合理进行多元分析——变量聚类分析

胡纯严[1]，胡良平[1,2*]

（1. 军事科学院研究生院，北京 100850；

2. 世界中医药学会联合会临床科研统计学专业委员会，北京 100029

*通信作者：胡良平，E-mail：lphu927@163.com）

【摘要】 本文目的是介绍与变量聚类分析有关的基本概念、计算方法、两个实例以及SAS实现。基本概念包括变量聚类分析、相似系数、变量聚类方法、类成分和类结构；计算方法涉及相似系数法计算过程和特征值法计算过程；两个实例涉及的资料分别是“60名正常男性10项指标的测定结果”和“36只兔子的7项指标测定结果”；借助SAS对两个实例中的定量资料进行了全面的变量聚类分析，并对输出结果给出了解释。

【关键词】 聚类统计量；聚类分析；相似系数；类成分；类结构

中图分类号：R195.1 文献标识码：A doi：10.11886/scjsws20230605003

Reasonably carry out multivariate analysis：variable clustering analysis

Hu Chunyan[1], *Hu Liangping*[1,2*]

（1. *Graduate School, Academy of Military Sciences PLA China, Beijing* 100850, *China*;

2. *Specialty Committee of Clinical Scientific Research Statistics of World Federation of Chinese Medicine Societies, Beijing* 100029, *China*

**Corresponding author: Hu Liangping, E-mail: lphu927@163.com*）

【**Abstract**】 The purpose of this article was to introduce the basic concepts, calculation methods, two examples and SAS implementation related to the variable cluster analysis. Basic concepts included variable cluster analysis, similarity coefficient, variable clustering methods, class composition and class structure. The calculation approaches involved the calculation process of the similarity coefficient method and the eigenvalue method. The data involved in the two examples were measurement results of 10 related indicators in 60 normal males and measurement results of 7 indicators in 36 rabbits. With the help of SAS software, a comprehensive variable cluster analysis was carried out on the quantitative data in the two cases, and a reasonable explanation was given for the output results.

【**Keywords**】 Cluster statistic; Cluster analysis; Similarity coefficient; Class component; Class structure

在医学研究等众多科学研究中，由于研究问题的复杂性，研究者往往需要观测很多定量指标的数值，以便对事物或现象的本质及其隐含的规律进行深入的了解和把握。面对多项定量指标，研究者常需要将它们分成不同的类，希望被聚在同一类中的变量具有某些相同的特性。本文将介绍多种变量聚类方法，并结合实例对聚类结果做出解释。

1 基本概念

1.1 变量聚类分析

变量聚类分析是“物以类聚”的一种统计分析方法，用于对事物类别及其类别的数量尚不清楚的情况下进行分类的场合[1-2]。具体地说，就是依据某种原理或规则，将全部变量划分成几类，分入同一类的变量被认为是彼此最密切或最接近的。

1.2 相似系数

变量聚类分析实质上是寻找一种能客观反映变量之间亲疏关系的统计量，然后根据这种统计量把变量分成若干类。变量聚类统计量通常以相似系数表示。相似系数有多种定义，反映定量变量之间密切程度的相似系数有相关系数和夹角余弦[3]；文献[4]提出将相关系数作如下调整，见式(1)和式(2)。

$$C_{ij}=|r_{ij}| \quad (1)$$

$$C_{ij}'=1+r_{ij} \quad (2)$$

在式(1)和式(2)中，r_{ij}为第i个与第j个变量之间的Pearson相关系数。对于同一个资料而言，基于式(1)与式(2)聚类的结果可能不一样。相似系数定义式的选择，取决于聚类结果能否在专业上做出解释。

1.3 变量聚类方法

从形式上来看，变量聚类方法大致可分为系统聚类法、分解法和动态聚类法。①系统聚类法：首先将n个元素（样品或变量）看成n类，然后将性质最接近（或相似程度最大）的两类合并为一个新类，得到$n-1$类，再从中找出最接近的两类加以合并，变成了$n-2$类，如此下去，最后所有的元素全聚在一类之中。②分解法：其程序与系统聚类相反，首先所有的元素均在一类，然后用某种最优准则将它们分为两类，再用同样准则将这两类各自分为两类，从中选1个使目标函数最符合要求者，这样由两类变成了三类。如此下去，一直分裂到每类中只有1个元素为止，有时即使是同一种聚类方法，因聚类形式（距离或相似度的定义方法）不同而有不同的停止规则。③动态聚类法：首先将n个元素大致分成若干类，然后用某种最优准则进行调整，一次又一次地调整，直至无法调整为止。

从计算角度来看，变量聚类方法大致可分为相似系数法和特征值法。①相似系数法：首先把每个变量视为一类，基于选定的相似系数定义式，计算出任意两类变量之间的相似系数值，将最大相似系数值对应的两类合并成一类，这样，类的个数就减少了一个。依此类推，直到所有变量都聚成一类时为止。②特征值法：首先把全部m个变量视为一类，基于相关矩阵计算其特征值和特征向量。若第一特征值除以m所得的贡献率大于事先设定的停止分裂的标准，则表明全部变量属于同一类，停止分裂；反之，需要继续分裂，此时，需要把一类划分成两类。依此类推，直到所有子类都不需要继续分裂时为止。SAS/STAT中的varclus过程采取的是特征值法。

1.4 类成分

基于一个变量集合构造出相关矩阵或协方差矩阵（简称矩阵），求出矩阵的第一特征值及其特征向量，将k个原变量与特征向量的k个元素对应相乘并求和，见式（3）。

$$Z=v_1x_1+\cdots+v_kx_k \quad (3)$$

式（3）中，“Z”被称为“类成分”，它实际上就是第一主成分。显然，类成分也可以是第二主成分、第三主成分……

1.5 类结构

若用语言表达前文的式（3），即Z是原变量或标准化变量的线性组合。同理，可以写出类成分的线性组合，见式（4）。

$$x_i=c_{i1}Z_1+\cdots+c_{ik}Z_k \quad (4)$$

式（4）中，x_i代表第i个原变量或标准化变量；c_{ik}代表与第i个类变量对应的系数。

2 计算方法

2.1 相似系数法计算过程

选定相似系数定义式，计算任意两个变量之间的相似系数值，将具有最大相似系数的两个或多个变量聚成一类，这样，类数至少会减一。并类原则是选最相似的两类合为一类，若最相似的有多类，则把它们同时合为一类。未并类间的相似性不作改变，但要重新计算新类与其他未并类之间的相似性，之后再按以上做法并类，并类一次，至少减少一类。直到所有变量合并成一个大类为止。

在上述的聚类过程中，最关键的问题在于如何计算“变量（特指1个变量）”与“类（至少包含2个变量）”之间的相似系数、“类”与“类”之间的相似系数。这涉及“最小相似系数法”“最大相似系数法”和“折中法或平均法”。因篇幅所限，详见文献[4]。

2.2 特征值法计算过程

SAS/STAT中varclus过程的算法既具有分裂性，又具有迭代性。默认情况下，proc varclus以单个类中的所有变量开始。然后重复以下步骤：①选择一个类进行拆分。根据指定的选项，选定的类具有由其类成分解释的最小变化百分比（使用比例=选项）或与第二个主成分相关的最大特征值（使用maxeigen=选项）。②通过找到前两个主成分，执行正交旋转（特征向量上的原始四次最大旋转[5]），并将每个变量分配给与其具有较高平方相关性的旋转成分，将所选类拆分为两个类。③将变量迭代地重新分配给类，以最大化类成分所占的方差。用户可以要求重新分配算法来维护类的层次结构。

当满足以下任一条件时，该过程将停止拆分：①类的数量大于或等于由maxclusters=选项指定的最大类数量；②每个类都满足由proportion=选项（解释的变异百分比）或maxeigen=选项（第二特征值）或两者指定的停止标准。

默认情况下，当每个类只有一个大于1的特征值时，varclus过程停止分裂，从而满足确定单个底层维度的充分性的最流行标准。

变量到类的迭代重新分配分两个阶段进行。

第一阶段是最近邻成分排序(NCS)阶段,原理上类似于Anderberg[6]描述的最近邻质心排序算法。在每次迭代中,计算类成分,并将每个变量分配给与其具有最高平方相关性的成分。第二阶段是搜索阶段,涉及搜索算法,检验每个变量,了解将其分配给不同的类是否会增加解释的方差。如果在搜索阶段重新分配了一个变量,那么在检验下一个变量之前,将重新计算所涉及的两个类成分。NCS阶段比搜索阶段快得多,但更有可能被局部最优捕获。

如果使用主成分,NCS阶段则是一种交替最小二乘法,并且收敛迅速。对于大量变量来说,搜索阶段可能非常耗时。但是,如果使用默认的初始化方法,搜索阶段很少能够显著改善NCS阶段的结果,因此,搜索需要很少的迭代。如果使用随机初始化,则NCS阶段可能被局部最优捕获,搜索阶段可以从该局部最优中逃脱。

如果使用质心成分,NCS阶段则不是交替最小二乘法,并且可能不会增加所解释的方差;因此,默认情况下,它被限制为一次迭代。

用户可以让varclus过程通过限制变量的重新分配来进行分层聚类,从而使类保持树结构。在这种情况下,当一个类被拆分时,两个结果类中的一个类中的变量可以重新分配给拆分后的另一个类,但不能重新分配给不属于原始类(被拆分的类)的类。

3　实例与SAS实现

3.1　问题与数据结构

3.1.1　两个实际问题及数据

【例1】为研究人脑老化的严重程度,某研究者测定了60名不同年龄的正常男性10项指标,包括年龄、图片记忆、数字广度记忆、图形顺序记忆、心算位数、心算时间、规定时间内穿孔数、步距、步行时双下肢夹角、步速,测定结果见表1[7]。试对这些指标作变量聚类分析。

【例2】某研究测定了36只兔子的7项指标,包括尿钠浓度(mmoL/L)、渗透清除率(mL/min)、尿钠排出量(mmoL/ min)、尿量(mL/min)、尿渗透压[mOsm/(kg·H_2O)]、尿与血浆渗透压之比、游离水清除率(mL/min)。欲通过聚类分析,减少指标以节省人力物力。7项指标两两之间的相关系数见表2[3]。

表1　60名正常男性10项指标的测定结果

Table 1　Measurement results of 10 indicators in 60 normal males

编　号	年　龄	图片记忆	数字广度记忆	图形顺序记忆	心算位数	心算时间	规定时间内穿孔数	步　距	步行时双下肢夹角	步　速
1	16	17	9	14	5.14	4	9	54	35.32	3.92
2	18	12	8	14	3.57	5	11	46	30.66	3.30
…	…	…	…	…	…	…	…	…	…	…
59	78	9	7	4	8.20	2	4	13	9.44	8.91
60	79	13	5	1	9.50	0	6	38	25.53	3.24

表2　7项指标之间的相关系数矩阵

Table 2　Correlation coefficients matrix among 7 indicators

指　标	X_1	X_2	X_3	X_4	X_5	X_6	X_7
X_1	1						
X_2	0.936	1					
X_3	0.995	0.896	1				
X_4	0.974	0.977	0.949	1			
X_5	0.610	0.490	0.621	0.612	1		
X_6	0.440	0.367	0.441	0.477	0.749	1	
X_7	0.705	0.890	0.640	0.773	0.150	0.715	1

注:X_1为尿钠浓度;X_2为渗透清除率;X_3为尿钠排出量;X_4为尿量;X_5为尿渗透压;X_6为尿与血浆渗透压之比;X_7为游离水清除率

3.1.2　对数据结构的分析

例1中,研究者测定了60名不同年龄的正常男性10项指标,故这是一个单组设计10元定量资料。

例2中,研究者测定了36只兔子7项定量指标的取值,故这是一个单组设计7元定量资料。

3.1.3　创建SAS数据集

分析例1资料,设所需要的SAS数据步程序如下:

```
data a1;
infile 'c:\saspal\llhyj.dat';
input age tj sg ts xx xs ck bj jj bs @@;
run;
```

【SAS程序说明】infile语句的含义是打开c盘saspal文件夹中数据文件llhyj.dat，通过下面的input语句读取10个定量变量的数值。数据文件llhyj.dat中包含表1中60行10列数据，以文本格式存储，数据的第一行没有变量名。

分析例2资料，设所需SAS数据步程序如下：

```
data a2(type=corr);
infile cards missover;
input _name_ $ x1-x7;
_type_='corr';
if _name_='n' then _type_='n'; else _type_='corr';
cards;
n 36
X1  1
X2  0.936  1
X3  0.995  0.896  1
X4  0.974  0.977  0.949  1
X5  0.610  0.490  0.621  0.612  1
X6  0.440  0.367  0.441  0.477  0.749  1
X7  0.705  0.890  0.640  0.773  0.150  0.715  1
;
run;
```

3.2 用SAS实现统计分析

3.2.1 分析例1中的资料

设所需要的SAS程序如下[8]：

```
/*采用3种聚类方法，聚成2类*/
proc varclus data=a1 maxclusters=2;
var age tj sg ts xx xs ck bj jj bs;
run;
proc varclus data=a1 centroid maxclusters=2;
var age tj sg ts xx xs ck bj jj bs;
run;
proc varclus data=a1 hi maxclusters=2;
var age tj sg ts xx xs ck bj jj bs;
run;
```

【SAS程序说明】以上SAS程序调用了3次varclus过程，分别采用了3种变量聚类方法：第一次调用时，采用默认的主成分分析法；第二次调用时，采用centroid（质心）聚类法；第三次调用时，采用hi（系统）聚类法。这3种聚类方法都将10个变量聚成2类。

在上面程序中“axclusters=”选项的等号后面依次填入3、4、5、6、7、8，就可实现采用3种变量聚类方法将10个变量依次聚成3、4、5、6、7、8类的目标。

【SAS输出结果及解释】为节省篇幅，下面给出采用3种变量聚类方法将10个变量依次聚成2、3、4、5、6、7、8类的输出结果，见表3。

表3 采用3种变量聚类方法将10个变量分别聚成2至8类的结果

Table 3 Results of clustering 10 variables into 2 to 8 categories using 3 variable clustering methods

聚类数	各类中包含的变量		
	主成分聚类法	质心聚类法	系统聚类法
2	{age,ck,bj,jj,bs} {tj,sg,bs,xx,xs}	{tg,sg,ts,xs,ck,bj,jj} {age,xs,bs}	{age,ck,bj,jj,bs} {tj,sg,bs,xx,xs}
3	{age,ck,bj,jj,bs}{tj,xx,xs}{sg,ts}	{tj,ck,bj,jj}{age,xx,bs}{sg,ts,xs}	{age,ck,bj,jj,bs}{tj,xx,xs}{sg,ts}
4	{bj,jj,bs}{xx,xs}{sg,ts}{age,tj,ck}	{tj,ck,bj,jj}{age,xx,bs}{sg,ts}{xs}	{bj,jj,bs}{tj,xx,xs}{sg,ts}{age,ck}
5	{bj,jj,bs}{xx,xs}{ts}{age,tj,ck}{sg}	{tj,ck}{xx}{bj,jj}{sg,ts,xs}{age,bs}	{bj,jj,bs}{tj,xx}{sg,ts}{age,ck}{xs}
6	{bj,jj,bs}{xs}{ts}{age,tj,ck}{sg}{xs}	{tj,ck}{xx}{bj,jj}{sg,ts}{age,bs}{xs}	{bj,jj,bs}{tj,xx}{ts}{age,ck}{xs}{sg}
7	{bj,jj,bs}{xx}{ts}{tj,ck}{sg}{xs}{age}	{tj,ck}{xx}{bj,jj}{ts}{age,bs}{xs}{sg}	{bj,jj,bs}{tj}{ts}{age,ck}{xs}{sg}{xx}
8	{bj,jj,bs}{xx}{ts}{ck}{sg}{xs}{age}{tj}	{tj,ck}{xx}{bj,jj}{ts}{bs}{xs}{sg}{age}	{bj,jj,bs}{tj}{ts}{age}{xs}{sg}{xx}{ck}

由表3可知，无论将10个变量聚成几类，3种聚类分析方法所得结果不尽相同，但在多数场合下，主成分聚类法与系统聚类法的聚类结果都比较接近。

结论：采用3种聚类方法将10个定量变量聚成多个不同的类，聚类结果不尽相同。结合专业知识可知，表3中“聚类数=4”且“系统聚类法”聚类的结果比较合理，即第一类包含bj（步距）、jj（步行时双下肢夹角）、bs（步速）这三个与走步有关的变量；第二类包含tj（图片记忆）、xx（心算位数）、xs（心算时间）这三个与记忆和计算有关的指标；第三类包含sg（数字广度记忆）和ts（图形顺序记忆）这两个与记忆和计算有关的指标；第四类包含age（年龄）和ck（穿孔）这两个与视力和协调能力有关的指标。

3.2.2　分析例2中的资料

设所需要的SAS程序如下：

proc varclus data=a2(type=corr) maxc=2 outtree=tree;

var X1-X7;

run;

【SAS输出结果及解释】将全部变量聚成一类的结果见表4。

表4　全部变量聚成一类的聚类汇总

Table 4　Clustering summary of all variables aggregated into one class

聚类数	成员数	聚类变异	解释的变异	解释的比例	第二特征值
1	7	7	5.212	74.50%	1.065

注：解释的总偏差=5.212，比例为74.50%

由表4可知，若将7个变量视为一类，基于第一主成分可解释全部7个变量变异的74.50%；第二特征值1.065>1，这提示：将7个变量视为一类不合适，应该将它们拆分，形成两类。将全部变量聚成两类的结果见表5。

表5　全部变量聚成两类的聚类汇总

Table 5　Clustering summary of all variables aggregated into two classes

聚类数	成员数	聚类变异	解释的变异	解释的比例	第二特征值
1	5	5	4.509	90.20%	0.450
2	5	2	1.749	87.50%	0.251

注：解释的总偏差为6.258，比例为89.4%

由表5可知，将7个变量聚成2类后，第一类包含5个变量，能解释该类5个变量变异的90.20%，该类的第二特征值0.450<1，意味着该类不需要继续被拆分；第二类包含2个变量，能解释该类2个变量变异的87.50%，该类的第二特征值0.251<1，意味着该类不需要继续被拆分。两类共解释总变异的6.258（注：7个变量经标准化变换后的总变异为7），贡献率为89.40%。输出的具体内容见表6。

表6　10个变量聚成2类的结果

Table 6　Results of aggregating 10 variablesinto 2 categories

类编号	变　量	R方1	R方2	A/B
Cluster 1	x1	0.952	0.315	0.071
	x2	0.980	0.210	0.025
	x3	0.901	0.322	0.146
	x4	0.975	0.339	0.038
	x7	0.708	0.214	0.380
Cluster 2	x5	0.875	0.283	0.175
	x6	0.875	0.256	0.169

注："R方1"代表本类中的原变量与其类成分之间相关系数的平方；"R方2"代表本类中的原变量与相邻的另一类的类成分之间相关系数的平方；A=1－R方1，B=1－R方2

由表6可知，第一类包含x1、x2、x3、x4、x7这5个变量；第二类包含x5和x6这2个变量。第三列（R方1）上的数值是各行上的变量与自己所在类的类成分之间的相关系数的平方，此值越大，表明分类越合理；第四列（R方2）上的数值是各行上的变量与最靠近类的类成分之间的相关系数的平方，此值越小，表明分类越合理。最后一列上的数值是由第三列与第四列上的数值计算得到的结果，各行上第五列上的数值越小，表明该行上的变量被聚在相应的类中越合适。由此可知，x7被聚在第一类，其合适性较差；而x2和x4被聚在第一类是非常合适的。

标准化评分系数的计算结果见表7。表7呈现的是从标准化变量预测类成分的标准回归系数，若设C_1与C_2分别代表第一和第二类的类成分，则可以写出如下表达式，见式（5）和式（6）。

表7　标准化评分系数的计算结果

Table 7　Calculation results of standardized scoring coefficients

变量名	第一类系数	第二类系数
x1	0.216	0.000
x2	0.220	0.000
x3	0.210	0.000
x4	0.219	0.000
x5	0.000	0.535
x6	0.000	0.535
x7	0.186	0.000

$$C_1=0.216x_1+0.220x_2+0.210x_3+0.219x_4+0.186x_7 \quad (5)$$

$$C_2=0.535x_5+0.535x_6 \quad (6)$$

类结构系数的计算结果见表8。

表8　类结构系数的计算结果

Table 8　Calculation results of the class structure coefficients

变量名	第一类系数	第二类系数
x1	0.976	0.561
x2	0.990	0.458
x3	0.949	0.568
x4	0.988	0.582
x5	0.532	0.935
x6	0.506	0.935
x7	0.838	0.462

表8呈现的是以类成分线性表达每个标准化变量的系数，可以写出如下表达式，见式（7）。

$$\begin{cases} x_1=0.976C_1+0.561C_2 \\ \cdots\cdots \\ x_7=0.838C_1+0.462C_2 \end{cases} \quad (7)$$

结论：基于分裂法，可将例2中的7个定量变量聚成两类。第一类包含"x_1、x_2、x_3、x_4、x_7"5个变量；第二类包含"x_5和x_6"2个变量。

4 讨论与小结

4.1 讨论

从“变量聚类”的字面意思来看，就是要把性质相同且关系密切的变量聚成一类。然而，从统计计算的角度来考量，“性质相同”是无法判定的。虽然“关系密切”可在统计学上给出定义，但需区分正相关和负相关。因此，无论是基于相似系数法，还是基于特征值法，在定义聚类统计量时，要么取相关系数的绝对值作为相似系数，要么取相关系数的平方作为判定依据。这就意味着，在对定量变量进行聚类分析时，被聚在同一类的变量可能具有较大的正相关关系，也可能具有较大的负相关关系。

由本文例1的聚类结果可知，变量聚类分析的结果是不确定的。不确定性表现为3个方面：其一，聚类统计量的定义不同，聚类的结果可能不同；其二，采取的聚类方法不同，聚类的结果可能不同；其三，聚类的数目不同，聚类的结果可能不同。事实上，要想获得最合理的聚类结果，首先应产生出尽可能多的聚类结果；其次，结合具体问题、基本常识和专业知识，判断哪一个聚类结果可能是最合理的。

4.2 小结

本文介绍了与定量变量聚类分析有关的基本概念、计算方法、两个实例及其SAS实现。基本概念包括变量聚类分析、相似系数、变量聚类方法、类成分和类结构；计算方法涉及相似系数法计算过程和特征值法计算过程；两个实例的资料分别是“60名正常男性10项指标的测定结果”和“36只兔子的7项指标测定结果”；基于实例1，采用3种聚类分析方法将10个定量变量分别聚成2~8类；基于实例2，仅采用分裂法将7个定量指标聚成2类，并对输出结果进行了较详细的解释。

参考文献

[1] Johnson RA，Wichern DW. 实用多元统计分析[M]. 6版. 北京：清华大学出版社，2008：671-705.
Johnson RA，Wichern DW. Applied multivariate statistical analysis [M]. 6th edition. Beijing：Tsinghua University Press，2008：671-705.

[2] 颜虹. 医学统计学[M]. 北京：人民卫生出版社，2005：382-388.
Yan H. Medical statistics [M]. Beijing：People's Mental Publishing House，2005：382-388.

[3] 胡良平. 面向问题的统计学：(3) 试验设计与多元统计分析[M]. 北京：人民卫生出版社，2012：40-56.
Hu LP. Problem-oriented statistics：(3) experimental design and multivariate statistical analysis [M]. Beijing：People's Mental Publishing House，2012：40-56.

[4] 孙尚拱. 实用多变量统计方法与计算程序[M]. 北京：北京医科大学、中国协和医科大学联合出版社，1990：147-149.
Sun SG. Practical multivariate statistical methods and calculation programs [M]. Beijing：Beijing Medical University and China Union Medical University Joint Publishing House，1990：147-149.

[5] Harris CW，Kaiser HF. Oblique factor analytic solutions by orthogonal transformation [J]. Psychometrika，1964，29 (4)：347-362.

[6] Anderberg MR. Cluster analysis for applications[M]. New York：Academic Press，1973：11-135.

[7] 胡良平. 现代统计学与SAS应用[M]. 北京：军事医学科学出版社，1996：336-354.
Hu LP. Modern statistics and SAS applications [M]. Beijing：Military Medical Science Press，1996：336-354.

[8] SAS Institute Inc. SAS/STAT®15.1 user's guide[M]. Cary，NC：SAS Institute Inc，2018：10567-10598.

（收稿日期：2023-06-05）
（本文编辑：陈 霞）