**主成分分析（PCA）**

# 目录

方差

协方差

维度灾难

主成分分析（PCA）

一、标准化

二、计算协方差矩阵

三、计算出主成分

主成分是什么

怎么计算主成分

特征值和特征向量

四、主成分向量

五、将数据映射到新的主成分坐标系中

PCA的步骤

## 方差

方差是指一组数据中的各个数减去这组数据的平均数的平方和的平均数。

例：（1，2，3，4，5）

平均数=3

各个数与平均数差的平方和=10

方差=2

方差用于测算数值型数据的离散程度。

协方差

协方差用于度量各个维度偏离其均值的程度。

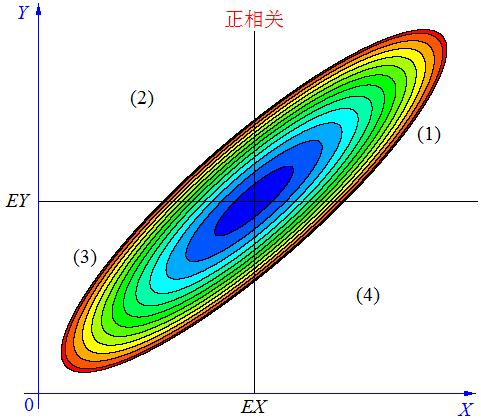
协方差的计算方法（以二维为例子）：

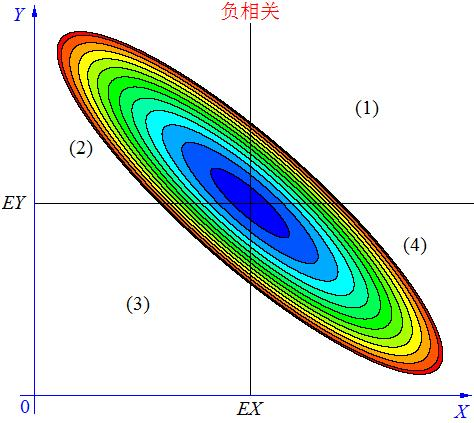
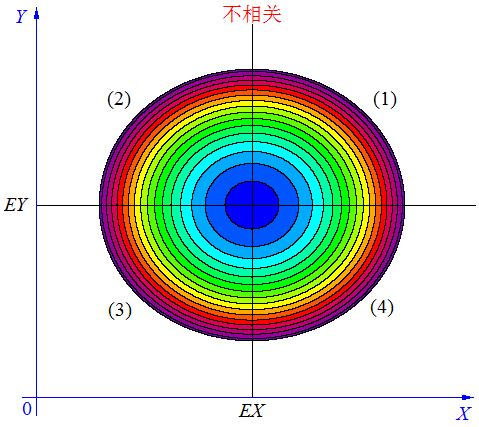


其中，表示均值。

协方差的结果有三种解释方式

当cov(X,Y)>0时，表明X和Y正相关；

当cov(X,Y)<0时，表明X和Y负相关；

当cov(X,Y)=0时，表明X和Y不相关。

协方差为0，两个随机变量不一定相互独立，而两个随机变量相互独立，协方差一定为0。

## 维度灾难

简单的说就是变量的个数多，如果变量个数增加，随之需要估计的参数个数也在增加

## 主成分分析（PCA）

PCA是一种降维方法，常用于对高维数据集作降维，它会将一个大的变量集合转化为更少的变量集合，同时保留大的变量集合中的大部分信息。

## 主成分分析的几个步骤如下：

### 一、标准化

为了让每一个维度对分析的结果造成同样的影响，我们需要对连续的初始变量的范围作标准化。

具体一点说就是因为我们后续的结果对数据的方差十分敏感，取值范围较大的维度会比相对较小的维度造成更大的影响（例如一个在1-100之间变化的维度对结果的影响，比一个0-1的更大），会导致一个偏差较大的结果，所以，将数据转化到相似的范围可以预防这个问题。

数据标准化的方法如下

 其中standard deviation表示标准差）

### 二、计算协方差矩阵

这一步是为了理解数据集中的变量是如何从平均值变化过来的，同时可以查看不同的特征之间又有什么关系，此时我们要计算协方差矩阵。

协方差矩阵是一个P\*P的对称矩阵（P是维度的数量）它涵盖了数据集中所有元组对初始值的协方差，例如一个拥有三个变量x,y,z和三个维度的数据集，协方差矩阵将是一个3\*3的矩阵（协方差的计算方法及含义见上文）：

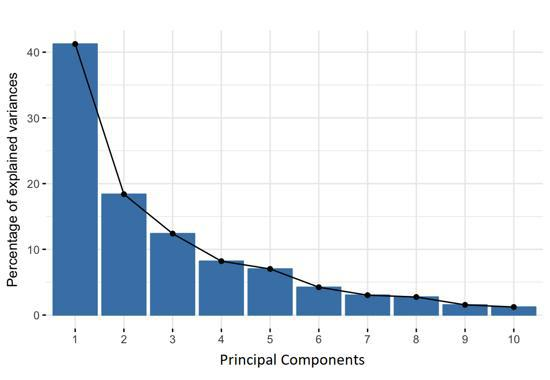
由于变量与自身的协方差等于他的方差，在主对角线上我们已将计算出了各个变量初始值的方差。协方差矩阵的每一个元组关于主对角线对称，这意味着上三角部分和下三角部分是相等的。

### 三、计算出主成分

#### 主成分是什么

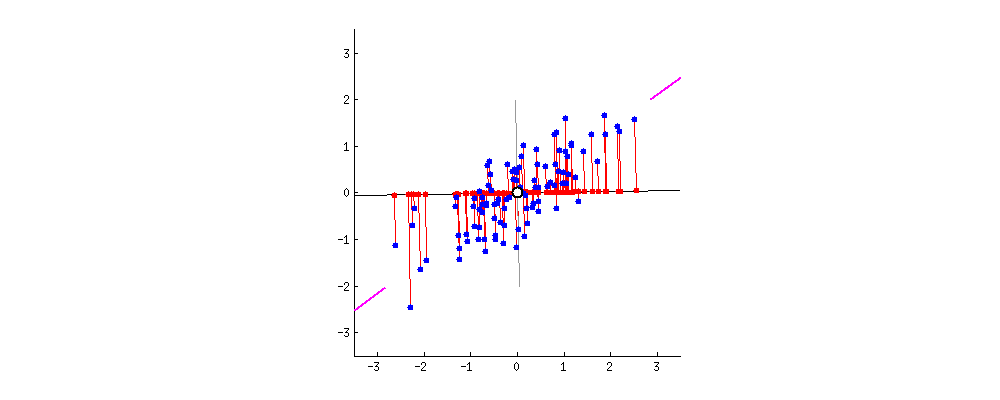
主成分是一个新的变量，他是初始变量的线性组合。新的变量之间是不相关的，第一个主成分中包含了初始变量的大部分信息，是初始变量的压缩和提取。

例如：虽然在一个 10 维的数据集中我们算出了 10 个主成分，但大部分的信息都会被压缩在第一主成分中，剩下的大部分信息又被压缩到第二主成分中，以此类推，我们得到了下面这张图：

从理论方面来说，主成分代表着蕴含 最大方差的方向。对于主成分来说，变量的方差越大，空间中点就越分散，空间中的点越分散，那么它包含的信息就越多。简单的讲，主成分就是一条更好的阐述数据信息的新坐标轴，因此我们更容易从中观测到差异。

#### 怎么计算主成分

有多少个变量就有多少个主成分，对于第一主成分来说沿着对应的坐标轴变化意味着有最大的方差，例如用下列的散点图表示：

你能直接猜测出主成分应该是沿着哪一个方向吗，大致是图中紫色线的方向。（因为它穿过了原点，而且数据映射在这条线上后，有着最大方差（各点与原点距离的均方））

第二个主成分也是这样计算的，它与第一主成分不相关（即为互相垂直）表示了下一个最大方差的方向。

重复上面的步骤，直到我们从原始数据中计算出所有的主成分。

#### 特征值和特征向量

特征值和特征向量通常成对出现，每一个特征向量对应一个特征值，他们各自的数量相等，等于原始数据的维度，例如有三个变量就会有三个特征向量与三个特征值。

协方差矩阵的特征向量其实就是一些列的坐标轴，将数据映射到这些坐标轴之后，我们将会得到最大的方差（这意味着更多的信息），他们就是我们要求的主成分，特征值其实就是特征向量的系数，它代表了每个特征向量包含了多少信息量。

### 四、主成分向量

主成分向量仅仅是一个矩阵，里面有那些我们决定保留的特征向量。这是数据降维的第一步，我们只是要在n个变量中保留p个特征向量（成分）我们把数据映射到新的坐标轴上时，最后数据将只有p个维度。

### 五、将数据映射到新的主成分坐标系中

我们将使用那些从协方差矩阵中算出来的特征向量形成主成分矩阵，并将原始数据映射到主成分矩阵对应的坐标轴上 —— 这就叫做主成分分析。具体的做法便是用原数据矩阵的转置乘以主成分矩阵的转置。

## PCA的步骤

1）将原始数据按列组成n行m列矩阵X

2）将X的每一行（代表一个属性字段）进行零均值化，即减去这一行的均值

3）求出协方差矩阵

4）求出协方差矩阵的特征值及对应的特征向量

5）将特征向量按对应特征值大小从上到下按行排列成矩阵，取前k行组成矩阵P

6）Y=PX即为降维到k维后的数据