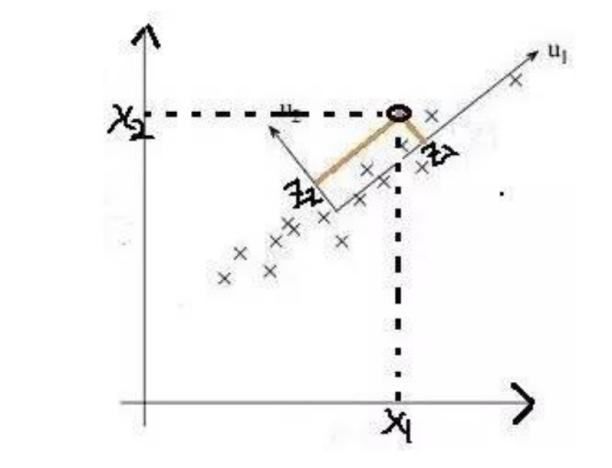
# PCA理论分析

PCA是一种常见的降维方法，目的在于不损失大量信息的情况下，减少特征向量的维度来降低高维度给计算机的存储和计算带来的极大负担。本文先给出PCA的实现步骤，再叙述PCA的原理。

下面先二维降一维为例，对PCA做一个直观的解释：

假设训练样本如下图中的“x”所示，标准直角坐标系为（1,0）和（0,1），PCA的目的：找到一组新的标准正交坐标系向量和，使得原训练样本通过坐标变换后得到新的坐标，其中足够小（可以忽略不计），留下，丢掉，这样就达到了降维的目的，同时注意要求新的坐标系向量要是正交的，目的也是尽量保留原始的有用信息。

**PCA实现的步骤：**

给出一组训练样本（未作特殊说明，向量均为列向量）：



1. 去均值（Mean Normalization）:



1. 计算样本的协方差矩阵：



3、对C进行特征值分解或者奇异值分解，得到N个特征向量，再对特征向量进行标准化，得到最终的特征向量；

4、取最大的d个特征值所对应的特征向量组成的矩阵作为新的坐标向量；

5、通过下式（坐标变换）：



得到降维的d维特征向量。

**PCA原理：**

为什么协方差矩阵C的特征向量可以作为新的坐标系？

首先构建一个准则：N维向量通过PCA降维，得到维向量：



使损失的信息越少越好。

如何衡量损失的信息？我们让通过新坐标矩阵进行还原，即，得到向量（维度为），计算与之间的欧式距离，距离越小代表损失的信息越少。

通过上面的准则构建的目标函数如下所示：



即转化为一个优化问题：



由拉格朗日乘子法可得拉格朗日函数为：



由：



可得：



由此可见：为协方差矩阵的特征向量是上式成立的充分条件，即当为协方差矩阵的特征向量时，上式成立。