**读书报告**

61518424 王贵涛

1. **问题与解答**

我提出问题：

1. 狄利克雷分布的参数a代表什么，先验分布的a怎么选取？

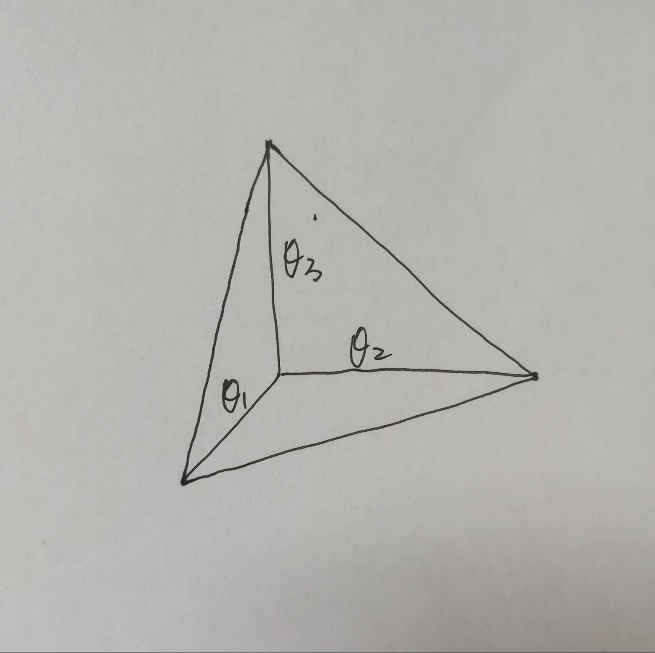
讨论结果：这里a是超参数，事先给定，描述了狄利克雷分布。a可以理解为引入的一种不存在的假设计数，类似于在没有实验数据的情况下的一种假设，其选择很多都是根据实际应用的经验来确定的。

别人提出的问题：

1. 怎么理解'狄利克雷分布θ存在于(k-1)维单纯形上'？

我的解答：满足∑θ=1，θ≥0的点就是一个单纯形，相当于在一个k维坐标系上加个限制条件，实际上降了一维。





1. LDA和PLSA有什么区别和联系？

我的解答：参考书上P391。可以认为LDA是PLSA (概率潜在语义分析)的扩展，相同点是两者都假设话题是单词的多项分布，文本是话题的多项分布。不同点是LDA使用狄利克雷分布作为先验分布，而PLSA不使用先验分布（或者说假设先验分布是均匀分布），两者对文本生成过程有不同假设；学习过程LDA基于贝叶斯学习，而PLSA基于极大似然估计。LDA的优点是，使用先验概率分布，可以防止学习过程中产生的过拟合。

1. **下周计划安排**

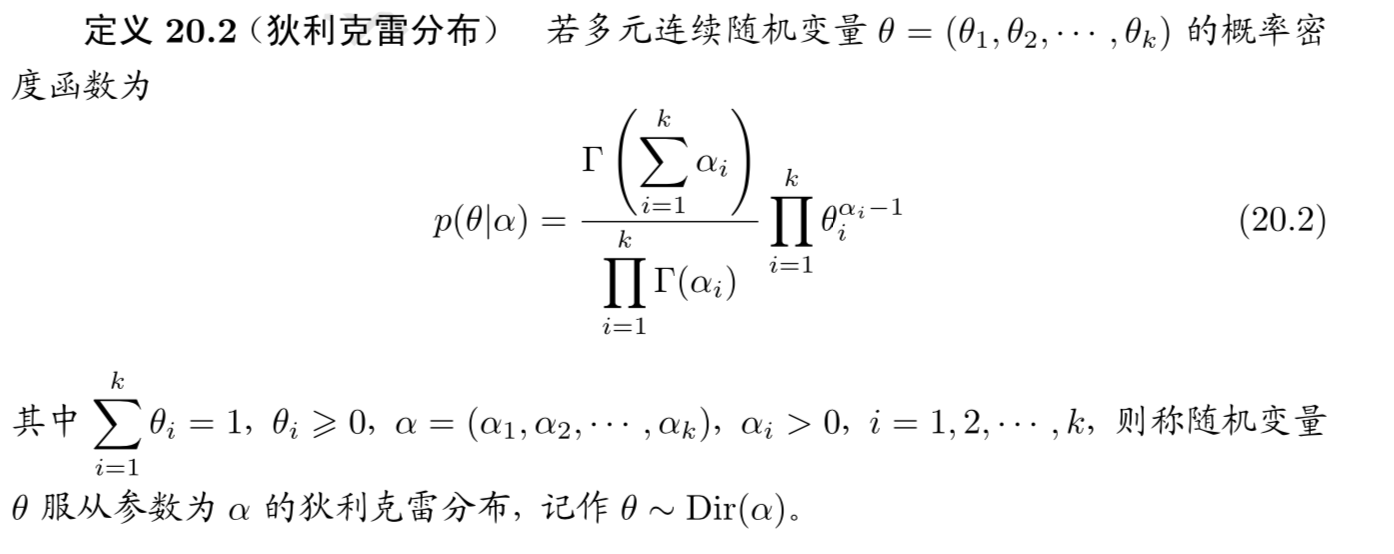
看完第二十一章并负责一部分的讲解。

1. **读书收获**

潜在狄利克雷分配(latent Dirichlet allocation, LDA)，作为基于贝叶斯学习的话题模型，是潜在语义分析、概率潜在语义分析的扩展，LDA 在文本数据挖掘、图像处理、生物信息处理等领域被广泛使用。

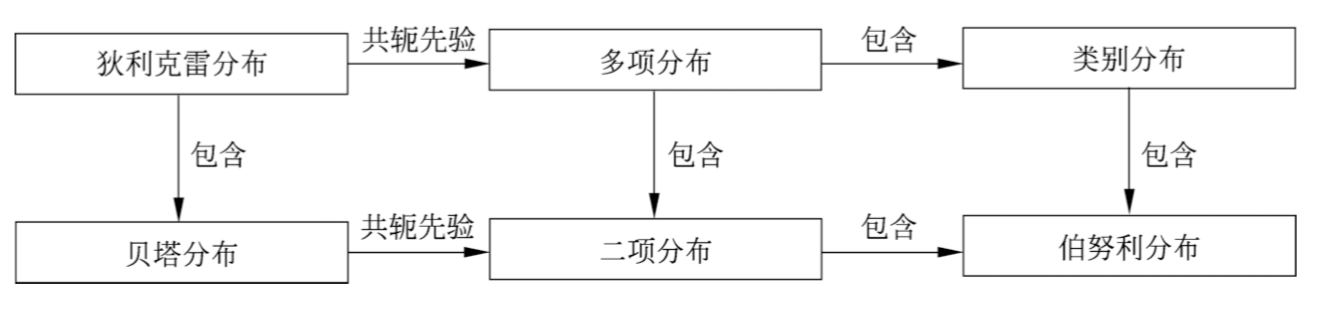
LDA模型是文本集合的生成概率模型，假设每个文本由话题的一个多项分布表示，每个话题由单词的一个多项分布表示。特别假设文本的话题分布的先验分布是狄利克雷分布，话题的单词分布的先验分布也是狄利克雷分布。先验分布的导入使LDA 能够更好地应对话题模型学习中的过拟合现象。

**狄利克雷分布**

****

狄利克雷分布θ存在于k-1维单纯形上。

各种概率分布之间的关系如下：



狄利克雷分布属于指数分布族。狄利克雷分布是多项分布的共扼先验。

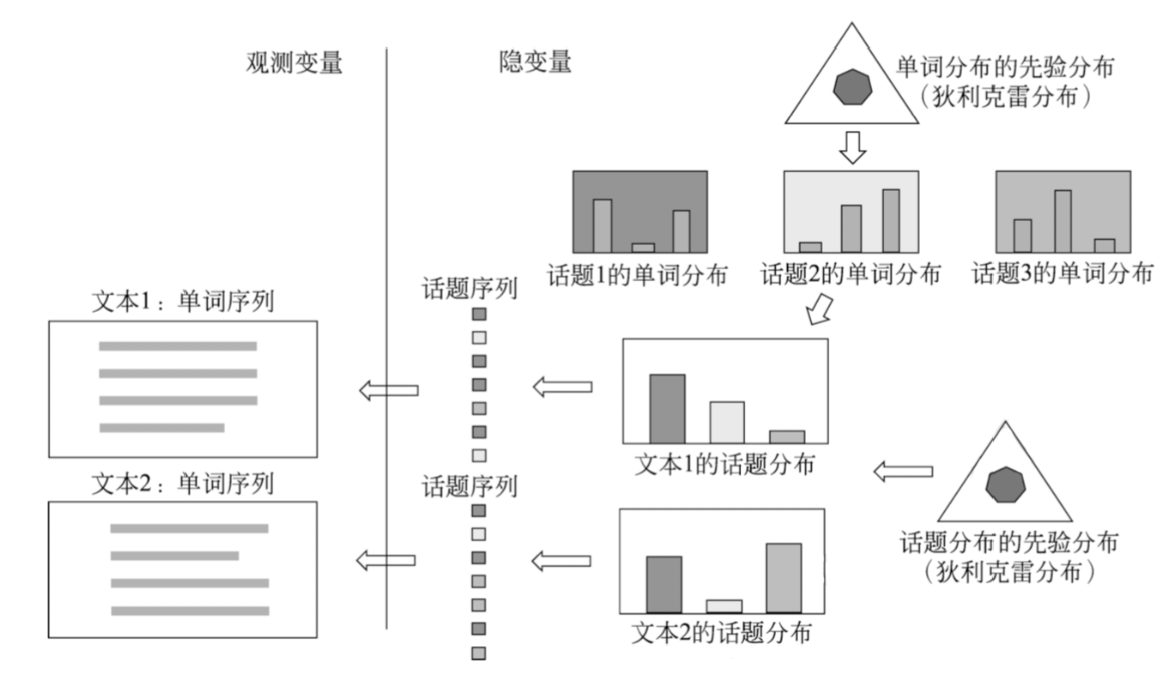
**潜在狄利克雷分配模型**

潜在狄利克雷分配（LDA）是文本集合的生成概率模型。

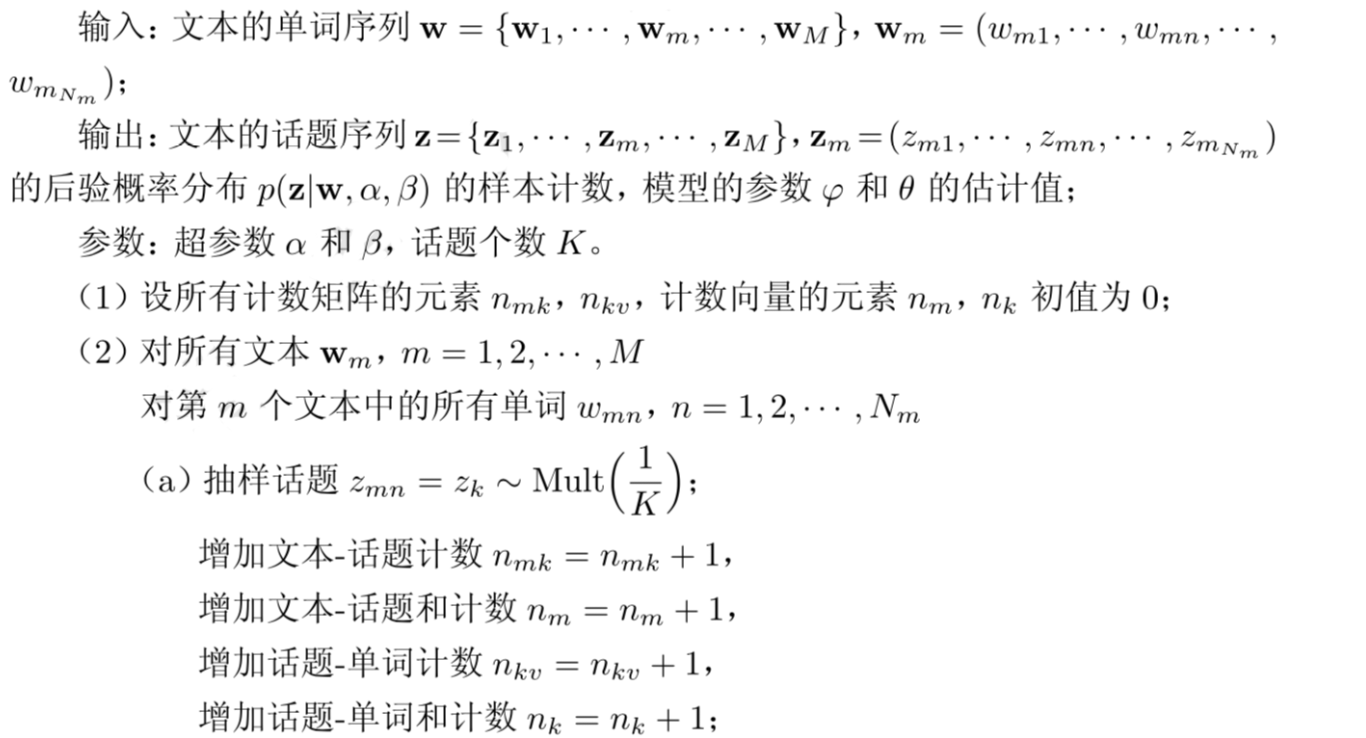
模型假设话题由单词的多项分布表示，文本由话题的多项分布表示，单词分布和话题分布的先验分布都是狄利克雷分布。文本内容的不同是由于它们的话题分布不同。

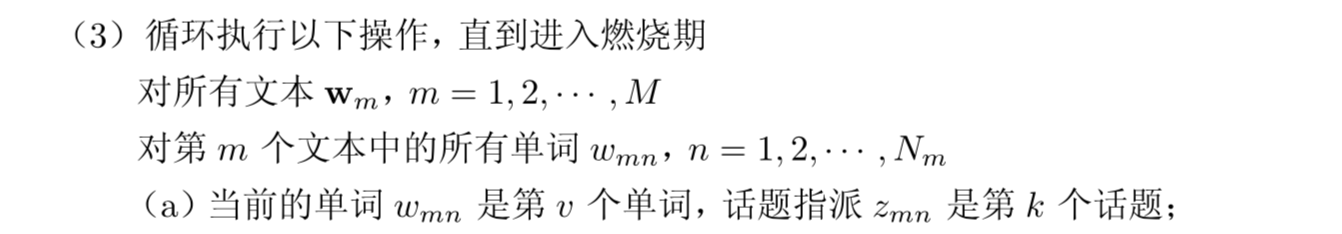
首先，基于单词分布的先验分布（狄利克雷分布）生成多个单词分布，即决定多个话题内容；之后，基于话题分布的先验分布（狄利克雷分布）生成多个话题分布，即决定多个文本内容；然后，基于每一个话题分布生成话题序列，针对每一个话题，基于话题的单词分布生成单词，整体构成一个单词序列，即生成文本；重复这个过程生成所有文本。

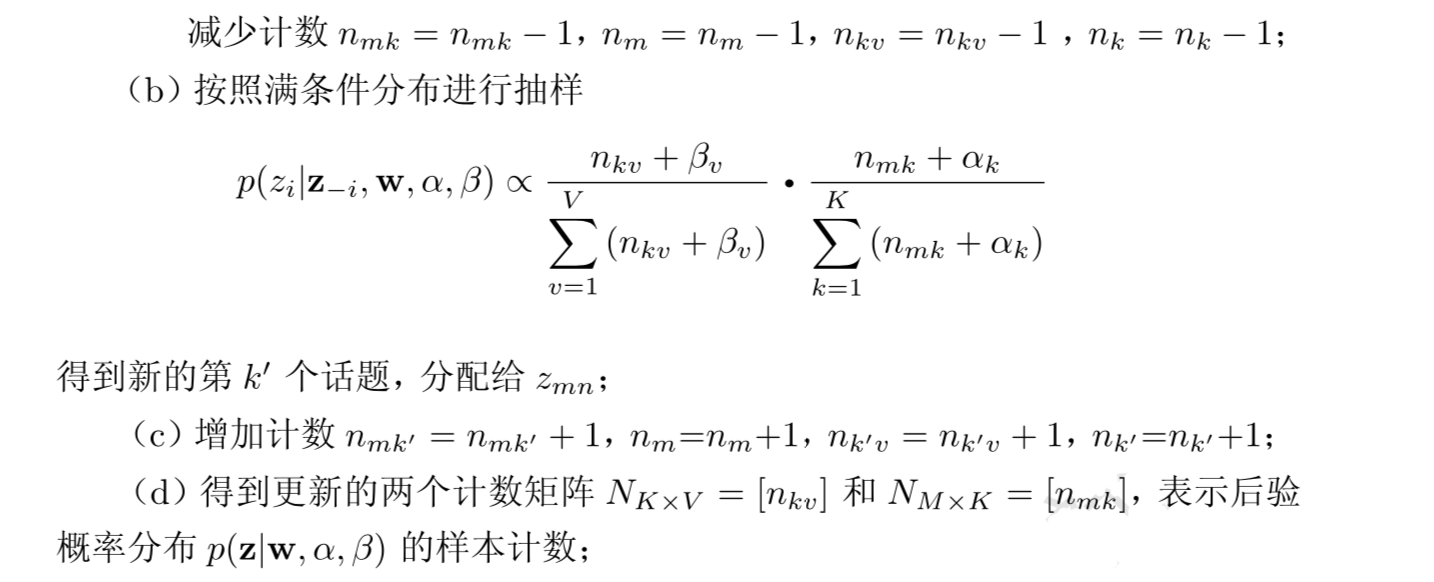
文本的单词序列是观测变量，文本的话题序列是隐变量，文本的话题分布和话题的单词分布也是隐变量。

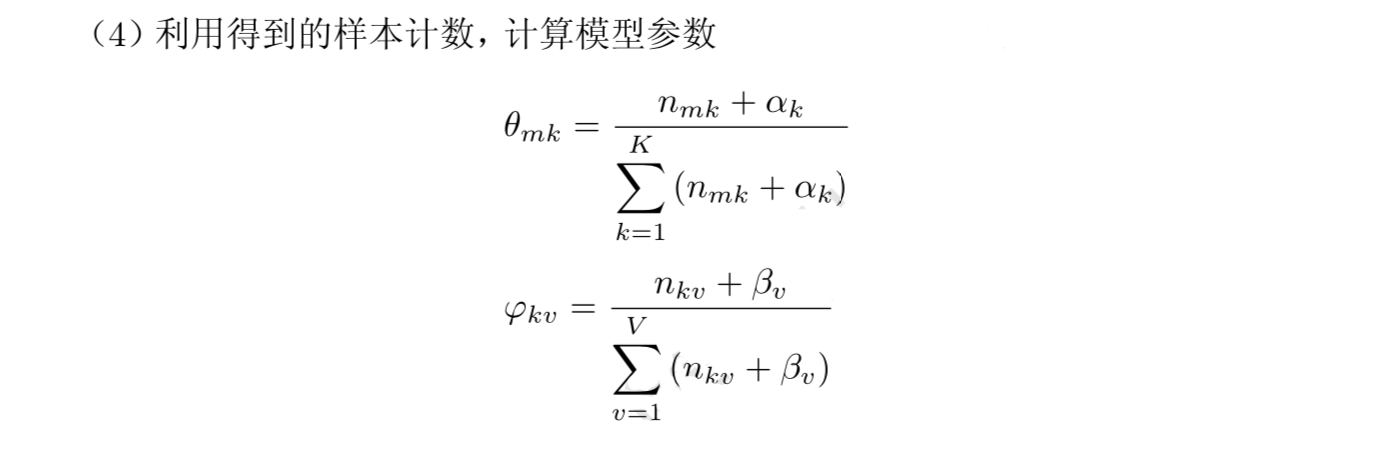


**LDA吉布斯抽样算法**









**LDA与PLSA的关系**

可以认为LDA是PLSA（概率潜在语义分析）的扩展，相同点是两者都假设话题是单词的多项分布，文本是话题的多项分布。不同点是LDA使用狄利克雷分布作为先验分布，而PLSA不使用先验分布（或者说假设先验分布是均匀分布），两者对文本生成过程有不同假设；学习过程LDA基于贝叶斯学习，而PLSA基于极大似然估计。LDA的优点是，使用先验概率分布，可以防止学习过程中产生的过拟合。

**LDA的变分EM算法**

