**第二十章读书报告**

09118223 吴亦珂

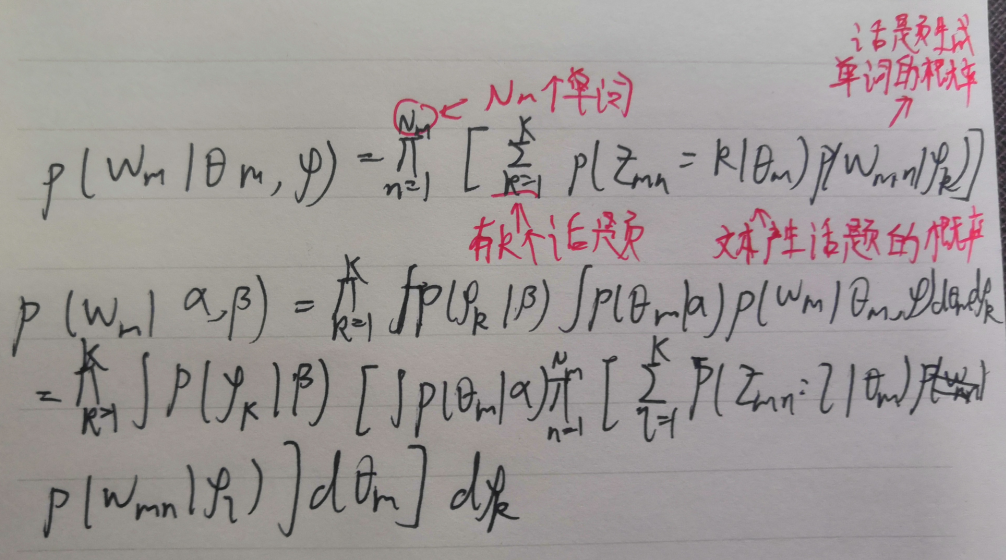
读书进度：二十章读完

1. **问题列表**

（我提出）P391页，为什么说PLSA使用的先验分布是均匀分布？从哪个地方可以看出LDA是基于贝叶斯的学习？

讨论结果：因为在PLSA中主题分布以及单词分布视为以确定的，并不需要使用先验概率生成，因此可以看做是均匀分布或者根本不使用先验分布。而对于LDA，该算法并不认为主题和词的分布是唯一的，而是有多种可能，LDA需要使用两个狄雷克雷分布去确定，，所以这相当于基于贝叶斯的学习。

（我提出）P395（20.17）和（20.18）是如何推导的？

（我提出）P398（20.29）该如何退出？和为什么需要减去当前的计数？

讨论结果：因为满条件分布求解时应该出去当前的单词，所以要将当前单词的计数减去。

（别人提出）LDA与潜在语义分析和概率潜在语义分析之间有什么区别和联系？  
讨论结果：LDA就是在PLSA的基础上再加上一个潜在狄利克雷的先验分布。两者很大的不同在于PLSA是直接确定分布，相当于分布是确定的，而LDA则需要用狄利克雷分布生成。但是两者的基本思想度可以看成是由文本到话题再到单词。

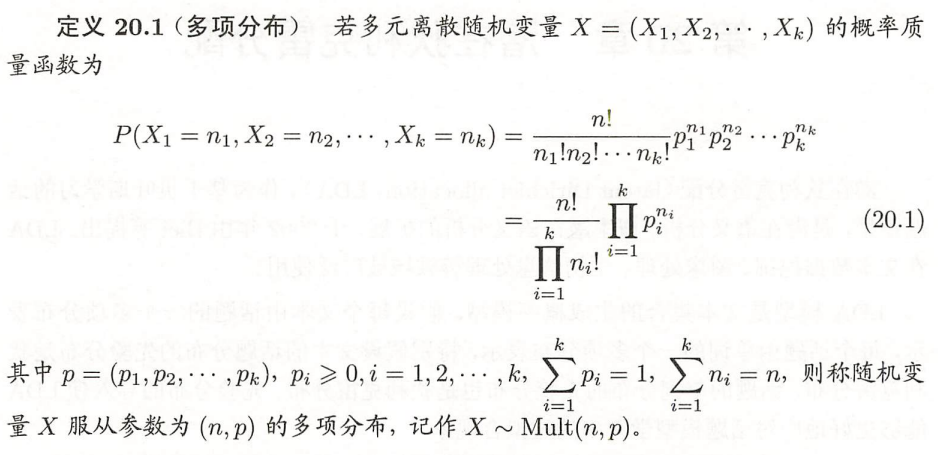
（别人提出）LDA中，狄利克雷分布的重要性是什么？能否使用其他分布来充当先验分布？  
讨论结果：对数据可以采用多项式分布拟合，但是我们希望先验分布和后验分布的形式应该是一样的，即共轭分布。而多项式分布的共轭分布是狄雷克雷分布，因此可以使用狄雷克雷分布作为先验分布。

（别人提出）吉布斯抽样和变分EM推理分别适合在什么情况使用？  
讨论结果：变分推理用的是EM算法，EM算法是肯定收敛的，因而收敛性一定可以得到保证。吉布斯抽样迭代次数较多但是实现较为简单，但变分推理的准确性以及效率更高。

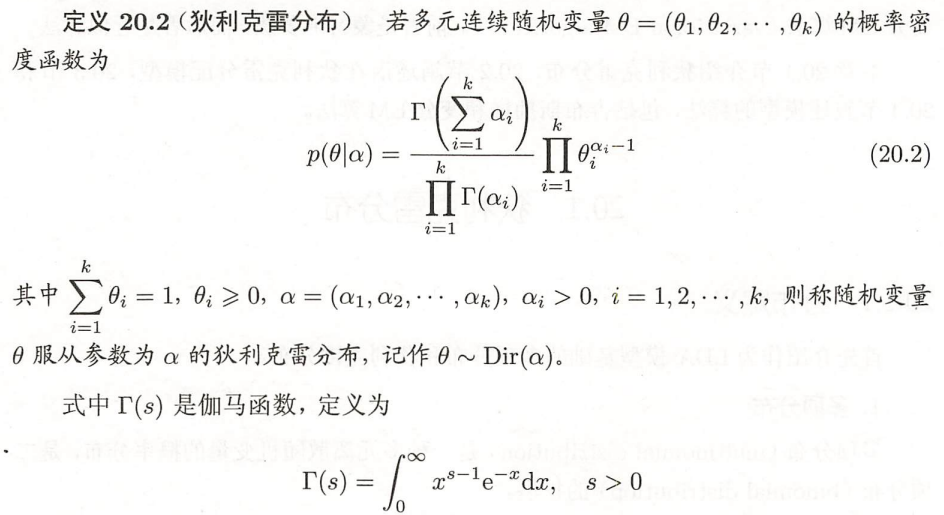
1. **读书收获**

20.1 狄利克雷分布

多项式分布：

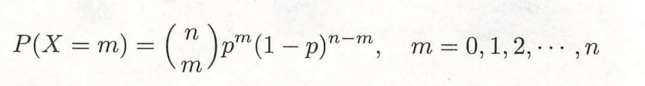


狄利克雷分布：

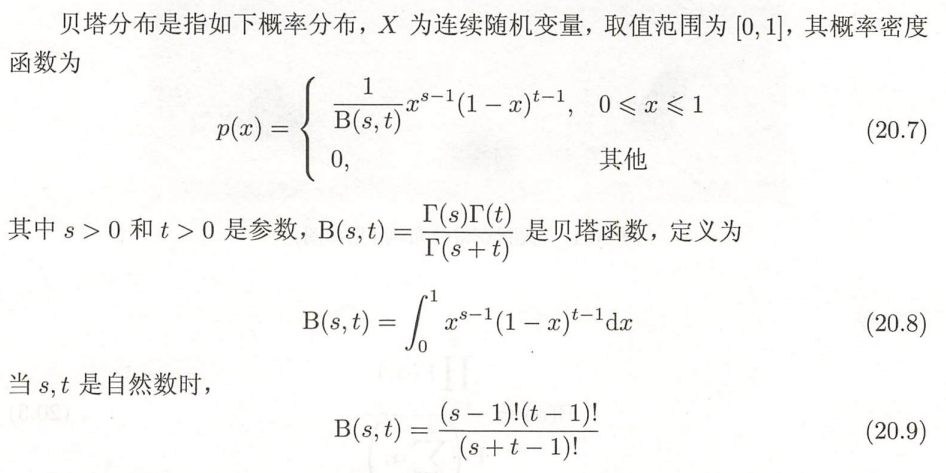


二项分布：

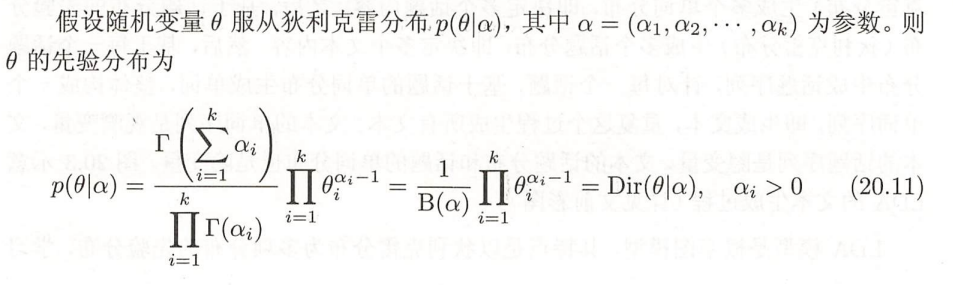
二项分布是多项分布的特殊情况，相当于只有两项：

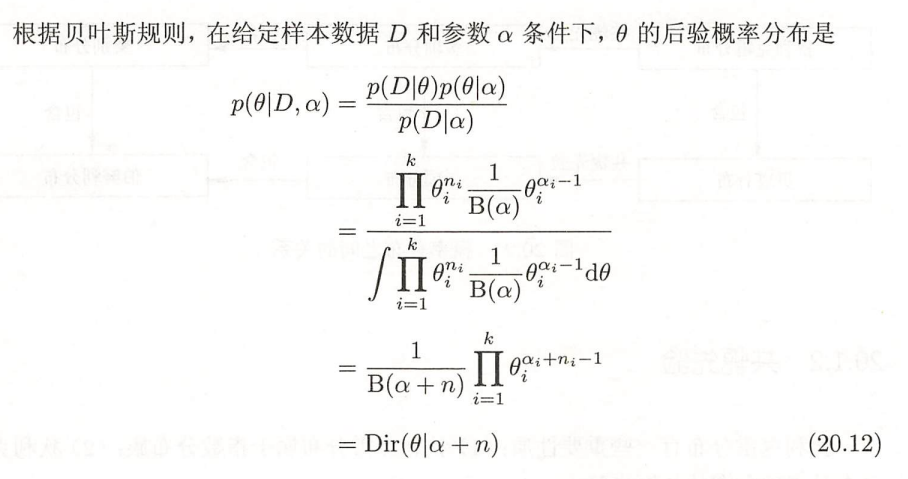


贝塔分布：



共轭先验：如果后验分布与先验分布属于同类，则先验分布与后验分布称为共轭分布，先验分布称为共轭先验。所以多项式的先验分布是狄利克雷分布则其后验分布也是狄利克雷分布。





所以，先验分布和后验分布都是狄利克雷分布。

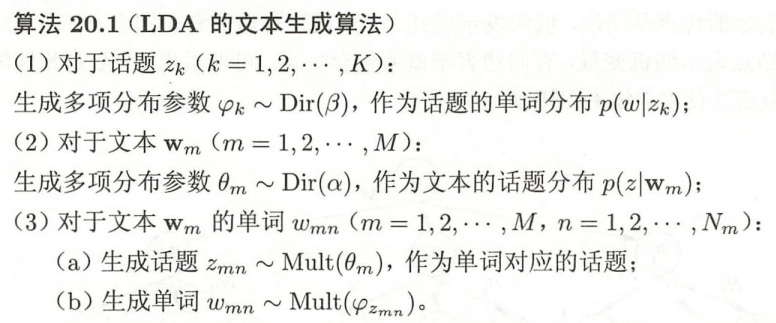
LDA文本集合生成过程：

设单词集合W，文本集合D，话题集合Z，狄利克雷分布的超参数为α和β。

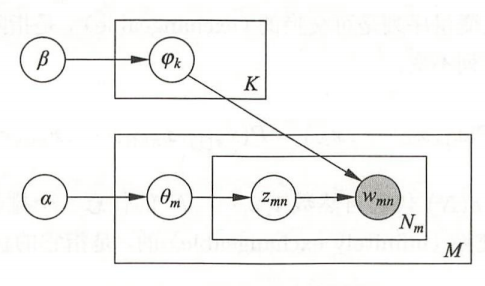
按照狄利克雷分布Dir(β)随机生成一个参数向量，服从Dir(β)，作为话题的单词分布p(w|)。

依照相同的方法，按照狄利克雷分布Dir(α)随机生成一个参数向量，服从Dir(α)，作为文本的话题分布p(z|)。

按照多项分布，由文本随机生成一个话题。然后按照多项分布，随机生成一个单词。

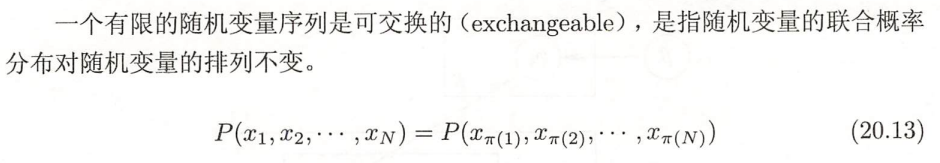


概率图模型：

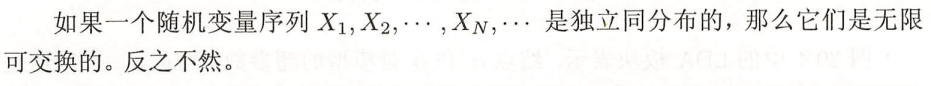


由β生成话题的单词分布的参数，重复K次，因为有K个话题。由α生成文本的话题分布的参数，重复M次，因为有M个文本。对于每个话题分布，产生话题，并且还要产生单词。因为一篇文本有个单词，因此需要次生成。

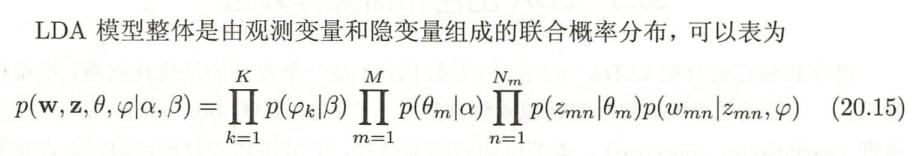
可交换的：

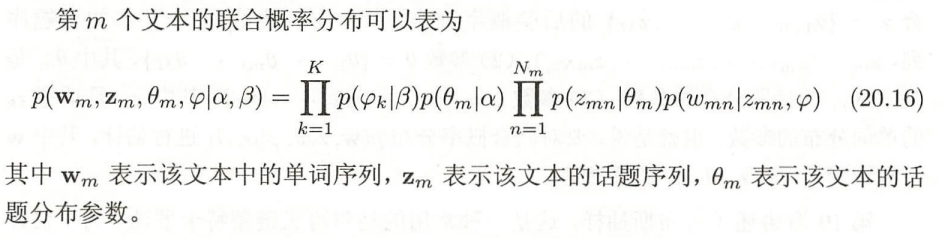


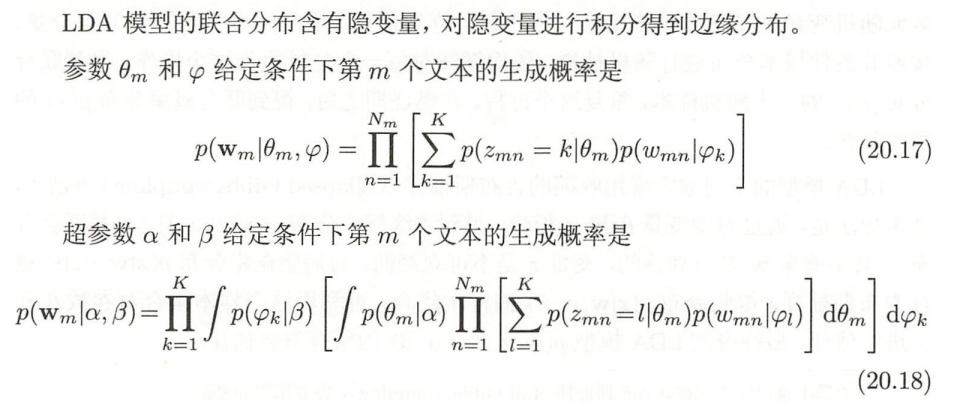
无限可交换的：任意一个有限子序列都是可交换的。

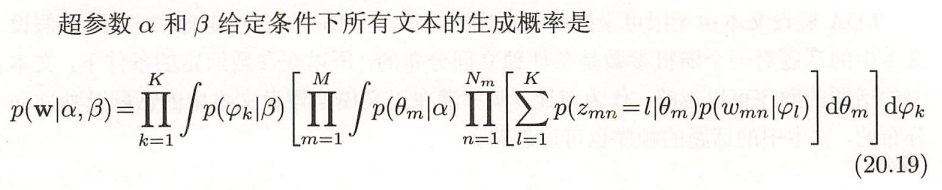


LDA假设文本由无限可交换的话题序列组成。

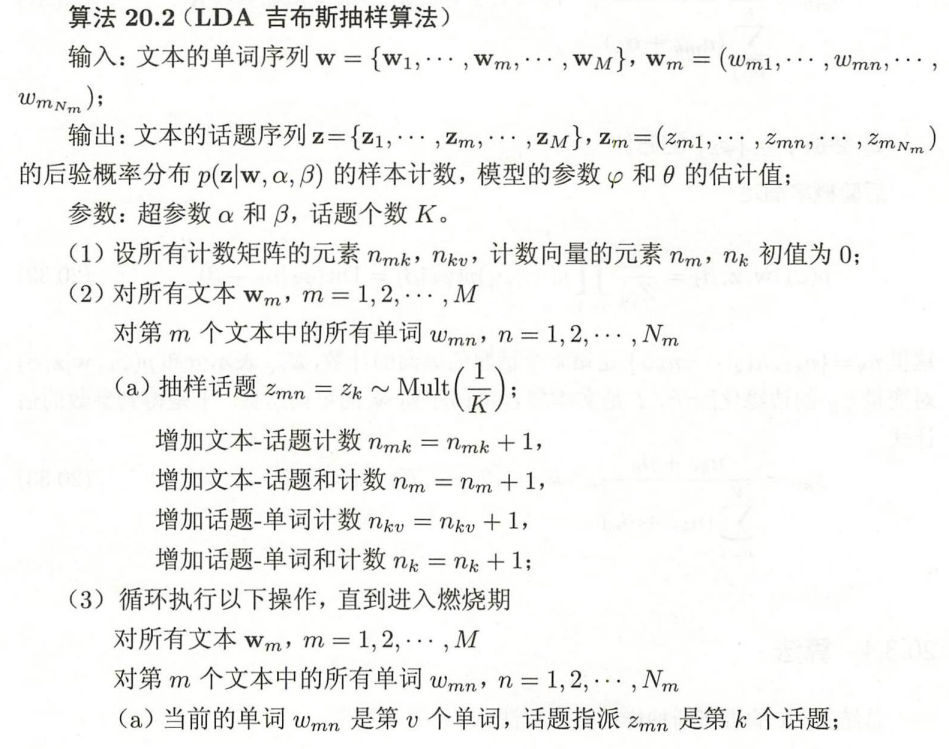


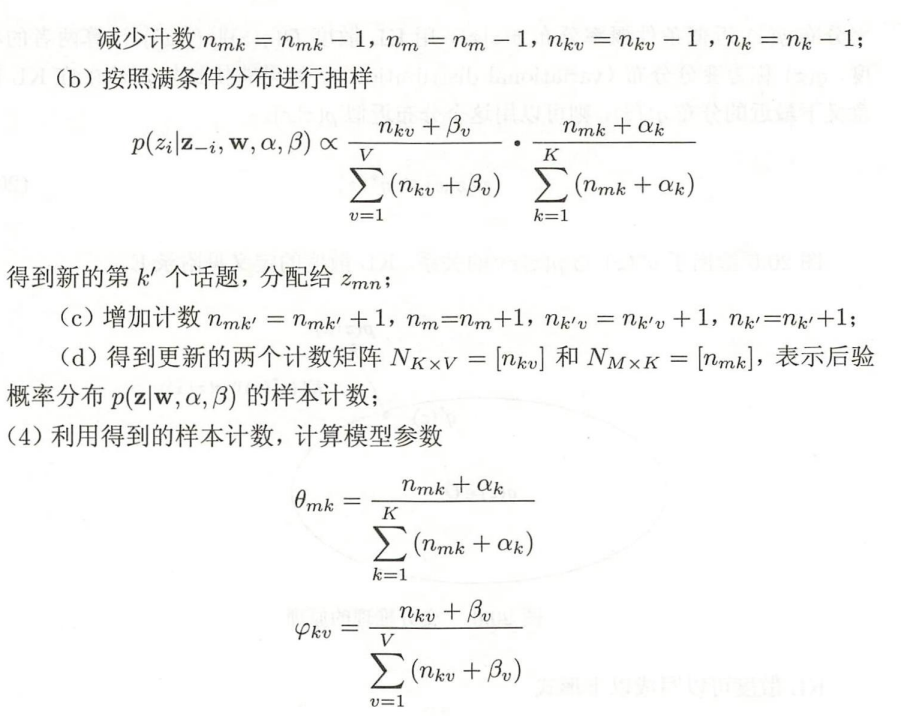






20.3 LDA的吉布斯抽样算法





20.4 LDA的变分EM算法

