返回总目录

目　　录

第九部分
集 群 分 析

第 42 章　SAS 系统内九种集群分析程序概述

42.1　什么是集群分析

集群分析是一些分类方法的统称。它的目的是将变量或观察体予以分类，也就是把相似的变量或观察体归纳成一个集群 (Cluster)。在分类的过程中，分类的标准完全是自生的，也就是由数据本身决定的，因此集群法不像鉴别法 (见第 37 章说明) 要仰赖外在的或预知的标准来分类。

42.2　四类常用的集群分析方式

集群分析的方法可分下列四大类：

■ 相斥式集群法 (Disjoint Clustering)

此法将每一个被分类的变量或观察体分到一个且唯一的一个集群中。

■ 层次式集群法 (Hierarchical Clustering)

此法从最基层的集群 (即每一个变量/观察体代表一个集群) 开始，逐步将这些集群合并而演变成一个最大的集群 (亦即将所有的变量/观察体都归属于同一个集群)。

■ 重叠式集群法 (Overlapping Clustering)

此法允许一个变量/观察体同时隶属于两个或两个以上的集群。

■ 模糊式集群法 (Fuzzy Clustering)

此法利用模糊集合论 (Fuzzy Set Theory)，将变量/观察体属于每一个集群的程度以概率来表示。它不像前三种方法以 0 (不属于该集群) 或 1 (属于该集群) 来表示变量 / 观察体属于一个集群的程度。模糊集群法可以是相斥式，阶层式，或重叠式的。

42.3　何种数据适用于集群分析

下列两种数据适用于集群分析：

(1) 一个正方矩阵，其行与列均代表被分类的观察体 (或变量)。矩阵的元素则代表行列间的相似性 (Similarity) 或距离 (Distance)。相关系数的矩阵就是一个代表的例子。

(2) 一个长方形的多变量矩阵，其行代表变量，列代表观察体。无论是变量或观察

体都可以用集群法分类。譬如说老师用的学生成绩簿，这本成绩簿的行代表学生各次考试的成绩 (即变量)，其列代表学生的学号 (即观察体)。所以，成绩簿所记载的数据就是一个多变量矩阵。

42.4　九种集群分析程序概述

九种集群分析的 SAS 程序简介如下：

■ PROC CLUSTER

执行层次式集群法，有十一种运算方式。输入资料文件的数据必须是距离矩阵或观察体的坐标矩阵；因此，相关系数矩阵必须先转换成距离矩阵才能被处理。

■ PROC FASTCLUS

利用 K-平均数法 (K-Means) 对变量 / 观察体进行相斥式集群分析。此程序最适用于含十万笔以上的大型资料。

■ PROC VARCLUS

对变量作阶层式或相斥式的分类。

■ PROC TREE

用来画集群法的分类图，此图称为树形图 (Dendrogram) 或现象图 (Phenogram)。数据首先必须用 PROC CLUSTER 或 PROC VARCLUS 处理 ，然后将处理过后的数据送入 PROC TREE 制图。PROC TREE 的结果是另一个 SAS 资料文件，其内容包括了集群的成员与成员所属的阶层 (这样的结果只可从层次式集群法获得!)。

■ PROC IPFPHC

把一个交换流程图 (Transaction Flow) 的元素分类以便形成阶层式的集群。有关这个程序的指令，请查阅 SUGI Supplemental Library User's Guide (1983 年版或最新版)。

■ PROC OVERCLUS

从相似数据的矩阵中找出重叠式的集群。有关这个程序的指令，请查阅 SUGI Supplemental Library User's Guide (1983 年版或最新版)。

除此之外，下列程序可用来处理输入资料文件的结构，使其适用于集群法：

■ PROC ACECLUS

利用数据的坐标矩阵来预测集群内的共变异数，所测出的典型变量值 (Canonical Variable Score) 可用作下一步的集群分析 (见第 47 章的介绍)。

■ PROC PRINCOMP

如第 34 章所介绍，此程序是用来进行主成份分析的，其输出值是主成份值。

■ PROC STANDARD

将所有变量依指定的平均数与变异数标准化，详细内容请见第 10 章。

集群分析常见的别名

集群分析常见的别名如下：

数理分类法 (Numerical Taxonomy)、 Q 分析法 (Q-Analysis)、分节法 (Partitioning)、拓扑法 (Typology)、自由原型识别法 (Unsupervised Pattern Recognition)、分类法 (Classification)、系统方法 (Systematics)、团摺法 (Clumping)、计程学 (Taximetrics)、 分类描述学 (Taxonorics)、花序分类学 (Botryology)、形状分类学 (Morphometrics)、疾病描述学 (Noxography)、疾病分类学 (Nosology)、菊状分类学 (Aciniformics) 及集群分类法 (Agminatics) 等。

参考书

适用于初学者的参考书有二：

(1) Everitt (1980)

(2) Massart & Kaufman (1983)

其它重要参考书有：

(1) Anderberg (1973)

(2) Sneath & Sokal (1973)

(3) Duran & Odell (1974)

(4) Hartigan (1975)

[Hartigan 的书包括许多执行集群分析的 FORTRAN 程序]

(5) Spath(1980)

(6) Tittrington，Smith，& Makov(1985)

(7) McLachlan & Baoford (1988)

专业论文有四：

(1) Milligan (1980)

(2) Milligan & Cooper (1983)

(3) Cooper & Milligan (1984)

[以上三文探讨集群法的不变属性]

(4) Blashfield & Aldenderfer (1978)

[广泛介绍集群法之相关文献]

42.5　如何进行变量的集群分析

在 SAS 系统里，读者可以用两种统计的方法来进行变量的集群与分类。一种方法是因子分析法，另一种是阶层式或相斥式的集群法。

因子分析法相当于重叠式的集群法，它的结果常是模糊不清的集群。欲避免此结果，读者可以用阶层式/相斥式的集群法 (PROC VARCLUS) 来做变量的分类。最理想的分析方法是将因子分析法及阶层式/相斥式集群法合并，以便检查是否有重叠的集群。如果必须要有重叠的集群，则可用因子分析法；否则，用阶层式 / 相斥式集群法。

下面的两行程序便是用阶层式 / 相斥式集群法处理因子分析的输出文件，以便进行变量的集群：

```
PROC FACTOR R=PROMAX SCORE OUTSTAT=VAR;
PROC VARCLUS INITIAL=INPUT PROPORTION=0;
```

在上面的例子中，PROC VARCLUS 利用因子分析的输出值 (亦即因子分数据数) 导出一个相关系数的正方矩阵，以便进行阶层式/相斥式的变量集群分析。PROPORTION=0 的功用在于防止 VARCLUS 程序把既已形成的集群加以分裂。

42.6　如何进行个体的集群分析

SAS 有两种集群法的程序可以用来执行个体的集群分析：它们是 PROC FASTCLUS 和 PROCCLUSTER。

PROC FASTCLUS 适用于大型的资料文件，而且规定读者事先决定集群的数目。

PROC CLUSTER 则不要求读者事先决定集群的数目，但分析的过程耗时。一般而言，读者可先用 FASTCLUS 程序把一组数据大致分成五十个以内的集群，然后用 CLUSTER 程序进行阶层式的集群分类。

42.7　个体集群法的比较

在统计学中，有几种常见的个体集群法。这一段所要讨论的主题便是这几种个体集群法究竟孰优孰劣。

研究理论

过去有不少研究利用完全随机的数字来比较十一种个体集群法的优劣 (见 Milligan 1981 年的论文)。一般而言，这些研究的结果均显示均连法 (Average Linkage) 与华滋最小变异数法 (Ward's Minimum Variance) 最优，而单连法 (Single Linkage) 最劣。但实际上每一种方法各有其优劣点。

这十一种个体集群法都各有其特性，而其适用性则视输入资料文件中集群的大小，数据分布的形状/密度而定。下表可帮助读者决定当用何种个体集群法来进行分类的分析：

在输入资料文件中，若	则我们建议你采用
各集群之成员数相近，	●K-平均数法或华滋最小变异数法
各集群内的变异数相近，	●均连法
集群内成员的分布形状是长型，椭圆型，或不规则型，	●单连法或密连法(Density Linkage)
集群内成员的分布形状趋向圆曲线型，	●下列六方法中的任何一种：双连法，中数法，臻连法，ML 法,弹性法，及马氏法 (详细内容见第 43 章第 43.2 节)

建议

我们建议读者最好同时采用两种以上的集群法来做个体的分类，其中的一种方法最好是密连法。这样，读者可以参考比较分析结果，而达成较客观的结论。

例 1

如果输入资料文件内三个集群之间的距离够大 (亦即集群易于分辨)，则即便使用所谓“最劣”的单连法也可找出这三个相当分散的集群。在这种情况下，不用密连法亦无妨，请看下面示范的单连法分析结果：

程　序

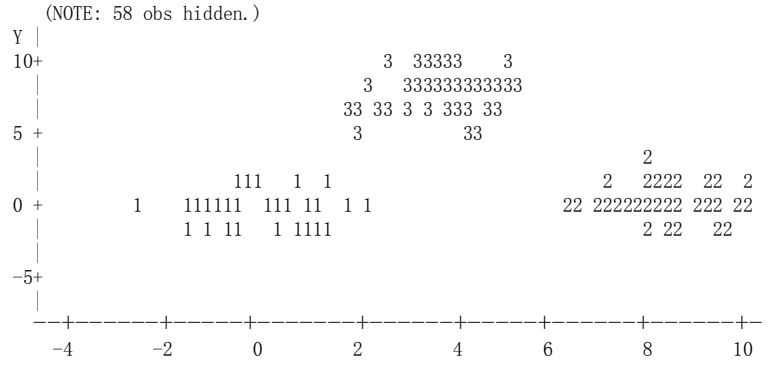
```
DATA COMPACT;
    KEEP X Y; N=50;  SCALE=1;
    MX=0; MY=0; LINK GENERATE;
    MX=8; MY=0; LINK GENERATE;
    MX=4; MY=8; LINK GENERATE;       STOP;
    GENERATE:
        DO I=1 TO N;
           X=RANNOR(1)*SCALE+MX;
           Y=RANNOR(1)*SCALE+MY;       OUTPUT;
        END;          RETURN;
PROC CLUSTER DATA=COMPACT OUTTREE=TREE METHOD=SINGLE NOPRINT;
PROC TREE NOPRINT OUT=OUT N=3;   COPY X Y;
PROC PLOT; PLOT Y*X=CLUSTER;
    TITLE 'SINGLE LINKAGE CLUSTER ANALYSIS';
    TITLE2 'OF DATA CONTAINING WELL-SEPARATED,COMPACT CLUSTERS';
RUN;
```

结　果

报表 42.1　单连法的分析结果

SINGLE LINKAGE CLUSTER ANALYSIS
OF DATA CONTAINING WELL-SEPARATED,COMPACT CLUSTERS

Plot of Y*X. Symbol is value of CLUSTER.



例 2

如果我们把上例中三个集群的距离拉近，然后利用 K-平均数法 (以 PROC FASCLUS 执行) 以及五种个体集群法来分类 (以 PROC CLUSTER 执行)，则各集群法所导出的结果迥异。这五种个体集群法是华滋最小变异数法，均连法，重心法 (Centroid Method)，双连法 (Two-Stage DensityMethod) 及单连法。

程　序

```
DATA CLOSER;  KEEP X Y;
     N=50; SCALE=1;
     MX=0; MY=0; LINK GENERATE;
     MX=3; MY=0; LINK GENERATE;
     MX=1; MY=2; LINK GENERATE;  STOP;
     GENERATE: DO I=1 TO N;
               X=RANNOR(9)*SCALE+MX;
               Y=RANNOR(9)*SCALE+MY;
               OUTPUT;
               END;
             RETURN;
PROC FASTCLUS DATA=CLOSER OUT=OUT  MAXC=3  NOPRINT;
PROC PLOT; PLOT Y*X=CLUSTER;
     TITLE 'FASTCLUS ANALYSIS';
     TITLE2 'OF DATA CONTAINING POORLY-SEPARATED,COMPACT CLUATERS';
PROC CLUSTER DATA=CLOSER OUTTREE=TREE METHOD=WARD NOPRINT;
PROC TREE NOPRINT OUT=OUT  N=3;  COPY X Y;
PROC PLOT; PLOT Y*X=CLUSTER;
     TITLE 'WARD"S MINIMUM VARIANCE CLUSTER ANALYSIS';
     TITLE2 'OF DATA CONTAINING POORLY-SEPARATED,COMPACT CLUSTERS';
PROC CLUSTER DATA=CLOSER OUTTREE=TREE METHOD=AVERAGE NOPRINT;
PROC TREE NOPRINT OUT=OUT  N=3 DOCK=5;  COPY X Y;
PROC PLOT; PLOT Y*X=CLUSTER;
     TITLE 'AVERAGE LINKAGE CLUSTER ANALYSIS';
     TITLE2 'OF DATA CONTAINING POORLY-SEPARATED,COMPACT CLUSTERS';
PROC CLUSTER DATA=CLOSER OUTTREE=TREE METHOD=CENTROID NOPRINT;
PROC TREE NOPRINT OUT=OUT  N=3 DOCK=5;  COPY X Y;
PROC PLOT; PLOT Y*X=CLUSTER;
     TITLE 'CENTROID CLUSTER ANALYSIS';
     TITLE2 'OF DATA CONTAINING POORLY-SEPARATED,COMPACT CLUSTERS';
PROC CLUSTER DATA=CLOSER OUTTREE=TREE METHOD=TWOSTAGE K=10 NOPRINT;
PROC TREE NOPRINT OUT=OUT  N=3 ;  COPY X Y;
PROC PLOT; PLOT Y*X=CLUSTER;
```

```
    TITLE 'TWO-STAGE DENSITY LINKAGE CLUSTER ANALYSIS';
    TITLE2 'OF DATA CONTAINING POORLY-SEPARATED,COMPACT CLUSTERS';
PROC CLUSTER DATA=CLOSER OUTTREE=TREE METHOD=SINGLE NOPRINT;
PROC TREE DATA=TREE NOPRINT OUT=OUT   N=3   DOCK=5;   COPY X Y;
PROC PLOT; PLOT Y*X=CLUSTER;
    TITLE 'SINGLE LINKAGE CLUSTER ANALYSIS';
    TITLE2 'OF DATA CONTAINING POORLY-SEPARATED,COMPACT CLUSTERS';
RUN;
```

结　果

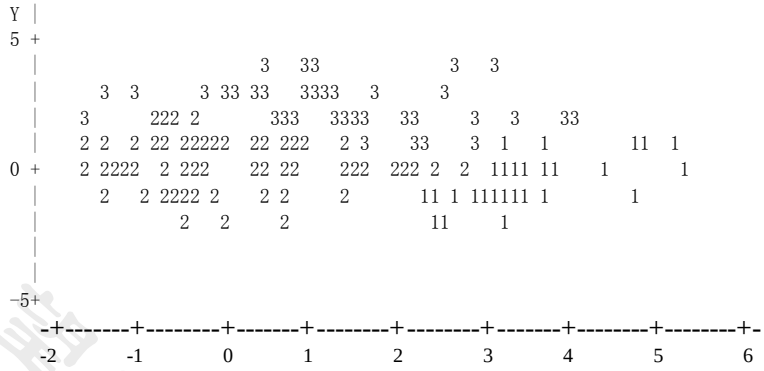
分析结果之比较显示，K-平均数法与华滋最小变异数法所产生的结果最接近真实的集群 (见报表 42.2)。

报表 42.2　六个体集群法的比较

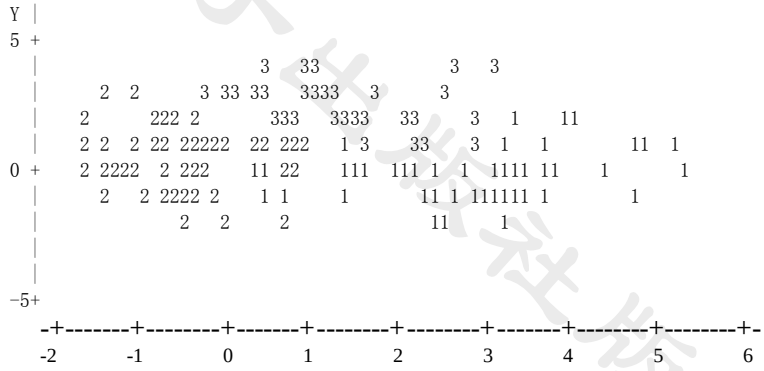
```
                FASTCLUS ANALYSIS  (K-平平均数法)
              OF DATA CONTAINING POORLY-SEPARATED,COMPACT CLUATERS
                   Plot of Y*X.  Symbol is value of CLUSTER.
    (NOTE: 33 obs hidden.)
Y |
5 +
  |                      2   22             2   2
  |      3  3      2 22 22   2222   2       2
  |    3      333 3       232   2222   22     2   2    22
  |    3 3  3 33 33333  33 333    3 2     22    2  1   1         11  1
0 +    3 3333  3 333    33 33     111  111 1  1  1111 11    1        1
  |      3    3 3333 3    3 3      1        11 1 111111 1         1
  |             3   3     3               11      1
  |
  |
-5+
  -+-------+--------+-------+--------+--------+-------+--------+--------+-
  -2       -1        0       1        2        3       4        5        6
                                      X

             WARD'S MINIMUM VARIANCE CLUSTER ANALYSIS  (华滋法)
           OF DATA CONTAINING POORLY-SEPARATED,COMPACT CLUSTERS
                Plot of Y*X.  Symbol is value of CLUSTER.
    (NOTE: 33 obs hidden.)
Y |
5 +
  |                      3   33             3   3
  |      2  2      3 33 33   3333   3       3
  |    2      222 2       333   3333   33     3   3    33
  |    2 2  2 22 22222  22 222    1 3     33    3  1   1         11  1
0 +    2 2222  2 222    22 22     111  111 1  1  1111 11    1        1
  |      2    2 2222 2    2 2      1        11 1 111111 1         1
  |             2   2     1               11      1
  |
  |
-5+
  -+-------+--------+-------+--------+--------+-------+--------+--------+-
  -2       -1        0       1        2        3       4        5        6
                                      X
```

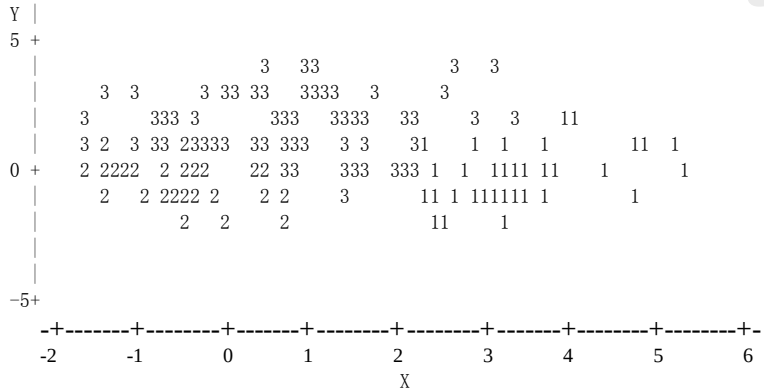
AVERAGE LINKAGE CLUSTER ANALYSIS (均连法)
OF DATA CONTAINING POORLY-SEPARATED,COMPACT CLUSTERS
Plot of Y*X. Symbol is value of CLUSTER.
(NOTE: 33 obs hidden.)

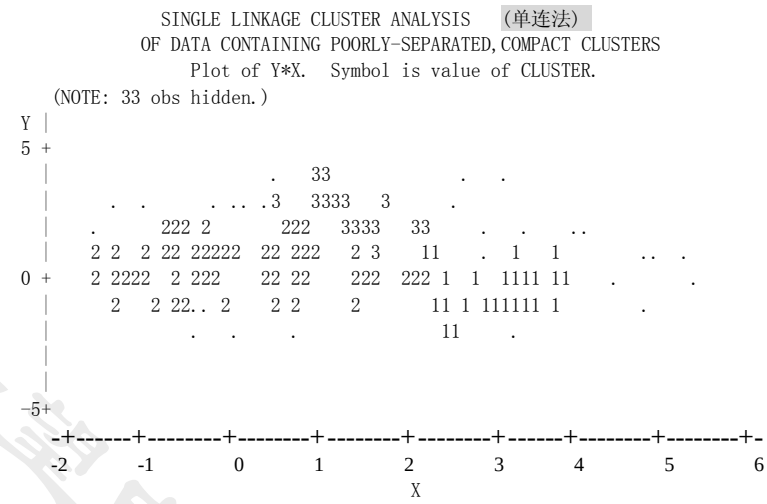


CENTROID CLUSTER ANALYSIS (重心法)
OF DATA CONTAINING POORLY-SEPARATED,COMPACT CLUSTERS
Plot of Y*X. Symbol is value of CLUSTER.
(NOTE: 33 obs hidden.)



TWO-STAGE DENSITY LINKAGE CLUSTER ANALYSIS (双连法)
OF DATA CONTAINING POORLY-SEPARATED,COMPACT CLUSTERS
Plot of Y*X. Symbol is value of CLUSTER.
(NOTE: 33 obs hidden.)





42.8　如何决定集群 (Clusters) 的个数

在集群法里，要决定集群的个数比在因子分析法里决定因子的个数更加困难。下面提供三点意见来帮助读者做决定。

执行 K 个最近邻集群法

Wong 与 Schaack (1982) 所提出的 K 个最近邻集群法是目前为止最成功的鉴别法。它的优点是结果可靠，统计假设少，计算省时省力。其计算过程如下：

(1) 读者先自行决定集群之成员数，以 K 值代表。

(2) 根据这个 K 值，此法推测资料文件内到底有多少个聚集点 (Modes)。

(3) 根据 (2) 所推测的聚点数，读者另订一 个 K 值。

(4) 根据新定的 K 值，此法再推测聚集点，看聚点数是否改变。

(5) 若聚集点的数目一直维持在一个常数，则我们可下结论说这个资料文件中大概就有这种个集群。

若聚集点的数目一直在改变， 则重复 (2)—(4) 的步骤。

利用 CLUSTER 程序里的指令 METHOD=DENSITY 与不同的 K 值配合，就可将此法应用在数据上。请看下面指令的示范：

```
PROC CLUSTER METHOD=DENSITY K=15;
```

R2 值

执行变量的集群时，R^2 值愈大，则分类愈成功，亦即集群的数目愈正确。因此，读者最好作一系列的尝试，再由各结果里取 R^2 值最大的分类结果。

CCC 指标

如果我们愿意接纳均等分配的虚无假设，则 Sarle (1983) 所提出的 CCC 指标 (Cubic Clustering Criterion) 可将母群中集群的聚集点 (如正方型的集群) 找出。有关 CCC 之程序撰写及解释请看第 44 章 (PROC FASTCLUS 程序) 的例子。

第 43 章　阶层式集群分析：统计程序 PROC CLUSTER

43.1　阶层式集群法的分类过程

阶层式集群法 (以下简称阶集法) 以数据间的距离或相近程度为集群分类的根据。阶集法的分类是一个循序渐进的过程：由最小的集群 (只含一个数据单元) 开始，将最相近的两个集群合并为一新集群，逐渐归并，最后所有的数据会并成一个最大的集群 (包括所有的数据单元)，即分析的结果。

43.2　SAS 系统内的十一种阶集法

SAS 目前有十一种阶集法可用来形成观察体的集群。这十一种方法是：

(1) 均连法　(Average Linkage Method)，
(2) 重心法　(The Centroid Method)，
(3) 近邻法　(Nearest Neighborhood) 又称
　　单连法　(Single Linkage Method)，
(4) 远邻法　(Furthest Neighborhood) 又称
　　臻连法　(Complete Linkage Method)，
(5) 密连法　(Density Linkage Method)，
(6) ML(EML) 法(Maximum Likelihood Method)，
(7) 弹性 β 法 (The Flexible-Beta Method)，
(8) 马氏法　(McQuitty's Method)，
(9) 中数法　(Median Method)，
(10) 双连法　(Two-Stage Density Method)，及
(11) 华滋法　(Ward's Minimum Variance Method)。

43.3　十一种阶集法的运算方式

根据 Anderberg (1973) 的分类，阶集法的运算方式分三种，即：数据储存式，距离储存式，及距离分类式。下表列举了 PROC CLUSTER 程序中十一种阶集法及它们运算的方式。

阶集法名称 (代号)	数据储存	距离储存	距离分类
均连法 (AVE)	×	×	
重心法 (CEN)	×	×	
单连法 (SIN)		×	

臻连法	(COM)		×	
中数法	(MED)		×	
密连法	(DEN)			×
ML 法	(EML)	×		
弹性β法	(FLE)		×	
马氏法	(MCQ)		×	
双连法	(TWO)			×
华滋法	(WAR)	×	×	

上述十一种阶集法中，有八种只用一种运算法；其余三种 (即均连、重心及华滋法) 用到两种运算法。这是因为这三种方法的输入数据可以是坐标值 (用数据储存)，亦可以是欧氏距离 (用距离储存) 的缘故。若读者手边的资料属于相似性资料 (如相关系数矩阵)，则必须将它们先转换成距离资料 (如用 1 减去相关系数平方)，否则 PROC CLUSTER 无法直接处理或分析相似性资料。

注：SAS 6.06 或 6.07 版的 CLUSTER 程序允许读者将数据自磁盘空间或存储器中读进来。计算过程中所产生的距离亦可储存在磁盘或存储器内。

43.4　如何撰写 PROC CLUSTER 程序

CLUSTER 程序只执行阶层式集群法。输入的资料可以是数据点的坐标，也可以是数据点间的欧氏距离。至于输出的结果则包括：各集群所属的阶层，每一阶层中所组成的新集群，集群中元素间的最大距离等。这些输出的结果可一种包括在另一个输出资料文件内以供其它的程序 (如 PROCTREE) 作进一步的处理或制树形图。PROC CLUSTER 含七道指令，它们的格式如下：

```
PROC CLUSTER    选项串;
    VAR         变量名称串;
    ID          变量名称;
    COPY        变量名称串;
    FREQ        变量名称;
    RMSSTD      变量名称;
    BY          变量名称串;
```

这七道指令中，以 PROC CLUSTER 选项串；最重要，它是本程序精华之所在，另外六道指令则可有可无。

指令 #1　PROC CLUSTER 选项串;

包含下列四种选项及 (5) NOTIE 选项：

(1) DATA=　　输入资料文件名称

(2) OUTTREE=　输出资料文件名称

(3) METHOD=　　一种阶集法的名字

(4) 用来控制集群分析过程打印的八个选项。

(5) NOTIE

这几种选项中，以第三种选项 METHOD= 最重要，一定要加以界定。其余各选项则可有可无，现依序说明于下面：

(1) DATA=输入资料文件名称

此选项告诉 SAS 到底要用那一个资料文件来进行集群分析。若读者省略此选项，则 SAS 会自动找出在此程序之前最后形成的资料文件，并对它执行集群分析。

如果输入资料文件含欧氏距离 (即 TYPE=DISTANCE)，则此资料文件一定要是一个对角线对称的正方矩阵。如果输入资料文件不是欧氏距离，则 SAS 假设数据是欧几里得空间中的坐标，由这些坐标值先导出欧氏距离，然后才进行集群分析。一般而言，无论输入的数据是坐标或是欧氏距离，SAS 的各种阶集法所导出的结果是完全相同的。

(2) OUTTREE=输出资料文件名称

读者可用此选项将阶集法的分析结果存入一个适当的输出资料文件中，以供制作树形图 (制图所须的统计程序是 PROC TREE；见第 46 章的说明)。这个输出资料文件的命名必须遵循双重命名的原则 (例：OUTTREE=OUT.DATA)。

如果读者省略此选项，则 SAS 会自动为输出资料文件命名 (如 DATA1，DATA2，... 等)，而此资料文件将会在关机后消失。

如果读者欲保留输出资料文件，但不想对它再作进一步的分析，则写 OUTTREE=_NULL_。

(3) METHOD=一种阶集法的名字

请读者注意，METHOD=(或 M=) 这一个选项非常重要，不可以省略。它的功用是告诉 SAS 在十一种阶集法中使用那一种方法来执行集群分析。所以一个 CLUSTER 程序中，只能有一个 METHOD= 选项。如果读者想用三种阶集法来分析同一个输入数据组，则你必须用三个 CLUSTER 程序来分别包括这三个 METHOD= 选项。

选项 METHOD= 之撰写

以下是十一种阶集法的选项撰写、简称及有关之次级选项。这十一种方法按其字母顺序排列分述如下 (括号内是简称)：

(A) METHOD=AVERAGE (M=AVE)

要求 SAS 执行均连法的集群分析。当输入资料文件是欧氏距离时，SAS 会自动先将这些距离加以平方，然后用均连法来分析此平方距离。若读者想分析原距离 (而非平方距离)，则你必须同时使用次级选项 NOSQUARE [见后面 (a) 的说明]。

(B) METHOD=CENTROID (M=CEN)

要求 SAS 执行重心法的集群分析。若输入资料文件含欧氏距离，此法的 SAS 处理过程与均连法相同，所以读者也可能会用到次级选项 NOSQUARE[见后面 (a) 的说明]。

(C) METHOD=COMPLETE (M=COM)

要求 SAS 执行臻连法 (亦作远邻法) 的集群分析。此法应与 TRIM=的次级选项合用 [见后面 (h) 的说明]。

(D) METHOD=DENSITY (M=DEN)

要求 SAS 执行密连法的集群分析。密连法是一个通称，它代表一系列无参数的集群法。读者必须用次级选项 K= 或 R= 或 HYBRID [见后面 (b)，(c)，(d) 的说明]来决定一个特定的无参数分析法。另外 MODE= 及 DIM= 两次级选项 [见(e)，(f)的说明] 也可与此法并用。

(E) METHOD=EML (M=EML)

要求 SAS 执行 ML 法 (意即最大可能率法) 的集群分析。如选用此法，则输入数据必须是坐标值。此法只适用于多元常态分配，等值变异数，而且不等混合比例 (Unequal Mixing Proportions) 的数据群。此法需与 PENALTY= 次级选项并用 [见后面 (i) 的说明]。

(F) METHOD=FLEXIBLE (M=FLE)

要求 SAS 执行弹性 β 法的集群分析，与 BETA= 次级选项并用[见后面 (g)的说明]。

(G) METHOD=MCQUITTY (M=MCQ)

要求 SAS 执行马氏法的集群分析。其理论基础是马氏的均连法，详细内容见 McQuitty (1966)。

(H) METHOD=MEDIAN (M=MED)

要求 SAS 执行中数法的集群分析。若输入资料文件是欧氏距离，此法的处理过程与均连法相同，所以读者也可能会用到次级选项 NOSQUARE [见后面 (a) 的说明]。

(I) METHOD=SINGLE (M=SIN)

要求 SAS 执行单连法 (亦作近邻法) 的集群分析。为了避免连锁式的集群效果，此法应与次级选项 TRIM = [见后面 (h) 的说明] 并用。

(J) METHOD=TWOSTAGE (M=TWO)

要求 SAS 执行双连法的集群分析。读者必须用 K= 或 R= 或 HYBRID 次级选项[见后面 (b)，(c)，(d) 的说明] 来指定一个特定的密度推测法。另外 MODE= 及 DIM=两个次级选项 [见后面 (e)，(f) 的说明] 也可与此法并用。

(K) METHOD=WARD (M=WAR)

要求 SAS 执行华滋法的集群分析。若输入数据是欧氏距离，则其处理过程与均连法相同，所以读者可能也会用到次级选项 NOSQUARE [见下面 (a) 的说明]。另外，此法应与次级选项 TRIM= 合用 [见后面 (h) 的说明]。

有关 METHOD= 之次级选项

在上一节中，我们曾提到九种有关 METHOD= 之次级选项。在本节内，我们将按照这九个次级选项所属的阶集法，分别叙述之：

(a) NOSQUARE

用来抑止数据点间的欧氏距离被平方。这个次级选项可与四种集群法并用：均连法(M=AVE)、重心法 (M=CEN)、中数法 (M=MED)，及华滋法 (M=WAR)。

若读者选用密连法 (M=DEN) 或双连法 (M=TWO)，则必须选择下列三个次级选项之一，以界定一个推测密度的方法 (在 SAS 6.06 版下内设的密度等于最大值

100)。

请读者注意，你必须在下列三个选项中选择一项，而且只能选择其一：

(b) K=n

n 是一个介于 2 与观察体个数之间的整数，如：K=4。K 值是集群的成员数，用来做 K 个最近邻集群法之推测。[请参照 TRIM= 及 MODE= 两个次级选项。]

(c) R=n

n 值代表集群的半径。[请参照 TRIM= 的次级选项。]

(d) HYBRID

此次级选项要求 SAS 执行王氏 (Wong，1982) 杂交集群法。此法利用 K-平均数法将其它集群法的分析结果再做一次分析，以找出可能的集群。输入数据组必须是下列这两种之一：

- PROC FASTCLUS 的输出资料文件 (见第 44 章的说明)。
- 另一个 PROC CLUSTER 的输出资料文件。这个资料文件中必须含算术平均数、次数和每一集群的标准差等资料。请读者注意，HYBRID 次级选项不可与另一个次级选项 TRIM= 合用。

当读者选用密连法 (M=DEN) 或双连法 (M=TWO) 时，还可以选取下列两个次级选项：

(e) MODE=n

适当两个集群联合时，每一个集群至少要含 n 个成员才可成为"模范" 集群。如果已在前面指定 K 值，则 SAS 自动设定 MODE 值等于 K 值；如果未在前面指定 K 值，则 SAS 自动设定 MODE=2。若程序内含 FREQ 指令或输入资料文件内含 FREQ 变量，则 n 值与即将要联合的集群内实际观察体的个数作比较，而非与集群内次数的总和作比较。

(f) DIM=n

界定计算密度预估值时所用的向量轴数目，与 M=DEN 或 M=TWO 以及 TRIM= 的次级选项并用。若读者省略此次级选项，而输入数据是坐标值，则 SAS 自动设定 n 值等于变量的数目；若数据是距离，则 SAS 会自动将 n 值定为 1。

当读者选用弹性 β 法 (M=FLE) 时，可以同时选用下面的次级选项：

(g) BETA=n

决定 METHOD=FLE 中 β 参数的值。此值必须小于 1，通常介于 0 与 -1 之间。若省略此次级选项，则 SAS 自动将 n 值定为-0.25。

当读者选用臻连法 (M=COM)、单连法 (M=SIN) 或华滋法 (M=WAR) 时，可以同时界定下列的次级选项：

(h) TRIM=p

很重要的一个次级选项，是用来剔除数据中过于分散的劣质数据 (Outliers)。被剔除的数据收集在 OUTTREE= 资料文件内，其在 _FREQ_ 变量上的值等于零。可与单连法 (M=SIN) 合用以防止连锁式的集群效果，从而产生较佳的集群分析。p 值可以小于 1 (解释成百分比，如 TRIM=.20 表示剔除 20% 的劣质数据)，p 值也可以大于 1 (加上 % 就成了百分比)。一般而言，TRIM=10 应该是一个合理的

估计。TRIM=的次级选项必须与 K=n 或 R=n 之一合用，但不可与 HYBRID 合用。当 TRIM=p 与 M=COM 或 M=WAR 合用时，可避免分析结果受劣质数据影响而扭曲真象。若与 M=SIN 合用，则可避免连锁式的集群效果。其它的阶集法也可与 TRIM=p 合用而产生较佳的分析结果。请注意：如果 M=DEN 或 M=TWO 已与 HYBRID 合用，则不可再与 TRIM=p 并用。

(i) PENALTY= 正实数

界定惩罚系数 (Penalty Coefficient)，此系数用以调整由于集群大小不等所导致的偏差。内设值是 2。

(4) 用来控制集群分析过程打印的八个选项 (括号内是简称)：

(甲) PRINT=n (P=n)

印出阶集分析法中前 n 个集群形成的过程。n 值可以是 0 (表示不印出集群形成的过程)，或是大于 0 的整数。内设值 (Default Value) 是印出整个阶层式集群分析的过程。

(乙) NOPRINT

指示 SAS 不必印出集群分析的结果。

(丙) NOID

指示 SAS 不必印出各阶层中新形成的集群识别代号。

(丁) NONORM

防止数据被标准化。当此选项与 M=WAR 并用时，其集群间的变异数可以免受总变异数的标准化。此选项对下列三种集群法无影响力：M=DEN，M=EML 或 M=TWO。

(戊) PSEUDO

指示 SAS 印出近似的 F 值和 t^2 值。只有当输入数据是坐标值 (然而 METHOD 不等于 SINGLE) 或阶集法是 AVE，CEN 或 WAR 时，才可选用此选项。

(己) RMSSTD

指示 SAS 印出每一个集群中的标准差。和 PSEUDO 的规定一样，只有当输入数据是坐标值或阶集法是 AVE，CEN 或 WAR 时，才可以用 RMSSTD 选项。

(庚) RSQUARE (RSQ)

指示 SAS 印出 R 平方值和半净相关系数的平方值。当输入数据是坐标值时，RSQUARE 可与 M=AVE 或 M=CEN 合用。

(辛) CCC

指示 SAS 印出 CCC 指标 (Cubic Clustering Criterion) 及在均等虚无假设下的 R2 近似值。此选项只适用于坐标值的输入数据，不适合与单连法 (METHOD=SINGLE)并用。

(5) NOEIGEN

抑制 CCC 指标之特性根的计算。此选项只适用于坐标数据，或变量数目过于庞大的资料文件，因为 NOEIGEN 可节省电脑计算的时间。

(6) NOTIE

要求 CLUSTER 程序在执行的过程里，不考虑重叠的数据。此选项适用于比较精密的资料而且它会节省许多电脑的存储器与分析时间。

(7) SIMPLE (或 S)

要求印出描述性统计值，此选项只适用于坐标数据。

(8) STANDARD

将变量标准化，使其平均数等于 0，标准差等于 1。此选项只适用于坐标数据。

两个 PROC CLUSTER 的例子

由于 PROC CLUSTER 指令是整套程序之精华，在此特地举一些示范的例子供读者参考。

例 1

```
PROC CLUSTER DATA=MILEAGES
     OUTTREE=OUT_DATA
     METHOD=AVERAGE NOSQUARE TRIM=10
     NOID;
```

例 1 中，PROC CLUSTER 指示 SAS 执行阶层式集群法。

DATA=MILEAGES 告诉 SAS 分析 MILEAGES 这个输入资料文件；OUTTREE=OUT_DATA 告诉 SAS 将分析的结果存入一个叫 OUT_DATA 的输出资料文件中；第三行要求 SAS 执行均连法，指示 SAS 不可将资料文件内的距离平方，并剔除 10% 的劣质数据。最后，NOID 选项指示 SAS 不必印出集群分析过程中新形成之集群的识别代号。此程序以分号结尾。

例 2

```
PROC CLUSTER METHOD=DENSITY HYBRID;
```

例 2 的程序中只指定了 METHOD 选项，而省略了 DATA=，OUTTREE= 和 PRINT=n 选项，所以 SAS 会做这样的处理：将 PROC CLUSTER 之前最后形成的资料文件引进，对它做密连法的集群分析 (并指明用王氏杂交集群法)；再将整个分析的过程在报表上印出；然其结果只被存入一个暂时的文件，故在关机后这个文件也会跟着消失。

指令 #2　VAR 变量名称串;

VAR 指令列出资料文件内用来作集群分析的变量名称串。若省略此指令，则 SAS 会自动将其它指令中未曾用过的所有数值变量找出，加以分析。VAR 指令的重要性仅次于 PROC CLUSTER。通常，在一个 CLUSTER 程序中可以只含 PROC CLUSTER 及 VAR 指令，而省略其它的指令。

指令 #3　ID 变量名称串;

ID 指令指示 SAS 在报表上及输出资料文件内，将每一个观察体编号以便于识别。若省略此指令，则 SAS 自动将各观察体编号，以 OBn 代表 (n 值等于所有观察体的个

数)。

指令 #4　COPY 变量名称串;

COPY 指令会将输入资料文件的变量复印到输出资料文件里 (亦即指令 OUTTREE= 所界定的输出资料文件)。

指令 #5　FREQ 变量名称;

FREQ 变量的值 (正整数) 代表观察体重复出现的次数。CLUSTER 程序根据这些值，自动将输入资料文件中的每一变量重复计算 (如 10 次，或 28 次)。

如果省略 FREQ 指令，但其输入资料文件中包含一个叫 _FREQ_ 的变量，则 _FREQ_ 变量的作用与上述 FREQ 指令相同。若读者既省略 FREQ 指令，而其输入资料文件中又无 _FREQ_ 变量，则 SAS 会自动认定各数据的出现次数为 1。另外，HYBRID 次级选项必须与 FREQ 指令或 _FREQ_变量合用。

若输入资料文件中的数据代表一个集群 (例如，将 PROC FASTCLUS 的输出资料文件转成 PROCCLUSTER 的输入资料文件)，则 FREQ 指令的值就是各集群所包含的成员数目。

指令 #6　RMSSTD 变量名称;

在下列情况中，我们建议读者一定要在 CLUSTER 程序中包括 RMSSTD 指令以取得精确的统计值：

(1) 当你选用均连、重心或华滋法执行集群分析，而且

(2) 输入资料文件具备下列三个条件：

- ●数据的坐标值代表集群的平均数 (例如，将 PROC FASTCLUS 的输出资料文件转成 PROCCLUSTER 的输入资料文件);
- ●含 _FREQ_ 变量;
- ●含一个代表集群内标准差的变量。

读者必须在 RMSSTD 指令中指出代表标准差的变量名字。并且，RMSSTD 指令须与 FREQ 指令合用。另外，RMSSTD 指令也可被输入资料文件中的 _RMSSTD_ 变量所取代。以外，HYBRID 次级选项必须与 RMSSTD 指令或 _RMSSTD_ 变量并用。

指令 #7　BY 变量名称串;

SAS 依据此指令所列举的变量将资料文件分成几个小的资料文件，然后对每一个小的资料文件分别执行阶集法的分析。 当读者选用此指令时，资料文件内的数据必须先依 BY 变量串的值做由小到大的重新排列。这个步骤可藉 PROC SORT 达成。

43.5　输出资料文件的进一步说明

在前面 43.4 节里，我们曾经简单的介绍了 PROC CLUSTER 指令中的 OUTTREE 选项。现在让我们对该输出资料文件做更进一步的说明。

OUTTREE= 输出资料文件通常含 2N-1 个观察体，N 代表原输入资料文件的观察体个数。新添的 N-1 个观察体代表阶层式集群分析所产生的树节 (Tree Node)。树节也就是分析过程中每一阶层所产生的集群。

OUTTREE= 资料文件内含的变量有下列数据：

(1) BY，　　BY 指令中的变量名称串。

(2) _NAME_，　　代表树节。当树节是一个观察体时，其值由该观察体的识别代号来表示 (这个识别代号可由 ID 指令来界定或由内设的 OB_n 界定)。当树节是一个集群时，其值由集群的名字来表示 (如 CL2 或 CL7)。

(3) _PARENT_，　　树形图上每一个树节上面一层的树节名称。

(4) _NCL_，　　集群的总个数。

(5) _FREQ_，　　每一个阶层集群的成员数。

(6) _HEIGHT_，　　相邻两阶层间树节的距离 (或相似的程度)。这个值也就是树形图上的纵坐标。

(7) ID，　　任何观察体的识别代号。

(8) COPY　　COPY 指令中的变量名称串。

若输入资料是坐标值，而且集群分析的方法是 AVERAGE，CENTROID，或 WARD，则下列两个变量也被纳入输出资料文件：

(9) _DIST_，　　两集群相连之前其平均数间的欧氏距离。

(10) _AVLINK_，　两集群相连之前其成员间的平均欧氏距离。

当输入资料文件内含坐标值 (或当 METHOD=AVE，CEN，WAR) 时，下列的变量自动被纳入输出资料文件：

(11) _RMSSTD_，集群内成员间距离的标准差。

(12) _SPRSQ_，　当两个集群相连后，集群内可被解释的变异数百分比减低的程度 (也等于半净相关系数的平方)。

(13) _RSQ_，　　复相关系数的平方。

(14) _PSF_，　　Pseudo F 值。

(15) _PST2_，　　Pseudo t2 值。

当 METHOD=EML 时，下列的变量自动出现在输出资料文件内：

(16) _LNLR_，　　对数可能比 (Log Likelihood Ratio)。

当 METHOD=DEN 时，下列两个变量可出现在输出资料文件内：

(17) _DENS_，　　集群内的最大密度，与 K= 或 R= 次级选项并提。

(18) _MODE_，　　集群内可能隐含的凝聚力强的小集群 (Modal Cluster) 之个数。

当 METHOD=TWO 时，只有上述 _MODE_ 出现在输出资料文件内。

最后，若输入资料文件包含坐标值，则下列变量会出现在输出资料文件中：

(19) 坐标轴的名称。

(20) _ERSQ_，　　在均等分配的假设下，求出之复相关系数平方的近似值。

(21) _RATIO_，　　等于 (1-_ERSQ_)/(1-_RSQ_)。

(22) _LOGR_，　　上述 _RATIO_ 的自然对数。

(23) _CCC_，　代表 CCC 指标 (或作 Cubic Clustering Criterion)。

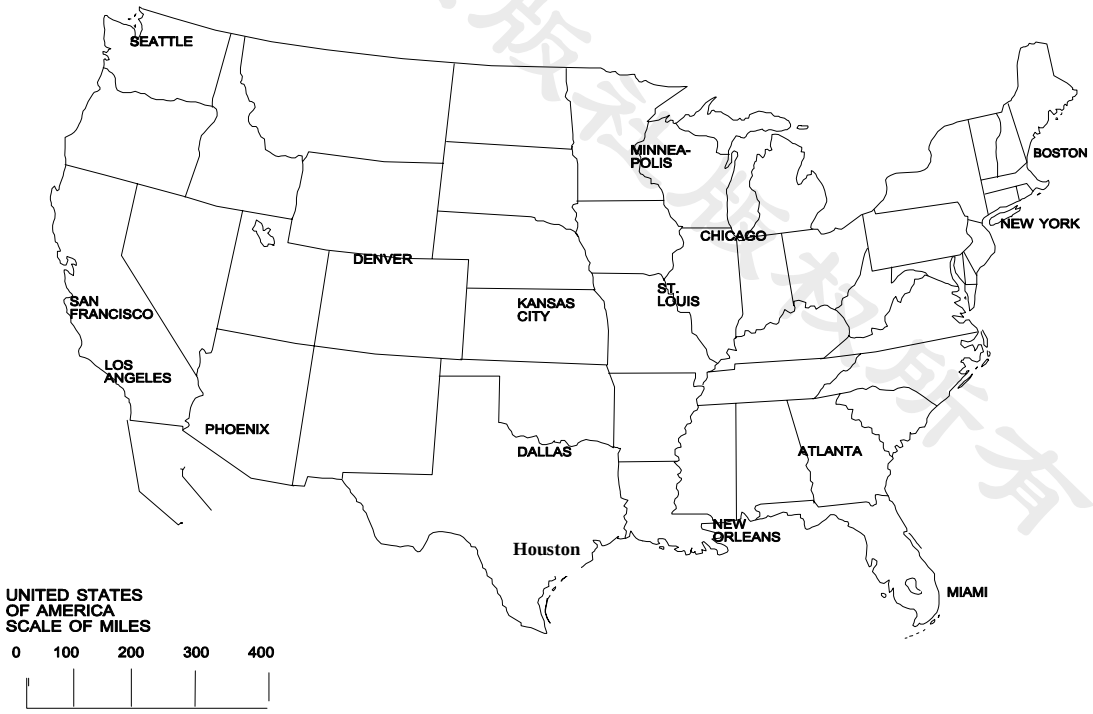
43.6 范　例

例一：美国十个城市的分类

这个例子根据两个城市间的航空距离将美国十个大城市作分类。这十个城市是：亚特兰大(Atlanta)、芝加哥 (Chicago)、丹佛 (Denver)、休斯顿 (Houston)、洛杉矶 (Los Angeles)、迈阿密 (Miami)、纽约 (New York)、旧金山 (San Francisco)、西雅图 (Seattle) 及华府 (Washington D.C.)。输入资料是欧氏距离。此例利用了 CLUSTER 程序中六种集群法来分析。在十一种方法中，ML 法只能处理含坐标值的输入数据，故不能用于此例。臻连法与弹性β法所分析出来的集群结果与华滋法很类似，故只采用华滋法。同理，此例采用均连法、重心法而省略马氏法及中数法。

一般而言，此六种集群法分析的结果显示东、西两岸的大城市各自形成两大集群；而丹佛、休斯顿似乎应另成一个集群。解释此结果时，读者应参阅下图所示美国十个城市的原址。

地　图



程　序

```
    TITLE 'CLUSTER ANALYSIS OF FLYING MILEAGES BETWEEN 10 AMERICAN CITIES';
    DATA MILEAGES(TYPE=DISTANCE);
          INPUT (ATLANTA CHICAGO DENVER HOUSTON LOSANGEL
                MIAMI NEWYORK SANFRAN SEATTLE WASHDC) (5.)
                @55 CITY $ 15.;
         CARDS;
       0                                                    ATLANTA
     587    0                                              CHICAGO
    1212  920    0                                        DENVER
     701  940  879    0                                   HOUSTON
    1936 1745  831 1374    0                             LOS ANGELES
     604 1188 1726  968 2339    0                        MIAMI
     748  713 1631 1420 2451 1092    0                   NEW YORK
    2139 1858  949 1645  347 2594 2571    0              SAN FRANCISCO
    2182 1737 1021 1891  959 2734 2408  678    0         SEATTLE
     543  597 1494 1220 2300  923  205 2442 2329    0    WASHINGTON D.C.
;
    PROC CLUSTER DATA=MILEAGES METHOD=AVERAGE PSEUDO; ID CITY;
PROC TREE HORIZONTAL SPACES=2; ID CITY;
    PROC CLUSTER DATA=MILEAGES METHOD=CENTROID PSEUDO; ID CITY;
PROC TREE HORIZONTAL SPACES=2; ID CITY;
    PROC CLUSTER DATA=MILEAGES METHOD=DENSITY K=3; ID CITY;
    PROC TREE HORIZONTAL SPACES=2; ID CITY;
    PROC CLUSTER DATA=MILEAGES METHOD=SINGLE;   ID CITY;
    PROC TREE HORIZONTAL SPACES=2; ID CITY;
PROC CLUSTER DATA=MILEAGES METHOD=TWOSTAGE K=3;   ID CITY;
    PROC TREE HORIZONTAL SPACES=2; ID CITY;
    PROC CLUSTER DATA=MILEAGES METHOD=WARD PSEUDO;   ID CITY;
    PROC TREE HORIZONTAL SPACES=2; ID CITY;
    RUN;
```

结　果

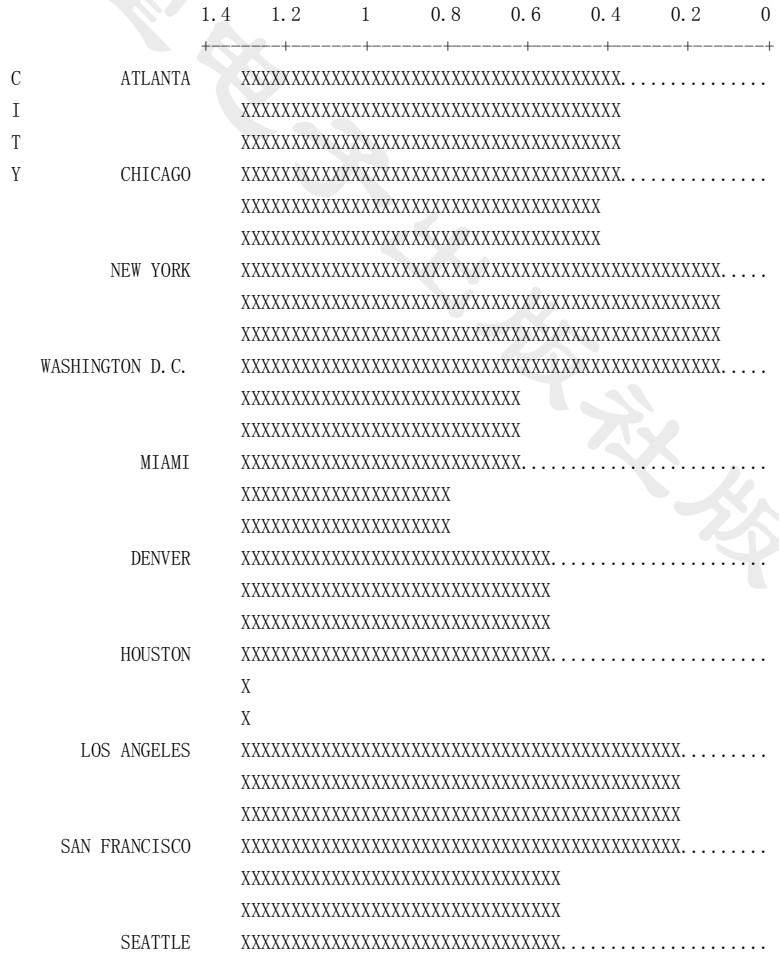
报表 43.1　美国十个城市的分类

```
CLUSTER ANALYSIS OF FLYING MILEAGES BETWEEN 10 AMERICAN CITIES
      Average Linkage Cluster Analysis  (均连法)
Root-Mean-Square Distance Between Observations   = 1580.242

                                                        T
                                Pseudo Pseudo Norm RMS i
```

NCL	Clusters Joined		FREQ	F	t**2	Dist	e
9	NEW YORK	WASHINGTON D.C.	2	66.7	.	0.129727	
8	LOS ANGELES	SAN FRANCISCO	2	39.2	.	0.219587	
7	ATLANTA	CHICAGO	2	21.7	.	0.371462	
6	CL7	CL9	4	14.5	3.4	0.414859	
5	CL8	SEATTLE	3	12.4	7.3	0.525534	
4	DENVER	HOUSTON	2	13.9	.	0.556244	
3	CL6	MIAMI	5	15.5	3.8	0.618457	
2	CL3	CL4	7	16.0	5.3	0.800540	
1	CL2	CL5	10	.	16.0	1.296665	

Average Distance Between Clusters



Centroid Hierarchical Cluster Analysis (重心法)

Root-Mean-Square Distance Between Observations = 1580.242

NCL	Clusters Joined		FREQ	Pseudo F	Pseudo t**2	Norm Cent Dist	T i e

9	NEW YORK	WASHINGTON D.C.	2	66.7	.	0.129727
8	LOS ANGELES	SAN FRANCISCO	2	39.2	.	0.219587
7	ATLANTA	CHICAGO	2	21.7	.	0.371462
6	CL7	CL9	4	14.5	3.4	0.365246
5	CL8	SEATTLE	3	12.4	7.3	0.513937
4	DENVER	CL5	4	12.4	2.1	0.533679
3	CL6	MIAMI	5	14.2	3.8	0.574270
2	CL3	HOUSTON	6	22.1	2.6	0.609053
1	CL2	CL4	10	.	22.1	1.173036

Distance Between Cluster Centroids

```
                1.2       1       0.8       0.6       0.4       0.2         0
                +--------+--------+--------+--------+--------+--------+
C          ATLANTA  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.................
I                   XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
T                   XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
Y          CHICAGO  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.................
                    XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
                    XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
          NEW YORK  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......
                    XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
                    XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
    WASHINGTON D.C. XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......
                    XXXXXXXXXXXXXXXXXXXXXXXXXXXX
                    XXXXXXXXXXXXXXXXXXXXXXXXXXXX
             MIAMI  XXXXXXXXXXXXXXXXXXXXXXXXXXXX.........................
                    XXXXXXXXXXXXXXXXXXXXXXXXXXXX
                    XXXXXXXXXXXXXXXXXXXXXXXXXXXX
           HOUSTON  XXXXXXXXXXXXXXXXXXXXXXXXXXXX.........................
                    X
                    X
            DENVER  XXXXXXXXXXXXXXXXXXXXXXXXXXXX.........................
                    XXXXXXXXXXXXXXXXXXXXXXXXXXXX
                    XXXXXXXXXXXXXXXXXXXXXXXXXXXX
       LOS ANGELES  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX..........
                    XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
                    XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
     SAN FRANCISCO  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX..........
                    XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
                    XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
            SEATTLE  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
```

Density Linkage Cluster Analysis (密连法)　K = 3

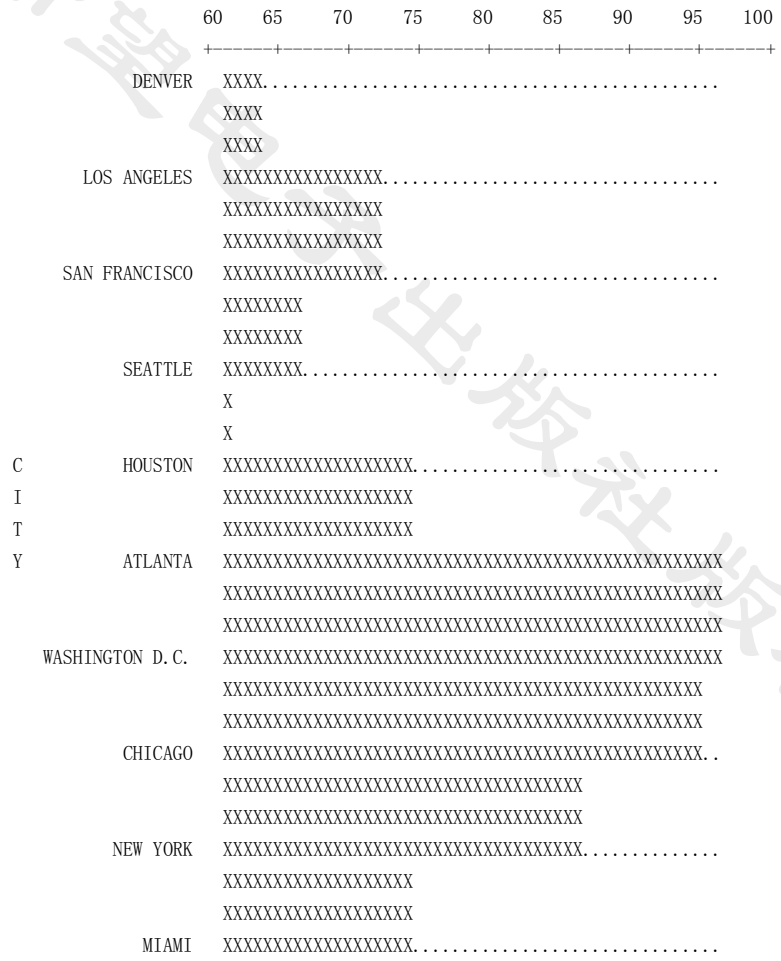
NCL	------Clusters Joined------		FREQ	Normalized Fusion Density	Normalized Maximum Density in Each Cluster Lesser	Greater	Tie
9	ATLANTA	WASHINGTON D.C.	2	96.1062	92.5043	100.0000	
8	CL9	CHICAGO	3	95.2632	90.9548	100.0000	

7	CL8	NEW YORK	4	86.4650	76.1571	100.0000
6	CL7	HOUSTON	5	74.0791	61.7747	100.0000 T
5	CL6	MIAMI	6	74.0791	58.8299	100.0000
4	LOS ANGELES	SAN FRANCISCO	2	71.9682	65.3430	80.0885
3	CL4	SEATTLE	3	66.3409	56.6215	80.0885
2	CL3	DENVER	4	63.5088	61.7747	80.0885
1	CL5	CL2	10	61.7747 *	80.0885	100.0000

* indicates fusion of two modal or multimodal clusters

2 modal clusters have been formed.

Cluster Fusion Density



Single Linkage Cluster Analysis (单连法)

Mean Distance Between Observations = 1417.133

Number of Clusters	Clusters Joined	Frequency of New Cluster	Normalized Minimum Distance	Tie

9	NEW YORK	WASHINGTON D.C.	2	0.144658
8	LOS ANGELES	SAN FRANCISCO	2	0.244861
7	ATLANTA	CL9	3	0.383168
6	CL7	CHICAGO	4	0.414216
5	CL6	MIAMI	5	0.426213
4	CL8	SEATTLE	3	0.478431
3	CL5	HOUSTON	6	0.494661
2	DENVER	CL4	4	0.586395
1	CL3	CL2	10	0.620266

Minimum Distance Between Clusters

```
                0.7     0.6      0.5      0.4      0.3      0.2      0.1       0
              +-------+-------+-------+-------+-------+-------+-------+-------+
C          ATLANTA       XXXXXXXXXXXXXXXXXXXX.............................
I                        XXXXXXXXXXXXXXXXXXXX
T                        XXXXXXXXXXXXXXXXXXXX
Y         NEW YORK       XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.........
                         XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
                         XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
   WASHINGTON D.C.       XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.........
                         XXXXXXXXXXXXXXXXXX
                         XXXXXXXXXXXXXXXXXX
             CHICAGO     XXXXXXXXXXXXXXXXXX.............................
                         XXXXXXXXXXXXXXXXXX
                         XXXXXXXXXXXXXXXXXX
               MIAMI     XXXXXXXXXXXXXXXXX..............................
                         XXXXXXXXXXXX
                         XXXXXXXXXXXX
             HOUSTON     XXXXXXXXXXX....................................
                         X
                         X
              DENVER     XXXX...........................................
                         XXXX
                         XXXX
         LOS ANGELES     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX...............
                         XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
                         XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
       SAN FRANCISCO     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX..............
                         XXXXXXXXXXXXXX
                         XXXXXXXXXXXXXX
             SEATTLE     XXXXXXXXXXXXXX.................................
```

Two-Stage Density Linkage Clustering （双连法） K = 3

NCL	Clusters Joined	FREQ	Normalized Fusion Density	Normalized Maximum Density in Each Cluster: Lesser	Normalized Maximum Density in Each Cluster: Greater	Tie

9	ATLANTA	WASHINGTON D.C.	2	96.1062	92.5043	100.0000
8	CL9	CHICAGO	3	95.2632	90.9548	100.0000
7	CL8	NEW YORK	4	86.4650	76.1571	100.0000
6	CL7	HOUSTON	5	74.0791	61.7747	100.0000 T
5	CL6	MIAMI	6	74.0791	58.8299	100.0000
4	LOS ANGELES	SAN FRANCISCO	2	71.9682	65.3430	80.0885
3	CL4	SEATTLE	3	66.3409	56.6215	80.0885
2	CL3	DENVER	4	63.5088	61.7747	80.0885

2 modal clusters have been formed.

NCL	Clusters Joined		FREQ	Normalized Fusion Density	Normalized Maximum Density in Each Cluster Lesser	Greater	T i e
1	CL5	CL2	10	61.7747	80.0885	100.0000	

Cluster Fusion Density

```
                  0         20        40        60        80        100
                  +---------+---------+---------+---------+---------+
         DENVER   XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.................
                  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
                  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
     LOS ANGELES  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.............
                  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
                  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
   SAN FRANCISCO  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.............
                  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
                  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
         SEATTLE  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.................
                  X
                  X
C        HOUSTON  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.............
I                 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
T                 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
Y        ATLANTA  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
                  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
                  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
  WASHINGTON D.C. XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
                  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
                  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
         CHICAGO  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.
                  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
                  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
        NEW YORK  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....
                  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
                  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
```

```
MIAMI XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....

Ward's Minimum Variance Cluster Analysis (华滋法)

Root-Mean-Square Distance Between Observations  = 1580.242
```

NCL	Clusters Joined		FREQ	SPRSQ	RSQ	PSF	PST2	Tie
9	NEW YORK	WASHINGTON D.C.	2	0.001870	0.9981	66.7	.	
8	LOS ANGELES	SAN FRANCISCO	2	0.005358	0.9928	39.2	.	
7	ATLANTA	CHICAGO	2	0.015332	0.9774	21.7	.	
6	CL7	CL9	4	0.029646	0.9478	14.5	3.4	
5	DENVER	HOUSTON	2	0.034379	0.9134	13.2	.	
4	CL8	SEATTLE	3	0.039131	0.8743	13.9	7.3	
3	CL6	MIAMI	5	0.058629	0.8157	15.5	3.8	
2	CL3	CL5	7	0.148757	0.6669	16.0	5.3	
1	CL2	CL4	10	0.666901	0.0000	.	16.0	

```
                                Semi-Partial R-Squared

                    0.7     0.6      0.5      0.4      0.3      0.2      0.1       0
                    +-------+-------+-------+-------+-------+-------+-------+
C           ATLANTA XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.
I                   XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
T                   XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
Y           CHICAGO XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.
                    XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
                    XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
          NEW YORK  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
                    XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
                    XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
    WASHINGTON D.C. XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
                    XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
                    XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
             MIAMI  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....
                    XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
                    XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
            DENVER  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX...
                    XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
                    XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
           HOUSTON  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX...
                    X
                    X
       LOS ANGELES  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
                    XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
                    XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
     SAN FRANCISCO  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
                    XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
```

```
            XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
SEATTLE     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX...
```

43.7　注 意 事 项

■　**遗漏数据的处理**

CLUSTER 程序对遗漏数据的处理视其输入数据的性质而定。若输入数据是坐标值，则任何有遗漏坐标的观察体将会完全从分析中剔除。若输入数据是欧氏距离，则遗漏值会中止分析的进行。

第 44 章　相斥式集群分析：统计程序 PROC FASTCLUS

44.1　PROC FASTCLUS 程序概述

相斥式集群法又称不重叠式集群法。顾名思义，这种集群法所产生的集群全是互相排斥的；也就是说每一个数据点只隶属于一个集群。因此通常我们不能用相斥式集群法的输出结果来画树形图。

相斥式集群法只适合用来分析大型的数据，亦即输入资料文件包括一百到十万个数据点。若输入资料文件内所含的数据点少于一百个，则相斥式集群法的分析会受到数据点先后排列顺序的影响，而产生不可靠的结果。

一般而言，读者必须事先决定集群的数目或集群的最小半径 (即集群内的点与重心间的距离)，而 PROC FASTCLUS 只要经过两三次的检验便可以找出集群的结构。

有关相斥式集群法的统计理论，请参阅 Hartigan (1975) 与 MacQueen (1967) 的著作。

44.2　如何撰写 PROC FASTCLUS 程序

PROC FASTCLUS 含六道指令，它们的格式如下：

```
PROC FASTCLUS   选项串;
     VAR        变量名称串;
     ID         变量名称;
     FREQ       变量名称;
     WEIGHT     变量名称;
     BY         变量名称串;
```

下面分别就这六个指令详加说明：

指令 #1　PROC FASTCLUS MAXCLUSTERS= 正整数或 RADIUS= 正实数据选项串;

PROC FASTCLUS 的选项大致可分为四类：第一类选项与输入 / 输出资料文件有关，第二类选项用来控制集群中心点的初选，第三类选项用来控制集群中心点的最后决定，第四类则包含其它的选项。下面起分别讨论各类选项：

第一类选项　下列五个选项与输入 / 输出资料文件有关：

(1) DATA=输入资料文件名称

指明对某一个输入资料文件执行相斥式集群分析。FASTCLUS 程序只接受坐标数据 (或原始数据文件)。若省略此选项，则 SAS 会自动找出在此程序之前最后形成的 SAS 资料文件，对它执行相斥式集群分析。

(2) SEED=SAS 的资料文件名称

指明从某一个输入资料文件里导出集群的中心点 (Seed)。此输入资料文件必须包含与 DATA=输入资料文件中相同的变量。若省略此选项，则集群的中心点将由 DATA=输入资料文件中导出。

(3) OUT=输出资料文件名称

为输出资料文件命名。这个输出资料文件包括所有输入资料文件的数据以及两个新变量：CLUSTER 及 DISTANCE。详细内容请见本章第 44.4 节有关 OUT=输出资料文件的进一步说明。

(4) MEAN=输出资料文件名称

为输出资料文件命名。这个输出资料文件包括集群的平均数和其它统计值。详细内容请见本章第 44.4 节有关 MEAN=输出资料文件的进一步说明。

(5) CLUSTER=变量名称

如果读者选用上述的 OUT= 或 MEAN= 选项，则输出资料文件中将会包含一个变量，代表集群的成员。选项 CLUSTER= 即为此一变量命名。若省略此选项，则 SAS 会自动设定 CLUSTER=CLUSTER。

第二类选项　下列四个选项控制集群中心点的初选 [选项 (1) 或 (2) 是必须的]：

(1) MAXCLUSTERS=正整数 (或 MAXC=正整数)

界定集群数目的最大值。若省略此选项，则 SAS 自动假设集群数的最大值是 100。

(2) RADIUS=正实数

设定一个距离准则以供 SAS 选择新的中心点。当一个观察体距原中心点的最近距离超过 RADIUS= 所定的标准时，这个观察体便有机会成为一个新的中心点。RADIUS 的内设值是 0。 若读者选用 REPLACE=RANDOM，则 RADIUS 选项不产生任何作用。

(3) REPLACE=FULL 或
REPLACE=PART 或
REPLACE=NONE 或
REPLACE=RANDOM

此选项决定集群中心点的取代方式。若读者定 REPLACE=FULL ，则集群中心点的取代由 MAXC= 及 RADIUS= 两选项来决定。若读者选 REPLACE=PART，则当数据点和任何一个中心点的距离必须大于任何两个既存中心点的距离时，初选的中心点机会被取代。若读者选 REPLACE=NONE，则初选的中心点维持不变，不被取代。若读者选 REPLACE=RANDOM ，则 SAS 会选择一组随机随机数点为集群的初选中心点。

(4) RANDOM=正整数

此选项与上述 REPLACE=RANDOM 联用，旨在起动随机随机数点的取样过程。

第三类选项　下列五个选项控制集群中心点的最后决定：

(1) DRIFT

经过上述中心点初选的过程后，每一数据点皆隶属于某一个集群。如果读者用 DRIFT (意即 “飘流”) 选项，则当 SAS 逐一处理某一集群内的数据点时，其

初选的中心点将随数据点的平均值改变而有变化 (故称“飘流”)。最后，集群中心点的值将是集群内所有数据点的平均值。

(2) STRICT=正整数 (或 STRICT)

这个选项的整数值设定一个距离准则。若某一个数据点与它最邻近集群之中心点的距离超过 STRICT= 所定的标准，则此数据点不能被归纳到任何一个既存的集群内。这一类的数据点将全被归纳在另一个集群内。此集群的名称是一个负整数，这个负整数的绝对值代表这些数据点最邻近的集群。若只用 STRICT 而不附加数值，则选项 RADIUS=不可省略，因为此时选项 RADIUS= 的整数值就成为距离的准则。

(3) MAXITER=正整数

此选项决定重复计算集群中心点的次数，其内设值是 1。在每一次重复计算的过程中，每一个数据点会被归纳到最邻近中心点的集群内。集群中心点的值也会被重新计算过，它永远是集群内各数据的平均值。

(4) CONVERGE=正有理数 (或 CONV= 正有理数)

此选项与上述的 MAXITER= 选项并用，其内设值是 .02。这个选项的目的在于决定何时终止重复计算的过程。当中心点的移动距离小于或等于选项 CONV=之数值与两个初选中心点间最小距离的乘积时，SAS 会终止重复计算。

(5) DELETE=正整数

若某一集群的成员数小于选项 DELETE= 所定的数值，则此集群的中心点被 SAS 删除，中心点的删除是在执行 DRIFT 选项及每一次循环计算之后。在最后一次集群形成后，中心点不再被删除。若读者省略此选项，则 SAS 保留所有的集群中心点。若读者设定 MAXITER=0，并且省略 DRIFT 选项，则选项 DELETE=无效。

第四类选项　除了上述三类选项外，还有八个选项可以在此一并讨论：

(1) LIST

要求列出原数据点 (或数据点 ID 的值)、数据点与其最终集群中心点的距离及数据点所属集群的名称。

(2) DISTANCE

要求 SAS 印出各集群中心点间的距离。

(3) SHORT

要求 SAS 不要印出各初选中心点的值、集群的平均数，或集群的变异数。

(4) SUMMARY

作用与选项 SHORT 相似，指示 SAS 抑止更多的统计值被印出。

(5) NOPRINT

此选项抑止打印所有的原始资料、统计结果，或集群分析的过程。

(6) IMPUTE

如果一个观察体在某一变量上有遗漏值 (Missing Value)，则 IMPUTE 选项指示 SAS 用该观察体所属之集群的中心点来取代此遗漏值。

(7) NOMISS

剔除任何有遗漏值的数据点。如果选项 IMPUTE 与 NOMISS 同时存在，则 NOMISS 选项无效。

(8) VARDEF=N 或

VARDEF=DF 或

VARDEF=WGT 或

VARDEF=WDF

此选项决定变异数或共变异数计算时所须分母的值。若 VARDEF=N，则分母是数据组内观察体的数目。若 VARDEF=DF，则分母是自由度，即 N-C，此处 C 代表集群的数目。若 VARDEF=WGT，则分母是 WEIGHT 变量的总和，此处，WEIGHT 代表观察体的比重。若 VARDEF=WDF，则分母是 WGT-C，此处，WGT 是 WEIGHT 变量的总和，而 C 是集群的数目。

指令 #2　VAR 变量名称串;

列出需要进行相斥式集群分析的数值变量。

指令 #3　ID 变量名称;

ID 变量可以是一个数值或是文字变量，主要用来鉴别报表上的输入数据。此指令与 PROCFASTCLUS 指令中的 LIST 选项合用。

指令 #4　FREQ 变量名称;

FREQ 变量的值是非零的正整数，代表每一个观察体在资料文件内重复出现的次数，这道指令可用来节省读者登录资料的时间。

指令 #5　WEIGHT 变量名称;

WEIGHT 变量是一个加权变量，功能与 FREQ 变量类似，其值是正有理数。这道指令决定观察体不同的比重，会改变集群平均数的计算。

指令 #6　BY 变量名称串;

FASTCLUS 程序依据此指令所列举的变量将资料文件分成几个小的资料文件，然后对每一个小资料文件分别执行分析。当读者选用此指令时，资料文件内的数据必须先按照 BY 变量串的值做由小到大的重新排列，这个步骤可藉 PROC SORT 达成。

下面我们要讨论 BY 指令与选项 SEED= 的关系。如果 BY 指令与选项 SEED= 并用，而且 SEED= 的资料文件内不含任何 BY 指令所列的变量，则整个 SEED= 的资料文件将用来当作 BY 指令的变量里每一集群的初选中心点。如果 SEED= 的资料文件含一部分 BY 指令的变量串，或是 BY 指令的某些变量与 SEED= 的资料文件不配合，则此差异会导致 FASTCLUS 程序印出警告讯息，并停止分析。最后，如果所有 BY 指令的变量都包含在 SEED= 资料文件内，则 SAS 就利用这些变量的观察体来设定集群的初选中心点；当这种现象发生时，SEED= 的资料文件和 DATA= 的资料文件中 BY 变量之集群的排列必须依同样的次序 (如：由小到大)，才可以执行相斥式的集群分析。

44.3 范　　例

例一：费氏紫罗兰的相斥式集群分析

这一组紫罗兰的数据是费氏于 1936 年在英国收集的 (Fisher，1936)。它常被用来当做范例以解释集群分析。这一组资料从三种不同属性的紫罗兰 (SETOSA=1，VERSICOLOR=2，和 VIRGINICA=3) 搜集而来。每种紫罗兰取五十个样本，然后测量它们花萼与花瓣的长与宽 (测量单位是厘米)。每一个样本包括五个数据，依序是花萼长、宽，花瓣长、宽以及属性号码。本例共进行两次相斥式集群分析：第一次取两个集群 (MAXC=2)，第二次取三个集群 (MAXC=3)，读者可比较两次分析的结果。另外，我们也利用 PROC FREQ，将紫罗兰依原属性以及集群分析的结果作列联表。

程　序

```
DATA IRIS;
    TITLE 'FISHER (1936) IRIS DATA';
    INPUT SEPALLEN SEPALWID PETALLEN PETALWID SPEC_NO @@;
    IF SPEC_NO=1 THEN SPECIES='SETOSA';
    ELSE IF SPEC_NO=2 THEN SPECIES='VERSICOLOR';
    ELSE SPECIES='VIRGINICA';
    LABEL SEPALLEN='SEPAL LENGTH IN MM.'
          SEPALWID='SEPAL WIDTH  IN MM.'
          PETALLEN='PETAL LENGTH IN MM.'
          PETALWID='PETAL WIDTH  IN MM.';
    CARDS;
       (原数据见第 38 章之例一)
;
PROC FASTCLUS DATA=IRIS MAXC=2 MAXITER=10 OUT=CLUS;
    VAR SEPALLEN SEPALWID PETALLEN PETALWID;
PROC FREQ;
    TABLES CLUSTER*SPECIES;
PROC FASTCLUS DATA=IRIS MAXC=3 MAXITER=10 OUT=CLUS;
    VAR SEPALLEN SEPALWID PETALLEN PETALWID;
PROC FREQ;
    TABLES CLUSTER*SPECIES;
RUN;
```

结　果

虽然三个集群 (MAXC=3) 的结果比两个集群 (MAXC=2) 更好，因为 R^2 值较高，

但若与原属性的种类相比，则两种解法都无法将第二与第三属性的紫罗兰分辨出来。只有第一种属性的紫罗兰易于归纳在第一个集群内。

因此两次的集群分析对第二与第三种属性的紫罗兰之分类无太大的助益。

报表 44.1　费氏紫罗兰的相斥式集群分析

FISHER (1936) IRIS DATA

FASTCLUS Procedure

Replace=FULL　Radius=0　Maxclusters=2　Maxiter=10　Converge=0.02

Initial Seeds

Cluster	SEPALLEN	SEPALWID	PETALLEN	PETALWID
1	43.0000	30.0000	11.0000	1.0000
2	77.0000	26.0000	69.0000	23.0000

Relative Change in Cluster Seeds

Iteration	Criterion	1	2
1	11.0638	0.1904	0.3163
2	5.3780	0.0596	0.0264
3	5.0718	0.0174	0.00766

Minimum Distance Between Initial Seeds = 70.85196　　Convergence criterion is satisfied.

Criterion Based on Final Seeds = 5.0417

Cluster Summary

Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Nearest Cluster	Centroid Distance
1	53	3.7050	21.1621	2	39.2879
2	97	5.6779	24.6430	1	39.2879

Statistics for Variables

Variable	Total STD	Within STD	R-Squared	RSQ/(1-RSQ)
SEPALLEN	8.280661	5.493128	0.562896	1.287784
SEPALWID	4.358663	3.703931	0.282710	0.394137
PETALLEN	17.652982	6.803310	0.852470	5.778291
PETALWID	7.622377	3.572004	0.781868	3.584390
OVER-ALL	10.692237	5.072913	0.776410	3.472463

Pseudo F Statistic =　513.92

Approximate Expected Over-All R-Squared =　0.51539

Cubic Clustering Criterion = 14.806

WARNING: The two above values are invalid for correlated variables.

Cluster Means

Cluster	SEPALLEN	SEPALWID	PETALLEN	PETALWID
1	50.0566	33.6981	15.6038	2.9057
2	63.0103	28.8660	49.5876	16.9588

Cluster Standard Deviations

Cluster	SEPALLEN	SEPALWID	PETALLEN	PETALWID
1	3.42735	4.39661	4.40428	2.10553
2	6.33689	3.26799	7.80058	4.15561

TABLE OF CLUSTER BY SPECIES

CLUSTER SPECIES

Frequency / Percent / Row Pct / Col Pct	SETOSA	VERSIC	VIRGIN	Total
1	50	3	0	53
	33.33	2.00	0.00	35.33
	94.34	5.66	0.00	
	100.00	6.00	0.00	
2	0	47	50	97
	0.00	31.33	33.33	64.67
	0.00	48.45	51.55	
	0.00	94.00	100.00	
Total	50	50	50	150
	33.33	33.33	33.33	100.00

FASTCLUS Procedure

Replace=FULL Radius=0 Maxclusters=3 Maxiter=10 Converge=0.02

Initial Seeds

Cluster	SEPALLEN	SEPALWID	PETALLEN	PETALWID
1	58.0000	40.0000	12.0000	2.0000
2	77.0000	38.0000	67.0000	22.0000
3	49.0000	25.0000	45.0000	17.0000

Relative Change in Cluster Seeds

Iteration	Criterion	1	2	3
1	6.7591	0.2652	0.3205	0.2985
2	3.7097	0	0.0459	0.0317
3	3.6427	0	0.0182	0.0124

Convergence criterion is satisfied.

Minimum Distance Between Initial Seeds = 38.23611

Criterion Based on Final Seeds = 3.6289

Cluster Summary

Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Nearest Cluster	Centroid Distance
1	50	2.7803	12.4803	3	33.5693
2	38	4.0168	14.9736	3	17.9718
3	62	4.0398	16.9272	2	17.9718

Statistics for Variables

Variable	Total STD	Within STD	R-Squared	RSQ/(1-RSQ)
SEPALLEN	8.280661	4.394883	0.722096	2.598359
SEPALWID	4.358663	3.248163	0.452102	0.825156
PETALLEN	17.652982	4.214314	0.943773	16.784895
PETALWID	7.622377	2.452436	0.897872	8.791618
OVER-ALL	10.692237	3.661982	0.884275	7.641194

Pseudo F Statistic = 561.63

Approximate Expected Over-All R-Squared = 0.62728

Cubic Clustering Criterion = 25.021

WARNING: The two above values are invalid for correlated variables.

Cluster Means

Cluster	SEPALLEN	SEPALWID	PETALLEN	PETALWID
1	50.0600	34.2800	14.6200	2.4600
2	68.5000	30.7368	57.4211	20.7105
3	59.0161	27.4839	43.9355	14.3387

Cluster Standard Deviations

Cluster	SEPALLEN	SEPALWID	PETALLEN	PETALWID
1	3.52490	3.79064	1.73664	1.05386
2	4.94155	2.90092	4.88590	2.79872
3	4.66410	2.96284	5.08895	2.97500

TABLE OF CLUSTER BY SPECIES

CLUSTER	SPECIES			
Frequency Percent Row Pct Col Pct	SETOSA	VERSIC	VIRGIN	Total
1	50 33.33 100.00 100.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	50 33.33
2	0 0.00 0.00 0.00	2 1.33 5.26 4.00	36 24.00 94.74 72.00	38 25.33
3	0 0.00 0.00 0.00	48 32.00 77.42 96.00	14 9.33 22.58 28.00	62 41.33
Total	50 33.33	50 33.33	50 33.33	150 100.00

例二：电脑自生的随机数之分析

本例所用的数据是由电脑自己衍生出来的。本例的重点在于示范如何利用 PROC FASTCLUS 指令中的 STRICT= 选项找出数据组中的劣质数据 (Outliers)，并且控制集群分析的结果使其不受这些劣质数据的影响。

程　序

```
TITLE 'USING FASTCLUS TO ANALYZE DATA OUTLIERS';
DATA X;DROP N;
       DO N=1 TO 100;
       X=RANNOR(12345)+2; Y=RANNOR(12345); OUTPUT;
       END;
    DO N= 1 TO 100;
       X=RANNOR(12345)-2; Y=RANNOR(12345); OUTPUT;
       END;
    DO N=1 TO 10;
       X=10*RANNOR(12345); Y=10*RANNOR(12345); OUTPUT;
       END;
RUN;
TITLE2 'PRELIMINARY FASTCLUS ANALYSIS WITH 20 CLUSTERS';
PROC FASTCLUS DATA=X MEAN=MEAN1 MAXC=20 MAXITER=0 SUMMARY;
    VAR X Y;
PROC PLOT DATA=MEAN1 VPERCENT=300;
    PLOT _GAP_*_FREQ_='G' _RADIUS_*_FREQ_='R'/OVERLAY
    VAXIS=0 TO 10 BY 2;
```

```
RUN;

DATA SEED;SET MEAN1;
    IF _FREQ_>5;
RUN;
TITLE2 'FASTCLUS ANALYSIS USING STRICT= TO OMIT OUTLIERS';
PROC FASTCLUS DATA=X SEED=SEED MAXC=2 STRICT=3.0 OUT=OUT MEAN=MEAN2;
    VAR X Y;
PROC PLOT DATA=OUT VPERCENT=300;
    PLOT Y*X=CLUSTER/VAXIS=-12.5 TO 12.5 BY 2.5
                HAXIS=-10 TO 17.5 BY 2.5;
RUN;
```

结　果

报表 44.2　电脑自生的随机数之分析

USING FASTCLUS TO ANALYZE DATA OUTLIERS

TRELIMINARY FASTCLUS ANALYSIS WITH 20 CLUSTERS

Replace=FULL Radius=0 Maxclusters=20 Maxiter=0

Cluster Summary

Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Nearest Cluster	Centroid Distance
1	8	0.4753	1.1924	19	1.7205
2	1	.	0	6	6.2847
3	44	0.6252	1.6774	5	1.4386
4	1	.	0	20	5.2130
5	38	0.5603	1.4528	3	1.4386
6	2	0.0542	0.1085	2	6.2847
7	1	.	0	14	2.5094
8	2	0.6480	1.2961	1	1.8450
9	1	.	0	7	9.4534
10	1	.	0	18	4.2514
11	1	.	0	16	4.7582
12	20	0.5911	1.6291	16	1.5601
13	5	0.6682	1.4244	3	1.9553
14	1	.	0	7	2.5094
15	5	0.4074	1.2678	3	1.7609
16	22	0.4168	1.5139	19	1.4936
17	8	0.4031	1.4794	5	1.5564
18	1	.	0	10	4.2514
19	45	0.6475	1.6285	16	1.4936
20	3	0.5719	1.3642	15	1.8999

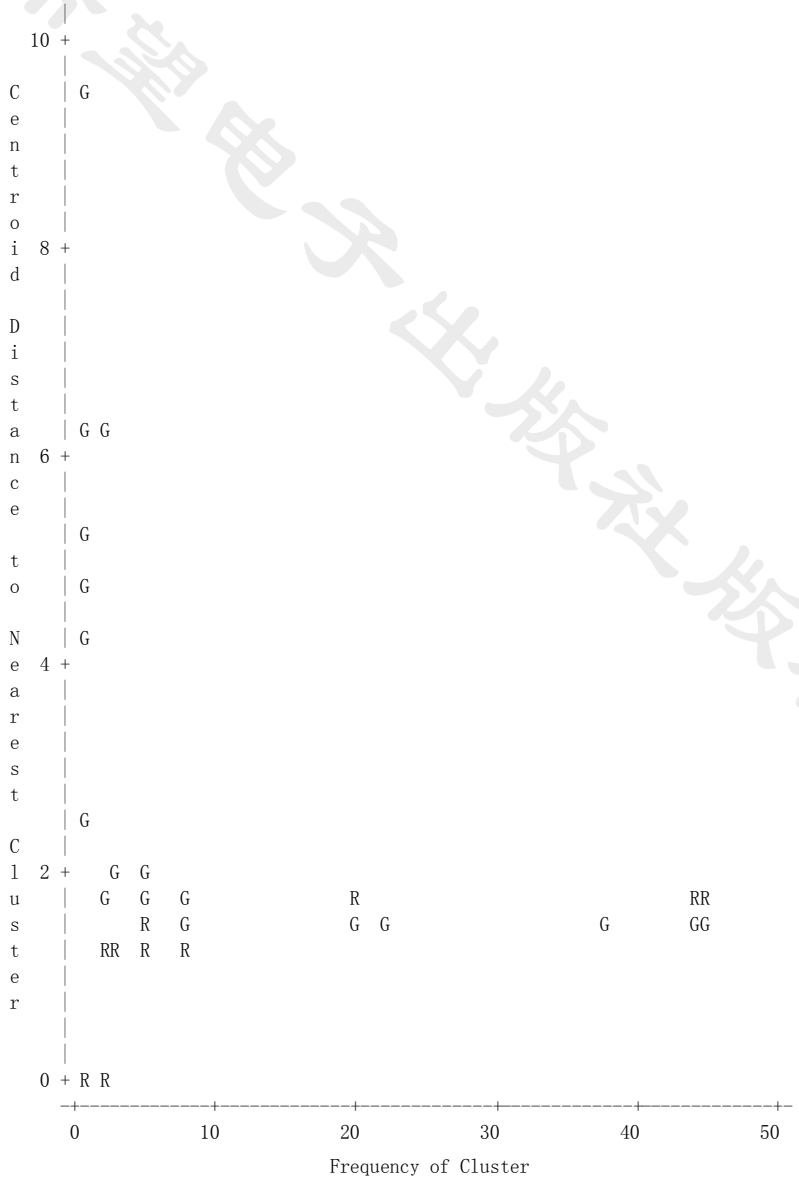
Statistics for Variables

Pseudo F Statistic =	207.58
Observed Over-All R-Squared =	0.95404
Approximate Expected Over-All R-Squared =	0.96103
Cubic Clustering Criterion =	-2.503

WARNING: The two above values are invalid for correlated variables.

Plot of _GAP_*_FREQ_. Symbol used is 'G'.

Plot of _RADIUS_*_FREQ_. Symbol used is 'R'.



NOTE: 12 obs hidden.

USING FASTCLUS TO ANALYZE DATA OUTLIERS

FASTCLUS ANALYSIS USING STRICT= TO OMIT OUTLIERS

Replace=FULL Radius=0 Strict=3 Maxclusters=2 Maxiter=1

Initial Seeds

Cluster	X	Y
1	2.79417	-0.06597
2	-2.02730	-2.05121

Criterion Based on Final Seeds = 0.95155

Cluster Summary

Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Nearest Cluster	Centroid Distance
1	99	0.9501	2.9589	2	3.7666
2	99	0.9290	2.8011	1	3.7666

12 Observation(s) were not assigned to a cluster
because the minimum distance to a cluster seed exceeded the STRICT= value.

Statistics for Variables

Variable	Total STD	Within STD	R-Squared	RSQ/(1-RSQ)
X	2.068537	0.870977	0.823609	4.669219
Y	1.021128	1.003520	0.039093	0.040683
OVER-ALL	1.631188	0.939589	0.669891	2.029303

Pseudo F Statistic = 397.74

Approximate Expected Over-All R-Squared = 0.60615

Cubic Clustering Criterion = 3.197

WARNING: The two above values are invalid for correlated variables.

Cluster Means

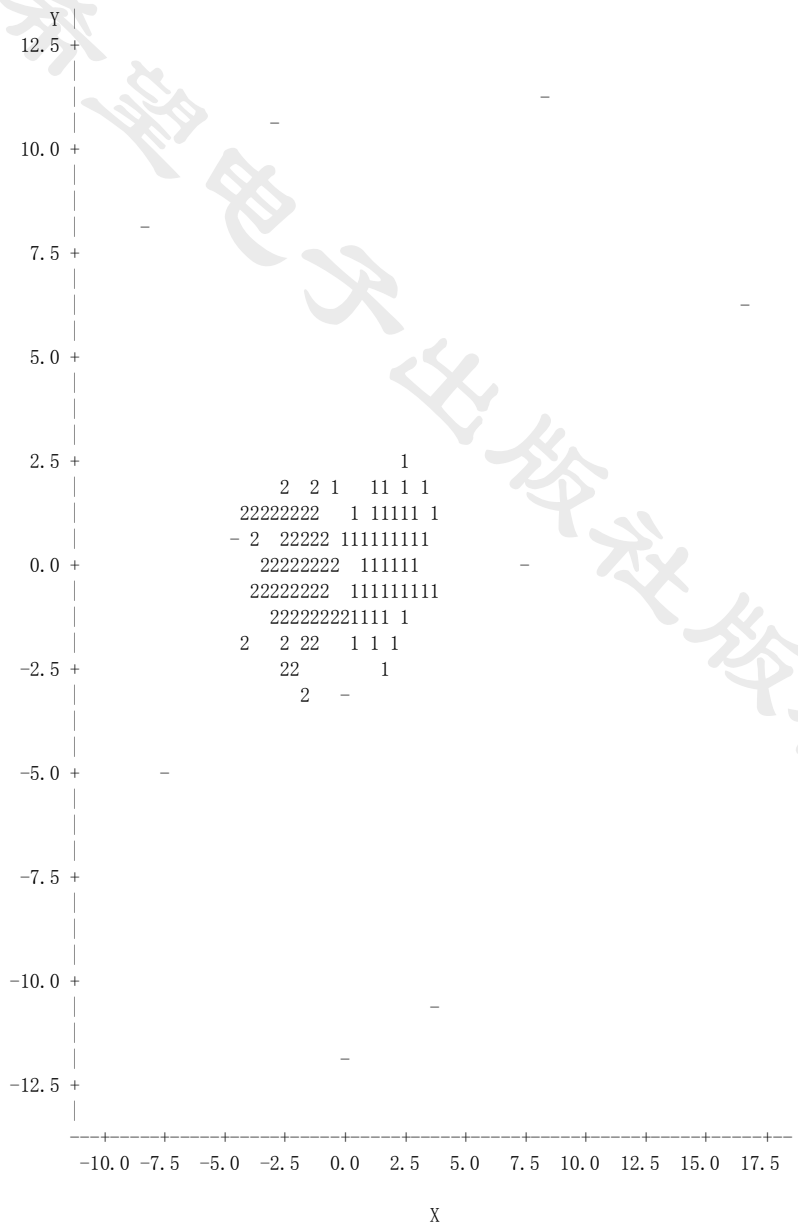
Cluster	X	Y
1	1.82511	0.14121

2	-1.91991	-0.26156

Cluster Standard Deviations

Cluster	X	Y
1	0.88955	1.00697
2	0.85200	1.00006

Plot of Y*X. Symbol is value of CLUSTER.



NOTE: 106 obs hidden. 1 obs were out of range.

44.4 注 意 事 项

■ 遗漏数据的处理

一个观察体在各变量上的值可能呈现下列三种情形：

(1) 完整，无任何遗漏数据，

(2) 部分完整，有部分遗漏数据，

(3) 不含任何数据，即全部是遗漏数据。

例如：一个资料文件可能含三个学生：甲、乙、丙 (三个观察体)，及两个变量 (期中考及期末考)。甲生两次考试都有参加，所以我们说甲生这个观察体所含的数据是完整的。乙生考了期中考，但期末考因故缺考，则我们说乙生这个观察体的数据是部分完整的(只有期中考成绩)，但亦有部分遗漏数据 (欠缺期末考成绩)。丙生因故两次考试皆缺席，则我们说丙生这个观察体不含任何数据，其变量值全部为遗漏数据。这个资料文件可以表示如下：

	变量数据串	
观　察　体	期中考	期末考
甲生 (无遗漏数据)	90	95
乙生 (含部分遗漏数据)	80	缺
丙生 (全部为遗漏数据)	缺	缺

SAS 在处理遗漏数据时，会自动删除那些完全没有数据的观察体 (如丙生)。若读者在 PROCFASTCLUS 指令中选用 NOMISS 选项，则那些有部分遗漏数据的观察体 (如乙生) 也会被剔除。另外，若观察体的数据不完整，则它不能成为任何集群的中心点。所以在这个例子，只有甲生有资格成为集群的中心点。

■ 选项 OUT＝输出资料文件的进一步说明

在前面 (指令 #1 的部分)，我们曾简单地介绍了选项 OUT=。现在我们要做进一步的说明。当读者在 PROC FASTCLUS 指令中选用选项 OUT=，则所得到的输出资料文件将包括下列变量：

(1) 原输入资料文件所含的变量串。

(2) 一个新的变量 DISTANCE，代表各观察体与其集群中心点的距离。

(3) 一个新的变量用来指明各观察体所属的集群。读者可用 PROC FASTCLUS 指令中 CLUSTER=选项来为此变量命名。此变量的值是由 1 到选项 MAXCLUSTER=所定的正整数。

如果读者又选用 IMPUTE 选项，则 OUT=的输出资料文件又多含一个新变量：

(4) _IMPUTE_，记载资料文件内每一个观察体到底遗漏了多少个变量的值。

■ 选项 MEAN＝输出资料文件的进一步说明

在前面指令 #1 的部分，我们曾简单地介绍了选项 MEAN=，现在要做进一步的说

明。当读者在 PROC FASTCLUS 指令中选用选项 MEAN=，则 SAS 会从每一个集群中挑出一个观察体来代表该集群。这些观察体与下列的变量就是 MEAN= 输出资料文件的数据：

(1) 所有的 BY 变量串。

(2) 一个新的变量用来指明各观察体所属的集群。读者可用 PROC FASTCLUS 指令中 CLUSTER=选项来替此变量命名，否则 SAS 会自动为它命名为 CLUSTER。

(3) _FREQ_，这个变量的值告诉我们每一集群内包含了多少个成员。

(4) _WEIGHT_，读者必须在原程序内包括指令 WEIGHT。

(5) 一个新的变量叫 _RMSSTD_，其值代表集群内的平均变异数。

(6) 一个新的变量叫 _RADIUS_，其值代表集群内成员与其中心点之间的最大距离。

(7) 一个新的变量叫 _GAP_，其值代表某一集群的平均数与其最邻近的另一个集群平均数之间的距离。

(8) 一个新的变量叫 _NEAR_，代表最邻近那个集群的名字。

(9) _VAR_ 变量串，其值代表各集群在这些变量上的平均数。

第 45 章　变量的集群分析：统计程序 PROC VARCLUS

45.1　PROC VARCLUS 程序概述

VARCLUS 程序根据相关系数或变异数/共变异数矩阵对变量执行相斥式或阶层式的集群分析。此程序与前述 CLUSTER 程序 (见第 43 章) 及 FASTCLUS 程序 (见第 44 章) 最大的不同在于分析的对象：VARCLUS 分析变量，而 CLUSTER 及 FASTCLUS 程序则分析观察体。

输入资料文件

VARCLUS 程序的输入资料文件可以是一个相关系数矩阵，也可以是一个变异数／共变异数的矩阵。当输入资料文件是一个相关系数矩阵时，所有的变量都具同等的重要性。当输入资料文件是一个变异数／共变异数矩阵时，变量的重要性随其变异数的增大而提升。

功能

VARCLUS 程序的主要功能是将一组数值变量归类到不重叠的或重叠的集群内。若集群内的变量可组成一个线性组合，通常这个线性组合是变量间的主成份 (Principal Component) 或重心成份 (Centroid Component)。基于这个概念，VARCLUS 程序分类的目的是增加这些线性组合在每一集群内所能解释的变异数。另外，VARCLUS 程序也可用来简化资料文件内的变量，使其不致过于繁复。比方说，一个心理测验中有五十道试题 (即五十个变量)，这五十道试题可用 VARCLUS 程序归并成五大类，每一大类代表一个子测验 (Subtest)。简化后变量的数目由 50 减到 5。学生在这些子测验上的分数，从统计的眼光看，也就是五个集群的重心成份 (Centroid Component)。

45.2　VARCLUS 程序的分析步骤

在分析开始时，所有的变量均属于一个大集群，然后 VARCLUS 程序按照下列的步骤进行归并：

第一步

VARCLUS 程序找出这一个大集群的第一与第二主成份 (亦称主轴向量)。这两个主成份经过正交的坐标转换 (亦即因子分析中常用的 Quartimax 方法) 后，变量被指定归入一个与其相关系数平方较高的主成份。如此原有的集群分裂成二。

第二步

两个 (或两个以上) 之中的一个集群被选中，照第一步的方法再分裂为二。这个被选中的集群通常拥有最大的第二特征值 (Eigen Value)，或者是拥有最小的可被集群向量解释的变异数百分比。

第三步

第一步骤与第二步骤不停的交互进行，直到集群内变量间的第二特征值或集群内可被解释的变异数百分比达到读者所预设的标准为止。

上述的步骤分为两大过程：第一个过程称为近邻主成份分类 (NCS 或作 Nearest Component Sorting)。在这个过程里，SAS 计算集群的主成份 (Cluster Component)，然后每一个变量被分配到与其相关系数平方最高的主成份里。第二个过程称为搜索 (Search)。在这个过程里，SAS 试着将一个变量分配到不同的集群内，看看这个新的分配是否会增加集群内可解释的变异数。如果可被解释的变异数百分比会增加，则 SAS 重新计算变量原属的集群与新集群的主成份。接下来，另一个变量也同样经历这个过程，直到所有的变量都被如此的试验过，而且所属新旧集群的主成份稳定下来为止。若比较这两个过程，我们不难发现 NSC 的手续较省时，但也较易陷入局部最优解 (Local Minimum) 的陷阱。

45.3　如何撰写 PROC VARCLUS 程序

PROC VARCLUS 含七道指令，它们的格式如下：

```
PROC VARCLUS    选项串;
    VAR         变量名称串;
    SEED        变量名称串;
    PARTIAL     变量名称串;
    WEIGHT      变量名称;
    FREQ        变量名称;
    BY          变量名称串;
```

指令 #1　PROC VARCLUS 选项串;

PROC VARCLUS 的选项大致可分为五类：第一类选项是为各资料文件命名，第二类选项控制集群的个数，第三类选项可控制集群形成的方法，第四类选项控制有关统计值的打印，第五类选项控制输出资料的打印。现在，让我们分别讨论这五类选项：

第一类选项　读者可以用这一类选项来为各资料文件命名：

(1) DATA=输入资料文件名称

指明对某一个资料文件做集群分析。若资料文件内包含的变量是相关系数 (CORR)、共变异数 (COV)、或因子分数 (FACTOR)，则必须在选项 DATA= 后用 (TYPE=) 指明，如：DATA=INPUT (TYPE=CORR)。 若读者省略此选项，则 SAS 会自动找出在此程序之前最后形成的 SAS 资料文件，对它执行集群分析。

(2) OUTSTAT=输出资料文件名称

为输出资料文件命名。这个资料文件包括平均数、标准差、相关系数、集群的线性组合系数、以及集群的结构。

(3) OUTTREE=输出资料文件名称

这一个选项可将集群分析的结果存入一个适当的资料文件内以供绘制树形图之用

(制树形图这一个步骤是藉 PROC TREE 达成，详细内容请见第 46 章说明)。当你界定 OUTTREE=选项时，SAS 会同时界定 HIERARCHY 选项。

第二类选项　读者可用下列的选项来控制集群的数目：

(1) MINCLUSTERS=正整数 (或 MINC=正整数)

读者可用这个选项来指示 SAS 最少要有几个集群。若你同时选用 INITIAL=RANDOM (或=SEED) 选项，则选项 MINC= 的内设值是 2。否则 SAS 假设集群的形成从一个最大的集群开始 (此时所有的变量都属于这个大集群)，不断地分裂，直到分裂的结果达到 PROPORTION= 或 MAXEIGEN=选项所定的标准为止。这样的分析过程类似阶层式集群法。

(2) MAXCLUSTERS=正整数 (或 MAXC=正整数)

读者可用此选项告诉 SAS 集群个数的上限。内设值是资料文件内被分析的变量总个数。若此正整数大于 2，则分析的结果类似阶层式集群法。

(3) PROPORTION=正有理数 (或 PERCENT=正有理数)

此选项规定集群主成份所能解释的变异数百分比。在界定这个选项时，请读者注意 PROPORTION=.75 与 PERCENT=75 的含义完全相同，所以只需这两者之一即可。一般而言，其内设值是 0。但若读者同时选用另一选项 CENTROID，则此选项的内设值是 .75。

(4) MAXEIGEN=正实数

此选项规定每一集群内第二特征值的最大可能值。一般而言，其内设值是 0。但当读者不把此选项与另外两个选项 PROPORTION= 及 MAXCLUSTER= 并用时，其内设值会随输入资料文件的成份而异。假如输入资料文件是相关系数矩阵，则此选项的内设值等于 1。如果输入资料文件是变异数 / 共变异数矩阵，则此内设值是变量间变异数的平均值。请注意：此选项不能与 CENTROID 选项并用。

第三类选项　读者可用下列选项控制集群形成的方法：

(1) COVARIANCE (或 COV)

指示 SAS 去分析一个变异数 / 共变异数的矩阵，而非相关系数矩阵。

(2) INITIAL= 方法名称

此选项规定 SAS 用一种方法展开集群分析。INITIAL 的方法有下列四种：RANDOM (随机法)、SEED (种籽法)、INPUT (输入法)，及 GROUP (群体法)，分别介绍如下：

INITIAL=RANDOM

此法将各变量随机分配到集群内。若读者选用此选项，然而没有同时选用 CENTROID 选项，则我们建议读者加用选项 MAXSEARCH=5。

INITIAL=SEED

以 SEED 指令中所界定的变量当做集群分析的初值，所以通常与 SEED 指令联用。然而，若读者省略 SEED 指令，则 SAS 自动取 VAR 指令中前 n 个变量当作初值，n 由 MINCLUSTERS=n (正整数) 决定。

INITIAL=INPUT

此法适用于五种特殊的输入资料文件，即：TYPE=CORR，UCORR，COV，UCOV，

或 FACTOR。

若 TYPE=FACTOR，则读者必须在输入资料文件内同时提供 _TYPE_='SCORE' 的因子分数据数。

INITIAL=GROUP

此法也仅用于 TYPE=CORR，COV 或 FACTOR 的输入资料文件。每一变量所属的集群由_TYPE_='GROUP' 决定。_TYPE_='GROUP' 的值由 1 到集群的总个数。

(3) CENTROID

此选项十分重要，许多其它的选项均受此选项影响。因为此选项导出集群的重心成份 (而非主成份)。重心成份通常是变量的未加权平平均数 (Unweighted Mean)。这种方法的分析结果并不能保证集群的重心成份与集群内变量的相关是最高的，此选项不能与 MAXEIGEN= 选项合用。

(4) MAXITER=正整数

此选项规定分析过程中循环 (Iteration) 的最高次数，内设值是 10。但当此选项与 CENTROID 选项合用时，内设值等于 1。

(5) MAXSEARCH=正整数

此选项规定搜索过程中循环的最高次数，内设值是 0。但当此选项与 CENTROID 选项合用时，内设值等于 10。

(6) HIERARCHY (或 HI)

此选项要求 SAS 进行阶层式集群法。若省略此选项，则 SAS 将执行相斥式集群法。

(7) MULTIPLEGROUP (或 MG)

此选项是下列所有选项串的简化：

MINC=1 MAXITER=0 MAXSEARCH=0 MAXEIGEN=0 PROPORTION=0 INITIAL=GROUP。

选用此选项时，输入资料文件必须是 TYPE=CORR，UCORR，COV，UCOV，SSCP，或 FACTOR 其中之一种，而且必须包含 _TYPE_='GROUP' 的变量。

(8) RANDOM=正整数

此选项与 REPLACE=RANDOM 联用，其目的在于启动 SAS 系统内的随机随机数表。若不选用此选项，则 VARCLUS 程序会根据电脑的时间来启动随机数表。

(9) NOINT

要求分析时不考虑截距 (或平均数)，由此选项所导出的 OUTSTAT= 输出资料文件会是一个 TYPE=UCORR 的资料文件。

(10) VARDEF=DF (或 N 或 WDF 或 WGT)

界定计算变异数或共变异数时所用的分母。其中，DF=自由度，也是此选项的内设值；N=观察体总数；WGT=加权后的观察体总数；WDF=WGT-1。

第四类选项　读者可以用下面介绍的选项来控制有关统计值的打印：

(1) SIMPLE (或 S)

这一个选项要求 SAS 印出每一变量的平均数与标准差。

(2) CORR (或 C)

这一个选项要求 SAS 印出变量间的相关系数矩阵。

第五类选项　读者可用第五类选项来控制输出资料的打印：

(1) SHORT

抑止集群结构、线性组合系数与集群间相关系数的印出。

(2) SUMMARY

除了总结论表外，其余资料一概不印出。

(3) NOPRINT

所有分析结果皆不印出。

(4) TRACE

印出每一次循环过程中各变量所属的集群。

指令 #2　VAR 变量名称串；

从输入资料文件中列举所有参与分析的数值变量名称串。

指令 #3　SEED 变量名称串；

此指令的名称可以用 SEED 或 SEEDS 来表示。其作用与 INITIAL=SEED 完全相同，旨在定出各集群的起始点。当读者选用指令 INITIAL= 的其它三种方法 (即 RANDOM，INPUT，或 GROUP)时，此指令无效。

指令 #4　PARTIAL 变量名称串；

如果读者有意使用净相关来进行集群分析，则必须在此指令中列举用来净化资料的变量名称串。

指令 #5　WEIGHT 变量名称；

一般而言，变量的重要性由其变异数的大小来决定。若读者想控制各变量的重要性，则可藉助 WEIGHT 指令；WEIGHT 变量的值即代表加权值。也可以用 WEIGHT 指令使每一被分析的变量有均等的重要性；此时，读者可界定加权值等于各变量变异数的倒数。

指令 #6　FREQ 变量名称；

与上述 WEIGHT 指令的功能类似，代表被分析变量的加权值。不过，FREQ 变量的值必须是正整数，WEIGHT 所界定的加权值则可以是正有理数。

指令 #7　BY 变量名称串；

此指令根据 BY 变量将输入资料文件分成几个小的资料文件，然后对每一个小资料文件分别进行集群分析。当读者使用此指令时，必须先将资料文件内的数据依 BY 变量串的值做由小到大的重新排列，这个步骤可藉 PROC SORT 来达成。

45.4 范　　例

例一：八个体型变量的集群分析

这个资料文件 (PHYS8)是一个相关系数矩阵，由 Harman 于 1976 年提供。资料文件内包含的变量是关于三百零五位学龄期女生的体型。八个体型变量分别是：身高(HEIGHT)、臂长(ARM_SPAN)、手腕到手肘的长度 (FOREARM)、小腿长度 (LOW_LEG)、体重 (WEIGHT)、Bitrochanteric 直径 (BIT_DIAM)、胸围 (GIRTH)、胸宽 (WIDTH) 等。这个资料文件就是这八个变量的相关系数矩阵。此范例中只执行一次 PROC VARCLUS 的分析，方法是阶层式的集群分析，其结果由 PROC TREE 以树形图表示。

程　序

```
DATA PHYS8(TYPE=CORR);
     TITLE 'EIGHT PHYSICAL VARIABLES MEASURED ON 305 SCHOOL GIRLS';
     TITLE2 'SEE PAGE 22 OF HARMAN:MODERN FACTOR ANALYSIS,3RD ED';
     LABEL HEIGHT='HEIGHT'
           ARM_SPAN='ARM SPAN'
           FOREARM='LENGTH OF FOREARM'
           LOW_LEG='LENGTH OF LOWER LEG'
           WEIGHT='WEIGHT'
           BIT_DIAM='BITROCHANTERIC DIAMETER'
           GIRTH='CHEST GIRTH'
           WIDTH='CHEST WIDTH';
           INPUT _NAME_ $ 1-8
               (HEIGHT ARM_SPAN FOREARM LOW_LEG WEIGHT BIT_DIAM GIRTH
                       WIDTH)
               (8.);_TYPE_='CORR';CARDS;
HEIGHT  1.0    .846    .805    .859    .473    .398    .301    .382
ARM_SPAN.846   1.0     .881    .826    .376    .326    .277    .415
FOREARM .805   .881    1.0     .801    .380    .319    .237    .345
LOW_LEG .859   .826    .801     1.0     .436   .329    .327    .365
WEIGHT  .473   .376    .380    .436     1.0    .762    .730    .629
BIT_DIAM .398  .326    .319    .329    .762    1.0     .583    .577
GIRTH    .301  .277    .237    .327    .730    .583   1.0     .539
WIDTH    382   .415    .345    .365    .629    .577    .539    1.0
;
PROC VARCLUS DATA=PHYS8 MAXC=8 SUMMARY OUTTREE=TREE;
PROC TREE PAGES=2; HEIGHT _PROPOR_;
RUN;
```

结　果

从树形图上，我们不难看出，这八个体型变量分属两个集群：第一个集群含体重、胸围、胸宽与 Bitrochanteric 直径变量，代表结实的体型。第二个集群含其余四个变量，代表瘦长的体型。

报表 45.1　八个体型变量的集群分析

SEE PAGE 22 OF HARMAN:MODERN FACTOR ANALYSIS,3RD ED

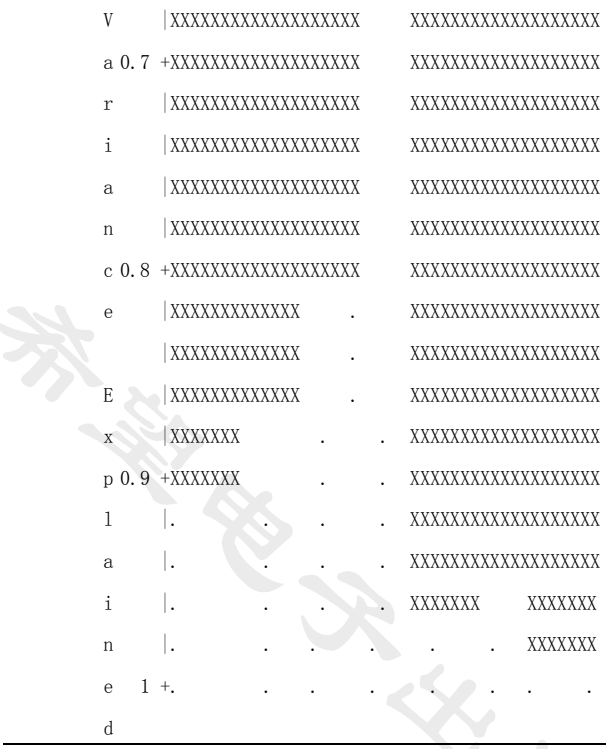
Oblique Principal Component Cluster Analysis

10000 Observations　　PROPORTION = 1

8 Variables　　MAXEIGEN = 0

Number of Clusters	Total Variation Explained by Clusters	Proportion of Variation Explained by Clusters	Minimum Proportion Explained by a Cluster	Maximum Second Eigenvalue in a Cluster	Minimum R-squared for a Variable	Maximum 1-R**2 Ratio for a Variable
1	4.672880	0.5841	0.5841	1.770983	0.3810	.
2	6.426502	0.8033	0.7293	0.476418	0.6329	0.4380
3	6.895347	0.8619	0.7954	0.418369	0.7421	0.3634
4	7.271218	0.9089	0.8773	0.238000	0.8652	0.2548
5	7.509218	0.9387	0.8773	0.236135	0.8652	0.1665
6	7.740000	0.9675	0.9295	0.141000	0.9295	0.2560
7	7.881000	0.9851	0.9405	0.119000	0.9405	0.2093
8	8.000000	1.0000	1.0000	0.000000	1.0000	0.0000

```
                    Name of Variable or Cluster
  P              B                             A
  r              I                          L  R  F
  o   W          T                    H     O  M  O
  p   E          _       G     W      E     W  _  R
  o   I          D       I     I      I     _  S  E
  r   G          I       R     D      G     L  P  A
  t   H          A       T     T      H     E  A  R
  i   T          M       H     H      T     G  N  M
  o  |
  n  |XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
 0.6 +XXXXXXXXXXXXXXXXXXX     XXXXXXXXXXXXXXXXXXX
  o  |XXXXXXXXXXXXXXXXXXX     XXXXXXXXXXXXXXXXXXX
  f  |XXXXXXXXXXXXXXXXXXX     XXXXXXXXXXXXXXXXXXX
     |XXXXXXXXXXXXXXXXXXX     XXXXXXXXXXXXXXXXXXX
```



45.5 注 意 事 项

■ 遗漏数据的处理

若观察体中所含的数据有部分 (或全部)遗漏，则 SAS 会自动剔除此观察体，不使其纳入分析。

■ 选项 OUTSTAT＝输出资料文件的进一步说明

此选项所导出的资料文件是以相关系数为主的 (即 TYPE=CORR)，其资料文件可用来作进一步的 VARCLUS 分析或成为线性组合分析 (即 PROC SCORE，见第 12 章) 的输入资料文件。此资料文件所包含的变量如下：

(1) BY 指令中的变量名称串。

(2) _NCL_，代表集群的个数。

(3) _NAME_，代表变量或集群的名字。

(4) 集群分析所包含的数值变量名称串。

(5) _TYPE_，代表统计值的种类，其值列举如下：

代号 (_TYPE_=)	定　义
MEAN	平均数
STD	标准差
USTD	未经平均数矫正过的标准差
N	观察体的个数
CORR	相关系数
MEMBERS	集群内成员的个数
VAREXP	集群内可解释的变异数
PROPOR	集群内可解释的变异数百分比
GROUP	每一变量所属的集群 (以数字代表)
RSQUARED	变量与集群主成份间相关系数的平方
SCORE	线性组合的标准化系数
USCORE	未经平均数矫正过的标准化系数，是 NOINT 选项的输出
STRUCTUR	集群的结构
CCORR	集群主成份间的相关系数
UCORR	未经平均数矫正过的相关系数矩阵，是选项 NOINT 界定的输出

■ 选项 OUTTREE＝输出资料文件的进一步说明

此选项所导出的资料文件包括：

(1) BY 指令中的变量名称串。

(2) _NAME_，　代表树形图上的节。当树节是一个变量时，该树节以此变量的名字来表示。当树节是一个集群时，则由集群的名字来表示 (如：CLUS2 或 CLUS7)。

(3) _PARENT_，　树形图上每一个树节上面一层的树节名称。

(4) _NCL_，　集群的总个数。

(5) _VAREXP_，　可被集群解释的变异数。

(6) _PROPOR_，　可被集群解释的变异数百分比。

(7) _MINPRO_，　可被集群解释之变异数的最低百分比。

(8) _MAXEIG_，　集群的第二大特征值值。

第 46 章　树形图：统计程序 PROC TREE

46.1　PROC TREE 程序概述

PROC TREE 统计程序的最主要功能在于绘制集群分析的树形图。树形图又可称为层次图 (Dendrogram) 或现象图 (Phenogram)。一般而言，PROC TREE 的输入资料文件是由 PROC CLUSTER 或 PROC VARCLUS 所提供的，而其输出的资料则包含树形图横切面的集群组合。

统计程序 PROC TREE 所绘制的树形图是依据强生氏 (Johnson) 在 1967 年文献中所建议的式样：

树的根在上，分枝朝下。也就是说：所有的变量 (或观察体) 原隶属于同一个集群 (这个大集群就是树的根)，然后这个最大集群分裂为二，再分裂为三，...。直到每一变量 (或观察体) 分属于单独的集群。

有关树形图的制作，可参考下列文献：

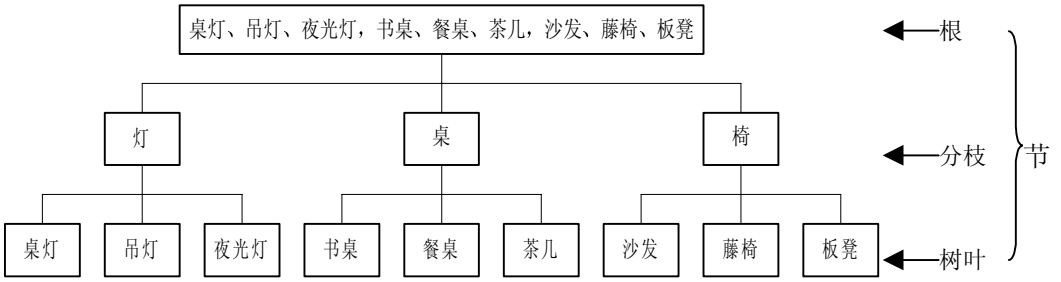
Duran 和 Odell (1974)，Hartigan (1975)，Everitt (1980)，或 Knuth (1973)。

46.2　有关树形图的专有名词

在我们做进一步的讨论之前，读者必须先认识六个有关树形图的专有名词：

根 (Root)	指树形图最顶端那个包含所有变量 (或观察体) 的最大之集群。
树叶 (Leaves)	指集群内的各变量或各观察体。
分枝 (Branch)	除根以外，任何含两个或两个以上变量或观察体的集群。
节点 (Node)	是根、树叶，及分枝的统称。
长辈 (Parent)	两个 (或两个以上) 集群的合集。
晚辈 (Children)	构成上述联集的集群。

由上述定义可知，树形图的根是最高的长辈 (故画在顶端)，而树叶是最低的晚辈 (故画在底部)。现在，我们举一个例子来说明这六个名词及其相互间的关系。假设我们有九件家俱：桌灯、吊灯、夜光灯、书桌、餐桌、茶几、沙发、藤椅，及板凳。它们经过集群分析后所得的结果如下图：



“灯” 是 “桌灯” 、“吊灯”，及“夜光灯” 的长辈，“桌灯”，“吊灯” 及 “夜光灯” 是 “灯” 的晚辈。其余的分枝可比照同理，界定长辈与晚辈之间的关系。

若集群分析中，每一个集群至多只可有两个晚辈，则其相对应的树形称为二项式树形图。PROCCLUSTER 的分析结果永远是二项式树形图。PROC VARCLUS 则可产生多项式的树形图。

46.3　如何撰写 PROC TREE 程序

PROC TREE 含八项指令，它们的格式如下：

```
PROC TREE       选项串;
    NAME        变量名称;
    PARENT      变量名称;
    HEIGHT      变量名称;
    ID          变量名称;
    COPY        变量名称串;
    FREQ        变量名称;
    BY          变量名称串;
```

若 PROC TREE 的输入资料文件是由 PROC CLUSTER 或 PROC VARCLUS 所提供的，则上述格式里只有 PROC TREE 选项串; 是必要的。

指令 #1　PROC TREE 选项串;

此指令的选项极多，现分述如下：

(1) DATA= 输入资料文件名称

为树形图的输入资料文件命名。若省略此选项，则 SAS 自动找出在此程序之前最后形成的 SAS 资料文件，将它定义成树形图的输入资料文件。

(2) OUT= 输出资料文件名称

为树形图的输出资料文件命名。树形图的输出资料文件含原输入资料文件的所有变量 (或观察体) 与其所属集群的名字。集群的名字以 CLUSTER 或 CLUSNAME 来表示。若读者选用此选项，则必须界定选项 NCLUSTERS=或 LEVEL=的值以表明阶集分析的层次(Hierarchical Level)。

(3) HEIGHT= NCL(或 N)

HEIGHT= HEIGHT(或 H)

HEIGHT= MODE (或 M)

HEIGHT= RSQ (或 R)

HEIGHT= LENGTH(或 L)

此选项决定树形图上纵轴的单位。选项的值可以是五种可能性之一，它们的定义如下：

NCL　　：表示树形图上每一节点 (即层次) 的集群个数，所以纵轴的单位是等距离的。见本章例一的第二个树形图。

HEIGHT　　：由 _HEIGHT_ 变量而来，此变量由输入资料文件提供。

MODE　　：由 _MODE_ 变量而来，表示集群的凝聚点或中心点。

RSQ　　：由 _RSQ_ 变量而来，表示集群内可解释的变异数百分比。

LENGTH　　：集群形成前所必经的长辈个数。

(4) 下面两个选项的定义恰好相反，所以只能选择其中一个选项来描写被研究的数据：

选项	定义
(a) SIMILAR (或 SIM)	表示 HEIGHT 的值是相近值，因此 HEIGHT 值愈大，则集群愈相似。
(b) DISSIMILAR (或 DIS)	与 SIMILAR 刚好相反，HEIGHT 值愈大，则集群间的相似性愈小。

(5) LEVEL= 正实数

此选项与选项 HEIGHT= 合用，指示 SAS 选择树形图上的一个横切面，然后印出由树根到那一个横切面之间的集群组合。比方说，读者将 HEIGHT=NCL 与 LEVEL=5 并用，则输出资料文件就只含五个相斥性的集群。若将 HEIGHT=RSQ 与 LEVEL=.9 并用，则输出资料文件内含复相关系数平方大于或等于 .9 的所有集群。

(6) NCLUSTERS (或 NCL 或 N)= 正整数

读者可用此选项指定输出资料文件内集群的数目。有时输出资料文件内所含集群的数目和此选项所指定的值不一致，造成这种现象的原因有四：

(甲) 树形图上的树叶 (即个别的变量或观察体) 少于此选项所设的值；

(乙) 输入资料文件内已含有多于此选项值的树；

(丙) 一棵多重层次的树 (Multi-Way Tree) 上恰巧没有一个层次含有这么多的集群；

(丁) 读者选用选项 DOCK= 剔除掉太多集群。

NCLUSTERS= 选项利用资料文件内 _NCL_变量来决定集群形成的次序。若资料文件内缺乏_NCL_ 变量，则可用 HEIGHT 选项或 HEIGHT 指令来代替。

(7) DOCK= 正整数

当某集群内成员的数目等于或小于此值时，SAS 判定该集群不存在。这类集群在变量 CLUSTER 或 CLUSNAME 上算成遗漏数据；而且这类集群也不算在 NCLUSTERS= 的个数据，此选项的内设值是 0。

(8) ROOT='NAME 变量的值'

如果读者不想印出整个树形图，可以利用此选项指明树形图上的一个节点，则 SAS 会以此节点为新树根，印出从此节点到树叶的部分。并且输出资料文件将会只包含这一部分的数据。

(9) SORT

根据集群形成的顺序，将各树节点以下的晚辈依 HEIGHT 变量做由小到大的排

列。

(10) DESCENDING (或 DES)

作用与上述 SORT 选项刚好相反，晚辈的排名是由大到小。

(11) MINHEIGHT (或 MINH)= 正实数

界定树形图上纵轴的最小值。见本章例二的树形图：MINH=0。

(12) MAXHEIGHT (或 MAXH)= 正实数

界定树形图上纵轴的最大值。

(13) SPACES (或 S)= 正整数

界定树形图上树节点与树节点之间的空隙。

(14) PAGES= 正整数

界定整个树形图 (由树根到树叶) 的长度，此长度以电脑报表纸的页数表示。

(15) POS= 正整数

界定树形图上纵轴参考点 (|) 的总个数。

(16) TICKPOS= 正整数

界定树形图的纵轴上两墨点 (+) 之间参考点的总个数。内设值是在 5 到 10 之间，见本章例二的树形图：TICKPOS=10。

(17) NTICK= 正整数

界定树形图的纵轴上墨点 (+) 的总个数。见本章例二的树形图：NTICK=9。

(18) INC= 正整数

界定树形图的纵轴上墨点的累积值。见本章例二的树形图：INC=0.0001。若读者选用选项 HEIGHT=NCL，则 INC=1。若 HEIGHT 选项的值不是 NCL，则 INC= 的值会随之调整。

(19) LEAFCHAR (或 LC)='C'

指定一个英文字母来代表树形图上的树叶，此字母必须用两个单引号括住，内设值是句号 (.)。

(20) TREECHAR (或 TC)='C'

指定一个英文字母来代表树节点，此字母必须用两个单引号括住，内设值是X。

(21) JOINCHAR (或 JC)='C'

指定一个英文字母来代表两片树叶的联集，此字母必须括在两个单引号内，内设值也是X。

(22) FILLCHAR (或 FC)='C'

指定一个英文字母来代表叶与叶之间的空隙，此字母必须括在两个单引号内，内设值是空白。

(23) LIST

印出所有的树节点，节点与节点之间的高度，每一个节点的长辈或晚辈。

(24) HORIZONTAL (或 HOR)

要求将树形图横印。如此，树节点会印在横轴 (或 X 轴) 上，资料文件内的变量 (或观察体) 则印在纵轴 (或 Y 轴) 上。当树形图的长度超过一页的报表纸或读者盼望在屏幕上显示树形图时，使用 HORIZONTAL 的指令更能表达集群分析的

结果。

(25) NOPRINT

抑止树形图以及其它相关资料的印出。

指令 #2　NAME 变量名称；

此指令旨在为树节点命名。NAME 变量与下述的 PARENT 变量共同决定树形图的结构。若读者省略此指令，则 SAS 会在输入资料文件中寻找 _NAME_变量来代替；若资料文件中没有 _NAME_ 变量，则 SAS 会发出错误讯息，并停止执行命令。

指令 #3　PARENT 变量名称；

为树节点中属于长辈的集群命名。这个名字的长度应与指令 NAME 的长度相等。若读者省略此指令，则 SAS 会在输入资料文件中寻找 _PARENT_变量来代替。若资料文件中没有_PARENT_变量，则 SAS 会发出错误讯息，并停止执行指令。

指令 #4　HEIGHT 变量名称；

此指令的作用与变量 _HEIGHT_ 或 PROC TREE 指令中的选项 HEIGHT= (见前面的叙述) 完全相同，故不另赘述。

指令 #5　ID 变量名称；

用来识别报表上树形图的树叶。ID 变量可以是一个英文字母，也可以是一个数字。若省略此指令，则 SAS 以指令 NAME 变量或输入资料文件中的 _NAME_ 变量来代替。

指令 #6　COPY 变量名称串；

读者可利用此指令指明一组变量，然后 SAS 会将这些变量的值复印在输出资料文件内。

指令 #7　FREQ 变量名称；

表示每一片树叶 (最小的集群) 所含的观察体数目。若省略此指令，SAS 会在输入资料文件中寻找 _FREQ_ 变量来代替。若资料文件中没有 _FREQ_ 变量，则 SAS 会自动设定每一片树叶只含一个观察体，而且每一个树节点的观察体总数即是其晚辈观察体的总个数。

指令 #8　BY 变量名称串；

读者可用 BY 指令界定一个或一组变量，然后要求 SAS 依此 (组) 变量对输入资料文件加以挑选整理，进行多次不同 (组) 的绘图。

46.4 范　　例

例一：根据动物的牙齿将哺乳类动物分类

此例的输入资料文件是三十二种哺乳类动物不同部位的牙齿及数目，这些部位分别是上／下门牙(V1/V2)，上／下犬齿 (V3/V4)、上／下前臼齿 (V5/V6) 以及上／下臼齿(V7/V8)。

资料输入后，此例先用均连法进行层次式的集群分析，再依据分析的结果绘制三次树形图：第一次的树形图选用 SORT 选项，并根据内设值的原理以均连的距离当树形图的纵轴。第二次的树形图抑止整体图形的印出，只挑出树形图上有六个集群的那一个层次，将此六个集群的结构纳入输出资料文件。最后按集群的形成顺序一一加以分类后，再以 PROC PRINT 将所属的集群印出。

程　序

```
OPTIONS PAGESIZE=60 LINESIZE=110;
DATA TEETH;  TITLE 'MAMMALS' 'TEETH';
    INPUT MAMMAL $ 1-16 @21 (V1-V8) (1.);
    LABEL V1='TOP INCISORS'     V2='BOTTOM INCISORS'  V3='TOP CANINES'
          V4='BOTTOM CANINES'   V5='TOP PREMOLARS'    V6='BOTTOM PREMOLARS'
          V7='TOP MOLARS'       V8='BOTTOM MOLARS';   CARDS;
BROWN BAT            23113333
MOLE                 32103333
SILVER HAIR BAT      23112333
PIGMY BAT            23112233
HOUSE BAT            23111233
RED BAT              13112233
PIKA                 21002233
RABBIT               21003233
BEAVER               11002133
GROUNDHOG            11002133
GRAY SQUIRREL        11001133
HOUSE MOUSE          11000033
PORCUPINE            11001133
WOLF                 33114423
BEAR                 33114423
RACCOON              33114432
MARTEN               33114412
WEASEL               33113312
WOLVERINE            33114412
```

```
BADGER                33113312
RIVER OTTER           33114312
SEA OTTER             32113312
JAGUAR                33113211
COUGAR                33113211
FUR SEAL              32114411
SEA LION              32114411
GREY SEAL             32113322
ELEPHANT SEAL         21114411
REINDEER              04103333
ELK                   04103333
DEER                  04003333
MOOSE                 04003333
;
OPTIONS PAGESIZE=60 LINESIZE=110;
PROC CLUSTER METHOD=AVERAGE STD PSEUDO NOEIGEN OUTTREE=TREE; ID MAMMAL;  VAR V1-V8;
PROC TREE SORT;
PROC TREE NOPRINT OUT=PART NCLUSTERS=6;   ID MAMMAL;   COPY V1-V8;
PROC SORT ; BY CLUSTER;
PROC PRINT UNIFORM;   ID MAMMAL;   VAR V1-V8;   FORMAT V1-V8 1.;  BY CLUSTER;  RUN;
```

结　果

报表 46.1　根据动物的牙齿将哺乳类动物分类

MAMMALSTEETH

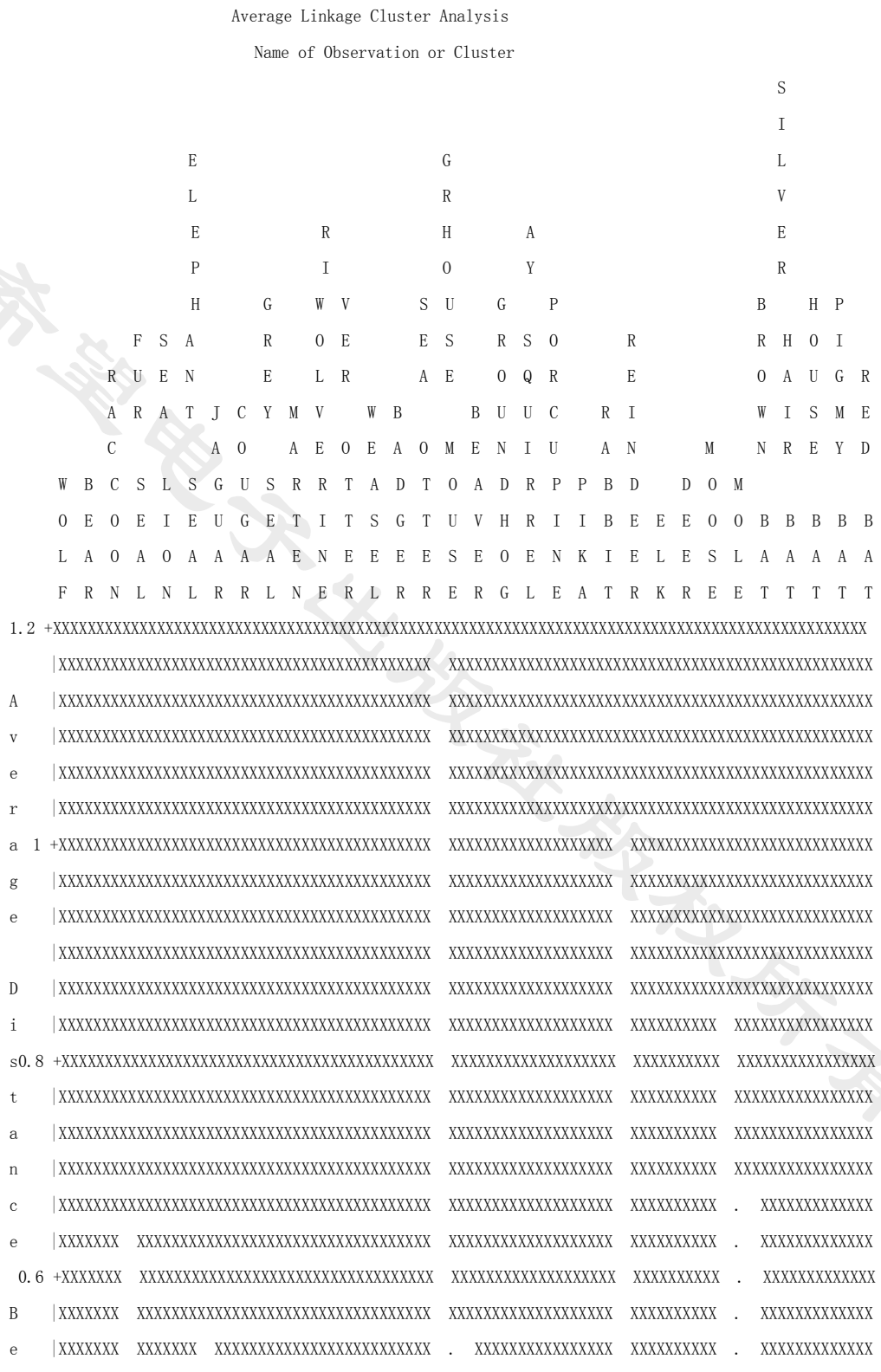
Average Linkage Cluster Analysis

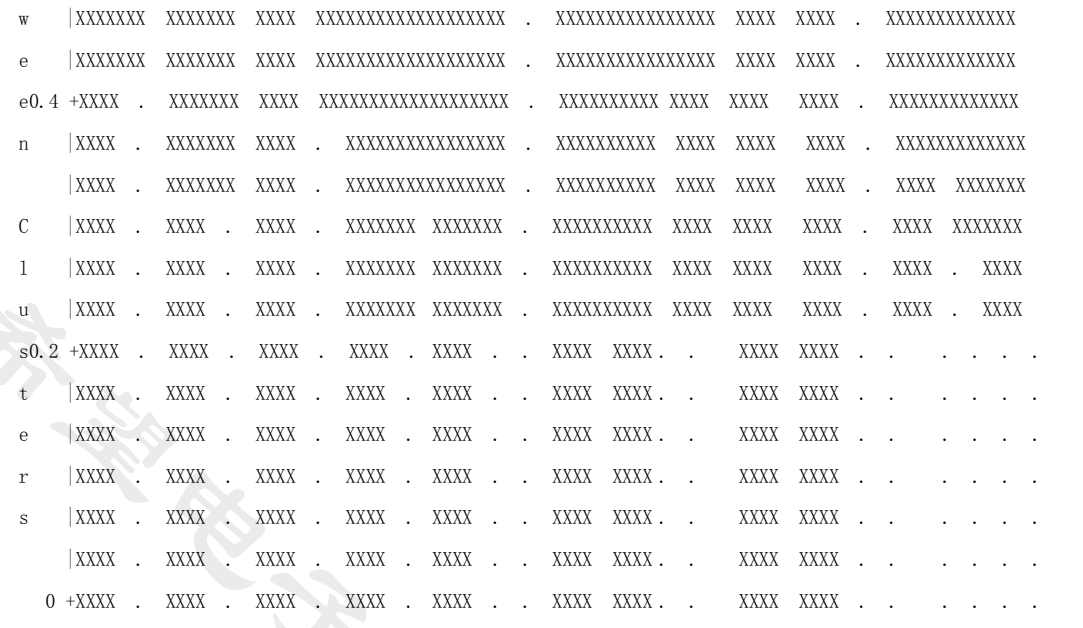
The data have been standardized to mean 0 and variance 1

Root-Mean-Square Total-Sample Standard Deviation =　　1

Root-Mean-Square Distance Between Observations　=　　4

Number of Clusters	Clusters Joined		Frequency of New Cluster	Pseudo F	Pseudo t**2	Normalized RMS Distance	Tie
31	BEAVER	GROUNDHOG	2	.	.	0.000000	T
30	GRAY SQUIRREL	PORCUPINE	2	.	.	0.000000	T
29	WOLF	BEAR	2	.	.	0.000000	T
28	MARTEN	WOLVERINE	2	.	.	0.000000	T
27	WEASEL	BADGER	2	.	.	0.000000	T
26	JAGUAR	COUGAR	2	.	.	0.000000	T
25	FUR SEAL	SEA LION	2	.	.	0.000000	T
24	REINDEER	ELK	2	.	.	0.000000	T
23	DEER	MOOSE	2	.	.	0.000000	
22	PIGMY BAT	RED BAT	2	281.19	.	0.228930	
21	CL28	RIVER OTTER	3	138.67	.	0.229221	
20	CL31	CL30	4	83.19	.	0.235702	T
19	BROWN BAT	SILVER HAIR BAT	2	76.71	.	0.235702	T
18	PIKA	RABBIT	2	73.24	.	0.235702	
17	CL27	SEA OTTER	3	67.37	.	0.246183	
16	CL22	HOUSE BAT	3	62.89	1.75	0.285937	
15	CL21	CL17	6	47.42	6.81	0.332845	
14	CL25	ELEPHANT SEAL	3	45.04	.	0.336177	
13	CL19	CL16	5	40.83	3.50	0.367188	
12	CL15	GREY SEAL	7	38.90	2.78	0.407838	
11	CL29	RACCOON	3	38.02	.	0.422997	
10	CL18	CL20	6	34.51	10.27	0.433918	
9	CL12	CL26	9	30.01	7.27	0.507122	
8	CL24	CL23	4	28.69	.	0.547281	
7	CL9	CL14	12	25.74	6.99	0.566841	
6	CL10	HOUSE MOUSE	7	28.32	4.12	0.579239	
5	CL11	CL7	15	26.83	6.87	0.662106	
4	CL13	MOLE	6	31.93	7.23	0.715610	
3	CL4	CL8	10	30.98	12.67	0.879851	
2	CL3	CL6	17	27.83	16.12	1.031622	
1	CL2	CL5	32	.	27.83	1.193815	


Average Linkage Cluster Analysis
Name of Observation or Cluster
SILVER
GREEN
BHP
RHOI
OAUGR
WISME
NREYD
1.2
A
v
e
r
a 1
g
e
D
i
s0.8
t
a
n
c
e
0.6
B
e
t



CLUSTER=1

MAMMAL	V1	V2	V3	V4	V5	V6	V7	V8
BEAVER	1	1	0	0	2	1	3	3
GROUNDHOG	1	1	0	0	2	1	3	3
GRAY SQUIRREL	1	1	0	0	1	1	3	3
PORCUPINE	1	1	0	0	1	1	3	3
PIKA	2	1	0	0	2	2	3	3
RABBIT	2	1	0	0	3	2	3	3
HOUSE MOUSE	1	1	0	0	0	0	3	3

CLUSTER=2

MAMMAL	V1	V2	V3	V4	V5	V6	V7	V8
WOLF	3	3	1	1	4	4	2	3
BEAR	3	3	1	1	4	4	2	3
RACCOON	3	3	1	1	4	4	3	2

CLUSTER=3

MAMMAL	V1	V2	V3	V4	V5	V6	V7	V8
MARTEN	3	3	1	1	4	4	1	2
WOLVERINE	3	3	1	1	4	4	1	2
WEASEL	3	3	1	1	3	3	1	2
BADGER	3	3	1	1	3	3	1	2
JAGUAR	3	3	1	1	3	2	1	1
COUGAR	3	3	1	1	3	2	1	1
FUR SEAL	3	2	1	1	4	4	1	1
SEA LION	3	2	1	1	4	4	1	1

RIVER OTTER	3	3	1	1	4	3	1	2
SEA OTTER	3	2	1	1	3	3	1	2
ELEPHANT SEAL	2	1	1	1	4	4	1	1
GREY SEAL	3	2	1	1	3	3	2	2

CLUSTER=4

MAMMAL	V1	V2	V3	V4	V5	V6	V7	V8
REINDEER	0	4	1	0	3	3	3	3
ELK	0	4	1	0	3	3	3	3
DEER	0	4	0	0	3	3	3	3
MOOSE	0	4	0	0	3	3	3	3

CLUSTER=5

MAMMAL	V1	V2	V3	V4	V5	V6	V7	V8
PIGMY BAT	2	3	1	1	2	2	3	3
RED BAT	1	3	1	1	2	2	3	3
BROWN BAT	2	3	1	1	3	3	3	3
SILVER HAIR BAT	2	3	1	1	2	3	3	3
HOUSE BAT	2	3	1	1	1	2	3	3

CLUSTER=6

MAMMAL	V1	V2	V3	V4	V5	V6	V7	V8
MOLE	3	2	1	0	3	3	3	3

例二：费氏紫罗兰的树形图

在第 44 与第 38 章内所提的费氏紫罗兰例子，在此再度出现，读者应参阅第 38 章或 44 章的例一以了解其资料文件结构，在此不再赘述。

资料文件输入后，先以 PROC CLUSTER 程序执行层次式的集群分析，然后以 PROC TREE 绘出分类的结果以及紫罗兰所属的集群类。

程　序

```
OPTIONS PAGESIZE=60 LINESIZE=110;
DATA IRIS;
    TITLE 'FISHER''S IRIS DATA';
    INPUT SEPALLEN SEPALWID PETALLEN PETALWID SPEC_NO @@;
    IF SPEC_NO=1 THEN SPECIES='SETOSA';
    ELSE IF SPEC_NO=2 THEN SPECIES='VERSICOLOR';
    ELSE IF SPEC_NO=3 THEN SPECIES='VIRGINICA';
    CARDS;
        (原数据刊登在第 38 章的例一)
;
PROC CLUSTER  DATA=IRIS METHOD=TWOSTAGE K=8 NOEIGEN PRINT=10;
    VAR SEPALLEN SEPALWID PETALLEN PETALWID;
```

```
    COPY SPECIES;
PROC TREE HORIZONTAL MINH=0 TICKPOS=10 NTICK=9 INC=0.0001;
    ID SPECIES;
RUN;
```

结　果

CLUSTER 程序以双连法先求出集群形成的先后顺序。报表上只印出十个集群的横切面。最后，用 PROC TREE 将整个分类的过程用 90 度旋转的树形图印出。

报表 46.2　费氏紫罗兰的树形图

FISHER'S IRIS DATA

Two-Stage Density Linkage Clustering

K = 8

Root-Mean-Square Total-Sample Standard Deviation = 10.69224

Number of Clusters	--Clusters Joined--		Frequency of New Cluster	Normalized Fusion Density	Normalized Maximum Density in Each Cluster Lesser	Greater	Tie
10	CL11	OB98	49	0.2879	0.1479	8.3678	
9	CL13	OB24	45	0.2802	0.2005	3.5156	
8	CL10	OB25	50	0.2699	0.1372	8.3678	
7	CL8	OB121	51	0.2586	0.1372	8.3678	
6	CL9	OB45	46	0.1412	0.0832	3.5156	
5	CL6	OB39	47	0.1070	0.0605	3.5156	
4	CL5	OB21	48	0.0969	0.0541	3.5156	
3	CL4	OB90	49	0.0715	0.0370	3.5156	

3 modal clusters have been formed.

Number of Clusters	--Clusters Joined--		Frequency of New Cluster	Normalized Fusion Density	Normalized Maximum Density in Each Cluster Lesser	Greater	Tie
2	CL7	CL3	100	2.6277	3.5156	8.3678	

```
                XXXX
VIRGIN          XXXX.....
                XXXX
VIRGIN          XXXX.....
                XXXX
VIRGIN          XXXX.....
                XXXX
VIRGIN          XXXX.....
                XXXX
VIRGIN          XXXX.....
                XXX
VIRGIN          XXX.....
                XXX
VERSIC          XXX.....
                XXX
VIRGIN          XXX.....
                XXX
VIRGIN          XXX.....
                XXX
VIRGIN          XXX.....
                XXX
VIRGIN          XXX.....
                XXX
VIRGIN          XXX.....
                XXX
VIRGIN          XXX.....
                XXX
VIRGIN          XXX.....
                XX
VIRGIN          XX.....
                XX
VIRGIN          XX.....
                XX
VIRGIN          XX.....
                XX
VIRGIN          XX.....
                XX
VIRGIN          XX.....
                XX
VIRGIN          XX.....
                XX
VIRGIN          XX.....
                X
VIRGIN          X.....
                X
VIRGIN          X.....
                X
VIRGIN          X.....
                X
VIRGIN          X.....
```

```
                X
VERSIC          XX.....
                XX
VERSIC          XX.....
                XX
VERSIC          XX.....
                XX
VERSIC          XXX.....
                XXX
VERSIC          XXX.....
                XXX
VERSIC          XXX.....
                XXX
VERSIC          XXX.....
                XXX
VERSIC          XXX.....
                XXX
VERSIC          XXX.....
                XXX
VERSIC          XXXX.....
                XXXX
VERSIC          XXXX.....
                XXXX
VERSIC          XXXX.....
                XXXX
VIRGIN          XXXX.....
                XXXX
VERSIC          XXXX.....
                XXXX
VERSIC          XXXX.....
                XXXX
VERSIC          XXXX.....
                XXXX
VERSIC          XXXXXX.....
                XXXXXX
VERSIC          XXXXXX.....
                XXXXXX
VERSIC          XXXXXX.....
                XXXXXX
VERSIC          XXXXXXX.....
                XXXXXXX
VERSIC          XXXXXXX.....
                XXXXXXX
VERSIC          XXXXXXXX.....
                XXXXXXXX
VERSIC          XXXXXXXXX.....
                XXXXXXXXX
VERSIC          XXXXXXXXX.....
                XXXXXXX
VERSIC          XXXXXXX.....
```

```
                XXXXXX
VERSIC          XXXXXX.....
                XXXXXX
VERSIC          XXXXXX.....
                XXXXX
VERSIC          XXXXX.....
                XXXX
VERSIC          XXXX.....
                XXXX
VERSIC          XXXX.....
                XXXX
VERSIC          XXXX.....
                XXXX
VERSIC          XXXX.....
                XXXX
VERSIC          XXXX.....
                XXXXX
VERSIC          XXXXX.....
                XXXXX
VERSIC          XXXXX.....
                XXXXX
VERSIC          XXXXX.....
                XXXXX
VERSIC          XXXXX.....
                XXXX
VERSIC          XXXX.....
                XXX
VERSIC          XXX.....
                XXX
VERSIC          XXX.....
                XXX
VERSIC          XXX.....
                XXX
VERSIC          XXX.....
                XXX
VERSIC          XXX.....
                XX
VIRGIN          XX.....
                XX
VERSIC          XX.....
                XX
VERSIC          XX.....
                X
VERSIC          X.....
                X
VERSIC          X.....
                X
VERSIC          X.....
                X
VERSIC          X.....
```

```
SETOSA           XXX......................................................................
                 XXX
SETOSA           XXXX.....................................................................
                 XXXX
SETOSA           XXXX.....................................................................
                 XXXX
SETOSA           XXXXXX...................................................................
                 XXXXXX
SETOSA           XXXXXX...................................................................
                 XXXXXX
SETOSA           XXXXXX...................................................................
                 XXXXXX
SETOSA           XXXXXX...................................................................
                 XXXXXX
SETOSA           XXXXXXX..................................................................
                 XXXXXXX
SETOSA           XXXXXXXXXXX..............................................................
                 XXXXXXXXXXX
SETOSA           XXXXXXXXXXX..............................................................
                 XXXXXXXXXXX
SETOSA           XXXXXXXXXXXX.............................................................
                 XXXXXXXXXXXX
SETOSA           XXXXXXXXXXXXXXXXX........................................................
                 XXXXXXXXXXXXXXXXX
SETOSA           XXXXXXXXXXXXXXXXXX.......................................................
                 XXXXXXXXXXXXXXXXXX
SETOSA           XXXXXXXXXXXXXXXXXX.......................................................
                 XXXXXXXXXXXXXXXXXX
SETOSA           XXXXXXXXXXXXXXXXXXXX.....................................................
                 XXXXXXXXXXXXXXXXXXXX
SETOSA           XXXXXXXXXXXXXXXXXXXX.....................................................
                 XXXXXXXXXXXXXXXXXXXX
SETOSA           XXXXXXXXXXXXXXXXXXXXX....................................................
                 XXXXXXXXXXXXXXXXXXXXX
SETOSA           XXXXXXXXXXXXXXXXXXXXXXX..................................................
                 XXXXXXXXXXXXXXXXXXXXXXX
SETOSA           XXXXXXXXXXXXXXXXXXXXXXXXXXXXX............................................
                 XXXXXXXXXXXXXXXXXXXXXXXXXXXXX
SETOSA           XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.........................................
                 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
SETOSA           XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX................................
                 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
SETOSA           XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX............................
                 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
SETOSA           XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX..........
                 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
SETOSA           XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
                 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
SETOSA           XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
                 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
```

```
SETOSA           XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX...........................
                 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
SETOSA           XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX...............................
                 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
SETOSA           XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................................
                 XXXXXXXXXXXXXXXXXXXXXXXXXXXXX
SETOSA           XXXXXXXXXXXXXXXXXXXXXXXXXXXXX...........................................
                 XXXXXXXXXXXXXXXXXXXXXXXXX
SETOSA           XXXXXXXXXXXXXXXXXXXXXXXXX...............................................
                 XXXXXXXXXXXXXXXXXXXXXXX
SETOSA           XXXXXXXXXXXXXXXXXXXXXXX.................................................
                 XXXXXXXXXXXXXXXXXXXXX
SETOSA           XXXXXXXXXXXXXXXXXXXXX...................................................
                 XXXXXXXXXXXXXXXXXXXX
SETOSA           XXXXXXXXXXXXXXXXXXXX....................................................
                 XXXXXXXXXXXXXXXXXXX
SETOSA           XXXXXXXXXXXXXXXXXXX.....................................................
                 XXXXXXXXXXXXXXXXXX
SETOSA           XXXXXXXXXXXXXXXXXX......................................................
                 XXXXXXXXXXXXXXXX
SETOSA           XXXXXXXXXXXXXXXX........................................................
                 XXXXXXXXXXX
SETOSA           XXXXXXXXXXX.............................................................
                 XXXXXXXXXXX
SETOSA           XXXXXXXXXXX.............................................................
                 XXXXXXXXXXX
SETOSA           XXXXXXXXXXX.............................................................
                 XXXXXX
SETOSA           XXXXXX..................................................................
                 XXXXXX
SETOSA           XXXXXX..................................................................
                 XXXXXX
SETOSA           XXXXXX..................................................................
                 XXXXX
SETOSA           XXXXX...................................................................
                 XXX
SETOSA           XXX.....................................................................
                 XXX
SETOSA           XXX.....................................................................
                 XXX
SETOSA           XXX.....................................................................
                 XX
SETOSA           XX......................................................................
                 XX
SETOSA           XX......................................................................
                 X
SETOSA           X.......................................................................
                 X
SETOSA           X.......................................................................
```

第 47 章　共变异数估计值的集群分析法：统计程序 PROC ACECLUS

47.1　PROC ACECLUS 程序概述

ACECLUS 是共变异数估计值的集群分析法 (Approximate Covariance Estimation for Clustering) 的简称。此分析法由 Art，Gnanadesikan，及 Kettenring 于 1982 年所提出。简而言之，本分析法的目的是利用集群内共变异数的近似估计值来执行集群分析。

ACECLUS 假设每一集群内的成员是由多元常态分配里随机抽样取来的，而且每一个多元常态分配的共变异数矩阵是相等的。

在进行 ACECLUS 分析之前，读者不须预知集群的个数或大小。因此本分析只能提供一个粗略的集群结构，其结果最好再用另外的集群分析程序 (如：PROC CLUSTER 或 PROC FASTCLUS) 来处理。

分析的过程

- 根据各观察体间的距离，将总变异数矩阵分成集群间与集群内的变异数矩阵。
- 所求得的集群内变异数矩阵只是近似值，但此近似值已足以用来计算集群分析法中集群之间的距离。
- 根据如此计算得到的距离，再经由一般的集群法设法取得最小的值，或在典型鉴别法内作组间的鉴别，即可完成集群分析。

47.2　对集群分析的贡献

根据 Everitt (1980) 的论文：大多数的集群分析法只适用于几何形态，如球形的集群，并不适用于椭圆形的集群。假如在这种椭圆形状的集群里，我们有办法利用线性转换将集群内的共变异数矩阵转变成球形的矩阵 (Spherical Matrix)，则一般的集群分析法便可以继续处理这个转换后的球形矩阵。

在线性转换的过程里，我们必须假定每一个椭圆形集群的形状与近似椭圆的程度是大同小异的。如此，经过线性转换后，集群内元素间的距离可以用变异数的倒数当做测量的单位。然而这个以线性转换来解决椭圆形集群的方法却不易付诸实行。因为在分析开始之前，集群的个数与大小均是未知数，如何能谈到集群的共变异数矩阵？于是 Art，Gnanadesikan 与 Kettenring 在 1982 年提出共变异数近似估计值的集群分析法，解决了这个难题。此法在本章中又称为 AGK 法，以纪念这三位创始人。

AGK 法对集群分析的贡献可由下页的表看出：

表 47.1　利用四集群法分析费氏(Fisher，1936)紫罗兰资料文件，所产生错误的分类以及无法分类的观察体个数

	集群法			
数类型	K-平均数法	华滋法	均连法	重心法
(1) 原始资料	16*	16*	25+12*	14*
(2) 经标准化后的数	25	26	33+4	33+4
(3) 两个标准化的主成份	29	31	30+9	27+32
(4) 四个标准化的主成份	39	27	32+7	45+11
(5) 经 ACECLUS (P=.32)转换后的数**	39	10+9	7+25	0
(6) 经 ACECLUS (P=.16)转换后的数	39	18+9	7+19	7+26
(7) 经 ACECLUS (P=.08)转换后的数	19	9	3+13	5+16
(8) 经 ACECLUS (P=.04)转换后的数	4	5	1+19	3+12
(9) 经 ACECLUS (P=.02)转换后的数	4	3	3	3
(10) 经 ACECLUS (P=.01)转换后的数	4	4	3	4
(11) 经 ACECLUS (P=.005) 转换后的数	4	4	4	4
(12) 典型相关变量	3	5	4	4+1

* 一个数目表该法所误分的观察体个数。若两个数目以加号分离，则加号前的数目表示该集群法误分的观察体个数，而加号后的数目表示无法分类的观察体个数。如：以华滋法来分析原始资料，有十六个误分的观察体，而以均连法来分析原始资料，则有二十五个误分及十二个无法分类的观察体。

**ACECLUS 中 P 值的定义请见本章第 47.3 节：PROC ACECLUS 指令的选项部分。

根据表 47.1 所示的分析结果，我们有下面几点建议：

(A) 对 ACECLUS 程序中选项 P = 的值须慎重加以选择。若读者有充分的时间，不妨多尝试几个不同的值。若 P 值选择合适，则 ACECLUS 所得的结果几乎趋近典型相关 (最佳的线性转换) 的结果。

(B) 一般读者所熟悉的线性转换，如：标准化或主成份的转换，有时会导致极不正确的分析结果。然而若将这些经过标准化转换的数据，用 ACECLUS 处理后，似乎又可以改善分类的结果。

总而言之，本章所介绍的 ACECLUS (或 AGK 法)，利多于弊。我们极力建议读者在任何集群法之前先用 ACECLUS 来处理数据。

47.3　如何撰写 PROC ACECLUS 程序

PROC ACECLUS 含五道指令，它们的格式如下：

```
PROC ACECLUS     选项串;
  VAR            变量名称串;
  FREQ           变量名称;
```

WEIGHT	变量名称；
BY	变量名称串；

一般而言，一个程序中只要采用 PROC ACECLUS 及 VAR 两道指令即可应付大部分研究者所分析的资料。

指令 #1　PROC ACECLUS 选项串；

本指令的选项可分六大类，现分别介绍如下：

第一类选项　下列三个选项与资料文件的界定有关：

(1) DATA=输入资料文件名称

指明到底对那一个资料文件执行分析。若省略此选项，则 SAS 会自动找出在此程序之前最后形成的 SAS 资料文件，对它执行分析。

(2) OUT=第一个输出资料文件的名称

这一个资料文件含所有的输入资料以及典型变量。典型变量的名称及数目的多少由选项 PREFIX= 及 N= 来决定。

(3) OUTSTAT= 第二个输出资料文件的名称

这一个资料文件与上述资料文件不同，它只含统计分析结果而且其结构属于 ACE (即 TYPE=ACE)。有关这些统计分析值的变量名称以及它们的定义，请见下页表的说明：

变量	定义
MEAN	各变量的平均数
N	参与分析的观察体总数
SUMWGT	由 WEIGHT 指令而来，代表所有加权比重值的总和，是一个常数。
CO	参与分析的各变量之间的共变异数矩阵 (以全部数据为计算的单位)。
ACE	参与分析的各变量之间的平均共变异数矩阵 (以集群内的数据为计算的单位)。
EIGENVAL	是矩阵 INV(ACE) * (COV-ACE) 的特性根。其数目由 PROC ACECLUS 指令中的选项 N= 决定。
SC0RE	特性向量。这些向量的名字与其所对应的典型变量的名字一样，皆由 PROCACECLUS 指令中的选项 PREFIX= 而来。向量的个数就是典型变量的个数。

除此之外，OUTSTAT 的资料文件还包括 BY 指令中所有的变量串 (由指令 BY 所提供)，以及另外两个新变量：_TYPE_ 和 _NAME_。 _TYPE_ 变量含这些统计分析的结果， 而 _NAME_ 变量含一些统计分析结果的名称。 OUTSTAT 所建立的资料文件适用于：

- 下一个 ACECLUS 程序的分析；
- SCORE 程序的分析 (以便计算典型变量的分数)；
- FACTOR 程序的分析 (采用 METHOD=SC0RE)，以便对典型变量作坐标轴转换。转换后的因子结构有助于解释数据间隐含的结构。

第二类选项　下列两个选项与典型变量有关：

(1) N=正整数

界定典型相关分析中典型变量的个数。内设值是参与分析的所有变量的总数。若N=0，则 SAS 不进行典型相关分析。

(2) PREFIX=英文名字

为典型变量命名。比方说 PREFIX=ABC，则典型变量的名字依序便是 ABC1，ABC2，ABC3，....等等。这个英文名字加上编号不得超过八个字母的长度，内设值是 CANn。

第三类选项　下列这个选项控制共变异数矩阵估计过程中的初值：

(1) INITIAL=FULL (F)　　或

INITIAL=DIAGONAL (D)　或

INITIAL=IDENTITY (I)　　或

INITIAL=INPUT=SAS 资料文件

指出集群内共变异数矩阵初步估计值的来源。

当 INITIAL=FULL 时，初值是总样本的共变异数矩阵。

当 INITIAL=DIAGONAL 时，初值是各集群样本的变异数 / 共变异数矩阵的对角线值。

当 INITIAL=IDENTITY 时，初值是单元矩阵。

当 INITIAL=INPUT=SAS 资料文件时，初值就是这个资料文件所提供的矩阵。

这个选项的内设值是如此决定的：

如果同时选用选项 METRIC=，则 INITIAL 的内设值就是选项 METRIC=的值。当省略选项 METRIC= 时，INITIAL 的内设值一般而言是 FULL。但若这个选择会导致一个非满秩的矩阵，则 ACECLUS 程序会再自动将内设值定为 DIAGONAL。

第四类选项　下列六个选项控制共变异数矩阵估计的过程：[选项(1)或(2)是必须的]

(1) THRESHOLD (或 T) = 正实数 (如 .5)

这个选项的值决定一对观察体是否有资格被纳入共变异数矩阵的估计。若这一对观察体之间的欧氏距离小于或等于此 T 值乘上所有个体与个体间距离之平方根 (Root Mean Square) 的乘积，则这一对观察体就有资格被纳入估计的过程。

(2) PROPORTION (或 PERCENT，或 P) = 正小数

界定样本中，到底有几对观察体会被纳入估计过程的百分比。在多元常态分配的假设下，ACECLUS 程序会根据 PROPORTION= 的值计算出一个底线值 (Threshold)。

(3) ABSOLUTE

界定选项 THRESHOLD= 中的 T 值或由 PROPORTION= 选项中导出的底线值是绝对的值，而非受平方根影响的相对值。读者必须在对初值的估计有相当的把握时，才可以选用这个选项。比方说：你在选项 INITIAL= 中，选用 INITIAL=INPUT= 上一个 ACECLUS 程序所产生的 OUTSTAT 资料文件，则你可以得到较优的初步估计值。

(4)　MAXITER=　正整数

界定循环估计的循环次数，内设值等于 10。

(5)　CONVERGE=　正小数

界定循环估计的收敛值。 内设值是 0.0001。 因此循环估计何时停止， 视 MAXITER=或 CONVERGE= 选项而定。只要达到其中一个选项的标准，则 ACECLUS 程序便停止估计的过程。

(6) SINGULAR=　极小的正实数

界定共变异数矩阵非满秩性的标准，内设值是 10 的 -4 次方。

第五类选项　下列四个选项控制报表打印的各式资料：

(1) PP

要求绘制估计过程中由最后一个循环所求得的欧氏距离图。此图的横轴表示距离的大小，纵轴表示各距离出现的概率。

(2) QQ

将上述 PP 选项中的欧氏距离作非线性的转换，然后印出转换后距离的次数表。此选项将占用极多的电脑运算时间，故读者应谨慎选用。

(3) SHORT

要求只印出循环估计的记录与其特性根。

(4) NOPRINT

不印出任何报表。

第六类选项　下列这个选项决定欧氏距离的度量衡单位：

(1) METRIC=FULL (F)　或

METRIC=DIAGONAL (D)　或

METRIC=IDENTITY (I)

此处 FULL，DIAGONAL 与 IDENTITY 的定义与本指令中选项 INITIAL= 的定义一致。一般而言，本选项的内设值是 FULL (如果样本含足够的观察体而且全部样本的共变异数矩阵是满秩的)，或 DIAGONAL (如果上述的条件不符合)。

指令 #2　VAR 变量名称串；

列出输入资料文件内所有参与分析的数值变量名称。若省略此指令，则 SAS 会找出本程序内其它指令未提及的所有数值变量，将它们纳入分析。

指令 #3　FREQ 变量名称；

这一个变量 (称为加权或次数变量) 的值应是一个正整数。这些正整数代表资料文件中各观察体重复出现的次数。

指令 #4　WEIGHT 变量名称；

这个指令的作用与上述 FREQ 指令相似。这个变量 (称为加权比重变量) 的值可以是一个正有理数。当观察体的变异数不等时，你可将变异数的倒数当作加权比重变量的值。如此，SAS 会调整不等的变异数，使其相等。

指令 #5　BY 变量名称串；

这个指令的作用是将输入资料文件按 BY 变量的值分成几个小资料文件，然后针对每一个小资料文件执行 ACECLUS 分析。在使用这个指令之前，输入资料文件应先经过 PROC SORT 的处理，使数据按 BY 变量串的值做由小到大的排列。

当你同时选用 BY 指令与选项 INITIAL=INPUT=SAS 资料文件 (暂命名为 OUTSTAT) 时，有三种可能发生的情况，分述如后：

第一种情况

若 OUTSTAT 资料文件内不含任何 BY 变量，则 OUTSTAT 资料文件的初步估计值成为各小资料文件分析时的初步估计值。

第二种情况

若 OUTSTAT 资料文件内只含一部分的 BY 变量，则 SAS 判断这种写法为错误，于是中断分析的过程。

第三种情况

若 OUTSTAT 资料文件内含所有的 BY 变量，则估计过程中，各小资料文件的初步估计值 (仍由 OUTSTAT 资料文件提供) 可以不同。在此，OUTSTAT 资料文件的数据仍然必须经过 PROC SORT 的重新排列。

47.4　范　　例

例一：费氏紫罗兰的集群分析

本资料文件 (IRIS) 的数据来自费契尔氏 (Fisher，1936)。它常被用来当做范例以解释集群分析。这一组资料是从三种不同属性的紫罗兰 (SETOSA=1，VERSICOLOR=2，及 VIRGINICA=3) 搜集而来的。每种紫罗兰各取五十个样本，然后测量它们花萼与花瓣的长与宽 (测量单位是厘米)。每一个观察体包括五个数据，依次是：花萼长，宽；花瓣长，宽；及属性号码。

在本范例中，先用 ACECLUS 程序转换数据，然后再用第 44 章所介绍的 FASTCLUS 程序对转换后的数据执行集群分析。请读者仔细比较本章与第 44 章分析的结果。

程　序

```
DATA IRIS;
    TITLE 'FISHER (1936) IRIS DATA';
    INPUT SEPALLEN SEPALWID PETALLEN PETALWID SPEC_NO @@;
    IF SPEC_NO=1 THEN SPECIES='SETOSA';
    ELSE IF SPEC_NO=2 THEN SPECIES='VERSICOLOR';
    ELSE SPECIES='VIRGINICA';
    LABEL SEPALLEN='SEPAL LENGTH IN MM.'  (花萼长)
          SEPALWID='SEPAL WIDTH  IN MM.'  (花萼宽)
```

```
                PETALLEN='PETAL LENGTH IN MM.'  (花瓣长)
                PETALWID='PETAL WIDTH  IN MM.'; (花瓣宽)
            CARDS;
             ( 原数据刊登在第 38 章的例一 )
        ;
        PROC ACECLUS DATA=IRIS OUT=ACE P=.02;
            VAR SEPALLEN SEPALWID PETALLEN PETALWID;
        PROC PLOT VPERCENT=200;  PLOT CAN2*CAN1=SPEC_NO;
        PROC FASTCLUS DATA=ACE MAXC=3 MAXITER=10 OUT=CLUS;  VAR CAN:;
        PROC FREQ;
            TABLES CLUSTER*SPECIES;
        RUN;
```

结　果

经 ACECLUS 程序转换数据的结果，找出四个典型变量。利用这四个典型变量的值以及三个集群的要求，FASTCLUS 程序在十次循环内就将一百五十株紫罗兰分得很正确。从 CLUSTER 程序对原属性类别的次数分配表看来，错分的紫罗兰只有五株：一株是属 VERSICOLOR 种，另四株属 VIRGINICA 种。

所以，利用 ACECLUS 执行 AGK 集群法的效果在此例中被肯定。

报表 47.1　费氏紫罗兰的集群分析

FISHER (1936) IRIS DATA

Approximate Covariance Estimation for Cluster Analysis

150 Observations　Proportion = 0.02

4 Variables　Converge = 0.001

Means and Standard Deviations

Variable	Mean	Std Dev	Label
SEPALLEN	58.433333	8.280661	SEPAL LENGTH IN MM.
SEPALWID	30.573333	4.358663	SEPAL WIDTH IN MM.
PETALLEN	37.580000	17.652982	PETAL LENGTH IN MM.
PETALWID	11.993333	7.622377	PETAL WIDTH IN MM.

COV: Total Sample Covariances

	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	68.56935	-4.2434	127.4315	51.62707
SEPALWID	-4.2434	18.99794	-32.9656	-12.1639
PETALLEN	127.4315	-32.9656	311.6278	129.5609
PETALWID	51.62707	-12.1639	129.5609	58.10063

Initial Within-Cluster Covariance Estimate = Full Covariance Matrix

Threshold = 0.334211

Iteration	RMS Distance	Distance Cutoff	Pairs Within Cutoff	Convergence Measure
1	2.828	0.945	408	0.465775
2	11.905	3.979	559	0.013487
3	13.152	4.396	940	0.029499
4	13.439	4.491	1506	0.046846
5	13.271	4.435	2036	0.046859
6	12.591	4.208	2285	0.025027
7	12.199	4.077	2366	0.009559
8	12.121	4.051	2402	0.003895
9	12.064	4.032	2417	0.002051
10	12.047	4.026	2429	0.000971

ACE: Approximate Covariance Estimate Within Clusters

	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	11.73343	5.475504	4.95389	2.029024
SEPALWID	5.475504	6.919926	2.421779	1.741252
PETALLEN	4.95389	2.421779	6.537464	2.353026
PETALWID	2.029024	1.741252	2.353026	2.051667

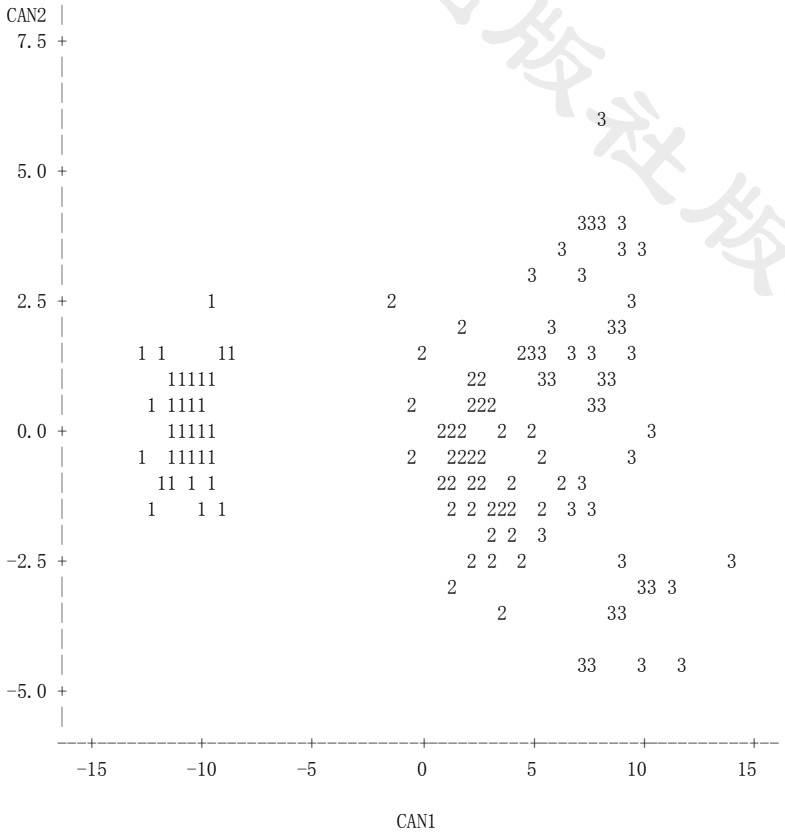
Eigenvalues of Inv(ACE)*(COV-ACE)

	Eigenvalue	Difference	Proportion	Cumulative
CAN1	63.7716	61.1593	0.936720	0.93672
CAN2	2.6123	1.5561	0.038372	0.97509
CAN3	1.0562	0.4167	0.015515	0.99061
CAN4	0.6395	.	0.009394	1.00000

Eigenvectors

	CAN1	CAN2	CAN3	CAN4	
SEPALLEN	-.012009	-.098074	-.059852	0.402352	SEPAL LENGTH IN MM.
SEPALWID	-.211068	-.000072	0.402391	-.225993	SEPAL WIDTH IN MM.
PETALLEN	0.324705	-.328583	0.110383	-.321069	PETAL LENGTH IN MM.
PETALWID	0.266239	0.870434	-.085215	0.320286	PETAL WIDTH IN MM.

Plot of CAN2*CAN1. Symbol is value of SPEC_NO.



NOTE: 35 obs hidden.

Replace=FULL Radius=0 Maxclusters=3 Maxiter=10 Converge=0.02

Initial Seeds

Cluster	CAN1	CAN2	CAN3	CAN4
1	-12.9510	-0.2516	1.8471	2.7075
2	13.8749	-2.5641	-0.4212	1.9412
3	-0.3181	0.5975	-2.8784	-0.8496

Minimum Distance Between Initial Seeds = 13.97481

Iteration	Criterion	Relative Change in Cluster Seeds 1	2	3
1	2.7950	0.2725	0.4188	0.3238
2	1.5265	0	0.0498	0.0438
3	1.4949	0	0.0153	0.0148

Convergence criterion is satisfied.

Criterion Based on Final Seeds = 1.492

Cluster Summary

Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Nearest Cluster	Distance Between Cluster Centroids
1	50	1.1016	5.2768	3	13.4326
2	47	1.9033	6.5797	3	5.8560
3	53	1.4337	5.5425	2	5.8560

Statistics for Variables

Variable	Total STD	Within STD	R-Squared	RSQ/(1-RSQ)
CAN1	8.048079	1.477330	0.966757	29.081409
CAN2	1.900612	1.866587	0.048431	0.050896
CAN3	1.433954	1.326865	0.155277	0.183820
CAN4	1.280440	1.278112	0.017007	0.017301
OVER-ALL	4.244987	1.505058	0.875982	7.063348

Pseudo F Statistic = 519.16

Approximate Expected Over-All R-Squared = 0.80391

Cubic Clustering Criterion = 5.148

WARNING: The two above values are invalid for correlated variables.

	Cluster Means				Cluster Standard Deviations			
Cluster	CAN1	CAN2	CAN3	CAN4	CAN1	CAN2	CAN3	CAN4
1	-10.6752	0.0671	0.2707	0.1116	0.95376	0.93194	1.39846	1.05822
2	8.3094	0.5051	0.5535	0.1347	1.70216	2.81910	1.29883	1.39941
3	2.7022	-0.5112	-0.7462	-0.2248	1.65818	1.41445	1.28135	1.35291

TABLE OF CLUSTER BY SPECIES

CLUSTER / Frequency / Percent / Row Pct / Col Pct	SETOSA	VERSIC	VIRGIN	Total
1	50	0	0	50
	33.33	0.00	0.00	33.33
	100.00	0.00	0.00	
	100.00	0.00	0.00	
2	0	1	46	47
	0.00	0.67	30.67	31.33
	0.00	2.13	97.87	
	0.00	2.00	92.00	
3	0	49	4	53
	0.00	32.67	2.67	35.33
	0.00	92.45	7.55	
	0.00	98.00	8.00	
Total	50	50	50	150
	33.33	33.33	33.33	100.00