**第十六章读书报告**

09118223 吴亦珂

读书进度：第十六章读完

1. **问题列表**

（我提出）书P313在用数据的协方差矩阵或相关矩阵求解PCA时，其实只要算到（4）就已经可以实现降维，后面那几步是不是可以没有，仅仅是为了对数据进行分析？第二种方法仅实现降维，后面几步并没有去算贡献率之类。

讨论结果：如果单单要实现降维，只要进行前面几步就可以完成。但一般来说后面几步也是非常重要的，可以对数据整体进行一个分析，获得对于数据整体更清楚的认识。同时，也是对k选择的一个检验，看看效果是否好。

（我提出）书上介绍了两种求解PCA的方法，为什么现在常用第二种方法，是因为第二种方法效率更高吗？因为一个求解特征值和特征向量，另一个需要进行SVD 。

讨论结果：应该来说进行SVD确实比特征值分解有优势。首先，特征值分解应该只能处理方阵，而SVD可以处理几乎所有矩阵。其次，SVD的效率更高。第一，SVD方法并不需要计算，因为这样可能导致矩阵中一些较小的数的缺失。第二，SVD的速度比特征值分解要快，可以充分利用协方差矩阵的性质。

（我提出）第二种方法的k值必须提前指定吗？这个算法好像不像第一种可以进行选择。因为直接进行截断奇异值分解，事先就必须先将k确定。

讨论结果：需要提前确定k的值，这样才可以进行截断奇异值分解。

（别人提出）样本的协方差矩阵和相关矩阵分别代表着什么？  
我的理解是这两个都是表示样本各个变量之间的线性相关关系。当均值为0，方差为1是两者一致。感觉这两者就是对变量间线性相关关系的一个度量。

（别人提出）主成分分析得出的结果除了可以用于聚类分析还有什么常见应用吗？  
我觉得只要是数据维数过大想要降低数据维数的场合应该都可以用PCA进行处理吧。比如SVM、线性回归应该都可以。

（别人提出）常用的两种PCA求解方法之间有什么区别与联系？哪一种效率更高？  
联系：两种解法其实都是求协方差矩阵的特征值与特征向量。区别：一个直接求，另一个构造一个矩阵利用SVD将两者求出。SVD确实效率更高。第一，SVD方法并不需要计算，因为这样可能导致矩阵中一些较小的数的缺失。第二，SVD的速度比特征值分解要快，可以充分利用协方差矩阵的性质。

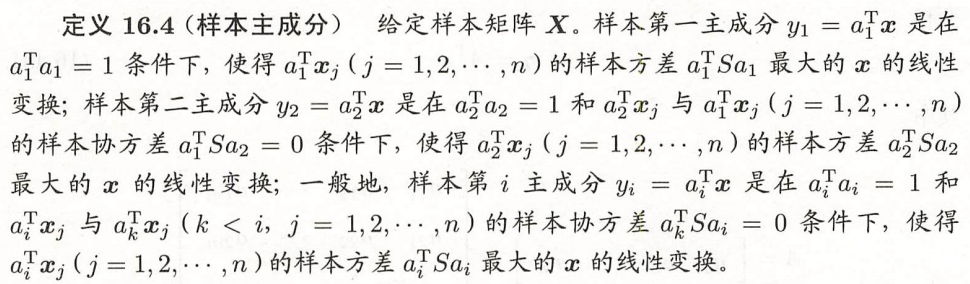
1. **下周读书计划**

下周计划阅读第十七章的前半部分内容。第十七章内容较为生疏，之前接触较少。

1. **读书收获**

16.2节主要介绍了样本主成分的性质以及PCA具体的实现方法。

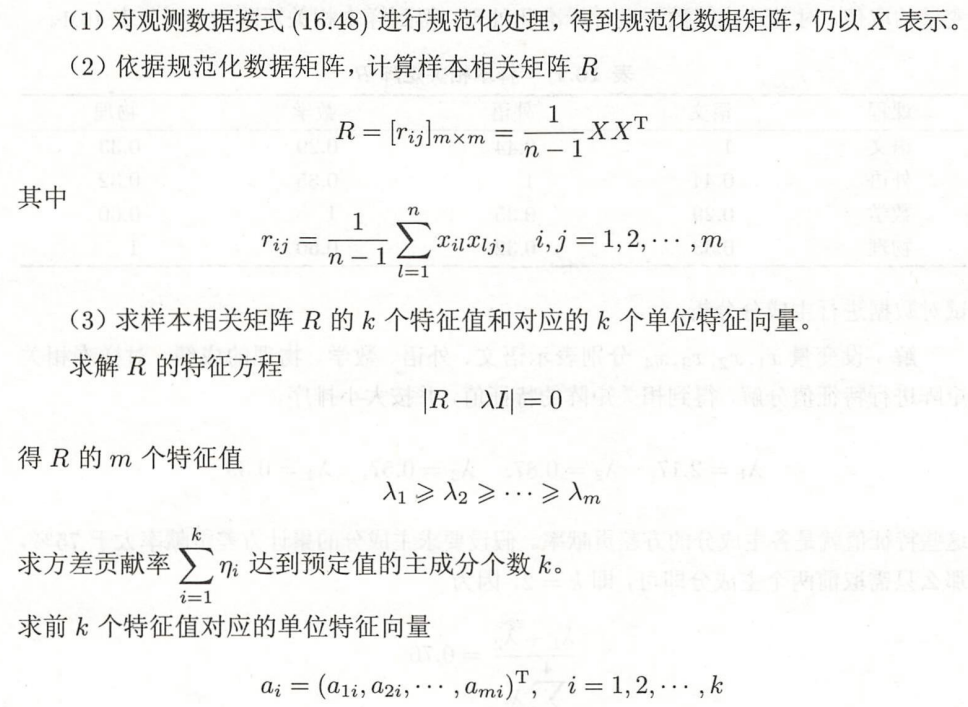
1. 样本主成分的定义以及相关概念：

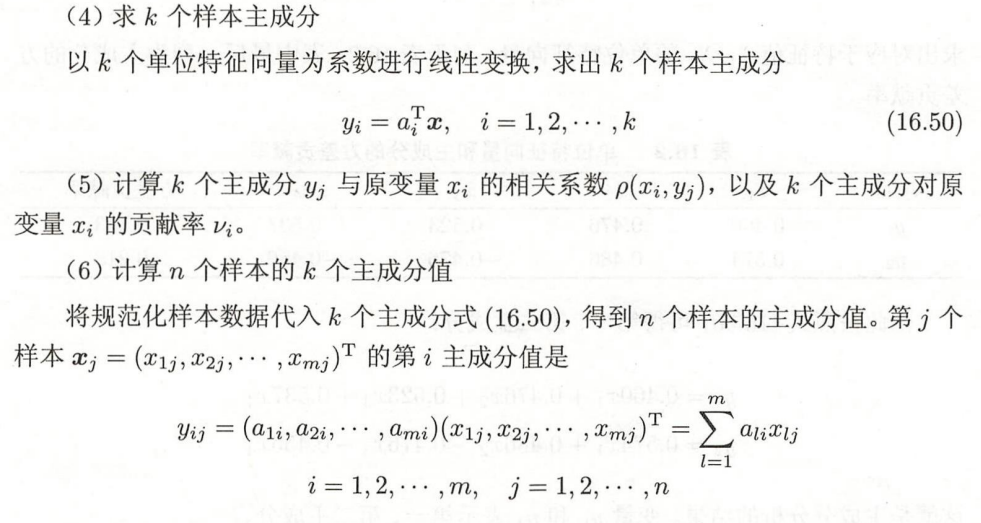


一般来说，在进行PCA之前需要将数据进行规范化，也就是使其均值为0，方差为1。这样一来，协方差矩阵与相关矩阵一致，均为。

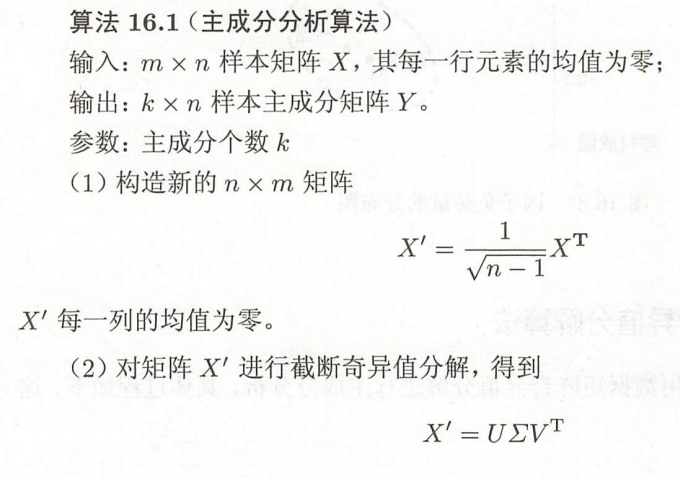
1. PCA的具体实现方法：

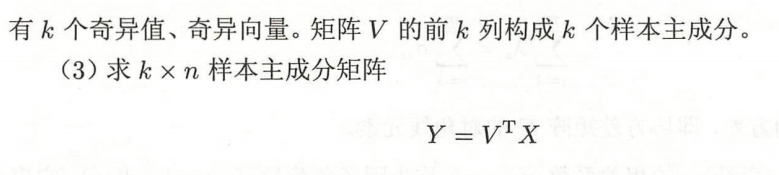
直接进行计算：即直接计算协方差矩阵（规范化后协方差矩阵与相关矩阵一致）的特征值以及特征向量，之后可以求解主成分的因子负荷量以及贡献率，来对样本数据进行进一步的分析。





考虑到求解PCA需要求解特征值以及特征向量，故此采用SVD的方法：





这两种方法第一种是传统求解方法，第二种是现在常用方法。主要是SVD可以适用于几乎所有矩阵，而特征值分解只能用于方阵，而且SVD的效率也相对较高。