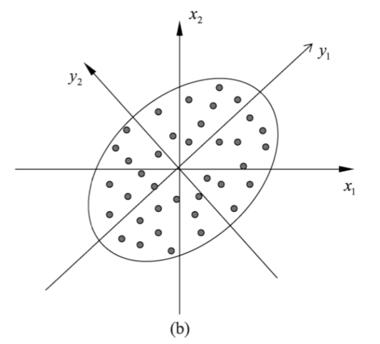
**读书报告**

61518424 王贵涛

1. **问题与解答**

我提出问题：

1. 在下图中，为什么说当知道其中一个变量y1的取值时，对另一个变量y2的预测是完全随机的，y2的取值范围难道不是跟y1有关吗？



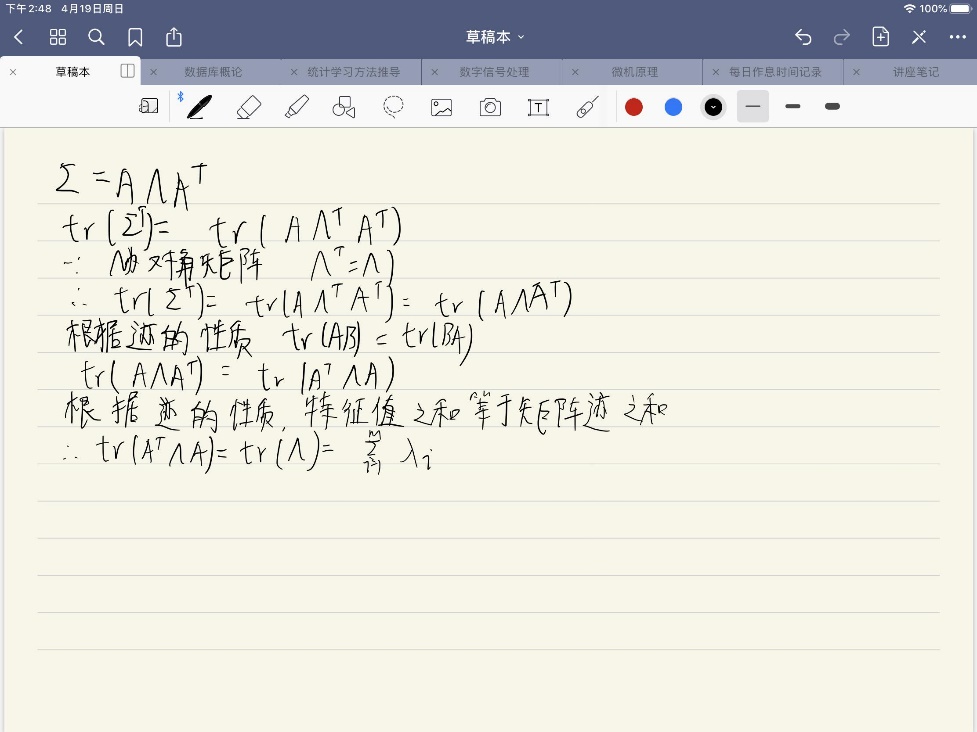
讨论结果：在知道y1的情况下，除了y2的取值范围，我们并没有得到任何其他关于y2的信息，y2的值是完全随机的，而如果是原始图中的x1，x2的话，知道x1，就能知道x2在哪个象限的可能性大，所以x1对x2有着一种预测的作用，但是y1就无法得出y2在哪个部分的概率大这种信息。

1. 主成分分析与奇异值分解有什么异同之处？

讨论结果：主成分分析与奇异值分解都实现了降维，主成分分析实现了坐标的变化。SVD有个好处，有一些SVD的实现算法可以不求先求出协方差矩阵，也能求出右奇异矩阵V。也就是说，PCA算法可以不用做特征分解，而是做SVD来完成。另一方面，注意到PCA仅仅使用了我们SVD的右奇异矩阵，没有使用左奇异矩阵。而左奇异矩阵可以用于行数的压缩，右奇异矩阵可以用于列数，也就是PCA降维。所以，有了SVD就可以得到两个方向的PCA。

1. 证明x的方差和等于y的方差和时用了矩阵的迹的什么性质？

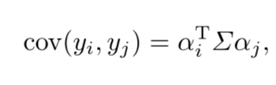
讨论结果：



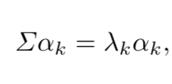
别人提出的问题：

1. 在主成分分析中，如何保证结果是线性无关的？

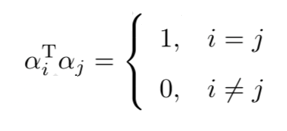
我的解答：通过协方差式子我们有：



计算最大方差时，我们得到：



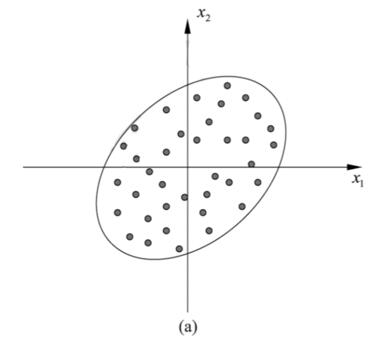
由于a的正交性：



因此cov(yi, yj)=aiλjaj=0，y的协方差为0，线性无关。

1. 为什么说16.1(a)中的x1和x2是线性相关的?

我的解答：如下图所示：



当x1取不同值时，x2的值也发生变化，而且变化的趋势与x1有关，大体是x1越大，x2也越大，因此他们是线性相关的。

1. 为什么需要将总体主成分和样本主成分分开讨论？

我的解答：在数据总体上进行的主成分分析称为总体主成分分析，在有限样本上进行的主成分分析称为样本主成分分析，总体主成分分析是样本主成分分析的基础，总体主成分分析，是定义在样本总体上的。在实际问题中，需要在观测数据上进行主成分分析，这就是样本主成分分析。总体和样本在统计学计算上有一定的区别。

1. 为什么主成分分析时需要对数据进行规范化，使其方差为1？

我的解答：在实际问题中，不同变量可能有不同的量纲，直接求主成分有时会产生不合理的结果。

1. **下周计划安排**

看完第十六章并参加讨论。

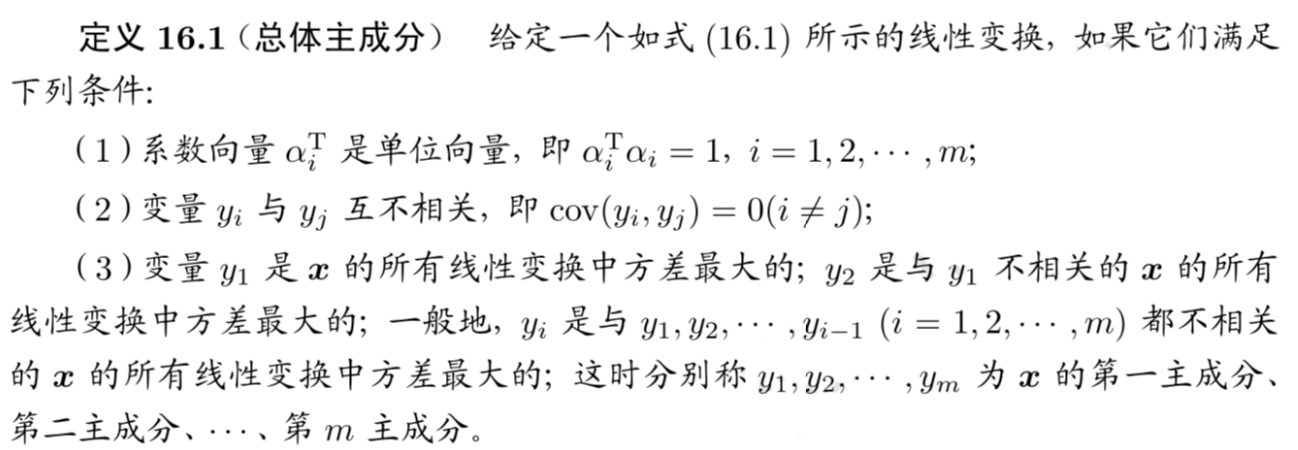
1. **读书收获**

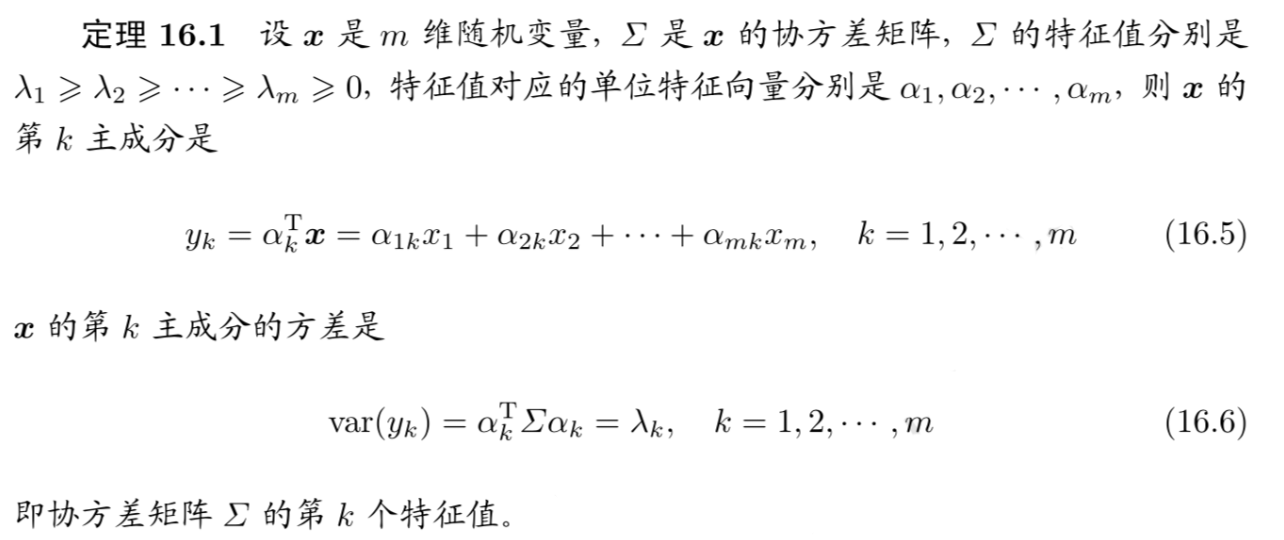
主成分分析（principal component analysis, PCA）是一种常用的无监督学习方法

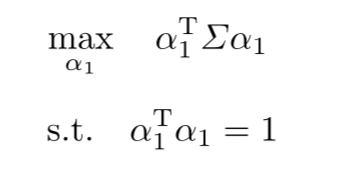
这一方法利用正交变换把由线性相关变量表示的观测数据转换为少数几个由线性无关变量表示的数据，线性无关的变量称为主成分。

主成分的个数通常小于原始变量 的个数，所以主成分分析属于降维方法。

主成分分析主要用于发现数据中的基本结构，即数据中变量之间的关系。



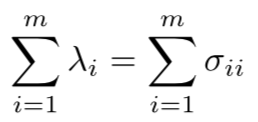
求第一主成分就是求解约束最优化问题：



总体主成分y的协方差矩阵是对角矩阵：



总体主成分y的方差之和等于随机变量x的方差之和，即：



主成分分析的主要目的是降维，所以一般选择k (k << m）个主成分（线性无关变量）来代替m个原有变量（线性相关变量），使问题得以简化，并能保留原有变量的大部分信息。