doi:10.3969/j.issn.1673-9833.2014.03.017

多属性数据聚类的一种因子分析新方法

张爱平，陈志彬

（湖南工业大学 理学院，湖南 株洲 412007）

摘 要：根据因子分析法的思想，用统计学的方法，建立多属性数据样本间的相似矩阵，探索求因子载荷矩阵的有效方法，实现多属性数据的样本聚类。文中的方法是因子分析法在聚类分析中的进一步推广与应用。

关键词：多属性；样本；相似矩阵；数据聚类

中图分类号：O212 **文献标志码**：A **文章编号**：1673-9833(2014)03-0083-05

A New Factor Analysis Method in Multiple Attribute Data Clustering

Zhang Aiping，Chen Zhibin

（School of Science，Hunan University of Technology，Zhuzhou Hunan 412007，China）

Abstract：According to the thought of factor analysis method established the similar matrix between multiple attribute data samples by statistical methods, and explored the effective solution to the factor loading matrix for the realization of sample clustering of multiple attribute data. This method is the factor analysis method further extended and applied in clustering analysis.

Keywords：multiple attribute; sample; similar matrix; data clustering

0 引言

因子分析法是通过建立数学模型，用线性代数的方法，研究多属性变量间的内在依赖关系。用少于属性变量个数的几个抽象变量表示被观测数据的基本结构，实现对被观测数据变量的降维，达到简化数据结构的目的。这几个抽象的变量通常被称为因子，每个因子综合地包含了多个属性变量的信息，是一些异于可观测原始变量的不可观测的潜在变量。因子分析的内容较丰富，常见的类型可以概括为两类[1-5]：一类是R型因子分析，另一类是Q型因子分析，前者是基于变量间的相关系数矩阵，后者则基于样本间的相似矩阵，两种矩阵均为非负定矩阵。这两种类型选择因子分析的对象和计算的出发点不同但方法类似。在实际问题中，由于被观测的样本数目 n 通常较大，因此Q型因子分析中的样本相似矩阵是一个阶数较高的 n 阶方阵，其计算量与 n^2 同阶且可能是非正定的；而求解样本相似矩阵的特征根与特征向量的计算量与 n^3 同阶。由于计算量随阶数 n 的增大而急剧增大，这就限制了以样本为变量的Q型因子分析法在经济学、生物学和社会学等领域中的应用。

为此，本文根据高阶样本相似矩阵与因子载荷矩阵的关系，通过间接地求解一个与高阶样本相似矩阵有联系的低阶矩阵的特征根与特征向量，探讨因子载荷矩阵的计算方法。

收稿日期：2014-03-10

基金项目：湖南省教育科学研究基金资助项目（10C0656），湖南省教育改革基金资助项目（288）

作者简介：张爱平（1967-），女，湖南冷水江人，湖南工业大学副教授，主要从事应用数学方面的教学与研究，E-mail：zaping@163.com

1 预备知识

文中所讨论的随机向量是具有 p 种属性变量的总体，用列向量$\boldsymbol{X}=\left(X_1,X_2,\cdots,X_p\right)^{\mathrm{T}}$表示。$n$ 个样本为源于总体 n 次观测所获得的 n 个 p 维数据组（$p\ll n$）；第 i 个样本用列向量$\boldsymbol{Y}_i=\left(y_{i1},y_{i2},\cdots,y_{ip}\right)^{\mathrm{T}}$表示，均值$\overline{Y}_i=\frac{1}{p}\sum_{k=1}^{p}y_{ik}$；$n$ 个样本的数据矩阵表示为$\boldsymbol{Y}=\left(y_{ij}\right)_{n\times p}$或用列向量表示为$\boldsymbol{Y}=\left(\boldsymbol{Y}_1,\boldsymbol{Y}_2,\cdots,\boldsymbol{Y}_n\right)^{\mathrm{T}}$。为了方便，在文中引入如下记号：

样本向量的均值

$$\overline{\boldsymbol{Y}}=\left(\frac{1}{p}\sum_{k=1}^{p}y_{1k},\frac{1}{p}\sum_{k=1}^{p}y_{2k},\cdots,\frac{1}{p}\sum_{k=1}^{p}y_{nk}\right)^{\mathrm{T}}=\left(\overline{Y}_1,\overline{Y}_2,\cdots,\overline{Y}_n\right)^{\mathrm{T}}。$$

样本向量的离差矩阵

$$\boldsymbol{S}=\sum_{i=1}^{p}\left(\boldsymbol{Y}_{(i)}-\overline{\boldsymbol{Y}}\right)\left(\boldsymbol{Y}_{(i)}-\overline{\boldsymbol{Y}}\right)^{\mathrm{T}}=\left(s_{ij}\right)_{n\times n},$$

式中：$\boldsymbol{Y}_{(l)}-\overline{\boldsymbol{Y}}=\left(y_{1l}-\overline{Y}_1,y_{2l}-\overline{Y}_2,\cdots,y_{nl}-\overline{Y}_n\right)^{\mathrm{T}}$；

$s_{ij}=\sum_{k=1}^{p}\left(y_{ik}-\overline{Y}_i\right)\left(y_{jk}-\overline{Y}_j\right)$。

样本向量的协方差矩阵

$$\boldsymbol{V}=\frac{1}{p-1}\boldsymbol{S}=\left(v_{ij}\right)_{n\times n},$$

式中$v_{ij}=\frac{1}{p-1}\sum_{k=1}^{p}\left(y_{ik}-\overline{Y}_i\right)\left(y_{jk}-\overline{Y}_j\right)$。

样本向量的相似阵

$$\boldsymbol{R}=\left(r_{ij}\right)_{n\times n},$$

式中$r_{ij}=\frac{s_{ij}}{\sqrt{s_{ii}}\sqrt{s_{jj}}}=\frac{v_{ij}}{\sqrt{v_{ii}}\sqrt{v_{jj}}}$。

令$z_{ij}=\frac{y_{ij}-\overline{Y}_i}{\sqrt{v_{ii}}}$，将样本数据矩阵$\boldsymbol{Y}=\left(y_{ij}\right)_{n\times p}$标准化后得到矩阵$\boldsymbol{Z}=\left(z_{ij}\right)_{n\times p}$，其列向量矩阵表示为$\boldsymbol{Z}=\left(\boldsymbol{Z}_1,\boldsymbol{Z}_2,\cdots,\boldsymbol{Z}_n\right)^{\mathrm{T}}$，其中$\boldsymbol{Z}_i=\left(z_{i1},z_{i2},\cdots,z_{iP}\right)^{\mathrm{T}}$，$i=1,2,\cdots,n$。

$\boldsymbol{Z}$的离差矩阵为$\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}$，协方差矩阵为$\frac{1}{p-1}\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}$，标准化后样本向量 $\boldsymbol{Z}$ 的协方差矩阵等于原样本向量 $\boldsymbol{Y}$ 的相似矩阵，即

$$\frac{1}{p-1}\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}=\boldsymbol{R}。$$

引理 1[6] 实对称矩阵的不同特征值的特征向量彼此正交。

引理 2[6] 对于 n 阶实对称矩阵 $\boldsymbol{B}$，必存在一个 n 阶正交矩阵 $\boldsymbol{P}$ 使得$\boldsymbol{P}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{P}=\boldsymbol{\Lambda}$（其中$\boldsymbol{\Lambda}$是以矩阵 $\boldsymbol{B}$ 的特征值为对角元素的对角矩阵），即实对称矩阵都可以对角化。

引理 3 对于实对称矩阵$\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}$与$\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}$有如下结论：

1）矩阵$\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}$与$\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}$矩阵的特征值为非负实数，且 2 个矩阵具有相同的正特征值。

2）若$\boldsymbol{\xi}_i$为矩阵$\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}$的非零特征值$\lambda_i$的单位特征向量，则$\boldsymbol{\eta}_i=\frac{1}{\sqrt{\lambda_i}}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\xi}_i$为矩阵$\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}$的非零特征值$\lambda_i$的单位特征向量。

3）若$\boldsymbol{\eta}_i$为矩阵$\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}$的非零特征值$\lambda_i$的单位特征向量，则$\boldsymbol{\xi}_i=\frac{1}{\sqrt{\lambda_i}}\boldsymbol{Z}\boldsymbol{\eta}_i$为矩阵$\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}$的非零特征值$\lambda_i$的单位特征向量。

证明 1）根据实对称矩阵$\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}$及$\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}$具有非负定性，由引理 2 即可得它们的特征值为非负实数。不妨设$\boldsymbol{\xi}_i$是矩阵$\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}$关于正特征值$\lambda_i$的单位特征向量，即$\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\xi}_i=\lambda_i\boldsymbol{\xi}_i$，则有

$$\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\xi}_i\right)=\lambda_i\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\xi}_i\right),\ \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\xi}_i\neq\boldsymbol{0}。\tag{1}$$

根据式（1），可推得矩阵$\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}$的特征多项式为

$$\left|\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}-\lambda_i\boldsymbol{I}\right|=0,\tag{2}$$

即λ_i是矩阵$\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}$的特征值。

同理由$\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}\boldsymbol{\eta}_i=\lambda_i\boldsymbol{\eta}_i$，可推得矩阵$\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}$的特征多项式为

$$\left|\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}-\lambda_i\boldsymbol{I}\right|=0,$$

即λ_i是矩阵$\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}$的特征值。

2）若$\boldsymbol{\xi}_i$是矩阵$\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}$关于$\lambda_i$的单位特征向量，则有

$$\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\xi}_i=\lambda_i\boldsymbol{\xi}_i,\ \boldsymbol{\xi}_i^{\mathrm{T}}\boldsymbol{\xi}_i=1。$$

于是得

$$\begin{cases}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}\left(\frac{1}{\sqrt{\lambda_i}}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\xi}_i\right)=\lambda_i\left(\frac{1}{\sqrt{\lambda_i}}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\xi}_i\right),\\ \left(\frac{1}{\sqrt{\lambda_i}}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\xi}_i\right)^{\mathrm{T}}\left(\frac{1}{\sqrt{\lambda_i}}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\xi}_i\right)=1。\end{cases}\tag{3}$$

取$\boldsymbol{\eta}_i=\frac{1}{\sqrt{\lambda_i}}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\xi}_i$，即$\boldsymbol{\eta}_i$为矩阵$\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}$相应于特征值$\lambda_i$的单位特征向量。

3）同理可证，若$\boldsymbol{\eta}_i$为非零特征值λ_i关于矩阵$\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}$的单位特征向量，则$\boldsymbol{\xi}_i=\frac{1}{\sqrt{\lambda_i}}\boldsymbol{Z}\boldsymbol{\eta}_i$为矩阵$\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}$相应于特征值$\lambda_i$的单位特征向量。

2 主要结论及证明

对于 p 种属性的 n 个样本，首先将原始数据矩阵标准化，得矩阵$\boldsymbol{Z}=\left(z_{ij}\right)_{n\times p}$，若用列向量表示，则记

为$\boldsymbol{Z}=(\boldsymbol{Z}_1,\boldsymbol{Z}_2,\cdots,\boldsymbol{Z}_n)^{\mathrm{T}}$，其中$\boldsymbol{Z}_i=(z_{i1},z_{i2},\cdots,z_{ip})^{\mathrm{T}}$，$i$=1，2，$\cdots$，$n$。

如果观测到的n个样本之间具有强相似性，则可依照样本相似性的大小将n个样本分组，使得同组的样本之间相似性较高，不同组的样本之间相似性较低，实现样本的聚类，并对类中样本所具有的共性进行分析和解析。

根据因子分析的思想，引入一组抽象公共因子变量$\boldsymbol{F}=(\boldsymbol{F}_1,\boldsymbol{F}_2,\cdots,\boldsymbol{F}_m)^{\mathrm{T}}$和一组特殊因子变量$\boldsymbol{\theta}=(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\cdots,\boldsymbol{\theta}_n)^{\mathrm{T}}$，将样本分量$\boldsymbol{Z}_i(i=1,2,\cdots,n)$近似地用因子变量$\boldsymbol{F}_i(i=1,2,\cdots,m)$的线性关系式表示。依据样本与$\boldsymbol{F}$的相似程度，近似地通过因子载荷矩阵用公共因子$\boldsymbol{F}_i(i=1,2,\cdots,m)$来描述位于同组样本具有的基本结构。样本$\boldsymbol{Z}=(\boldsymbol{Z}_1,\boldsymbol{Z}_2,\cdots,\boldsymbol{Z}_n)^{\mathrm{T}}$关于公共因子$\boldsymbol{F}$和特殊因子变量$\boldsymbol{\theta}$的数学模型为

$$\boldsymbol{Z}_i=a_{i1}\boldsymbol{F}_1+a_{i2}\boldsymbol{F}_2+\cdots+a_{im}\boldsymbol{F}_m+\boldsymbol{\theta}_i,\ i=1,2,\cdots,n,\quad(4)$$

且满足下列条件：

ⅰ）$m\leqslant n$；

ⅱ）$\mathrm{cov}(\boldsymbol{F},\boldsymbol{\theta})=0$，即协方差矩阵为零，公共因子与特殊因子不相关；

ⅲ）$\mathrm{cov}(\boldsymbol{Z}_i,\boldsymbol{Z}_i)=1$，$i$=1，2，$\cdots$，$n$，即样本的方差等于1；

ⅳ）$\mathrm{cov}(\boldsymbol{F}_i,\boldsymbol{F}_j)=\begin{cases}1, & i=j,\\ 0, & i\neq j,\end{cases}$ $i,j=1,2,\cdots,m$，即公共因子彼此不相关，$\boldsymbol{F}$的协方差矩阵为单位矩阵。

ⅴ）$\mathrm{cov}(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j)=\begin{cases}\sigma_i^2, & i=j,\\ 0, & i\neq j,\end{cases}$ $i,j=1,2,\cdots,n$即特殊因子彼此不相关，特殊因子$\boldsymbol{\theta}_i$的方差等于σ_i^2。

令式（4）中公共因子变量的系数矩阵与特殊因子变量的列矩阵分别为

$$\boldsymbol{A}=(\boldsymbol{A}_1,\boldsymbol{A}_2,\cdots,\boldsymbol{A}_m),$$

$$\boldsymbol{\theta}=(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\cdots,\boldsymbol{\theta}_n)^{\mathrm{T}}。$$

则模型（4）简化为

$$\boldsymbol{Z}=\boldsymbol{AF}+\boldsymbol{\theta},\quad(5)$$

其中系数矩阵$\boldsymbol{A}=(a_{ij})_{n\times m}$称为公共因子变量$\boldsymbol{F}$的载荷矩阵。

定理1　对于模型（5），在满足条件ⅰ）~ⅴ）时，有如下结论：

1）$\mathrm{cov}(\boldsymbol{Z}_i,\boldsymbol{F}_j)=a_{ij}$，$i$=1，2，$\cdots$，$n$，$j$=1，2，$\cdots$，$m$；

2）$a_{i1}^2+a_{i2}^2+\cdots+a_{im}^2+\sigma_i^2=1$，$i$=1，2，$\cdots$，$n$；

3）$\boldsymbol{Z}_i$与$\boldsymbol{F}_j$的相关系数r_{ij}=a_{ij}；

4）$\boldsymbol{R}=\boldsymbol{AA}^{\mathrm{T}}+\mathrm{diag}(\sigma_1^2,\sigma_2^2,\cdots,\sigma_n^2)$，其中$\boldsymbol{R}=\dfrac{1}{p-1}\boldsymbol{ZZ}^{\mathrm{T}}$，$\mathrm{cov}(\boldsymbol{\theta},\boldsymbol{\theta})=\mathrm{diag}(\sigma_1^2,\sigma_2^2,\cdots,\sigma_n^2)$。

证明　1）根据式（4）及条件ⅱ）有

$$\begin{aligned}\mathrm{cov}(\boldsymbol{Z}_i,\boldsymbol{F}_j)&=\mathrm{cov}(a_{i1}\boldsymbol{F}_1+a_{i2}\boldsymbol{F}_2+\cdots+a_{im}\boldsymbol{F}_m+\theta_i,\boldsymbol{F}_j)=\\ &\mathrm{cov}[(a_{i1},a_{i2}\cdots,a_{im})\boldsymbol{F},\boldsymbol{F}_j]+\mathrm{cov}(\boldsymbol{\theta}_i,\boldsymbol{F}_j)=\\ &(a_{i1},a_{i2},\cdots a_{im})\mathrm{cov}(\boldsymbol{F},\boldsymbol{F})(0,\cdots,1,0,\cdots,0)^{\mathrm{T}}=\\ &a_{ij}。\end{aligned}\quad(6)$$

2）根据式（4）及条件ⅱ）~ⅳ）有

$$\begin{aligned}&\mathrm{cov}(\boldsymbol{Z}_i,\boldsymbol{Z}_i)=\mathrm{cov}\left(\sum_{k=1}^{m}a_{ik}\boldsymbol{F}_k+\boldsymbol{\theta}_i,\sum_{k=1}^{m}a_{ik}\boldsymbol{F}_k+\boldsymbol{\theta}_i\right)=\\ &\mathrm{cov}[(a_{i1},a_{i2},\cdots,a_{im})\boldsymbol{F},(a_{i1},a_{i2},\cdots,a_{im})\boldsymbol{F}]+\mathrm{cov}(\boldsymbol{\theta}_i,\boldsymbol{\theta}_i)=\\ &(a_{i1},a_{i2},\cdots,a_{im})\mathrm{cov}(\boldsymbol{F},\boldsymbol{F})(a_{i1},a_{i2},\cdots,a_{im})^{\mathrm{T}}+\sigma_i^2=\\ &a_{i1}^2+a_{i2}^2+\cdots+a_{im}^2+\sigma_i^2=1。\end{aligned}$$

3）根据条件ⅳ）和式（6）~（7）得

$$r_{ij}=\frac{\mathrm{cov}(\boldsymbol{Z}_i,\boldsymbol{F}_j)}{\sqrt{\mathrm{cov}(\boldsymbol{Z}_i,\boldsymbol{Z}_i)}\sqrt{\mathrm{cov}(\boldsymbol{F}_j,\boldsymbol{F}_j)}}=a_{ij}。\quad(8)$$

4）由式（5）求$\boldsymbol{Z}$的协方差矩阵，可得

$$\mathrm{cov}(\boldsymbol{Z},\boldsymbol{Z})=\boldsymbol{A}\mathrm{cov}(\boldsymbol{F},\boldsymbol{F})\boldsymbol{A}^{\mathrm{T}}+\mathrm{cov}(\boldsymbol{\theta},\boldsymbol{\theta})。\quad(9)$$

根据等式

$$\mathrm{cov}(\boldsymbol{Z},\boldsymbol{Z})=\frac{1}{p-1}\boldsymbol{ZZ}^{\mathrm{T}}=\boldsymbol{R},\quad(7)$$

$$\mathrm{cov}(\boldsymbol{F},\boldsymbol{F})=\mathrm{diag}(1,1,\cdots,1),$$

$$\mathrm{cov}(\boldsymbol{\theta},\boldsymbol{\theta})=\mathrm{diag}(\sigma_1^2,\sigma_2^2,\cdots,\sigma_n^2),$$

将式（9）化为

$$\boldsymbol{R}=\boldsymbol{AA}^{\mathrm{T}}+\mathrm{diag}(\sigma_1^2,\sigma_2^2,\cdots,\sigma_n^2)。\quad(10)$$

推论1　在模型$\boldsymbol{Z}=\boldsymbol{AF}+\boldsymbol{\theta}$中，对于任意$m$阶正交矩阵$\boldsymbol{T}$，若$\boldsymbol{A}$=$\boldsymbol{A}^*\boldsymbol{T}$，则有

$$\boldsymbol{R}=\boldsymbol{A}^*\boldsymbol{A}^{*\mathrm{T}}+\mathrm{diag}(\sigma_1^2,\sigma_2^2,\cdots,\sigma_n^2)。$$

定理2　在式（10）中，令

$$\boldsymbol{R}^*=\boldsymbol{R}-\mathrm{diag}(\sigma_1^2,\sigma_2^2,\cdots,\sigma_n^2),$$

$$h_j=\boldsymbol{A}_j^{\mathrm{T}}\boldsymbol{A}_j\ (j=1,2,\cdots,m),$$

若满足条件$h_1\geqslant h_2\geqslant\cdots\geqslant h_m$，则有如下结论：

1）矩阵$\boldsymbol{R}^*$的特征值λ_i=h_i（i=1，2，$\cdots$，m）；

2）相应特征值λ_i的单位特征向量

$$\boldsymbol{t}_i=\frac{1}{\sqrt{\lambda_i}}\boldsymbol{A}_i\ (i=1,2,\cdots,m);$$

3）载荷矩阵$\boldsymbol{A}=(\sqrt{\lambda_1}\boldsymbol{t}_1,\sqrt{\lambda_2}\boldsymbol{t}_2,\cdots,\sqrt{\lambda_m}\boldsymbol{t}_m)$。

证明　令$\boldsymbol{R}^*=(r_{ij}^*)_{n\times n}$，由式（10）得

$$r_{ij}^*=\sum_{k=1}^{m}a_{ik}a_{jk}=\begin{cases}r_{ij}, & i\neq j,\\ r_{ii}-\sigma_i^2, & i=j,\end{cases}\ i,j=1,2,\cdots,n。\quad(11)$$

首先考虑在满足条件$r_{ij}^*=\sum_{k=1}^{m}a_{ik}a_{jk}$时，使$h_1=\boldsymbol{A}_1^{\mathrm{T}}\boldsymbol{A}_1$取得最大值，求向量$\boldsymbol{A}_1=(a_{11},a_{21},\cdots,a_{n1})^{\mathrm{T}}$的情形。

根据拉格朗日乘数法，构造目标函数

$$L(a_{11},a_{21},\cdots,a_{n1})=$$

$$\boldsymbol{A}_1^{\mathrm{T}}\boldsymbol{A}_1+\sum_{i=1}^{n}\sum_{j=1}^{n}u_{ij}\left(r_{ij}^{*}-\sum_{k=1}^{m}a_{ik}a_{jk}\right)。\quad(12)$$

由于$\boldsymbol{R}^{*}=\left(r_{ij}^{*}\right)_{n\times n}$是对称矩阵，在式（12）中有$u_{ik}=u_{ki}$，分别对$a_{i1}(i=1,2,\cdots,n)$求偏导数得

$$\frac{\partial L}{\partial a_{i1}}=2a_{i1}-2\sum_{j=1}^{n}u_{ij}a_{j1}=0\quad(i=1,2,\cdots,n),\quad(13)$$

$$\frac{\partial L}{\partial a_{ij}}=-2\sum_{k=1}^{n}u_{ik}a_{kj}=0\quad(j\neq 1;\ i=1,2,\cdots,n)。(14)$$

将式（13）和（14）改写为

$$\sum_{k=1}^{n}u_{ik}a_{kj}-\tau_{1j}a_{i1}=0(\ i=1,2,\cdots,n;\ j=1,2,\cdots,m),\quad(15)$$

式中$\tau_{1j}=\begin{cases}1, j=1,\\ 0, j\neq 1。\end{cases}$

用a_{i1}乘式（15）并对i求和，得

$$\sum_{k=1}^{n}\left(\sum_{i=1}^{n}u_{ik}a_{i1}\right)a_{kj}-\tau_{1j}\sum_{i=1}^{n}a_{i1}^{2}=0。\quad(16)$$

结合式（13），将式（16）化为

$$\sum_{k=1}^{n}a_{k1}a_{kj}-\tau_{1j}h_1=0(j=1,2,\cdots,m)。\quad(17)$$

用a_{ij}乘式（17）并对j求和，得

$$\sum_{k-1}^{n}a_{k1}\left(\sum_{j-1}^{m}a_{ij}a_{kj}\right)-\sum_{j-1}^{m}\tau_{1j}h_1a_{ij}=0。\quad(18)$$

结合式（11），由式（18）得

$$\left(r_{i1}^{*},r_{i2}^{*},\cdots,r_{in}^{*}\right)\begin{bmatrix}a_{11}\\ a_{21}\\ \vdots\\ a_{m1}\end{bmatrix}=h_1a_{i1}\ \ (i=1,2,\cdots,n)。\quad(19)$$

将式（19）的n个等式用矩阵表示即为

$$\left(\boldsymbol{R}^{*}-h_1\boldsymbol{I}\right)\boldsymbol{A}_1=\boldsymbol{0}。\quad(20)$$

式（20）表明，h_1是矩阵$\boldsymbol{R}^{*}$的最大特征值，$\boldsymbol{A}_1$是相应于h_1的特征向量，若设$\boldsymbol{R}^{*}$的最大特征值为λ_1，相应的单位特征向量为$\boldsymbol{t}_1$，则可得定理2中的结论：

$\lambda_1=h_1,\boldsymbol{t}_1=\dfrac{1}{\sqrt{\lambda_1}}\boldsymbol{A}_1$，或改写为$\boldsymbol{A}_1=\sqrt{\lambda_1}\boldsymbol{t}_1$。

由$\boldsymbol{R}^{*}$的谱分解式

$$\boldsymbol{R}^{*}=\boldsymbol{A}_1\boldsymbol{A}_1^{\mathrm{T}}+\sum_{i=1}^{m}\boldsymbol{A}_i\boldsymbol{A}_i^{\mathrm{T}}=\lambda_1\boldsymbol{t}_1\boldsymbol{t}_1^{\mathrm{T}}+\sum_{i=2}^{m}\lambda_i\boldsymbol{t}_i\boldsymbol{t}_i^{\mathrm{T}},$$

同理可得矩阵$\boldsymbol{R}^{*}$由大到小的其余$m-1$个特征值$\lambda_2\geqslant\lambda_3\geqslant\cdots\geqslant\lambda_m$，其中$\lambda_j=h_j\ \ (j=2,3,\cdots,m)$，以及相应的单位特征向量$\boldsymbol{t}_j=\dfrac{1}{\sqrt{\lambda_j}}\boldsymbol{A}_j$，或改写为$\boldsymbol{A}_j=\sqrt{\lambda_j}\boldsymbol{t}_j\ \ (j=2,3,\cdots,m)$。

定理2证毕。

由于样本相似矩阵$\boldsymbol{R}$是已知的，特殊因子$\boldsymbol{\theta}_i$的方差σ_i^2通常是一个待估计的量，等式

$$\boldsymbol{R}^{*}=\boldsymbol{R}-\mathrm{diag}\left(\sigma_1^2,\sigma_2^2,\cdots,\sigma_n^2\right)$$

中的矩阵$\boldsymbol{R}^{*}$为未知。因此定理2中由条件$\boldsymbol{R}^{*}=\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$求载荷矩阵$\boldsymbol{A}$不可行。在实际应用中通常将$\boldsymbol{R}^{*}$作近似处理，将由条件$\boldsymbol{R}^{*}=\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$求载荷矩阵$\boldsymbol{A}$的问题转化为在条件$\boldsymbol{R}\approx\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$下，由协方差矩阵$\dfrac{1}{p-1}\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}=\boldsymbol{R}$求载荷矩阵$\boldsymbol{A}$的近似矩阵$\boldsymbol{A}^{*}$，再利用回归的思想估计特殊因子的方差。因为样本离差矩阵$\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}$是一个$n$阶的对称矩阵，而$\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}$是一个$p$阶对称矩阵，样本数$n$通常远大于样本所具有的属性个数$p$，因此求$n$阶标准矩阵$\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}$的特征值和特征向量的计算量较求$p$阶矩阵$\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}$的特征值和特征向量复杂得多。为此，由矩阵$\dfrac{1}{p-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}$的特征值及特征向量得到$\dfrac{1}{p-1}\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}$的特征值及特征向量，即可减少求载荷矩阵$\boldsymbol{A}$的近似矩阵$\boldsymbol{A}^{*}$的计算量。

推论2 设矩阵$\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}$的$m$个非零特征值为$\lambda_i\ \ (i=1,2,\cdots,m)$，其排列的顺序由大到小，相应于第$i$个特征值$\lambda_i$的单位特征向量为$\boldsymbol{\eta}_i$。对于样本相似矩阵$\boldsymbol{R}=\dfrac{1}{p-1}\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}$，则有如下结论：

1）$u_i=\dfrac{\lambda_i}{p-1}\ \ (i=1,2,\cdots,m)$是矩阵$\boldsymbol{R}$的特征值；

2）矩阵$\boldsymbol{R}$与矩阵$\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}$具有相同的单位特征向量$\xi_i$，且$\xi_i=\dfrac{1}{\sqrt{\lambda_i}}\boldsymbol{Z}\boldsymbol{\eta}_i\ \ (i=1,2,\cdots,m)$；

3）载荷矩阵

$$\boldsymbol{A}^{*}=\left(\frac{1}{\sqrt{p-1}}\boldsymbol{Z}\boldsymbol{\eta}_1,\frac{1}{\sqrt{p-1}}\boldsymbol{Z}\boldsymbol{\eta}_2,\cdots,\frac{1}{\sqrt{p-1}}\boldsymbol{Z}\boldsymbol{\eta}_m\right)。$$

由引理3和定理2易证，故略去证明。

3 实例

例1 10名学生的数学与语文考试成绩见表1。

表1 学生成绩

Table 1 Student's score

样本序号	数学 y_{i1}	语文 y_{i2}	平均 $\overline{Y}_i$	样本序号	数学 y_{i1}	语文 y_{i2}	平均 $\overline{Y}_i$
1	77	64	70.5	6	66	52	59.0
2	67	65	66.0	7	77	72	74.5
3	80	74	77.0	8	83	61	72.0
4	74	84	79.0	9	86	41	63.5
5	78	62	70.0	10	65	84	74.5

以这 10 名学生作为样本观测点，共 10 个样本。第 i 个样本用 $\boldsymbol{Y}_i$ 表示，它是由数学成绩 y_{i1} 与语文成绩 y_{i2} 构成的二维数组，记为 $\boldsymbol{Y}_i=(y_{i1},y_{i2})^{\mathrm{T}}\ (i=1,2,\cdots,10)$。试

$$\boldsymbol{Z}=\begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix}^{\mathrm{T}},$$

从而得 $\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}=\begin{bmatrix} 5 & -5 \\ -5 & 5 \end{bmatrix}$。

矩阵 $\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}$ 有一个非零特征值 $\lambda=10$，对应的单位特征向量 $\boldsymbol{\eta}=\begin{bmatrix} -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix}^{\mathrm{T}}$；根据矩阵 $\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}$ 与矩阵 $\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}$ 特征值及特征向量之间的关系，得矩阵 $\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}$ 的单位特征向量

$$\boldsymbol{\xi}=\frac{1}{\sqrt{\lambda}}\boldsymbol{Z}\boldsymbol{\eta}=\frac{1}{\sqrt{10}}\begin{bmatrix} -1 & -1 & -1 & 1 & -1 & -1 & -1 & -1 & -1 & 1 \end{bmatrix}^{\mathrm{T}}。$$

于是得载荷矩阵

$$\boldsymbol{A}=\begin{bmatrix} -1 & -1 & -1 & 1 & -1 & -1 & -1 & -1 & -1 & 1 \end{bmatrix}^{\mathrm{T}},$$

提取的公共因子只有一个，即 $\boldsymbol{F}_1$，样本 $\boldsymbol{Z}=(\boldsymbol{Z}_1,\boldsymbol{Z}_2,\cdots,\boldsymbol{Z}_{10})^{\mathrm{T}}$ 关于公共因子 $\boldsymbol{F}_1$ 和特殊因子变量 $\boldsymbol{\theta}$ 的数学模型表示为 $\boldsymbol{Z}=\boldsymbol{A}\boldsymbol{F}_1+\boldsymbol{\theta}$。

根据定理 1 可知，第 i 个样本与公共因子 $\boldsymbol{F}_1$ 的相关系数 $r_{i1}=a_{i1}\ \ (i=1,2,\cdots,10)$ 见表 2。

表 2　相关系数分布

Table 2　The distribution of correlation coefficient

样本 $\boldsymbol{Z}_i$	$\boldsymbol{Z}_1$	$\boldsymbol{Z}_2$	$\boldsymbol{Z}_3$	$\boldsymbol{Z}_4$	$\boldsymbol{Z}_5$	$\boldsymbol{Z}_6$	$\boldsymbol{Z}_7$	$\boldsymbol{Z}_8$	$\boldsymbol{Z}_9$	$\boldsymbol{Z}_{10}$
相关系数 r_{i1}	−1	−1	−1	1	−1	−1	−1	−1	−1	1

表 2 表明，$\boldsymbol{Z}_4$ 和 $\boldsymbol{Z}_{10}$ 相关于公共因子 $\boldsymbol{F}_1$ 的正方向；$\boldsymbol{Z}_1$, $\boldsymbol{Z}_2$, $\boldsymbol{Z}_3$, $\boldsymbol{Z}_5$, $\boldsymbol{Z}_6$, $\boldsymbol{Z}_7$, $\boldsymbol{Z}_8$, $\boldsymbol{Z}_9$ 相关于公共因子 $\boldsymbol{F}_1$ 的反方向。因此，可将这 10 名学生分为 2 类，第一类由 4 号与 10 号学生组成；第二类为余下的 8 名学生组成。公共因子 $\boldsymbol{F}_1$ 的正方向表明学生的语文成绩优于数学成绩，反方向表明学生的语文成绩劣于数学成绩。

4　结语

对于 p 种属性的 n 个样本，样本相似矩阵 $\boldsymbol{R}=\frac{1}{p-1}\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}$ 是一个 n 阶的高阶对称矩阵，矩阵 $\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}$ 用因子分析法，按样本与因子相似的程度将这 10 名学生分类，且作出合理的解释。

解　由 $z_{ij}=\frac{y_{ij}-\overline{Y}_i}{\sqrt{v_{ii}}}$ 将原始数据标准化得矩阵

$$\boldsymbol{Z}=\begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix}^{\mathrm{T}},$$

的阶是一个比 $\boldsymbol{R}$ 矩阵阶数低得多的 p 阶对称矩阵。由推论 2 获得求载荷矩阵 $\boldsymbol{A}^*$ 的方法，其计算的复杂程度远低于通过 $\boldsymbol{R}$ 矩阵获得载荷矩阵 $\boldsymbol{A}^*$ 的方法。

参考文献：

[1] 何晓群. 多元统计分析[M]. 北京: 中国人民大学出版社，2012：142–144.
He Xiaoqun. Multivariate Statistical Analysis[M]. Beijing：China Renmin University Press，2012：142–144.

[2] 虞　欣，郑肇葆. 基于Q 型因子分析的训练样本的选择[J]. 测绘学报，2007，36(1)：67–71.
Yu Xin，Zheng Zhaobao. Selection of Training Samples Based on Q-Factor Analysis[J]. Acta Geodaetica et Cartographica Sinica，2007，36(1)：67–71.

[3] 殷瑞飞，朱建平. 关于利用因子分析方法对变量分类的探讨[J]. 统计与决策，2005(2)：20–21.
Yin Ruifei，Zhu Jianping. Using the Factor Analysis Method for the Classification Variables[J]. Statistics and Decision，2005(2)：20–21.

[4] 张秋瑾. 主成分分析法在多变量变动分析中的应用[J]. 数学的实践与认识，2012，42(17)：29–33.
Zhang Qiujin. The Application of Principal Component Analysis Method in Multivariate Analysis of Changes[J]. Mathematics in Practice and Theory，2012，42(17)：29–33.

[5] Ramsay J Q. Functional Components of Variation in Handwriting[J]. Journal of the American Statistic Association，2000，95(449)：9–15.

[6] 周　勇，朱　砾. 线性代数[M]. 上海: 复旦大学出版社，2012：129–131.
Zhou Yong，Zhu Li. Linear Algebra[M]. Shanghai：Fudan University Press，2012：129–131.

（责任编辑: 邓光辉）