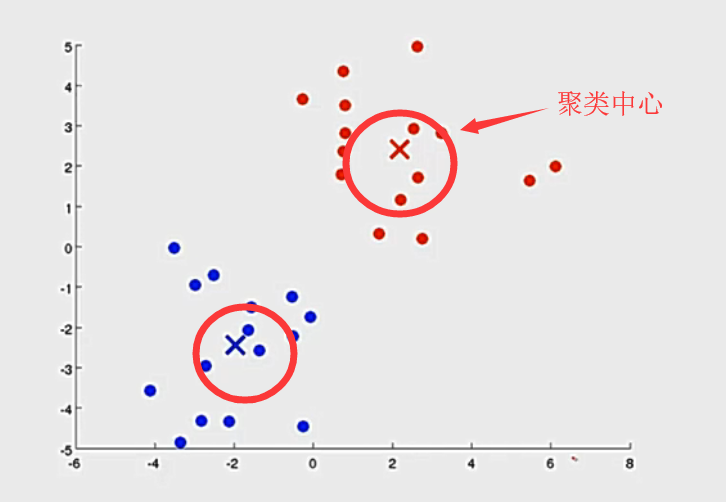
1、K-means（K均值）

（1）聚类中心：

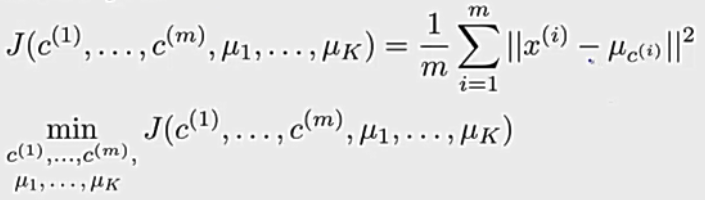


（2）原理：

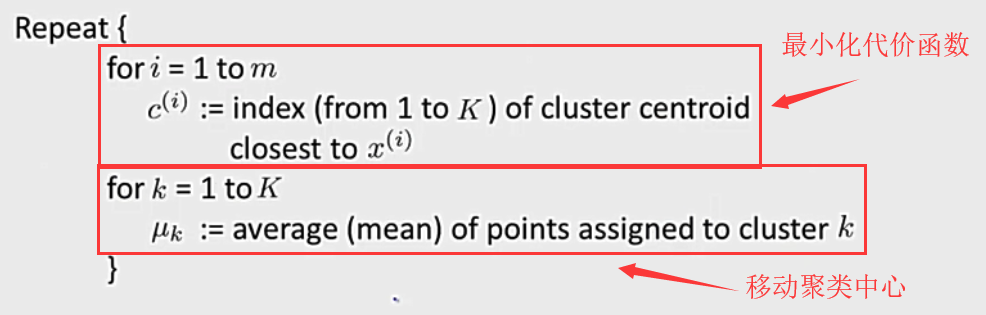
1、参数

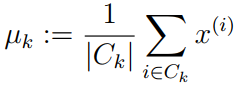


2、代价函数



3、执行步骤**（需要随机初始化聚类中心）**





找到最近的中心：

for i=1:length(idx)

distanse = pdist2(centroids,X(i,:));

% compute the distance(K,1) pdist2 is a good function

[C,idx(i)]=min(distanse); % find the minimum

end

移动聚类中心：

for i=1:K

centroids(i,:) = mean( X( find(idx==i) , :) ); %

end

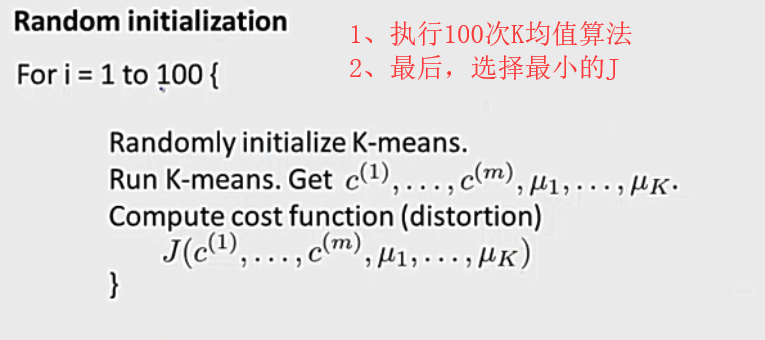
4、随机初始化

方式一：聚类中心个数K<样本数量m

方式二：随机选择聚类中心个数K

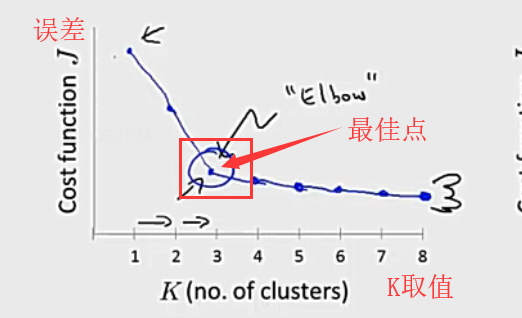
方式三：任意选取样本，作为聚类中心。

步骤：



5、选取最佳的聚类中心数量K：（肘部法则）

但该方法不一定有效果。



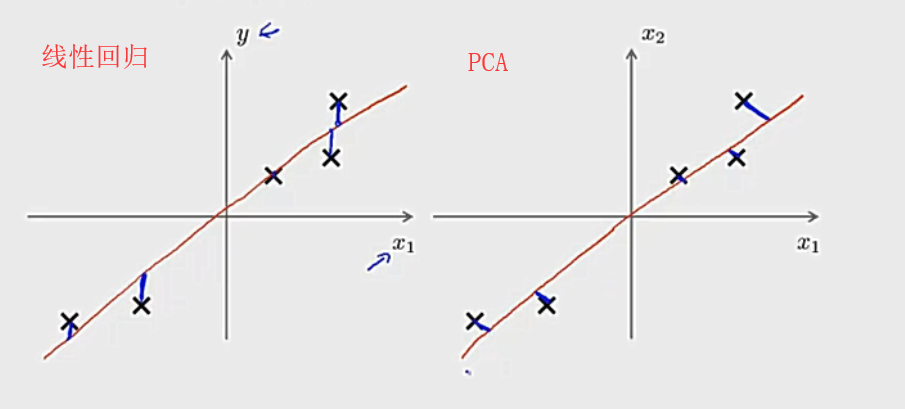
2、降维

（1）PCA：主成分分析法**（无监督学习）**

1、思想：找到一个平面（直线）等，使得数据投影在上面的距离最小

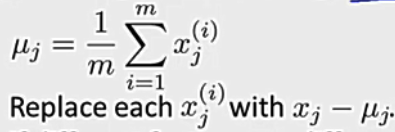
（投影距离：指的是正交投影，区别于线性回归）

2、PCA于线性回归的区别：



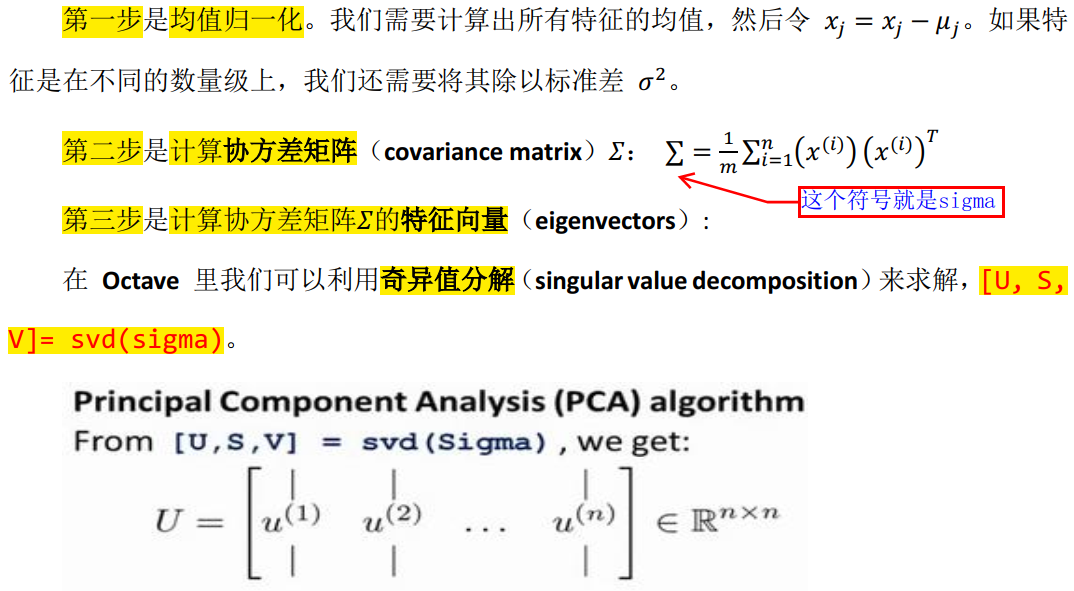
3、需要执行的步骤：

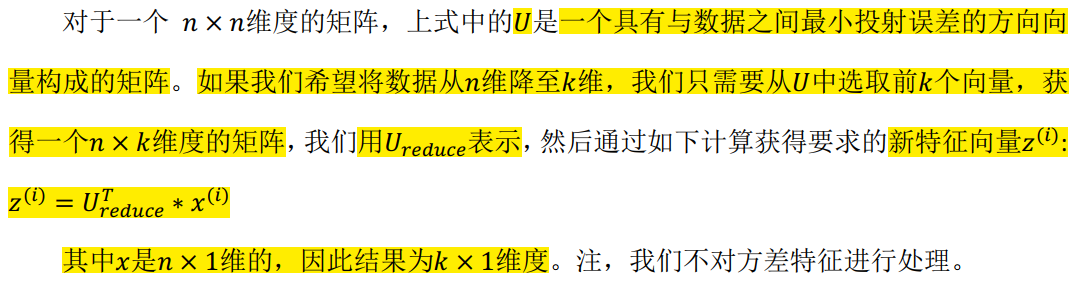
1）均值归一化：使得所有数据的均值为0



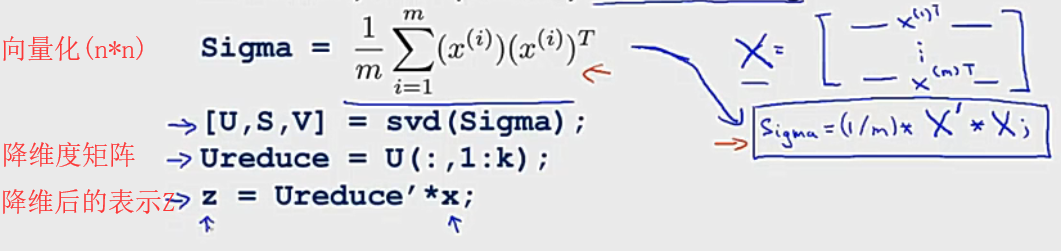
2）特征缩放：

4、降维步骤：



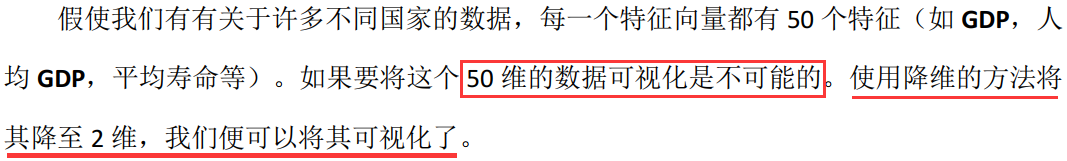


5、PCA公式：



**（PS：n为特征向量的维数）**

6、可视化



7、PCA使用：  
 1）可以提高算法的速度：由于特征值数量少了，计算速度快；

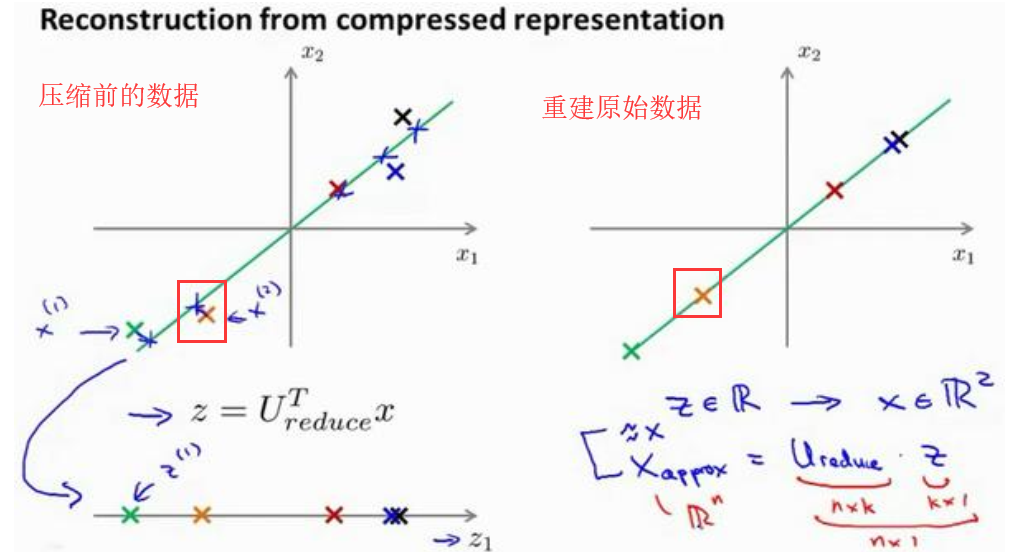
压缩数据；减小对内存、硬盘空间的使用。

2）不可以用于防止过拟合：即使特征数量可以减少，但是PCA不涉及数据标签y，有可能会损失有价值的数据。

（2）压缩重现



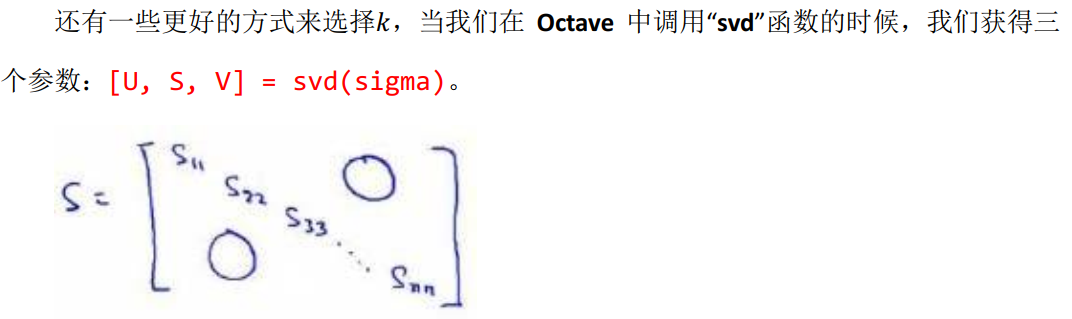
具体如下图所示：



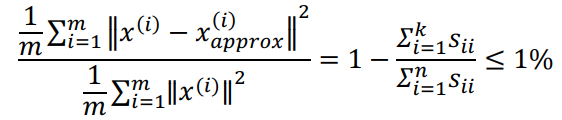
（3）主成分变量的数量（K）选择

1、方差保留的百分比

使用svd计算即可：平均均方误差/训练集均方误差



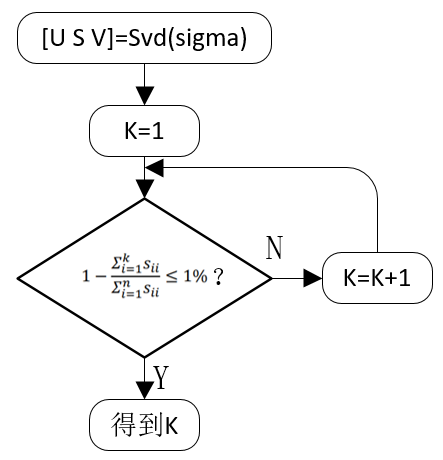
Eg：方差保留99%



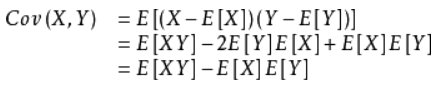
2、K的选择

选择上述满足方差保留比的K值即可。

寻找K的流程图如下：



补充知识：

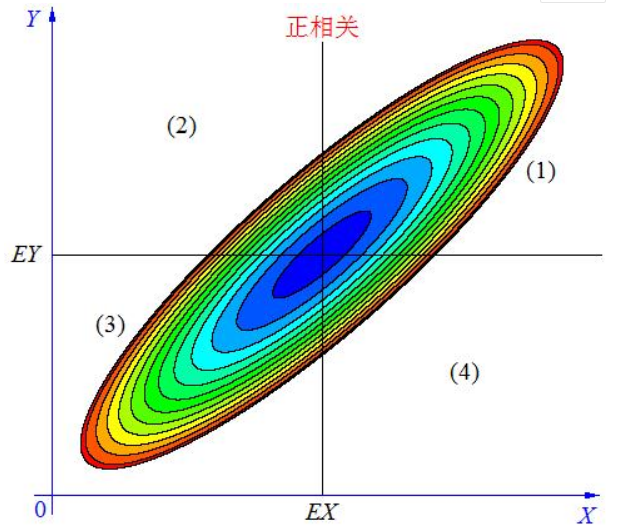
1、协方差：**协方差矩阵是一个对称的矩阵，且主对角线是各个维度上的方差。**

（1）含义：描述两个数据之间的关系

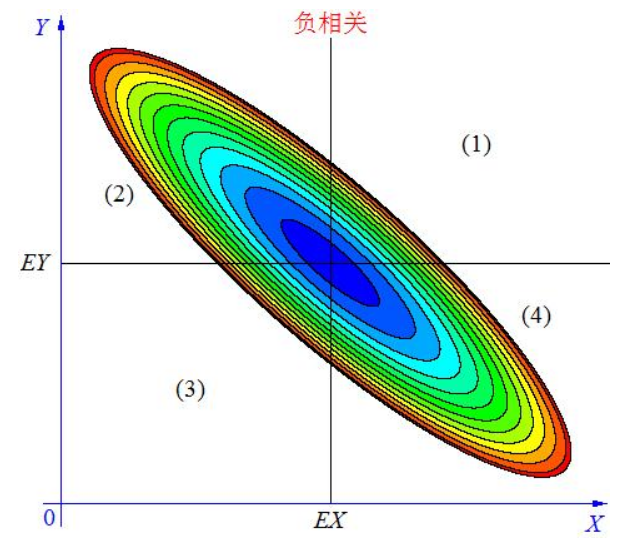
度量各个维度偏离其均值的程度。协方差的值如果为正值，则说明两者是正相关的(从协方差可以引出“相关系数”的定义)，结果为负值就说明负相关的，如果为0，也是就是统计上说的“相互独立”。

理解协方差矩阵的关键就在于牢记它计算的是不同维度之间的协方差，而不是不同样本之间。

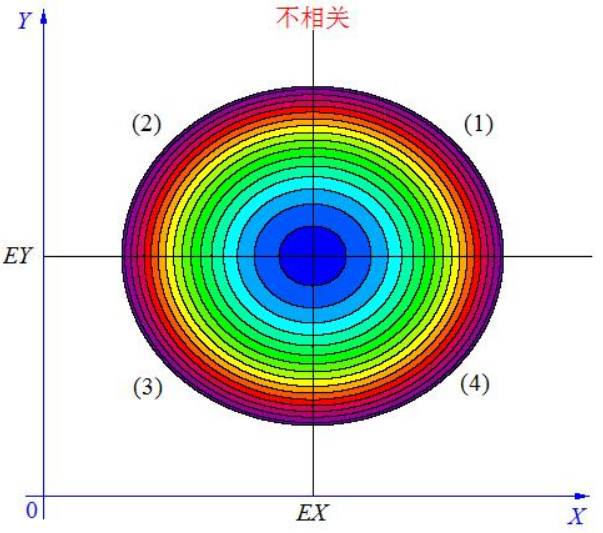
附：1：正相关



2：负相关



3：不相关



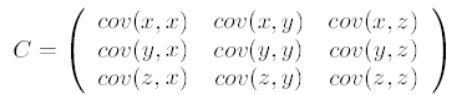
2、相关系数：



1、也可以反映两个变量变化时是同向还是反向，如果同向变化就为正，反向变化就为负。

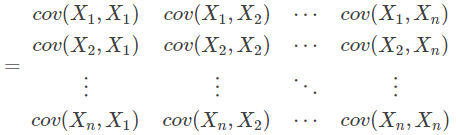
2、由于它是标准化后的协方差，因此更重要的特性来了：它消除了两个变量变化幅度的影响，而只是单纯反应两个变量每单位变化时的相似程度。

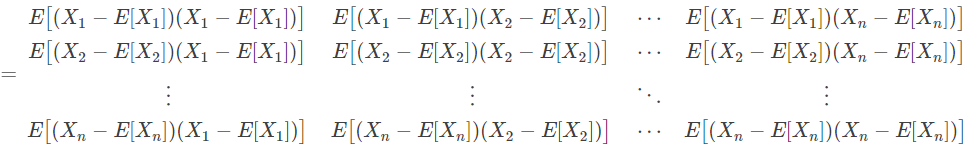
3、协方差阵：

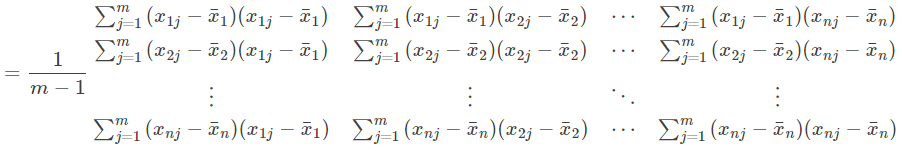


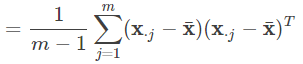
附：协方差矩阵推导：











但之前使用了均值归一化，故此处的均值=0，因此有下面的式子：