主成分分析(Principle Component analysis, PCA)

1. 背景介绍

对某一维随机变量做次观测，得到的样本矩阵即为，假设各观测样本近似分布于某一超平面附近，我们希望将其近似分解为该超平面的基向量的线性组合，将数据从原有坐标系转换到新的低维坐标系，进而实现数据降维且希望可以去除原有变量的相关性.

为找到一种合理的坐标变换实现数据降维，一方面希望减少分析指标数量，另一方面尽可能保留原有数据的信息. 在信号处理中，认为信号具有较大的方差，噪声有较小的方差，信噪比就是信号与噪声的方差比，通常希望其越大越好. 因此，我们衡量坐标变换效果的标准为数据在某一分量上的方差值，方差值越大则数据在该分量上的离散程度越大，即反应保留的信息更多. 因此我们选取使方差最大的方向为第一坐标轴方向，选取使方差次大的方向作为第二坐标轴方向，以此类推……选取使方差第大的方向作为第坐标轴方向.

1. 模型描述

给定样本矩阵.

设变量，对原有变量做变换，得新变量和新样本矩阵，则

其中，.

希望新变量满足：

① . 表示相关系数.

② 最大，次大，……第大.

又由于等价于，得如下优化问题：

1. 模型求解
   1. **拉格朗日乘子法**

为找使 最大且的. 令，构造拉格朗日函数如下：

由极值必要条件得：

得为矩阵的特征值，为对应的特征向量. 此时

若选取模为1的特征向量，则

即所要的为最大特征值所对应的模为1的特征向量，且等于的最大特征值.

同理，为找使 最大且的. 构造拉格朗日函数如下：

由极值必要条件得：

得为矩阵的特征值，为对应的特征向量. 此时

若选取模为1的特征向量，则

由特征向量相互正交得满足条件，且为矩阵的第二大特征值.

即所要的为第二大特征值所对应的模为1的特征向量，且等于的第二特征值.

以此类推，为第大特征值所对应的模为1的特征向量，且等于的第大特征值.

由此得到各主成分，其中. 将协方差矩阵的特征值从大到小排序，定义得分

其中，为协方差矩阵的第大的特征值.

根据得分大小判断降维效果，得分越高则保留原有信息越多. 通常选取使得分大于0.85的前几大特征值作为各主成分.

**算法描述**

* 1. 求;
  2. 求特征值与特征向量;
  3. 根据得分选取主成分.
  4. **SVD求解**

将原样本矩阵按行零均值化，得矩阵

则协方差矩阵可表示为

SVD(Singular Value Decomposition)，奇异值分解. 对任意矩阵，可将其唯一分解为：

其中为正交矩阵，只有对角位上有非零元素，且从大到小排列，这些对角元素称为奇异值.

奇异值分解的性质如下：

1. 为正交矩阵，即；
2. ，，即是的特征向量，是的特征向量，奇异值的平方是和的特征值；
3. 选取前个奇异值，的前列，的前行近似描述矩阵，即

根据，我们只需将做SVD分解，并选取右奇异矩阵的前列.

**算法描述**

1. 求零均值化矩阵;
2. 对作奇异值分解;
3. 选取右奇异矩阵的前列.
4. 模型推广

对原样本矩阵作主成分分析后，得到新样本矩阵和变换矩阵，希望根据变换矩阵将原样本矩阵分解为标准正交基的线性组合.

假设，则有

其中，.

由于的行向量相互正交，因此可通过下式求解.

则有

或对做SVD分解，得左奇异矩阵奇异值矩阵和右奇异矩阵. 由奇异值分解性质有

可根据上式将原样本矩阵近似表示为个正交基的线性组合.