**第6章 统计分析建模方法**

**一、复习题**

1、主成分分析法（方法1-基于协方差矩阵方法）

* 由X的协方差阵Σx，求出其特征根；
* 求出分别所对应的特征向量；
* 计算累积贡献率，给出恰当的主成分个数；
* 计算所选出的k个主成分的得分。将原始数据的中心化值。

2、主成分分析法（方法2-基于相关系数矩阵方法）

* 计算相关系数矩阵；
* 计算特征值与特征向量；
* 计算主成分贡献率及累计贡献率；
* 计算主成分载荷；
* 各主成分的得分。

3、聚类分析法（方法1-直接聚类法）

原理：先把各个分类对象单独视为一类，然后根据距离最小的原则，依次选出一对分类对象，并成新类。

* 性质1：如果其中一个分类对象已归于一类，则把另一个也归入该类；
* 性质2：如果一对分类对象正好属于已归的两类，则把这两类并为一类。

每一次归并，都划去该对象所在的列与列序相同的行。经过n-1次就可以把全部分类对象归为一类，根据归并的先后顺序作出聚类谱系图。

4、聚类分析法（方法2- k-means聚类算法）

* 选择一个含有随机选择样本的k个簇的初始划分，计算这些簇的质心。
* 根据距离把剩余的每个样本分配到距离它最近的簇质心的一个划分。
* 计算被分配到每个簇的样本的均值向量，作为新的簇的质心。
* 重复2,3步，直到k个簇的质心点不再发生变化或准则函数收敛。
* **注：聚类分析中要考虑相似性度量（** Lk范数**）和数据处理方法（4种方法）。**

**二、主成分分析方法（PCA**

**1.基础概念**

* 多变量问题是经常会遇到的。变量太多，无疑会增加分析问题的难度与复杂性.
* 问题：能否用较少的新变量尽可能多地保留原来较多的变量所反映的信息？
* 事实上，这种想法是可以实现的。因为在许多实际问题中，多个变量之间是具有一定的相关关系的。
* 主成分分析原理: 是把原来多个变量化为少数几个综合指标的一种统计分析方法，从数学角度来看，这是一种降维处理技术。
* 主成分分析方法就是综合处理这种问题的一种强有力的方法。
* 在力求数据信息丢失最少的原则下，对高维的变量空间降维，即研究指标体系的少数几个线性组合，并且这几个线性组合所构成的综合指标将尽可能多地保留原来指标变异方面的信息。这些综合指标就称为主成分。要讨论的问题是：

**(1) 如何作主成分分析?**

* **当分析中所选择的变量具有不同的量纲，变量水平差异很大，应该选择基于相关系数矩阵的主成分分析。**

**（2） 如何选择几个主成分。**

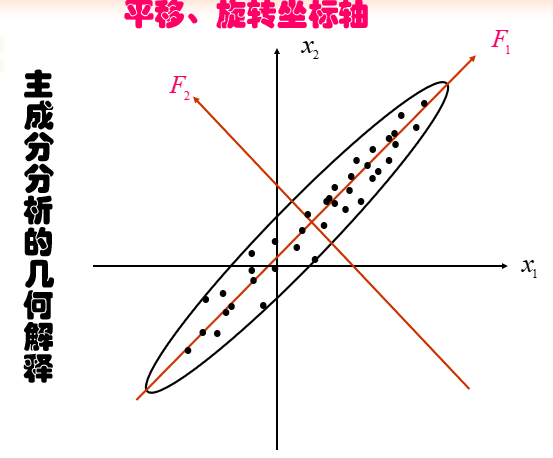
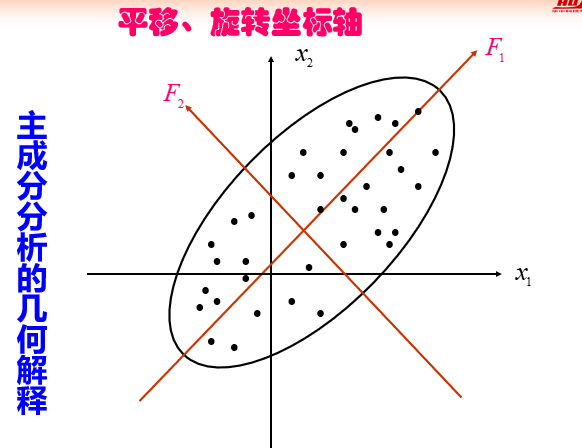
* **主成分分析的目的是简化变量，一般情况下主成分的个数应该小于原始变量的个数。关于保留几个主成分，应该权衡主成分个数和保留的信息。**

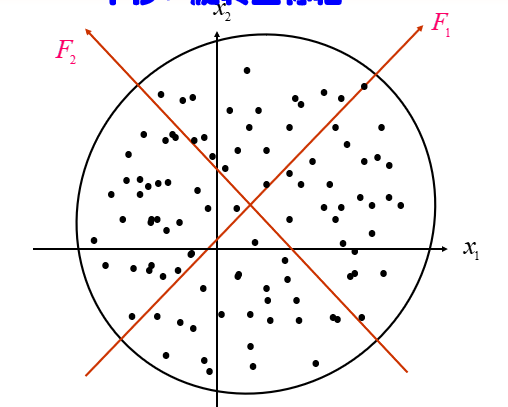
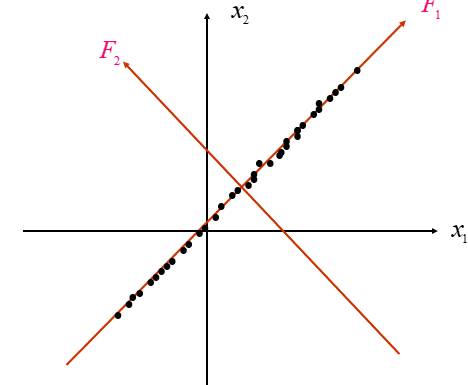
**（3）如何解释主成分所包含的几何意义或经济意义或其它。**

主成分分析就是试图在力保数据信息丢失最少的原则下，对这种多变量的数据表进行最佳综合简化，也就是说，对高维变量空间进行降维处理。很显然，识辨系统在一个低维空间要比在一个高维空间容易得多。

**2.几何解释**

先假定数据只有二维，即只有两个变量，它们由横坐标和纵坐标所代表；因此每个观测值都有相应于这两个坐标轴的两个坐标值；如果这些数据形成一个椭圆形状的点阵（这在变量的二维正态的假定下是可能的）



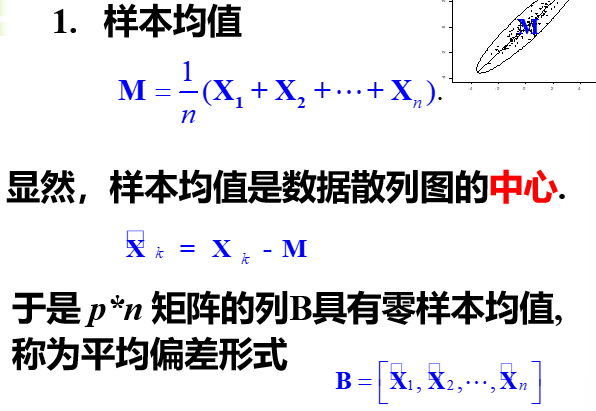
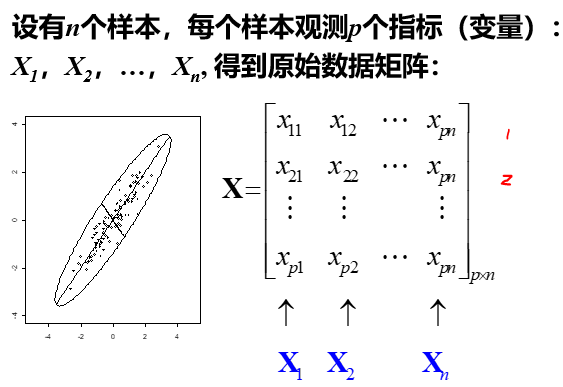


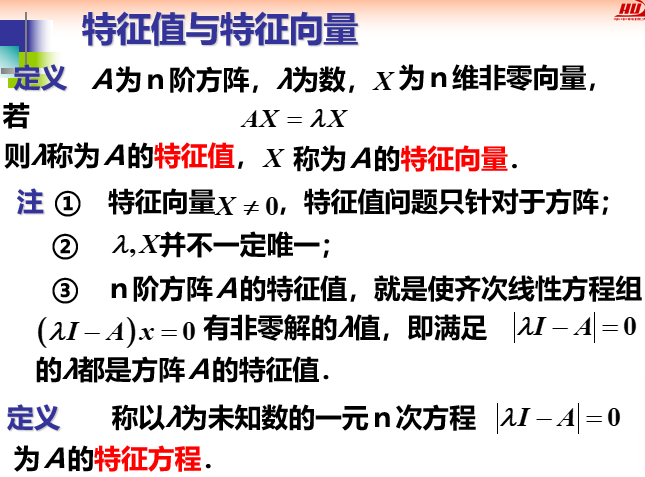
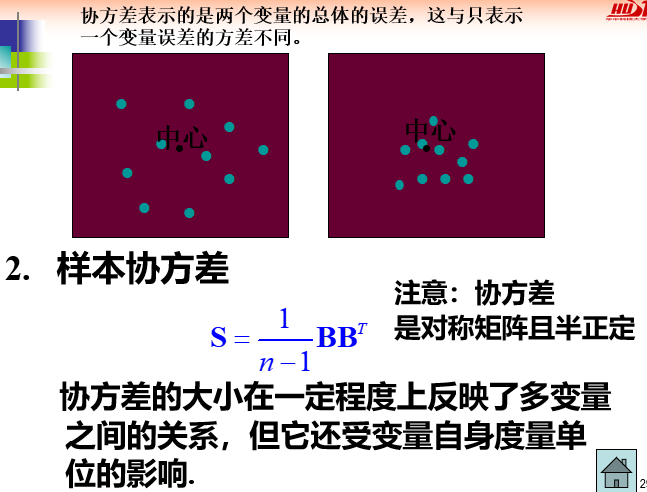
**PCA: 进一步解释**

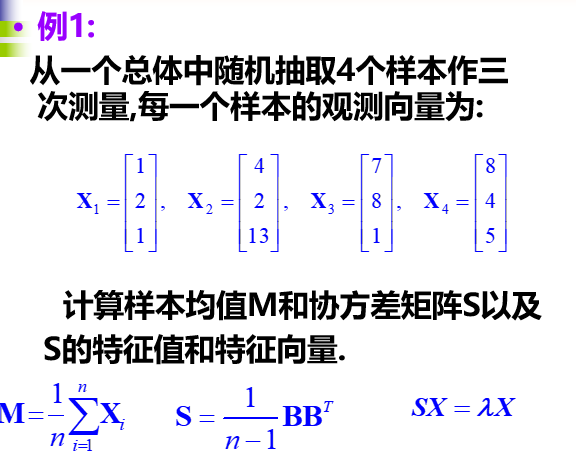
椭圆有一个长轴和一个短轴。在短轴方向上，数据变化很少；在极端的情况，短轴如果退化成一点，那只有在长轴的方向才能够解释这些点的变化了；这样，由二维到一维的降维就自然完成了。

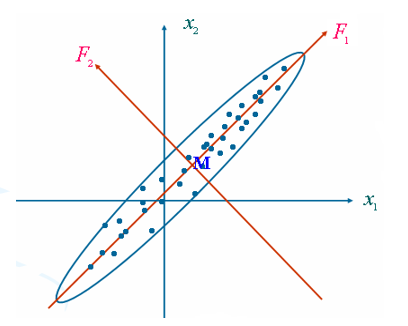
* 当坐标轴和椭圆的长短轴平行，那么代表长轴的变量就描述了数据的主要变化，而代表短轴的变量就描述了数据的次要变化。
* 但是，坐标轴通常并不和椭圆的长短轴平行。因此，需要寻找椭圆的长短轴，并进行变换，使得新变量和椭圆的长短轴平行。
* 如果长轴变量代表了数据包含的大部分信息，就用该变量代替原先的两个变量（舍去次要的一维），降维就完成了。
* 椭圆（球）的长短轴相差越大，降维也越有道理。
* 对于多维变量的情况和二维类似，也有高维的椭球，只不过无法直观地看见罢了。
* 首先把高维椭球的主轴找出来，再用代表大多数数据信息的最长的几个轴作为新变量；这样，主成分分析就基本完成了。
* 注意，和二维情况类似，高维椭球的主轴也是互相垂直的。这些互相正交的新变量是原先变量的线性组合，叫做主成分(principal component)。
* 正如二维椭圆有两个主轴，三维椭球有三个主轴一样，有几个变量，就有几个主成分。
* **选择越少的主成分，降维就越好。**
* **什么是标准呢？**
* **——这些被选的主成分所代表的主轴的长度之和占了主轴长度总和的大部分。有些文献建议，所选的主轴总长度占所有主轴长度之和的大约85%即可，其实，这只是一个大体的说法；具体选几个，要看实际情况而定。**

**3. 均值和协方差 特征值和特征向量**

****



 为了方便，我们在二维空间中讨论主成分的几何意义。 设有n个样本，每个样本有两个观测变量xl和x2，在由变量xl和x2 所确定的二维平面中，n个样本点所散布的情况如椭圆状。由图可以看出这n个样本点无论是沿着xl 轴方向或x2轴方向都具有较大的离散性，其离散的程度可以分别用观测变量xl 的方差和x2 的方差定量地表示。显然，如果只考虑xl和x2 中的任何一个，那么包含在原始数据中的信息将会有较大的损失。

如果我们将xl 轴和x2轴先平移，再同时按逆时针方向旋转θ角度，得到新坐标轴Fl和F2。Fl和F2是两个新变量。

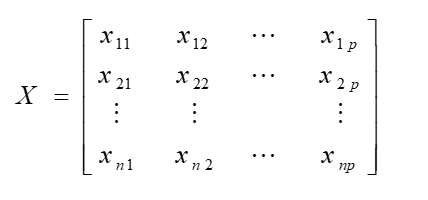
Fl，F2除了可以对包含在Xl，X2中的信息起着浓缩作用之外，还具有不相关的性质，这就使得在研究复杂的问题时避免了信息重叠所带来的虚假性。二维平面上的个点的方差大部分都归结在Fl轴上，而F2轴上的方差很小。Fl和F2称为原始变量x1和x2的综合变量。

F简化了系统结构，抓住了主要矛盾。

**4主成分分析**

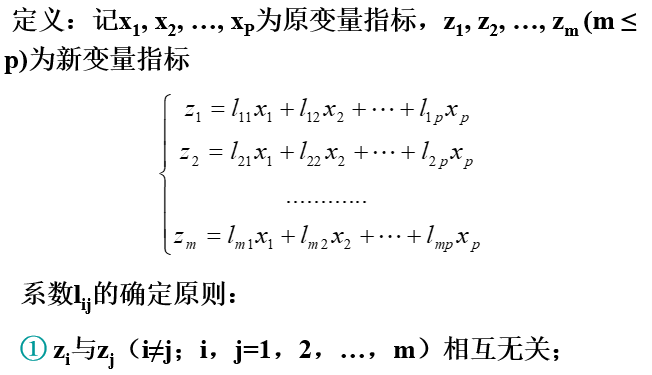
**基本原理：**

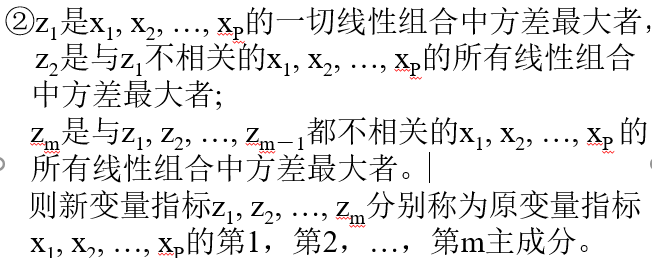
假定有*n*个数据样本，每个样本共有*p*个变量，构成一个*n*×*p*阶的数据矩阵



当*p*较大时，在*p*维空间中考察问题比较麻烦。为了克服这一困难，就需要进行降维处理。

**——用较少的几个综合指标代替原来较多的变量指标，而且使这些较少的综合指标既能尽量多地反映原来较多变量指标所反映的信息，同时它们之间又是彼此独立的。**



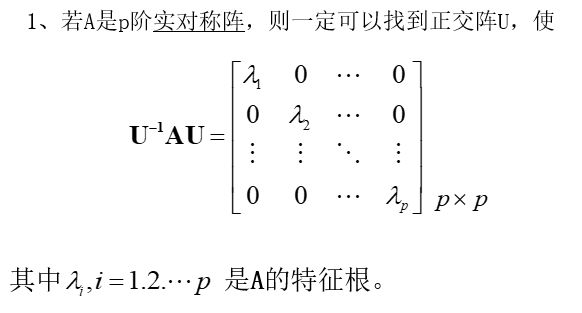
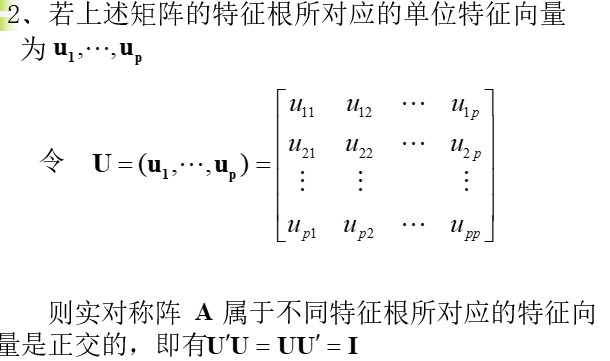


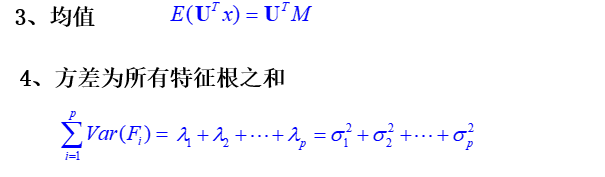
**主成分分析的实质就是确定原来变量xj (j=1, 2, …, p)在诸主成分zi (i=1, 2, …, m)上的荷载lij (i=1, 2, …, m; j=1, 2, …, p)。**

**从数学上可以证明，它们分别是相关矩阵m个较大的特征值所对应的特征向量。**

**PCA的性质：**

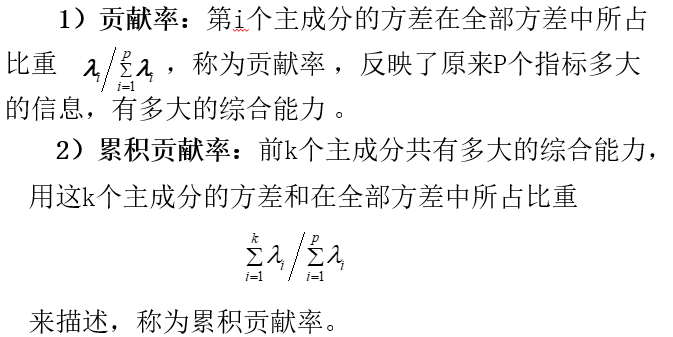
**1 两个线性代数的结论**

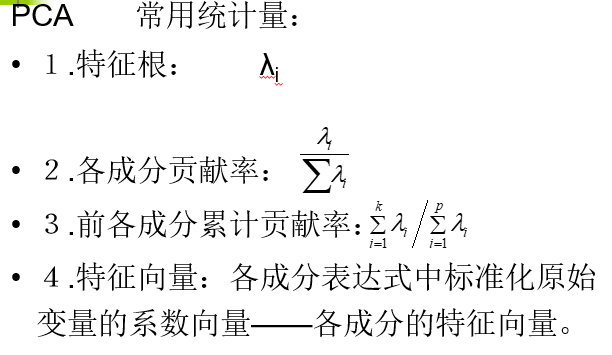
 

 说明主成分分析把P个随机变量的总方差分解成为P个不相关的随机变量的方差之和。

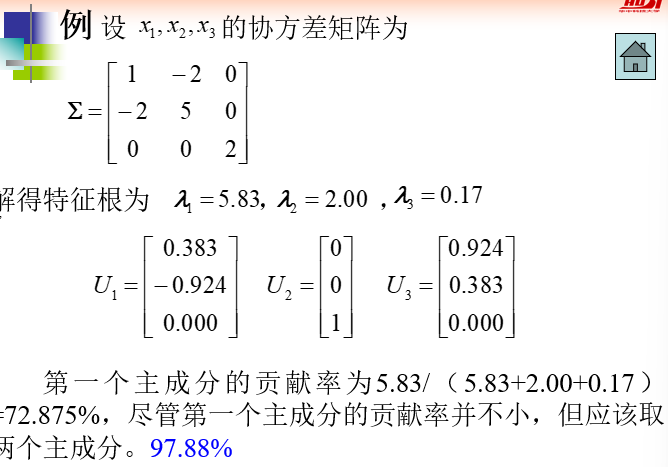
协方差矩阵Σ的对角线上的元素之和等于特征根之和。

**精度分析：**



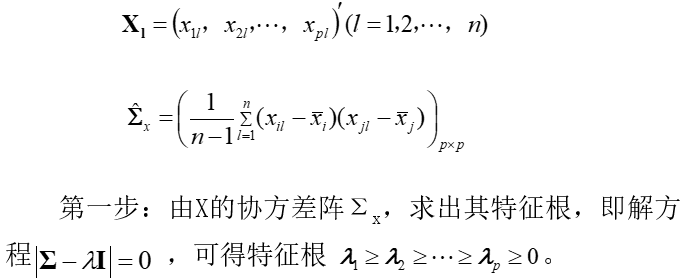


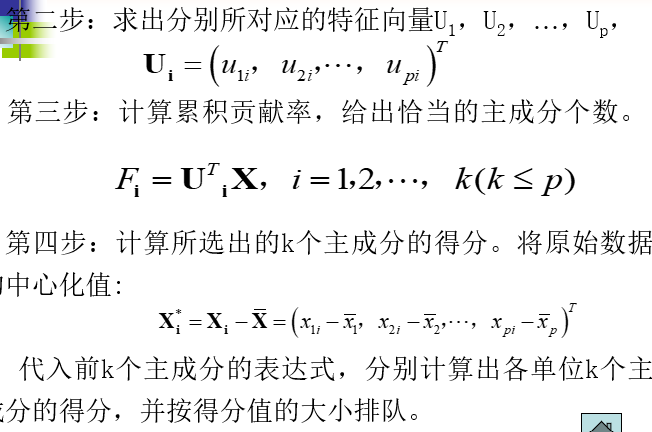
**我们进行主成分分析的目的之一是希望用尽可能少的主成分F1，F2，…，Fk（k≤p）代替原来的P个指标。到底应该选择多少个主成分，在实际工作中，主成分个数的多少取决于能够反映原来变量80%以上的信息量为依据，即当累积贡献率≥80%时的主成分的个数就足够了。最常见的情况是主成分为2到3个。**



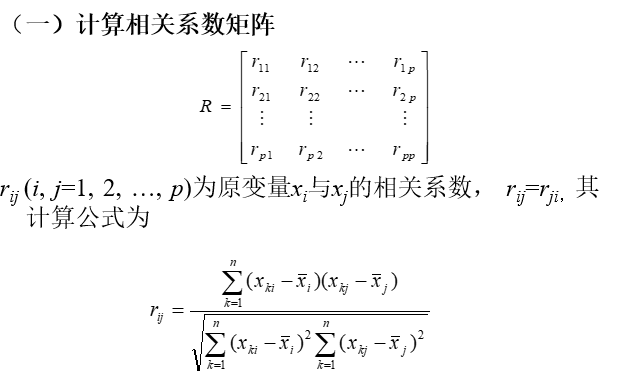
**主成分分析的步骤**

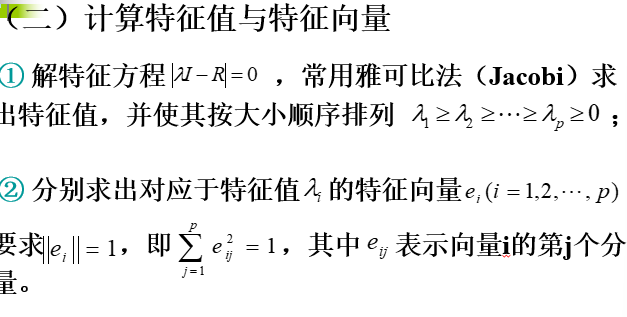
1. **基于协方差矩阵方法**

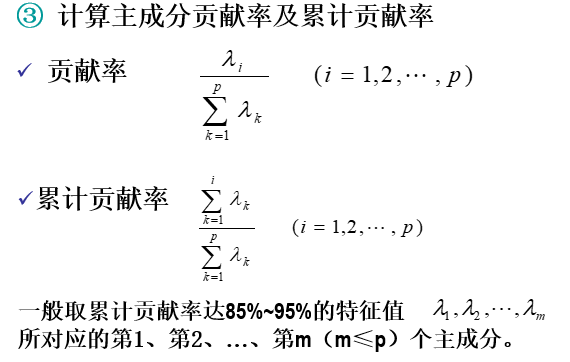
****

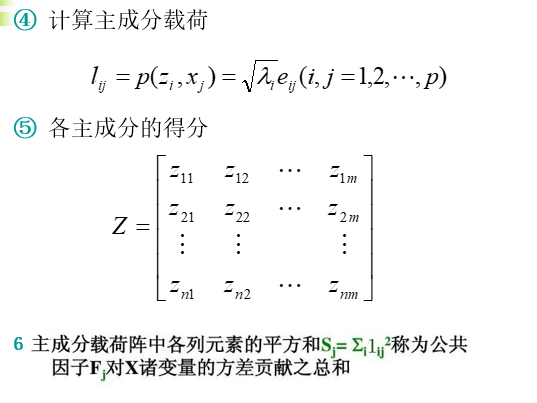
****

1. **基于相关系数矩阵方法**







****

**例题：**

**例：应收账款是指企业因对外销售产品、材料、提供劳务及其它原因，应向购货单位或接受劳务的单位收取的款项，包括应收销货款、其它应收款和应收票据等。出于扩大销售的竞争需要，企业不得不以赊销或其它优惠的方式招揽顾客，由于销售和收款的时间差，于是产生了应收款项。应收款赊销的效果的好坏，不仅依赖于企业的信用政策，还依赖于顾客的信用程度。由此，评价顾客的信用等级，了解顾客的综合信用程度，做到“知己知彼，百战不殆”，对加强企业的应收账款管理大有帮助。某企业为了了解其客户的信用程度，采用西方银行信用评估常用的5C方法，5C的目的是说明顾客违约的可能性。**

**1、品格（用X1表示），指顾客的信誉，履行偿还义务的可能性。企业可以通过过去的付款记录得到此项。**

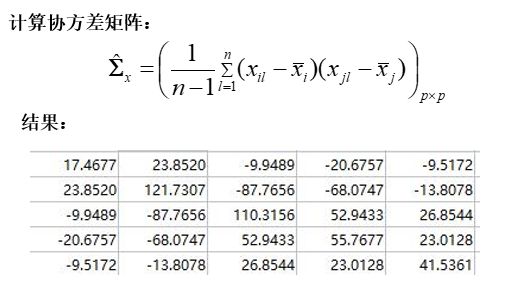
**2、能力（用X2表示），指顾客的偿还能力。即其流动资产的数量和质量以及流动负载的比率。顾客的流动资产越多，其转化为现金支付款项的能力越强。同时，还应注意顾客流动资产的质量，看其是否会出现存货过多过时质量下降，影响其变现能力和支付能力。**

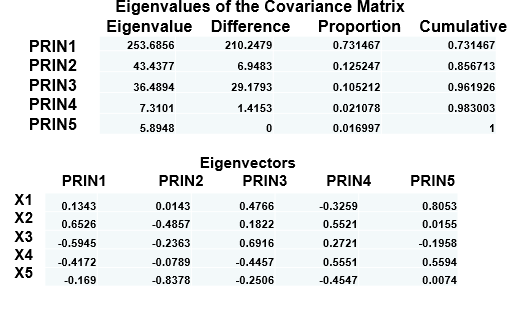
**3、资本（用X3表示），指顾客的财务势力和财务状况，表明顾客可能偿还债务的背景。**

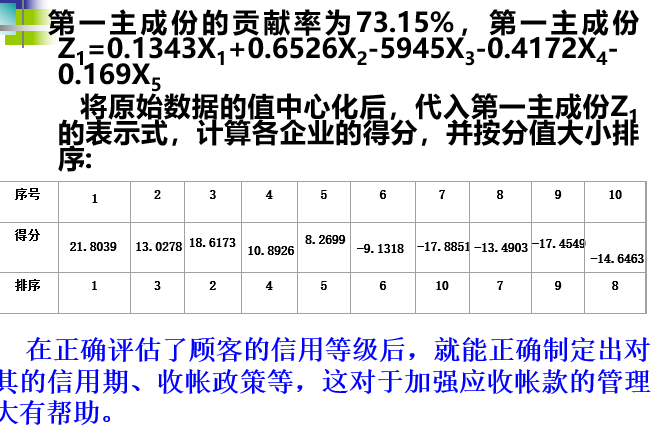
**4、附带的担保品（用X4表示），指借款人以容易出售的资产做抵押。**

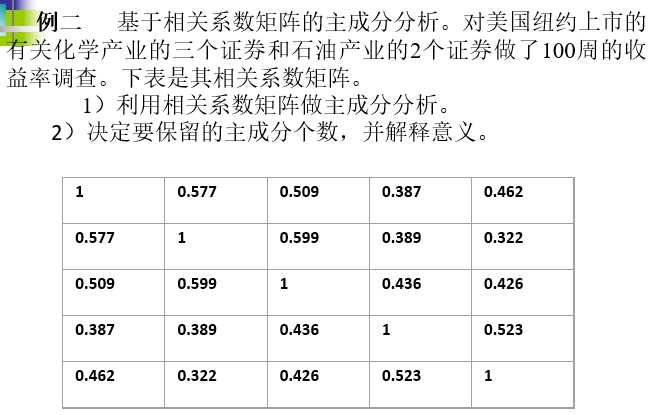
**5、环境条件（用X5表示），指企业的外部因素，即指非企业本身能控制或操纵的因素。**

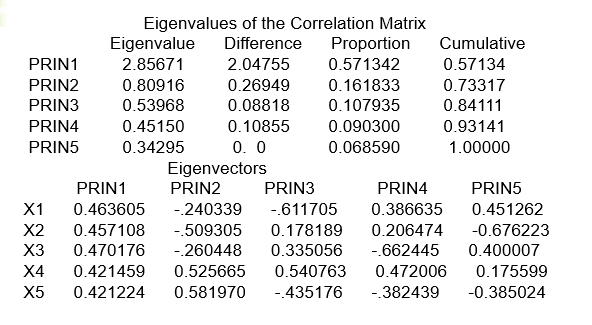
****

****

****

****





**主成分分析结论**

**根据主成分分析的定义及性质，我们已大体上能看出主成分分析的一些应用。概括起来说，主成分分析主要有以下几方面的应用。**

1．主成分分析能降低所研究的数据空间的维数。

用研究m维的Y空间代替p维的X空间(m＜p)，而低维的Y空间代替高维的x空间所损失的信息很少。即：使只有一个主成分Yl(即m＝1)时，这个Yl仍是使用全部X变量(p个)得到的。例如要计算Yl的均值也得使用全部x的均值。在所选的前m个主成分中，如果某个Xi的系数全部近似于零的话，就可以把这个Xi删除，这也是一种删除多余变量的方法。

2.多维数据的一种图形表示方法。

我们知道当维数大于3时便不能画出几何图形，多元统计研究的问题大都多于3个变量。要把研究的问题用图形表示出来是不可能的。然而，经过主成分分析后，我们可以选取前两个主成分或其中某两个主成分，根据主成分的得分，画出n个样品在二维平面上的分布况，由图形可直观地看出各样本在主分量中的地位。

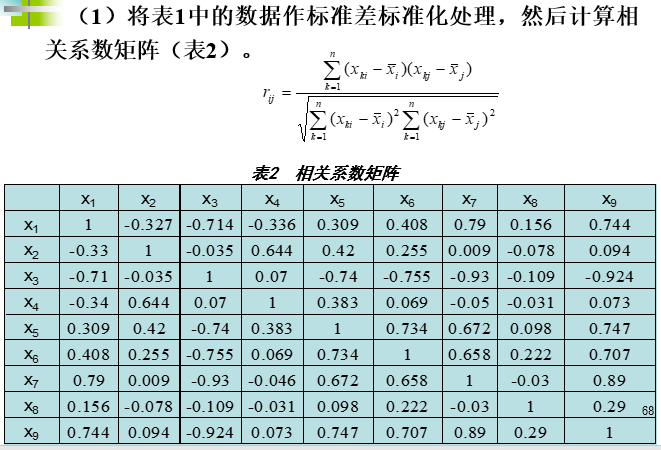
3．由主成分分析法构造回归模型。

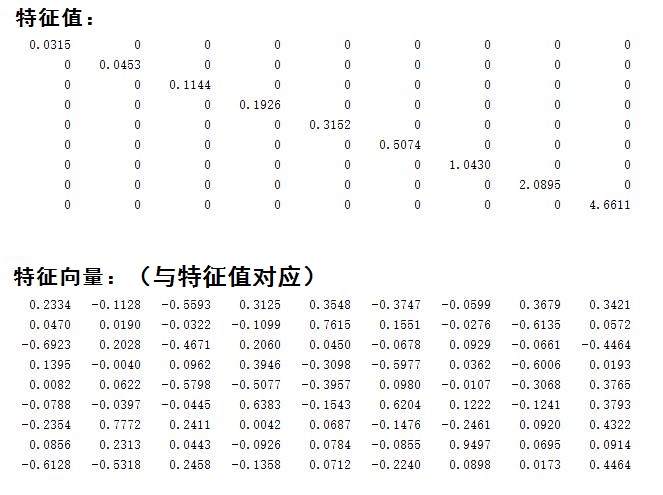
把各主成分作为新自变量代替原来自变量x做回归分析。

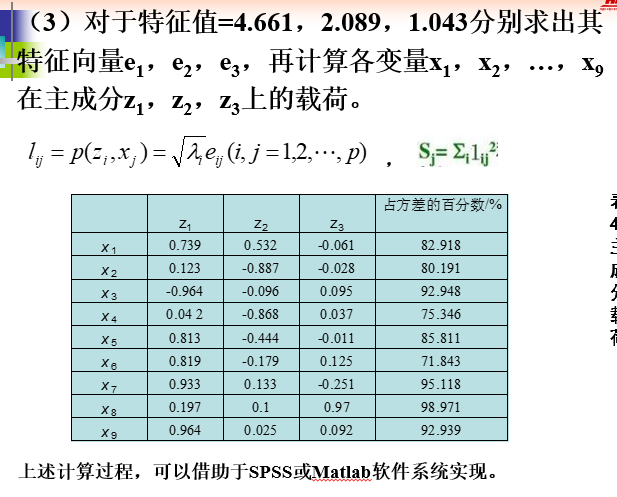
4．用主成分分析筛选回归变量。

回归变量的选择有着重的实际意义，为了使模型本身易于做结构分析、控制和预报，好从原始变量所构成的子集合中选择最佳变量，构成最佳变量集合。用主成分分析筛选变量，可以用较少的计算量来选择量，获得选择最佳变量子集合的效果。

**主成分分析方法应用实例：**







(1)第1主成分*z*1与*x*1，*x*5，*x*6，*x*7，*x*9呈现出较强的正相关，与*x*3呈现出较强的负相关，而这几个变量则综合反映了生态经济结构状况，因此可以认为第1主成分*z*1是生态经济结构的代表。

(2)第2主成分*z*2与*x*2，*x*4，*x*5呈现出较强的正相关，与*x*1呈现出较强的负相关，其中，除了*x*1为人口总数外，*x*2，*x*4，*x*5都反映了人均占有资源量的情况，因此可以认为第2主成分*z*2代表了人均资源量。

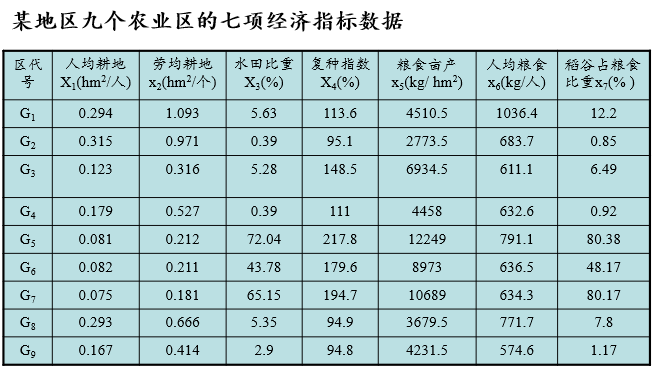
(3)第3主成分z3与x8呈现出的正相关程度最高，其次是x6，而与x7呈负相关，因此可以认为第3主成分在一定程度上代表了农业经济结构。

(4)另外，表4中最后一列（占方差的百分数），在一定程度上反映了3个主成分z1、z2、z3包含原变量（x1，x2，…，x9）的信息量多少。

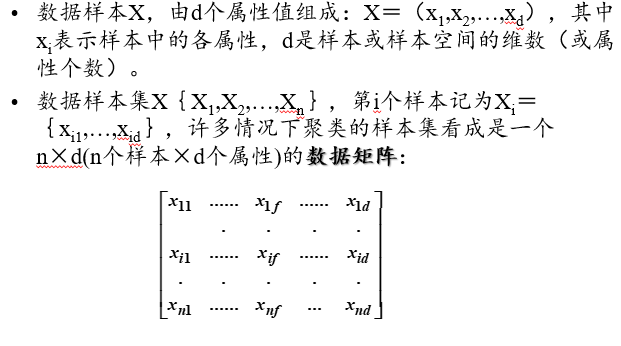
显然，用3个主成分z1、z2、z3代替原来9个变量（x1，x2，…，x9）描述农业生态经济系统，可以使问题更进一步简化、明了。

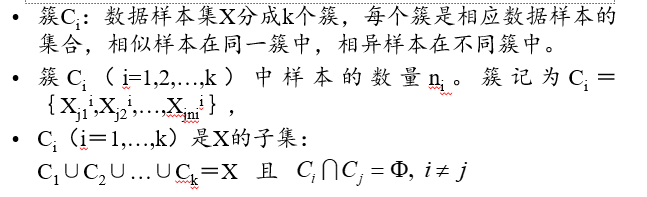
**三、聚类分析法**

* 对一个数据，既可以对变量(指标)进行分类(相当于对数据中的列分类)，也可以对观测值(事件，样品)来分类(相当于对数据中的行分类)。
* 当然，不一定事先假定有多少类，完全可以按照数据本身的规律来分类。
* 聚类分析（cluster analysis）：对变量（列）的聚类称为R型聚类，而对观测值（行）聚类称为Q型聚类。它们在数学上是无区别的。



* 如果想要对100个学生进行分类，而仅知道他们的数学成绩，则只好按照数学成绩分类；这些成绩在直线上形成100个点。这样就可以把接近的点放到一类。
* 如果还知道他们的物理成绩，这样数学和物理成绩就形成二维平面上的100个点，也可以按照距离远近来分类。
* 三维或者更高维的情况也是类似；只不过三维以上的图形无法直观地画出来而已。
* 上面的例子





* 用下面的特征来描述簇：

①簇的质心（centroid）:即样本的平均值，是簇的“中间值”（middle），但并不需要是簇中实际点。

令ni表示簇Ci中样本的数量，mi表示对应样本的均值

②簇的半径，是簇中两个点间的均方差的平方根。

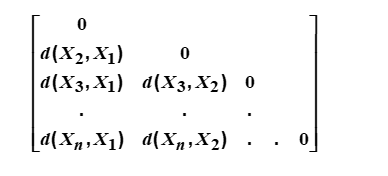
* 聚类定义：给定一数据样本集X｛X1,X2,…,Xn｝，根据数据点间的相似程度将数据集合分成k簇：｛C1,C2,…,Ck｝的过程称为聚类， ，
* 相似样本在同一簇中，相异样本在不同簇中。
* ***关于同一簇中的样本比来自不同簇的样本更为相似的判断问题***主要涉及以下两个独立的子问题：

**a.怎样度量样本之间的相似性；**

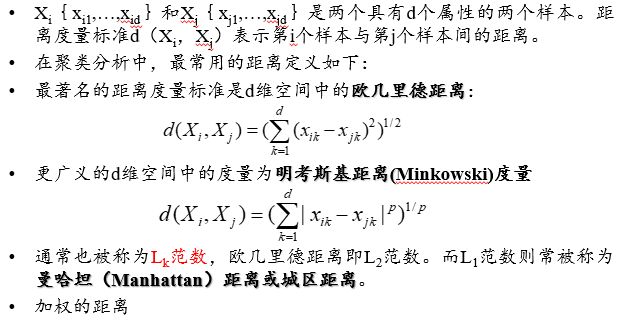
**b.怎样衡量对样本集的一种划分的好坏。**

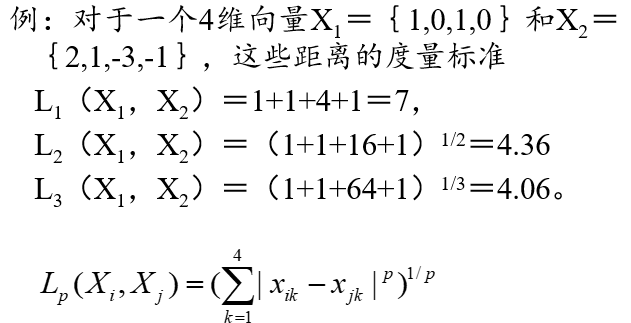
**1.相似性度量**

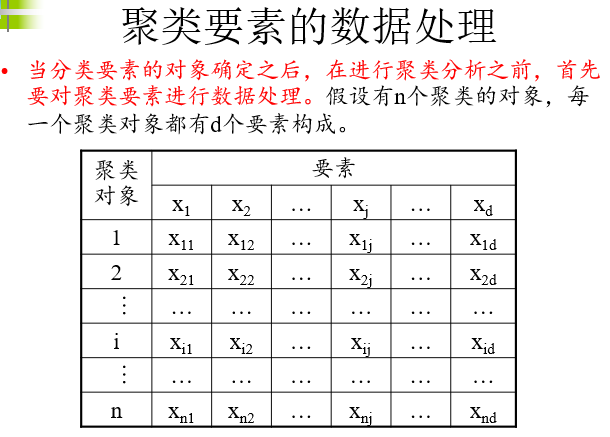
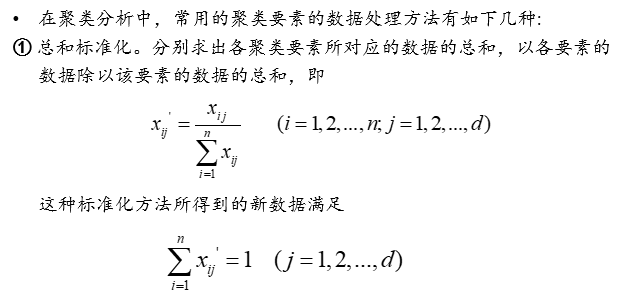
* 相异度矩阵（dissimilarity matrix）用来存储n个样本两两之间的相似性，表现形式是一个n×n维的矩阵：

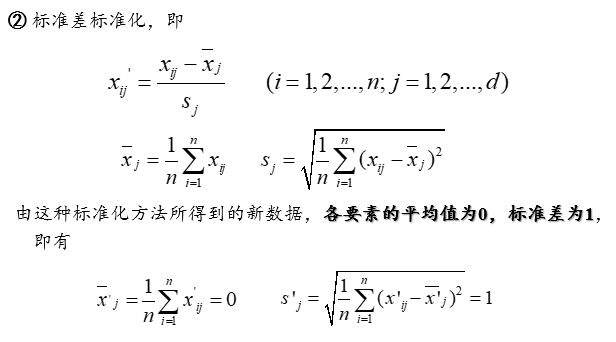


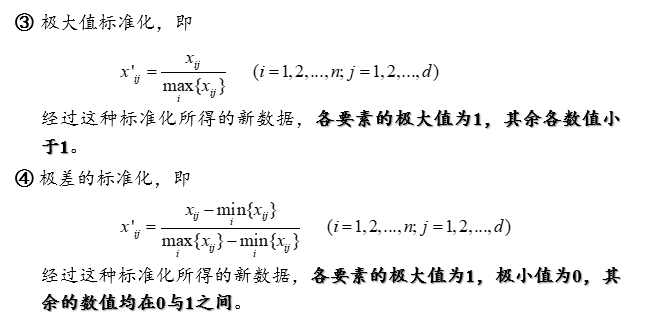
* d(Xi，Xj)是样本Xi和样本Xj间相异性的量化表示。
* 最明显的相似性度量是样本之间的距离。

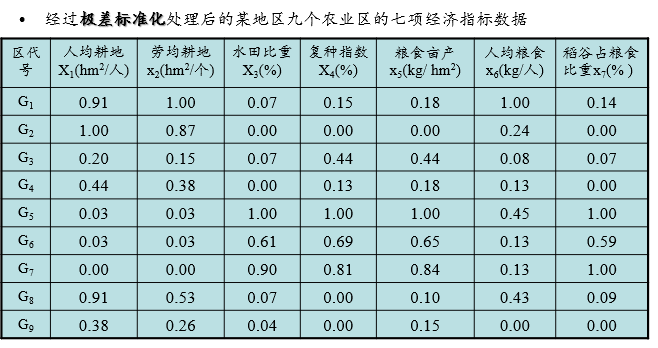
****

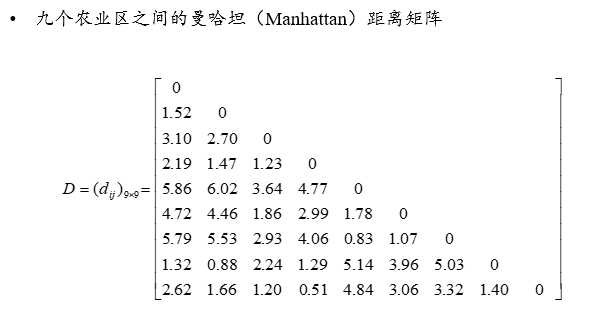
****

**** ****

****

****

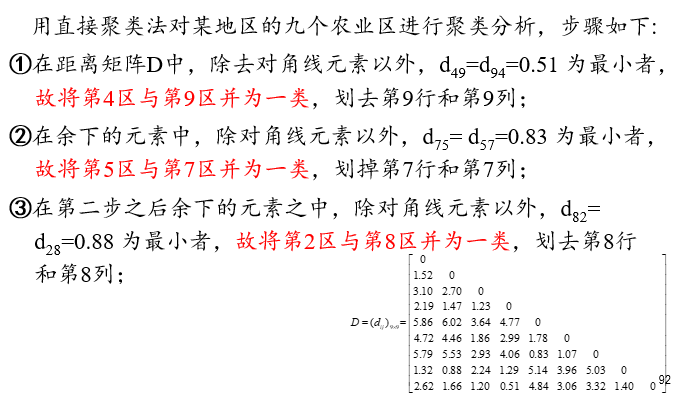
****

****

**直接聚类法:**

* **原理：先把各个分类对象单独视为一类，然后根据距离最小的原则，依次选出一对分类对象，并成新类。**
* **如果其中一个分类对象已归于一类，则把另一个也归入该类；**
* **如果一对分类对象正好属于已归的两类，则把这两类并为一类。**

**每一次归并，都划去该对象所在的列与列序相同的行。经过n-1次就可以把全部分类对象归为一类，这样就可以根据归并的先后顺序作出聚类谱系图。**

****

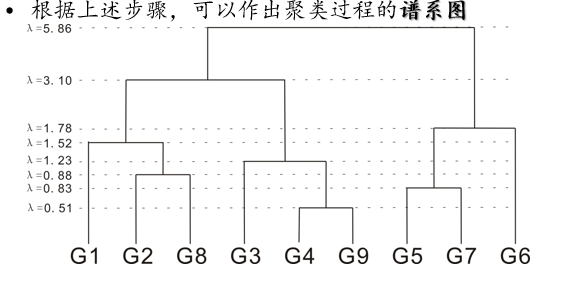
**④在第三步之后余下的元素中，除对角线元素以外，d43= d34=1.23 为最小者，故将第3区与第4区并为一类，划去第4行和第4列，此时，第3、4、9区已归并为一类；**

**⑤在第四步之后余下的元素中，除对角线元素以外，d21= d12=1.52 为最小者，故将第1区与第2区并为一类，划去第2行和第2列，此时，第1、2、8区已归并为一类；**

**⑥在第五步之后余下的元素中，除对角线元素以外，d65= d56=1.78 为最小者，故将第5区与第6区并为一类，划去第6行和第6列，此时，第5、6、7区已归并为一类；**

**⑦在第六步之后余下的元素中，除对角线元素以外，d31= d13=3.10 为最小者，故将第1区与第3区并为一类，划去第3行和第3列，此时，第1、2、3、4、8、9区已归并为一类；**

**⑧在第七步之后余下的元素中，除去对角线元素以外，只有d51= d15=5.86，故将第1区与第5区并为一类，划去第5行和第5列，此时，第1、2、3、4、5、6、7、8、9、区均归并为一类；**

****

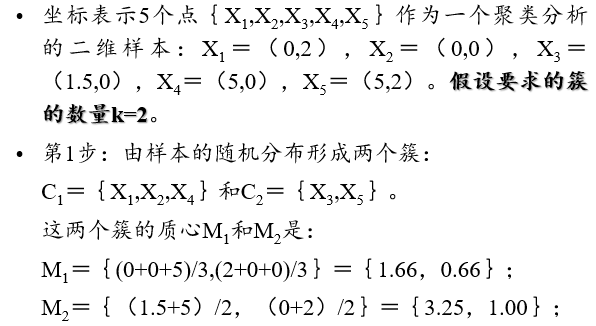
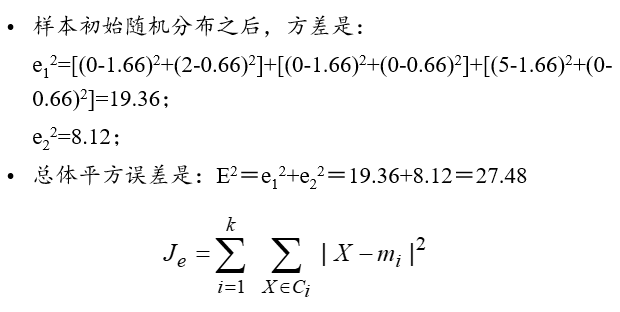
**k－means聚类算法**

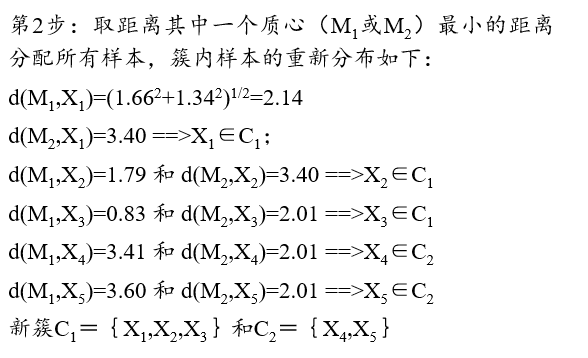
**1.选择一个含有随机选择样本的k个簇的初始划分，计算这些簇的质心。**

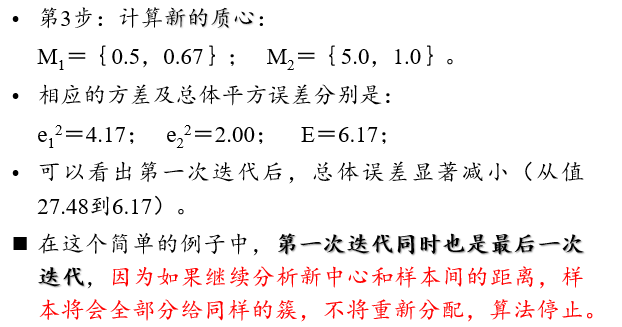
**2.根据距离把剩余的每个样本分配到距离它最近的簇质心的一个划分。**

**3.计算被分配到每个簇的样本的均值向量，作为新的簇的质心。**

**4.重复2,3直到k个簇的质心点不再发生变化或准则函数收敛。**

**** ****

****

****