0)$。根据 Dirichlet 分布的性质和式(3.1)可知，对于任意的 $T \subset \Theta$，有：

$$\mathbb{E}[G(T)] = G_0(T) \tag{3.2}$$

所以 G_0 表征了其中心，而 a_0 表征了其分散程度.

反之，如果满足 $G \sim DP(\alpha_0, G_0)$，则有式(3.1)成立.

共轭性 与Dirichlet分布类似, Dirichlet过程具有共轭性, 对于$G \sim DP(\alpha_0, G_0)$, 若观测到样本$\theta \sim G$，根据 Dirichlet 分布的共轭性，有:

$$(G(A_1), ..., G(A_r)|\theta \in A_k) \sim Dir(\alpha_0 G_0(A_1), ..., \alpha_0 G_0(A_k) + 1, ..., \alpha_0 G_0(A_r)) \quad (3.3)$$

从而可知

$$p(G|\theta_1, ..., \theta_N, \alpha, G_0) \sim DP(\alpha_0 + N, \frac{1}{\alpha_0 + N}(\alpha_0 G_0 + \sum_{i=1}^{N} \delta_{\theta_i})) \quad (3.4)$$

下面通过引入 stick-breaking 构造和中国餐馆过程，给出 DP 另外两个重要性质.

3.1.1 Stick-breaking 构造

Ferguson 证明了服从 DP 的测度是以概率 1 离散的 [46]，Sethuraman 给出了一种称为 stick-breaking 的方法，可以构造出服从 DP 的概率测度，直观的展示了这一性质.

按如下分布，生成两个独立的变量序列 $\boldsymbol{\beta}$ 和 $\boldsymbol{\theta}$:

$$\beta_k \sim Beta(1, \alpha_0), \theta_k \sim G_0 \quad (3.5)$$

根据这两个变量序列，定义随机分布 G:

$$\pi_k = \beta_k \prod_{l=1}^{k-1}(1 - \beta_l), G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k} \quad (3.6)$$

其中 $Beta(1, \alpha_0)$ 是 beta 分布,δ_θ 表示一个集中在点 θ 处的概率测度（即在 θ 处测度为 1 其他点测度均为 0 的一个概率测度)。Sethuraman 证明了通过这一过程构造出的概率测度 G 服从 $DP(\alpha_0, G_0)$。

这里需要注意，按照这一过程构造的序列 π 以概率 1 满足 $\sum_{k=1}^{\infty} \pi_k = 1$。在文献中，通常用 $\pi \sim GEM(\alpha_0)$ 表示 π 的构造。

3.1.2 中国餐馆过程

预测分布 对于一个服从 DP 分布的概率测度 G，当已经观察到 G 的一系列采样 $\phi_1, ...\phi_N$ 时，考察对下个采样 ϕ_{N+1} 的预测分布:

$$p(\phi_{N+1}|\phi_1, ..., \phi_N, \alpha_0, G_0) = \int p(\phi_{N+1}|G)p(G|\phi_1, ..., \phi_N)dG \quad (3.7)$$

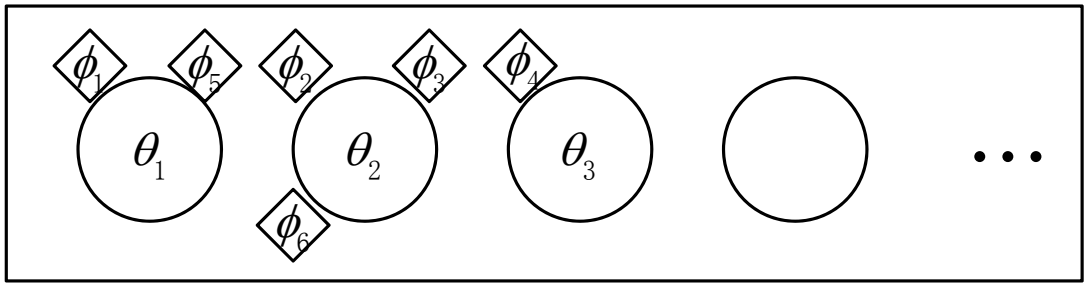

图 3.1: 中国餐馆过程示意图

上式右侧其实是 $\mathbb{E}[G|\phi_1,...,\phi_n]$，根据式(3.2)和式(3.4)，有:

$$p(\phi_{N+1}|\phi_1,...,\phi_N,\alpha_0,G_0)=\frac{1}{\alpha_0+N}(\sum_{i=1}^{N}\delta_{\phi_i}+\alpha_0 G_0) \tag{3.8}$$

这一预测分布对应的过程称为 Blackwell-MacQueen Urn 模型或者 Polya Urn 模型：假设有一个罐子，其中装有各种颜色的球 (颜色可以重复)，每次或者按等概率取出一个球，然后放回两个这一颜色的球，或者按照正比于 α 的概率往罐子内放入一个新球 (颜色可能是新的，也可能是罐子内已有)。

中国餐馆过程 上述的 Polya urn 模型中，ϕ 可能是重复的，故可以用 θ_1 到 θ_K 来表示 ϕ_1 到 ϕ_N 的所有可能的不同取值，进而重写式(3.8):

$$p(\phi_{N+1}|\phi_1,...,\phi_N,\alpha_0,G_0)=\frac{1}{\alpha_0+N}(\sum_{k=1}^{K}N_k\delta_{\theta_k}+\alpha_0 G_0) \tag{3.9}$$

其中 N_k 表示 $\phi_i=\theta_k$ 的个数。这个式子可以看出 DP 的聚类性质，即对于未观察的 ϕ_{N+1}，其和某些已有的 ϕ_k 取值相同的概率是严格大于 0 的概率。为了更好地表达此性质，这里定义一个用于指示类目的变量 $\boldsymbol{z}$，令 $z_i=k$ 表示 $\phi_i=\theta_k$，则有:

$$p(z_{N+1}=z|z_1,...,z_N,\alpha_0,G_0)=\frac{1}{\alpha_0+N}(\sum_{k=1}^{K}N_k\delta_k+\alpha_0\delta_{k^{new}})) \tag{3.10}$$

其中，k^{new} 表示一个新的类。Pitman 和 Dubins 受到旧金山唐人街里几乎可以坐下无限人的中式餐馆的启发，将上述这个过程比喻为中国餐馆过程 (Chinese restaurant process，CRP)：假设有一家可以容纳无限张桌子的中国餐馆，每张桌子记为 θ_k，每位顾客记为 ϕ_i，第 1 个顾客就坐于第 1 张桌子，第 i 个顾客或者按正比于 n_k(已经就坐于第 k 张桌子上的顾客数) 的概率就坐于第 k 张桌子，或者按正比于 α_0 的概率就坐于一张新桌子。图3.1直观的展示了这一过程。

s(n，m)	m=0	m=1	m=2	m=3	m=4
n = 0	1	0	0	0	0
n = 1	0	1	0	0	0
n = 2	0	1	1	0	0
n = 3	0	2	3	1	0
n = 4	0	6	11	6	1

表 3.1: 第一类 striling 数表

可以看出，DP 不仅具有聚类性质，而且他的聚类数是可以随着观察的增加而变化的，这正是其解决模型选择的关键所在。

Antoniak 定理 Antoniak[47] 证明了，如果知道当前的总样本数 n(顾客数)，其当前的不同成分个数 m(桌子数) 满足如下分布：

$$p(m|n,\alpha_0) \propto s(n,m)(\alpha_0)^m \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0+n)} \tag{3.11}$$

其中 s(n,m) 为第一类 Stirling 数。表3.1给出了 Stirling 数的一些值。其中 $s(0,0)=1, s(n,0)=0, s(n,n)=1$，当 $m>n$ 时，$s(n,m)=0$。其他情况下，$s(n+1,m)=s(n,m-1)+ns(n,m)$。

3.2 Dirichlet 过程混合

利用 Dirichlet 过程可以将完全相同的两个观察划分到一起，但这个性质不能直接用来做聚类。回顾2.2.4节介绍的混合模型，对于不同的观察，如果其对应的隐成分是一致的，就将其聚为一类，这样便可以用混合模型进行聚类分析。式(2.33)和式(2.34)给出了有限混合模型一种等价形式，不过其中观察 x_i 对应的隐成分变量 ϕ_i 的先验是一个含有 K 个成分的离散分布 G，这导致了需要人工确定 K 的取值。根据上一节可知，服从 DP 的分布是以概率 1 离散的无限维离散分布，故考虑用一个服从 DP 的 G 来替换式(2.33)和式(2.34)中的 G，即一个含有无限个成分的 G，得到：

$$\begin{aligned} G &\sim DP(\alpha_0, G_0) \\ \phi_i &\sim G \\ x_i &\sim F(\phi_i) \end{aligned} \tag{3.12}$$

这个模型称为 Dirichlet 过程混合模型 (Dirichlet process mixture model,DPMM)[47]。

回忆 stick-breaking 构造，G 可以写成 $\sum_{k=1}^{\infty}\pi_k\delta_{\theta_k}$，如果使用变量 z_i，令 $z_i=k$

指示 $\phi_i = \theta_k$，则可得到其等价表示:

$$\begin{aligned} \pi &\sim Gem(\alpha_0) \\ \theta_k &\sim G_0 \\ z_i &\sim \pi \\ x_i &\sim F(\theta_{z_i}) \end{aligned} \tag{3.13}$$

可以看出，通过 stick-breaking 构造得到的表示形式，相当于对式(2.31)和式(2.32)中的有限混合模型进行了从 K 维到无限维的扩展。此时，先验 G_0 仍然是离散的，从而保证每个点的概率质量 (probability mass) 都是正的，Dirichlet 过程的性质保证了能以严格的正概率采样出已经出现过的成分，并且可能采样出新的成分，这使得这一模型可以用来将进行聚类，并且类目个数可以从根据样本而变化。

3.2.1 推断方法

本文使用 Gibbs 采样的方法进行模型的推断 [48]，假设观察数据 $\boldsymbol{x} = \{x_1, ..., x_N\}$，需要采样的未知随机变量为 θ_k 和 ϕ_i。这里用对 z_i 的采样来代替对 ϕ_i 的采样，从而简化过程。参考式(2.44)，z_i 在其他变量条件下的后验概率为:

$$p(z_i = k|\boldsymbol{z}^{-i}, \boldsymbol{\theta}, \boldsymbol{x}) \propto p(z_i = k|\boldsymbol{z}^{-i}) p(x_i|\theta_k) \tag{3.14}$$

其中第一项可看做是 z_i 的先验，即中国参馆过程中的预测分布（式(3.10)）。第二项可看做是似然函数，即 $f(x_i|\theta_k)$。进而得到:

$$p(z_i = k|\boldsymbol{z}^{-i}, \boldsymbol{\theta}, \boldsymbol{x}) \propto \begin{cases} \alpha_0 f(x_i|\theta_k) & k = k^{new} \\ n_k^{-i} f(x_i|\theta_k) & k \text{ 是已经存在的类目.} \end{cases} \tag{3.15}$$

如果 G_0 是 $F(\theta_k)$ 的先验分布，则可以用 collapsed Gibbs 采样，加快采样的收敛速度:

$$p(z_i = k|\boldsymbol{z}^{-i}, \boldsymbol{x}) \propto p(z_i = k|\boldsymbol{z}^{-i}) p(x_i|\boldsymbol{x}^{-i}) \tag{3.16}$$

由于 Dirichlet 过程可以看做是分层 Dirichlet 过程模型单层情况下的特殊形式，所以关于 collapsed gibbs 采样的细节在后面讨论。另外，模型中含有一个超参数 α_0，关于 α_0 的更新算法也将在后面讨论。

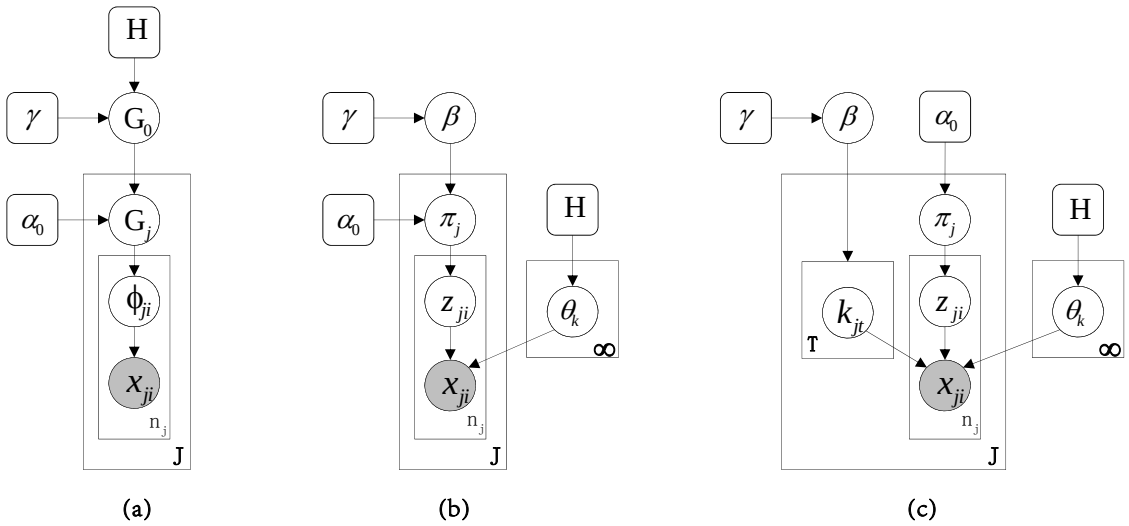

图 3.2: HDP 的图模型表示

3.3 分层 Dirichlet 过程

2.2.6小节讨论的 LDA 模型可以用于建模多组相关数据。在 LDA 中，混合成分的个数是需要人工设置，从而导致了模型选择的问题。这一节，利用 Dirichlet 过程，建立一个含有无限个混合成分的多层贝叶斯模型。

LDA 中假设每组数据是一个有限混合模型，这里假设每组数据都是一个 DPM 模型，第 j 组数据的 x_{ji} 对应的隐成分变量 ϕ_{ji} 服从 G_j 分布，而 G_j 服从一个 Dirichlet 过程 $DP(\alpha_0, G_0)$。注意，和 LDA 相同，G_i 和 G_j 能取值的点是共享的。而跟 LDA 不同的是，这里要求每个 G_j 都是以概率 1 离散的，即可以在无限个点取到值。所以基分布 G_0 不能是连续分布，否则每个 G_j 虽然是概率 1 离散的，却无法共享成分。而 G_0 是有限离散分布也不行，因为这样每个 G_j 都变成了有限离散的了，所以考虑让 G_0 也服从一个 Dirichlet 过程，这样 G_0 是概率 1 离散的，且每个 G_j 都共享这些离散的点，从而满足条件。所以令 G_0 服从一个中心参数为 γ，基分布为 H 的 Dirichlet 过程，即:

$$G_0 \sim DP(\gamma, H) \tag{3.17}$$

第 j 组数据的生成过程如下:

$$\begin{aligned} G_j &\sim DP(\alpha_0, G_0) \\ \phi_{ji} &\sim G_j \\ x_{ji} &\sim F(\phi_{ji}) \end{aligned} \tag{3.18}$$

这一模型称为分层 Dirichlet 过程模型 (Hierarchical Dirichlet processes model,HDP)[8]，

图3.2（a）给出了它的图模型表示。

3.3.1 Stick-Breaking 构造

因为 G_0 服从 $DP(\gamma, H)$，根据 stick-breaking 构造，G_0 可以写成：

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k} \tag{3.19}$$

其中 $\theta_k \sim H$，$\boldsymbol{\beta} = (\beta_i)_{i=1}^{\infty} \sim Gem(\gamma)$。

可知 G_0 在 $\boldsymbol{\theta} = (\theta_i)_{i=1}^{\infty}$ 处有值，而每个 G_j 是服从 $DP(\alpha_0, G_0)$ 的，所以每个 G_j 也在这些点处有值，G_j 可写为：

$$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\theta_k} \tag{3.20}$$

记 $\boldsymbol{\pi}_j = (\pi_{jk})_{k=1}^{\infty}$。因为当给定 G_0 时 G_j 是相互独立的，所以当给定 $\boldsymbol{\beta}$ 时 $\boldsymbol{\pi}_j$ 是相互独立的。下面分析 $\boldsymbol{\pi}_j$ 和 $\boldsymbol{\beta}$ 的关系。根据 Dirichlet 过程的定义，对于 Θ 上的任意可测划分 $(A_1, ..., A_r)$，有：

$$(G_j(A_1), ..., G_j(A_r)) \sim Dir(\alpha_0 G_0(A_1), ..., \alpha_0 G_0(A_r)) \tag{3.21}$$

令 $K_l = \{k : \theta_k \in A_l\}, l = 1, ..., r$，则有：

$$(\sum_{k \in K_1} \pi_{jk}, ..., \sum_{k \in K_r} \pi_{jk}) \sim Dir(\alpha_0 \sum_{k \in K_1} \beta_k, ..., \alpha_0 \sum_{k \in K_r} \beta_k) \tag{3.22}$$

成分变量 ϕ_{ji} 是服从 G_j 分布的，且以概率 π_{jk} 的概率取到 θ_k，和 Dirichlet 过程（式(3.13)）一样，用 $z_{ji} = k$ 来指示 $\phi_{ji} = \theta_k$，则 HDP 模型可以等价表示为：

$$\begin{aligned} &\boldsymbol{\beta} \sim Gem(\gamma) \\ &\boldsymbol{\pi}_j \sim DP(\alpha_0, \boldsymbol{\beta}) \\ &\theta_k \sim H \\ &z_{ji} \sim \boldsymbol{\pi}_j \\ &x_{ji} \sim F(\theta_{z_{ji}}) \end{aligned} \tag{3.23}$$

这一等价表示如图3.2（b）所示。

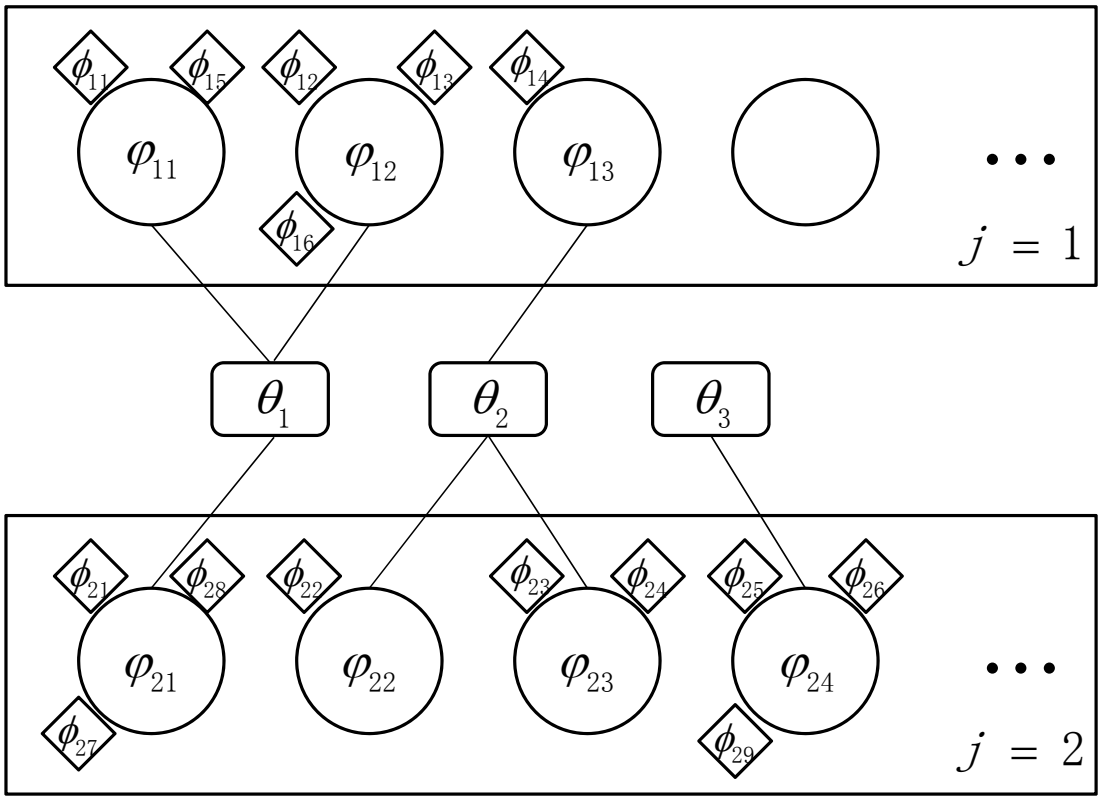

图 3.3: 连锁中国餐馆过程示意图

3.3.2 连锁中国餐馆过程

在讨论 Dirichlet 过程时，用中国餐馆过程来描述其采样的预测分布，而此处的 HDP 模型中含有两层 Dirichlet 过程，也可以用类似的过程来描述。其对应的图模型如图3.2（c）所示。

沿用节的定义，ϕ_{ji} 为第 j 组数据中 x_{ji} 对应的成分变量。这里令 $\theta_1, ..., \theta_K$ 表示全局已经出现过的成分变量，他们是独立同分布于 H 的，$\psi_{j1}, ..., \psi_{jT_j}$ 表示第 j 组数据中 T_j 个已经出现过的成分变量，他们是独立同分布于 G_0 的。

每个 ϕ_{ji} 是和一个 ψ_{jt} 关联的，而每个 ψ_{jt} 是和一个 θ_k 关联的。令 n_{jt} 表示和 ψ_{jt} 关联的 ϕ_{ji} 的个数，m_{jk} 表示第 j 组数据中和 θ_k 关联的 ψ_{jt} 的个数，$m_k = \sum_j m_{jk}$ 表示和 θ_k 关联的所有 ψ_{jt} 的个数。因为 G_j 和 G_0 都是服从 Dirichlet 过程的分布，ϕ_{ji} 和 ψ_{jt} 分别服从 G_j 和 G_0，所以根据式(3.9)可知:

$$p(\phi_{ji}|\phi_{j1}, ..., \phi_{ji-1}, \alpha_0, G_0) = \frac{1}{\alpha_0 + i - 1}(\sum_{t=1}^{T_j} n_{jt}\delta_{\psi_{jt}} + \alpha_0 G_0) \tag{3.24}$$

$$p(\psi_{jt}|\psi_{11}, \psi_{12}, ..., \psi_{21}, ..., \psi_{jt-1}, \gamma, H) = \frac{1}{\sum_k m_k + \gamma}(\sum_{k=1}^{K} m_k\delta_{\theta_k} + \gamma H) \tag{3.25}$$

如果仿照式(3.10)，用指示变量 $t_{ji}=t$ 表示 $\phi_{ji}=\psi_{jt}$，$k_{jt}=k$ 表示 $\psi_{jt}=\theta_k$，则得到：

$$p(t_{ji}|t_{j1},...,t_{ji-1},\alpha_0)\sim\sum_{t=1}^{T_j}n_{jt}\delta_t+\alpha_0\delta_{t^{new}} \tag{3.26}$$

$$p(k_{jt}|k_{11},k_{12},...,k_{21},...,kjt-1,\gamma)\sim\sum_{k=1}^{K}m_k\delta_k+\gamma\delta_{k^{new}} \tag{3.27}$$

上面的过程称为连锁中国餐馆过程 (Chinese restuarant franchise，CRF)。在这个比喻里面，假设一个中国餐馆连锁店有 J 家分店 (每家分店对应一组数据)，这些分店的菜单都是相同的且有无限种菜品 (组间共享无限个成分)，每个餐馆里每张桌子上只有一道菜，由第一个坐到这张桌子上的顾客点选，之后坐到这张桌子的所有顾客分享这一道菜。不同餐馆的桌子上可能会上同一道菜，而同一个餐馆上的不同桌子上也可能有同一道菜。模型里的 ϕ_{ji} 对应着顾客，ψ_{jt} 对应着桌子，θ_k 对应着菜。

当一位顾客进入第 j 家餐馆，他以某种概率坐在某张已经有人的桌子上，以剩余的概率坐在一张新的桌子上，这对应着式(3.24)，当他坐在一个新桌子上时，他会根据菜单中的菜在所有连锁店中的流行程度点一份新的菜，这对应着式(3.25)。其示意图如图3.3所示。

3.4 HDP 的推断

HDP 的推断需要利用估计推断方法，本文主要介绍基于 Gibbs sampling 的方法。

3.4.1 基于 CRF 的采样方法

在利用 CRF 描述的模型里，未知的随机变量是 θ_k，ψ_{jt} 和 ϕ_{ji}。这里利用 Gibbs 采样的方法，根据每个变量在其他变量上的条件概率每次采样单个变量。由于 θ_k，ψ_{jt} 和 ϕ_{ji} 都是分布的参数，如高斯分布的均值和方差，如果直接采样，会需要大量的存储空间。所以，可以只采样 θ_k，而对于 ψ_{jt} 和 ϕ_{ji}，则通过采样相应的指示变量 k_{jt} 和 t_{ji} 得到，由于 θ_k 的个数只有 K 个，采样过程更加有效。

这样，用于采样的状态空间变为了 $\boldsymbol{t}$，$\boldsymbol{k}$，$\boldsymbol{\theta}$，其中 $\boldsymbol{t}$ 的维数是固定的，即观察数，而 $\boldsymbol{k}$ 和 $\boldsymbol{\theta}$ 的维数是不固定的，所以采样空间是一个无限可数维度的. 不过，对于每一步采样而言，其维度是有限的。

采样 t 这里需要从 t_{ji} 在其他变量下的条件概率中采样 t_{ji}。根据条件概率公式，只需计算 t_{ji} 的条件先验分布和生成 x_{ji} 的似然，即可得到 t_{ji} 的条件后验分

布。

$$p(t_{ji}=t|\boldsymbol{t}^{-ji},\boldsymbol{k},\boldsymbol{\theta},\boldsymbol{x})\propto p(t_{ji}=t|\boldsymbol{t}^{-ji})p(x_{ji}|t_{ji},\boldsymbol{k},\boldsymbol{\theta}) \tag{3.28}$$

式中右手第一项即条件先验概率，根据中国餐馆过程可知，t_{ji} 等于一个已经存在的 t 的概率是正比于 n_{jt}^{-i} 的，而等于一个新的 t^{new} 的概率是正比于 α_0 的。当 t_{ji} 等于一个已经存在的 t 时，x_{ji} 的似然为 $f(x_{ji}|\theta_{k_{jt}})$。当 t_{ji} 等于 t^{new} 时，情况要稍微复杂一些，这时需要为餐馆 j 采样出一个新的菜的 $\psi_{jt^{new}}$，即采样出一个 $k_{jt^{new}}$，如果 $k_{jt^{new}}$ 是一个新的 k，即 $k_{jt^{new}}=K+1$，则还要从基分布中生成一个新的 θ，其过程如下

$$k_{jt^{new}}|\boldsymbol{k}\sim\sum_{k=1}^{K}\frac{m_k}{\sum_k m_k+\gamma}\delta_k+\frac{\gamma}{\sum_k m_k+\gamma}\delta_{k^{new}},\theta_{k^{new}}\sim H \tag{3.29}$$

此时 x_{ji} 的似然为 $f(x_{ji}|\theta_{k_{jt^{new}}})$。故:

$$p(t_{ji}=t|\boldsymbol{t}^{-ji},\boldsymbol{k},\boldsymbol{\theta},\boldsymbol{x})\propto\begin{cases}\alpha_0 f(x_{ji}|\theta_{k_{jt}}) & t=t^{new},\\ n_{jt}^{-i}f(x_{ji}|\theta_{k_{jt}}) & t\text{ 为已经有顾客的餐桌.}\end{cases} \tag{3.30}$$

此处可以对 $k_{jt^{new}}$ 进行平滑，即不是采样出一个新的 k 值，而是考虑在 $k_{jt^{new}}$ 所有可能取值下的期望. 此时似然函数为:

$$\boldsymbol{E}_{k_{jt^{new}}|\boldsymbol{k}}[f(x_{ji}|\theta_{k_{jt^{new}}})]=\sum_{k=1}^{K}\frac{m_k}{\sum_k m_k+\gamma}f(x_{ji}|\theta_k)+\frac{\gamma}{\sum_k m_k+\gamma}f(x_{ji}|\theta_{k^{new}}). \tag{3.31}$$

采样 k 采样 k_{jt} 和采样 t_{ji} 的过程类似，区别在于似然函数上，对于 t_{ji} 采样时，和 t_{ji} 相关的 x 只有 x_{ji}，而对于 k_{jt}，他的采样值会影响到第 j 个餐馆第 t 个桌上所有的 x_{ji}(即 $t_{ji}=t$ 的 x_{ji})，所以其似然部分为 $\prod_{i:t_{ji}=t}f(x_{ji}|\theta_k)$. 故:

$$p(k_{jt}=k|\boldsymbol{t},\boldsymbol{k}^{-jt},\boldsymbol{\theta},\boldsymbol{x})\propto\begin{cases}\gamma\prod_{i:t_{ji}=t}f(x_{ji}|\theta_k) & k=k^{new},\\ m_k^{-t}\prod_{i:t_{ji}=t}f(x_{ji}|\theta_k) & k\text{ 为已经有顾客点过的菜.}\end{cases} \tag{3.32}$$

采样 θ θ_k 的取值会影响到所有和 $\phi_{ji}=\theta_k$(即 $k_{jt_{ji}}=k$) 的 x_{ji} 的似然，所以 θ_k 的后验分布如下:

$$p(\theta_k|\boldsymbol{t},\boldsymbol{k},\boldsymbol{\theta}^{-k},\boldsymbol{x})\propto h(\theta_k)\prod_{ji:k_{jt_{ji}}=k}f(x_{ji}|\theta_k) \tag{3.33}$$

collapsed gibbs 采样 如果 H 是 $f(\theta)$ 的共轭先验，比如 H 是 Dirichlet 分布而 $f(\theta)$ 是多项分布，则可以利用边缘概率公式 θ 消去，利用 collapsed Gibbs 采样加

快采样的收敛速度。

首先考虑采样 t_{ji} 时需要计算的似然。当 θ_k 已知时，x_{ji} 的似然为 $f(x_{ji}|\theta_k)$，而当不对 θ_k 进行采样时，根据式(2.45)，则需要计算 x_{ji} 在 $\boldsymbol{x}^{-ji}$ 上的后验分布：

$$f_k(x_{ji}|\boldsymbol{x}^{-ji}) = \frac{\int f(x_{ji}|\theta_k)\prod_{j'i'\neq ji,z_{j'i'}=k} f(x_{j'i'}|\theta_k)H(\theta_k)d\theta_k}{\int \prod_{j'i'\neq ji,z_{j'i'}=k} f(x_{j'i'}|\theta_k)H(\theta_k)d\theta_k} \tag{3.34}$$

其中 $z_{ji} = k_{jt_{ji}}$. 若 $k = k^{new}$，此时的后验分布为 $\int f(x_{ji}|\theta_k)H(\theta_k)d\theta_k$.

当 t_{ji} 的采样为 t^{new} 时，可以根据式(3.4.1)采样出一个 $k_{jt^{new}}$，如果 $k_{jt^{new}}$ 是一个新的 k^{new}，则此时的似然为 $\int f(x_{ji}|\theta_k)H(\theta_k)d\theta_k$。

如果使用式(3.31)中的平滑算法，则此时该式变为：

$$\boldsymbol{E}_{k_{jt^{new}}|\boldsymbol{k}}[f(x_{ji}|\boldsymbol{x}^{-ji})] = \sum_{k=1}^{K}\frac{m_k}{\sum_k m_k+\gamma}f_k(x_{ji}|\boldsymbol{x}^{-ji}) + \frac{\gamma}{\sum_k m_k+\gamma}f_{k^{new}}(x_{ji}|\boldsymbol{x}^{-ji}) \tag{3.35}$$

在采样 k_{jt} 时，需要计算第 j 个餐馆第 t 个桌子上所有 x 的似然，记作 $\boldsymbol{x}_{jt}$，这时后验分布变成了：

$$f_k(\boldsymbol{x}_{jt}|\boldsymbol{x}^{-jt}) = \frac{\int \boldsymbol{x}_{jt}|\theta_k)\prod_{j'i'\neq ji,z_{j'i'}=k} f(x_{j'i'}|\theta_k)H(\theta_k)d\theta_k}{\int \prod_{j'i'\neq ji,z_{j'i'}=k} f(x_{j'i'}|\theta_k)H(\theta_k)d\theta_k} \tag{3.36}$$

若 $k_{jt} = k^{new}$，此时的后验分布为 $\int \prod_{t_{ji}=t} f(x_{ji}|\theta_k)H(\theta_k)d\theta_k$.

若 H 是参数为 η 的 Dirichlet 分布而 $f(\theta)$ 是多项分布，则式(3.36)变为：

$$f_k(\boldsymbol{x}_{jt}|\boldsymbol{x}^{-jt}) = \frac{\Gamma(n_k^{-\boldsymbol{x}_{jt}}+V\eta)}{\Gamma(n_k^{-\boldsymbol{x}_{jt}}+n^{\boldsymbol{x}_{jt}}+V\eta)}\frac{\prod_v\Gamma(n_{kv}^{-\boldsymbol{x}_{jt}}+n_v^{\boldsymbol{x}_{jt}}+V\eta)}{\prod_v\Gamma(n_{kv}^{-\boldsymbol{x}_{jt}}+V\eta)} \tag{3.37}$$

其中 $n_{kv}^{-\boldsymbol{x}_{jt}}$ 表示除了 $\boldsymbol{x}_{jt}$ 以外，满足 $x_{ji}=v$,$k_{ji}=k$ 的 x_{ji} 的个数。$n_k^{-\boldsymbol{x}_{jt}}$ 表示除了 $\boldsymbol{x}_{jt}$ 以外，满足 $k_{ji}=k$ 的 x_{ji} 的个数。$n_v^{\boldsymbol{x}_{jt}}$ 表示满足 $t_{ji}=t, x_{ji}=v$ 的 x_{ji} 的个数。$n^{\boldsymbol{x}_{jt}}$ 表示满足 $t_{ji}=t$ 的 x_{ji} 的个数。

而式(3.34)只是式(3.36)的一个特殊情况：

$$f_k(x_{ji}=v|\boldsymbol{x}^{-ji}) = \frac{n_{kv}^{-x_{ji}}+\eta)}{\Gamma(n_{kv}^{-x_{ji}}+V\eta)} \tag{3.38}$$

3.4.2 直接分配采样方法

除了从基于中国餐馆过程的表示得到采样算法，也可以从基于 stick-breaking 的表示 (式(3.13)) 得到 HDP 的另一种采样方法。这里直接用 z_{ji} 来表示 x_{ji} 和 θ_k 的

对应关系，故本文称其为直接分配采样方法 [8, 49]。

根据模型，需要采样的变量有 $\boldsymbol{z}$，$\boldsymbol{\pi}$，$\boldsymbol{\beta}$ 和 $\boldsymbol{\theta}$. 由于 $\boldsymbol{\pi}_j$ 服从 Dirichlet 分布，而其是多项分布 (z_{ji} 服从一个参数为 $\boldsymbol{\pi}_j$ 的多项分布) 的共轭先验，类似于式(2.44)，将 $\boldsymbol{\pi}$ 积分掉，只对 $\boldsymbol{z}$，$\boldsymbol{\beta}$ 以及 $\boldsymbol{\theta}$ 进行 Gibbs 采样。为了使得采样过程可行，只考虑当前模型中已有的 K 个混合成分 (有限个)，而对于其他所有仍未出现的成分 (无限个)，并不显式表示出每一个成分，而是用单个变量来记录。这里将模型中已有的 K 个混合成分对应的 $\boldsymbol{\beta}$ 记为 β_1 到 β_K，而将其他所有没有出现的成分记为 $\beta_u = \sum_{k=K+1}^{L} \beta_k$。这样采样过程中的 $\boldsymbol{\beta}$ 为 $(\beta_1, ..., \beta_K, ...\beta_u)$。当采样到一个新的混合成分时，对 K 增加 1，而对用来记录未出现成分的变量进行更新。

采样 $\boldsymbol{z}$ 根据 Dirichlet 分布和多项分布的共轭性质，参考式(3.38)，得到 z_{ji} 在 $\boldsymbol{z}^{-ji}$ 和 $\boldsymbol{\beta}$ 条件下的先验分布:

$$f_k(z_{ji} = k|\boldsymbol{z}^{-ji}, \boldsymbol{\beta}) = \frac{n_{jk}^{-z_{ji}} + \alpha_0\beta_k}{n_j^{-z_{ji}} + K\alpha_0\beta_k} \tag{3.39}$$

该式分母对于不同的 k 值是个常量，故可以省去，根据 x_{ji} 的条件分布，得到 z_{ji} 的后验概率为:

$$f_k(z_{ji} = k|\boldsymbol{z}^{-ji}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{x}) \propto (n_{jk}^{-z_{ji}} + \alpha_0\beta_k) f(x_{ji}|\theta_k) \tag{3.40}$$

采样 $\boldsymbol{\beta}$ 利用 Dirichlet 分布和多项分布的共轭性质，得到:

$$p(\boldsymbol{z}|\boldsymbol{\beta}) = \prod_{j=1}^{J} \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + n_j)} \prod_{k=1}^{K} \frac{\Gamma(\alpha_0\beta_k + n_{jk})}{\Gamma(\alpha_0\beta_k)} \tag{3.41}$$

这个式子可以看做是 $\boldsymbol{\beta}$ 的似然函数，注意其中 $\boldsymbol{\beta}$ 出现在 Γ 函数中，所以联合 $\boldsymbol{\beta}$ 的先验得到的 $p(\boldsymbol{\beta}|\boldsymbol{z})$ 并不是一个指数函数族分布，难以从中直接采样出 $\boldsymbol{\beta}$。所以下面通过引入一个辅助变量，构造出一个满足指数族分布的后验概率来采样 $\boldsymbol{\beta}$。

对于式(3.41)中的第二项，可以展开为关于 $\alpha_0\beta_k$ 的多项式函数:

$$\frac{\Gamma(\alpha_0\beta_k + n_{jk})}{\Gamma(\alpha_0\beta_k)} = \prod_{m_{jk}=1}^{n_{jk}} (m_{jk} - 1 + \alpha_0\beta_k) = \sum_{m_{jk}=0}^{n_{jk}} s(n_{jk}, m_{jk})(\alpha_0\beta_k)_{jk}^{m} \tag{3.42}$$

其中 $s(n_{jk}, m_{jk})$ 为第一类 Stirling 数，$\boldsymbol{m} = (m_{jk})$ 是引入的一个辅助变量。

这里构造一个关于 $\boldsymbol{z}$,$\boldsymbol{m}$ 和 $\boldsymbol{\beta}$ 的分布:

$$q(\boldsymbol{z},m,\beta)=\frac{\Gamma(\gamma)}{\Gamma(\gamma_r)\Gamma(r_u)}(\sum_{j=1}^{J}\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0+n_j)})\beta_u^{\gamma_u-1}\prod_{k=1}^{K}\beta_k^{\gamma_r-1}\prod_{j=1}^{J}(\alpha_0\beta_k)^{m_{jk}}s(n_{jk},m_{jk}) \tag{3.43}$$

可以验证，$\sum_m q(\boldsymbol{z},\boldsymbol{m},\boldsymbol{\beta})=p(\boldsymbol{z},\boldsymbol{\beta})$, 所以从分布 q 采样得到的 $\boldsymbol{\beta}$ 是和从分布 $p(\boldsymbol{\beta},\boldsymbol{z})$ 中采样的 $\boldsymbol{\beta}$ 是一致的，从而建立了一个基于辅助变量 $\boldsymbol{m}$ 的采样方法来采样 $\boldsymbol{\beta}$:

$$q(m_{jk}=m|\boldsymbol{z},m_{-jk},\beta|)\propto s(n_{jk},m)(\alpha_0\beta_k)^m \tag{3.44}$$

$$q(\boldsymbol{\beta}|\boldsymbol{z},m)\propto\beta_u^{\gamma-1}\prod_{k=1}^{K}\beta_k^{\sum_j m_{jk}-1} \tag{3.45}$$

回忆 CRF 构造中已经使用了 m_{jk} 来表示第 j 个餐馆中上了第 k 道菜的桌子的个数。这里的辅助变量也使用的相同的记号，其实两个记号一致并不是巧合，他们正是同一个变量，但是由于在直接分配的方法中，是直接采样的 z_{ji}，因为有可能同一餐馆的两个不同桌子都是点的第 k 道菜，所以无法计算出 m_{jk}，而只能去计算它的概率。下面给出证明。

考虑第 j 个餐馆的采样过程，并且只考虑顾客被安排到点了第 k 道菜的桌子 (即所有 $k_{jt}=k$ 的桌子 t) 的情况，这可以看做是第 j 个餐馆的完整采样过程的一个子情况，其先验概率如下:

$$p(t_{ji}=t|\boldsymbol{t}^{-ji},k_{jt}=k,\boldsymbol{\beta},\boldsymbol{x})\propto\begin{cases}\alpha_0\beta_k & t=t^{new},\\ n_{jt}^{-i} & t\text{ 为已经有顾客的餐桌.}\end{cases} \tag{3.46}$$

可以看出，这个子情况可以等价为一个中心参数为 $\alpha_0\beta_k$ 的 Dirichlet 过程对应的中国餐馆过程。根据 Antoniak 定理（式(3.11)），对于一个 Dirichlet 过程，如果知道当前的总采样数，则可以得到一个关于其当前的成分个数的分布。上述过程的总采样数为 $n_{jk}=\sum_{k_{jt}=k}n_{jt}$，其成分个数则是餐馆 j 中所有 $k_{jt}=k$ 的桌子个数，即 m_{jk}, 从而有:

$$p(m_{jk}=m|\text{采样数为}n_{jk},\alpha_0\beta_k)=s(n_{jk},m)(\alpha_0\beta_k)^m\frac{\Gamma(\alpha_0\beta_k)}{\Gamma(\alpha_0\beta_k+n_{jk})} \tag{3.47}$$

此式和式(3.44)一致, 说明辅助变量 m_{jk} 和中国 CRF 中的 m_{jk} 是一致的.

这也说明，当采用基于 CRF 的采样方法时，如果需要采样 $\boldsymbol{\beta}$，可以直接使用 m_{jk}。

这里再介绍另一种推导 $\boldsymbol{\beta}$ 采样公式的方法。假设当前已经有的 K 个不同的菜品将 $\boldsymbol{\Theta}$ 划分为 $\{\theta_1,\theta_2,..\theta_K,\theta_{\tilde{k}}\}$, 其中 $\theta_{\tilde{k}}=\boldsymbol{\Theta}\backslash\bigcup_{k=1}^{K}\{\theta_k\}$ 是当前还没出现过的参数。

根据 DP 的定义，有

$$\begin{aligned}(G_0(\theta_1),...,G_0(\theta_K),G_0(\theta_{\tilde{k}})) &\sim Dir(\gamma H(\theta_1),...,\gamma H(\theta_K),\gamma H(\theta_{\tilde{k}})) \\ &\sim Dir(0,...,0,\gamma)\end{aligned} \tag{3.48}$$

注意，其中 H 是连续空间上的测度，所以在参数点上的测度为 0。假设第 j 个餐馆第 t 个桌子上的菜是 ψ_{jt}, 则有

$$p(G_0(\theta_1),...,G_0(\theta_K),G_0(\theta_{\tilde{k}})|\{\psi_{jt}\}),\gamma) \propto Dir(\sum_j m_{j1}...\sum_j m_{jK},\gamma) \tag{3.49}$$

根据定义，可知 $(G_0(\theta_1),...,G_0(\theta_K),G_0(\theta_{\tilde{k}}))$ 就是 $\boldsymbol{\beta}$。这样得到和式(3.45)一致的采样公式。

采样 $\boldsymbol{\theta}$ 对于 $\boldsymbol{\theta}$ 的采样和基于 CRF 的采样方法中的一致:

$$p(\theta_k|\boldsymbol{t},\boldsymbol{k},\boldsymbol{\theta}^{-k},\boldsymbol{x}) \propto h(\theta_k)\prod_{ji:z_{ji}=k} f(x_{ji}|\theta_k) \tag{3.50}$$

同样，和基于 CRF 的采样方法一样，如果 H 是 $f(\theta)$ 的共轭先验，也可以进一步将 $\boldsymbol{\theta}$ 积分掉，进而只对 $\boldsymbol{z}$，$\boldsymbol{\beta}$ 进行 Gibbs 采样。

采样 $\boldsymbol{\pi}$ 前面在采样 $\boldsymbol{z}$ 和 $\boldsymbol{\beta}$ 时将 $\boldsymbol{\pi}$ 积分掉来加快采样速度，但是有些时候需要采样 $\boldsymbol{\pi}$ 的值，比如在对多文档建模时 $\boldsymbol{\pi}$ 可以用来表征文档在主题层的表示[18]。根据共轭性质，其后验分布如下:

$$p(\boldsymbol{\pi}|\boldsymbol{z},\beta) \propto \prod_{j=1}^{J}{\pi_{ju}}^{\alpha_0\beta_u-1}\prod_{k=1}^{K}\pi_{jk}^{\alpha_0\beta_k+n_{jk}-1} \tag{3.51}$$

3.4.3 超参数的更新

模型含有两个超参数 γ 和 α_0，无论是用哪种采样方法，都需要对其进行更新。这里为这两个参数加上先验，然后根据参数的后验概率进行采样。

α_0 的采样 根据式(3.47), 可以得到:

$$p(T_1,...,T_J|\alpha_0,n_1,...,n_J) = \prod_{j=1}^{J} s(n_j,T_j){(\alpha_0)}^{T_j}\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0+n_j)} \tag{3.52}$$

这里 T_j 是第 j 个餐馆的桌子个数，n_j 是第 j 个餐馆当前观察的总数. 当观察

到 T_j 时，X 是独立于 α_0 的，所以上式可以看做求 α_0 后验时需要的似然函数，在联合 α_0 的先验，就可以得到 α_0 的后验概率，从而可以进行采样更新。

假设 α_0 的先验是一个参数为 a 和 b 的 $gamma$ 分布，则 α_0 的后验为:

$$p(\alpha_0|T_1,...,T_j,a,b) \propto {\alpha_0}^{a-1+\sum_{j=1}^{J}T_j} e^{\alpha_0 b} \prod_{j=1}^{J} s(n_j,T_j) \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0+n_j)} \tag{3.53}$$

由于 α_0 出现在 $\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0+n_j)}$ 中，导致这个后验概率不服从指数族分布。这里使用 West 等人提出的一种基于辅助变量的方法 [50, 51, 52]，使得 α_0 在辅助变量下的后验是服从指数族分布的，从而可以进行采样更新。

首先考虑 $\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0+n_j)}$，可以将其展开成:

$$\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0+n_j)} = \int_0^1 {w_j}^{\alpha_0}(1-w_j)^{n_j-1}(1+\frac{n_j}{\alpha_0})dw_j \tag{3.54}$$

这里定义辅助变量 $\boldsymbol{w}=(w_j)_{j=1}^J$ 和 $\boldsymbol{s}=(s_j)_{j=1}^J$，其中 w_j 在 $[0,1]$ 之间取值而 s_j 是一个只能取 $0,1$ 的二值变量。定义分布:

$$q(\alpha_0,\boldsymbol{w},s) \propto {\alpha_0}^{a-1+\sum_{j=1}^{J}T_j} e^{\alpha_0 b} \prod_{j=1}^{J} {w_j}^{\alpha_0}(1-w_j)^{n_j-1}(\frac{n_j}{\alpha_0})^{s_j} \tag{3.55}$$

可以验证，积分掉 $\boldsymbol{w}$ 和 $\boldsymbol{s}$，得到的正是 α_0 的后验分布:

$$\int q(\alpha_0,\boldsymbol{w},s)d\boldsymbol{w}d\boldsymbol{s} = p(\alpha_0|T_1,...,T_j,a,b) \tag{3.56}$$

从而，可以对三个变量分别进行采样, 得到 α_0 的采样方法:

$$\begin{aligned} &q(\alpha_0|\boldsymbol{w},s) \propto {\alpha_0}^{a-1+\sum_{j=1}^{J}T_j-s_j} e^{\alpha_0(b-\sum_{j=1}^{J}\ln w_j)} \\ &q(w_j|\alpha_0) \propto {w_j}^{\alpha_0}(1-w_j)^{n_j-1} \quad j\in\{1,...,J\} \\ &q(s_j|\alpha_0) \propto (\frac{n_j}{\alpha_0})^{s_j} \quad j\in\{1,...,J\} \end{aligned} \tag{3.57}$$

从三个变量的后验概率形式可以看出，α_0 服从 $gamma$ 分布，w_j 服从 dirichlet 分布，s_j 服从 bernoulli 分布。

γ 的采样 对于参数 γ，可以得到:

$$p(K|\gamma,T) = s(T,K)\gamma_K \frac{\Gamma(\gamma)}{\Gamma(\gamma+T)} \tag{3.58}$$

其中 $T = \sum_j T_j$ 是总桌子数，而 K 是模型中的不同成分个数，即不同种类菜的个数。这个式子和式(3.52)一致，从而可以使用与上面更新 α_0 完全一样的方法来更新 γ。

注意，一般先对变量采样多次，然后更新一遍参数.

4 新闻广播故事分割建模

本章主要研究在广播新闻的语音识别结果上的故事分割，这属于一个自然语言处理的任务。在建模自然语言处理任务时，根据任务的不同，对于数据的可交换性假设也不同，如在文本分类的任务中，通常假设单个文本段内的单词是可交换的，即没有顺序关系，而在分词的任务中，则假设字之间是有顺序关系的，从而对不同的任务建立的模型区别很大。而广播新闻故事分割任务，则介于二者之间，通常先将新闻先切分为一些文本块，在建模时，假设文本块内部是可交换的，而文本块之间是不可交换的。

4.1 故事分割任务分析

故事分割就是将一段文本流分割成一系列段落，使得同一段落内有一定的相关性，而不同段落间有一定的差异性。本文主要研究广播新闻故事分割，此处文本流对应了未分割的完整新闻，段落对应了每个独立新闻故事 (story)，我们的任务是得到这些独立的新闻故事，即确定故事的边界。

在故事分割建模中，一般先根据文本的标点信息将文本流切割为一些子块 (block) [1]，然后在这些子块组成的数据流上进行分割任务。这里假设每个子块内的单词是可交换的，所以可以用一个词频向量来表示。

本文将一段新闻中的不同故事记作 $S_j, j \in 1...J$, 其中 J 表示整段新闻中的故事个数。x_i 表示第 i 个子块，$\boldsymbol{x}_j$ 表示第 j 个故事里的所有子块构成的集合.

4.2 基线系统

4.2.1 TextTiling

通常，同一个故事内部用词相似，比如篮球新闻播报过程中，可能反复提及三分、助攻等词，而政治新闻中则重复出现重申、强调等词，所以可以合理的假设同一个故事 S_j 内部的子块 $\boldsymbol{x}_j$ 在词频上具有较大相似性，而不同故事 S_j 和 S_i 的子块之间具有较大差异性。所以一个直观的分割方法就是定义一个度量相似性的函数，然后计算相邻的 x_i 和 x_{i+1} 之间的相似性，如果其低于某个阈值，就认为

[1] 在新闻广播的分割中，由于数据是语音识别的结果，无法获取标点信息，所以本文利用语音激活检测 (Voice Activity Detection,VAD) 的结果来切割子块。

其之间差异性过大，即分属两个不同的故事。

不过这种简单的方法容易被一些噪声干扰，为了增加鲁棒性，hearst 引入了深度值的概念，即在相似度曲线上，只将处于波谷的点作为边界候选点，然后计算该点到左右两个相邻的波峰点的下降值之和，称之为深度值，如果某点的深度值大于阈值，就认为该点是故事边界。这一方法称为 TextTiling 方法 [13]。

4.2.2 隐语义分析

一般的相似度度量采用词频分布上的余弦函数或者交叉熵, 不过词频分布向量的维度很大，这种度量方法的结果不佳。如果认为每个故事都是由一系列主题混合而成，即不同故事间共享主题，但是每个故事在不同主题上的权重不同，则可以用主题成分上的权重作为特征。这正是2.2.6小节中的 LDA 模型。如果不为混合参数加一个先验，则得到一个非贝叶斯模型，称为 PLSA 模型。

当得到主题成分时，就可以在 x_i 上计算其对应的主题分布，如果使用主题分布作为 x_i 的特征，这样就将一个高维的表示变为一个 K 维的表示。

不过这一方法是有监督的，因为需要提供一份已标注好故事边界且和待分割的文本领域相同的语料来训练出主题成分，不然共享相同的主题这一假设就不成立。[2]

4.2.3 全局最优算法

在 textiling 中，根据相邻点的相似度来寻找边界，这是一种局部的贪心算法。考虑建立一个全局上最优的算法，即对于某种能够衡量段内相似度的度量函数 sim，期望得到一个划分 $\boldsymbol{\pi}$，使得全局的得分最大:

$$\boldsymbol{\pi} = argmax_{\boldsymbol{\pi}} \sum_{j=1}^{J} sim(\boldsymbol{x}_j) \tag{4.1}$$

可以用下面的动态规划算法求的最优解:

$$\begin{cases} f(1) = sim(\boldsymbol{x}_1) \\ f(i) = \max\limits_{k \le i} \{f(k-1) + sim(\boldsymbol{x}_{ki})\} \\ b(i) = \arg\max\limits_{k \le i} \{f(k-1) + sim(\boldsymbol{x}_{ki})\} \end{cases} \tag{4.2}$$

其中 $f(i)$ 是从 u_1 到 u_i 间的文本对应的最优得分，$b(i)$ 记录最优边界的回溯值。

[2]LDA 经常用于无监督建模，因为对于 LDA 而言语料本身已经是划分好的，这里说的有监督，是指整个系统

然而，如果直接使用得分最高的划分，其分割结果和启发式的相似度测量函数的参数有关，如果参数设定不当，则效果较差，所以需要指定一个估计的分段个数 n，此时动态算法要稍微复杂一些，变为:

$$\begin{cases} f(1,i) = sim(\boldsymbol{x}_{1i}) \\ f(n,i) = \max\limits_{n \le k \le i} \{f(n,k-1) + sim(\boldsymbol{x}_{ki})\} & 1 < n \le N \\ b(n,i) = \arg\max\limits_{n \le k \le i} \{f(n,k-1) + sim(\boldsymbol{x}_{ki})\} & 1 \le n \le N \end{cases} \tag{4.3}$$

其中 $f(n,i)$ 是从 u_1 到 u_i 间的文本分成 n 段对应的最优得分，$b(n,i)$ 记录最优边界的回溯值。

4.3 贝叶斯概率模型分割

本文考虑从概率模型出发，为故事分割任务建立一个一致的概率模型，将问题的求解转化为对概率模型中参数的估计和推断问题。假设一段文本流由数个连续的故事组成，故事 j 中的单词是从一个与故事 j 相关的离散分布 θ_j 中生成的，假设不同故事之间都是可以交换的，则 θ_j 是独立同分布于一个概率分布的，取一个参数为 λ 的共轭 Dirichlet 分布。设 x_i 对应的故事为 z_i，则对于某个划分 Z 模型的似然为:

$$p(X|Z,\theta) = \prod_{j=1}^{J} p(X_j|\lambda) \tag{4.4}$$

其中 $X_j = \{x_t|t : z_t = j\}$，$p(X_j|\lambda) = \int p(X_j|\theta_j)p(\theta_j|\lambda)d\theta_j$

由于 $\{x_t|t : z_t = j\}$ 在观察到 θ_j 的情况下是条件独立的，再根据 Dirichlet 分布的性质，可得整段文本的最大似然函数。这一模型可以看做是 Mincut 框架的一种特例，和上述全局最后算法的区别在于其段内相似度的度量函数不是手工设定的，而是从概率建模中自然推导出来的，即:

$$\frac{\Gamma(V\lambda_0)\prod_v \Gamma(n_j^v + \lambda_0)}{\Gamma(n_j + V\lambda_0)\Gamma^V(\lambda_0)} \tag{4.5}$$

这一模型称为基于贝叶斯概率模型分割的方法。

4.4 依赖于距离的中国餐馆过程

假设一段新闻是从如下过程中采样生成的:

首先，采样出每个子块的主题 z_i。然后对每个子块，根据其对应的主题，采样出该子块内的单词。假设子块内的单词是可交换的，他们独立同分布于一个跟

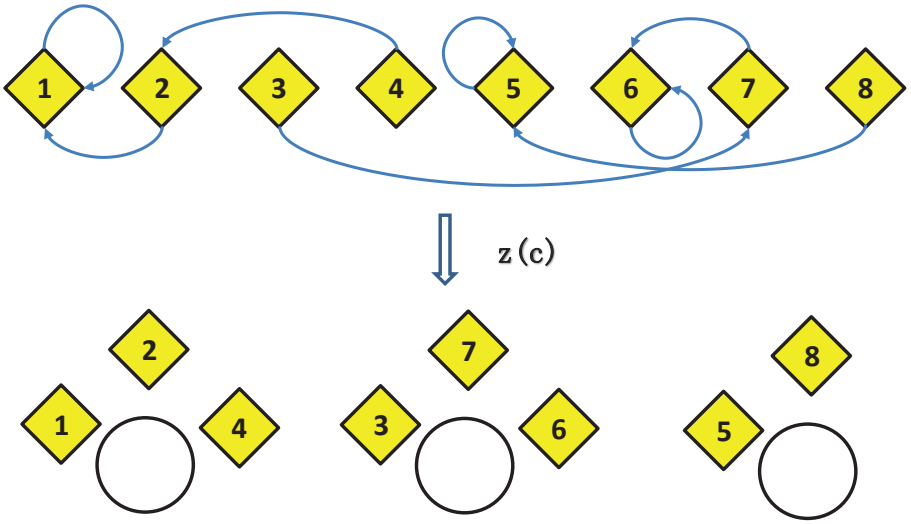

图 4.1: 依赖于距离的中国餐馆过程 (distance dependent CRP,dd-CRP) 示意图.

主题相关的分布。如果假设子块之间也是可交换的，那么用 DP 作为 z 的先验，得到的正是一个 DPM 模型，其中不同的主题对应了中国餐馆过程中的不同桌子，不同的子块对应着不同的顾客。

但是由于同一个故事的几个子块是连续的，虽然可以假设同个故事中的几个子块之间可以交换，可是不同故事间的子块是不能交换的，所以这里并不能用 DP 作为 z 的先验，必须寻找一种具有不可交换性的分布。这里考虑中国餐馆过程的一种变形，并不直接建立起顾客到餐桌的划分，而是建立顾客和顾客之间的关系。如果引入一个顾客间的距离关系 (比如越熟悉的顾客或者是拥有更多相同喜好的顾客之间距离就越小)，认为对于一个新来的顾客，他会选择和与自己距离小的顾客坐在一起, 用 $c_i = j$ 表示 i 顾客选择了和顾客 j 坐在一起，这样通过 $c_i, i = 1...N$ 也可以得到顾客到餐桌的一种划分 $z_i, i = 1...N$。如图4.1中所示，菱形代表顾客圆形代表桌子，顾客之间的箭头代表着顾客选择和谁坐在一起。在这个图中，第一个顾客 c_1 只能选择和自己坐在一起，顾客 c_2 选择和 c_1 坐在一起，顾客 c_4 选择和 c_2 坐在一起，没有其他的顾客和他们坐在一起，这样 c_1 ，c_2，c_4 坐在同一张座子上。这一过程的定义如下:

$$p(c_i = j|D, f, \alpha) \propto \begin{cases} f(d_{ij}) & \text{if } j \neq i \\ \alpha & \text{if } j = i \end{cases} \tag{4.6}$$

其中 D 是一个顾客之间距离测度的集合,d_{ij} 是顾客 c_i 和 c_j 间的距离，$f(d)$ 是一个非增函数，因为顾客间的距离越大，坐在一起的概率就越小。上述这一模型称依赖于距离的中国餐馆过程 (distance dependent Chinese restaurant process ,dd-CRP)。如果顾客之间的距离和他们的顺序是有关系的，那么这个模型具有

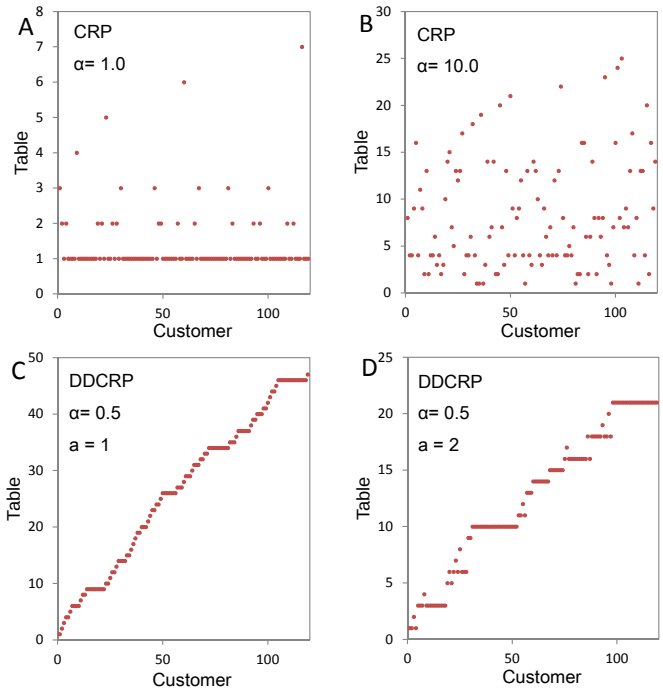

图 4.2: 可以看到，从 CRP 中的采样 (A 和 B) 的统一聚类点较分散，没有时间上的连续性，从而不具有分割性质。而 dd-CRP 中的采样 (C 和 D 图) 则是把相邻的点聚成一类，从而具有很好的分割性质。

不可交换性。本文用这个过程作为故事分割中子块主题的先验分布。若使用最简单的时序距离和窗口函数，即对于 $j < i$ 有 $d_{ij} = i - j$，对 $j > i$ 有 $d_{ij} = \infty$，$f(d) = 1[d <= a]$，a 是窗口大小，则根据生成过程可知，子块 i 的主题要么和前 a 个子块中的的一致，要么是一个新主题。这相当于为 z 增加了一个具有分段性质的先验。而 CRP 只能增加一个具有聚类性质却无分段性质的先验。见图4.2.

4.4.1 推断方法

对于建立的上述模型，本文用吉布斯采样的方法，根据每个 c_i 在其他变量上的条件概率依次采样。这里 G_0 选 Dirichlet 分布，是 ϕ_k 的共轭先验，从而每一步采样时可以将 ϕ_k 积分掉，利用 collapsed Gibbs 采样提高收敛速度:

$$p(c_i|c_{-i}, x_{1:N}, \theta, G_0) \propto p(c_i|\theta)p(x_{1:N}|z(c_{1:N}), G_0). \tag{4.7}$$

其中第一项是 dd-CRP 提供的先验，第二项是似然项。因为同一个餐桌上的观察值是独立同分布的，所以似然项是根据不同的餐桌因子化的，即:

$$p(x_{1:N}|z(c_{1:N}), G_0) = \prod_{k=1}^{|z(c)|} p(x_{z^k(c)}|G_0) \tag{4.8}$$

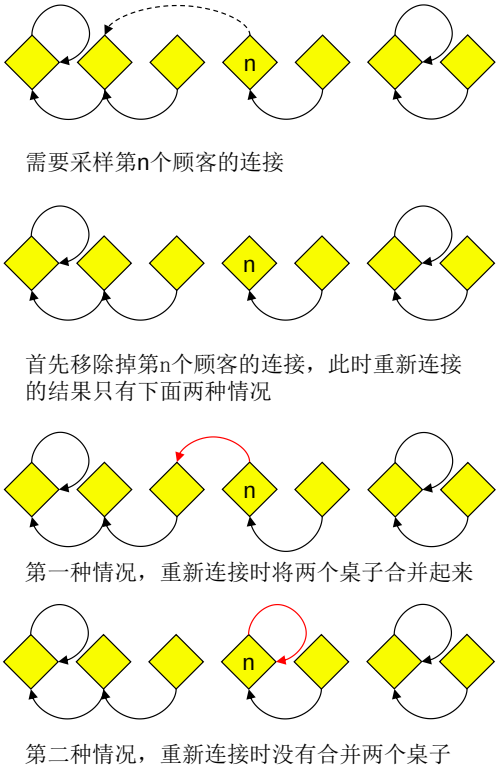

图 4.3: dd-CRP 单次 gibbs 采样时的情况

这里 $|z(c)|$ 是桌子的总个数，$z^k(c)$ 是第 k 个桌子上所有的顾客对应的索引集合。

在采样 c_i 时，先移除掉从顾客 i 指出的箭头。根据 c_i 的重新取值对最终划分的影响，这里只有两类可能的情况，要么新的 c_i 将某两个桌子合并到一起，要么对于划分没有影响 (图4.3)。

对于第一种情况，令 l 和 m 代表被合并的两个桌子的标号，则关于 z^l 和 z^m 的两个似然因子项被合并为一个关于的 $z^l(c) \cup z^m(c)$ 因子。而其他的因子不会改变。所以似然为：

$$p(x_{z^l(c)\cup z^m(c)}|G_0) \prod_{k\neq m,l} p(x_{z^k(c)}|G_0) \tag{4.9}$$

对于第二种情况，所有的因子都不会改变，从而得到：

$$\begin{aligned} &p(c_i|c_{-i}, x_{1:N}, \theta, G_0) \\ &\propto \begin{cases} p(c_i|\theta)\Delta(x,z,G_0) & c_i \text{ joins } l \text{ and } m \\ p(c_i|\theta) & \text{otherwise} \end{cases} \end{aligned} \tag{4.10}$$

其中

$$\Delta(x,z,G_0) = \frac{p(x_{z^l(c)\cup z^m(c)}|G_0)}{p(x_{z^l(c)}|G_0)p(x_{z^m(c)}|G_0)} \tag{4.11}$$

方法	F1	方法	F1
dd-CRP	0.7357	PLSA-DP-CE	0.6815
BayesSeg	0.7137	TextTiling	0.5341

表 4.1: 实验结果

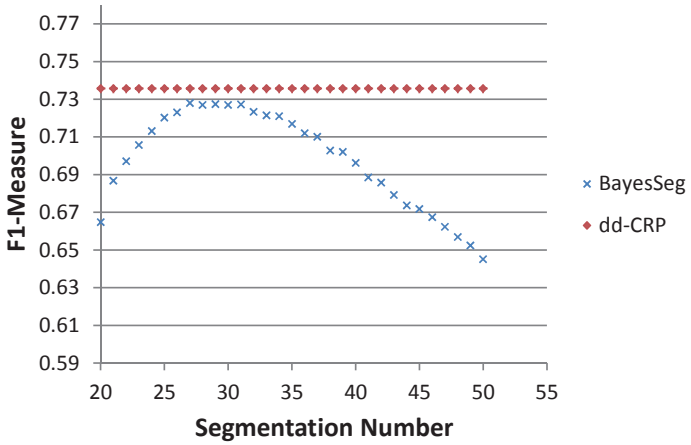

图 4.4: BayeSeg 受到分割个数的严重影响，如果设置不当，会导致结果下降很多，而 dd-CRP 则可以通过自动发现合适的分割个数来避免这一问题。

第 k 个因子项的似然为:

$$p(x_{z^k(c)}|G_0) = \int p(x_{z^k(c)}|\phi_k)p(\phi_k|G_0)d\phi_k \tag{4.12}$$

4.5 实验与分析

4.5.1 实验设置

本章的实验数据使用 TDT2 中英文广播新闻语料库[3]中的的 VOA 新闻，对本文提出的方法进行评估。语料来源于 1998 年 2 月至 7 月的美国之音英文广播新闻，共含 111 个新闻段，时长 111 小时。语料含有 LVCSR 识别文本及手工标注的过的真实边界。

数据预处理 对于所有的文本，使用 Porter 算法 [53] 对其进行处理，将单词进行 stem 化，并且去除掉停用词 (stop word)。然后将文本分割成不重合的固定长度的文本块。对于每个文本块，将其转化为词频表示，作为模型的观察特征。

评价方法 使用自然语言处理领域常用的 F1 测度作为实验的评价方法，其定义如下:

$$F1 = \frac{2 * rec * pre}{rec + pre} \tag{4.13}$$

[3]http://www.ldc.upenn.edu/Projects/TDT2

其中，pre 表示正确率 (precision), 即找到的正确边界数占找到的总边界数的比例。rec 表示召回率 (recall)，即找到的正确边界数占真实故事边界数的比例。其定义如下:

$$pre = \frac{N_c}{N_d}, rec = \frac{N_c}{N_g} \tag{4.14}$$

其中，N_c 是检测到的边界正确的个数，N_r 是检测到的所有边界，N_g 是真实的边界数，即手工标注的边界个数。根据 TDT2 国际评测标准，如果检测到的边界和真实边界相距在 15 秒内，即可认为是正确的边界。

参数设置 基于依赖于距离的中国餐馆过程的方法由 Dirichlet 过程的中心参数 α，基分布的参数 λ_0，以及窗口大小 a 三个参数控制。α 用于控制为顾客分配一个新桌子的概率。α 越大则分段的长度期望越小，分段的个数越多。λ_0 用于对词频的平滑，相当于语言模型中的加 n 平滑方法，λ_0 越大，平滑度越高，两个子块之间越难区分。窗口大小 a 用来作为依赖关系的距离阈值。$a = 1$ 时，每个子块只考虑和其前一个子块的相关性，当 $a = \infty$ 时，此时 dd-CRP 先验退化为序列 crp。对于参数 α 和 λ_0，可以为其分别添加一个无信息先验，然后利用一个类似于 EM 算法的过程来进行更新。在 E 步，固定参数的值，通过桌子的采样得到分段边界。在 M 步，固定分段边界，计算更新两个参数需要的统计量。对于 λ_0，利用最大后验方法进行更新。此时可以将模型看做是 crp 的结果，从而利用3.4.3小节中采样超参数的方法对 α 进行更新。

4.5.2 结果分析

实验结果参见表4.1。可以看出，本文提出的方法比几种基线系统方法结果有了明显的提高。相比于基线系统中最好的 PLSA-DP-CE，BayesSeg 和 dd-CRP 分别提高了 3.2% 和 5.4%。同样是概率模型，利用 dd-CRP 的非参数特性来释放固定分割数的约束，可以获得比 BayesSeg 更好的结果。如图4.4所示，BayesSeg 受到分割个数的影响很大，如果设置不当，结果会明显下降，dd-CRP 则没有这一问题，。

另外，基于先进行主题映射再分割的方法，如果训练集和测试集的来源不同，结果会下降很多。Lu 研究了非同源时的情况 [54]，PLSA-TT-CE 下降了 10% 以上，而基于概率统一建模的方法，并不存在这一问题，这也是本文方法的另一个优势。

5 基于非参数模型的类音素分割

本章主要研究类音素分割任务，这是语音研究中的一项重要任务。类音素分割和故事分割任务有一个重要的区别，即故事分割中的故事是独立不重复的，而音素单元是重复出现的。因此，本文针对类音素分割任务的这一特殊性，建立了不同于上一章的分割模型。

5.1 类音素分割任务分析

在故事分割任务中，使用了 ASR 的结果进行分割，但是分割效果与识别率有很大的关系，如果识别错误率较高，对后续的分割任务影响较大。如果考虑直接从音频特征寻找相似的语义单元，这样便能避免识别错误率的问题[1]。一般的做法是可将任务分为词级的任务和子词级的任务进行考虑。

词级的任务是寻找语音序列中重复出现的一些具有语义的模式，类似于识别结果中的单词。子词级任务可以看做是寻找最小的发音单元，作为构成更高层次的语义单元的基本单元。如果可以很好地完成子词级任务，那么对于词级的模式发现，可以在子词单元构成的数据流上利用动态时间规整 (dynamic time warp,DTW) 相关的算法进行确定性匹配，找到共现的模式。这样，最终的故事分割使用的特征可以是以这些模式为维度构成的频率向量，从而避免了语音识别错误率带来的问题。

这里的子词级任务，也称为声学单元分割，由于这里声学单元类似于语音学中定义的音素概念，也称为类音素分割。有些文献也称为音素分割，不过由于音素单元具有语言学的意义，可能某个音素是有多个发音构成的，或者某些音素在一般人发音时并不区分，而这里的子词级分割主要是服务于上层任务的，所以和语言学上的音素往往并不等价。

如果不考虑样本在时间上的相关性，可以直接用有限混合模型来进行聚类，或者用 Dirichlet 过程混合建立无限混合模型。当得到不同观察对应的聚类后，只要将连续属于同一类目的观察当做一个音素，就可以得到一组划分。然而，如果这样做，并没有充分利用到语音数据本身的特殊性，即语音数据有很强的时序性。一方面，同一个音素对应了连续多个帧，它们的特征是相似的。另一方面，音素之间的转换具有统计意义上的差别性。如果假设帧之间是条件独立的，

[1]虽然音频层面的相似单元避免了识别错误，但是也丢失语音识别中的文本语义，使得不同语义的序列被混淆为统一模式，目前基于这一思路的方法仍在研究中，

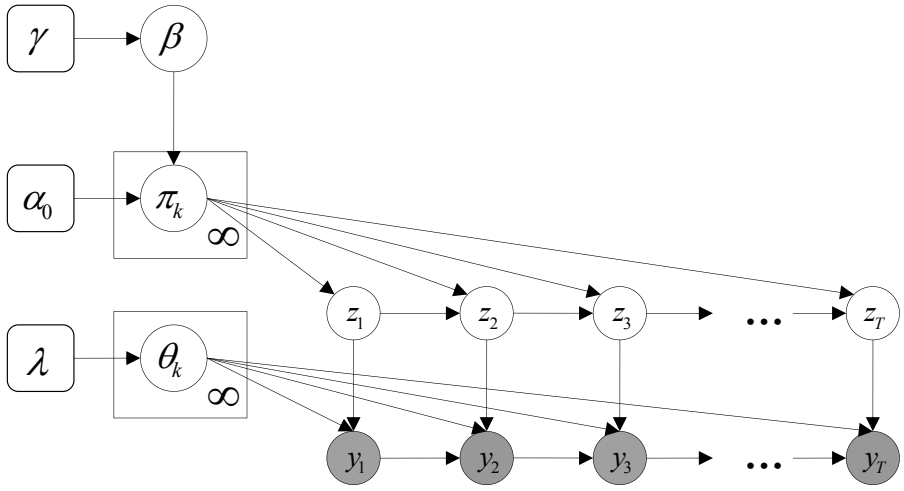

图 5.1: HDP-HMM 的有向图表示

那么就丢失了这两点先验知识。所以这里考虑用隐马尔科模型来建模帧间的时序关系，这样就可以建模音素之间的转换关系。对于音素个数未知问题，可以利用 DP 的性质，建立非参数模型以自适应的发现合适的个数，上述这一模型称为分层 Dirichlet 过程隐马尔科夫模型 (hierarchical Dirichlet process hidden Markov model,HDP-HMM) [2]。

不过，无限状态隐马尔科夫模型和传统的有限状态隐马尔科夫模型相比有一个重要的区别。由于可以有无限个状态，考虑下面这种极端情况：每一帧样本都对应一个不同的状态，即模型状态数和帧数一样大，而对于所有的 t，从 z_{t-1} 到 z_t 的概率都是 1，其他跳转都是 0，那么这一参数是参数空间中使得模型似然最大的，但这显然不是所期望的。这里对无限状态 HMM 加入一个状态持续变量进行约束，使得模型有一个状态自跳转的先验，从而解决这一问题。

5.2 无限状态隐马尔科夫模型

回顾2.2.5小节中介绍的隐马尔科夫模型，其中的状态个数是固定的，π_j 表示第 j 个状态的跳转概率，即 $z_t \sim \pi_{z_{t-1}}$。回顾在3.3节介绍的 HDP 模型，π_j 表示第 j 组（第 j 个餐馆）的参数，考虑一个含有 K 组数据 K 个共享成分（即 K 个餐馆 K 种菜品）且 $K \to \infty$ 的 HDP，如果把 HMM 中的 K 个状态看作是 HDP 中的 K 个不同的组，则可以直接将 HDP 中介绍的模型和推断方法用在 HMM 中，建立起一

[2]回顾之前的 DPMM 模型和 dd-CRP 模型，前者有较好的聚类性质，后者有较好的分割性质，而隐马尔科夫模型，即具有聚类的性质，又具有分割的性质。

个含有无限个状态的 HMM，参考式(3.23)，有：

$$\begin{aligned}
&\boldsymbol{\beta} \sim Gem(\gamma) \\
&\boldsymbol{\pi}_j \sim DP(\alpha_0, \boldsymbol{\beta}) \\
&\theta_k \sim H \\
&z_t \sim \boldsymbol{\pi}_{z_{t-1}} \\
&x_t \sim F(\theta_{z_t})
\end{aligned} \tag{5.1}$$

这一模型称之为基于 HDP 的无限状态 HMM(HDP-HMM)(图5.1)。

5.3 HDP-HMM 的推断方法

回顾之前介绍的直接赋值采样方法，将 $\boldsymbol{\pi}$ 和 $\boldsymbol{\theta}$ 积分掉，并引入一个辅助变量 $\boldsymbol{m}$，后来证明了 $\boldsymbol{m}$ 正是 CRF 中的桌子数. 和 HDP 一样，这里需要采样的变量仍然是 $\boldsymbol{z}$，$\boldsymbol{\beta}$ 和 $\boldsymbol{m}$，采样方法几乎完全一样，唯一的区别在对 $\boldsymbol{z}$ 的采样。

在 HDP 中采样，当观察到 β 时，z_{ji} 只跟 j 中的 z 相关，和其他的 z 是条件独立的。而此时，当观察到 β，虽然不同的 j 中的 z 被 β 阻塞了，但是由于 z 之间存在连接，所以情况发生了变化。此时当观察到 β,z_{t-1}，z_{t+1} 时，z_t 和其他的 z 是条件独立的，即 $p(z_t = k|\boldsymbol{z}^{-t}, \beta, \alpha_0) = p(z_t = k|z_{t-1}, z_{t+1}, \beta, \alpha_0)$。故其采样公式3.39

$$p(z_t = k|\boldsymbol{z}^{-t}, \beta, \alpha_0) \propto \begin{cases} (\alpha_0\beta_k + n^{-t}_{z_{t-1}k})(\frac{\alpha_0\beta_{z_{t+1}} + n^{-t}_{kz_{t+1}} + \delta(z_{t-1},k)\delta(z_{t+1},k)}{\alpha_0 + n^{-t}_{k} + \delta(z_{t-1},k)} & k \in 1, ..., K \\ \alpha_0\beta_k\beta_{z_{t+1}} & k = k \end{cases} \tag{5.2}$$

5.4 模型的改进

简单的将 HMM 用 HDP 先验扩展为一个无限状态模型存在着一个严重的问题。由于状态个数不受限制，用多个状态来表示单状态可能会有更大的似然，这是一种过拟合的情况。因为状态一般具有持续性，考虑为模型增加一个倾向于自跳转的先验知识进行规整，将式(5.1)中 π 的分布变为：

$$\boldsymbol{\pi}_j \sim DP(\alpha_0 + \kappa, \frac{\alpha_0\boldsymbol{\beta} + \kappa\delta_j}{\alpha_0 + \kappa}) \tag{5.3}$$

这里为自跳转的概率增加了一个大小为 κ 的量，得到的新模型称为 sticky-HDP-HMM[55]。其有向图表示见图5.2。

这里使用一种类似于连锁中国餐馆过程的表述来帮助理解这一模型。假设每

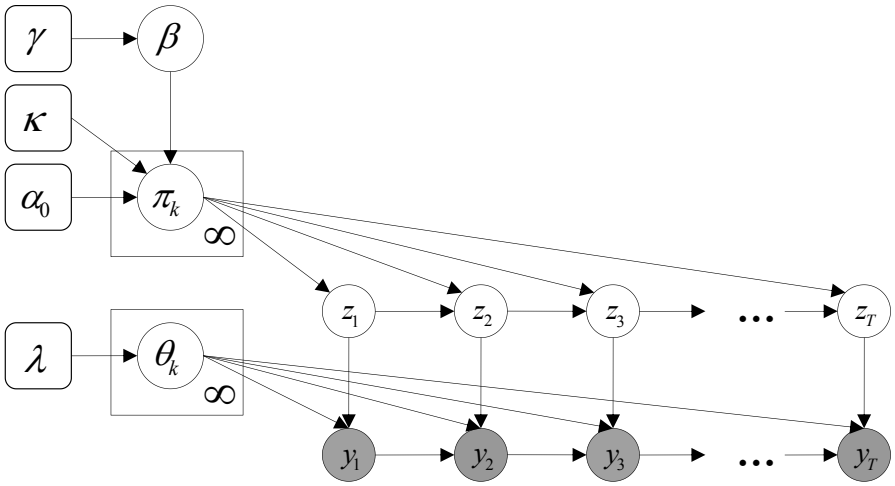

图 5.2: sticky-HDP-HMM 的有向图表示

个餐馆都有一个特色菜，第 j 个餐馆的特色菜就是第 j 道菜，注意第 j 道菜并不是 j 餐馆特有的，其他餐馆也有这道菜，但是味道要略差一些。另外，顾客的血缘关系会影响到他们的就餐习惯，若父母吃的是第 j 道菜，则孩子会去第 j 个餐馆 (第 j 道菜是其特色菜) 用餐，且更倾向于选择吃第 j 道菜。

将 z_t 看做父母，将 z_{t+1} 看做孩子。若 $z_t = j$, 则 z_{t+1} 会进入第 j 个餐厅，再根据 π_j 选择菜。

5.4.1 Sticky-HDP-HMM 的采样

参考式(3.47)，在 sticky-HDP-HMM 中，$\alpha_0\beta_k$ 变成了 $\alpha_0\beta_k + \kappa\delta_{(j,k)}$，从而对 $\boldsymbol{m}$ 的采样公式为:

$$p(m_{jk} = m|n_{jk}, \alpha_0, \boldsymbol{\beta}, \kappa) = s(n_{jk}, m)(\alpha_0\beta_k + \kappa\delta_{(j,k)})^m \frac{\Gamma(\alpha_0\beta_k + \kappa\delta_{(j,k)})}{\Gamma(\alpha_0\beta_k + \kappa\delta_{(j,k)} + n_{jk})} \quad (5.4)$$

考虑第 j 个餐馆第 t 个桌子上的菜，因为引入了 κ，其分布变为:

$$k_{jt} \sim \frac{\alpha_0\boldsymbol{\beta} + \kappa\delta_j}{\alpha_0 + \kappa} \quad (5.5)$$

辅助变量 m 实际上和 k 有关，而这里 k 的分布有了变化，所以得到的 $\boldsymbol{m}$ 并不能

直接用来采样 $\boldsymbol{\beta}$，这里引入辅助变量 $\bar{k}_{jt}$ 和 w_{jt} 来表示 k_{jt}:

$$\begin{aligned} &\bar{k}_{jt} \sim \boldsymbol{\beta} \\ &w_{jt} \sim Ber(\frac{\kappa}{\alpha_0 + \kappa}) \\ &k_{jt} \propto \begin{cases} \bar{k}_{jt} & w_{jt} = 0 \\ j & w_{jt} = 1 \end{cases} \end{aligned} \tag{5.6}$$

这样可以用从 $\bar{k}_{jt}$ 得到的 $\bar{\boldsymbol{m}}$ 来采样 β。而只要采样出 w_{jt}，就可以从 $\boldsymbol{m}$ 求得 $\bar{\boldsymbol{m}}$。

若 $k_{jt} \neq j$，则 $w_{jt} = 0$，所以只需要计算 $k_{jt} = j$ 的 w_{jt}。由式(5.6), 根据贝叶斯公式，得到 w_{jt} 的后验概率:

$$p(w_{jt}|k_{jt} = j, \beta, \kappa, \alpha_0) \propto \begin{cases} \beta_j(\frac{\alpha_0}{\alpha_0+\kappa}) & w_{jt} = 0 \\ \frac{\kappa}{\alpha_0+\kappa} & w_{jt} = 1 \end{cases} \tag{5.7}$$

这样，$\bar{\boldsymbol{m}}$ 为:

$$\bar{m}_{jk} = \begin{cases} m_{jk} & j \neq k \\ m_{jj} - \sum_{t=1}^{m_{jj}} w_{jt} & j = k \end{cases} \tag{5.8}$$

5.5 采样方法的改进

对于样本中两段时间上分隔的但是对应同一状态的序列，模型很可能先将他们分别对应到两个状态，然后随着学习过程逐渐将其混合成同一个状态，这一过程的速度称为混合率。由于 sticky 模型的自跳转偏置，使得两段同状态序列混合速率较慢。直接赋值采样按照坐标轴逐个更新，混合速率很慢，所以考虑利用模型的马尔科夫性，对状态序列进行 block 采样。仿照隐马尔科夫模型的前向后向算法，可以进行有效的 block 采样，但是为了利用这一方法，必须要确定 $\boldsymbol{\theta}$ 和 $\boldsymbol{\pi}$ 的值，即也需要对这两个变量进行采样，而不是像之前那样将其积分掉。

由于模型中的状态数不确定，有限 HMM 的前向后向算法不能适用，所以这里考虑对 GEM 的一种估计，用一个 L 个分量的 Dirichlet 分布来估计 DP, 称为 DP 的自由度 L 的弱极限估计:

$$GEM_L(\alpha) = Dir(\frac{\alpha_0}{L}, ..., \frac{\alpha_0}{L}) \tag{5.9}$$

其中 L 是一个大于期望状态数的值，这种估计形式，会学到一个状态数随数据变化，以 L 为上界的模型。

根据式(5.9)这一估计形式，有:

$$\begin{aligned}\beta &\sim Dir(\frac{\gamma}{L},...,\frac{\gamma}{L})\\ \pi_j &\sim Dir(\alpha_0\beta_1,...,\alpha_0\beta_j+\kappa,...,\alpha_0\beta_L)\end{aligned} \tag{5.10}$$

可以根据其后验分布进行采样:

$$\begin{aligned}\beta|\boldsymbol{z},\bar{\boldsymbol{m}} &\sim Dir(\frac{\gamma}{L}+\bar{m}_{\cdot 1},...,\frac{\gamma}{L}+\bar{m}_{\cdot L})\\ \pi_j|\boldsymbol{z},\beta &\sim Dir(\alpha_0\beta_1+n_{j1},...,\alpha_0\beta_j+\kappa+n_{jj},...,\alpha_0\beta_L+n_{jL})\end{aligned} \tag{5.11}$$

对于 θ 的采样也很简单, 只要根据先验和似然计算出后验即可:

$$\theta_j \sim p(\theta|\{y_t|z_t=j,\lambda\}) \tag{5.12}$$

这里对 $z_{1:T}$ 进行联合采样，根据模型的马尔科夫结构，可将 $z_{1:T}$ 的联合后验分布分解成因子相乘的形式:

$$\begin{aligned}p(z_{1:T}|y_{1:T},\boldsymbol{\pi},\boldsymbol{\theta}) = p(z_T|z_{T-1},y_{1:T},\boldsymbol{\pi},\boldsymbol{\theta})p(z_{T-1}|z_{T-2},y_{1:T},\boldsymbol{\pi},\boldsymbol{\theta})\\ ...p(z_2|z_1,y_{1:T},\boldsymbol{\pi},\boldsymbol{\theta})p(z_1|y_{1:T},\boldsymbol{\pi},\boldsymbol{\theta})\end{aligned} \tag{5.13}$$

这样进行分解后，可先从 $p(z_1|y_{1:T},\boldsymbol{\pi},\boldsymbol{\theta})$ 中采样出 z_1, 然后根据 z_1 的取值从 $p(z_2|z_1,y_{1:T},\boldsymbol{\pi},\boldsymbol{\beta},\boldsymbol{\theta})$ 中采样出 z_2, 以此类推，采样出整个 $z_{1:T}$。z_t 的条件分布可以写成:

$$p(z_t|y_{1:T},\boldsymbol{\pi},\boldsymbol{\theta}) \propto p(z_t)p(y_t|\theta_{z_t})\sum_{z_{t+1:T}}\prod_{t'=t+1}^{T}p(z_{t'}|\pi_{z_{t'-1}})p(y_{t'}|\theta_{z_{t'}}) \tag{5.14}$$

如果令 $m_{t,t-1}(z_{t-1}) = \sum_{z_{t:T}}\prod_{t'=t}^{T}p(z_{t'}|\pi_{z_{t'-1}})p(y_{t'}|\theta_{z_{t'}})$, 则有:

$$p(z_t|y_{1:T},\boldsymbol{\pi},\boldsymbol{\theta}) \propto p(z_t)p(y_t|\theta_{z_t})m_{t+1,t}(z_t) \tag{5.15}$$

根据 $m_{t,t-1}(z_{t-1})$ 的定义，可以得到一个 $m_{t,t-1}(z_{t-1})$ 和 $m_{t+1,t}(z_t)$ 间的递推关

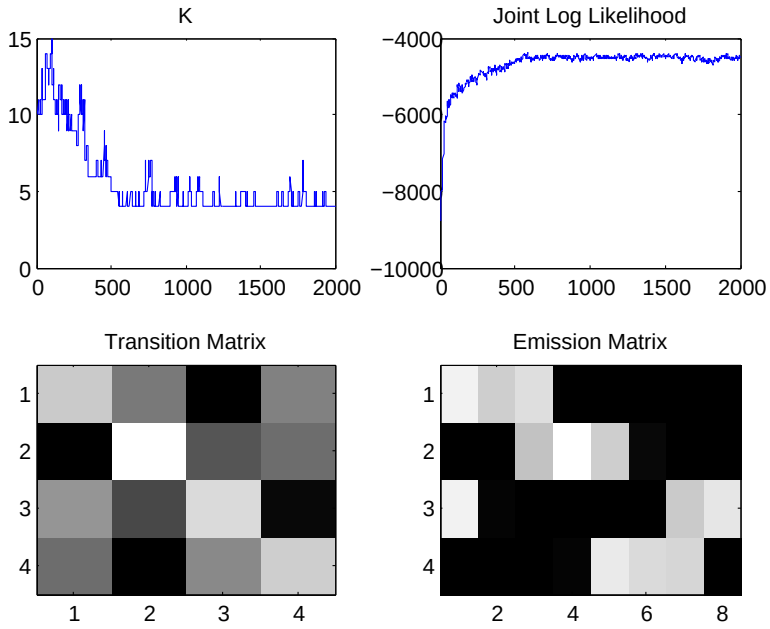

图 5.3: 2000 个样本点的人工数据实验结果

系式:

$$\begin{aligned} m_{t,t-1}(z_{t-1}) &= \sum_{z_{t:T}} \prod_{t'=t}^{T} p(z_{t'}|\pi_{z_{t'-1}}) p(y_{t'}|\theta_{z_{t'}}) \\ &= \sum_{z_t} p(z_t|\pi_{z_{t-1}}) p(y_t|\theta_{z_t}) \sum_{z_{t+1:T}} \prod_{t'=t+1}^{T} p(z_{t'}|\pi_{z_{t'-1}}) p(y_{t'}|\theta_{z_{t'}}) \\ &= \sum_{z_t} p(z_t|\pi_{z_{t-1}}) p(y_t|\theta_{z_t}) m_{t+1,t}(z_t) \end{aligned} \tag{5.16}$$

其中 $m_{T+1,T}(z_T) = 1$，这样，只需要从马尔科夫链的尾部向头部递推的计算出 $m_{t,t-1}(z_{t-1})$, 然后在从头部向尾部依次采样各个 z_t。这里称 $m_{t,t-1}(z_{t-1})$ 为从 z_t 到 z_{t-1} 传播的后向消息。很容易证明，该方法是置信传播算法的一个特例。

5.6 实验结果与分析

本文首先在人工生成的数据上验证算法的有效性，然后将其应用到类音素分割任务中。

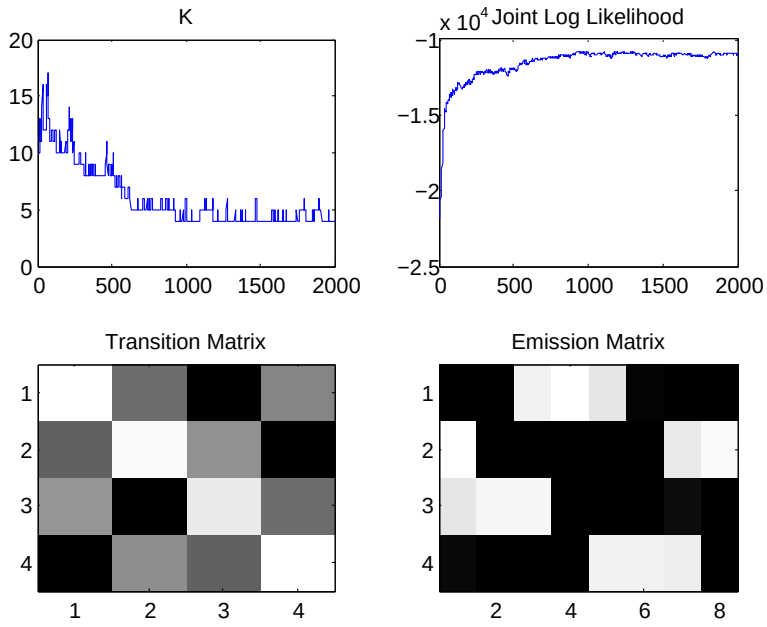

图 5.4: 5000 个样本点的人工数据实验结果。可以看出，由于数据量增多，统计意义更加充分，此时的结果和真实的参数更加接近。

5.6.1 仿真实验

仿真实验中，按照如下的模型生成含有 4 个隐状态、发射概率是 8 维的离散分布的序列数据:

$$A = \begin{pmatrix} 0.5 & 0.3 & 0.2 & 0.0 \\ 0.0 & 0.5 & 0.3 & 0.2 \\ 0.2 & 0.0 & 0.5 & 0.3 \\ 0.3 & 0.2 & 0.0 & 0.5 \end{pmatrix} E = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \\ \frac{1}{3} & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} \end{pmatrix} \tag{5.17}$$

其中 A 是状态跳转概率矩阵，E 中第 i 行对应第 i 个状态的离散发射概率。

利用 HDP-HMM 对数据进行建模和推断，对于 500 个样本点的数据，迭代 2000 次，结果如图5.3。可以看出，随着算法的迭代进行，模型学习到的主题个数 (K) 逐渐接近真实个数，最终可以收敛到真实的主题个数。在状态转移概率和发射概率的结果图中，白色代表了 0，黑色代表了 1，其他颜色根据灰度值代表了 0 到 1 之间的值。可以看出，模型可以有效地推断出这两组参数，不过和真实的参数却有一些偏差。这是因为数据太少的缘故，4 状态对应了 16 种跳转情况，只生成 500 的采样点肯定是不够的。如果生成更多的数据，进行实验，可以得到更好的结果。如图5.4，可见由于数据量增多，统计量更加充分，此时灰色块的颜色更加接近。

方法	正确率	召回率	F1
Dusan[30]	0.668	0.752	0.708
Qiao[31]	0.769	0.775	0.769
HDP-HMM	0.718	0.824	0.767

表 5.1: 分割实验结果

5.6.2 类音素分割实验结果

实验使用 TIMIT[56] 数据库中的语料，该数据库包含多个说话人的语音信息，每个说话人录制了多句话。语料文件命名为 people-sentence.type，其中，people 是不同说话人的标识，sentence 是不同语句的标识[3],type 代表相应的文件类型，包括音频源文件，音素抄本，语句抄本等。本实验选用其中 500 句话作为实验数据。

该语料的采样频率为 16K，即每秒采样 16000 个点。帧长 25ms，帧移 10ms，帧间重叠为 15ms。对于每一帧数据，本实验提取 13 维的 MFCC 特征作为样本点。

分割分析 实验对本文中的方法与其他几种最新的方法进行比较，结果见表5.6.2。其中第一行是 Dusan 提出的一种无监督分割方法 [30]，这一方法也没有假定分段的个数。第二行是 Qiao 提出的一种半监督的方法 [31]。相比于第一行的无监督方法，本文方法的结果在各项指标上均有很大的提高，精确度、召回率和 F1 值分别提高了 5%、7.2% 和 5.9%。作为无监督的方法，HDP-HMM 和半监督方法 [31] 的 F1 值非常接近. 另外，本文的方法具有较好的召回率，这说明本文的方法在真实边界位置匹配的更好。

聚类分析 HDP-HMM 模型在建模时就考虑到了数据的聚类特征，因此，除了对单元分割的结果进行评估，本文也对其聚类的效果进行分析，如图5.5，展示了语料 FKAA0-SX298 上的实验结果，图中底端是分割聚类的结果，中间是音频原始信号，上端是真实的音素标注。上端的白线表示真实音素边界，相同颜色块对应同一个的音素，下端的白线表示结果中的边界，相同颜色块对应同一个的聚类。相同的聚类由于许多音素的发音十分类似，如 ix，iy 和 ih，er 和 axr，无论是人耳的听音，还是从音频信号上观察都非常相近，算法将这些相似的音素合并成一类，所以聚类的个数会比真实音素个数少一些。本文分析了多组语料上的结果（图5.6展示了语料 FCJF0-SA1 上的结果），直观的研究聚类状态和真实音素之间的关系，表5.6.2给出了一些相关性明显的聚类，可以看出，和 i(音一) 相近的发音 ix，iy，ih 都被聚成到 64 类，s，z 等音被聚到 77 类，en，n，ng 等鼻音被聚到 66 类。几乎所有的语料，h#（静音段）都被聚为 22 类，而一些清辅音如 k，f，q 也被分到 22 类，这是因为这些清辅音能量较小，不容易区分开。不过这种更加粗

[3]其中所有说话人都录制了 SA1 和 SA2 的语句。

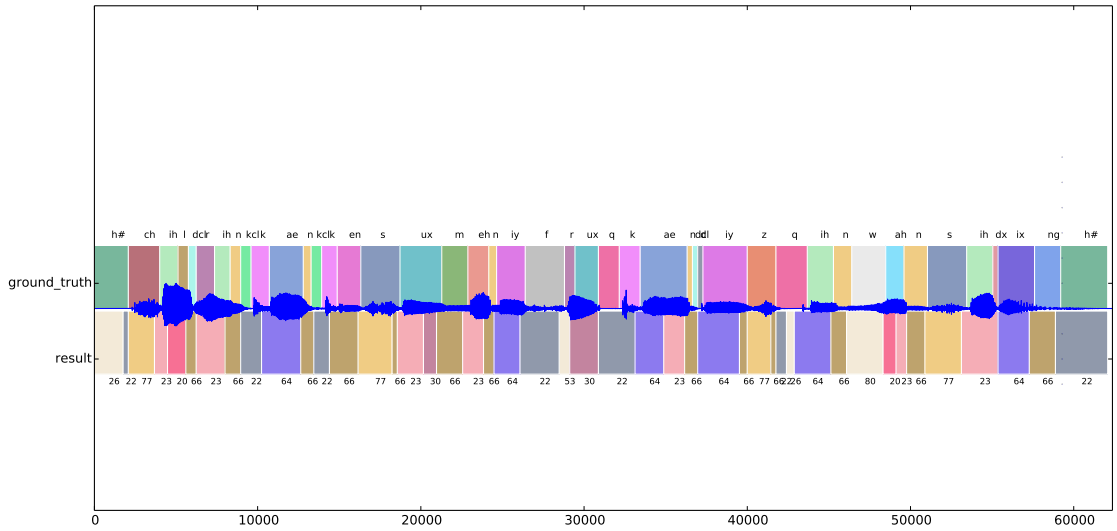

图 5.5: 语料 FKAA0-SX398 上的实验结果，其中横坐标以采样点为单位。这段话的抄本为“Children can consume many fruit candies in one sitting”。语料的采样频率为 16K，包含 62362 个采样点。由于 mfcc 特征文件丢弃了最后的一部分静音段，所以图中的标注和实验结果序列均比语音序列短一些。

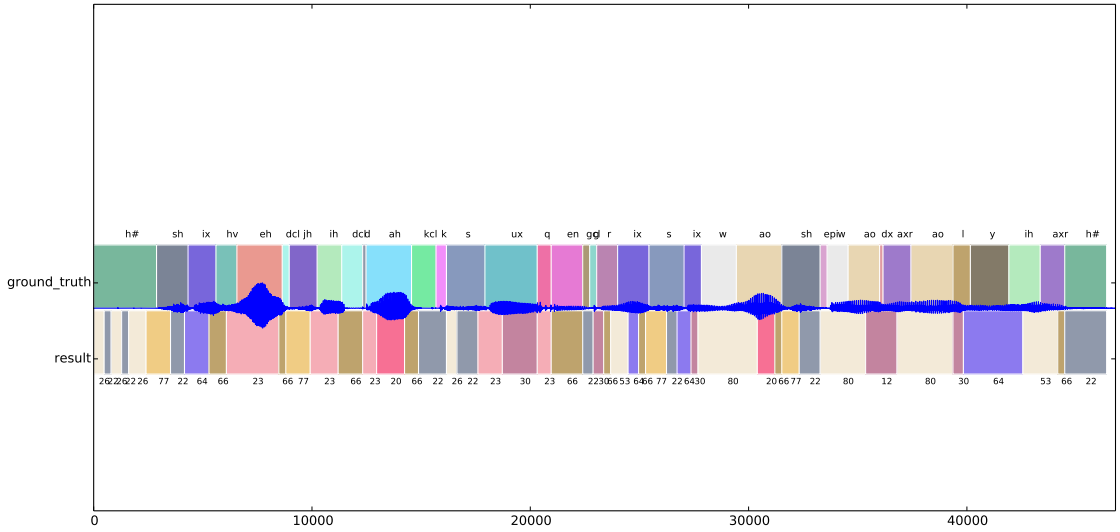

图 5.6: 语料 FCJF0-SA1 上的实验结果。这段话的抄本为“She had your dark suit in greasy wash water all year”。.

聚类类目	对应的音素
64	iy,ih,ix
77	s,z
66	en,n,ng
22	h#,k,f,q
80	ao,aa

表 5.2: 实验的一些聚类结果，可以看出，HDP-HMM 找到的类目相较于真实音素个数少一些, 不过却具有很明显的物理意义，即同一族的音素被聚到了一类

粒度的聚类对边界的寻找的影响并不严重，因为大部分的边界位于元音和辅音的交界处。从图5.5中可以看出，大部分的音素边界都被找到。

6 总结与展望

本文对狄利克雷过程相关的方法进行了详细的研究，包括将其应用在经典的概率模型如混合模型，超混合模型和时序模型上的方法，以及相关的基于采样的推断算法。通过对经典的参数模型加入狄利克雷过程先验，可以有效地解决人工模型选择的繁杂任务。

本文主要研究了狄利克雷过程在时序模型上的具体应用任务，利用依赖于距离的中国餐馆过程对故事分割任务进行建模，并与多个主流的基线系统比较，实验表明其具有非常好的结果。而另一种对时序数据进行建模的思路是建立马尔科夫模型，本文通过引入分层狄利克雷过程先验，对隐马尔科夫模型进行非参数化，并将其应用在类音素发现的任务上，得到了较好的实验结果。

文中主要使用的是基于 gibbs 采样的方法，这对于大数据问题会存在一些效率问题。面对互联网上海量的数据，需要进一步研究更加有效的算法。目前，关于狄利克雷过程相关的一些变分方法的研究非常重要，也是相关学术研究中的热点之一，所以下一步考虑利用变分方法来加快模型的求解速度。另一方面，对于许多任务，数据是在不断更新增加的，所以在线算法也显得非常重要，这也是下一步需要研究的重点。

对于语料中的多篇文档，实际上本身就有一个边界信息，但是本文在建模时并没有利用上。在故事分割任务中，由于没有去建模出每个段落之间共享特征这个性质，所以对每个新闻语料单独处理是合理的。但是，对于音素发现的任务，每篇文档（在这个语料里即每一句话）之间其实是条件独立的，简单的将所有文档拼接成一个连续的序列，就损失了一些天然的性质。为了建模这一条件独立性质，可以利用贝塔过程 [57] 相关的理论进行建模，这也是后续研究的一个重要内容。

参考文献

[1] Koller D, Friedman N. Probabilistic graphical models: principles and techniques [M]. MIT press, 2009.

[2] Jordan M I. An introduction to probabilistic graphical models. 2003.

[3] 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.

[4] Kohavi R, et al. A study of cross-validation and bootstrap for accuracy estimation and model selection [C]. In IJCAI. 1995: 1137–1145.

[5] Burnham K P, Anderson D R. Model selection and multimodel inference: a practical information-theoretic approach [M]. Springer, 2002.

[6] Hastie T, Tibshirani R, Friedman J, et al. The elements of statistical learning [M]. Springer, 2009.

[7] Rasmussen C E. Gaussian processes for machine learning [J]. 2006.

[8] Teh Y W, Jordan M I, Beal M J, et al. Hierarchical dirichlet processes [J]. Journal of the American Statistical Association. 2006, 101 (476): 1566–1581.

[9] Teh Y W. Dirichlet process [M] // Teh Y W. Encyclopedia of machine learning. Springer, 2010: 2010: 280–287.

[10] Gershman S J, Blei D M. A tutorial on Bayesian nonparametric models [J]. Journal of Mathematical Psychology. 2012, 56 (1): 1–12.

[11] Hauptmann A G, Witbrock M J. Story segmentation and detection of commercials in broadcast news video [C]. In Research and Technology Advances in Digital Libraries, 1998. ADL 98. Proceedings. IEEE International Forum on. 1998: 168–179.

[12] Allan J. Topic Detection and Tracking: Event-Based Information Organization [J]. Springer. 2002, 12.

[13] Hearst M A. TextTiling: Segmenting text into multi-paragraph subtopic passages [J]. Computational linguistics. 1997, 23 (1): 33–64.

[14] Banerjee, Satanjeev, I R A. A TextTiling based approach to topic boundary detection in meetings [C]. In Proc. Interspeech.ISCA. 2006.

[15] Wang X, Xie L, Ma B, et al. Phoneme Lattice based TextTiling towards Multilingual Story Segmentation [C]. In Proc. Interspeech. 2010: 1305–1308.

[16] Malioutov I, Barzilay R. Minimum cut model for spoken lecture segmentation [C]. In Proc. ACL. 2006: 25–32.

[17] Hofmann T. Probabilistic latent semantic indexing [C]. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. 1999: 50–57.

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

[19] Choi F Y Y, Wiemer-Hastings P, Moore J. Latent semantic analysis for text segmentation [C]. In Proc. EMNLP. 2001: 109–117.

[20] Hall D, Jurafsky D, Manning C D. Studying the history of ideas using topic models [C]. In Proc. EMNLP. 2008: 363–371.

[21] Chien J-T, Chueh C-H. Topic-Based Hierarchical Segmentation [J]. IEEE Transactions on Audio, Speech, and Language Processing. 2012, 20 (1): 55–66.

[22] Lu M, Leung C-C, Xie L, et al. Probabilistic latent semantic analysis for broadcast news story segmentation [C]. In Proc. Interspeech. 2011: 1301–1304.

[23] Lu X, Leung C-C, Xie L, et al. Broadcast news story segmentation using latent topics on data manifold [C]. In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. 2013: 8465–8469.

[24] Xie L, Zheng L, Liu Z, et al. Laplacian eigenmaps for automatic story segmentation of broadcast news [J]. Audio, Speech, and Language Processing, IEEE Transactions on. 2012, 20 (1): 276–289.

[25] Bengio Y. Learning deep architectures for AI [J]. Foundations and trends® in Machine Learning. 2009, 2 (1): 1–127.

[26] Park A, Glass J R. Towards unsupervised pattern discovery in speech [C]. In Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on. 2005: 53–58.

[27] Glass J. Towards unsupervised speech processing [C]. In Information Science, Signal Processing and their Applications (ISSPA), 2012 11th International Conference on. 2012: 1–4.

[28] Aversano G, Esposito A, Esposito A, et al. A new text-independent method for phoneme segmentation [C]. In Proc the 44th IEEE Midwest Symposium on Circuits and Systems. 2001: 516–519.

[29] Estevan Y P, Wan V, Scharenborg O. Finding maximum margin segments in speech [C]. In Proc. ICASSP. 2007.

[30] Dusan S, Rabiner L R. On the relation between maximum spectral transition positions and phone boundaries. [C]. In INTERSPEECH. 2006.

[31] Qiao Y, Shimomura N, Minematsu N. Unsupervised optimal phoneme segmentation: objectives, algorithm and comparisons [C]. In Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on. 2008: 3989–3992.

[32] Scharenborg O, Wan V, Ernestus M. Unsupervised speech segmentation: An analysis of the hypothesized phone boundaries [J]. The Journal of the Acoustical Society of America. 2010, 127 (2): 1084–1095.

[33] Torbati A H H N, Picone J, Sobel M. Speech Acoustic Unit Segmentation Using Hierarchical Dirichlet Processes [J]. 2013.

[34] Lee C-y, Glass J. A nonparametric Bayesian approach to acoustic model discovery [C]. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. 2012: 40–49.

[35] Bishop C M, et al. Pattern recognition and machine learning [M]. springer New York, 2006.

[36] Heinrich G. Parameter estimation for text analysis [R]. 2005.

[37] Murphy K P. Conjugate Bayesian analysis of the Gaussian distribution [J]. def. 2007, 1: 16.

[38] Shafer G R, Shenoy P P. Spatial interaction and the statistical analysis of lattice systems [J]. Journal of the Royal Statistical Society, Series B. 1974, 36: 192–223.

[39] Clifford P. Markov random fields in statistics [M]. Oxford University Press, 1990.

[40] Shafer G R, Shenoy P P. Probability propagation [J]. Annals of Mathematics and Artificial Intelligence. 1990, 2: 327–351.

[41] Rabiner L. A tutorial on hidden Markov models and selected applications in speech recognition [J]. Proceedings of the IEEE. 1989, 77 (2): 257–286.

[42] Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference [M]. Morgan Kaufmann, 1988.

[43] De Finetti B. Funzione caratteristica di un fenomeno aleatorio [J]. 1931.

[44] Gilks W R. Markov chain monte carlo [M]. Wiley Online Library, 2005.

[45] Chib S, Greenberg E. Understanding the metropolis-hastings algorithm [J]. The American Statistician. 1995, 49 (4): 327–335.

[46] Ferguson T S. A Bayesian analysis of some nonparametric problems [J]. Annals of Statistics. 1973, 1 (2): 209–230.

[47] Antoniak C E, et al. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems [J]. The annals of statistics. 1974, 2 (6): 1152–1174.

[48] Neal R M. Markov chain sampling methods for Dirichlet processes mixture models [J]. Computational linguistics. 2000, 9 (2): 249–265.

[49] Heinrich G. Infinite LDA [J]. Implementing the HDP with Minimum Code complexity. Technical report. 2011.

[50] West M. Hyperparameter estimation in Dirichlet process mixture models [M]. Duke University, 1992.

[51] Escobar M D, West M. Bayesian density estimation and inference using mixtures [J]. Journal of the american statistical association. 1995, 90 (430): 577–588.

[52] Rasmussen C E. The infinite Gaussian mixture model. [C]. In NIPS. 1999: 554–560.

[53] Porter M. The Porter Stemming Algorithm [J]. 2009.

[54] Lu M. 基于潜在语义和主题建模的广播新闻 [J]. 2012.

[55] Fox E B, Sudderth E B, Jordan M I, et al. An HDP-HMM for systems with state persistence [C]. In Proceedings of the 25th international conference on Machine learning. 2008: 312–319.

[56] Garofolo J S, Consortium L D, et al. TIMIT: acoustic-phonetic continuous speech corpus [M]. Linguistic Data Consortium, 1993.

[57] Fox E B, Sudderth E B, Jordan M I, et al. Sharing Features among Dynamical Systems with Beta Processes. [C]. In NIPS. 2009: 549–557.

科研成果发表

论文发表

1.Chao Yang, Lei Xie and Xiangzeng Zhou,"Unsupervised broadcast news story segmentation using distance dependent Chinese restaurant process",IEEE International Conference on Acoustics, Speech, and Signal Processing,FLORENCE, ITALY,2014(EI 索引，语音研究顶级国际会议)

科研工作:

1. 国家自然科学基金面上项目《基于 DBN 协同建模的中文及跨语种语音结构事件检测研究》（61175018）

致谢

随着毕业论文的尘埃落定，我的校园生活也接近尾声。借此机会，对所有帮助过我的师长、亲人、同学、朋友表达我真挚的感谢。

首先，我要衷心的感谢我的导师谢磊教授。他有着严谨的治学理念和精益求精的态度，将大部分时间都奉献给了科研事业和对学生的指导。正是在他一步步耐心引导下，我才接触到这一有趣且富有挑战性的研究领域并逐渐掌握了正确的研究思路和方法。本论文的完成过程是极为不易的，在论文选题，文献阅读，理论研究，实验分析，论文撰写每个阶段，谢老师都在不遗余力的指导和鼓励着我，可以说这篇论文里处处都有他的心血。在此，谨向谢老师表达我我最衷心的感谢，您的教诲我会永远铭记。

感谢蒋冬梅教授和付中华教授。两位老师的踏实的研究态度、渊博的学科知识，一直深深地感染着我并将使我受益终生。另外，两位老师的幽默风趣和乐观豁达也对我影响颇深。

感谢我的师兄师姐: 田霄海，赵文淮，卢咪咪，郑李磊，史倩，吴杰，周详增，赵亚丽，孙乃才，李冰锋。他们毫无保留的将自己的学习经验和方法与我分享，使我在研究领域得到了更快的进步。尤其要感谢周详增师兄，他仔细阅读了我的论文并提出了宝贵的修改意见。

感谢教研室的同学：牛建伟，路晓明，高新远，原帅，李龙飞，陈爱华，唐玲，刘洋。这两年多来大家相互鼓励，相互帮助，一起经历了科研和生活中的辛酸与甘甜，是我人生中一段美好的回忆。感谢许军海同学，协助我完成了本文部分实验内容。感谢于佳和陈宏杰两位同学，他们为我的论文修改提出了宝贵的意见。

感谢秦巧玲同学一直以来对我的关心、照顾和理解。

最后，要感谢我的父母这些年对我含辛茹苦的养育和教导，他们一直默默地为我奉献着最无私的爱，我也愿早日报答这一切。

西北工业大学业

学位论文知识产权声明书

本人完全了解学校有关保护知识产权的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属于西北工业大学。学校有权保留并向国家有关部门或机构送交论文的复印件和电子版。本人允许论文被查阅和借阅。学校可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。同时本人保证，毕业后结合学位论文研究课题再撰写的文章一律注明作者单位为西北工业大学。

保密论文待解密后适用本声明。

学位论文作者签名：_____　　指导教师签名：_____

年　月　日　　年　月　日

西北工业大学

学位论文原创性声明

秉承学校严谨的学风和优良的科学道德，本人郑重声明：所呈交的学位论文，是本人在导师的指导下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和致谢的地方外，本论文不包含任何其他个人或集体已经公开发表或撰写过的研究成果，不包含本人或其他已申请学位或其他用途使用过的成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式表明。

本人学位论文与资料若有不实，愿意承担一切相关的法律责任。

学位论文作者签名：_____

年　月　日