**35. 聚类分析**

**（一）概述**

聚类分析，相当于“物以类聚”，用于对事物的类别面貌尚不清楚，甚至在事前连总共有几类都不能确定的情况下对数据进行分类。

而判别分析，必须事先知道各种判别的类型和数目，并且要有一批来自各判别类型的样本，才能建立判别函数来对未知属性的样本进行判别和归类。

聚类分析是把分类对象按一定规则分成组或类，这些组或类不是事先给定的而是根据数据特征而定的。在同类的对象在某种意义上倾向于彼此相似，而在不同类里的这些对象倾向于不相似。根据这种相似性的不同定义，聚类分析也有不同的方法。

聚类分析分为：对样品的聚类，对变量的聚类。

样品聚类：其统计指标是类与类之间距离，把每一个样品看成空间中的一个点，用某种原则规定类与类之间的距离，将距离近的点聚合成一类，距离远的点聚合成另一类。

变量聚类：其统计指标是相似系数，将比较相似的变量归为一类，而把不怎么相似的变量归为另一类，用它可以把变量的亲疏关系直观地表示出来。

**（二）原理**

**一、距离和相似系数**

**1. 距离**

设有n组样品，每组样品有p个变量的数据如下：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 变量  样品 | V1 | V2 | … | Vp |
| X1 | *x*11 | *x*12 | … | *x*1p |
| X2 | *x*21 | *x*22 | … | *x*2p |
| … | … | … | … | … |
| Xn | *x*n1 | *x*n2 | … | *x*np |

例如，Xi到Xj的闵科夫斯基距离定义为：



q=2时为欧几里得距离；还有马氏距离：

dij = (Xi-Xj)T*S*-1(Xi-Xj)

其中，Xi=(*x*i1, …, *x*ip )，*S*-1为n个样品的p×p的协方差矩阵的逆矩阵。

注：马氏距离考虑了观测变量之间的相关性和变异性（不再受各指标量纲的影响）。

**距离选择的基本原则：**

（1）要考虑所选择的距离公式在实际应用中有明确的意义。如欧氏距离就有非常明确的空间距离概念。马氏距离有消除量纲影响的作用。

（2）要综合考虑对样本观测数据的预处理和将要采用的聚类分析方法。如在进行聚类分析之前已经对变量作了标准化处理，则通常就可采用欧氏距离。

（3）应根据研究对象的特点不同做出具体分折。实际中，聚类分析前不妨试探性地多选择几个距离公式分别进行聚类，然后对聚类分析的结果进行对比分析，以确定最合适的距离测度方法。

**2. 相似系数**

变量间的相似性，可以从它们的方向趋同性或“相关性”进行考察，从而得到“夹角余弦法”和“相关系数”两种度量方法。

（1）夹角余弦

Xi=(*x*i1, …, *x*ip )和Xj作为p维向量，按通常内积中夹角的概念来定义



（2）相关系数



显然 |*c*ij| ≤1，借助于相似系数，可以定义变量之间的距离。例如，采用非相似测度距离为dij=1-|*c*ij|, 或dij2=1- *r*ij2.

**二、类的特征、类间的距离**

1. 类的定义



给定某阈值T，① 若dij ≤ T，对任意i, j∈G；或者

② 若(∑j∈G dij)/(k-1) ≤ T，对任意i∈G；或者

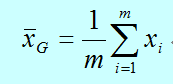
③ 若对任意i∈G，都存在j∈G使得dij ≤ T；

则称G为一个类。条件从强到弱的顺序为：①②③

2. 类的特征

设类G的元素为*x*1, …, *x*m, m为G中样品数。描述类的特征有：

（1）均值或重心



（2）样品协方差阵



（3）直径



3. 类间的距离

设Gp和Gq中分别有p和q个样品，它们的重心分别记为和.

（1）最短距离



它不对类的形状加以限制，保证了对拉长和不规则类的检测，例如(b)(c)(e)(d)形式的类；但它却牺牲了恢复压缩类的性能，另外它也趋向于在分开主要类之前去掉分布的尾部。

（2）最长距离



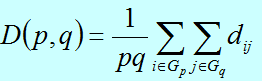
严重地倾向于产生直径粗略相等的类，而且可能被异常值严重地扭曲，例如(a).

（3）重心法距离



即两个重心之间欧氏距离的平方。重心法在处理异常值上比其他谱系方法更稳健，但是在其他方便不如Ward或类平均距离法的效果好。

（4）类平均距离



Gp和Gq中所有两个样品对之间距离的平均。类平均距离法趋向于合并具有较小偏差的类，而且稍微有点倾向与产生相同方差的类。例如(a).

（5）Ward最小方差法或Ward离差平方和距离



D表示类的直径，也有



若样品间的距离采用欧氏距离，上式可表示为



该方法在每次合并类Gp和Gq为Gp+q时，总是选择使得合并后的Dw(p,q)值最小的类Gp和Gq（最小方差法）. 合并后增加的最小方差Dw(p,q)除以合并后总的离差平方和TSS的比值（即半偏R2）的统计意义是容易解释的。

Ward方法趋向于合并具有少量观察的类，并倾向于形成具有大约相同数目观察的类，一般是在多元正态混合型、等球形协方差、等抽样概率假设下合并类。例如(a). Ward方法对异常值也很敏感。

（6）密度估计法

例如(b). 使用一种基于密度估计的新的非相似测度d\*来计算样品*x*i和*x*j的近邻关系；②然后根据基于d\*方法计算的距离，采用最小距离法进行聚类。有三种不同的密度估计法：k最近邻估计法、均匀核估计法、Wong混合法（适用于大数据集而不适用于小数据集）。

（7）两阶段密度估计法

在密度估计法中，众数类经常在尾部所有点聚类之前就已经被合并掉了。两阶段密度估计法对密度估计法这一缺点进行了修正，以确保众数类被合并之前，所有点能被分配到众数类中去。

在第一阶段中，互不相交的众数类被生成，只有在两个类中至少有一个类中的样品个数少于proc cluster过程的选项“*mode*=值”时才能合并。在第二阶段中，把前面形成的众数类再按最短距离法谱系地合并（某一类能够完全地包含在别的类里，但在类之间不允许其他形式的重迭）。

（8）最大似然法



最大似然估计法是在三个基本假设：多元正态混合分布、等球面协方差和不等抽样概率的前提下，使得在谱系的每个水平上似然值最大。

该方法除了偏向于生成不等大小的类外，与Ward最小方差法很相似。proc cluster过程的偏度选项penalty=p值可用于调整偏向的程度。修正公式可将公式中的2改为p值。

**三、类的统计量**

（1）类Gp的均方根标准差



即类内的离差平方和（定义为直径），点越分散直径越大。

（2）R2



R2用于评价每次合并成m个类时的聚类效果，R2越大说明聚类效果（能被分开）越好。R2=1，说明类被完全分开；当所有样品最后合并成一类时，R2=0，说明类被完全混合在一起了，分不开。

R2的值总是随着分类个数的减少而变小。那么如何根据R2的值来确定所有样品应分为几类最合适呢？

首先，最合适分类的R2的值不能太小，最好能达到0.7以上；其次，不能孤立地看合并后R2值的大小，应该看R2值的变化。类的合并总是使R2值变小，如果这种变小是均匀的，合并是应该的，但当某次合并使R2值减少很多，这次合并就不应该，即我们找到了最合适的分类数。

例如，从10类合并到5类时，R2值每次减少大约在0.01到0.02之间，R2值减小到0.8，从5类合并到4类时，R2值减少了0.3，即R2=0.8-0.3=0.5，这时，我们就可以得出分5类是最合适的。

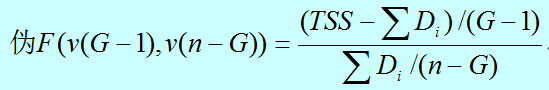
（3）半偏R2统计量

合并类Gp和类Gq为类Gm时，可以用半偏R2统计量评价这次合并的效果：



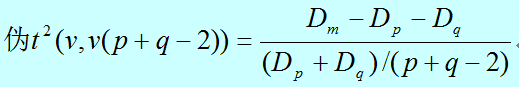
表示合并后类内离差平方和的增量（=上次R2-本次R2）。

（4）伪F统计量



其中，G为聚类的个数，n为观察总数，v为观察样品的变量维数。该统计量用于评价分为G个聚类的效果。伪F值越大越表示这些观察可显著地分为G个类。

（5）伪t2统计量



该统计量用以评价合并类Gp和类Gq为类Gm的效果。该值很大说明合并后离差平方和的增加量相对于原来两类的类内离差平方和很大。这表明上一次合并的两个类Gp和Gq是很分开的，也就是说上一次聚类的效果是很好的。否则，这两个类应该在再合并时的离差平方和的增加量很小。

**四、类数的确定**

至今无太满意的结果。

（1）一般情况下，是看每个变量的R2值及所有变量的累积R2值，观察这些R2值与聚类个数所画的折线图形，以此来判断；

（2）通过对数据的典型变量（可用proc candisc典型判别分析选出）做散点图来判断；

（3）立方聚类准则CCC；

通过数据模拟研究，发现三个最好的准则是：①伪F统计量；②伪t2统计量；③立方聚类准则CCC。

在cluster过程中这三个准则都被应用和输出，而在fastclus过程中仅伪F统计量和CCC统计量被应用和输出。

建议寻找这三种统计量之间的一致性，即：CCC统计量和F统计量的局部峰值所对应的聚类数，与这个聚类数伪t2统计量的一个较小值和下一个聚类数的一个较大伪t2统计量相吻合。还必须强调这些准则仅仅适用于紧凑的或略微拉长的类，也较好地适合于略微正态分布的类。

**（三）聚类的方法**

**一、系统聚类法**

目前使用最多的一种方法。基本步骤是：

（1）首先各样品自成一类，这样对组样品就相当于有类（即一类包括一个样品）；

（2）计算各类间的距离（可选8种距离见前文），将其中最近的两类进行合并；

（3）计算新类与其余各类的距离，再将距离最近的两类合并；

（4）重复上述的步骤，直到所有的样品都聚为一类时为止。

注意：

① 在整个聚类的过程中，如果在某一步的距离矩阵中最小元素不止一个时，则可以将其同时合并；

② 系统聚类法由于每次都计算距离，计算量（内存占用）较大；

③ 研究表明：最好综合特性的聚类方法为类平均法或Ward最小方差法，而最差的则为最短距离法。Ward最小方差法倾向于寻找观察数相同的类。类平均法偏向寻找等方差的类。具有最小偏差的聚类方法为最短距离法和密度估计法。拉长的或无规则的类使用最短距离法比其他方法好。最没有偏见的聚类方法为密度估计法。

**二、K均值聚类法（动态聚类法）**

基本步骤：

（1）选择K个观察作为“凝聚点”或称类的中心点，作为这些类均值的第一次猜测值，将所有的样品分成K个初始类；

（2）通过欧氏距离将某个样品划入离中心最近的类中，并对获得样品与失去样品的类，重新计算中心坐标；

（3）重复步骤2，直到所有的样品都不能再分配时为止（反复迭代，直到再也无法降低样本与重心离差之和为止）。

优点：速度快省内存；缺点：需要事先确定类数。

注意：该聚类法对异常点很敏感，通常会把异常点分配为单独的一个类（可用于检测异常点proc fastclus）。

**（三）SAS实现**

SAS提供了5个聚类过程，即cluster，fastclus，modeclus、varclus和tree过程。

cluster为系统聚类过程，可使用十一种聚类方法进行谱系聚类，包括重心法、Ward离差平方和法和欧氏平均距离法等。

fastclus为动态聚类过程，使用 K-均值算法寻找不相交的聚类，适宜于大样本分析，观察值可多达10万个。

modeclus为动态聚类过程，使用非参数密度估计法寻找不相交的聚类。

varclus过程可用于系统或动态聚类，通过斜交多组分量分析对变量作“谱系的”和“不相交的”两种聚类。

cluster过程、fastclus过程和modeclus过程用于对样品聚类，varclus过程用于对变量聚类。tree过程将cluster或varclus过程的聚类结果画出树形结构图或谱系图。

**一、PROC CLUSTER过程步**

基本语法：

proc cluster data = 数据集　<可选项>;

var 变量列表;

id 变量;

freq 变量;

copy 变量列表;

rmsstd 变量;

by 变量列表;

说明：

（1）可选项

outtree=输出数据集

供tree过程调用，用来输出聚类结果的树状图；

method=算法

ward（离差平方和法），average（类平均法），centroid（重心法），complete（最长距离法），single（最短距离法），median（中间距离法），density（密度法），flexible（可变类平均法），twostage（两阶段密度法），eml（最大似然法），mcquitty（相似分析法）；

standard/std——对变量实施标准化；

nonorm——阻止距离被正态化成均数为1或均方根为1；

nosquare——阻止过程在*method*=average/centroid/median/ward方法中距离数据被平方；

mode=*n*——当合并两个类时，规定对被指定的众数类中的每个类至少有*n*个成员，该选项只能在*method*=density/twostage时使用；

penalty=p——指定用于*method*=eml中的惩罚系数（p>0, 默认p=2）；

trim=p——要求从分析中删去那些概率密度估计较小的点（0≤p<100, 被当作百分比），在*method*=ward/complete时，因为类可能被异常值严重地歪曲，最好使用这个选项（也可用于*method*= single）；

dim=n——用于*method*=density/twostage时指定使用的维数（n≥1），若是坐标数据，缺省值为变量个数；若是距离数据，缺省值为1；

hybrid——要求用Wong混合聚类方法，其中密度用k均值法的初始聚类分析中的均值计算得到。这个选项只能在*method*=density/ twostage时使用；

k=*n*——指定k最近邻估计法中近邻的个数（2≤n<观察数）；

r=*n*——指定均匀核密度估计法的支撑球半径（n>0）；

notie——阻止cluster过程在聚类历史过程中检查每次产生的类间最小距离连结（ties）的情况，此选项可以减少过程执行的时间和空间；

rsquare——输出R2和半偏R2；

rmsstd——输出每一类的均方根标准差；

ccc——输出在均匀的原假设下判断聚类分成几类合适的立方聚类准则统计量ccc和近似期望值R2；同时输出选项rsquare有关的R2和半偏R2；此选项不适合于*method*=single（容易删掉分布的结尾部分）；

pseudo——输出伪F统计量(PSF)和伪t2统计量(PST2)，当分类数目不同时，它们有不同的取值；

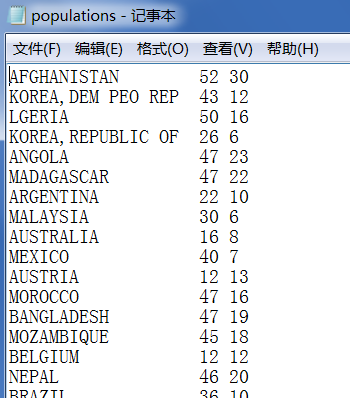
simple——输出简单统计数；

在输出报表中，可以根据输出的ccc、psf和pst2统计量确定多少分类数较合适，当ccc和psf值出现峰值所对应的分类数较合适，而pst2值是在出现峰值所对应的分类数减1较合适。

copy语句——指定输入数据集中的一些变量拷贝到outtree=的输出数据集中；

rmsstd语句——当输入数据集中的坐标数据代表类的均值时，定义表示均方根标准差变量，通常与freq语句中的变量配合使用。

**例1**用cluster和tree过程1976年74个国家每100000人口的出身率和死亡率数据（C:\MyRawData\populations.txt）进行聚类分析。



**代码：**

**data** vital;

infile 'C:\MyRawData\populations.txt';

title 'CRUDE BIRTH AND DEATH RATES IN 1976';

input country $19. birth death;

**run**;

**proc** **print** data = vital;

**run**;

**proc** **plot** formchar='|-----|--' hpct=**100** vpct=**100**;

plot death\*birth / hpos=**86** vpos=**26**;

**run**;

**proc** **cluster** data = vital out=tree method = average print = **15** ccc pseudo;

var birth death;

title 'Cluster Analysis of Birth and Death Rates in 74 Countries';

**run**;

**proc** **tree** data = tree noprint out = out ncl = **8**;

copy birth death;

**run**;

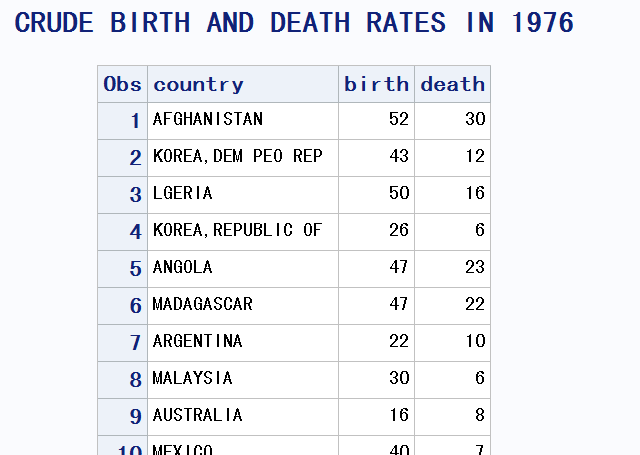
**proc** **plot** formchar='|-----|--';

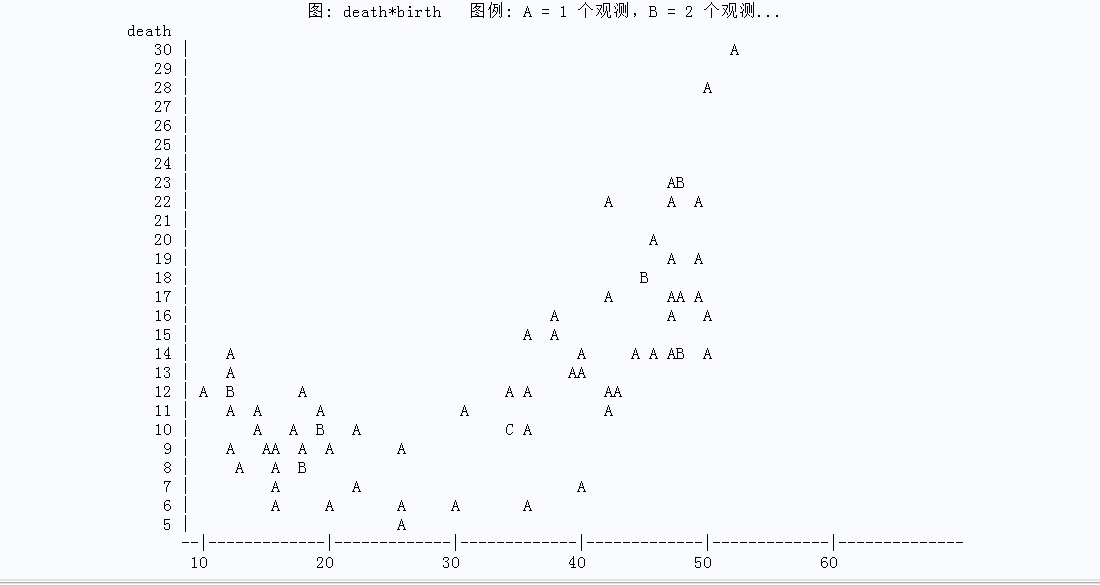
plot death\*birth=cluster / hpos=**86** vpos=**26**;

title2 'Plot of Clusters Form Method = Average';

**run**;

**运行结果及说明：**





A表示此处有一个散点，而B表示此处有二个重叠散点，C表示此处有三个重叠散点；直观上看散点图可以区分成三类：低出身率和低死亡率的国家（富裕发达的国家），高出身率和高死亡率的国家（贫穷落后的国家），中出身率和低死亡率的国家（中等发展中的国家）。



输出包括协方差阵的特征值，两相邻特征值之差，各特征值占总方差的百分比，和累计百分比。



聚类是从74国家作为74类，先聚类成73=74-1类开始，直至将所有国家聚为１类的聚类过程和统计数。

“print=15”，聚类数列只输出显示聚类成15类后各合并过程；

“连接聚类列”下的两列为每一次聚类成1个新类的2个样品，标有ob*n*表示是原始样品中的第n号样品，标有cl*n*表示是在上面聚类过程中已经聚成的第n类；

例如，第14新类（聚类数=14）由前面形成的第20旧类（cl20）和原始样品的第8号（ob41）合并产生，共有4个（freq=4）原始样品；同时也可求出前面形成的cl20旧类有4-1=3个原始样品；最后由旧的cl2类和cl3类合并形成1类，其中cl2类由cl4类和cl5类合并，cl3类由cl6类和cl14类合并。

频数列为新类中所含的样品数；

半偏R2，表示每一次合并对信息的损失程度；

R2，表示累计聚类结果；

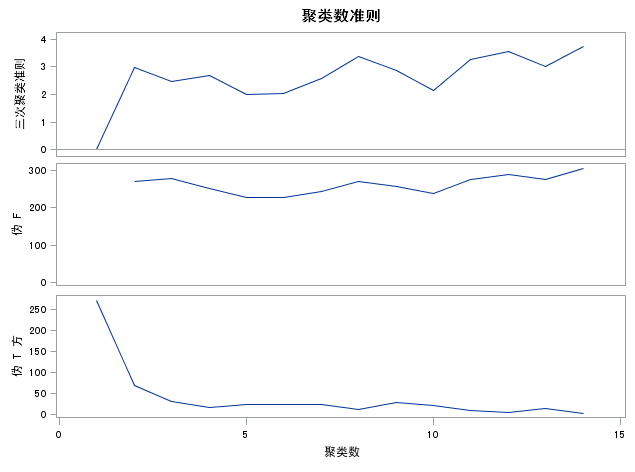
近似期望R2为在均匀零假设下近似期望的R2；

ccc、psf、psf2列为3个判断分成几类较合适的统计量；

Norm RMS Dist列为距离被正态化后两样品或类间的平均距离；

Tie（连接）列里用T来指出最小距离的一个连结，而空白表明没有连结；

“out=tree”将聚类过程和聚类结果及相应的统计量输出到tree数据集中。



关于如何选择合适的聚类数。主要方法是查看“三次聚类准则值CCC”在*ncl*聚类数为多少时达到峰值，以及“伪F值PSF”和“伪t2值PST2”在*ncl*聚类数为多少时达到峰值。同时可以通过再合并成新类时R2值的减少最多（即每一次合并时信息的损失程度“半偏R2值”最大）来验证，还可以直接通过树状图决定分类数。

通常CCC、PSF、PST2值在不同的*ncl*处有多处峰值，为便于比较和观测，将CCC、PSF和PST2统计量按聚类数大小顺序连接成曲线（则必须先对tree数据集中表示聚类数的字段\_ncl\_进行排序），在同一张图中绘制相应的散布图。**代码：**

**data** vital;

infile 'C:\MyRawData\populations.txt';

input country $19. birth death;

**run**;

**proc** **cluster** data = vital out=tree method = average print = **15** ccc pseudo;

var birth death;

title 'Cluster Analysis of Birth and Death Rates in 74 Countries';

**run**;

**proc** **print** data=tree;

title 'tree';

**run**;

**proc** **sort** data=tree;

by \_ncl\_;

goptions reset=global gunit=pct cback=white border htitle=**6** htext=**3** ftext=swissb colors=(back);

**proc** **gplot** data=tree;

plot \_ccc\_\*\_ncl\_ / haxis = axis1 vaxis = axis2;

plot2 \_psf\_\*\_ncl\_ \_pst2\_\*\_ncl\_ / vaxis = axis3 overlay;

symbol1 i=join v='C' l=**3** h=**2.5** w=**1.5**;

symbol2 i=join v='F' l=**1** h=**2.5** w=**1.5**;

symbol3 i=join v='T' l=**1** h=**2.5** w=**1.5**;

axis1 label = ('Number of Clusters') order=(**1** to **15** by **1**) ;

axis2 label = ('CCC') order=(**0** to **6** by **0.5**);

axis3 label = ('PSF and PST2') order=(**0** to **300** by **25**);

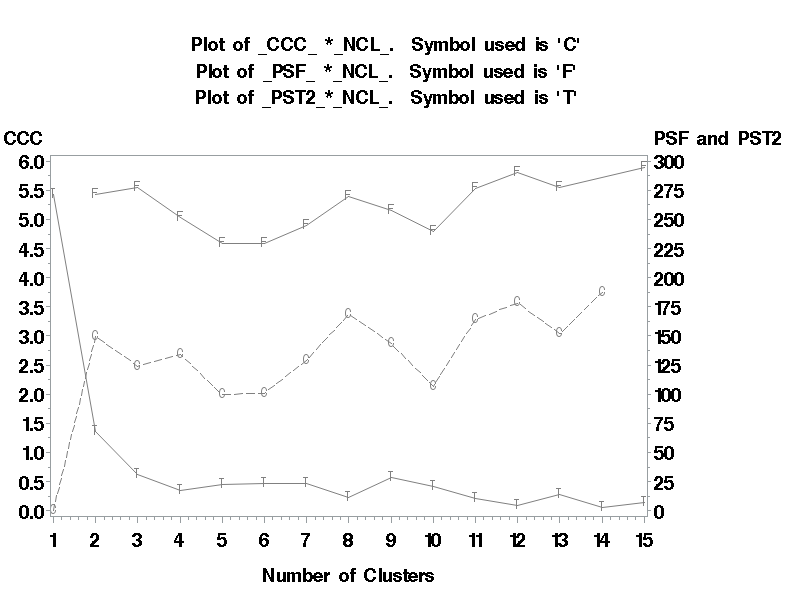
title2 "Plot of \_CCC\_ \*\_NCL\_. Symbol used is 'C' ";

title3 "Plot of \_PSF\_ \*\_NCL\_. Symbol used is 'F' ";

title4 "Plot of \_PST2\_\*\_NCL\_. Symbol used is 'T' ";

**run**;

**运行结果：**

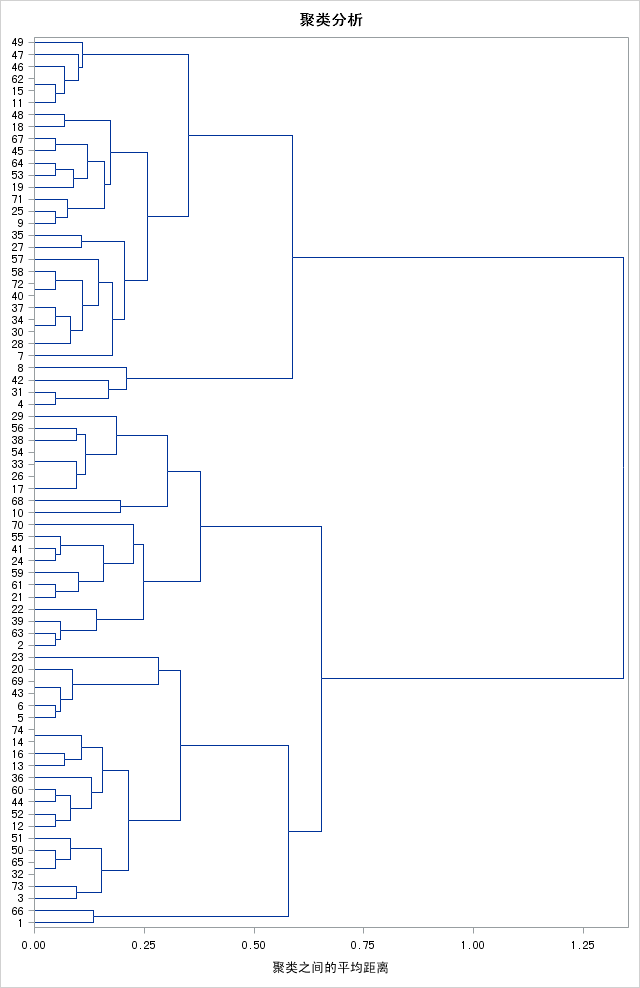


CCC值在ncl=2, 4, 8, 12处有峰值，这些是立方聚类准则认为可考虑取的分类数，但ncl=8时更陡些；

psf在ncl=3, 8, 12处有峰值，这些是根据伪F统计量可考虑取的分类数，但ncl=8时更陡些；

pst2在ncl=5, 9, 13处有峰值，通常伪t2统计量认为取再减1的分类较合适，或者直接取伪t2统计量最小值处的分类数，因为再次合并将会有比较大伪t2统计量，所以根据伪t2统计量可考虑的分类数为4, 8, 12;

从半偏R2值分析，从12类合并成11类时，有较大的信息损失值0.003198，从8类合并成7类时，有较大的信息损失值0.009841，这项指标特别指出从3类合并成2类时，有仅次于分成1类的信息损失值0.096203，半偏R2指标值支持分成2, 3, 8, 12类；



最后，从带有聚类过程和合并距离的树状图中，也能较直观地分成2, 3, 4, 8类；

综合结论，可以把74国家分成2类、3类、4类、8类、12类，尤其要注意的是分成8类，因为这种分类已从聚类运算结果分析中显示了良好的指标性能。

注：完全分开的球状类，适合用average（类平均法）；

没有完全分开的球状类，适合用ward（离差平方和法）；

大小和离散程度不等的多元正态类，适合用eml（最大似然法）；

拉长的多元正态类，适合用twostage, k=10（两阶段密度法）；

中心相同大小不等的圆环类，适合用single（最短距离法）。

**二、PROC FASTCLUS过程步**

选择“凝聚点”，根据距离接近的原则归类，适合对大样品数据做快速的“粗”分类。

初始 “凝聚点”可以指定或由过程自动选择；可以指定最大的“凝聚点”个数，及“凝聚点”之间的必须分隔开的最小距离。

注意：fastclus过程对异常点很敏感，通常会把异常点分配为单独的一个类，因此也是检测异常点的有效过程。

基本语法：

proc fastclus data = 数据集 <可选项>;

var 变量列表;

id变量;

freq变量;

weight变量;

by 变量列表;

可选项：

seed=输入数据集——指定该数据集作为初始“凝聚点”；

mean=输出数据集——包含每个类的均值和一些统计量；

out=输出数据集——包含原始数据和分类标志的聚类结果；

maxclusters/maxc=*n*——指定允许的最大聚类数（默认为100）；

radius =*t*——为选择新的“凝聚点”指定最小距离准则；当一个观察点与已有“凝聚点”的最小距离都大于t值时，该观察可考虑用来作为一个新的“凝聚点”；默认t=0；选择一个好的t值，可以得到好的分类结果（注意，必须规定maxclusters =或radius =中的一个）；

replace= *full*/*part/none/random*——指定“凝聚点”的替换方式；默认为full（用统计方法替换）；part为仅当观察点与最近的“凝聚点”距离大于“凝聚点”之间的最小距离时，那么有一个老的“凝聚点”被替换；none为禁止“凝聚点”的替换；random为从完整的观察样本中随机地选择一组样本作为初始“凝聚点”；

drift——每当一个观察归入最靠近的“凝聚点”所在类后，此类的“凝聚点”都要被该类中现有观察的均值所替代；因而“凝聚点”不断发生变化；

convergec/conv=*c*——指定收敛的判断准则，c≥0，默认为0.02.当“凝聚点”改变的最大距离小于或等于初始“凝聚点”之间的最小距离乘以c值时，循环过程结束；

maxiter= *n*——指定重新计算“凝聚点”所需的最大迭代次数，默认为１；

strict=*s*——当一个观察同其最近“凝聚点”之间的距离大于s值时，此观察不归入此类；

distance——输出类均值之间的距离；

list——要求列出所有观察的id变量值，观察所归入类的类号，及观察与最终“凝聚点”之间的距离；

注意：在采用fastclus过程之前，若变量值的单位不一致则必须对变量预先用standard过程转换成标准分布，例如

**proc** **standard** mean=**0** std=**1** out=abc;

var x1 x2;

或由factor过程产生因子得分，然后由fastclus过程加以聚类（weight）。

**三、PROC VARCLUS过程步**

该过程属于变量聚类，是基于相关矩阵或协方差矩阵，对数值变量进行不相交或谱系分类。类的划分通过计算每类第一主成份或重心成分的最大方差而确定，因此，同每一类有联系的是该类中这些变量的线性组合。

VARCLUS过程可用来作为变量压缩的方法。对于含有很多变量的变量集，常常用信息损失很少的部分变量集替代。若采用相关矩阵的信息，则所有变量都平等；当引用协方差矩阵分析时，某变量有较大方差，该变量则较为重要。VARCLUS过程生成的输出数据集，可由SCORE过程计算出每类的得分。

基本步骤：若没有为过程提供初始分类，则把所有变量看成一个类，然后

（1）首先挑选一个将被分离的类：或者用它的类分量所解释的方差百分比最小，或者同第二主成份有关的特征值为最大；

（2）把选中的类分成两个类：先计算出两个主成份，再进行斜交旋转，并把每个变量分配到旋转分量对应的类里，分配原则为使变量与这个主成份的相关系数最大；

（3）变量重新分类：通过迭代，变量被重新分配到新类中，使得由这些类分量所解释的方差最大（可能要求保持谱系结构）；

（4）当每一类满足规定的准则时，过程停止迭代。准则：或者是每个类分量所解释的方差百分比，或者是每一类的第二特征值（默认准则：每类只要有一个特征值大于1，则过程停止）。

基本语法：

proc varclus data = 数据集 <可选项>;

var变量列表;

partial 变量列表;

seed 变量列表;

freq 变量;

weight变量;

by 变量列表;

可选项：

outstat=输出数据集——包含均值、标准差、相关系数、类得分系数和聚类结构；

outtree=输出数据集——包含聚类过程的树状结构信息，供tree过程调用；

minc =*n*——最小聚类个数（默认为ｌ）；

maxc= *n*——最大聚类个数；

maxeigen= *n*——规定每一类中第二特征值所允许的最大值；

percent= *n*——指定类分量必须解释的方差百分比；

centroid——使用重心成份法而不是主成份法；

maxiter= *n*——规定在交替最小二乘法阶段中的最大迭代次数。

maxsearch= *n*——指定在搜索阶段最大迭代次数；

cov——用协方差矩阵聚类；

hi——要求在不同层次的类构成谱系聚类结构；

initial=group/input/random/seed——规定初始化类的方法；

corr——输出相关系数；

simple——输出均值和标准差。

short——不输出类结构、得分系数和类间相关阵；

summary——只输出最后的汇总表；

trace——列出在迭代过程中每个变量所归入的类。

**四、PROC TREE过程步**

该过程利用CLUSTER过程和VARCLUS过程生成的数据集来绘制树状结构图（可以按垂直或水平方向输出）。

TREE过程可以把输入数据集中的任何数值变量都能够用来规定这些类的高度，还可根据用户的要求生成一个输出数据集，其中包含一个变量，其值用以标识在这个树里指定水平上不相交的类。

基本语法：

proc tree data = 由cluster过程和varclus过程生成的数据集数据集 <可选项>;

name 变量;

parent 变量 ;

height 变量;

id 变量;

copy 变量列表;

freq 变量;

by 变量列表;

可选项：

out=输出数据集——输出绘制树形结构图的有关数据到指定数据集；

level=*n*——对于out=的输出数据集，规定确定不相交类的树状图水平（层次）；

ncl =*n*——规定在out=的输出数据集中所希望的类个数；

dock=*n*——当某个类中的对象（观察或变量）的个数小于或等于*n*时，在out=的输出数据集里把该类中这些对象的变量cluster和clusname的值置为丢失值，默认n=0；

root=“名称值”——如果不想输出整个树状图，规定想输出的子树根的name变量的值；

hor——要求树状图的取向为水平方向，且树根在左边。默认为垂直方向，树根在上部；

height*=*h/l/m/n——规定在树状图中用以确定高度轴的常规变量，分别指\_height\_变量、根到自己节点的路径长度、\_mode\_变量、\_ncl\_变量；

maxh=*n*——指定在高度轴上输出的最大值；

minh*=n*——指定在高度轴上打印的最小值；

ntich*=n*——指定在高度轴上刻度之间的间隔个数；

pages*=n*——规定这个树状图被展开的页数；

pos*=n*——指定在高度轴上打印位置的个数；

spaces*=n*——规定在打印输出中对象之间的空格数；

tickpos*=n*——指定在高度轴上每个刻度间隔打印位置的个数；

fillchar*=*“字母”——规定没有连成一类的树叶之间的打印字符。缺省值为空格；

joinchar*=*“字母”——规定已连成一类的树叶之间的打印字符，缺省值为X；

leafchar=“字母”—— 规定表示没有子辈的类的打印字符，缺省值为“.”;

treechar=“字母”—— 规定表示有子辈的类的打印字符，缺省值为“X”;

sort——按照类的形成顺序，用height变量对每个节点的子辈排序；

des——把选项sort的排列顺序反过来；

list——列出这个树中所有节点，并且打印高度、父辈及每个节点的子辈；

noprint——只创建out=的输出数据集而不绘制树状图；

graphics——要求在graph窗口中输出高分辨率的树状图，类的合并用连接线归纳表示；

name语句——规定一个字符或数值变量，用以标识每个观察代表的节点，name变量同parent变量联合确定树的结构（缺省时寻找\_name\_变量）；

parent语句——规定一个字符或数值变量，用以标识每个观察的父辈节点（缺省时寻找\_parent\_变量）；

height语句——规定一个数值变量用于定义这个树中每个节点（类）的高度，高度变量由选项height=规定；

copy语句——把语句中列出的一个或几个变量复制到out=的输出数据集中；

id语句——id变量可以是字符或数值变量，用以在打印输出树状图中识别对象。

注：对于大样本数据，可以用PROC CLUSTER过程训练聚类样本，确定分类变量和分类数目，再用PROC FASTCLUS过程聚类。