# [主成分分析-最大方差解释](http://www.cnblogs.com/jerrylead/archive/2011/04/18/2020209.html)(来源于网络)

## 1. 问题

真实的训练数据总是存在各种各样的问题：

1、 比如拿到一个汽车的样本，里面既有以“千米/每小时”度量的最大速度特征，也有“英里/小时”的最大速度特征，显然这两个特征有一个多余。

2、 拿到一个数学系的本科生期末考试成绩单，里面有三列，一列是对数学的兴趣程度，一列是复习时间，还有一列是考试成绩。我们知道要学好数学，需要有浓厚的兴趣，所以第二项与第一项强相关，第三项和第二项也是强相关。那是不是可以合并第一项和第二项呢？

3、 拿到一个样本，特征非常多，而样例特别少，这样用回归去直接拟合非常困难，容易过度拟合。比如北京的房价：假设房子的特征是（大小、位置、朝向、是否学区房、建造年代、是否二手、层数、所在层数），搞了这么多特征，结果只有不到十个房子的样例。要拟合房子特征->房价的这么多特征，就会造成过度拟合。

4、 这个与第二个有点类似，假设在IR中我们建立的文档-词项矩阵中，有两个词项为“learn”和“study”，在传统的向量空间模型中，认为两者独立。然而从语义的角度来讲，两者是相似的，而且两者出现频率也类似，是不是可以合成为一个特征呢？

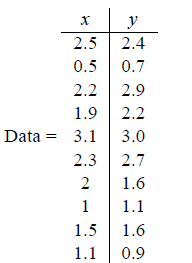
5、 在信号传输过程中，由于信道不是理想的，信道另一端收到的信号会有噪音扰动，那么怎么滤去这些噪音呢？

下面探讨一种称作主成分分析（PCA）的方法来解决部分上述问题。PCA的思想是将n维特征映射到k维上（k<n），这k维是全新的正交特征。这k维特征称为主元，是重新构造出来的k维特征，而不是简单地从n维特征中去除其余n-k维特征。

## 2. PCA计算过程

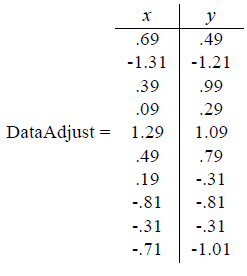
首先介绍PCA的计算过程：

假设我们得到的2维数据如下：

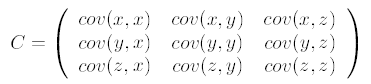
[](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182110393017.png)

行代表了样例，列代表特征，这里有10个样例，每个样例两个特征。可以这样认为，有10篇文档，x是10篇文档中“learn”出现的TF-IDF，y是10篇文档中“study”出现的TF-IDF。也可以认为有10辆汽车，x是千米/小时的速度，y是英里/小时的速度，等等。

第一步分别求x和y的平均值，然后对于所有的样例，都减去对应的均值。这里x的均值是1.81，y的均值是1.91，那么一个样例减去均值后即为（0.69,0.49），得到

[](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182110402112.png)

第二步，求特征协方差矩阵，如果数据是3维，那么协方差矩阵是

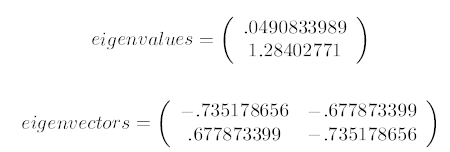
[](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182110404031.png)

这里只有x和y，求解得

[clip_image004[4]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182110417586.png)

 对角线上分别是x和y的方差，非对角线上是协方差。协方差大于0表示x和y若有一个增，另一个也增；小于0表示一个增，一个减；协方差为0时，两者独立。协方差绝对值越大，两者对彼此的影响越大，反之越小。

 第三步，求协方差的特征值和特征向量，得到

[](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182110413965.png)

 上面是两个特征值，下面是对应的特征向量，特征值0.0490833989对应特征向量为[clip_image007[4]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182110418392.png)，这里的特征向量都归一化为单位向量。

第四步，将特征值按照从大到小的顺序排序，选择其中最大的k个，然后将其对应的k个特征向量分别作为列向量组成特征向量矩阵。

这里特征值只有两个，我们选择其中最大的那个，这里是1.28402771，对应的特征向量是[clip_image009[6]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182110412504.png)。

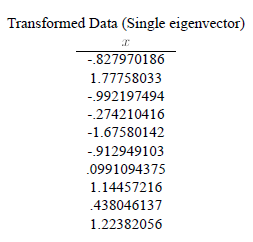
第五步，将样本点投影到选取的特征向量上。假设样例数为m，特征数为n，减去均值后的样本矩阵为DataAdjust(m\*n)，协方差矩阵是n\*n，选取的k个特征向量组成的矩阵为EigenVectors(n\*k)。那么投影后的数据FinalData为

[clip_image011[4]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182110424979.png)

这里是



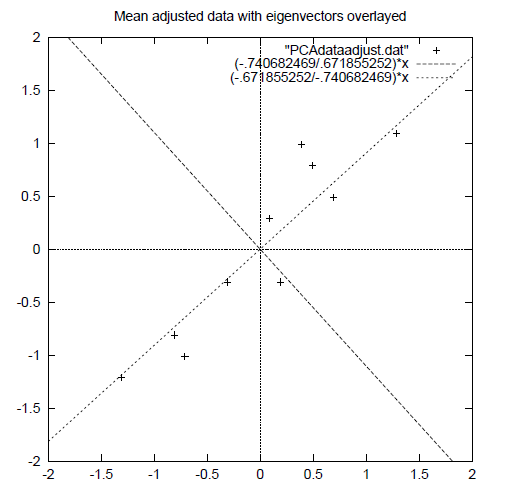
得到结果是

[](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182110427737.png)

这样，就将原始样例的n维特征变成了k维，这k维就是原始特征在k维上的投影。

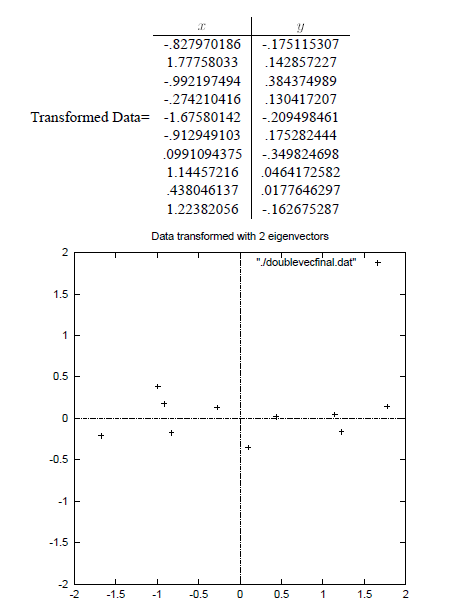
上面的数据可以认为是learn和study特征融合为一个新的特征叫做LS特征，该特征基本上代表了这两个特征。

上述过程有个图描述：

[](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182110437355.png)

正号表示预处理后的样本点，斜着的两条线就分别是正交的特征向量（由于协方差矩阵是对称的，因此其特征向量正交），最后一步的矩阵乘法就是将原始样本点分别往特征向量对应的轴上做投影。

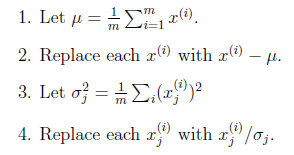
如果取的k=2，那么结果是

[](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182110442489.png)

这就是经过PCA处理后的样本数据，水平轴（上面举例为LS特征）基本上可以代表全部样本点。整个过程看起来就像将坐标系做了旋转，当然二维可以图形化表示，高维就不行了。上面的如果k=1，那么只会留下这里的水平轴，轴上是所有点在该轴的投影。

这样PCA的过程基本结束。在第一步减均值之后，其实应该还有一步对特征做方差归一化。比如一个特征是汽车速度（0到100），一个是汽车的座位数（2到6），显然第二个的方差比第一个小。因此，如果样本特征中存在这种情况，那么在第一步之后，求每个特征的标准差[clip_image016[6]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182110459599.png)，然后对每个样例在该特征下的数据除以[clip_image016[7]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182110469849.png)。

归纳一下，使用我们之前熟悉的表示方法，在求协方差之前的步骤是：

[](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182110465879.png)

其中[clip_image019[6]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182110473828.png)是样例，共m个，每个样例n个特征，也就是说[clip_image019[7]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182110485714.png)是n维向量。[clip_image021[4]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182110499551.png)是第i个样例的第j个特征。[clip_image023[4]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182110496662.png)是样例均值。[clip_image025[4]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182110506911.png)是第j个特征的标准差。

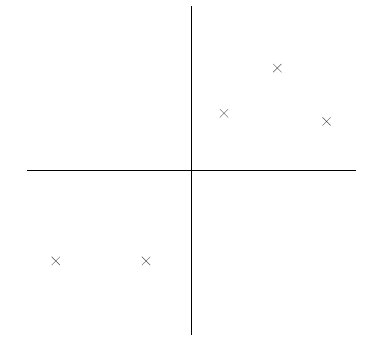
整个PCA过程貌似及其简单，就是求协方差的特征值和特征向量，然后做数据转换。但是有没有觉得很神奇，为什么求协方差的特征向量就是最理想的k维向量？其背后隐藏的意义是什么？整个PCA的意义是什么？

## 3.最大方差理论

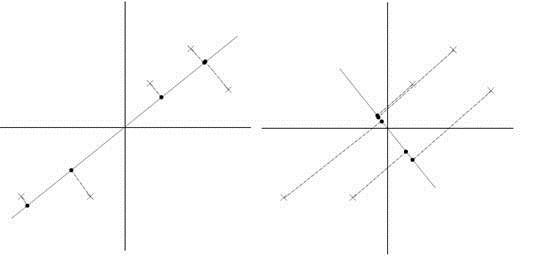
在信号处理中认为信号具有较大的方差，噪声有较小的方差，信噪比就是信号与噪声的方差比，越大越好。如前面的图，样本在横轴上的投影方差较大，在纵轴上的投影方差较小，那么认为纵轴上的投影是由噪声引起的。

因此我们认为，最好的k维特征是将n维样本点转换为k维后，每一维上的样本方差都很大。

比如下图有5个样本点：（已经做过预处理，均值为0，特征方差归一）

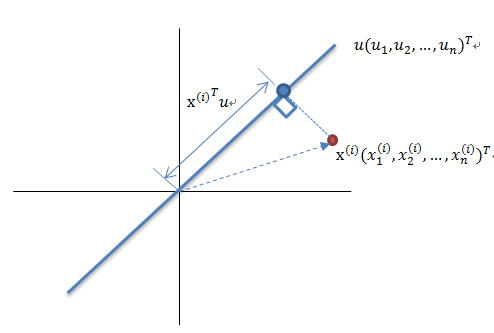
[](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182110512941.png)

下面将样本投影到某一维上，这里用一条过原点的直线表示（前处理的过程实质是将原点移到样本点的中心点）。

[](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182110514304.jpg)

假设我们选择两条不同的直线做投影，那么左右两条中哪个好呢？根据我们之前的方差最大化理论，左边的好，因为投影后的样本点之间方差最大。

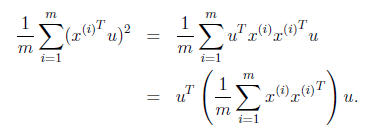
这里先解释一下投影的概念：

[](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182110524172.png)

红色点表示样例[clip_image037[14]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182110531216.png)，蓝色点表示[clip_image037[15]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182110548957.png)在u上的投影，u是直线的斜率也是直线的方向向量，而且是单位向量。蓝色点是[clip_image037[16]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182110559971.png)在u上的投影点，离原点的距离是[clip_image039[4]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182110554921.png)（即[clip_image030[4]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182110567397.png)或者[clip_image041[4]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182110566458.png)）由于这些样本点（样例）的每一维特征均值都为0，因此投影到u上的样本点（只有一个到原点的距离值）的均值仍然是0。

回到上面左右图中的左图，我们要求的是最佳的u，使得投影后的样本点方差最大。

由于投影后均值为0，因此方差为：

[](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182110572804.png)

中间那部分很熟悉啊，不就是样本特征的协方差矩阵么（[clip_image037[17]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/20110418211058787.png)的均值为0，一般协方差矩阵都除以m-1，这里用m）。

用[clip_image044[10]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182110582116.png)来表示[clip_image046[4]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182110592606.png)，[clip_image048[6]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182110591145.png)表示[clip_image050[4]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182111006095.png)，那么上式写作

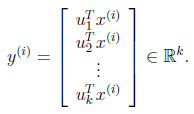
[clip_image052[4]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182111009443.png)

由于u是单位向量，即[clip_image054[4]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182111004393.png)，上式两边都左乘u得，[clip_image056[4]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182111017392.png)

即[clip_image058[4]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182111018754.png)

We got it！[clip_image044[11]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182111026180.png)就是[clip_image048[7]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/20110418211103782.png)的特征值，u是特征向量。最佳的投影直线是特征值[clip_image044[12]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182111038523.png)最大时对应的特征向量，其次是[clip_image044[13]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182111043125.png)第二大对应的特征向量，依次类推。

因此，我们只需要对协方差矩阵进行特征值分解，得到的前k大特征值对应的特征向量就是最佳的k维新特征，而且这k维新特征是正交的。得到前k个u以后，样例[clip_image037[18]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182111052502.png)通过以下变换可以得到新的样本。

[](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182111054945.png)

其中的第j维就是[clip_image037[19]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182111063799.png)在[clip_image061[4]](http://images.cnblogs.com/cnblogs_com/jerrylead/201104/201104182111074812.png)上的投影。

通过选取最大的k个u，使得方差较小的特征（如噪声）被丢弃。