合理进行多元分析——非度量型多维尺度分析

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】 本文目的是介绍与非度量型多维尺度分析有关的基本概念、计算方法、两个实例以及SAS实现。基本概念包括非度量型、应用场合、基本思想和应力系数；计算方法涉及定义研究对象之间的次序关系和Kruskal算法的基本步骤；两个实例分别为“英文字母错误识别调查结果”和“6种糖果相似性的调查结果”；借助SAS软件，对两个实例中的数据分别进行了非度量型多维尺度分析，并对SAS输出结果做出了解释。

【关键词】 非度量型；相似性；相异性；应力系数；单调回归转换

中图分类号：R195.1 **文献标识码**：A doi：10.11886/scjsws4合理进行多元分析——非度量型多维尺度分析

Reasonably carry out multivariate analysis: nonmetric multidimensional scaling 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*: *lphu*927@163.com）

【Abstract】 The purpose of this article was to introduce the basic concepts, calculation methods, two examples and SAS implementation related to the nonmetric multidimensional scaling analysis. Basic concepts included nonmetric, application occasions, basic ideas and stress coefficients. The calculation methods involved the definition of the order relationship between the research objects and the basic steps of the Kruskal algorithm. The data in the two examples were "survey results on misrecognition of English letters" and "survey results on the similarity of 6 candies". With the help of SAS software, nonmetric multidimensional scaling analysis was carried out on the data in the two examples, and an explanation was made for the output results of SAS.

【Keywords】 Nonmetric; Similarity; Dissimilarity; Stress coefficient; Monotone regression transformation

在大量的实际问题中，人们获得的相似性或相异性数据所反映的仅仅是研究对象之间近似程度的顺序关系，而非具体的数值大小，这时，度量型多维尺度分析将不再适用，需要采用非度量型多维尺度分析。本文将介绍非度量型多维尺度分析的基本概念、计算方法、应用实例以及SAS实现。

1 基本概念

1.1 非度量型

所谓非度量型，就是所获得的每个数据本身并非具有高度的精确性，数据之间彼此的数量大小是不准确的，只知道它们之间的先后顺序关系。

1.2 应用场合

非度量型多维尺度分析适用于无法获得研究对象之间精确的相似性或相异性数据，仅能得到它们之间等级或次序关系数据的情形。

1.3 基本思想

应用最广的是Shepard于1962年提出的非度量型多维尺度模型[1-2]，其基本思想是将研究对象之间的相似性或相异性数据视为点间距离的单调函数，在保持原始数据次序关系的基础上，用新的相同次序的数据替换原始数据，然后采用度量型多维尺度分析。

1.4 应力系数

应力系数是依据输入数据与输出结果之间的吻合程度、评价模型对资料的拟合效果的统计量[3-4]。

2 计算方法

2.1 定义研究对象之间的次序关系

设δ_{ij}代表第i个对象与第j个对象之间的相似性或相异性的数量，$X_1,\cdots,X_n$代表p维空间中的n个点，第i个点X_i的坐标为$(x_{i1},x_{i2},\cdots,x_{ip})$，任意两点$X_i$和$X_j$之间的距离为$\hat{d}_{ij}$。将所有研究对象之间的$\delta_{ij}$按从小到大的顺序排列，见式(1)。

$$\delta_{i_1j_1}\leqslant\delta_{i_2j_2}\leqslant\cdots\leqslant\delta_{i_mj_m},m=\frac{1}{2}n(n-1) \quad (1)$$

对象i和对象j所对应的δ_{ij}在该排列中的次序可看作是δ_{ij}的秩。在非度量型多维尺度分析中，就是要寻找与研究对象对应的点X_i和X_j，使得X_i和X_j之间的距离$\hat{d}_{ij}$也有如上的次序[即式(1)中的次序]，见式(2)。

$$\hat{d}_{i_1j_1}\leqslant\hat{d}_{i_2j_2}\leqslant\cdots\leqslant\hat{d}_{i_mj_m} \quad (2)$$

在非度量型多维尺度分析中，就是要使研究对象之间相似性或相异性δ_{ij}的次序关系与低维空间中各点距离$\hat{d}_{ij}$的次序关系相匹配。

2.2 Kruskal算法的基本步骤

在求解非度量型多维尺度模型的不同算法中，Kruskal的算法应用最广，包括以下五个步骤[5-6]。

第一步，确定空间的维数k，给出$X_1,\cdots,X_n$的初始值，也就是各点坐标的初始值，该初始值可以采用古典解或随机初始值。

第二步，根据各点的坐标计算两点之间的距离$\hat{d}_{ij}$。

第三步，根据δ_{ij}以及上一步计算的距离$\hat{d}_{ij}$，采用最小二乘单调回归计算出d^*_{ij}，由此将δ_{ij}转换为d^*_{ij}。

第四步，以d^*_{ij}为基础，通过最小化应力函数重新计算各点的坐标。应力函数的定义有多种形式，其中原应力(Raw stress)、应力-1(Stress-1)、应力-2(Stress-2)分别定义如下，见式(3)、式(4)和式(5)。

$$\sigma_r=\sum_{i=1}^{n}\sum_{j=1}^{n}(d^*_{ij}-\hat{d}_{ij})^2 \quad (3)$$

$$\text{Stress}-1=\sigma_1=\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}(d^*_{ij}-\hat{d}_{ij})^2\Big/\sum_{i=1}^{n}\sum_{j=1}^{n}\hat{d}^2_{ij}} \quad (4)$$

$$\text{Stress}-2=\sigma_2=\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}(d^*_{ij}-\hat{d}_{ij})^2\Big/\sum_{i=1}^{n}\sum_{j=1}^{n}(\hat{d}_{ij}-\bar{d})^2} \quad (5)$$

式(5)中，$\bar{d}$为距离$\hat{d}_{ij}$的平均值。应力函数同样可以应用于度量型多维尺度分析，只需将以上各式中的d^*_{ij}用$f(\delta_{ij})$(映射函数)代替。

在应力函数中，d^*_{ij}是已知的，由前文第三步算得，$\hat{d}_{ij}$以及与之相关的各点坐标是未知的，需要进行估计，这时不再使用第一步中的初始值。最小化应力函数就可以估计出各点的坐标，估计过程可以使用梯度法。

第五步，估计出各点新的坐标值后，如果满足收敛标准，则计算停止，该值即为最终的估计值；如果不满足收敛标准，则以该估计值为初始值，返回二步重新进行整个计算，如此反复进行，直至收敛标准满足为止。

Kruskal算法是一个二重迭代过程，其目的是找到使应力尽可能小的k维空间中的n个点。应力的取值大小预示拟合效果，但尚无统计检验方法给出精准的判断。在实践中，可参考以下经验作出判断：当Stress-1=0时，说明拟合完美；当0<Stress-1≤2.5%时，说明拟合非常好；当2.5%<Stress-1≤5%时，说明拟合较好；当5%<Stress-1≤10%时，说明拟合一般；当10%<Stress-1≤20%时，说明拟合较差。

3 实例与SAS实现

3.1 问题与数据结构

3.1.1 2个实际问题及数据

【例1】初学英语时，字母常被认错，表1的数据是在一次调查中获得的结果，“列”表示的字母被认作“行”表示的字母的次数[1]。试根据该数据对这些字母之间的相似程度进行分析。

表1 英文字母错误识别调查结果

Table 1 Survey results of English letter error recognition

字母	错误识别的次数								
	b	d	p	q	g	m	n	v	w
b	.	.	.	.	.	.	.	.	.
d	23	.	.	.	.	.	.	.	.
p	15	5	.	.	.	.	.	.	.
q	5	10	16	.	.	.	.	.	.
g	5	10	15	16	.	.	.	.	.
m	2	0	1	1	2	.	.	.	.
n	1	1	0	2	2	21	.	.	.
v	1	0	1	1	2	5	21	.	.
w	0	1	0	1	3	15	5	32	.

【例2】调查消费者对6种糖果的看法，让他们对任意两种糖果之间的相似性打分，1分表示最相似，15分表示最不相似。其中一位消费者的评价结果见表2[1]。请选择合适的统计分析方法对该资料进行分析。

表2 6种糖果相似性的调查结果

Table 2 Survey results of similarity among six candies

糖果类型	相异性分值					
	A	B	C	D	E	F
A	0	.	.	.	.	.
B	2	0	.	.	.	.
C	13	12	0	.	.	.
D	4	6	9	0	.	.
E	3	5	10	1	0	.
F	8	7	11	14	15	0

3.1.2 对数据结构的分析

表1用下三角矩阵的形式呈现了不同英文字母之间的相似性情况，该表中数据越大，说明两个英文字母越相似；数据越小，说明两个英文字母越不相似，所以此资料属于相似性数据。

表2用下三角矩阵的形式呈现了不同糖果之间的相似性情况，较大的数值表示两种糖果之间的差别越大或近似程度越小，较小的数值表示近似程度越大，所以该数据是相异性数据。

3.1.3 创建SAS数据集

分析例1资料，设所需SAS数据步程序如下：

```
data a1;
input letter $ x1-x9;
cards;
b . . . . . .
d 23 . . . . . .
p 15 5 . . . . . .
q 5 10 16 . . . . . .
g 5 10 15 16 . . . . .
m 2 0 1 1 2 . . . .
n 1 1 0 2 2 21 . . .
v 1 0 1 1 2 5 21 . .
w 0 1 0 1 3 15 5 32 .
;
run;
```

分析例2资料，设所需SAS数据步程序如下：

```
data a2;
input letter $ x1-x6;
cards;
A 0 . . . . .
B 2 0 . . . .
C 13 12 0 . . .
D 4 6 9 0 . .
E 3 5 10 1 0 .
F 8 7 11 14 15 0
;
run;
```

3.2 用SAS实现统计分析

3.2.1 分析例1的资料

设所需要的SAS过程步程序如下[7]：

```
proc mds data=a1 level=ordinal pfinal similar
out=aaa outfit=bbb outres=ccc;
id letter;
run;
proc print data=aaa;run;
proc print data=bbb;run;
proc print data=ccc;run;
```

【SAS程序说明】建立数据集a1，本资料中数据矩阵的主对角线元素都是缺失值。在MDS过程步中，选项level=ordinal表示进行非度量型分析。选项similar指定这里的数据为相似性数据，因为本资料中评分的值越大，说明两个英文字母越相似。需要注意的是，由于主对角线元素是缺失值，这里必须写出similar选项，否则，结果将会有很大的不同。

【SAS输出结果及解释】迭代计算输出结果很多，此处从略。最后一行内容为“拟合劣度统计量或应力系数”，其数值为0.024<0.025，说明模型对资料的拟合效果非常好。基于模型算出的二维拟合构图中与各英文字母对应的2个坐标轴上的坐标见表3。

表3 基于模型算出的二维拟合构图中2个坐标轴上的坐标

Table 3 Coordinates on two coordinate axes in the two-dimensional configuration calculated based on the model

字 母	dim1	dim2	字 母	dim1	dim2
b	1.25	0.23	m	−1.55	0.24
d	1.30	0.25	n	−1.56	0.11
p	1.35	−0.12	v	−1.58	−0.13
q	1.25	−0.20	w	−1.57	−0.21
g	1.11	−0.18			

表3是不同英文字母在二维空间中所对应的点的坐标，也就是拟合构图或感知图中各点的横坐标与纵坐标。这部分结果是由pfinal选项所产生的，默认状态下它们不会被输出。基于非度量型多维尺度分析所产生的拟合构图见图1。由图1可看出，(b,d)，(g,q,p)，(m,n)，(v,w)，这4个括号内的英文字母彼此之间是非常相似的，极易出现识别错误。

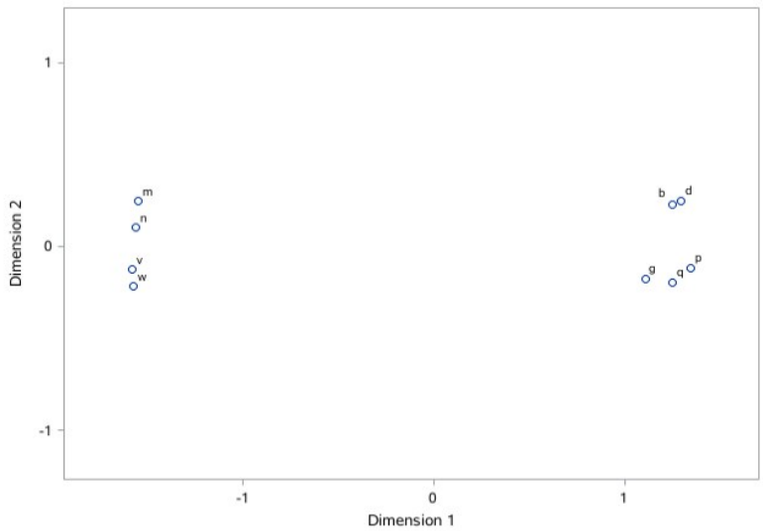

图1　9个英文字母彼此之间关系的拟合构图

Figure 1　Configuration of the relationship between 9 English alphabet.

非度量型多维尺度分析模型对资料的拟合效果见图2。横坐标代表二维空间中两点之间的距离，纵坐标代表经过最小二乘单调回归转换后的相似性数据。本资料中所有散点都集中在直线的两端，但散点离直线的距离都很小，说明模型对资料的拟合效果非常好。

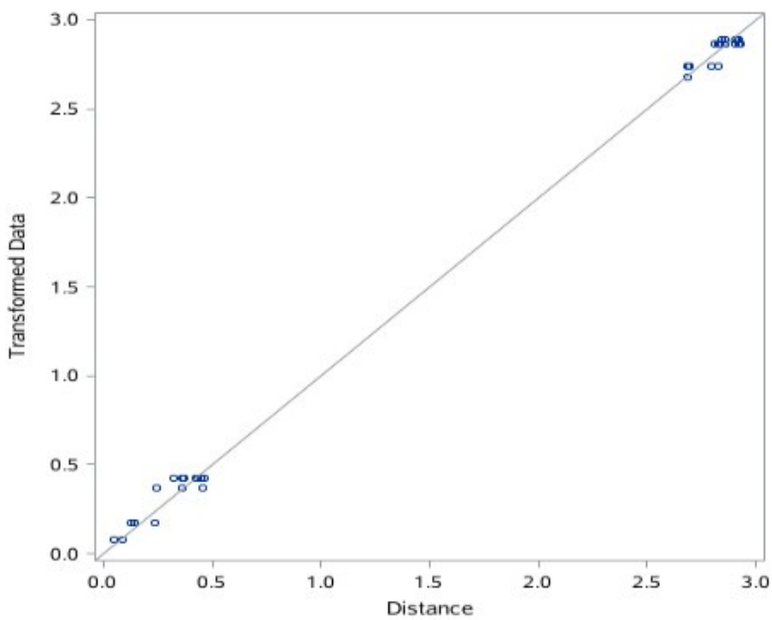

图2　多维尺度分析模型对9个不同英文字母识别错误资料拟合效果散布图

Figure 2　Scatter plot of the fitting effect of multidimensional scaling analysis model on the recognition error data of 9 different English alphabet

3.2.2　分析例2的资料

设所需要的SAS过程步程序如下[7]：

```
proc mds data=a2 level=ordinal coef=identity fit=
1 formula=1
pfinal out=aaa outfit=bbb outres=ccc;
id letter;
run;
proc print data=aaa;run;
proc print data=bbb;run;
proc print data=ccc;run;
```

【SAS程序说明】首先建立数据集a2，字符型变量letter表示糖果种类，变量x1-x6代表不同糖果之间的相异性评分。使用MDS过程实现非度量型多维尺度分析，proc mds语句表示调用MDS过程。该语句中的选项level=ordinal规定使用非度量型多维尺度分析，默认状态下采用的也是非度量型多维尺度分析；coef=identity表示采用未加权模型；fit=1表示使用距离本身；formula=1规定使用Kruskal应力-1

的公式计算应力系数；pfinal规定输出模型中各项参数的最终估计值。

【SAS输出结果及解释】例2资料迭代计算的过程和结果见表4。结果显示，经过5次迭代，收敛标准得到满足，模型最终收敛。拟合劣度标准为0.003<0.025，说明模型对资料的拟合效果非常好。

基于模型算出的二维拟合构图中与6种糖果对应的两个坐标轴上的坐标见表5。表5是6种糖果在二维空间中所对应的点的坐标，也就是拟合构图中各点的横坐标与纵坐标。这部分结果是由pfinal选项所产生的，默认状态下它们不会被输出。

由此模型计算得到的拟合构图见图3。(A、B)，(D、E)，C，F，形成了4种类型，即A与B这两种糖果接近；D与E这两种糖果接近；而C、F与其他四种都相差很大。

反映模型对资料拟合效果的散布图见图4。几乎所有的散点都在一条直线上，说明模型对资料的拟合效果非常好。

表4　基于模型拟合例2资料的迭代计算过程和结果

Table 4　Iterative calculation process and results based on the model fitting example 2 data

迭代	类型	拟合劣度准则	准则中的更改	收敛测度	
				单调	梯度
0	Initial	0.179	.	.	.
1	Monotone	0.011	0.169	0.141	0.760
2	Gau-New	0.007	0.004	.	.
3	Monotone	0.005	0.002	0.005	0.704
4	Gau-New	0.003	0.001	.	0.011
5	Gau-New	0.003	0.000	.	0.000

表5　基于模型算出的二维拟合构图中2个坐标轴上的坐标

Table 5　Coordinates on two coordinate axes in the two-dimensional configuration calculated based on the model

糖果	dim1	dim2	糖果	dim1	dim2
A	0.15	0.95	D	1.08	−0.21
B	−0.12	0.79	E	1.27	0.01
C	−0.41	−1.85	F	−1.98	0.30

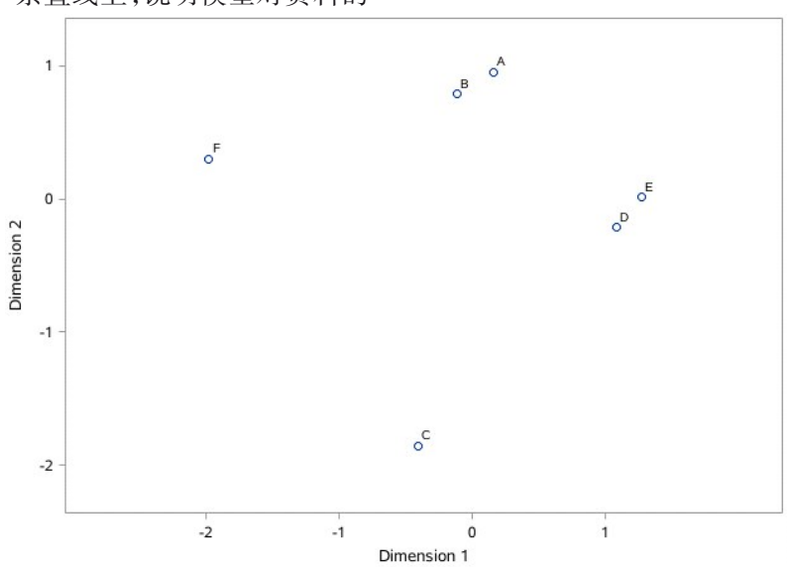

图3　6种糖果彼此之间关系的拟合构图

Figure 3　Configuration of the relationship between six different candies

4　讨论与小结

4.1　讨论

4.1.1　低维空间维数的确定

在非度量型多维尺度分析中，可以通过制作应力与维数k的图形来实现维数的确定[8-9]。在图形中，应力会随着维数的增加而下降，若找到一个k，下降趋势到这一点开始接近水平状态，即形成一个“肘”形曲线，这个k便是“最佳”维数。与度量型多维尺度分析相同，实际应用中空间的维数通常不会超过三维，使用最多的仍然是二维空间。

4.1.2　两种类型多维尺度分析的选择

虽然非度量型多维尺度分析利用的只是研究对象之间近似程度的次序关系，但是，当定量的近似数据不可靠、而其中的顺序关系可靠时，采用非度量方法比度量方法得到的结果更接近实际。所以，在选择这两种类型的多维尺度分析时，应根据数据的实际情况做出合理的判断。

4.2　小结

本文介绍了与非度量型多维尺度分析有关的基本概念、计算方法、两个实例以及使用SAS实现的方法。基本概念包括非度量型、应用场合、基本思

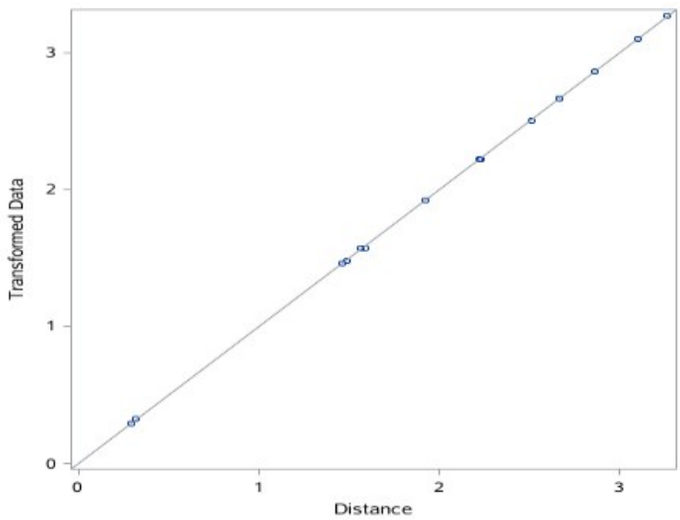

图4 多维尺度分析模型对6种糖果资料拟合效果散布图

Figure 4 Scatter plot of the fitting effect of multidimensional scaling analysis model on six kinds of candy data

想和应力系数;计算方法涉及定义研究对象之间的次序关系和Kruskal算法的基本步骤;两个实例分别为“英文字母错误识别调查结果”和“6种糖果相似性的调查结果”;借助SAS软件,对两个实例中的数据分别进行了非度量型多维尺度分析,并对SAS输出结果做出了解释。

参考文献

[1] 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: 125-134, 303-317.
Hu LP. Problem-oriented statistics: (3) experimental design and multivariate statistical analysis [M]. Beijing: People's Mental Publishing House, 2012: 125-134, 303-317.

[2] 余锦华, 杨维权. 多元统计分析与应用[M]. 广州: 中山大学出版社, 2005: 251-268.
Yu JH, Yang WQ. Multivariate statistical analysis and application [M]. Guangzhou: Sun Yat-sen University Press, 2005: 251-268.

[3] 张润楚. 多元统计分析[M]. 北京: 科学出版社, 2006: 288-311.
Zhang RC. Multivariate statistical analysis[M]. Beijing: Science Press, 2006: 288-311.

[4] 万崇华, 罗家洪. 高级医学统计学[M]. 北京: 科学出版社, 2014: 199-217.
Wan CH, Luo JH. Advanced medical statistics [M]. Beijing: Science press, 2014: 199-217.

[5] 李卫东. 应用多元统计分析[M]. 北京: 北京大学出版社, 2008: 313-330.
Li WD. Applied multivariate statistical analysis [M]. Beijing: Peking University Press, 2008: 313-330.

[6] 何晓群. 多元统计分析[M]. 2版. 北京: 中国人民大学出版社, 2008: 227-254.
He XQ. Multivariate statistical analysis[M]. 2nd edition. Beijing: China Renmin University Press, 2008: 227-254.

[7] SAS Institute Inc. SAS/STAT®15.1 user's guide[M]. Cary, NC: SAS Institute Inc, 2018: 2514-2578, 2997-3216.

[8] Armitage P, Colton T. Encyclopedia of biostatistics [M]. 2nd edition. New York: John Wiley & Sons, Inc, 2005: 3635-3643.

[9] 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: 706-715.
Johnson RA, Wichern DW. Applied multivariate statistical analysis [M]. 6th edition. Beijing: Tsinghua University Press, 2008: 706-715.

(收稿日期:2023-07-26)
(本文编辑:陈 霞)