1. **主成分分析**

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，这些**线性无关的变量**成为**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**。

**对线性相关的理解：**

在书上的示意图中，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？

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

**16.1 总体主成分分析**

**16.1.1 基本想法**

对给定数据进行规范化，使得数据每一变量的**平均值是0，方差为1**，之后对数据进行**正交变换**，原来由线性相关变量表示的数据，通过正交变换变成由若干个线性无关的新变量表示的数据。

**方差**：表示**信息量**的大小

对原坐标系中数据进行主成分分析等价于对坐标轴进行**旋转变换**

在新坐标系里，数据中的变量y1和y2是线性无关的，当**知道其中一个变量y1的取值的时候，对另一个变量的预测完全是随机的**

旋转中样本点到原点的距离的平方和保持不变

在数据总体(population)上进行的主成分分析成为**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，前者是后者的基础

**16.1.2 定义和导出**

**总体主成分**定义：

给定一个线性变换，如果它满足: ①**系数向量是单位向量** ②变量**yi与yj互不相关** ③**变量y1是x的所有线性变换中方程最大的**。Y2是不与y1相关的x的所有线性变换中方差最大的。一般的，yi是与它之前的成分都不相关的x的所有线性变换中方差最大的，这时分别称y1,y2...ym为x的第一、第二...第m主成分。

第②第③个定义给出了一个迭代求主成分的过程

**16.1.3 主要性质**

定理16.1： x的第k主成分的方差是协方差矩阵的第k个特征值

推导时可以从第k-1主成分触发递推证明第k个主成分的情况

推论16.1：成为x主成分的充要条件

性质：

1. 总体主成分y的**协方差矩阵是对角矩阵**
2. 总体主成分y的方差之和等于随机变量x的方差之和
3. 第k个主成分yk与变量xi的相关系数r(yk,xi)成为因子的**负荷量(factor loading),** 它表示**第k个主成分yk与变量xi的相关关系**
4. 第k个主成分yk与m各变量的因子负荷量满足关系
5. m个主成分与第i个变量xi的因子负荷量满足

**16.1.4 主成分个数**

定理16.2 对于任意正整数q，1≤q≤m，考虑正交线性变换y=BTx，其中y是q维向量，BT是q×m矩阵，则**y的协方差矩阵的迹在B=Aq时取得最大值，**其中取值Aq由正交矩阵A的前q列组成

**矩阵的迹：**在线性代数中，一个n×n矩阵A的**主对角线（从左上方至右下方的对角线）上各个元素的总和被称为矩阵A的迹（或迹数）**，一般记作tr(A)。

证明主要是从迹的角度进行变换

当A的前q列取x的前q个主成分时，能最大限度的保留原有的变量方差信息

定理16.2、16.3从方差的角度定义了主成分y1...yk对于**信息的贡献率，**实际使用中只需要这个贡献率达到一定的阈值即可