****协方差，相关系数，PCA旋转****

****统计学的基本概念****

学过概率统计的孩子都知道，统计里最基本的概念就是样本的均值，方差，或者再加个标准差。首先我们给你一个含有n个样本的集合，依次给出这些概念的公式描述，这些高中学过数学的孩子都应该知道吧，一带而过。

[协方差矩阵的概念及matlab计算](http://photo.blog.sina.com.cn/showpic.html" \l "blogid=4aa4593d01012am3&url=http://s8.sinaimg.cn/orignal/4aa4593d4b5d3b510c6b7" \t "http://blog.sina.com.cn/s/_blank)  
****均值：****[协方差矩阵的概念及matlab计算](http://photo.blog.sina.com.cn/showpic.html" \l "blogid=4aa4593d01012am3&url=http://s8.sinaimg.cn/orignal/4aa4593d4b5d3b912eef7" \t "http://blog.sina.com.cn/s/_blank)  
  
  
****标准差：****[协方差矩阵的概念及matlab计算](http://photo.blog.sina.com.cn/showpic.html" \l "blogid=4aa4593d01012am3&url=http://s11.sinaimg.cn/orignal/4aa4593d078952c55192a" \t "http://blog.sina.com.cn/s/_blank)  
  
  
****方差：****[协方差矩阵的概念及matlab计算](http://photo.blog.sina.com.cn/showpic.html" \l "blogid=4aa4593d01012am3&url=http://s15.sinaimg.cn/orignal/4aa4593d4b5d3bc21467e" \t "http://blog.sina.com.cn/s/_blank)

很 显然，均值描述的是样本集合的中间点，它告诉我们的信息是很有限的，而标准差给我们描述的则是样本集合的各个样本点到均值的距离之平均。以这两个集合为 例，[0，8，12，20]和[8，9，11，12]，两个集合的均值都是10，但显然两个集合差别是很大的，计算两者的标准差，前者是8.3，后者是 1.8，显然后者较为集中，故其标准差小一些，标准差描述的就是这种“散布度”。之所以除以n-1而不是除以n，是因为这样能使我们以较小的样本集更好的 逼近总体的标准差，即统计上所谓的“无偏估计”。而方差则仅仅是标准差的平方。

****为什么需要协方差？****

上 面几个统计量看似已经描述的差不多了，但我们应该注意到，**标准差和方差一般是用来描述一维数据的**，但现实生活我们常常遇到含有多维数据的数据集，最简单的 大家上学时免不了要统计多个学科的考试成绩。面对这样的数据集，我们当然可以按照每一维独立的计算其方差，但是通常我们还想了解更多，比如，一个男孩子的 猥琐程度跟他受女孩子欢迎程度是否存在一些联系啊，嘿嘿~协方差就是这样一种用来度量两个随机变量关系的统计量，我们可以仿照方差的定义：

[协方差矩阵的概念及matlab计算](http://photo.blog.sina.com.cn/showpic.html" \l "blogid=4aa4593d01012am3&url=http://s5.sinaimg.cn/orignal/4aa4593d4b5d3bf8af144" \t "http://blog.sina.com.cn/s/_blank)

来度量各个维度偏离其均值的程度，标准差可以这么来定义：

[协方差矩阵的概念及matlab计算](http://photo.blog.sina.com.cn/showpic.html" \l "blogid=4aa4593d01012am3&url=http://s16.sinaimg.cn/orignal/4aa4593d4b5d3c217c2ef" \t "http://blog.sina.com.cn/s/_blank)  
  
协方差的结果有什么意义呢？如果结果为正值，则说明两者是正相关的(从协方差可以引出“相关系数”的定义)，也就是说一个人越猥琐就越受女孩子欢迎，嘿嘿，那必须的~结果为负值就说明负相关的，越猥琐女孩子越讨厌，可能吗？如果为0，也是就是统计上说的“相互独立”。

从协方差的定义上我们也可以看出一些显而易见的性质，如：

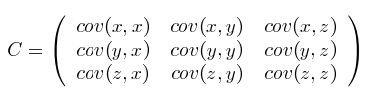
[协方差矩阵的概念及matlab计算](http://photo.blog.sina.com.cn/showpic.html" \l "blogid=4aa4593d01012am3&url=http://s10.sinaimg.cn/orignal/4aa4593d4b5d3c580e689" \t "http://blog.sina.com.cn/s/_blank)  
  
****协方差多了就是协方差矩阵****

上一节提到的猥琐和受欢迎的问题是典型二维问题，而协方差也只能处理二维问题，那维数多了自然就需要计算

[协方差矩阵的概念及matlab计算](http://photo.blog.sina.com.cn/showpic.html" \l "blogid=4aa4593d01012am3&url=http://s9.sinaimg.cn/orignal/4aa4593d4b5d3c9c812e8" \t "http://blog.sina.com.cn/s/_blank)

多个协方差，比如n维的数据集就需要计算个协方差，那自然而然的我们会想到使用矩阵来组织这些数据。给出协方差矩阵的定义：

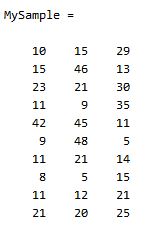
[协方差矩阵的概念及matlab计算](http://photo.blog.sina.com.cn/showpic.html" \l "blogid=4aa4593d01012am3&url=http://s2.sinaimg.cn/orignal/4aa4593d4b5d3cb9ef231" \t "http://blog.sina.com.cn/s/_blank)  
  
这个定义还是很容易理解的，我们可以举一个简单的三维的例子，假设数据集有三个维度，则协方差矩阵为

[](http://photo.blog.sina.com.cn/showpic.html" \l "blogid=4aa4593d01012am3&url=http://s11.sinaimg.cn/orignal/4aa4593d4b5d3cf7f22fa" \t "http://blog.sina.com.cn/s/_blank)  
可见，协方差矩阵是一个对称的矩阵，而且对角线是各个维度上的方差。

****Matlab协方差实战****

上面涉及的内容都比较容易，协方差矩阵似乎也很简单，但实战起来就很容易让人迷茫了。必须要明确一点，****协方差矩阵计算的是不同维度之间的协方差，而不是不同样本之间的。****这个我将结合下面的例子说明，以下的演示将使用Matlab，为了说明计算原理，不直接调用Matlab的cov函数(蓝色部分为Matlab代码)。

首先，随机产生一个10\*3维的整数矩阵作为样本集，10为样本的个数，3为样本的维数。  
MySample = fix(rand(10,3)\*50)

[](http://photo.blog.sina.com.cn/showpic.html" \l "blogid=4aa4593d01012am3&url=http://s15.sinaimg.cn/orignal/4aa4593d4b5d3d3ea20ae" \t "http://blog.sina.com.cn/s/_blank)

根据公式，计算协方差需要计算均值，那是按行计算均值还是按列呢，我一开始就老是困扰这个问题。前面我们也特别强调了，协方差矩阵是计算不同维度间的协方差，要时刻牢记这一点。样本矩阵的每行是一个样本，每列为一个维度，所以我们要****按列计算均值****。为了描述方便，我们先将三个维度的数据分别赋值：

dim1 = MySample(:,1);  
dim2 = MySample(:,2);  
dim3 = MySample(:,3);

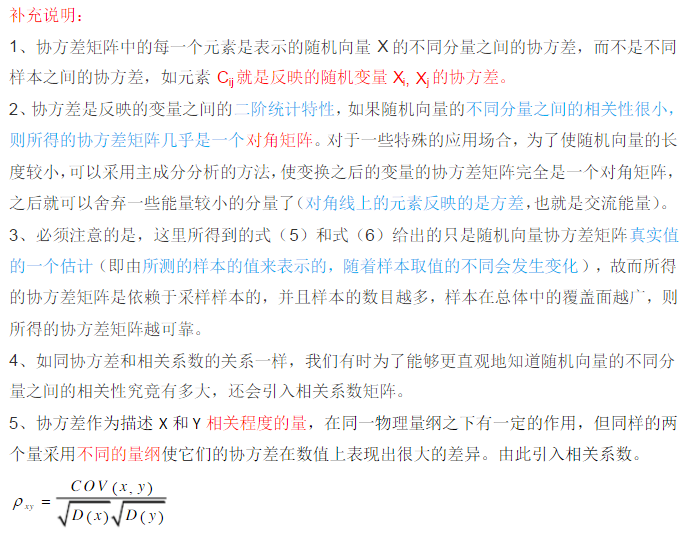
计算dim1与dim2，dim1与dim3，dim2与dim3的协方差：

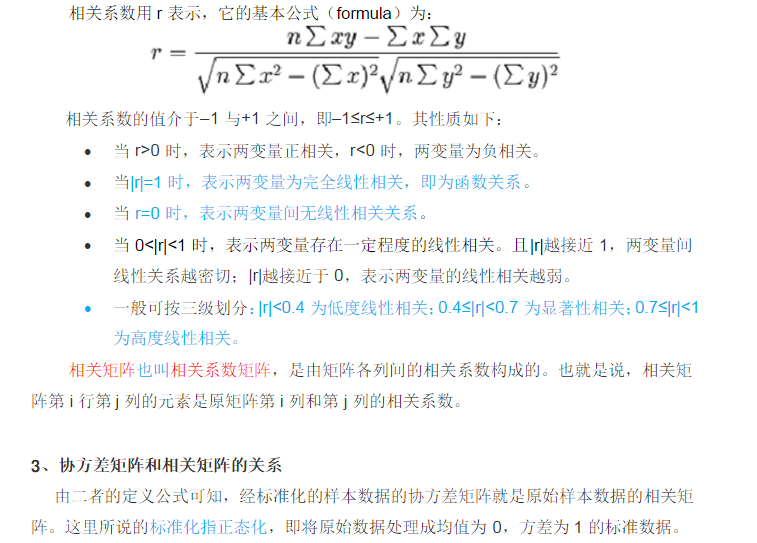
sum( (dim1-mean(dim1)) .\* (dim2-mean(dim2)) ) / ( size(MySample,1)-1 ) % 得到  74.5333  
sum( (dim1-mean(dim1)) .\* (dim3-mean(dim3)) ) / ( size(MySample,1)-1 ) % 得到  -10.0889  
sum( (dim2-mean(dim2)) .\* (dim3-mean(dim3)) ) / ( size(MySample,1)-1 ) % 得到  -10\*\*\*000

搞清楚了这个后面就容易多了，协方差矩阵的对角线就是各个维度上的方差，下面我们依次计算：

std(dim1)^2 % 得到   108.3222  
std(dim2)^2 % 得到   260.6222  
std(dim3)^2 % 得到  94.1778

这样，我们就得到了计算协方差矩阵所需要的所有数据，调用Matlab自带的cov函数进行验证：

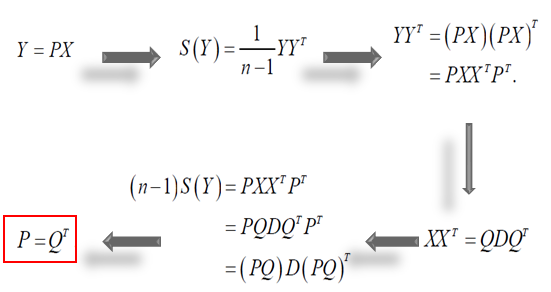




PCA其实就是找到一个选择变换的矩阵，使得原始空间的数据投影到变换后的空间具有较大的方差（方差越大，所包含的信息越大）。PCA的全部工作简单点说，就是对原始的空间中顺序地找一组相互正交的坐标轴，第一个轴是使得方差最大的，第二个轴是在与第一个轴正交的平面中使得方差最大的，第三个轴是在与第1、2个轴正交的平面中方差最大的，这样假设在N维空间中，我们可以找到N个这样的坐标轴，我们取前r个去近似这个空间，这样就从一个N维的空间压缩到r维的空间了，但是我们选择的r个坐标轴能够使得空间的压缩使得数据的损失最小。

3、由上述可知，变换后的数据在所在空间没有了相关关系，因为他们是正交的或者是一维的，因此可以认为变换后的矩阵其协方差矩阵或者相关系数矩阵是一个对角阵，即除对角上的数外，其余位置的数均为0，PCA旋转变换矩阵推导的想法便是基于此。

4、假设矩阵X，有变换矩阵P，变换后的矩阵为Y=PX，则



其中IMG_257为特征分解。

5、这个变换矩阵就是X的协防差矩阵的特征向量Q的转置，到此这个PCA的旋转变换矩阵已经找到。