返回总目录

目　　录

第七部分
多变量的分析

第 33 章　SAS 系统内四种多变量分析程序概述

33.1　四种多变量分析的统计程序

本章将简要地介绍四种多变量分析的统计程序，即主成份分析 (PRINCOMP)，传统式因子分析 (FACTOR)，典型相关分析 (CANCORR) 和多次元尺度分析 (MDS)。这四种统计程序的功能在于找寻多个变量之间的关系或简化数据的复杂性。这些变量并不一定得视为自变量或因变量。其中，主成份分析、传统式因子分析以及多次元尺度分析都是对一组变量作分析，而典型相关则是对两组变量作分析。SAS 还有其它的统计程序可以执行多变量的统计分析，如：CATMOD、变异数分析、回归分析、集群分析及鉴别分析等。

若读者熟悉在 SAS 旧版的环境下执行这些程序，则建议直接参考附录 D 有关这些程序增进的简介。

下面分别介绍这四种程序。

■ PRINCOMP 程序（主成份分析）

对同一组观察体的多个变量执行主成份分析。主成份分析的目的是找出一组变量之间互相依赖的程度，将这些线性相关以主成份值表示。其分析的结果包括未经标准化及标准化后的主成份值。这些主成份值可以代替变量的原始数据，进行进一步的分析、处理，如：制图，执行回归分析或集群分析。值得读者注意的是：主成份分析 (Principal Component Analysis) 与主轴因子分析 (Principal Axis Common Factor Analysis) 不是同义词。

■ FACTOR 程序（传统式因子分析）

对同一组观察体内的多个变量执行上述的主成份分析及传统式因子分析。因子分析法还附带有因子的坐标转换，以取得最大的诠释效果。其分析结果可以是标准化的主成份值，也可以是传统因子分析的值。传统式因子分析的目的在寻求一小群隐藏的变量以解释原变量之间的相关，和主成份分析不同的是这一小群隐藏的变量不直接由原变量间的线性组合导出。一般国内教科书将因子分析翻译成 "因素分析"；因此，对本书读者而言，这两个名词实系同义词。

■ CANCORR 程序（典型相关分析）

对两组变量执行典型相关分析，其分析的结果是典型变量值。典型相关分析的目的是藉一小群有最高组间相关的组内变量之线性组合 (又称向量) 来解释并概述两组变量之间的关系。构成向量的变量多少并没有限制，若某个向量中只含一个变量，则典型相关的作用与回归分析或皮尔森相关系数类似。

■ MDS 程序（多次元尺度分析）

MDS 是 Multidimensional Scaling 的简称，它代表一系列的分析法，其目的在于从一组距离矩阵中找出观察体 (或变量或刺激词) 的坐标。如此，读者可藉图形的视觉效果来检视点与点之间的关系以及潜在向度的意义。

33.2　主成份分析和传统式因子分析的比较

如上所述，FACTOR 程序除了涵盖 PRINCOMP 程序，并且包括了另外几种常用的因子分析法。当读者使用 FACTOR 程序时，若不指明用那一种分析法，则主成份分析便是 FACTOR 程序的内设值。FACTOR 程序产生的主成份值是经过标准化的，然而 PRINCOMP 程序所产生的主成份值是未经标准化的。不过，读者也可额外地要求 PRINCOMP 算出标准化的主成份值。

与 FACTOR 程序相比，PRINCOMP 程序的优点如下：

(1) 最适用于变量多，但主成份少的大型资料文件；可节省电脑处理时间。
(2) 易于使用。
(3) 输入资料文件可以是一个净相关系数矩阵或一个净共变异数矩阵。

与 PRINCOMP 程序相比，FACTOR 程序的优点如下：

(1) 产生的分析结果较 PRINCOMP 程序广泛，包括：误差值的检定，因子坐标转换的角度，及特性根由大到小的排列等。
(2) 包含好几种坐标转换的理论。
(3) 其输出矩阵较易了解。
(4) 所涵盖的因子分析法较完全。PRINCOMP 程序只有一种分析法，即：主成份分析法，然而 FACTOR 程序内有九种分析法供你选择。

第 34 章　主成份分析：统计程序 PROC PRINCOMP

34.1　PROC PRINCOMP 程序概述

读者可用 PRINCOMP 程序对输入资料文件执行主成份分析。其输入资料文件可以是原始数据，也可以是一个相关系数矩阵，或是一个变异数／共变异数矩阵。输出资料则包括特性根、特性向量及 (未经) 标准化的主成份值。

主成份分析是一个多变量的统计程序，可用来检定多个数值变量之间的关系。主成份分析除了用来概述变量间的关系外，还可用来削减回归或集群分析中变量的数目。它的主要目的是求出一组变量的线性组合 (即主成份)，这些线性组合就是原变量矩阵的特性向量。每一个向量的内乘积就是该向量对原变量群能解释的变异数百分比。这些特性向量之间应该是彼此线性独立的。

主成份分析首由皮尔森氏 (Pearson) 于 1901 年提出。其后经过赫德林氏 (Hotelling，1933) 的发扬。有关其应用可见罗氏 (Rao，1964)，古氏及隆斯氏 (Cooley and Lohnes，1971) 和干那氏 (Gnanadesikan，1977) 的着作。

34.2　如何撰写 PROC PRINCOMP 程序

PROC PRINCOMP 含六道指令，它们的格式如下：

```
PROC PRINCOMP    选项串;
      VAR        变量名称串;
      PARTIAL    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，只须用到前两个指令，亦即 PROC PRINCOMP 以及 VAR。

指令 #1　PROC PRINCOMP 选项串;

有下列十个选项可供选择：

(1) DATA=输入资料文件名称

指明到底对那一个 SAS 资料文件执行 PROC PRINCOMP 的分析。这个输入资料文件可以是原始数据，也可以是一个相关系数矩阵 (TYPE=CORR 或 UCORR)，或是一个变异数／共变异数矩阵 (TYPE=COV 或 UCOV)，或 TYPE=FACTOR，SSCP，ESP 等不同形式的资料文件。若省略此选项，则 SAS 会自动找出在此程序之前最后形成的资料文件，对它执行主成份分析。

(2) OUT=输出资料文件名称

这一个输出资料文件包括输入资料文件的数据以及主成份值。

(3) OUTSTAT=输出资料文件名称

这一个输出资料文件包含下列的统计值：算术平均数、标准差、观察体的总数、相关系数 (或变异数／共变异数)、特性根和特性向量。它们的代号与定义如下：

代号_(_TYPE_)	定　　义
MEAN	每一变量的平均数
STD	每一变量的标准差
N	观察体的总个数
CORR	每一变量与自己或其它的变量之间的相关系数
COV	每一变量与自己或其它的变量之间的共变异数
EIGENVAL	特性根。当选项 N= 界定成份个数少于实际导出的个数，则以 N= 界定的个数为准，其余的主成份以遗漏值 (.) 表示。
SCORE	特性向量 (这些向量值一般是用来计算主成份值或被输送到 FACTOR 程序作因子坐标的转换)
SUMWGT	加权值的总和，若读者在程序中包括了 PARTIAL 指令而且定 VARDEF=WDF，则 SUMWGT 的值是加权值的总和减去 PARTIAL 变量串的自由度。当 SUMWGT 与 N 值相同时，SUMWGT 的变量不会被纳入 OUTSTAT= 输出资料文件内。

(4) NOINT

要求相关矩阵或变异数／共变异数矩阵不针对平均数作校正，也就是说主成份分析不包括截距。

(5) COVARIANCE (或 COV)

要求以变异数／共变异数矩阵为分析的数据。若省略此选项，则此统计分析将以相关系数矩阵为依据。

(6)　N=正整数

界定主成份的总数。

(7) STANDARD(或 STD)

要求 OUT=输出资料文件中含标准化的主成份值。若省略此选项，则输出资料文件中将含未经标准化的主成份值 (这些值的变异数等于特性根的值)。

(8) PREFIX=主成份的名字

为主成份命名。内设值是 PRIN1，PRIN2，... PRINn，n 为正整数，主成份的名字 (包括字母及数字) 不得超过八个字母或数字。

(9) NOPRINT

不印出分析的结果。

(10) VARDEF=DF (或 N 或 WGT 或 WDF)

界定计算变异数与共变异数时所用的分母。DF 代表自由度，是此选项的内设值；N 是样本总数；WGT 是加权后的样本总数；WDF 则是 (WGT-1)。

指令 #2　VAR 变量名称串;

指明对那些数值变量作主成份分析。若省略此指令，则本程序内其它指令里未曾提到的所有数值变量均将被纳入分析。

指令 #3　PARTIAL 变量名称串;

此指令指明一组变量，它们的值将会从其它的变量中净化出来。净化后的变量值所形成的矩阵是净相关系数矩阵而非相关系数矩阵。若读者在程序中同时界定 OUT= 或 OUTSTAT= 输出资料文件名，则此输出资料文件也会含净化后的残差变量 (Residual Variable)。这些残差变量的命名原则是 R_ 加上 VAR 指令所界定之变量名称的前六个字母。所以，如果 VAR 指令含 X， Y， Z 三个变量，则其所对应的残差变量就是 R_X，R_Y，R_Z 了。

指令 #4　FREQ 变量名称;

此变量的值代表资料文件内各观察体重复出现的次数。所以，计算自由度时，将以这个变量的总值为依据。

指令 #5　WEIGHT 变量名称;

当输入资料文件内各观察体的变异数不等时，读者常须依这些不等变异数的倒数指派不同的加权值以区分各观察体的重要性。这些加权值可被存入一个 WEIGHT 变量内，以代表各观察体的加权值。

指令 #6　BY 变量名称串;

此指令指示 SAS 将输入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进行主成份分析。当读者选用此指令时，输入资料文件内的数据必须先依 BY 指令里所列举的变量值作从小到大的排列，这个步骤可藉 PROC SORT 达成。

34.3　范　　例

例一：一月和七月的气温分析

本例的输入资料文件 (TEMPERAT) 是美国六十四个城市一月与七月的平均日温。分析过程首先用 PROC PLOT 画出原始数据的分配图。然后用 PRINCOMP 程序执行主成份分析求出两个主轴 (PRIN1，PRIN2)。由于一月的温差较大而且选用 COV 选项，使得一月在第一主成份上的负荷量较重。最后用 PROC PLOT 画出两个主成份上各城市的负荷量。读者可同时参阅第一次与第二次 PLOT 程序所求得的两个图表，来归纳出第一与第二主成份是原坐标轴旋转 30 度的结果。

程　序

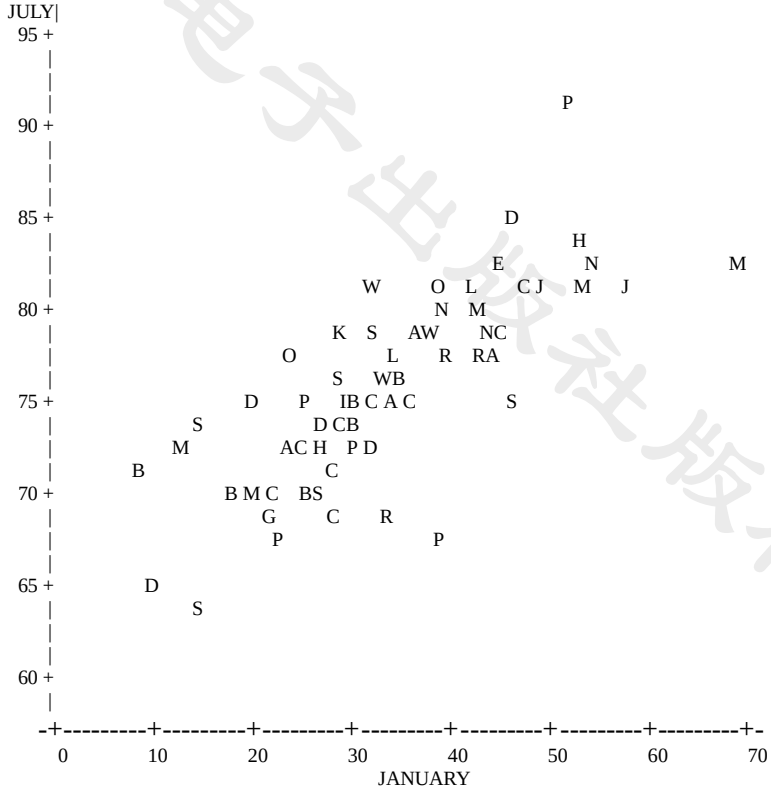
```
DATA TEMPERAT;
    LENGTH CITY $ 16;
    TITLE 'Mean Temperature in January and July for Selected Cities';
    INPUT CITY $ :16. JANUARY :4.1 JULY :5.1 @@;
CARDS;
Mobile          51.2 81.6 Concord          20.6 69.7
Phoenix         51.2 91.2 Atlantic_City    32.7 75.1
Little_Rock     39.5 81.4 Albuquerque      35.2 78.7
Sacramento      45.1 75.2 Albany           21.5 72.0
Denver          29.9 73.0 Buffalo          23.7 70.1
Hartford        24.8 72.7 New_York         32.2 76.6
Wilmington      32.0 75.8 Charlotte        42.1 78.5
Washington_DC   35.6 78.7 Raleigh          40.5 77.5
Jacksonville    54.6 81.0 Bismarck         8.2 70.8
Miami           67.2 82.3 Cincinnati       31.1 75.6
Atlanta         42.4 78.0 Cleveland        26.9 71.4
Boise           29.0 74.5 Columbus         28.4 73.6
Chicago         22.9 71.9 Oklahoma_City    36.8 81.5
Peoria          23.8 75.1 Portland_OR      38.1 67.1
Indianapolis    27.9 75.0 Philadelphia     32.3 76.8
Des_Moines      19.4 75.1 Pittsburgh       28.1 71.9
Wichita         31.3 80.7 Providence       28.4 72.1
Louisville      33.3 76.9 Columbia         45.4 81.2
New_Orleans     52.9 81.9 Sioux_Falls      14.2 73.3
Porland_ME      21.5 68.0 Memphis          40.5 79.6
Baltimore       33.4 76.6 Nashville        38.3 79.6
Boston          29.2 73.3 Dallas           44.8 84.8
Detroit         25.5 73.3 El_Paso          43.6 82.3
Sault_Ste_Marie 14.2 63.8 Houston          52.1 83.3
Duluth          8.5 65.6 Salt_Lake_City    28.0 76.7
Minneapolis     12.2 71.9 Burlington       16.8 69.8
Jackson         47.1 81.7 Norfolk          40.5 78.3
Kansas_City     27.8 78.8 Richmond         37.5 77.9
St_Louis        31.3 78.6 Spokane          25.4 69.7
Great_Falls     20.5 69.3 Charleston_WV    34.5 75.0
Omaha           22.6 77.2 Milwaukee        19.4 69.9
Reno            31.9 69.3 Cheyenne         26.6 69.1
```

```
;
PROC PLOT;
        PLOT JULY*JANUARY=CITY / VPOS=31;    RUN;
PROC PRINCOMP COV OUT=PRIN;
        VAR JULY JANUARY;    RUN;
PROC PLOT;
        PLOT PRIN2*PRIN1=CITY / VPOS=19;
        TITLE2    'Plot of Principal Components';    RUN;
```

结　果

报表 34.1　一月和七月的气温分析

Mean Temperature in January and July for Seleted Cities
Plot of JULY*JANUARY. Symbol is value of CITY.



NOTE: 3 obs hidden.

Principal Component Analysis

64 Observations
2 Variables

Simple Statistics

	JULY	JANUARY
Mean	75.60781250	32.09531250
Std	5.12761910	11.71243309

Covariance Matrix

	JULY	JANUARY
JULY	26.2924777	46.8282912
JANUARY	46.8282912	137.1810888

Total Variance = 163.47356647

Eigenvalues of the Covariance Matrix

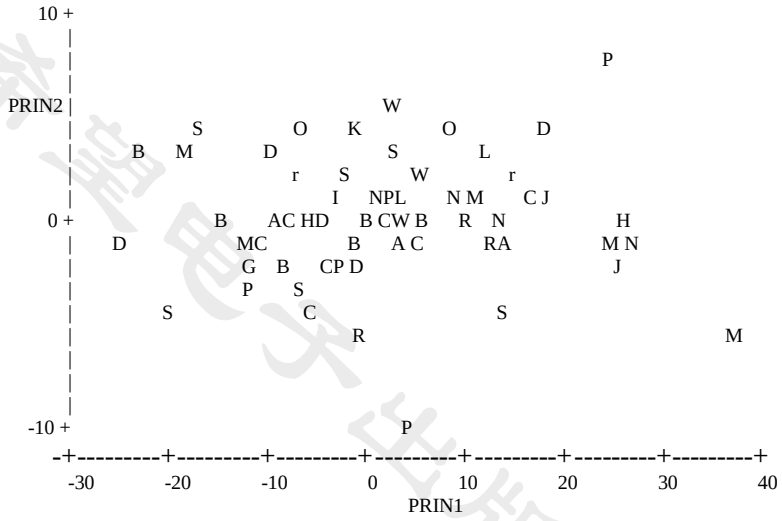
	Eigenvalue	Difference	Proportion	Cumulative
PRIN1	154.311	145.148	0.943948	0.94395
PRIN2	9.163		.0.056052	1.00000

Eigenvectors

	PRIN1	PRIN2
JULY	0.343532	0.939141
JANUARY	0.939141	-.343532

Plot of Principal Components

Plot of PRIN2*PRIN1. Symbol is value of CITY.



NOTE: 4 obs hidden.

例二：犯罪率的分析

本例的输入资料文件 (CRIME) 是一个五十个观察体乘以七个变量的原始数据矩阵，它包含了美国五十个州在七种犯罪项目上的发生频率。这七种罪名分别是：谋杀(MURDER)、强暴 (RAPE)、抢劫 (ROBBERY)、骚扰 (ASSAULT)、夜间偷窃(BURGLARY)、窃盗 (LARCENY) 及偷车 (AUTO)。这样一个大型的资料文件可以用主成份分析法简化到只用两个或三个特性向量就可以圆满地表示。

程　序

```
DATA CRIME;
    TITLE 'Crime Rates per 100,000 Population by State';
    INPUT STATE $ 1-14 MURDER 18-21 RAPE 23-26 ROBBERY 28-32 ASSAULT 34-38
          BURGLARY 40-45 LARCENY 47-52 AUTO 53-59; CARDS;
Alabama        14.2 25.2  96.8 278.3 1135.5 1881.9  280.7
Alaska         10.8 51.6  96.8 284.0 1331.7 3369.8  753.3
Arizona         9.5 34.2 138.2 312.3 2346.1 4467.4  439.5
Arkansas        8.8 27.6  83.2 203.4  972.6 1862.1  183.4
California     11.5 49.4 287.0 358.0 2139.4 3499.8  663.5
Colorado        6.3 42.0 170.7 292.9 1935.2 3903.2  477.1
```

Connecticut	4.2	16.8	129.5	131.8	1346.0	2620.7	593.2
Delaware	6.0	24.9	157.0	194.2	1682.6	3678.4	467.0
Florida	10.2	39.6	187.9	449.1	1859.9	3840.5	351.4
Georgia	11.7	31.1	140.5	256.5	1351.1	2170.2	297.9
Hawaii	7.2	25.5	128.0	64.1	1911.5	3920.4	489.4
Idaho	5.5	19.4	39.6	172.5	1050.8	2599.6	237.6
Illinois	9.9	21.8	211.3	209.0	1085.0	2828.5	528.6
Indiana	7.4	26.5	123.2	153.5	1086.2	2498.7	377.4
Iowa	2.3	10.6	41.2	89.8	812.5	2685.1	219.9
Kansas	6.6	22.0	100.7	180.5	1270.4	2739.3	244.3
Kentucky	10.1	19.1	81.1	123.3	872.2	1662.1	245.4
Louisiana	15.5	30.9	142.9	335.5	1165.5	2469.9	337.7
Maine	2.4	13.5	38.7	170.0	1253.1	2350.7	246.9
Maryland	8.0	34.8	292.1	358.9	1400.0	3177.7	428.5
Masssachusetts	3.1	20.8	169.1	231.6	1532.2	2311.3	1140.1
Michigen	9.3	38.9	261.9	274.6	1522.7	3159.0	545.5
Minnesota	2.7	19.5	85.9	85.8	1134.7	2559.3	343.1
Mississippi	14.3	19.6	65.7	189.1	915.6	1239.9	144.4
Missouri	9.6	28.3	189.0	233.5	1318.3	2424.2	378.4
Montana	5.4	16.7	39.2	156.8	804.9	2773.2	309.2
Nebraska	3.9	18.1	64.7	112.7	760.0	2316.1	249.1
Nevada	15.8	49.1	323.1	355.0	2453.1	4212.6	559.2
New Hampshire	3.2	10.7	23.2	76.0	1041.7	2343.9	293.4
New Jersey	5.6	21.0	180.4	185.1	1435.8	2774.5	511.5
New Mexico	8.8	39.1	109.6	343.4	1418.7	3008.6	259.5
New York	10.7	29.4	472.6	319.1	1728.0	2782.0	745.8
North Carolina	10.6	17.0	61.3	318.3	1154.1	2037.8	192.1
North Dakota	0.9	9.0	13.3	43.8	446.1	1843.0	144.7
Ohio	7.8	27.3	190.5	181.1	1216.0	2696.8	400.4
Oklahoma	8.6	29.2	73.8	205.0	1288.2	2228.1	326.8
Oregon	4.9	39.9	124.1	286.9	1636.4	3506.1	388.9
Pennsylvania	5.6	19.0	130.3	128.0	877.5	1624.1	333.2
Rhode Island	3.6	10.5	86.5	201.0	1489.5	2844.1	791.4
South Carolina	11.9	33.0	105.9	485.3	1613.6	2342.4	245.1
South Dakota	2.0	13.5	17.9	155.7	570.5	1704.4	147.5
Tennessee	10.1	29.7	145.8	203.9	1259.7	1776.5	314.0
Texas	13.3	33.8	152.4	208.2	1603.1	2988.7	397.6
Utah	3.5	20.3	68.8	147.3	1171.6	3004.6	334.5

```
Vermont          1.4 15.9  30.8 101.2 1348.2 2201.0  265.2
Virginia         9.0 23.3  92.1 165.7  986.2 2521.2  226.7
Washington       4.3 39.6 106.2 224.8 1605.6 3386.9  360.3
West Virginia    6.0 13.2  42.2  90.9  597.4 1341.7  163.3
Wisconsin        2.8 12.9  52.2  63.7  846.9 2614.2  220.7
Wyoming          5.4 21.9  39.7 173.9  811.6 2772.2  282.0
;
PROC PRINCOMP OUT=CRIMCOMP;
RUN;
PROC SORT;      BY PRIN1;
PROC PRINT;     ID STATE;
      VAR PRIN1 PRIN2 MURDER RAPE ROBBERY ASSAULT BURGLARY LARCENY
                      AUTO;
      TITLE2 'States Listed in Order of Overall Crime Rate';
      TITLE3 'As Determined by the First Principal Component';
PROC SORT;      BY PRIN2;
PROC PRINT;     ID STATE;
      VAR PRIN1 PRIN2 MURDER RAPE ROBBERY ASSAULT BURGLARY LARCENY
                      AUTO ;
      TITLE2 'States Listed in Order of Property Vs. Violent Crime';
      TITLE3 'As Determined by the Second Principal Component';
RUN;
PROC PLOT;      PLOT PRIN2*PRIN1=STATE / VPOS=31;
      TITLE2 'Plot of the First Two Principal Components';
PROC PLOT;      PLOT PRIN3*PRIN1=STATE / VPOS=26;
      TITLE2 'Plot of the First and Third Principal Components';
RUN;
```

结　果

由初步的分析结果看来，前两个主成份加起来便可以解释 76% 的变异数。若再加上第三个主轴，则百分比升到 87%，但第四个及以后的主成份便没有这么显著的影响 (见报表 34.2a)。第一个主成份代表一般犯罪率的高低，它的特性向量在这七个变量上差不多。第二个主成份似乎在犯罪类型中分出财物偷窃和暴力犯罪的不同。第三主成份的解释则不甚清楚。为了诠释这些主成份的意义，可将原始数据依各主成份的值重新排列，然后印出整理过后的数据 (见报表 34.2b)。

另一种有效的方法是将各州主成份的值以坐标图表示，然后试着去了解各区 (如：中西部，东南部) 在坐标图上的分布 (见报表 34.2c)。现举一例说明如何在坐标图上识别各州。如：第一图上有四个 "A" 开头的州名，即：Alabama，Arkansas，Alaska 和 Arizona。在这四州中，Alabama 的位置最靠近横轴，其坐标值是 (-.0499，-2.0961)。请读者同时参

阅坐标值与图形，以便识别各州在犯罪率上的分析。

报表 34.2a　犯罪率的分析 — 初步结果

Crime Rates per 100,000 Population by State

Principal Component Analysis

50 Observations

7 Variables

Simple Statistics

	MURDER	RAPE	ROBBERY	ASSAULT	BURGLARY	LARCENY	AUTO
Mean	7.444000000	25.73400000	124.0920000	211.3000000	1291.904000	2671.288000	377.5260000
Std	3.866768941	10.75962995	88.3485672	100.2530492	432.455711	725.908707	193.3944175

Correlation Matrix

	MURDER	RAPE	ROBBERY	ASSAULT	BURGLARY	LARCENY	AUTO
MURDER	1.0000	0.6012	0.4837	0.6486	0.3858	0.1019	0.0688
RAPE	0.6012	1.0000	0.5919	0.7403	0.7121	0.6140	0.3489
ROBBERY	0.4837	0.5919	1.0000	0.5571	0.6372	0.4467	0.5907
ASSAULT	0.6486	0.7403	0.5571	1.0000	0.6229	0.4044	0.2758
BURGLARY	0.3858	0.7121	0.6372	0.6229	1.0000	0.7921	0.5580
LARCENY	0.1019	0.6140	0.4467	0.4044	0.7921	1.0000	0.4442
AUTO	0.0688	0.3489	0.5907	0.2758	0.5580	0.4442	1.0000

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
PRIN1	4.11496	2.87624	0.587851	0.58785
PRIN2	1.23872	0.51291	0.176960	0.76481
PRIN3	0.72582	0.40938	0.103688	0.86850
PRIN4	0.31643	0.05846	0.045205	0.91370
PRIN5	0.25797	0.03593	0.036853	0.95056
PRIN6	0.22204	0.09798	0.031720	0.98228
PRIN7	0.12406	0.017722	1.00000	

Eigenvectors

	PRIN1	PRIN2	PRIN3	PRIN4	PRIN5	PRIN6	PRIN7
MURDER	0.300279	-.629174	0.178245	-.232114	0.538123	0.259117	0.267593
RAPE	0.431759	-.169435	-.244198	0.062216	0.188471	-.773271	-.296485
ROBBERY	0.396875	0.042247	0.495861	-.557989	-.519977	-.114385	-.003903
ASSAULT	0.396652	-.343528	-.069510	0.629804	-.506651	0.172363	0.191745
BURGLARY	0.440157	0.203341	-.209895	-.057555	0.101033	0.535987	-.648117
LARCENY	0.357360	0.402319	-.539231	-.234890	0.030099	0.039406	0.601690
AUTO	0.295177	0.502421	0.568384	0.419238	0.369753	-.057298	0.147046

报表 34.2b　犯罪率的分析—第一与第二主成份值的排列

Crime Rates per 100,000 Population by State
States Listed in Order of Overall Crime Rate
As Determined by the First Principal Component

STATE	PRIN1	PRIN2	MURDER	RAPE	ROBBERY	ASSAULT	BURGLARY	LARCENY	AUTO
NorthDakota	-3.96408	0.38767	0.9	9.0	13.3	43.8	446.1	1843.0	144.7
SouthDakota	-3.17203	-0.25446	2.0	13.5	17.9	155.7	570.5	1704.4	147.5
WestVirginia	-3.14772	-0.81425	6.0	13.2	42.2	90.9	597.4	1341.7	163.3
Iowa	-2.58156	0.82475	2.3	10.6	41.2	89.8	812.5	2685.1	219.9
Wisconsin	-2.50296	0.78083	2.8	12.9	52.2	63.7	846.9	2614.2	220.7
NewHampshire	-2.46562	0.82503	3.2	10.7	23.2	76.0	1041.7	2343.9	293.4
Nebraska	-2.15071	0.22574	3.9	18.1	64.7	112.7	760.0	2316.1	249.1
Vermont	-2.06433	0.94497	1.4	15.9	30.8	101.2	1348.2	2201.0	265.2
Maine	-1.82631	0.57878	2.4	13.5	38.7	170.0	1253.1	2350.7	246.9
Kentucky	-1.72691	-1.14663	10.1	19.1	81.1	123.3	872.2	1662.1	245.4
Pennsylvania	-1.72007	-0.19590	5.6	19.0	130.3	128.0	877.5	1624.1	333.2
Montana	-1.66801	0.27099	5.4	16.7	39.2	156.8	804.9	2773.2	309.2
Minnesota	-1.55434	1.05644	2.7	19.5	85.9	85.8	1134.7	2559.3	343.1
Mississippi	-1.50736	-2.54671	14.3	19.6	65.7	189.1	915.6	1239.9	144.4
Idaho	-1.43245	-0.00801	5.5	19.4	39.6	172.5	1050.8	2599.6	237.6
Wyoming	-1.42463	0.06268	5.4	21.9	39.7	173.9	811.6	2772.2	282.0
Arkansas	-1.05441	-1.34544	8.8	27.6	83.2	203.4	972.6	1862.1	183.4
Utah	-1.04996	0.93656	3.5	20.3	68.8	147.3	1171.6	3004.6	334.5
Virginia	-0.91621	-0.69265	9.0	23.3	92.1	165.7	986.2	2521.2	226.7
NorthCarolina	-0.69925	-1.67027	10.6	17.0	61.3	318.3	1154.1	2037.8	192.1
Kansas	-0.63407	-0.02804	6.6	22.0	100.7	180.5	1270.4	2739.3	244.3
Connecticut	-0.54133	1.50123	4.2	16.8	129.5	131.8	1346.0	2620.7	593.2
Indiana	-0.49990	0.00003	7.4	26.5	123.2	153.5	1086.2	2498.7	377.4
Oklahoma	-0.32136	-0.62429	8.6	29.2	73.8	205.0	1288.2	2228.1	326.8
RhodeIsland	-0.20156	2.14658	3.6	10.5	86.5	201.0	1489.5	2844.1	791.4
Tennessee	-0.13660	-1.13498	10.1	29.7	145.8	203.9	1259.7	1776.5	314.0
Alabama	-0.04988	-2.09610	14.2	25.2	96.8	278.3	1135.5	1881.9	280.7
NewJersey	0.21787	0.96421	5.6	21.0	180.4	185.1	1435.8	2774.5	511.5
Ohio	0.23953	0.09053	7.8	27.3	190.5	181.1	1216.0	2696.8	400.4
Georgia	0.49041	-1.38079	11.7	31.1	140.5	256.5	1351.1	2170.2	297.9
Illinois	0.51290	0.09423	9.9	21.8	211.3	209.0	1085.0	2828.5	528.6
Missouri	0.55637	-0.55851	9.6	28.3	189.0	233.5	1318.3	2424.2	378.4
Hawaii	0.82313	1.82392	7.2	25.5	128.0	64.1	1911.5	3920.4	489.4
Washington	0.93058	0.73776	4.3	39.6	106.2	224.8	1605.6	3386.9	360.3
Delaware	0.96458	1.29674	6.0	24.9	157.0	194.2	1682.6	3678.4	467.0
Masssachusett	0.97844	2.63105	3.1	20.8	169.1	231.6	1532.2	2311.3	1140.1
Louisiana	1.12020	-2.08327	15.5	30.9	142.9	335.5	1165.5	2469.9	337.7
NewMexico	1.21417	-0.95076	8.8	39.1	109.6	343.4	1418.7	3008.6	259.5
Texas	1.39696	-0.68131	13.3	33.8	152.4	208.2	1603.1	2988.7	397.6
Oregon	1.44900	0.58603	4.9	39.9	124.1	286.9	1636.4	3506.1	388.9

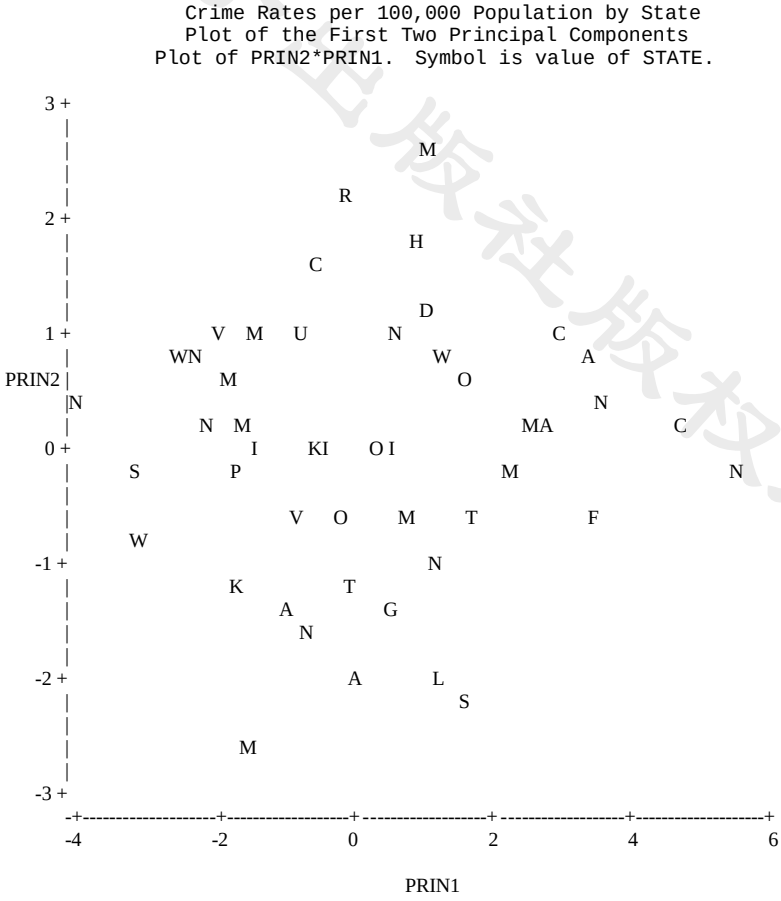
SouthCarolina	1.60336	-2.16211	11.9	33.0	105.9	485.3	1613.6	2342.4	245.1
Maryland	2.18280	-0.19474	8.0	34.8	292.1	358.9	1400.0	3177.7	428.5
Michigen	2.27333	0.15487	9.3	38.9	261.9	274.6	1522.7	3159.0	545.5
Alaska	2.42151	0.16652	10.8	51.6	96.8	284.0	1331.7	3369.8	753.3
Colorado	2.50929	0.91660	6.3	42.0	170.7	292.9	1935.2	3903.2	477.1
Arizona	3.01414	0.84495	9.5	34.2	138.2	312.3	2346.1	4467.4	439.5
Florida	3.11175	-0.60392	10.2	39.6	187.9	449.1	1859.9	3840.5	351.4
NewYork	3.45248	0.43289	10.7	29.4	472.6	319.1	1728.0	2782.0	745.8
California	4.28380	0.14319	11.5	49.4	287.0	358.0	2139.4	3499.8	663.5
Nevada	5.26699	-0.25262	15.8	49.1	323.1	355.0	2453.1	4212.6	559.2

Crime Rates per 100,000 Population by State
States Listedin Orderof Property Vs. Violent Crime
As Determined by the Second Principal Component

STATE	PRIN1	PRIN2	MURDER	RAPE	ROBBERY	ASSAULT	BURGLARY	LARCENY	AUTO
Mississippi	-1.50736	-2.54671	14.3	19.6	65.7	189.1	915.6	1239.9	144.4
SouthCarolina	1.60336	-2.16211	11.9	33.0	105.9	485.3	1613.6	2342.4	245.1
Alabama	-0.04988	-2.09610	14.2	25.2	96.8	278.3	1135.5	1881.9	280.7
Louisiana	1.12020	-2.08327	15.5	30.9	142.9	335.5	1165.5	2469.9	337.7
NorthCarolina	-0.69925	-1.67027	10.6	17.0	61.3	318.3	1154.1	2037.8	192.1
Georgia	0.49041	-1.38079	11.7	31.1	140.5	256.5	1351.1	2170.2	297.9
Arkansas	-1.05441	-1.34544	8.8	27.6	83.2	203.4	972.6	1862.1	183.4
Kentucky	-1.72691	-1.14663	10.1	19.1	81.1	123.3	872.2	1662.1	245.4
Tennessee	-0.13660	-1.13498	10.1	29.7	145.8	203.9	1259.7	1776.5	314.0
NewMexico	1.21417	-0.95076	8.8	39.1	109.6	343.4	1418.7	3008.6	259.5
WestVirginia	-3.14772	-0.81425	6.0	13.2	42.2	90.9	597.4	1341.7	163.3
Virginia	-0.91621	-0.69265	9.0	23.3	92.1	165.7	986.2	2521.2	226.7
Texas	1.39696	-0.68131	13.3	33.8	152.4	208.2	1603.1	2988.7	397.6
Oklahoma	-0.32136	-0.62429	8.6	29.2	73.8	205.0	1288.2	2228.1	326.8
Florida	3.11175	-0.60392	10.2	39.6	187.9	449.1	1859.9	3840.5	351.4
Missouri	0.55637	-0.55851	9.6	28.3	189.0	233.5	1318.3	2424.2	378.4
SouthDakota	-3.17203	-0.25446	2.0	13.5	17.9	155.7	570.5	1704.4	147.5
Nevada	5.26699	-0.25262	15.8	49.1	323.1	355.0	2453.1	4212.6	559.2
Pennsylvania	-1.72007	-0.19590	5.6	19.0	130.3	128.0	877.5	1624.1	333.2
Maryland	2.18280	-0.19474	8.0	34.8	292.1	358.9	1400.0	3177.7	428.5
Kansas	-0.63407	-0.02804	6.6	22.0	100.7	180.5	1270.4	2739.3	244.3
Idaho	-1.43245	-0.00801	5.5	19.4	39.6	172.5	1050.8	2599.6	237.6
Indiana	-0.49990	0.00003	7.4	26.5	123.2	153.5	1086.2	2498.7	377.4
Wyoming	-1.42463	0.06268	5.4	21.9	39.7	173.9	811.6	2772.2	282.0
Ohio	0.23953	0.09053	7.8	27.3	190.5	181.1	1216.0	2696.8	400.4
Illinois	0.51290	0.09423	9.9	21.8	211.3	209.0	1085.0	2828.5	528.6
California	4.28380	0.14319	11.5	49.4	287.0	358.0	2139.4	3499.8	663.5
Michigen	2.27333	0.15487	9.3	38.9	261.9	274.6	1522.7	3159.0	545.5
Alaska	2.42151	0.16652	10.8	51.6	96.8	284.0	1331.7	3369.8	753.3
Nebraska	-2.15071	0.22574	3.9	18.1	64.7	112.7	760.0	2316.1	249.1
Montana	-1.66801	0.27099	5.4	16.7	39.2	156.8	804.9	2773.2	309.2
NorthDakota	-3.96408	0.38767	0.9	9.0	13.3	43.8	446.1	1843.0	144.7
NewYork	3.45248	0.43289	10.7	29.4	472.6	319.1	1728.0	2782.0	745.8

Maine	-1.82631	0.57878	2.4	13.5	38.7	170.0	1253.1	2350.7	246.9
Oregon	1.44900	0.58603	4.9	39.9	124.1	286.9	1636.4	3506.1	388.9
Washington	0.93058	0.73776	4.3	39.6	106.2	224.8	1605.6	3386.9	360.3
Wisconsin	-2.50296	0.78083	2.8	12.9	52.2	63.7	846.9	2614.2	220.7
Iowa	-2.58156	0.82475	2.3	10.6	41.2	89.8	812.5	2685.1	219.9
NewHampshire	-2.46562	0.82503	3.2	10.7	23.2	76.0	1041.7	2343.9	293.4
Arizona	3.01414	0.84495	9.5	34.2	138.2	312.3	2346.1	4467.4	439.5
Colorado	2.50929	0.91660	6.3	42.0	170.7	292.9	1935.2	3903.2	477.1
Utah	-1.04996	0.93656	3.5	20.3	68.8	147.3	1171.6	3004.6	334.5
Vermont	-2.06433	0.94497	1.4	15.9	30.8	101.2	1348.2	2201.0	265.2
NewJersey	0.21787	0.96421	5.6	21.0	180.4	185.1	1435.8	2774.5	511.5
Minnesota	-1.55434	1.05644	2.7	19.5	85.9	85.8	1134.7	2559.3	343.1
Delaware	0.96458	1.29674	6.0	24.9	157.0	194.2	1682.6	3678.4	467.0
Connecticut	-0.54133	1.50123	4.2	16.8	129.5	131.8	1346.0	2620.7	593.2
Hawaii	0.82313	1.82392	7.2	25.5	128.0	64.1	1911.5	3920.4	489.4
RhodeIsland	-0.20156	2.14658	3.6	10.5	86.5	201.0	1489.5	2844.1	791.4
Masssachusetts	0.97844	2.63105	3.1	20.8	169.1	231.6	1532.2	2311.3	1140.1

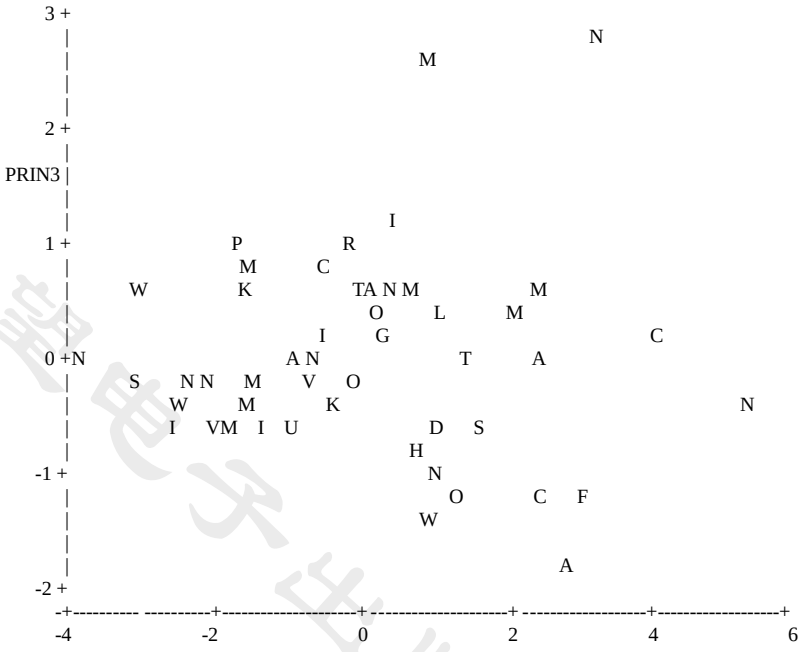
报表 34.2c　犯罪率的分析 — 第一与第二主成份，第一与第三主成份的坐标图



NOTE: 2 obs hidden.

Crime Rates per 100,000 Population by State
Plot of the First and Third Principal Components

Plot of PRIN3*PRIN1. Symbol is value of STATE.



NOTE: 1 obs hidden.

第 35 章　因子分析：统计程序 PROC FACTOR

35.1　因子分析法中的“因子”一词指什么

许多人对因子分析法中所指的 “因子” 一词不甚了解，本节特就此说明之。因子分析法中提到两种因子：共同因子 (又称共因子) 和独特因子。 这两种因子都是指一个(或一组) 假设的、抽象的变量。

所谓共同因子，指一个假设的、抽象的变量，它可用来解释两个或两个以上的原始变量。然而独特因子则指一个假设的、抽象的变量，它只能用来解释一个原始的变量，与其它变量完全无关。

如上所述，因子指假设的、抽象的变量。它的功能在于诠释原始变量之间的关系或结构。然而主成份是指原始变量间的线性组合，它的功能在于简化原有的变量群。

35.2　共因子分析法的模型

共因子分析法的模型允许每一变量有一独特因子，所以：

$Y_{ij} = X_{i1}b_{1j} + X_{i2}b_{2j} +...+ X_{iq}b_{qj} + E_{ij}$

其中，Y_{ij} = 第 i 个观察体在第 j 个变量上的值

X_{ik}= 第 i 个观察体在第 k 个共因子上的值

b_{kj}= 被第 k 个共因子用来预测第 j 个变量的回归系数，又称因子负荷量 (FactorLoading)

Ei_j= 第 i 个观察体在第 j 个独特因子上的值

q = 共同因子的总数

这个模型的两项假设如下：

- 独特因子之间是互相独立的
- 共因子与独特因子之间是互相独立的

35.3　PROC FACTOR 程序概述

■ 因子分析及坐标的转换

PROC FACTOR 可以对输入资料文件执行许多种不同的共因子分析及主成份分析，也可将分析的结果经过坐标的转换，以利于诠释。

■ 输入资料文件

PROC FACTOR 的输入资料文件可以是多变量数据、一个相关系数矩阵、一个变异数 / 共变异数矩阵、因子型态 (Factor Pattern)，或是一个因子分数系数 (Factor Score Coefficient) 的矩阵。FACTOR 程序也接受其它程序的输出资料文件，所以输入资料文件变化很多，详情见本章的第 35.6 节。

■ 因子提炼的方法

FACTOR 程序提供九种因子提炼的方法供读者选用，这九种方法将在介绍选项 METHOD= 中详加解释。另外，FACTOR 程序也提供了六种方法来预估变量间的共通性，见选项 PRIORS= 的说明。

■ 因子坐标的转换

一般而言，因子坐标的转换可分正交及斜交两大类。FACTOR 程序提供了八种坐标转换的方法供读者选择，见选项 ROTATE= 的说明。

■ 输出资料文件

FACTOR 程序所产生的输出资料文件不止一个，它们分别在选项 OUTSTAT= 中逐一说明。

35.4　因子分析法的历史背景

共因子分析由史氏 (Spearman) 于 1904 年首创。读者可参阅古德氏 (Gould，1981) 及金氏与穆勒氏 (Kim and Mueller，1978) 的书籍以便对分析法的目的及模型有初步的认识。较深入的讨论可参看慕雷克 (Mulaik，1972) 与哈门 (Harman，1976) 的着作。

35.5　如何撰写 PROC FACTOR 程序

PROC FACTOR 含七道指令，它们的格式如下：

```
PROC FACTOR      选项串;
     PRIORS      变量共通性的预估值;
     VAR         变量名称串;
     PARTIAL     变量名称串;
     FREQ        变量名称;
     WEIGHT      变量名称;
     BY          变量名称串;
```

通常读者只须用到 PROC FACTOR 及 VAR 两道指令。

指令 #1　PROC FACTOR 选项串；

PROC FACTOR 的选项可分下列五大类讨论：第一类选项与资料文件的界定有关，第二类与因子提炼有关，第三类与因子坐标的转换有关，第四类选项控制报表的印出，第五类含其它选项。

第一类选项　下列四选项与资料文件的界定有关：

(1) DATA=输入资料文件名称

为输入资料文件命名。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，对它执行因子分析。

(2) TARGET=资料文件名称

这一个资料文件内含有 Procrustes 坐标转换后理想的值，必须与 ROTATE=PROCRUSTES 选项并用。

(3) OUT=输出资料文件名称

这一个输出资料文件包括原输入资料文件的观察值以及因子分数 (Factor Score)。这些值以 FACTOR1，FACTOR2 等表示，读者必须同时用 NFACTOR=选项界定因子个数上限。

(4) OUTSTAT=第二个输出资料文件名称

这一个输出资料文件较上述 OUT=输出资料文件详尽。下页的表是 OUTSTAT 文件所含因子分析的各项统计值之代号及它们的定义，有些概念会在后面的章节中进一步解释。

代号 (_TYPE_=)	定 义
MEAN	变量的平均数
STD	变量的标准差
N	观察体的总数
CORR	相关系数矩阵，矩阵内各横列的变量名字以 _NAME_ 表示。
IMAGE	映象系数矩阵 (Image Coefficient Matrix)，矩阵内各横列的变量名字以 _NAME_ 表示。
IMAGECOV	映象的共变异数矩阵 (Image Covariance Matrix)，矩阵内各横列的变量名字以 _NAME_ 表示。
COMMUNAL	各变量共通性的最终估计值
PRIORS	各变量共通性的预估值
WEIGHT	变量的加权值
EIGENVAL	特性根
UNROTATE	未经坐标转换的因子系数型态
RESIDUAL	独特因子的相关系数矩阵
TRANSFOR	坐标转换矩阵
FCORR	共因子间的相关系数矩阵
PATTERN	因子系数的型态
RCORR	坐标轴间的相关系数矩阵

REFERENC	参考结构矩阵 (Reference Structure Matrix)
STRUCTUR	因子结构矩阵 (Factor Structure Matrix)
SCORE	共因子分数的系数 (可输入 SCORE 程序以便产生共因子分数。见第 12 章的例一)
USCORE	未经平均数矫正过的共因子分数的系数

第二类选项　下列十一个选项与因子提炼有关：

(1) METHOD=因子提炼的方法 (简写为 M=)

一般而言，此选项的内设值是 MEHTOD=PRINCIPAL。但当输入资料文件是 TYPE=FACTOR 的情况下，内设值是 METHOD=PATTERN。下列九种因子提炼的方法可供读者选用：

M=PRINCIPAL (或 PRIN 或 P)	此选项的因子提炼方法视选项 PRIORS= 而定。当此选项不与 PRIORS= 并用，或与 PRIORS=ONE 并用时，它的因子提炼方法是主成份分析法。否则它的因子提炼法是主轴因子分析法 (Principal Axis Common Factor Analysis)。
M=PRINIT	界定循环式主轴因子分析 (Iterative Principal Axis Method)。
M=ULS (或 U)	界定未加权的最小误差平方之因子分析 (Unweighted Least Squares Method)。
M=ALPHA (或 A)	界定阿尔法因子分析 (Alpha Factor Analysis)。
M=ML (或 M)	界定最大可能率因子分析，此法要求一个满秩的相关系数矩阵。
M=HARRIS (或 H)	界定哈里斯氏 (Harris) 于 1962 年提出的 $S^{-1}RS^{-1}$ 主轴分析。此处，S 是变量的变异数 / 共变异数矩阵，R 是变量间的相关系数矩阵。此法要求一个满秩的相关系数矩阵。
M=IMAGE (或 I)	针对映象共变异数矩阵作主成份分析 (Principal Component Analysis of Image Covariance Matrix)。此法要求一个满秩的相关系数矩阵。请读者注意比法与凯斯 (Kaiser，1963；1970；1974) 所提的映象分析 (Image Analysis) 无关。
M=PATTERN	从输入资料文件 (其 TYPE=FACTOR，CORR，或 COV) 内取得因子负荷量矩阵。若因子之间有线性相关，则其间的相关系数也必须同时输入 (TYPE='FCORR' 的数据)。
M=SCORE	从输入资料文件 (其 TYPE=FACTOR，CORR，或 COV) 内取得因子分数的系数。这个输入资料文件必须同时包括变量间的相关系数或其变异数 / 共变异数矩阵。

(2) PRIORS=变量共通性的预估值

PRIORS=ONE (或 O)	设定所有共通性的预估值为 1。
PRIORS=MAX (或 M)	取每一变量与其它变量的最大相关系数绝对值为该变量共通性的预估值。
PRIORS=SMC (或 S)	取每一变量与其它变量的复相关平方值为该变量共通性的预估值。
PRIORS=ASMC(或 A)	将上述的复相关 (SMC) 调整，使其总和等于最大相关系数绝对值的总和。而共通性预估值将与此值成正比 (Cureton，1968)。
PRIORS=INPUT(或 I)	如果输入资料文件的 TYPE=FACTOR，则读者可选用此选项。SAS 会进入资料文件内寻找 _TYPE_='PRIORS' 或_TYPE_='COMMUNAL'的变量，此变量的第一个观察值就成为共通性的预估值。
PRIORS=RANDOM(或 R)	随机取 0 与 1 之间的任何值为共通性的预估值。

下表列出因子提炼方法与共通性预估值的内设值之配对：

因子提炼的方法	共通性预测值的内设值
METHOD=	PRIORS=
PRINCIPAL	ONE
PRINIT	ONE
ALPHA	SMC
ULS	SMC
ML	SMC
HARRIS	(不适用)
IMAGE	(不适用)
PATTERN	(不适用)
SCORE	(不适用)

(3) RANDOM=正整数

起始随机随机数表的起始值，与选项 (2) PRIORS=RANDOM 联用。

(4) MAXITER=正整数

界定 METHOD=PRINIT， ULS， ALPHA 或 ML 等因子分析法中循环分析的次数。内设值是 30。

(5) CONVERGE (或 CONV)= 正实数

界定 METHOD=PRINIT， ULS， ALPHA 或 ML 等因子分析法中循环分析的收敛值。它的定义是两次循环所求得变量之共通性的差距。当这个差距小于此选项所定的值时，循环分析停止。内设值是 .001。

(6) COVARIANCE (或 COV)

要求 FACTOR 程序对变异数／共变异数矩阵 (而非相关系数矩阵) 执行因子分析。此选项必须与 METHOD=PRINCIPAL， PRINIT， ULS 或 IMAGE 适用。

(7) WEIGHT

要求 FACTOR 程序对一个经过加权调整的相关系数矩阵或变异数 / 共变异数矩阵执行因子分析。选用此项时，必须同时满足下列的条件：

●METHOD=PRINCIPAL， PRINIT， ULS 或 IMAGE。

●输入资料文件的 TYPE=CORR， COV， UCORR， UCOV 或 FACTOR。

●各变量的加权值由 _TYPE_='WEIGHT' 提供。

下面三个选项都可用来决定因子的总数。如果读者在下面三选项中同时选用两个或三个选项，则 SAS 会自动挑选最小的值。

(8) NFACTORS (或 NFACT，或 N)=正整数

界定因子个数的上限，内设值是所有被分析变量的总个数。

(9) PROPORTION (或 PERCENT，或 P)=百分比(正实数不带 % 符号)

界定一个共因子至少要能解释的变量之变异数百分比。内设值是 1 (即百分之百)。此选项不可与 METHOD=PATTERN 或 SCORE 合用。

(10) MINEIGEN (或 MIN)=最小特性根的值

要求 SAS 保留特性根大于此选项所设定的那些因子。此选项不可与 METHOD=PATTERN 或 SCORE 合用。一般而言，其内设值是 0，若读者对未加权过的相关系数矩阵进行因子分析，则其内设值等于 1。但如果读者同时省略 NFACTORS=、PROPORTION= 及 MINEIGEN= 三选项时，SAS 会依下面的原则，自行设定 MINEIGEN 的值。

当 METHOD=	则 MINEIGEN 的值为：
ALPHA 或 HARRIS	1
IMAGE	映象的总变异数 (Total Image Variance) 变量的总个数
其它 的方法，而且	经过加权调整的总变异数
PRIORS=1	变量的总个数

一般而言，当共通性的预估值超过 1 时， METHOD=PRINIT， ULS， ALPHA 和 ML 立刻停止分析的过程，并设因子的总个数为 0。下列两个选项可以让分析过程恢复：

(11) HEYWOOD (或 HEY)

将大于 1 的变量共通性重新调整为 1。如此，分析可以继续进行。

(12) ULTRAHEYWOOD (或 ULTRA)

改变规定，允许变量的共通性大于 1。此选项极可能导致不合理的分析结果，因此应慎重使用之。

第三类选项　下列六个选项与坐标转换有关：

(1) ROTATE (或 R)=坐标转换法，

有八种方法可供选择：

R=VARIMAX (或 V)　界定最大变异数转换法

R=QUARTIMAX (或 Q)　界定四次方最大值转换法

R=EQUAMAX (或 E)　界定平衡最大值坐标转换法

R=ORTHOMAX	界定标准正交转换法，其加权值来自选项 GAMMA=。
R=HK	界定哈雷斯-凯斯 (Harris-Kaiser) 坐标转换法。可与选项 HKPOWER= 合用，它界定特性根开方的次数，以便调整特性向量的元素。
R=PROMAX (或 P	界定最优斜交转换法，可与选项 PREROTATE= 及 POWER= 合用。若读者没有特别指明，则 PRER-OTATE=VARIMAX 而且 POWER=3。
R=PROCRUSTES	界定斜交转换法，与选项 TARGET= 合用。
R=NONE (或 N)	不执行任何坐标转换，是 R= 的内设值。

(2) GAMMA=正整数

决定标准正交转换的程度。此选项只能和 ROTATE=ORTHOMAX 或 PREROTATE=ORTHOMAX 合用。

(3) HKPOWER (或 HKP)=正实数

界定特性根开方的次数，开方过的特性根是用来调整特性向量的元素，与 ROTATE=HK 合用，也可以和 ROTATE=QUARTIMAX，VARIMAX，EQUA-MAX 或 ORTHOMAX 合用。此选项的值常介于 0 与 1 之间，内设值为 0。当此值等于 1 时，所得的结果就是最大变异数转换法的结果。

(4) POWER=正整数

界定 ROTATE=PROMAX 方法中所需的次方数，以达到理想的矩阵形态。内设值是 3。

(5) PREROTATE (或 PRE)=坐标转换法

只适合与 ROTATE=PROMAX 选项适用。它用来界定初步坐标转换的方法，内设值是 VARIMAX。此选项不可是 PROMAX 或 PROCRUSTES。当 METHOD=PATTERN 时，PREROTATE 必须是 NONE。

(6) NORM=KAISER (或 WEIGHT、或 COV、或 RAW、或 NONE)

为因子系数矩阵的列 (Row)界定一个标准化的方法。内设值是 KAISER。若 NORM=KAISER，则采凯斯氏标准化之法。若 NORM=WEIGHT，各列以 Cureton-Mulaik 法来标准化。若 NORM=COV，则各列将反映出变量与因子的共变异数 (而非相关系数)。若 NORM=NONE 或 RAW，则不进行任何标准化。

第四类选项　下面的十六个选项可控制报表印出：

(1) SIMPLE (或 S)

印出平均数与标准差。

(2) CORR (或 C)

印出相关系数或净相关系数的矩阵。

(3) MSA

印出每一对变量之间 (控制其它变量后) 的净相关矩阵，此矩阵又称为负反映象系数矩阵 (Negative Anti-Image Correlations Matrix)。且印出 Kaiser 的抽样合宜度 (Measure of Sampling Adequacy) (Kaiser，1970；Cerny and Kaiser，1977)。

(4) SCREE

将特性根由大到小排列后以图形显示，此图形称为 SCREE PLOT。

(5) EIGENVECTORS (或 EV)

印出特性向量。

(6) PRINT

印出输入资料文件中有关因子负荷量，因子分数系数的数据。此选项只适合与 METHOD=PATTERN 或 SCORE 联用。

(7) RESIDUALS (或 RES)

印出残差矩阵及其净相关系数矩阵。残差矩阵等于原始相关系数矩阵减去由因子型态导出的估计值矩阵。

(8) PREPLOT

印出尚未经过坐标转换的因子负荷量矩阵。

(9) PLOT

印出经过坐标转换后的因子负荷量矩阵。

(10) NPLOT=正整数 (如 4)

印出前几 (4) 个最重要的因子负荷量矩阵。最小值是 2 ，内设值等于所有因子的总个数。

(11) SCORE

印出因子分数系数，这些系数可再输入 PROC SCORE 以便计算各观察体因子分数。

(12) ALL

印出除 PLOT 及 NPLOT 以外其它 选项所产生的报表。

(13) REORDER (或 RE)

重新排列因子系数矩阵的列，使那些在第一因子上负荷量的绝对值高的变量排在前面几列，以协助解释因子的含意。

(14) ROUND

将相关系数及因子负荷量的值乘以 100 后，四舍五入成为整数。

(15) FLAG=正整数

与 ROUND 选项合用，目的是注明较重要的因子。所以因子负荷量大于 FLAG= 所订的值，将以星号标明之。内设值是矩阵内元素的标准差。

(16) FUZZ=正实数

凡相关系数或因子负荷量的绝对值小于 FUZZ= 所定的值，均会以句号 (表遗漏数据) 取代之，以简化矩阵。

第五类选项　其它选项：

(1) NOINT

不使用截距，故相关系数或共变异数不会对平均数作矫正。

(2) NOCORR

与 METHOD=PATTERN 或 SCORE 联用。阻止相关系数矩阵被纳入 OUTSTAT= 输出资料文件内。当资料文件含许多变量但很少因子时，此选项可以大大地减少

电脑的负荷。

(3) SINGULAR (或 SING)=正实数

界定一个矩阵不满秩的标准，内设值是 10 的 -8 次方。

(4) VARDEF=分母

界定变异数与共变异数计算时所用的分母，有四种选择：

N	：观察体的总个数。
DF	：上述 N 值减 i (未使用 PARTIAL 指令) 或 N-p-i (使用 PARTIAL 指令)。在此，i 等于 0 (如果界定 NOINT 选项) 或 1 (如果未界定 NOINT 选项)，p 等于 PARTIAL 指令中界定的变量个数。
WEIGHT (或 WGT)	：加权值之总和。
WDF	：WGT-i (未使用 PARTIAL 指令) 或 WGT-p-i (使用 PARTIAL 指令)。i，p 的定义同 DF。

VARDEF 的内设值是 DF。

指令 #2　PRIORS 变量共通性的预估值；

此预估值应介于 0 与 1 之间。预估值的数目应与 VAR 指令中所列变量的个数相对应。

请看如下的程序：

```
PROC FACTOR;
    VAR X Y Z ;
    PRIORS .7 .8 .9 ;
```

则 X 变量的共通性预估值是 .7，

Y 变量的共通性预估值是 .8，

Z 变量的共通性预估值是 .9。

指令 #3　VAR 变量名称串；

列举所有参与因子分析的变量名称。若省略此指令，则本程序内其它指令中未曾提到的所有数值变量均将被纳入因子分析内。

指令 #4　PARTIAL 变量名称串；

此指令指定一组变量，其值将从其余的变量中净化出来。如此所得的值构成净相关系数矩阵，而非相关系数矩阵。

指令 #5　FREQ 变量名称；

此变量的值代表资料文件内观察体重复出现的次数。

指令 #6　WEIGHT 变量名称；

当输入资料文件内各观察体的变异数不等时，读者可用这些不等的变异数之倒数，来指派不同的加权值，以区分各观察体的重要性。WEIGHT 变量的值就代表这些加权值。

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 (如：男/女，年龄组等)。然后对每一个小的资料文件单独进行因子分析。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。

35.6　五种合乎语法的输入资料文件形式

下面的例子示范几种输入资料文件的活泼形式：

例 1

```
PROC CORR DATA=RAW OUT=CORREL;     * 产生 TYPE=CORR 的资料文件;
PROC FACTOR DATA=CORREL METHOD=ML; * 最大可能率因子分析法;
PROC FACTOR DATA=CORREL;            * 主成份分析;
```

例 2

```
PROC FACTOR DATA=RAW OUTSTAT=FACT; * 主成份分析;
PROC FACTOR ROTATE=VARIMAX;        * VARIMAX 坐标转换法;
PROC FACTOR ROTATE=QUARTIMAX;      * QUARTIMAX 坐标转换法;
```

例 3

```
DATA CORREL (TYPE=CORR);
    _TYPE_='CORR';
    INPUT _NAME_ $ X Y Z ;
    CARDS;
X  1.0   .    .
Y   .7  1.0   .
Z   .5   .4  1.0
;
PROC FACTOR;
```

注：相关系数矩阵经 ML 法分析后，其样本数都会设定为 100,000。读者必须参考 Harman 等书籍内的公式将正确的样本数套入 χ^2 检定，否则报表上的 χ^2 值与其显著度却成为无意义的。

例 4

```
DATA PAT (TYPE=FACTOR);
```

```
        _TYPE_='PATTERN';
        INPUT _NAME_ $ X Y Z ;
        CARDS;
FACTOR1 .5 .7 .3
FACTOR2 .8 .2 .8
;
PROC FACTOR ROTATE=PROMAX PREROTATE=NONE;
```

例 5

```
DATA PAT (TYPE=FACTOR);
    INPUT _TYPE_ $ _NAME_ $ X Y Z ;
    CARDS;
PATTERN FACTOR1   .5   .7  .3
PATTERN FACTOR2   .8   .2  .8
FCORR   FACTOR1  1.0   .2  .
FCORR   FACTOR2   .2  1.0  .
;
PROC FACTOR ROTATE=PROMAX PREROTATE=NONE;
```

35.7　范　　例

下面三个因子分析的示范，都是用同一套输入资料文件 (SOCECON)。这一组资料代表美国洛杉矶市十二个社区的社会、经济情况 (以五个变量表示)。这五个社经变量是人口 (POP)，区民教育程度 (SCHOOL)，雇佣情形 (EMPLOY)，社会服务机构 (SERVICES) 及房价 (HOUSE)。

例一：主成份分析(Principal Component Analysis)

资料文件输入后，首先以主成份法加以分析 (因无 PRIORS= 选项，故 METHOD= 的内设值是主成份分析)。分析的结果并没有经过坐标转换。结果显示两个主成份因子似乎足以解释 93.4% 的总变异数。

程　序

```
DATA SOCECON;
     TITLE 'Five Socioeconomic Variables';
     TITLE2 'See Page 14 of Harman: Modern Factor Analysis,3rd ED';
     INPUT POP SCHOOL EMPLOY SERVICES HOUSE;
     CARDS;
5700 12.8 2500 270 25000
1000 10.9  600  10 10000
3400  8.8 1000  10  9000
```

```
3800 13.6 1700 140 25000
4000 12.8 1600 140 25000
8200  8.3 2600  60 12000
1200 11.4  400  10 16000
9100 11.5 3300  60 14000
9900 12.5 3400 180 18000
9600 13.7 3600 390 25000
9600  9.6 3300  80 12000
9400 11.4 4000 100 13000
;
PROC FACTOR DATA=SOCECON SIMPLE CORR;
     TITLE3 'Principal Component Analysis';
RUN;
```

结　果

主成份分析的结果找出两个主成份，它们对应的特性根均大于 1 (分别是 2.873314 与 1.796660)。这两个主成份可以解释的变异数百分比高达 93.40%。

当我们检视因子形态时，发现五个变量在第一主成份上的负荷量均是大于 0.50 的正值，在第二主成份的负荷量则有正有负。由于这个因子形态未经坐标轴的转换，我们只能说第一主成份是一个一般性的社经成份。然而，第二主成份则与社区经济较有关。这是因为 POP 与 EMPLOY 在第二主成份上的负荷量为正值，而区民教育程度与房价的负荷量是负值。

最后，五个变量中以人口数 (POP) 的因子共通性最高，达 98.7826%。

报表 35.1　主成份分析 (Principal Component Analysis)

Five Socioeconomic Variables

See Page 14 of Harman: Modern Fctor Analysis, 3rd ED

Principal Component Analysis

Means and Standard Deviations from 12 observations

	POP	SCHOOL	EMPLOY	SERVICES	HOUSE
Mean	6241.66667	11.4416667	2333.33333	120.833333	17000.0
Std Dev	3439.99427	1.78654483	1241.21153	114.927513	6367.53

Correlations

	POP	SCHOOL	EMPLOY	SERVICES	HOUSE

POP	1.00000	0.00975	0.97245	0.43887	0.02241
SCHOOL	0.00975	1.00000	0.15428	0.69141	0.86307
EMPLOY	0.97245	0.15428	1.00000	0.51472	0.12193
SERVICES	0.43887	0.69141	0.51472	1.00000	0.77765
HOUSE	0.02241	0.86307	0.12193	0.77765	1.00000

Initial Factor Method: Principal Components

Prior Communality Estimates: ONE

Eigenvalues of the Correlation Matrix: Total = 5 Average=1

	1	2	3	4	5
Eigenvalue	2.873314	1.796660	0.214837	0.099934	0.015255
Difference	1.076654	1.581823	0.114903	0.084679	
Proportion	0.5747	0.3593	0.0430	0.0200	0.0031
Cumulative	0.5747	0.9340	0.9770	0.9969	1.0000

2 factors will be retained by the MINEIGEN criterion.

Initial Factor Method: Principal Components

Factor Pattern

	FACTOR1	FACTOR2
POP	0.58096	0.80642
SCHOOL	0.76704	-0.54476
EMPLOY	0.67243	0.72605
SERVICES	0.93239	-0.10431
HOUSE	0.79116	-0.55818

Variance explained by each factor

FACTOR1	FACTOR2
2.873314	1.796660

Final Communality Estimates: Total = 4.669974

POP	SCHOOL	EMPLOY	SERVICES	HOUSE
0.987826	0.885106	0.979306	0.880236	0.937500

例二：主轴因子分析 (Principal Axis Common Factor Analysis)

对同一套输入资料文件执行主轴因子分析。因变量的共通性预估值是 SMC，故 METHOD 选项的内设值是主轴因子分析。因子的坐标转换采 PROMAX 法。这个例子要求重新排列因子负荷量矩阵的变量次序 (REORDER 选项)。最后将分析结果放入一个叫 FACT_ALL 的输出资料文件内。

程　序

```
PROC FACTOR DATA=SOCECON PRIORS=SMC MSA SCREE RESIDUAL PREPLOT
     ROTATE=PROMAX REORDER PLOT
     OUTSTAT=FACT_ALL;
     TITLE3 'Principal Factor Analysis Promax Rotation';
PROC PRINT;
     TITLE3 'Factor Output Data Set';
RUN;
```

结　果

分析的结果显示有两大类的因子 (或说有两个共因子，见报表 35.2)。一类因子以 HOUSE 与 SCHOOL 为主，另一类因子以 POP 与 EMPLOY 为主。这两类因子可解释成经济因子 (FACTOR 1) 与人口因子 (FACTOR 2)。这两类因子经过 VARIMAX 或 PROMAX 的坐标转换后显明出来。但变量 SERVICES 一直介于这两个因子之间无法被纳入某一个因子内。若读者想加强上述两个因子并且相对的减轻 SERVICES 在任一因子上的负荷量时，可试着采用 HK 转换法。此法所需输入资料文件可由例二的输出资料文件提供。唯一必须修饰的地方是删除 _TYPE_='PATTERN' 与_TYPE_='FCORR' 的观察体，并将 _TYPE_='UNROTATE' 改成 'PATTERN'。

请看下面的示范，并请读者试运行这个程序：

```
DATA FACT2(TYPE=FACTOR);
     SET;
    IF _TYPE_='PATTERN'|_TYPE_='FCORR' THEN DELETE;
    IF _TYPE_='UNROTATE' THEN _TYPE_='PATTERN';
PROC FACTOR ROTATE=HK NORM=WEIGHT REORDER PLOT;
   TITLE3 'Harris-Kaiser Rotation with Cureton-Mulaik Weights';
RUN;
```

报表 35.2　主轴因子分析 (Principal Axis Common Factor Analysis)

Five Socioeconomic Variables

See Page 14 of Harman: Modern Fctor Analysis, 3rd ED

Principal Component Analysis

Principal Factor Analysis Promax Rotation

Initial Factor Method: Principal Factors

Partial Correlations Controlling all other Variables

	POP	SCHOOL	EMPLOY	SERVICES	HOUSE
POP	1.00000	-0.54465	0.97083	0.09612	0.15871
SCHOOL	-0.54465	1.00000	0.54373	0.04996	0.64717
EMPLOY	0.97083	0.54373	1.00000	0.06689	-0.25572
SERVICES	0.09612	0.04996	0.06689	1.00000	0.59415
HOUSE	0.15871	0.64717	-0.25572	0.59415	1.00000

Kaiser's Measure of Sampling Adequacy: Over-all MSA = 0.57536759

POP	SCHOOL	EMPLOY	SERVICES	HOUSE
0.472079	0.551588	0.488511	0.806644	0.612814

Prior Communality Estimates: SMC

POP	SCHOOL	EMPLOY	SERVICES	HOUSE
0.968592	0.822285	0.969181	0.785724	0.847019

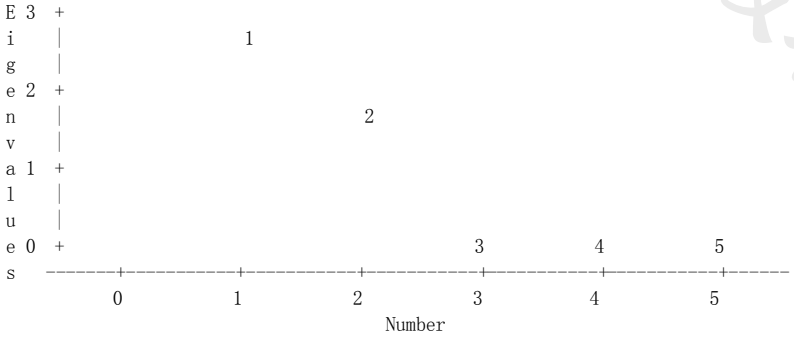
Eigenvalues of the Reduced Correlation Matrix:

Total = 4.39280116　Average = 0.87856023

	1	2	3	4	5
Eigenvalue	2.734301	1.716069	0.039563	-0.024523	-0.072608
Difference	1.018232	1.676506	0.064086	0.048084	
Proportion	0.6225	0.3907	0.0090	-0.0056	-0.0165
Cumulative	0.6225	1.0131	1.0221	1.0165	1.0000

2 factors will be retained by the PROPORTION criterion.

Scree Plot of Eigenvalues



Factor Pattern

	FACTOR1	FACTOR2
SERVICES	0.87899	-0.15847
HOUSE	0.74215	-0.57806
EMPLOY	0.71447	0.67936

Variance explained by each factor

FACTOR1	FACTOR2
2.734301	1.716069

Final Communality Estimates: Total = 4.450370

SCHOOL	0.71370	-0.55515	POP	SCHOOL	EMPLOY	SERVICES	HOUSE
POP	0.62533	0.76621	0.978113	0.817564	0.971999	0.797743	0.884950

Residual Correlations With Uniqueness on the Diagonal

	POP	SCHOOL	EMPLOY	SERVICES	HOUSE
POP	0.02189	-0.01118	0.00514	0.01063	0.00124
SCHOOL	-0.01118	0.18244	0.02151	-0.02390	0.01248
EMPLOY	0.00514	0.02151	0.02800	-0.00565	-0.01561
SERVICES	0.01063	-0.02390	-0.00565	0.20226	0.03370
HOUSE	0.00124	0.01248	-0.01561	0.03370	0.11505

Root Mean Square Off-diagonal Residuals: Over-all = 0.01693282

POP	SCHOOL	EMPLOY	SERVICES	HOUSE
0.008153	0.018130	0.013828	0.021517	0.019602

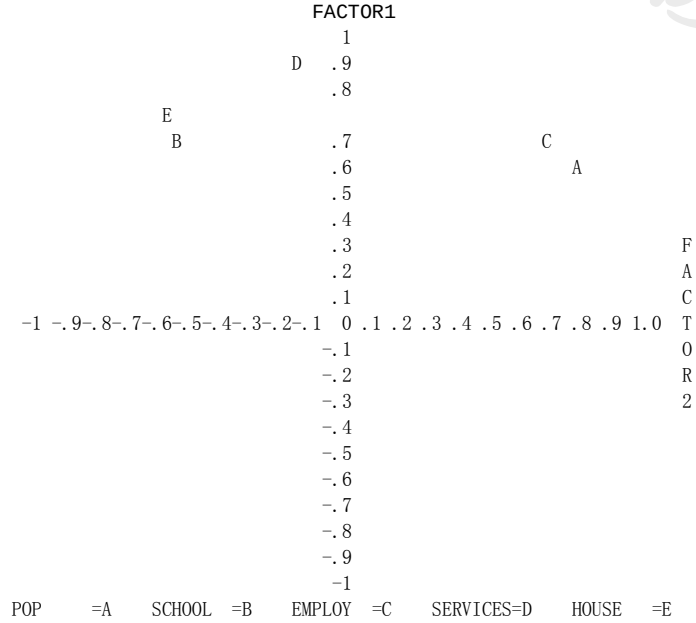
Partial Correlations Controlling Factors

	POP	SCHOOL	EMPLOY	SERVICES	HOUSE
POP	1.00000	-0.17693	0.20752	0.15975	0.02471
SCHOOL	-0.17693	1.00000	0.30097	-0.12443	0.08614
EMPLOY	0.20752	0.30097	1.00000	-0.07504	-0.27509
SERVICES	0.15975	-0.12443	-0.07504	1.00000	0.22093
HOUSE	0.02471	0.08614	-0.27509	0.22093	1.00000

Root Mean Square Off-diagonal Partials: Over-all = 0.18550132

POP	SCHOOL	EMPLOY	SERVICES	HOUSE
0.158508	0.190259	0.231818	0.154470	0.182015

Plot of Factor Pattern for FACTOR1 and FACTOR2



Prerotation Method: Varimax

Orthogonal Transformation Matrix

	1	2
1	0.78895	0.61446
2	-0.61446	0.78895

Rotated Factor Pattern

	FACTOR1	FACTOR2
HOUSE	0.94072	-0.00004
SCHOOL	0.90419	0.00055
SERVICES	0.79085	0.41509
POP	0.02255	0.98874
EMPLOY	0.14625	0.97499

Variance explained by each factor

FACTOR1	FACTOR2
2.349857	2.100513

Final Communality Estimates: Total = 4.450370

POP	SCHOOL	EMPLOY	SERVICES	HOUSE
0.978113	0.817564	0.971999	0.797743	0.884950

Plot of Factor Pattern for FACTOR1 and FACTOR2

```
                                FACTOR1
                                  1
                                  E
                                 .B
                                 .8          D
                                 .7
                                 .6
                                 .5
                                 .4
                                 .3                        C  F
                                 .2                           A
                                 .1                           C
-1 -.9-.8-.7-.6-.5-.4-.3-.2-.1  0 .1 .2 .3 .4 .5 .6 .7 .8 .9 A.0T
                                -.1                           O
                                -.2                           R
                                -.3                           2
                                -.4
                                -.5
                                -.6
                                -.7
                                -.8
                                -.9
                                 -1
```

POP =A　SCHOOL =B　EMPLOY =C　SERVICES=D　HOUSE =E

Rotation Method: Promax

Target Matrix for Procrustean Transformation

	FACTOR1	FACTOR2
HOUSE	1.00000	-0.00000
SCHOOL	1.00000	0.00000
SERVICES	0.69421	0.10045
POP	0.00001	1.00000

Procrustean Transformation Matrix

	1	2
1	1.04117	-0.09865
2	-0.10572	0.96303

Normalized Oblique Transformation Matrix

	1	2

EMPLOY	0.00326	0.96793

1	0.73803	0.54202
2	-0.70555	0.86528

Inter-factor Correlations

	FACTOR1	FACTOR2
FACTOR1	1.00000	0.20188
FACTOR2	0.20188	1.00000

Reference Axis Correlations

	FACTOR1	FACTOR2
FACTOR1	1.00000	-0.20188
FACTOR2	-0.20188	1.00000

Rotated Factor Pattern (Std Reg Coefs)

	FACTOR1	FACTOR2
HOUSE	0.95558	-0.09792
SCHOOL	0.91842	-0.09352
SERVICES	0.76053	0.33932
POP	-0.07908	1.00192
EMPLOY	0.04799	0.97509

Reference Structure (Semipartial Correlations)

	FACTOR1	FACTOR2
HOUSE	0.93591	-0.09590
SCHOOL	0.89951	-0.09160
SERVICES	0.74487	0.33233
POP	-0.07745	0.98129
EMPLOY	0.04700	0.95501

Variance explained by each factor eliminating other factors

FACTOR1	FACTOR2
2.248089	2.003020

Factor Structure (Correlations)

	FACTOR1	FACTOR2
HOUSE	0.93582	0.09500
SCHOOL	0.89954	0.09189
SERVICES	0.82903	0.49286
POP	0.12319	0.98596
EMPLOY	0.24484	0.98478

Variance explained by each factor ignoring other factors

FACTOR1	FACTOR2
2.447349	2.202280

Final Communality Estimates: Total = 4.450370

POP	SCHOOL	EMPLOY	SERVICES	HOUSE
0.978113	0.817564	0.971999	0.797743	0.884950

Plot of Reference Structure for FACTOR1 and FACTOR2

Reference Axis Correlation = -0.2019 Angle = 101.6471

```
                                FACTOR1
                                   1
                                E
                                B .9
                                  .8
                                  .7         D
                                  .6
                                  .5
                                  .4
                                  .3                              F
                                  .2                              A
                                  .1                          C   C
   -1 -.9-.8-.7-.6-.5-.4-.3-.2-.1  0 .1 .2 .3 .4 .5 .6 .7 .8 .9 1.0 T
                                 -.1                              O
                                 -.2                          A   R
```

−.3　　2
−.4
−.5
−.6
−.7
−.8
−.9
−1

POP　=A　SCHOOL　=B　EMPLOY　=C　SERVICES=D　HOUSE　=E

Five Socioeconomic Variables

See Page 14 of Harman: Modern Fctor Analysis, 3rd ED

Factor Output Data Set

OBS	_TYPE_	_NAME_	POP	SCHOOL	EMPLOY	SERVICES	HOUSE
1	MEAN		6241.67	11.4417	2333.33	120.833	17000.00
2	STD		3439.99	1.7865	1241.21	114.928	6367.53
3	N		12.00	12.0000	12.00	12.000	12.00
4	CORR	POP	1.00	0.0098	0.97	0.439	0.02
5	CORR	SCHOOL	0.01	1.0000	0.15	0.691	0.86
6	CORR	EMPLOY	0.97	0.1543	1.00	0.515	0.12
7	CORR	SERVICES	0.44	0.6914	0.51	1.000	0.78
8	CORR	HOUSE	0.02	0.8631	0.12	0.778	1.00
9	COMMUNAL		0.98	0.8176	0.97	0.798	0.88
10	PRIORS		0.97	0.8223	0.97	0.786	0.85
11	EIGENVAL		2.73	1.7161	0.04	-0.025	-0.07
12	UNROTATE	FACTOR1	0.63	0.7137	0.71	0.879	0.74
13	UNROTATE	FACTOR2	0.77	-0.5552	0.68	-0.158	-0.58
14	RESIDUAL	POP	0.02	-0.0112	0.01	0.011	0.00
15	RESIDUAL	SCHOOL	-0.01	0.1824	0.02	-0.024	0.01
16	RESIDUAL	EMPLOY	0.00514	0.02151	0.02800	-0.00565	-0.01561
17	RESIDUAL	SERVICES	0.01063	-0.02390	-0.00565	0.20226	0.03370
18	RESIDUAL	HOUSE	0.00124	0.01248	-0.01561	0.03370	0.11505
19	PRETRANS	FACTOR1	0.78895	-0.61446	.	.	.
20	PRETRANS	FACTOR2	0.61446	0.78895	.	.	.
21	PREROTAT	FACTOR1	0.02255	0.90419	0.14625	0.79085	0.94072
22	PREROTAT	FACTOR2	0.98874	0.00055	0.97499	0.41509	-0.00004
23	TRANSFOR	FACTOR1	0.73803	-0.70555	.	.	.
24	TRANSFOR	FACTOR2	0.54202	0.86528	.	.	.
25	FCORR	FACTOR1	1.00000	0.20188	.	.	.
26	FCORR	FACTOR2	0.20188	1.00000	.	.	.
27	PATTERN	FACTOR1	-0.07908	0.91842	0.04799	0.76053	0.95558
28	PATTERN	FACTOR2	1.00192	-0.09352	0.97509	0.33932	-0.09792
29	RCORR	FACTOR1	1.00000	-0.20188	.	.	.
30	RCORR	FACTOR2	-0.20188	1.00000	.	.	.
31	REFERENC	FACTOR1	-0.07745	0.89951	0.04700	0.74487	0.93591
32	REFERENC	FACTOR2	0.98129	-0.09160	0.95501	0.33233	-0.09590

33	STRUCTUR	FACTOR1	0.12319	0.89954	0.24484	0.82903	0.93582
34	STRUCTUR	FACTOR2	0.98596	0.09189	0.98478	0.49286	0.09500

例三：最大可能率的因子分析 (Maximum Likelihood Factor Analysis)

对 SOCECON 资料文件执行三次最大可能率因子分析 (METHOD=ML)。分析过程包括单因子 (N=1)、双因子 (N=2) 及三因子 (N=3) 的解法。每一解法均以 Chi-Square 检定检验统计的显著度。

最后，我们还是觉得双因子的结论最合适。三因子的解法太多余，而单因子解法并不足以解释许多变量之间的关系。

程　序

```
PROC FACTOR DATA=SOCECON METHOD=ML HEYWOOD N=1;
     TITLE3 'Maximun-Likelihood Factor Analysis with One Factor';
PROC FACTOR DATA=SOCECON METHOD=ML HEYWOOD N=2;
     TITLE3 'Maximun-Likelihood Factor Analysis with Two Factors';
PROC FACTOR DATA=SOCECON METHOD=ML HEYWOOD N=3;
     TITLE3 'Maximun-Likelihood Factor Analysis with Three Factors';
RUN;
```

结　果

所以，经过这三次因子分析检定，我们一致的结论是双因子的解法最适宜。

报表 35.3　最大可能率因子分析 (Maximum Likelihood Factor Analysis)

Five Socioeconomic Variables

See Page 14 of Harman: Modern Fctor Analysis, 3rd ED

Maximun-Likelihood Factor Analysis with One Factor (单因子的解法)

Initial Factor Method: Maximum Likelihood

Prior Communality Estimates: SMC

POP	SCHOOL	EMPLOY	SERVICES	HOUSE
0.968592	0.822285	0.969181	0.785724	0.847019

Preliminary Eigenvalues: Total = 76.1165859 Average = 15.223317

	1	2	3	4	5
Eigenvalue	63.701009	13.054719	0.327639	-0.347281	-0.619501
Difference	50.646289	12.727080	0.674920	0.272220	
Proportion	0.8369	0.1715	0.0043	-0.0046	-0.0081
Cumulative	0.8369	1.0084	1.0127	1.0081	1.0000

1 factors will be retained by the NFACTOR criterion.

Iter	Criterion	Ridge	Change	Communalities

```
1     6.54292     0.000   0.10330    0.93828 0.72227 1.00000 0.71940 0.74371
2     3.12327     0.000   0.72885    0.94566 0.02380 1.00000 0.26493 0.01487
                    Convergence criterion satisfied.
               Significance tests based on 12 observations:
Test of H0: No common factors.                    Test of H0: 1 Factors are sufficient.
     vs HA: At least one common factor.                 vs HA: More factors are needed.
Chi-square = 54.252    df = 10    Prob>chi**2 = 0.0000      Chi-square = 24.466    df = 5    Prob>chi**2 = 0.0002
          Chi-square without Bartlett's correction = 34.35596878
          Akaike's Information Criterion = 24.35596878
          Schwarz's Bayesian Criterion = 21.931435531
          Tucker and Lewis's Reliability Coefficient = 0.120231384
                    Squared Canonical Correlations
                                FACTOR1
                               1.000000
       Eigenvalues of the Weighted Reduced Correlation Matrix:
                          Total = 0  Average = 0
                          1          2          3          4          5
     Eigenvalue      .          1.927160 -0.228313 -0.792956 -0.905891
     Difference      .          2.155473  0.564643  0.112935
         Factor Pattern
                  FACTOR1              Variance explained by each factor
       POP        0.97245                                FACTOR1
       SCHOOL     0.15428                    Weighted    17.801063
       EMPLOY     1.00000                    Unweighted   2.249260
       SERVICES   0.51472
       HOUSE      0.12193
            Final Communality Estimates and Variable Weights
    Total Communality: Weighted = 17.801063   Unweighted = 2.249260
                        POP     SCHOOL     EMPLOY   SERVICES      HOUSE
     Communality     0.945656  0.023803   1.000000  0.264935   0.014866
     Weight         18.401165  1.024384   .         1.360424   1.015090

       Maximun-Likelihood Factor Analysis with Two Factors (双因子的解法)
                  Prior Communality Estimates: SMC
                  POP     SCHOOL     EMPLOY   SERVICES      HOUSE
             0.968592   0.822285   0.969181   0.785724   0.847019
   Preliminary Eigenvalues:  Total = 76.1165859  Average = 15.223317
```

	1	2	3	4	5
Eigenvalue	63.701009	13.054719	0.327639	-0.347281	-0.619501
Difference	50.646289	12.727080	0.674920	0.272220	
Proportion	0.8369	0.1715	0.0043	-0.0046	-0.0081
Cumulative	0.8369	1.0084	1.0127	1.0081	1.0000

2 factors will be retained by the NFACTOR criterion.

Iter	Criterion	Ridge	Change	Communalities				
1	0.34312	0.000	0.04710	1.00000	0.80672	0.95058	0.79348	0.89412
2	0.30722	0.000	0.03068	1.00000	0.80821	0.96023	0.81048	0.92480
3	0.30679	0.000	0.00629	1.00000	0.81149	0.95948	0.81677	0.92023
4	0.30674	0.000	0.00218	1.00000	0.80985	0.95963	0.81498	0.92241
5	0.30673	0.000	0.00071	1.00000	0.81019	0.95955	0.81569	0.92187

Convergence criterion satisfied.

Significance tests based on 12 observations:

Test of H0: No common factors.　　Test of H0: 2 Factors are sufficient.

vs HA: At least one common factor.　　vs HA: More factors are needed.

Chi-square = 54.252　df = 10　Prob>chi**2 = 0.0000　　Chi-square = 2.198　df = 1　Prob>chi**2 = 0.1382

Chi-square without Bartlett's correction = 3.3740529531

Akaike's Information Criterion = 1.3740529531

Schwarz's Bayesian Criterion = 0.8891463034

Tucker and Lewis's Reliability Coefficient = 0.7292200071

Squared Canonical Correlations

FACTOR1	FACTOR2
1.000000	0.951889

Eigenvalues of the Weighted Reduced Correlation Matrix:

Total = 19.7853157　Average = 4.94632893

	1	2	3	4	5
Eigenvalue	.	19.785314	0.543185	-0.039771	-0.503412
Difference	.	19.242129	0.582956	0.463641	
Proportion	.	1.0000	0.0275	-0.0020	-0.0254
Cumulative	.	1.0000	1.0275	1.0254	1.0000

Factor Pattern

	FACTOR1	FACTOR2
POP	1.00000	0.00000
SCHOOL	0.00975	0.90003
EMPLOY	0.97245	0.11797
SERVICES	0.43887	0.78930
HOUSE	0.02241	0.95989

Variance explained by each factor

	FACTOR1	FACTOR2
Weighted	24.432971	19.785314
Unweighted	2.138861	2.368353

Final Communality Estimates and Variable Weights

Total Communality: Weighted = 44.218285　Unweighted = 4.507214

	POP	SCHOOL	EMPLOY	SERVICES	HOUSE
Communality	1.000000	0.810145	0.959571	0.815603	0.921894
Weight	.	5.268294	24.724667	5.425646	12.799679

Maximun-Likelihood Factor Analysis with Three Factors (三因子的解法)

Prior Communality Estimates: SMC

POP	SCHOOL	EMPLOY	SERVICES	HOUSE
0.968592	0.822285	0.969181	0.785724	0.847019

Preliminary Eigenvalues: Total = 76.1165859 Average = 15.223317

	1	2	3	4	5
Eigenvalue	63.701009	13.054719	0.327639	-0.347281	-0.619501
Difference	50.646289	12.727080	0.674920	0.272220	
Proportion	0.8369	0.1715	0.0043	-0.0046	-0.0081
Cumulative	0.8369	1.0084	1.0127	1.0081	1.0000

3 factors will be retained by the NFACTOR criterion.

WARNING: Too many factors for a unique solution.

Iter	Criterion	Ridge	Change	Communalities				
1	0.17980	0.031	0.05014	0.96081	0.84184	1.00000	0.80175	0.89716
2	0.00164	0.031	0.06784	0.98081	0.88713	1.00000	0.79559	0.96500
3	4.13501E-6	0.031	0.00939	0.98195	0.88603	1.00000	0.80498	0.96751
4	2.53532E-8	0.031	0.00063	0.98202	0.88585	1.00000	0.80561	0.96735

Converged, but not to a proper optimum.

Try a different 'PRIORS' statement.

Significance tests based on 12 observations:

Test of H0: No common factors.
vs HA: At least one common factor.
Chi-square = 54.252　df = 10　Prob>chi**2 = 0.0000

Test of H0: 3 Factors are sufficient.
vs HA: More factors are needed.
Chi-square = 0.000　df = -2　Prob>chi**2 = .

Chi-square without Bartlett's correction = 2.7888512E-7

Akaike's Information Criterion = 4.0000002789

Schwarz's Bayesian Criterion = 4.9698135785

Tucker and Lewis's Reliability Coefficient = 0

Squared Canonical Correlations

FACTOR1	FACTOR2	FACTOR3
1.000000	0.975189	0.689446

Eigenvalues of the Weighted Reduced Correlation Matrix:

Total = 41.5254193　Average = 10.3813548

	1	2	3	4	5
Eigenvalue	.	39.305483	2.220057	0.000087	-0.000207
Difference	.	37.085426	2.219969	0.000295	
Proportion	.	0.9465	0.0535	0.0000	-0.0000
Cumulative	.	0.9465	1.0000	1.0000	1.0000

Factor Pattern

	FACTOR1	FACTOR2	FACTOR3
POP	0.97245	-0.11233	-0.15409
SCHOOL	0.15428	0.89108	0.26083
EMPLOY	1.00000	-0.00000	0.00000
SERVICES	0.51472	0.72416	-0.12766
HOUSE	0.12193	0.97227	-0.08473

Variance explained by each factor

	FACTOR1	FACTOR2	FACTOR3
Weighted	54.611524	39.305483	2.220057
Unweighted	2.249260	2.276344	0.115254

Final Communality Estimates and Variable Weights

Total Communality: Weighted = 96.137063　Unweighted = 4.640858

	POP	SCHOOL	EMPLOY	SERVICES	HOUSE
Communality	0.982017	0.885852	1.000000	0.805643	0.967347
Weight	55.606690	8.760719	.	5.144426	30.625108

第 36 章　典型相关分析：统计程序 PROC CANCORR

36.1　何谓典型相关

典型相关的功能在于分析两组变量间的关系。这两组变量的数目可以不止一个或不等。当两组都只含一个变量时，典型相关就是皮尔森相关；当一组含一个变量，另一组含多个变量时，典型相关就是复相关；当两组都含多个变量时，就是典型相关。所以皮尔森相关及复相关都是典型相关的特例。

36.2　PROC CANCORR 程序概述

CANCORR 程序可用来执行典型相关分析、典型净相关分析、及典型冗余分析 (Canonical Redundancy Analysis)。分析的结果包括 (未经过) 标准化的典型相关系数、所有典型变量与原始变量之间的相关系数，以及典型变量值。

为了了解 CANCORR 程序如何找出两组变量之间的关系，读者必须先熟悉三个名词：典型变量、典型相关，及典型系数。典型变量乃各组变量所形成的线性组合。典型相关指这两个线性组合 (即典型变量) 之间的相关。典型相关的程度以典型系数 (又称典型加权值) 表示。

分析时，CANCORR 程序从各组内找出具有最高典型相关的一对典型变量，称为第一典型变量。

然后 CANCORR 程序再找出具有次高典型相关的另一对典型变量，称为第二典型变量，如此重复多次。重复的次数等于含较少变量那一组的变量数目。CANCORR 程序假设这两组变量中至少有一组是从多元常态分配中随机选出的，因此利用 F 检定来检验典型相关是否为 0。

请读者注意：各对典型变量 (即：第一对典型变量，第二对典型变量...) 之间必须是独立无关的。另外，组内每一个典型变量只可和另一组内相对应的典型变量有相关，而必须和本组内及另一组内的其它典型变量独立无关。

CANCORR 程序也可用来执行净典型相关分析，有关其理论基础，请参阅 Cooley 与 Lohns (1971) 或 Timm (1975)。

36.3　如何撰写 PROC CANCORR 程序

PROC CANCORR 含七道指令，它们的格式如下：

PROC CANCORR	选项串；
VAR	变量名称串；
WITH	变量名称串；
PARTIAL	变量名称串；
FREQ	变量名称；
WEIGHT	变量名称；
BY	变量名称串；

研究者通常只需 PROC CANCORR，VAR，和 WITH 指令。请读者注意，不可省略 WITH 指令。

指令 #1　PROC CANCORR 选项串；

PROC CANCORR 的选项可分五大类来讨论：第一类选项与资料文件的界定有关，第二类选项可用来控制报表的打印，第三类选项界定报表上变量的命名与计算过程的有关事宜，第四类选项与回归分析有关，第五类选项与回归分析中产生的统计量有关。

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

为输入资料文件命名。这个资料文件可以包含原始变量的数据，也可以是一个相关系数矩阵(TYPE=CORR 或 UCORR)，或是一个变异数／共变异数矩阵(TYPE=COV 或 UCOV)，甚或 TYPE=SSCP 或 TYPE=FACTOR 的资料文件。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，对它执行典型相关分析。

(2) OUT=输出资料文件

这一个输出资料文件包括原输入资料文件的数据以及典型变量值。当输入资料文件是一个相关系数或是一个变异数／共变异数矩阵时，不可用此选项。

(3) OUTSTAT= 第二个输出资料文件

这一个输出资料文件包括典型相关分析的各式结果，它们的代号与定义如下表所示：

代　号 (_TYPE_)	定　义
MEAN	变量的平均数
STD	变量的标准差
N	观察体的总个数
SUMWGT	加权值的总和，是选用 WEIGHT 指令的结果
CORR	相关系数矩阵
CANCORR	典型相关系数
SCORE	标准化的典型系数
RAWSCORE	未经标准化的典型系数
STRUCTURE	典型结构
RSQUARED	复相关系数平方

ADJRSQ	矫正过后的复相关系数平方
UCLRSQ	复相关系数平方的 95% 信赖区间之上限
LCLRSQ	复相关系数平方的 95% 信赖区间之下限
F	F 检定的统计值
PROBF	上述 F 检定值的统计显著度
T	t 检定的统计值
PROBT	上述 t 检定值的统计显著度
CORRB	回归系数估计值间的相关系数
STB	标准化后的回归系数
B	未经标准化的回归系数
SEB	回归系数的标准误差
PCORR	净相关系数
SQPCORR	净相关系数的平方
SPCORR	半净相关系数
SQSPCORR	半净相关系数的平方

第二类选项　下列七个选项可用来控制报表的打印：

(1) SIMPLE (或 S)

印出平均数与标准差。

(2) CORR (或 C)

印出原始变量之间的相关系数矩阵。

(3) REDUNDANCY (或 RED)

印出典型冗余分析的统计值，其值可用来探讨原始变量被典型变量解释的百分比。

(4) ALL

印出所有的统计值。

(5) SHORT

只印出典型相关系数与其 F 检定的显著度。

(6) NOPRINT

不印出分析的结果。

(7) NCAN=正整数 (如 2)

印出前几对如(2)典型变量的所有统计值。

第三类选项　下列八个选项可用来界定报表上变量的命名与计算过程的有关事宜：

(1) EDF=正整数

若输入资料文件的数据是某一个回归分析的结果，则读者可用此选项界定 F 检定中分母的自由度，内设值是有效观察体的总数减 1。

(2) RDF=正整数

若输入资料文件的数据是某一个回归分析的结果，则读者可用此选项界定 F 检定中分子的自由度，内设值是原回归分析中自变量的数目。

(3) NOINT

规定在典型相关分析与回归分析的模型中不包括截距。

(4) SINGULAR (或 SING)=正小数 (P)

检查各矩阵是否为满秩的矩阵。此值必须小于 1，大于 0。内设值是 10 的 -8 次方。若某一变量与其它变量的复相关平方大于或等于 1-P，则此变量将自动从典型相关分析中剔出，此时这个变量的典型系数便等于 0。

(5) VPREFIX (或 VP)= 典型变量的名字

为 VAR 指令所导出的典型变量命名。若读者指定 VP=ABC，则第一、第二典型变量就是 ABC1，ABC2。内设值是 V1，V2....等。

(6) WPREFIX (或 WP)= 典型变量的名字

为 WITH 指令所导出的典型变量命名。若读者指定 WP=XYZ，则第一、第二典型变量就是 XYZ1，XYZ2。内设值是 W1，W2 ....等。

(7) VNAME (或 VN)='VAR 变量名称'

读者可用此选项在报表上为 VAR 指令中所列的变量串命名。字数限在四十个字母之内，且用单引号括之。

(8) WNAME (或 WN)='WITH 变量名称'

读者可用此选项在报表上为 WITH 指令中所列的变量串命名。同样地，字数限在四十个字母之内，用单引号括起来。

第四类选项　下列四个选项与回归分析有关：

(1) VDEP

(2) WREG

这一对选项要求将 VAR 的变量串当作因变量，将 WITH 变量串当作自变量，进行多变量的复回归分析 (Multivariate Multiple Regression)。

(3) WDEP

(4) VREG

与上述 (1)、(2) 的作法刚好相反，这一对选项要求将 WITH 的变量串当作因变量，将 VAR 变量串当作自变量，进行多变量的复回归分析。

第五类选项　下列十三个选项与回归分析中产生的统计量有关：

(1) ALL

要求报表印出所有的统计量并收集在 OUTSTAT= 输出资料文件内 (若你同时界定此选项)。然而若你又界定了 NOPRINT 的选项，则统计量不会在报表上印出来 ，它们只被归入 OUTSTAT= 的输出文件内。

(2) INT

要求回归分析的模型考虑截距，这个选项应与 B，SEB，T 或 PROBT 同时联用。

(3) B

要求计算且打印回归分析的系数。

(4) SEB

要求计算且打印回归系数的标准误差。

(5) T

针对每一个回归系数，执行 t 检定 (在此，t=B/SEB)。

(6) PROBT

界定上述 t 检定 (即选项 T) 的统计显著度。

(7) STB

要求计算且打印标准化的回归系数。

(8) SMC

要求打印回归分析后产生的复相关系数平方与 F 检定的结果。

(9) CORRB

打印出回归系数间的相关系数矩阵。

(10) PCORR

要求打印净相关系数。

(11) SPCORR

要求打印半净相关系数。

(12) SQPCORR

要求打印净相关系数的平方，亦即上述 (10) 的平方。

(13) SQSPCORR

要求打印半净相关系数的平方，亦即上述 (11) 的平方。

指令 #2　VAR 变量名称串；

典型相关分析的两组变量中的第一组变量名称串。若省略此指令，则在本程序内其它指令未曾提到的所有数值变量将构成第一组变量。

指令 #3　WITH 变量名称串；

读者一定要在 CANCORR 程序中使用这一道指令，不可省略。此指令列举典型相关分析的两组变量中的第二组变量名称串。

指令 #4　PARTIAL 变量名称串；

若读者想执行典型净相关分析，则可用此指令列出变量名称串。其值将由第一，第二组变量中净化出来。然后，SAS 再对净化过后的两组变量执行典型净相关分析。

指令 #5　FREQ 变量名称；

此变量的值代表资料文件中各观察体重复出现的次数。

指令 #6　WEIGHT 变量名称；

这个变量的值是正实数，代表观察体的比重或加权值。

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典型相关分析。 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。

36.4 范　　例

例一：健身俱乐部成员的分析

本例的输入资料文件 (FIT) 包括两组变量的数据。第一组变量包括三个生理变量即体重 (WEIGHT)、腰围 (WAIST) 与脉搏速度 (PULSE)。第二组变量包括三个运动变量即：拉单杠成绩 (CHINS)，仰卧起坐次数 (SITUPS) 及跳高成绩 (JUMPS)。这些数据由一家健身俱乐部的二十位中年男性会员所提供。

程　序

```
DATA FIT;

     INPUT WEIGHT WAIST PULSE CHINS SITUPS JUMPS ;
     CARDS;
191 36 50  5 162  60
189 37 52  2 110  60
193 38 58 12 101 101
162 35 62 12 105  37
189 35 46 13 155  58
182 36 56  4 101  42
211 38 56  8 101  38
167 34 60  6 125  40
176 31 74 15 200  40
154 33 56 17 251 250
169 34 50 17 120  38
166 33 52 13 210 115
154 34 64 14 215 105
247 46 50  1  50  50
193 36 46  6  70  31
202 37 62 12 210 120
176 37 54  4  60  25
157 32 52 11 230  80
156 33 54 15 225  73
138 33 68  2 110  43
;
PROC CANCORR DATA=FIT ALL
```

```
        VPREFIX=PHYS VNAME='Physiological Measurements'
        WPREFIX=EXER WNAME='Exercise';
        VAR WEIGHT WAIST PULSE;  WITH CHINS SITUPS JUMPS;
        TITLE 'Middle-Aged Men in a Health Fitness Club';
        TITLE2 'Data Courtesy of Dr.A.C. Linnerud,NC State Univ';
    RUN;
```

结　果

(1) 两组变量间一对一的皮尔森系数属于中等。最高的相关系数是介于腰围和仰卧起坐间 (-.6456)。各组内变量的相关较高。在生理变量组内，最高的相关系数是体重和腰围 (.8702)。在运动变量内，最高的相关系数是仰卧起坐和拉单杠的成绩 (.6957)。

(2) 第一典型相关系数 (也是最高系数) 等于 .7956。经 F 检定后，其显著度在 .0635 上下，故无法下任何有力的结论。其余两个次要的典型系数都没有达到任何显著度。所以归纳起来，只能勉强算有一个典型系数。

(3) 因为各变量测量的单位不大相同，所以应采用标准化的典型系数来定义典型变量。

(4) 构成第一对典型变量的两个组内线性组合是：
(1.5793) 腰围 + (-0.7754) 体重 + (-.0591) 脉搏，及
(0.7164) 跳高 + (-1.0540) 仰卧起坐 + (-.3495) 拉单杠。
其中脉搏与拉单杠的系数近乎 0，可以忽略之。

(5) 上述的系数似乎影射体重与跳高是两个中介变量；它们的目的在于强化腰围与仰卧起坐之间的负相关。

(6) 最后冗余分析的结果指出第一典型变量对另一组变量的解释程度偏低：只有 28.54% 的生理变量以及 25.84% 的运动变量可以被第一对典型变量所互相解释。

(7) 由于样本的个数太少，上述分析的结果与解释应加以慎重的诠释。

报表 36.1　健身俱乐部成员的分析

Middle-Aged Men in a Health Fitness Club

Data Courtesy of Dr.A.C. Linnerud,NC State Univ

Means and Standard Deviations

3 Physiological Measurements	Variable	Mean	Std Dev
3 Exercise			
20 Observations	WEIGHT	178.600000	24.690505
	WAIST	35.400000	3.201973
	PULSE	56.100000	7.210373

CHINS	9.450000	5.286278
SITUPS	145.550000	62.566575
JUMPS	70.300000	51.277470

Correlations Among the Original Variables

Correlations Among the Physiological Measurements

	WEIGHT	WAIST	PULSE
WEIGHT	1.0000	0.8702	-0.3658
WAIST	0.8702	1.0000	-0.3529
PULSE	-0.3658	-0.3529	1.0000

Correlations Among the Exercise

	CHINS	SITUPS	JUMPS
CHINS	1.0000	0.6957	0.4958
SITUPS	0.6957	1.0000	0.6692
JUMPS	0.4958	0.6692	1.0000

Correlations Among the Original Variables

Correlations Between the Physiological Measurements and the Exercise

	CHINS	SITUPS	JUMPS
WEIGHT	-0.3897	-0.4931	-0.2263
WAIST	-0.5522	-0.6456	-0.1915
PULSE	0.1506	0.2250	0.0349

Canonical Correlation Analysis

	Canonical Correlation	Adjusted Canonical Correlation	Approx Standard Error	Squared Canonical Correlation
1	0.795608	0.754056	0.084197	0.632992
2	0.200556	-.076399	0.220188	0.040223
3	0.072570	.	0.228208	0.005266

Eigenvalues of INV(E)*H
= CanRsq/(1-CanRsq)

	Eigenvalue	Difference	Proportion	Cumulative
1	1.7247	1.6828	0.9734	0.9734
2	0.0419	0.0366	0.0237	0.9970
3	0.0053	.	0.0030	1.0000

Test of H0: The canonical correlations in the current row
and all that follow are zero

	Likelihood Ratio	Approx F	Num DF	Den DF	Pr > F
1	0.35039053	2.0482	9	34.22293	0.0635
2	0.95472266	0.1758	4	30	0.9491
3	0.99473355	0.0847	1	16	0.7748

Multivariate Statistics and F Approximations

S=3　M=-0.5　N=6

Statistic	Value	F	Num DF	Den DF	Pr > F
Wilks' Lambda	0.35039053	2.0482	9	34.22293	0.0635
Pillai's Trace	0.67848151	1.5587	9	48	0.1551
Hotelling-Lawley Trace	1.77194146	2.4938	9	38	0.0238
Roy's Greatest Root	1.72473874	9.1986	3	16	0.0009

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

Raw Canonical Coefficients for the Physiological Measurements

	PHYS1	PHYS2	PHYS3
WEIGHT	-0.031404688	-0.076319506	-0.007735047
WAIST	0.49324167560	.3687229894	0.1580336471
PULSE	-0.008199315	-0.032051994	0.1457322421

Raw Canonical Coefficients for the Exercise

	EXER1	EXER2	EXER3
CHINS	-0.066113986	-0.071041211	-0.245275347
SITUPS	-0.016846231	0.0019737454	0.0197676373
JUMPS	0.0139715689	0.0207141063	-0.008167472

Standardized Canonical Coefficients for the Physiological Measurements

	PHYS1	PHYS2	PHYS3
WEIGHT	-0.7754	-1.8844	-0.1910
WAIST	1.5793	1.1806	0.5060
PULSE	-0.0591	-0.2311	1.0508

Standardized Canonical Coefficients for the Exercise

	EXER1	EXER2	EXER3
CHINS	-0.3495	-0.3755	-1.2966
SITUPS	-1.0540	0.1235	1.2368
JUMPS	0.7164	1.0622	-0.4188

Canonical Structure

Correlations Between the Physiological Measurements and Their Canonical Variables

Correlations Between the Exercise and Their Canonical Variables

	PHYS1	PHYS2	PHYS3		EXER1	EXER2	EXER3
WEIGHT	0.6206	-0.7724	-0.1350	CHINS	-0.7276	0.2370	-0.6438
WAIST	0.9254	-0.3777	-0.0310	SITUPS	-0.8177	0.5730	0.0544
PULSE	-0.3328	0.0415	0.9421	JUMPS	-0.1622	0.9586	-0.2339

Canonical Structure

Correlations Between the Physiological Measurements and the Canonical Variables of the Exercise

	EXER1	EXER2	EXER3
WEIGHT	0.4938	-0.1549	-0.0098
WAIST	0.7363	-0.0757	-0.0022
PULSE	-0.2648	0.0083	0.0684

Correlations Between the Exercise and the Canonical Variables of the Physiological Measurements

	PHYS1	PHYS2	PHYS3
CHINS	-0.5789	0.0475	-0.0467
SITUPS	-0.6506	0.1149	0.0040
JUMPS	-0.1290	0.1923	-0.0170

Canonical Redundancy Analysis

Raw Variance of the Physiological Measurements

	Explained by Their Own Canonical Variables		Explained by The Opposite Canonical Variables		
	Proportion	Cumulative Proportion	Canonical R-Squared	Proportion	Cumulative Proportion
1	0.3712	0.3712	0.6330	0.2349	0.2349
2	0.5436	0.9148	0.0402	0.0219	0.2568
3	0.0852	1.0000	0.0053	0.0004	0.2573

Raw Variance of the Exercise

	Explained by Their Own Canonical Variables		Explained by The Opposite Canonical Variables		
	Proportion	Cumulative Proportion	Canonical R-Squared	Proportion	Cumulative Proportion
1	0.4111	0.4111	0.6330	0.2602	0.2602
2	0.5635	0.9746	0.0402	0.0227	0.2829
3	0.0254	1.0000	0.0053	0.0001	0.2830

Standardized Variance of the Physiological Measurements

	Explained by Their Own Canonical Variables		Explained by The Opposite Canonical Variables	
	Cumulative Proportion	Canonical Proportion	R-Squared	Cumulative Proportion
1	0.4508	0.4508	0.6330	0.2854	0.2854
2	0.2470	0.6978	0.0402	0.0099	0.2953
3	0.3022	1.0000	0.0053	0.0016	0.2969

Standardized Variance of the Exercise

	Explained by Their Own Canonical Variables		Explained by The Opposite Canonical Variables	
	Cumulative Proportion	Canonical Proportion	R-Squared	Cumulative Proportion
1	0.4081	0.4081	0.6330	0.2584	0.2584
2	0.4345	0.8426	0.0402	0.0175	0.2758
3	0.1574	1.0000	0.0053	0.0008	0.2767

Squared Multiple Correlations Between the Physiological Measurements and the First 'M' Canonical Variables of the Exercise

M	1	2	3
WEIGHT	0.2438	0.2678	0.2679
WAIST	0.5421	0.5478	0.5478
PULSE	0.0701	0.0702	0.0749

Squared Multiple Correlations Between the Exercise and the First 'M' Canonical Variables of the Physiological Measurements

M	1	2	3
CHINS	0.3351	0.3374	0.3396
SITUPS	0.4233	0.4365	0.4365
JUMPS	0.0167	0.0536	0.0539

第 37 章　多次元尺度法：统计程序 PROC MDS

37.1　PROC MDS 程序概述

MDS 是 Multidimensional Scaling 的简称，中文的翻译是“多次元尺度法”一般而言，MDS 代表一系列分析法，其目的在于从一组距离矩阵中找出观察体的坐标。简而言之，MDS 的分析法是解析几何原理的相反。现举一例说明：假设在一个平面上有两点 A 与 B，A 的坐标是(1,9)，B 的坐标是 (-4,0)，因此两点间的欧几里得距离是 $\sqrt{[1-(-4)]^2+(9-0)^2}=10.3$。反过来说，若我们只知道 A 与 B 间的距离是 10.3，如何能确定 A 与 B 点的位置，亦即它们的坐标值呢? 这个问题的解答就是 MDS。

在执行 MDS 的分析过程中，读者有许多的选择。比方说，欧几里得的距离计算只是其中的一种定义，其它 的定义可能导出的坐标值就不尽相同了。其次，距离与数据间吻合的程度也可有多种不同的定义。最后，输入的资料文件内除了含观察体两两之间的距离资料外，也可含其它的变量，如时间 (早上、中午、下午等)。如此，就形成一个三元的矩阵。读者可试着用个别差异的 MDS 模型来解释这一类型的数据。

以下是有关 MDS 分析方法的名词解释：

距离数据 (Proximity Data)

两个观察点 (如北京、南非) 或刺激词 (如红色、橘色) 间的距离长短。这个距离可以是相似性的 (Similarity) 或相异性的 (Dissimilarity)。相似性的数据代表两物之间相近或类似的程度。比方说，社会学家所研究的同侪指标 (Sociometric) 就属于相似性数据。此外，两个种族间通婚的频率或两城市间彼此通电话的频繁程度都可算是相似性数据。

相异性数据是上述相似性数据的反面，亦即数据值愈大，其所代表的距离也愈远。最典型的相异性数据就是两点间的欧几里得距离，或两城市间飞行的距离等。若一位研究者请一群受试判断一组联业间相异的程度，其所使用的测验工具是 1-5 的相异度量表。受试者以此量表判断两两职业 (如医生与护士) 间差异的程度，则所收集的数据就是相异性数据。

格局 (Configuration)

观察点在欧几里得空间或加权欧氏空间内的坐标值。MDS 程序利用非线性最小误差法 (Nonlinear Least Squares) 根据输入资料文件内的距离资料来估计各坐标值。

向量系数 (Dimension Coefficient)

这个名词来自加权欧氏空间模型 (Weighted Euclidean Model)，与 Carroll 与 Chang (1970) 的个别差异模型 (INDSCAL) 有关。向量系数与团体欧氏空间 (Group Euclidean Space)相乘后，就产生个人的欧氏空间 (Individual Euclidean Space)。向量系数 (如 0.5) 是 INDSCAL 模型中个人加权值 (Subject Weight) 的平方根。

转换函数 (Transformation Function)

这个函数的作用在于将输入资料文件内的距离数据与报表上格局中两两观察体间的

距离作对应的关系。视读者对选项 LEVEL= 的界定，转换函数可以是一个回归的模型或计量的模型。若函数的形式是回归的模型，则其公式如下：

fit (输入数据)＝fit (输出格局中距离的转换)+误差

若函数以计量模型的形式呈现，则公式如下：

fit (输入数据的转换)＝fit (输出格局中的距离)+误差

在此，

fit 为由选项 FIT= 所界定的函数，由读者自订

转换公式等于由选项 LEVEL= 界定的对应函数，其形式可以是线性的，仿射的 (Affine)，指数函数或单调递增或递减函数。

误差是公式左边与右边平衡时所需的差。MDS 程序假设误差的分配是常态分配，彼此之间是统计独立的。在这些假设的条件下，最小误差平方的估计法所导出的参数估计值是极合理的。

若读者想更深一层了解 MDS，可参考下列的书籍：

(1) Kruskal 与 Wish (1978)。

(2) Arabie，Carroll 与 DeSarbo (1987)。

(3) Young (1987) [含数个 MDS 实际应用的例子]。

(4) Torgerson (1958) [解释 MDS 的理论基础与心理计量学的发展]。

(5) Schiffman，Reynolds，与 Young (1981)。

37.2　MDS 程序基本功能的示范

在这一节里，我们将示范 MDS 如何在很短的时间内把美国十大城市间的航空距离转变成报表上的地图格局。这十个城市分别是：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)。本节所采用的数据与第 43 章第 6 节的例一完全相同，读者可参考该例附的美国地图及十大城市的原址。

由于输入资料文件内的航空距离与两两城市间的欧氏距离十分接近，故程序中测量度设定为绝对的 (LEVEL=ABSOLUTE)，这些选项导致上一节中所提到的转换函数不产生任何作用。

分析的结果会产生每一个城市在二度空间内的坐标值。这些坐标值收集在一个输出资料文件内，然后经由 PLOT 程序绘出图来。

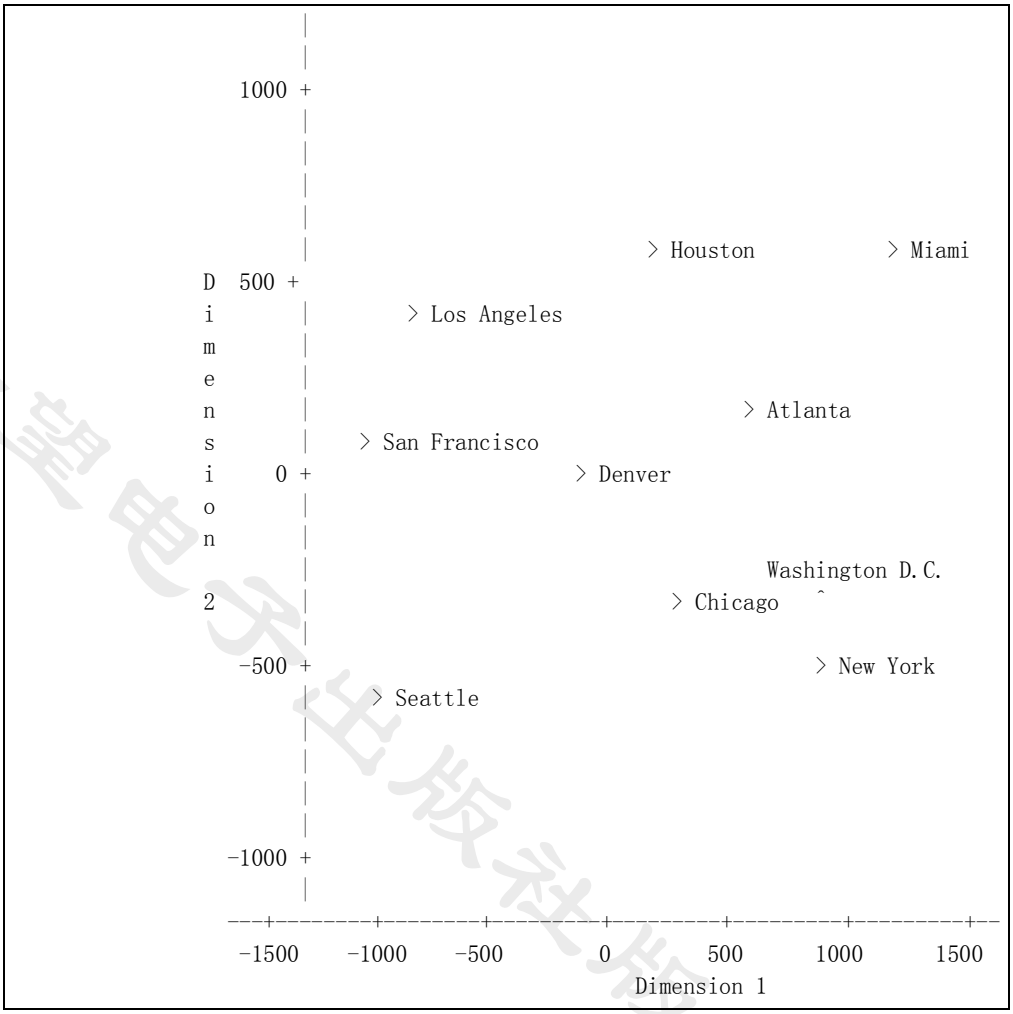
程序如下所示：

```
OPTIONS LS=78 PS=60 NODATE;
TITLE 'MDS analysis of flying mileages between 10 American cities';
DATA MDS;
  INPUT (ATLANTA CHICAGO DENVER HOUSTON LOSANGEL MIAMI NEWYORK SANFRAN
        SEATTLE WASHDC) (5.) @56 CITYNAME $ 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 MDS LEVEL=ABSOLUTE OUT=OUT;
    ID CITYNAME;
RUN;
PROC PLOT DATA=OUT VTOH=1.7;
    PLOT DIM2 * DIM1 $ CITYNAME/HAXIS=BY 500 VAXIS=BY 500;
    WHERE _TYPE_='CONFIG';
RUN;
```

MDS 程序与 PLOT 分析的结果如下：

```
            MDS analysis of flying mileages between 10 American cities

                Multidimensional Scaling:  Data=WORK.MDS
Shape=TRIANGLE Cond=MATRIX Level=ABSOLUTE Coef=IDENTITY Dim=2 Formula=1 Fit=1

            Gconverge=0.01 Maxiter=100 Over=1 Ridge=0.0001

                          Badness-of-Fit     Change in     Convergence
   Iteration    Type          Criterion     Criterion         Measure
   ------------------------------------------------------------------
           0    Initial        0.003273             .        0.856171
           1    Lev-Mar        0.001689      0.001584        0.005128

               Convergence criterion is satisfied.

               MDS analysis of flying mileages between 10 American cities

                 Plot of DIM2*DIM1$CITYNAME.  Symbol points to label.
```



MDS 的结果只含分析的过程 (只需一次的循环距离估计过程) 以及不适合度检定 (Badness-of-fit=0.001689)。两者都显示输入数据与报表上分析的结果十分吻合。

PLOT 程序所产生的格局是 MDS 结果的图形表示。读者不难发现十个城市间的距离是正确的，东西两岸的位置也正确， 然而南北的方向正好颠倒了。这样的结果在 MDS 分析中很难避免，因为输入资料的数据只提供两两城市间的距离而非东南西北的方向。因此，读者可视需要将报表上的格局作适度的调整。

此外，格局内横轴与纵轴的比例是 1.7 比 1。这种安排由 PLOT 程序内选项 VTOH= 来控制。其它的选项则在第 8 章内有详细说明。

37.3　如何撰写 PROC MDS 程序

PROC MDS 含七道指令，其中只有 PROC MDS 是必需的，不可省略。其余的六道指令，则可有可无，以下列举 PROC MDS 的六道指令：

```
PROC MDS 选项串;
    VAR 变量名称串;
    INVAR 变量名称串;
    ID (或 OBJECT) 变量名称;
    WEIGHT 变量名称串;
    BY 变量名称串;
```

指令 #1　PROC MDS 选项串;

在此指令后，有下列选项：

(1) DATA=输入资料文件名称

指明对那一个 SAS 资料文件执行分析。MDS 程序只接受正方矩阵形式的资料，每一个正方矩阵由一位受试者提供。 因此，输入资料文件内含一至多个同等大小的正方矩阵，代表一位受试者 (或团体平均的) 或两位以上受试者对一组刺激 (或观察体) 间距离的判断。数据的性质可以之相似性的或相异性的，然而内设值是相异性数据，如距离。因此，若你的数据属于相似性的，则可加上 SIMILAR 选项让 MDS 程序对这一类数据作适当的处理。一般而言，若输入矩阵的对角线元素大过任何其它 的非对角线值，则 MDS 自动视该数据为相似性数据数据矩阵可以是对称或非对称的矩阵，由选项 SHAPE= 决定。矩阵的大小由 VAR 指令中界定的变量数目决定，而且必须是一个正方矩阵。

(2) SIMLAR (或 SIM 或 SIM= 最大值)

此选项用来将相似性数据转变成相异性数据，以便进行 MDS 的分析。转变的过程如下：取数据中最大的值或 SIM= 所界定的最大值，然后将其它的数据从最大值中减去。如此，相似性数据就转成相异性数据，MDS 分析过程就此开始。

(3) CUTOFF=实数 (如 10)

此选项界定数据中有效数据的下限，凡小于此下限的数据 (如 9 < 10)，则自动被视为遗漏数据。内设值是是 0。

(4) SHAPE=T (或 TRI 或 TRIANGLE 或 TRIANGULAR) 或
SHAPE=S (或 SQU 或 SQUARE)

这个选项界定数据矩阵是否为对称的正方矩阵。若 SHAPE=SQUARE，则表示输入矩阵内右上角与左下角的元素不尽相同，MDS 程序应该将整体的矩阵储存在记忆内并纳入分析过程中。

SQUARE 是另一个选项 CONDITION=ROW 的内设值。

若 SHAPE=TRIANGLE，则 MDS 程序在存储器内只储存左下角的数据矩阵不过，整个矩阵会先输入，然后 MDS 取对角线上下对称的值，将它们平均，以此平均值为代表，存入存储器内并纳入分析中。这个平均值也就是 OUTRES= 资料文件内的原数据 (DATA)。TRIANGLE 是这个选项的内设值。

(5) CONDITION=U (或 UN) 或
CONDITION=R (或 ROW) 或

CONDITION=M (或 MAT 或 MATRIX 或 S， SUB， SUBJECT)

这个选项决定数据矩阵的条件性 (Conditionality)，内设值是 MATRIX。

若 CONDITION=UN，则整个输入数据矩阵被视为一个分割 (Partition) 单位。若 CONDITION=ROW，则每一个正方矩阵的横列被视为一个分割 (Partition) 单位，而且每一分割单位在 MDS 分析过程中会单独被转换函数所处理。若 CONDITION=ROW 而且读者不界定 SHAPE= 的选项，则每一个数据矩阵以正方矩阵的形式储存在存储器内。此正方矩阵可以是不对称的。

若 CONDITION=MATRIX，则每一个数据矩阵被视为一个分割单位。若 CONDITION=UN 或 MATRIX 而且读者不界定 SHAPE= 的选项则只有一个正方数据矩阵被储存在内存内。

(6) LEVEL=A (或 ABS 或 ABSOLUTE) 或

LEVEL=R (或 RAT 或 RATIO) 或

LEVEL=I (或 INT 或 INTERVAL) 或

LEVEL=L (或 LOG 或 LOGINTERVAL) 或

LEVEL=O (或 ORD 或 ORDINAL)

界定数据的测量度 (Lewl of Measurement)；由此导出转变函数的性质。内设值定为 LEVEL=ORDINAL。

若 LEVEL=ABSOLUTE，数据值视为绝对的，因此不经过任何函数的转换。

若 LEVEL=PATIO，则转换函数是一回归模型公式，其作用是将格局上两点间的距离乘以斜率参数，以便与输入的资料相对应。

若 LEVEL=INTERVAL，则转换函数仍然是一个回归模型公式，其作用是将格局上两点间的距离乘以斜率再加上截距参数，以便与输入的资料相对应。这种转换函数的性质是仿射的。

若 LEVEL=LOGINTERVAL，则转换函数属于指数函数的一种，其作用是将格局上两点的距离作平方或三次方的转换，然后再乘以斜率。其结果最后与输入的资料相比较。

若 LEVEL=ORDINAL，则转换函数将输入资料作单调递增地转变，保留数据中相对距离大小的关系而非实际的大小。因此，ORDINAL 所导致的转换函数属于计量的模型，其参数估计的方法是最小误差平方法。

(7) UNTIE

与上述选项 LEVEL=ORDINAL 联用，使输入数据中相持的数据在报表的格局上与不等的距离相对应。若不使用此选项，则相持的输入数据在报表上仍与同等的距离相对应。

(8) FIT=D (或 DIS 或 DISTANLE) 或

FIT=S (或 SQU 或 SQUARED) 或

FIT=L (或 LOG) 或

FIT=正整数

界定一个转换函数，也就是第 37.1 节公式中提到的 fit 函数。这个函数的性质由读者事先设定，内设值是 FIT=DISTANLE 或 FIT=1，其效果等于将输入的距

离资直接与格局上的距离作一对一的对应。

若 FIT=SQUARED 或 FIT=2，则输入资料与格局上的距离都先平方，然后才作一对一的对应。平方后的结果会使距离大的数据对参数估计的影响力较距离小的更深远。

若 FIT=LOG 或 FIT=0，则输入资料与格局上的距离都先经过对数的的转换，然后才作一对一的对应。对数转换的结果会使得距离小的数据对参数估计的影响力较距离大的更深远，这个现象与上述 FIT=SQUARED 的效果刚好相反。

一般而言，FIT=n 代表输入数据与格局上距离的指数乘幂。因此若数据中含负值，则 n 的值必须是 0 或正值。

(9) EPSILON (或 ESP)=正小数 (如 0.5)

这个值必须介于 0 与 1 之间，其作用是避免不能解决的分析状况，如距离除以 0，或负值被开根号等。内设值等于 10 的 -12 次方。在计算过程中，MDS 程序将 EPSILON 乘以第一轮循环分析所求得的距离之平方，这两个数的相乘积称为△。然后将 △ 附加于每一个距离的平方值后，再开根号，如下式所示：

$$\text{距离} = \sqrt{\text{平方距离} + \triangle};$$

如此所求得的距离被纳入下一轮的循环估计中。

(10) DECIMALS (或 DEC)=正整数

此选项界定数估计值在报表上打印时所保留的有效位数，内设值等于 2。

(11) DIM (或 DIMENS 或 DIMENSION)=正整数 (或 n TO m)

此选项界定 MDS 分析结果的格局向度。内设值是 MDS=2，亦即二度空间的格局是 MDS 程序的内设向度。若你想得到多重向度的格局解，则可界定 DIM=1 TO 3，如此表的结果就含一度空间，二度空间，以及三度空间的解。一般而言，四度或以上的向度解不但繁复，而且失去其空间视觉的效益。

(12) COEF=I (或 IDEN 或 IDENTITY) 或

COEF=D (或 DIAG 或 DIAGONAL)

此选项界定分析结果的格局是否依个别受试者作适当的调整。若 COEF=IDENTITY，则格局上两点间的欧氏距离不时个别受试者作调整，这是本选项的内设值。

若 COEF=DIAGONAL，则 MDS 程序执行 Carroll 与 Chang 在 1970 年提出的个别差异多次元尺度法。因此，每一位受试者的格局均可不同。两受试者间的差异在于对每一向度的重要性不同。不同的重要性由此选项界定的矩阵之对角线元素决定。

(13) ALTERNATE=N (或 NO 或 NONE) 或

ALTERNATE=S (或 SUBJECT 或 M 或 MATRIX) 或

ALTERNATE=ROW (或 R=正整数 n)

这个选项界定分析算法所使用的系统。MDS 程序的分析等法称为交替最小平方误差法(Alternating Least Squares Method)，因此这个选项是用来决定交替推算过程中参数前后调整的频繁程度。若数据不算太多，电脑记忆储存空间足够，则交替

调整的次数就频繁；反之，交替调整的次数降低。以下按照所需的内存之多少分别说明这个选项的值：若 ALTERNATE=NONE，则参数的值在每一次循环估计过程自动彼此调整。这种算法的界定只适用于数据较少的资料文件。

若 ALTERNATE=MATRIX，则参数的估计值先针对第一位受试者的资料作调整，然后再一齐针对第二位受试者的资料作调整。按此顺序进行直到每一位受试者的资料都被一一考虑过。最后，才调整横轴与纵轴的尺度单位以及无条件的转换式。这个选项值适用于受试者多然而观察个体 (或刺激词) 数目少的资料文件。

若 ALTERNATE=ROW，则有关受试者参数的估计与上述 MATRIX 一样，然而 ALTERNATE=ROW 将观察体 (或刺激词) 分成小集合。针对每一小集合的观察体，交替运算系统执行无条件式的参数估计。这个选项值适用于含大量观察体的资料。

若 ALTERNATE=ROW=n (如 7)，则 MDS 程序以 7 为单位将资料文件先分成含七横列的小集合，然后进行参数的估计。

(14) INITIAL (或 IN)=SAS 资料文件

此选项界定一个含参数初值的输入资料文件。若省略此选项，则初值由数据本身产生。

(15) INAV=D (或 DATA) 或

INAV=S (或 SSCP)

此选项参定各点在格局上坐标的初值。

若 INAV=DATA，则 MDS 程序计算所有受试者数据的加权平均，由此导出坐标的初值。

若 INAV=SSCP，则 MDS 程序首先估计各受试者数据矩阵中的遗漏数据，将矩阵转变成内积 (Scalar Product)，然后求所有内积的非加权平均，由此导出坐标的初值。

(16) RANDOM 或

RANDOM=seed

这个选项界定坐标的初值为随机数。随机数表的起始可藉 seed (任何实数) 来控制。

(17) FORMULA=0 (或 OLS) 或

FORMULA=1 (或 USS) 或

FORMULA=2 (或 CSS)

这个选项界定分析过程中的不适合度 (Badness-of-fit Criterion)。公式 1 与 2 等于 Kruskal 与 Wish 于 1978 年提出的公式 1，2。若 FIT=LOG，则此选项的内设值是 FORMULA=2；在任何其它情况下，内设值都为 FORMULA=1。

当 FORMULA=0 时，MDS 程序以回归模型来导出格局上各点的位置。回归模型参数的值是根据最小误差平方法估计出来的，其不适合度等于误差均方的平方根。请读者注意，FORMULA=0 不可与 LEVEL=ORDINAL 联用。

当 FORMULA=1 时，不适合度的值会经过数据之 USS (Uncorrected sum of squares，亦即未经平均数矫正过的平方和) 的标准化。

当 FORMULA=1 与选项 FIT=DISTANCE 以及 LEVEL=CRDINAL 联用时，不适合度的公式与 Kruskal 的 Stress 1 公式完全相等。当 FORMULA=1 与选项 FIT=SQUARED 以及 LEVEL=ORDINAL 联用时，不适合度的公式与 Young 的 S-Stress 1 公式完全相等。一般而言，不适合度的值相当于$\sqrt{1-R^2}$，在此，R= 原距离数据与估计距离之间的相关系数。请读者注意：FORMULA=1 的设定不可与选项 FIT=LOG 联用。

当 FORMULA=2 时，不适合度的值会经过数据之 CSS (Corrected sum of squares，亦即经过平均数矫正过的平方和) 的标准化。当 FORMULA=2 与选项 FIT=DISTANCE 以及 LEVEL=ORDINAL 联用时，不适合度的公式与 Kruskal 的 Stress 2 公式完全相等。当 FORMULA=2 与选项 FIT=SQUARED 以及 LEVEL=ORDINAL 联用时，不适合度的公式与 Young 的 S-Stress 2 公式完全相等。一般而言，FORMULA=2 定义下的不适合度值等于$\sqrt{1-R^2}$，在此，R= 原距离数据与估计距离之间的相关系数，两者都事先经过平均数的矫正。FORMULA=2 的定义最适用于展开 (Unfolding) 的分析。

(18) MINCRIT (或 CRITMIN)=n

这个选项界定不适合度收敛的标准。当不适合度值小于或等于 n 时，循环估计的过程停止，内设值等于 10 的 -6 次方。

(19) MAXITER=n

界定循环估计的次数，内设值等于 100 次。

(20) NEGATIVE

此选项容许模型中的斜率参数或幂指数为负值，此选项只可与选项 LEVEL=RATIO， INTERVAL，或 LOGINTERVAL 联用。

(21) NOPHIST (或 NOP)

抑止循环估计的过程在报表上印出来。

(22) NONORM

抑止参数的估计算被标准化。

(23) PDATA

要求在报表上打印出每一个数据矩阵。

(24) PFINAL

要求在报表上打印出参数估计的终值。

(25) PFIT

要求在报表上印出不适合度值以及 R 值。

(26) PINIT

要求在报表上印出参数的初值。

(27) PITER

要求在报表上印出每一次循环估计过程所导出的参数估计值。

(28) PININ

要求在报表上印出由 INITIAL=SAS 资料文件提供的参数初值。

(29) PINAVDATA

要求在报表上印出利用 INAV=DATA 导出的加权和以及数据矩阵的加权平均。

(30) PTRANS

要求在报表上印出转换函数的参数估计值。若 LEVEL=ORDINAL，则此选项无效。

(31) PCONFIG

要求在报表上印出格局上每点的坐标。

(32) PCOEF

要求在报表上印出各向度的加权值。

(33) OUT= 输出资料文件

界定一个 SAS 输出资料文件，内含 MDS 模型中参数所有的估计值以及不适合度的值。然而当读者只要求保留参数中某些估计值时，OUT= 资料文件内就只储存这些估计值而已。

(34) OUTFIT= 输出资料文件

界定一个 SAS 输出资料文件，内含每一分割单位以及整体数据的适合度 (Goodness-of-fit) 及不适合度的值。

(35) OUTRES= 输出资料文件

界定一个 SAS 输出资料文件，内含有效数据的原值，根据 MDS 模型导出的估计值及这两者间的误差等。

(36) OCEF

要求将各向度的加权值纳入 OUT= 输出资料文件内。

(37) OCONFIG

要求将格局上各点的坐标值纳入 OUT= 输出资料文件内。

(38) OCRIT

要求将不适合度的值纳入 OUT= 输出资料文件内。

指令 #2　VAR 变量名称串；

这道指令的目的在于指明输入资料文件内被分析的变量名称。一般而言，凡数值变量都可被纳入分析过程中。由于 MDS 的输入数据必须是一个正方矩阵，因此这些变量名称代表矩阵的横列。

若两位或以上受试者的正方矩阵连续地读进输入资料文件内，则 VAR 指令界定的变量代表矩阵的直行。

指令 #3　INVAR 变量名称串；

这道指令界定在 INITIAL=SAS 资料文件内有关参数初值的变量名称。按内设的语法，第一个变量代表格局的第一向度，第二个变量代表第二向度等，以此类推。若省略此指令，则向度的名称自动以 DIM 1，DIM 2，...，DIM n 表示；在此，n=最高的向度。

指令 #4　ID（或 OBJECT 或 OBJ）变量名称；

这道指令指认 DATA= 输入资料文件内一个变量，其值是用来标明数据中的观察体或刺激词，如第 37.2 节示范例题中的变量 CITYNAME。若省略此指令，则数值变量 (或 VAR 指令界定的变量) 名称就自动成为观察体或刺激词的代号。

指令 #5　WEIGHT 变量名称串；

这道指令旨在界定加权变量串的名称。加权变量的个数应与 VAR 指令所界定的变量个数完全相等，若省略此指令，则所有加权值视为均等。而且排列顺序也完全一致。加权变量的值是用来推导加权欧氏空间模型的参数值。若省略此指令，则所有加权值视为均等。

指令 #6　BY 变量名称串；

MDS 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，然后对每一个小的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 变量串的值作由小到大的重新排列，这个步骤可藉 PROC SORT 来达成。或者，资料文件可先经由 PROC DATASETS 处理，将分组的代号附加在每一个观察体旁。如此，读者可直接使用这个分组代号来撰写 BY 的指令。

有关这种处理法的详细介绍，可参考附录 C.9 节。

37.4　范　　例

例一：教育从业人员对学生能力的多次元尺度分析

这一个例题是根据笔者在美国印第安那大学教授统计课时所收集的一组资料。这组资料所要探讨的问题是：「到底资优学生能力之间有无相关? 这些能力的背后有没有潜在的能力向度? 若有，到底是那些?」

学生能力的类别分为十四种：记忆力 (Memory 或 V1)，音乐 (Music 或 V2)，科学能力(Science 或 V3)，艺术 (Arts 或 V4)，领导能力 (Leadership 或 V5)，关读能力 (Reading 或 V6)，电脑能力 (Computers 或 V7)，领悟力 (Comprehension 或 V8)，词汇 (Vocabulary 或 V9)，动机 (Motivation 或 V10)，解决问题本领 (Problem Solving 或 V11)，机械能力 (Mechanical Ability 或 V12)，独立自主力 (Independence 或 V13)，及推理能力 (Reasoning 或 V14)。这十四个能力的类别由 Guskin，Peng，Majd-Jabbari (1988) 的研究中简化出来的。

笔者将这十四个能力列在问卷上，请 26 位主修社会科学博士学位的学生以自由分类的方式将这十四种能力项目进行归类，然后再使用 MDS 程序分析其结果。

下面程序首先将 26 位受试者归类的结果转变成距离数据，这个步骤藉 ARRAY 及 DO...END 指令来完成。接下来 PRINT 程序将距离数据再打印出来以便检验数据是否正确。MDS 与 PLOT 程序的撰写与第 37.2 节的示范例题十分类似，唯一不同的是：本例的解是三维向度空间的解。所以，PLOT 的指令要求绘制两个图形：一是第一与第二向度

的平面图，二是第一与第三向度的平面图。每一能力在图形上以 * 号表示，而且图形的四边以直线框住。

程　序

```
OPTIONS LS=80 PS=60;
TITLE ' MDS analysis of 14 abilities';
DATA GROUP (TYPE=DISTANCE);
    INPUT (V1-V14)(3.0) @43 ABILITY $ 25.;
    LABEL V1='MEMORY'
          V2='MUSIC'
          V3='SCIENCE'
          V4='ART'
          V5='LEADERSHIP'
          V6='READING'
          V7='COMPUTER'
          V8='COMPREHENSION'
          V9='VOCABULARY'
          V10='MOTIVATION'
          V11='PROBLEM SOLVING'
          V12='MECHANICAL ABILITY'
          V13='INDEPENDENCE'
          V14='REASONING';
ARRAY V{14} V1-V14;
DO I=1 TO 14;
  V{I}= 26.0-V{I}; END; DROP I;
CARDS;
26                                          MEMORY
 4 26                                       MUSIC
13  1 26                                    SCIENCE
 2 25  1 26                                 ART
 2  0  1  0 26                             LEADERSHIP
17  5  7  4  1 26                          READING
13  2 17  1  2  5 26                       COMPUTER
15  3 11  3  2 21 10 26                    COMPREHENSION
16  4  6  3  4 21  5 20 26                 VOCABULARY
 6  9  1  9 15  3  5  4  4 26              MOTIVATION
12  0 17  0  5  5 21 12  5 10 26           PROBLEM SOLVING
 6  5 10  6  0  1 14  3  2  5 11 26        MECHANICAL ABILITY
 4 11  2 10 15  4  3  5  5 14  5  3 26     INDEPENDENCE
13  1 20  1  5  8 20 14  8  7 23 13  8 26 REASONING
;
PROC PRINT;
PROC MDS LEVEL=ORDINAL  DIMENS=3 OUT=OUT;
```

```
    ID ABILITY;
RUN;
PROC PLOT DATA=OUT VTOH=1.7;
    PLOT (DIM2 DIM3) *DIM1 = '*' $ ABILITY/BOX HAXIS=BY .5 VAXIS=BY .5;
    WHERE _TYPE_='CONFIG';
TITLE2 'Plot of Configuration';
RUN;
```

结　果

分三部分：第一部分是 PRINT 程序的输出，目的在于检验 INPUT 步骤中数据的转换是否正确。第二部分是 MDS 分析的结果，三维空间的解十分圆满，因为不适度的值极小 (0.050849)。

第三部分由 PLOT 程序产生，也是 MDS 的视觉显示。欲了解这三个向度所代表的潜在能力，我们可检视各向度上极端的能力：

向度

	1	2	3
正极	音乐 艺术	阅读 词汇	机械 音乐 艺术
负极	科学	机械	领导能力

由这个归纳的表看来，第一向度代表艺术 / 逻辑能力的分野，第二向度代表语文 / 操作能力的分野，第三向度则是个人的 / 社交能力的分野。

报表 37.1　教育从业人员对学生能力的多次元尺度分析

（第一部分）

MDS analysis of 14 abilities

OBS	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	ABILITY
1	0	.	.	.	.	.	.	.	.	.	.	.	.	.	MEMORY
2	22	0	.	.	.	.	.	.	.	.	.	.	.	.	MUSIC
3	13	25	0	.	.	.	.	.	.	.	.	.	.	.	SCIENCE
4	24	1	25	0	.	.	.	.	.	.	.	.	.	.	ART
5	24	26	25	26	0	.	.	.	.	.	.	.	.	.	LEADERSHIP
6	9	21	19	22	25	0	.	.	.	.	.	.	.	.	READING
7	13	24	9	25	24	21	0	.	.	.	.	.	.	.	COMPUTER
8	11	23	15	23	24	5	16	0	.	.	.	.	.	.	COMPREHENSION
9	10	22	20	23	22	5	21	6	0	.	.	.	.	.	VOCABULARY
10	20	17	25	17	11	23	21	22	22	0	.	.	.	.	MOTIVATION
11	14	26	9	26	21	21	5	14	21	16	0	.	.	.	PROBLEM SOLVING
12	20	21	16	20	26	25	12	23	24	21	15	0	.	.	MECHANICAL ABILITY
13	22	15	24	16	11	22	23	21	21	12	21	23	0	.	INDEPENDENCE
14	13	25	6	25	21	18	6	12	18	19	3	13	18	0	REASONING

（第二部分）

Multidimensional Scaling: Data=WORK.GROUP
Shape=TRIANGLE Cond=MATRIX Level=ORDINAL Coef=IDENTITY Dim=3 Formula=1 Fit=1
Mconverge=0.01 Gconverge=0.01 Maxiter=100 Over=2 Ridge=0.0001

		Convergence Measures	
Badness-of-Fit	Change in		

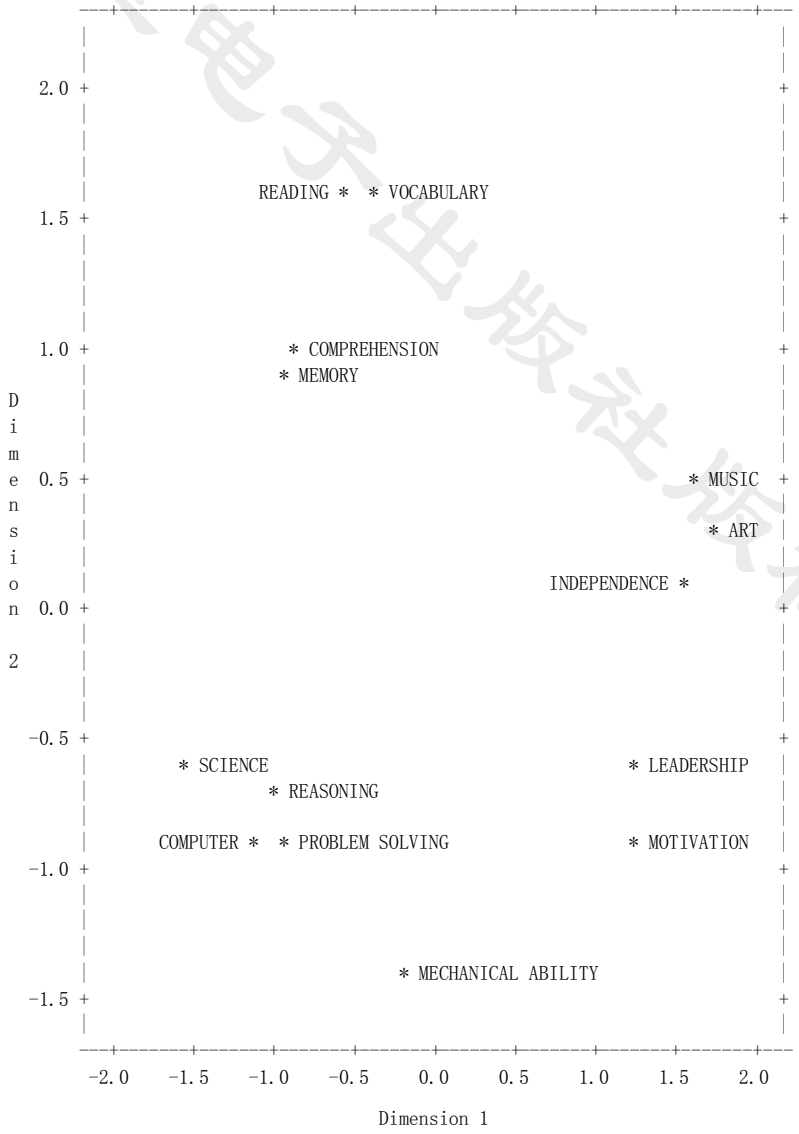
Iteration	Type	Criterion	Criterion	Monotone	Gradient
0	Initial	0.088303	.	.	.
1	Monotone	0.071722	0.016581	0.050706	0.555613
2	Gau-New	0.058699	0.013023	.	.
3	Monotone	0.055782	0.002917	0.018117	0.269793
4	Gau-New	0.055379	0.000403	.	.
5	Monotone	0.052497	0.002882	0.016840	0.205336
6	Gau-New	0.052451	0.000046057	.	.
7	Monotone	0.051530	0.000920	0.009861	0.158977
8	Gau-New	0.050857	0.000673	.	0.016571
9	Gau-New	0.050849	0.000008156	.	0.003663

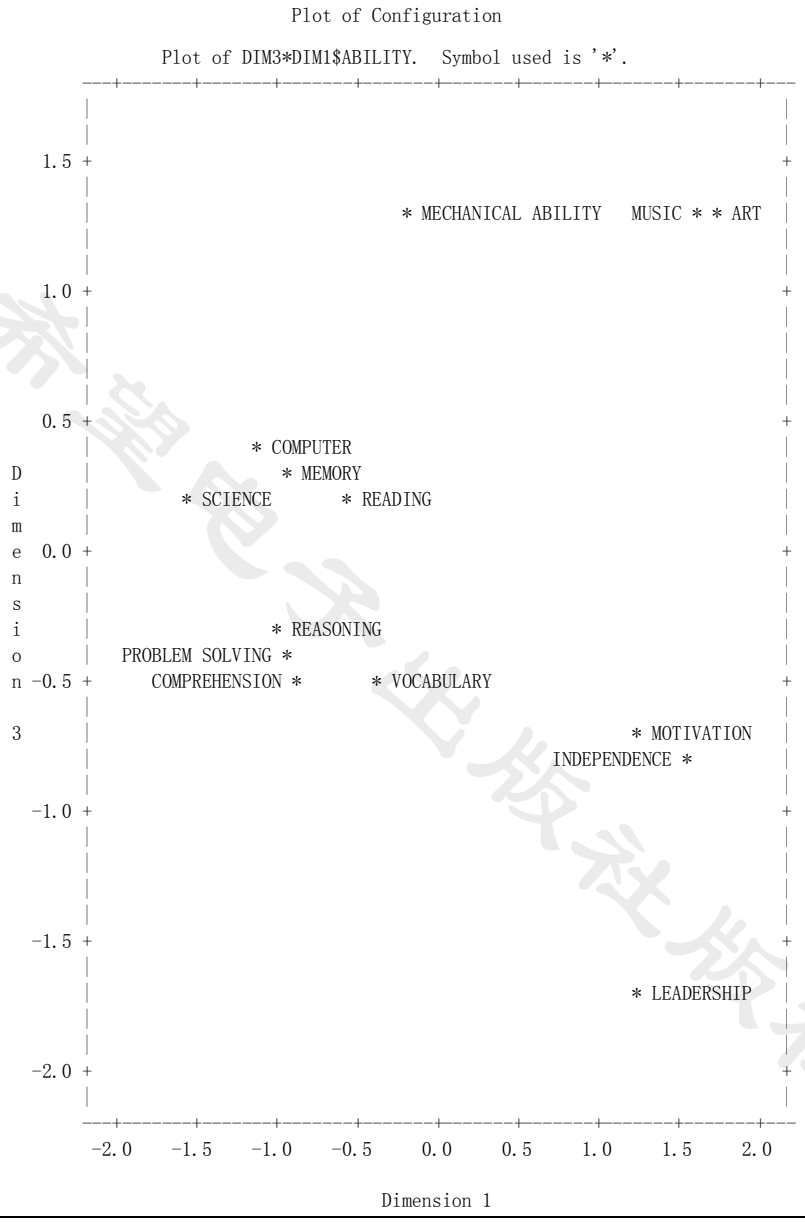
Convergence criteria are satisfied.

（第三部分）

Plot of Configuration

Plot of DIM2*DIM1$ABILITY. Symbol used is '*'.





37.5　注 意 事 项

■遗漏数据的处理

MDS 程序对数据矩阵中缺漏的相似性或相异性的数据值采取忽视的态度。同理，缺漏的加权值以 0 看待。INITIAL=SAS 资料文件内也可含遗漏数据，然而其数目不可太多。

■选项 OUT= 输出资料文件的进一步说明

此选项所产生的资料文件是以 MDS 模型中参数的估计值以及不适合度为主。此资料文件所含的所有变量如下：

(1) BY 指令中的变量名称串。

(2) _DIMENS_ 变量，其值等于格局的总向度。

(3) DIM1，DIM2，...，DIMn 等变量；在此，n= 格局的总向度。

(4) _TYPE_ 变量，其值如下所述：

CONTIG=观察体或刺激词在格局上的坐标

DIAGCOF=向度的加权值或系数，由选项 COEF=DIAGONAL 导出

INTERCEPT=截距参数的估计值

SLOPE=斜率参数的估计值

POWER=指数参数的估计值

CRITERION=不适合度的值

(5) _LABEL_ 变量，或 ID 指令中提及的识别变量名称，用来指认观察体或刺激词。否则，这个变量的值等于遗漏数据，以句号 (.) 表示。

(6) _NAME_，是一个文字变量，其长度不起过八个字母，其值代表观察体 (或刺激词) 以及格局之向度的名称。否则，这个变量的值等于遗漏数据，以圆点 (.) 表示。

■选项 OUTFIT= 输出资料文件的进一步说明

此选项所产生的资料文件是以每一分割单位或整系数据的适合度及不适合度的值为主。此资料文件所含的所有变量如下：

(1) BY 指令中的变量名称串。

(2) _DIMENS_ 变量，其值等于格局的总向度。

(3) N，有效数据的个数。

(4) WEIGHT，每一分割单位的加权值。

(5) CRITER，不适合度的值。

(6) _LABEL_ 变量，或 ID 指令中提及的识别变量名称，用来指认观察体或刺激词 (当 CONDITION=ROW)。否则，这个变量的值等于遗漏数据，以句号 (.) 表示。

(7) _NAME_，是一个文字变量，其长度不超过八个字母。当 CONDITION=ROW 时，这个变量的值代表观察体 (或刺激词) 的名称或格局之向度名称。否则，这个变量的值等于遗漏数据，以句号 (.) 表示。

(8) DISCORR，若 LEVEL=ORDINAL，则此变量的值是转换后的输入数据与格局上距离之间的相关系数。否则，DISLORR 代表输入数据与格局上转换后之距离之间的相关系数。

(9) FITCORR，是转换后的原数据与格局上转换后之距离之间的相关系数。

■选项 OUTRES= 输出资料文件的进一步说明

此选项所产生的资料文件以有效数据的原值，根据 MDS 模型导出的估计值及两者间的误差值等为主。此资料文件所含的所有变量如下：

(1) BY 指令中的变量名称串。
(2) _DIMENS_ 变量，其值等于格局的总向度。
(3) _ROW_ 变量，其值是数据矩阵横列的名字。这个名字由 ID 指令界定的变量串或 INPUT 指令提及的数值变量而来。
(4) _COL_ 变量，其值是数据矩阵直行的名字。这个名字由 ID 指令界定的变量串或 INPUT 指令提及的数值变量而来。
(5) DATA，原数据矩阵中的距离测量。
(6) TRANDATA，当 LEVEL=ORDINAL 时，原始数据的最佳转换数据 (Optimally Transformed datum)。
(7) DISTANCE，根据 MDS 模型导出的两点间的距离。
(8) TRANDIST，当 LEVEL≠ORDINAL 或 ABSOLUTE 时，上述 (7) 距离的最佳转换距离值(Optimally Transformed Distance)。
(9) FITDATA，根据选项 FIT= 的界定，将上述 (6) 的转换数据再加转换后的数据。
(10) FITDIST，根据选项 FIT= 的界定，将上述 (7) 的转换距离再加转换后的距离。
(11) RESIDUAL，残差，等于 FITDATA 减去 FITDIST。
(12) WEIGHT，各数据点的加权值，也就是 WEIGHT 指令所界定之加权变量下的值。

■参数估计值正规化 (Normalization)

在多次元尺度法的分析过程中，参数的估计值并非 MDS 模型的唯一解，因此可经各式的正规化转换而不致影响不适合度的大小。换言之，MDS 程序所估计出来的参数值都是经过正规化转换的。正规化的结果会使得每一向度上各点坐标的平均为 0，而且点与点间距离的等级与输入数据的大小成正比。

若 COEF=IDENTITY，则格局的定向与主轴的方向相同，而且坐标值经过正规化之后，其均方的平方根等于 1。然而当 LEVEL=ABSOLUTE 时，坐标值视为绝对的，因此，其均方的平方根不一定等于 1。

若 COEF=DIAGONAL，则每一向度分别经过正规化，使得每一向度上的坐标值之平均方根等于 1。若 LEVEL≠ABSOLUTE 而且 CONDITION=UN，则向度系数经过正规化后，其平均方根等于 1。若 LEVEL≠ABSOLUTE 而且 CONDITION≠UN，则坐标系数以每一位受试者为单位，经过正规化转换使其平均方根为 1。

若 LEVEL=ORDINAL，则正规化过程会尽量删除有关截距、斜率，或幂级的参数。

■选项 INITIAL=SAS 资料文件的进一步说明

此选项所界定的 SAS 资料文件含参数的初值，其结构与前述的 OUT= 输出资料文件完全相同。一般而言，读者可直接将前一次 MDS 分析的结果 (亦即储存在 OUT= 资料文件内的数据) 当作下一次 INITIAL= 的资料。

INITIAL= 资料文件所需的变量是向度变量，其名字是 DIM1， DIM2， ...，等，或 INVAR 选项所界定的变量名称串。这些变量的值直接代表格局上点的坐标值。一般而言，向度的系数或转换函数的参数不被纳入 INITIAL= 的资料文件内。