多元统计分析在宏观经济分析中的应用

谢江宏[1],李雪梅[2,3],王生原[2]
(1.山西省电力公司 科技信息部,山西 太原 030001;
2.清华大学 计算机科学与技术系,北京 100084;
3.山西大学 工程学院信息工程系,山西 太原 030013)

摘 要:研究多元统计分析的理论,利用主成分分析和聚类分析的方法对区域经济指标体系进行分析和综合,找出实质体的数量特征和内在统计规律性。通过实际的历史数据进行演算,证实与当时的客观实际情况相吻合,为决策部门衡量本地区的经济发展,制定科学决策提供了有利的支持。

关键词:多元统计分析;主成分分析;聚类分析
中图分类号:TP311 **文献标识码**:A

Application of multivariate statistical analysis in the macro-economy analysis

XIE Jiang Hong[1], LI Xue Mei[2,3], WANG Sheng Yuan[2]
(1.Department of Science & Information Technology, Shanxi Electric Power Co., Taiyuan 030001, China;
2.Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China;
3.Department of Information Engineering, Engineering College of Shanxi University, Taiyuan 030013, China)

Abstract: With the theory of multivariate statistical analysis, principal components analysis and cluster analysis are applied into the analysis and synthesis of the indices system of the regional economy in order to find the quantity feature and inherent statistical characteristic of the studied objects. The operation on real historical data shows a result that conforms to the situation in that time, which is very helpful for the local government to evaluate the development of area economy and make a decision scientifically.

Key words: multivariate statistical analysis; principal components analysis; cluster analysis

统计方法是科学研究的一种重要工具,其应用颇为广泛。在工业、农业、经济、生物和医学等领域的实际问题中,常常需要处理多个变量的观测数据。因此,对多个变量进行综合处理的多元统计分析方法显得尤为重要。随着电子计算机技术的普及,以及社会、经济和科学技术的发展,过去被认为具有数学难度的多元统计分析方法,已越来越广泛地应用于实际工作中。

主成分分析[1]是一种常用的多元统计分析方法,相对于其他统计学方法,更强调用数据本身来指导分析过程,而不是依赖事先给定的某些假设。主要目的是希望用较少的变量解释原始资料中的大部份变异,期望能将许多相关性很高的变量转化成彼此互相独立的变量,从中选取较原始变量个数少且能解释大部份资料中变异的几个新变量(降低原始变量的维数),也就是所谓的主成分,而这几个主成分也就成为用来解释资料的综合性指标。

聚类分析[2]是研究事物分类的一种方法,是认识和探索事物内在联系的一种手段。聚类分析源于许多研究领域,包括数据挖掘、统计学、机器学习和模式识别等,并作为一个独立的工具来获得数据分布的情况,概括出每个簇的特点,或者集中注意力对特定的某些簇进行分析。聚类就是将数据对象分组成为多个类或簇,划分的原则是在同一个簇中的对象之间有高度的相似度,而不同簇中的对象差别较大。聚类分析通常被用作最初的分析工具,可以使数据挖掘具备识别群这一功能,它的流程通常是首先对数据进行图形描述,再用量化方法来描

述数据的特征。

1 设计思想

1.1 主成分分析

主成分分析主要应用于简化观测系统，将原始因子变换为新因子，把多个单项指标转化为最少数量的综合指标。其设计思想[3]是通过对每个变量的实际观测值的协方差矩阵进行计算，依次提取方差贡献最大的各个主成分，以达到选择、浓缩和提炼变量的目的。主成分分析中的因子分析所涉及的计算与此类似，是研究一组样品之间的相关关系的一种统计方法，即对于一组具有复杂的相关关系的样品，可以通过研究其相关矩阵的内部结构，找出若干个对这组样品起着支配作用的独立的新因子(实际上是原始变量在通常的、或者是最小二乘意义上的线性组合)，用这些独立的新因子(称为公因子或主因子的数目往往比原始变量的数目要少)来表达所有观测数据，既极少损失总的关于原始变量的相关信息，又合理解释了包含在原始变量(样品)的相关性，简化了观测系统，抓住了影响所有观测数据的主要矛盾。

传统的一些综合评价方法在选择权数时有很大的主观随意性，而用主成份方法综合评价经济效益，既避免了信息量的重复，又克服权数选择的人为性。可以方便地得到全面、客观的评价结果。此方法已被我国许多统计工作者应用到实际工作中，正在产生积极的效果。

1.2 聚类分析

聚类分析的思想来自于方差分析，是由 Ward 于 1936 年提出，1967 年经 Orloci 等人发展建立起来的一种系统聚类方法[4-5]。具体做法是在一批样品的多个观测指标中找出能度量样品(或指标)之间相似程度的统计量，构成一个对称的相似性矩阵，在此基础上进一步找寻各样品(或指标)之间或样品组合之间的相似程度，按相似程度的大小把样品(或指标)逐一归类，进行比较。具体做法就是先将 N 个样品各自视为一类，然后计算确定样本之间、类与类之间的距离，选择距离最小的一对样本合并成一个新类，计算包括新类在内的其余各类的距离，再将距离最近的两类合并，这样每次减少一类，直至所有的样品都成为一类为止。

在宏观经济的分析研究中根据经济指标体系的多个指标值，找出一些能够度量样品相似程度的指标，以这些指标为划分类型的依据，使一些相似程度较大的区域聚合为一类，再将另一些彼此之间相似程度较大的聚合为另一类，直到把所有区域都聚合完毕，形成一个由小到大的分类系统，最后将整个分类系统绘成一张聚类图，并结合因子分析的评价结果和实际情况具体分类。

2 数学模型及算法实现

2.1 主成分分析

选择所确立的宏观经济指标作为样品的原始数据组成矩阵，设有 N 个地区，并各观测 P 个指标变量，其原始矩阵为：

$$X=\begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ & \cdots & \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

其中，x_{ij} 表示第 i 个地区的第 j 个指标的值($i=1,2,\cdots,n;j=1,2,\cdots,p$)。

对原始数据进行标准化处理，形成标准化矩阵：

$$x_{ij}'=(x_{ij}-\bar{x}_j)/s_j \quad (i=1,2,\cdots,p;j=1,2,\cdots,n)$$

其中：

均值 $\bar{x}_j=\frac{1}{n}\sum_{i=1}^{n}x_{ij} \quad (j=1,2,\cdots,p)$

方差 $s_j=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2} \quad (j=1,2,\cdots,p)$

假定经过变换后为 X，则 X 的元素 x_{ij} 的均值为 0，方差为 1，各单项指标具有相同度量尺度和一致的变化范围。然后计算相关矩阵及其特征根，选取主成分。

相关矩阵 R 是一个对称矩阵 $R=(r_{ij})$

其中，$r_{ij}=\frac{1}{n}\sum_{k=1}^{n}x_{ki}x_{kj} \quad (i,j=1,2,\cdots,p)$

在此基础上利用雅可比法求 R 的全部特征根 λ_i(由大到小排列)及相应的特征向量 a_i，全部特征根 $\lambda_1>\lambda_2>\cdots>\lambda_p$ 均大于等于零，算出每一特征值对总体方差的贡献率及累积贡献率总和为 1。

根据实际情况确定累积贡献率大于某一特定值来确定主成分个数 m，这样就由若干个单项指标变换得到个别几项综合指标。

假定选取了 $m=2$ 个公因子之后，则取载荷矩阵的前 $m=2$ 列为初始因子载荷矩阵 A，计算初始因子载荷矩阵：

$$A=\begin{pmatrix} a_{11} & \cdots & a_{1m} \\ a_{21} & \cdots & a_{2m} \\ & \cdots & \\ a_{n1} & \cdots & a_{nm} \end{pmatrix}$$

其中，$a_{ij}=a_i\sqrt{\lambda_j} \quad (i=1,2,\cdots,n;j=1,\cdots,m)$

初始因子载荷矩阵 A 的因子载荷 a_{ij} 反映的是初始因子(变量 i)对新因子(主成分 j)的载荷强度(即相关程度)，为了促使初始因子(变量 i)对新因子(主成分 j)上的载荷分布向 0 或 1 两极分化，使处于中间状态的载荷强度趋于消失，需要对载荷矩阵 A 实施方差极大化旋转，从而获得经济含义鲜明的主特征分析。所以对载荷矩阵进行方差极大化正交旋转，首先计算公共因子方差：

$$h_i^2=\sum_{j=1}^{m}a_{ij}^2 \quad (i=1,2,\cdots,n;j=1,2)$$

之后求正交因子解——方差极大正交因子旋转，用 h_i 除 A 的各个元素将因子载荷矩阵正规化，再将 m 个因子轴

两两组合进行旋转，共旋转 $m(m-1)/2$ 次。第 r 个和第 s 个公共因子旋转后的载荷由 $B=A\times T$ 决定。

其中：

$$T=\begin{pmatrix}\cos\varphi & -\sin\varphi\\ \sin\varphi & \cos\varphi\end{pmatrix}$$

$$\mathrm{tg}\,4\varphi=r'/\delta$$

$$r'=2\sum_{i=1}^{n}2a_{ir}a_{i\delta}(a_{ir}^{2}-a_{ir}^{2})^{2}-2\sum_{i=1}^{p}2a_{ir}a_{i\delta}\sum_{i=1}^{p}(a_{ir}^{2}-a_{ir}^{2})/p$$

$$\delta=\sum_{i=1}^{n}2a_{ir}a_{i\delta}[(a_{ir}^{2}-a_{i\delta}^{2})^{2}]-(a_{ir}a_{i\delta})^{2}-\left\{\left[\sum_{i=1}^{k}(a_{ir}^{2}-a_{i\delta}^{2})\right]^{2}-\left(\sum_{i=1}^{n}(2a_{ir}a_{i\delta})\right)^{2}\right\}/p$$

$$(r=1,2,\cdots,(m-1),s=r+1,\cdots,m)$$

得正交变换 $T_1=T_{12}\ T_{13}\ \cdots\ T_{rs}\ \cdots\ T_{(m-1)m}$，旋转因子载荷矩阵 $B_1=A\times T_1$ 及因子载荷平方的方差为：

$$V=\frac{1}{p}\sum_{j=1}^{m}\sum_{i=1}^{n}(b_{ij}/h_i)^4-\sum_{j=1}^{m}\left(\sum_{i=1}^{n}b_{ij}^{2}/h_i^{2}\right)^{2}$$

以 B_1 作为新的初始因子载荷矩阵，重复上一步，直至最后 2 次的 V 之差绝对值小于所要求的精度要求为止。将最后求得的旋转因子载荷矩阵进行正规化还原，得 $G=(b_{ij}h_i)$ 即为所求的正交因子解。

通过以上方法求得各单项指标对综合指标的相关程度的相关系数都在 70%以上，说明这一主特征反映了国民经济发展的总体规模和实力水平，因此可称第一主特征为“总体规模”指标；第二主特征与人均水平、经济效益指标明显相关，而与总量指标相关程度较小。说明这一主特征反映了国民经济发展中的经济社会效益状况，因此第二主特征称为“综合效益”指标。

利用公式计算各地区总体规模和综合效益得：

$F=AR^{-1}X$（其中，A 为载荷矩阵，R 为相关矩阵，X 为原始矩阵）可计算出各个样本（$N=28$ 个地区）在新因子（$P=31$ 个指标）上的得分，根据得分排出如图 1 所示的位次。从图中可以很直观地看出某一地区在总体规模和综合效益两方面在全国所处的位置，根据某一时间段内位次的变化就可以分析其中的原因。

2.2 聚类分析

选择宏观经济指标作为样品的原始数据组成矩阵，设有 N 个样品，并对其观测 P 个变量，其原始矩阵为：

$$X=\begin{pmatrix}x_{11} & \cdots & x_{1n}\\ x_{21} & \cdots & x_{2n}\\ & \cdots & \\ x_{p1} & \cdots & x_{pn}\end{pmatrix}$$

其中，x_{ij} 表示第 i 个变量的第 j 个观测值（$i=1,2,\cdots,p$；$j=1,2,\cdots,n$）。

对原始数据进行标准化，形成标准化矩阵

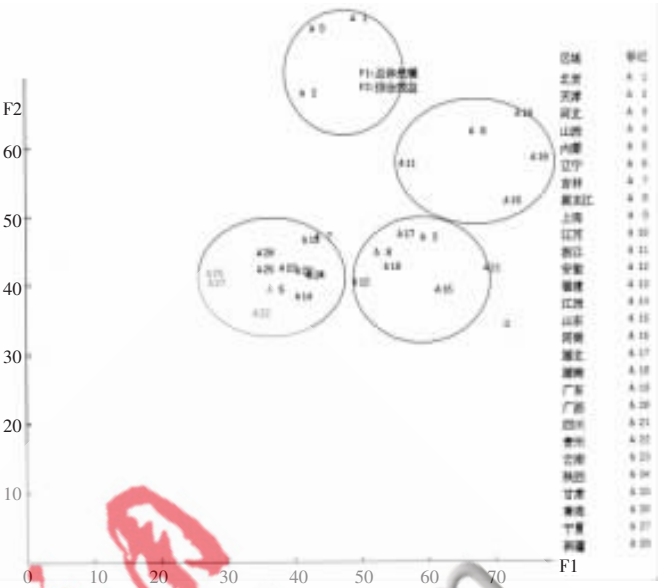

图 1 主成分图

$$x_{ij}=(x_{ij}-x_i')/s_i\quad(i=1,2,\cdots,p;j=1,2,\cdots,n)$$

其中：

$$x_i'=\frac{1}{N}\sum_{i=1}^{\ell}x_{ij}$$

$$s_i=\sqrt{\frac{1}{N-1}\sum_{j=1}^{N}(x_{ij}-x_i')^{2}}$$

计算相似系数矩阵，选出最大相似系数样品组。把对应的一组样品加权平均：

$$x_{ij}=(n_1x_{j1}+n_2x_{j2})/(n_1+n_2)$$

形成一个新的样品点，其中 n_1、n_2 分别为已组合过的样品组中样品的个数，x_{j1}、x_{j2} 为相应的数据。用新的样品点代替原来的一对样品点。

对新形成的样品数据与其余样品数据重新计算相似矩阵，以代替原相似矩阵，再找出新相似矩阵中最大系数的对应样品组。如此重复以上步骤，直到将所有样品都归类完毕为止。

按下列原则连接成谱系图：

(1)若 2 个样品在已经连结成组的组中没有出现过，则把它们联结成 1 个新组。

(2)若 2 个样品中有 1 个在某组中出现过，另一个就加入该组。

(3)若 2 个样品都在同一组中，这对样品不再分组。

(4)若 2 个样品都已在不同组中出现过，则把 2 组连接在一起。

通过选用 28 个地区的实际数据，并利用此方法进行聚类分析，得出了对总体规模指标和综合效益指标进行综合后的聚类谱系图，如图 2 所示。

本文是通过主成分分析和聚类分析对我国区域经济发展进行比较分析的一个应用实例，采用了上世纪

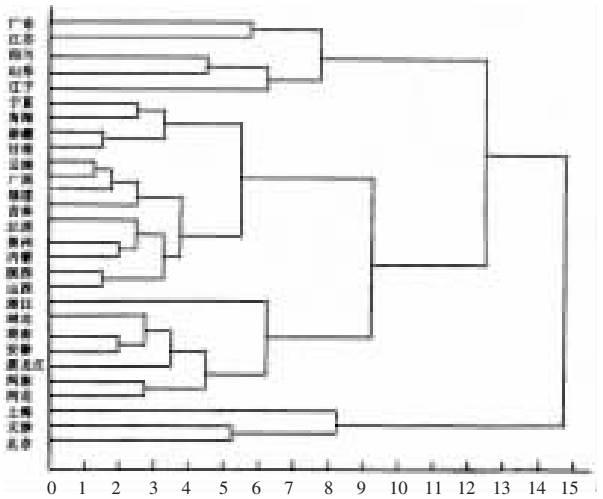

图 2 宏观经济聚类分析图

80 年代的历史数据进行了比较分析，从结果中可以看出完全符合当时我国区域宏观经济的发展状况。总之，应用主成分分析和聚类分析的方法可以在运算结果的基础上，对各省区经济发展战略模式和经济发展总体水平进行综合性的比较、分析和评价，为制定决策提供科学的依据。该方法一直在区域宏观经济分析系统中使用，也得到了有关方面的一致好评。

参考文献

[1] 张润楚.多元统计分析.北京：科学出版社，2006.

[2] 朱玉全，杨鹤标，孙蕾，等.数据挖掘技术[M].南京：东南大学出版社，2006.

[3] 何晓群.现代统计分析方法与应用.北京：中国人民大学出版社，1998.

[4] 雷钦礼.经济管理多元统计分析[M].北京：中国统计出版社，2002.

[5] 陈正昌，多变量分析方法[M].北京：中国税务出版社，2005.

（收稿日期：2009-01-05）