DOI:10.13203/j.whugis1991.01.008

第16卷 第1期 武汉测绘科技大学学报 Vol.16 No.1
1991年3月 Journal of Wuhan Technical University of Surveying and Mapping Mar. 1991

因子分析用于对单元分类时与聚类分析的比较

吴 纪 桃

摘 要

本文分四种情况将因子分析用于分类时与聚类分析进行了比较，弄清了它们之间的联系和可能发生的差异，并导出了三个控制这种差异的不等式，对一些以前仅有直观感觉的事实得出了理论上的依据。

【关键词】 因子分析；聚类分析；公因子方差；阈值

在专题制图中，常常需要对样品单元进行数字分析，定量地确定各样品单元之间的亲疏关系，并按此关系进行分类。以前常用的方法是聚类分析。近来越来越多地使用因子分析来对样品单元进行分类，在近期出版的国内外专题地图集上，可以看到用因子分析法得出的各种分类地图、区划图等。但是，用因子分析与用聚类分析方法得出的结果会有较大差别吗？在分类的阈值确定后，这种差别最大不超过什么范围？这种差别与什么因素有关？这些问题对因子分析在制图中进一步的应用来讲有着显著的实际意义，但至今尚未见到有关的讨论。下面试对以上问题作一些探讨。

1 用单元间的平方和——交叉和矩阵作因子分析

设 $\underset{p\times n}{Y}$ 是原始数据阵，其中 p 是指标数，n 是制图区域中的地区单元数。$\underset{p\times n}{X}$ 是经过行标准化得到的标准化数据阵，也即：若 $X=(x_{ij})$，$Y=(y_{ij})$，$\bar{y}_i$ 是 Y 的第 i 行数据的平均值，S_i 是 Y 的第 i 行数据的离差平方和被 n 除：

$$S_i^2=\frac{1}{n}\sum_{j=1}^{n}(y_{ij}-\bar{y}_i)^2$$

则

$$x_{ij}=\frac{y_{ij}-\bar{y}_i}{S_i} \qquad i=1,2,\cdots,p, \qquad j=1,2,\cdots,n$$

收稿日期：1990－02－23

单元间的平方和——交叉和矩阵为$\frac{1}{p}X'X$，用它来反映单元间的差异。对此短阵作因子分析，设已求得 $\underset{*\times r}{A}$，$\underset{*\times *}{\Delta}$，使得：

$$\frac{1}{p}X'X = AA' + \Delta . \tag{1}$$

这里 $\Delta=\begin{bmatrix}\sigma_1^2 & & 0\\ & \ddots & \\ 0 & & \sigma_*^2\end{bmatrix}$ 对角线元素是特殊因子方差，A 是载荷矩阵。

若记 $X=\{X_1\cdots X_*\}$，$A=\begin{bmatrix}a'_{(1)}\\ \vdots \\ a'_{(*)}\end{bmatrix}$

则(1)式可写成：

$$(\frac{1}{p}x_ix_j) = (a'_{(i)}a_{(j)}) + \Delta = (a'_{(i)}a_{(j)} + \sigma_i^2\delta_{ij}) \tag{2}$$

$$\delta_{ij}=\begin{cases}1 & i=j\\ 0 & i\neq j\end{cases}$$

$x_1,\cdots,x_*$ 代表制图区域中的几个单元点，$a_{(1)},a_{(2)},\cdots,a_{(*)}$就是几个单元点在新的坐标系(即因子轴)下的坐标。

1.1 用单元间的夹角余弦作为分类的相似系数

设分类的阈值为 γ，即当类与类间的相似系数大于等于 γ 则并为一类，小于 γ 则分为二类。类间的相似系数取元素间的相似系数的最大者。

这样，在原空间中，第 i 与第 j 单元间的相似系数为：

$$\cos\theta_{ij} = \frac{x_i' x_j}{\| x_i \| \cdot \| x_j \|}$$

在因子空间中，第 i 与第 j 单元间的相似系数为：

$$\cos_f\theta_{ij} = \frac{a'_{(i)}a_{(j)}}{\| a_{(i)} \| \cdot \| a_{(j)} \|}$$

由(2)式：

$$\cos_f\theta_{ij} = \frac{\frac{1}{p}x_ix_j}{\sqrt{\frac{1}{p}\| x_i \|^2 - \sigma_i^2}\cdot\sqrt{\frac{1}{p}\| x_j \|^2 - \sigma_j^2}} = \frac{x_i'x_j}{\sqrt{\| x_i \|^2 - p\sigma_i^2}\ \sqrt{\| x_j \|^2 - p\sigma_j^2}}$$

由此式可以看出，在因子空间中，所有的单元间的夹角余弦都比在原空间中的夹角余弦在绝对值上增大了。

由此式还可看出，如果所有的特殊因子方差 $\sigma_i(i=1,\cdots,n)$ 都很小，接近于零，则 $\cos_f\theta_{ij}$就与 $\cos\theta_{ij}$相差很小，所以此时用因子分析所作的分类结果与直接聚类分析的结果是一致的，否则用因子分析所作的分类结果就与直接聚类分析的结果不一致。那么对于给定的阈值 γ，$\sigma_i(i=1,\cdots,n)$应小到什么程度，才能保证这两种方法得出的结果一致。

为方便起见，这里只讨论相似系数非负的情况。

若 $\cos\theta_{ij}\geqslant\gamma$，即在聚类分析时将第 i,j 单元分为同一类，由于 $\cos_f\theta_{ij}\geqslant\cos\theta_{ij}$，所以有：

$$\cos_f\theta_{ij} \geqslant \gamma$$

即在因子空间中也将第 i,j 单元分为同一类。

这说明在原空间中用聚类分析分为同类的单元在因子空间中仍然保持同类，此时对 $\sigma_i(i=1,\cdots,n)$ 没有要求。

若 $\cos\theta_{ij}<\gamma$，即在聚类分析时将第 i,j 单元分为不同类，要想在因子空间中也使这两单元分为不同类，必须保持：

$$\cos_f\theta_{ij} < \gamma$$

$$\Longleftrightarrow \frac{\cos_f\theta_{ij}}{\cos\theta_{ij}} < \frac{\gamma}{\cos\theta_{ij}}$$

$$\Longleftrightarrow \frac{x_i'x_j}{\sqrt{\| x_i \|^2 - p\sigma_i^2}\ \sqrt{\| x_j \|^2 - p\sigma_j^2}} \Big/ \frac{x_i'x_j}{\| x_i \| \cdot \| x_j \|} < \frac{\gamma}{\cos\theta_{ij}}$$

整理得：

$$\sqrt{\left(1 - p\frac{\sigma_i^2}{\| x_i \|^2}\right)\left(1 - p\frac{\sigma_j^2}{\| x_j \|^2}\right)} > \frac{\cos\theta_{ij}}{\gamma} \tag{3}$$

也就是说，要使用因子分析方法分类在第 i 与第 j 单元上有与聚类分析同样的结果，σ_i^2,σ_j^2 必须满足(3)式。从(3)式中分析得知，σ_i^2,σ_j^2 的取值大小与 $\cos\theta_{ij}$ 与 γ 之比的大小有关。若 $\cos\theta_{ij}/\gamma$ 很小，说明 $\cos\theta_{ij}$ 比 γ 小很多，第 i,j 单元间的“亲密”程度与规定的“不亲密”的标准 γ 比差很多，此时 σ_i^2,σ_j^2 可适当大一点；如 $\cos\theta_{ij}/\gamma$ 较大，甚至接近 1，说明第 i,j 单元间“亲密”程度趋于临界值 γ，此时稍有误差就会引起结果的改变，所以，应取 σ_i^2,σ_j^2 很小。这个结论和我们直观想象也是相吻合的。

1.2 用单元间的欧氏距离作分类的相似系数

设第 i,j 单元间的距离的平方为：

$$d_{ij}^2 = \| x_i - x_j \|^2$$

在因子空间中，第 i,j 单元间的距离的平方为：

$$d_{ij}^2(f) = \| a_{(i)} - a_{(j)} \|^2 = \| a_{(i)} \|^2 + \| a_{(j)} \|^2 - 2a_{(i)}'a_{(j)}$$

由(2)式：

$$\begin{aligned} d_{ij}^2(f) &= \frac{1}{p}\| x_i \|^2 + \frac{1}{p}\| x_j \|^2 - 2\frac{1}{p}x_i'x_j - (\sigma_i^2 + \sigma_j^2) \\ &= \frac{1}{p}\| x_i - x_j \|^2 - (\sigma_i^2 + \sigma_j^2) \\ &= \frac{1}{p}d_{ij}^2 - (\sigma_i^2 + \sigma_j^2) \end{aligned}$$

由上式可以看出，如果 σ_i^2,σ_j^2 很小可略去不计，则 $d_{ij}^2 = pd_{ij}^2(f)$。此时虽然在因子空间中两单元的距离发生了变化，但只是将每两个单元间的距离缩小到 $\frac{1}{\sqrt{p}}$ 倍，而并不改变单元间的点位关系，故此时使用因子分析所作的分类结果与聚类分析的结果应是一致的。如果 σ_i^2,σ_j^2 较大，则不能忽略，有：

$$pd_{ij}^2(f) < d_{ij}^2。$$

与 1.1 节中情形类似，此时在因子空间中所作的分类有可能将原来聚类分析时分为不同类的两个单元分为同类。为此，必须将 σ_i^2,σ_j^2 限制在一定小的范围内。

设 d^2 为分类的阈值，则当 $d_{ij}^2>d^2$ 时，第 i,j 单元分为不同类，反之分为同类。现设 $d_{ij}^2>d^2$，即在原空间中，i,j 两单元在不同的类，要使在因子空间中这两单元也分为不同类，需使

$$pd_{ij}^2(f) > d^2$$

即：

$$d_{ij}^2 - pd_{ij}^2(f) < d_{ij}^2 - d^2$$

整理得：

$$p(\sigma_i^2 + \sigma_j^2) < d_{ij}^2 - d^2 \tag{4}$$

故 σ_i^2,σ_j^2 要满足(4)式所限的范围，才能使聚类分析与因子分析得到同样的结果。分析(4)式可知 σ_i^2,σ_j^2 的大小限制是依 d_{ij}^2 与 d^2 的差而定的：$d_{ij}-d^2$ 大，，则 σ_i^2,σ_j^2 也可相应大而不改变分类结果；$d_{ij}^2-d^2$ 小，则 σ_i^2,σ_j^2 就应很小。对于选定的 d^2，所有大于 d^2 的 d_{ij}^2 中的最小者若记为 D^2，则在因子分析时若每两两单元的特殊因子方差满足：

$$p(\sigma_i^2 + \sigma_j^2) < D^2 - d^2$$

则可断言在此因子空间中分类的结果与聚类分析的结果是一致的。

2 用单元间夹角余弦阵作因子分析

设原始数据 $\underset{n\times m}{Y}$，同样作标准化得到 $\underset{n\times m}{X}$。由于标准化过程是对 Y 的行所做的，也就是对各变量做的(在应用上对 Y 的列往往不宜做标准化)，所以在上一段中用的“平方和——交叉和”矩阵 $\frac{1}{P}X'X$ 的对角线元素不是 1，不利于解释。因此在应用上对单元分类更常用的是用单元间的夹角余弦阵作因子分析。

设单元间夹角余弦阵

$$R = \left(\frac{x_i' x_j}{\| x_i \| \cdot \| x_j \|}\right)$$

且经过因子分析已求得一组 $\underset{n\times r}{A}$，$\underset{n\times n}{\Delta}$，使得

$$R = AA' + \Delta$$

记

$$A = \begin{pmatrix} a'_{(1)} \\ \vdots \\ a'_{(n)} \end{pmatrix} \qquad \Delta = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2 \end{pmatrix}$$

上式可写成：

$$\left(\frac{x_i' x_j}{\| x_i \| \cdot \| x_j \|}\right) = (a'_{(i)}a_{(j)}) + \Delta \tag{5}$$

2.1 用单元间的夹角余弦作分类的相似系数

这时，第 i,j 单元间的相似系数为：

$$\cos\theta_{ij} = \frac{x_i' x_j}{\| x_i \| \cdot \| x_j \|}$$

而在因子空间中，第 i,j 单元间的相似系数为：

$$\cos_f\theta_{ij}=\frac{a'_{(i)}a_{(j)}}{\|a_{(i)}\|\cdot\|a_{(j)}\|}$$

由(5)式可得：

$$\cos_f\theta_{ij}=\frac{\cos\theta_{ij}}{\sqrt{1-\sigma_i^2}\sqrt{1-\sigma_j^2}}$$

由此式可知，当 σ_i^2,σ_j^2 很小几乎为零时，用因子分析分类的结果与聚类分析的结果是一致的；当 σ_i^2,σ_j^2 较大时，$\cos_f\theta_{ij}>\cos\theta_{ij}$，这时就有可能将原来分在不同类的单元分为同类了。

设 γ 为分类的阈值，与 1.1 节情形类似，这里也只讨论 $\cos\theta_{ij}$ 非负的情况。看看 σ_i^2,σ_j^2 应有什么限制能保证聚类分析时分为不同类的单元在因子空间中也分为不同类。

现设 $\cos\theta_{ij}<\gamma$，即在聚类分析时将第 i,j 单元分为不同类，要使在因子空间中也将这两单元分为不同类，须：

$$\cos_f\theta_{ij}<\gamma$$

也即：

$$\sqrt{1-\sigma_i^2}\cdot\sqrt{1-\sigma_j^2}>\frac{\cos\theta_{ij}}{\gamma}\tag{6}$$

也就是说，σ_i^2,σ_j^2 若满足(6)式，就能保证这两种分类方法在第 i,j 单元上的结果是一致的。

从分析(6)式还可知：σ_i^2,σ_j^2 的大小是与 $\cos\theta_{ij}$ 与 γ 的比值的大小有关的。若在聚类分析中，类间的界限明显，趋于临界值 γ 的 $\cos\theta_{ij}$ 很少或没有，这时就可以对所有两两单元 i,j 取 $\cos\theta_{ij}$ 中最大者，记为 L，那么只要所有的 σ_i^2,σ_j^2 满足：

$$\sqrt{1-\sigma_i^2}\cdot\sqrt{1-\sigma_j^2}>\frac{L}{\gamma}\tag{7}$$

就能保证在因子空间中的分类结果与聚类分析的结果一致。若类间差别不大，取出的 L 离 γ 就很近，L/γ 很小，σ_i^2,σ_j^2 也必须很小。这样，在因子分析中就不容易使每两两 σ_i^2,σ_j^2 都满足(7)式，从而就会导致与聚类分析不一致的结果。

2.2 用单元间的欧氏距离作分类的相似系数

这时，第 i,j 单元间的距离的平方：

$$d_{ij}^2=\|x_i-x_j\|^2$$

在因子空间中，第 i,j 单元间的距离平方为：

$$d_{ij}^2(f)=\|a_{(i)}-a_{(j)}\|^2=\|a_{(i)}\|^2+\|a_{(j)}\|^2-2a'_{(i)}a_{(j)}$$

由(5)式有：

$$d_{ij}^2(f)=1-\sigma_i^2+1-\sigma_j^2-2\frac{x_i'x_j}{\|x_i\|\cdot\|x_j\|}$$

$$=\left\|\frac{x_i}{\|x_i\|}-\frac{x_j}{\|x_j\|}\right\|^2-(\sigma_i^2+\sigma_j^2)$$

由上式可看出，即使 σ_i^2,σ_j^2 很小或接近于零，因子空间中两单元间的距离也和原空间中的两单元间的距离不相等，单元间点位关系发生变化，此时用因子分析分类与直接用聚类分析分类就在结果上可能有较大差别，在应用中应加以注意。

3 结束语

综上所述，在我们讨论的四种情形中，前三种情形都能在 $\sigma_1^2 \cdots \sigma_r^2$ 满足一定的条件时保证因子分析与聚类分析的分类结果一致，而且 $\sigma_1^2 \cdots \sigma_r^2$ 小的程度是依原空间中类与类间的差别而定的。总体上看，因子分析与聚类分析之间的差异均发生在一些属性不太确定的过渡性单元上。因此，在应用中，使用因子分析对样品单元分类，其结果与聚类分析是不会有本质差别的。

在地图制图中，我们经常需要将分类过程与依据在平面上用散点图或三角形图表表示出来，但 $x_1 \cdots x_r$ 是 P 维点，当变量个数 $P \geqslant 4$ 时，就不能做到这一点。而在因子空间中，$a_{(1)} \cdots a_{(n)}$ 是 r 维的，r 一般较小(2 或 3)，所以可将 n 个单元点 $a_{(1)} \cdots a_{(n)}$ 描在平面上或三角形图表中，在图中用肉眼就立刻观察出分类的结果。由于因子轴一般具有实际意义，所以从图上还得到分类的依据。从这个意义上讲，在地图制图中用因子分析方法对地区单元进行分类是合理可行的。

参 考 文 献

[1] 张尧庭，方开泰. 多元统计分析引论. 科学出版社，1982.

[2] 张克权. 专题制图数学制图模型及其解释. 武汉测绘科技大学讲义，1985.

[3] Granadesikan R. Methods for Statistical Data Analysis of Multivariate Observations. John Wiley & Sons, Inc. 1977.

Comparision between Factor Analysis and Classification Analysis

Wu Jitao

Abstract

In this paper, factor analysis, when being used for classification, is compared with classification analysis under four situations. Some connections and differences between the two methods are found out and three inequalities which restrict the differences are deduced. Theroretical proof is given here to some facts which were felt true.

【Key words】 factor analysis; classification analysis; communality threshold