

主成分和因子分析在城镇消费支出中的应用比较

荣文静

(成都理工大学信息管理学院, 四川 成都 610059)

摘要:主成分分析和因子分析都是简化数据结构(降维问题)的重要方法,二者既有区别也有联系。文章将从数据分析的角度,采用信息的概念,介绍这两种方法;同时,通过这两种方法对全国城镇消费支出数据资料进行分析和比较,以便在实际应用时选择更实用更合理的方法,对经济现象进行更有效的分析和评价。

关键词:主成分分析;因子分析;城镇消费;SPSS

中图分类号:O 213

文献标识码:A

引言

在对全国城镇消费支出数据进行分析时,为了尽可能全面反映评价对象的整体情况,需要选取恰当的、客观的评价指标。由于城镇消费支出数据的复杂性,综合评价通常涉及到多指标,这不仅会增加评价的工作量,而且会因评价指标间的相关性造成评价信息相互干扰,从而难以客观地反映评价对象的真实属性。在实际工作中,需要精简指标,将原来的指标重新组合成一组相互无关的综合指标以此来尽可能多地反映原来指标信息量,主成分分析与因子分析为解决此类问题提供了很好的方法。究竟选择那种分析方法更加理想呢,本文就这个问题结合实例进行了分析比较。

1 主成分分析和因子分析原理

设 p 维随机向量 $X' = (X_1, X_2, \dots, X_p)$ 的协方差矩阵为 Σ ,

$$\Sigma = \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) & \dots & Cov(X_1, X_p) \\ Cov(X_2, X_1) & Var(X_2) & \dots & Cov(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_p, X_1) & Cov(X_p, X_2) & \dots & Var(X_p) \end{pmatrix}$$

$$= \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp} \end{pmatrix}$$

Σ 的 p 个特征值为 $\lambda_1, \lambda_2, \lambda_3, \dots, \lambda_p$ ($\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \dots \geq \lambda_p$), 对应的 p 个单位特征向量为 e_1, e_2, \dots, e_p 。

1.1 主成分分析

主成分分析是多元统计分析中简化数据结构(降维问题)的一种重要方法,是指将某些较复杂的数据结构中相互依赖的变量变成互不相关的,使问题得到简化而损失的信息又不太多。

设样本数据矩阵为:

$$X = (x_1, x_2, \dots, x_n)' = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

则样本协方差矩阵 S 和样本相关矩阵 R 分别为:

$$S = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})' = (s_{ij})_{p \times p}$$

$$R = (r_{ij})_{p \times p}$$

其中

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p)'$$

$$s_{ij} = \frac{1}{n-1} \sum_{i=1}^n (x_{ii} - \bar{x}_i)(x_{ii} - \bar{x}_i)'$$

$$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}} \sqrt{s_{jj}}} \quad (i, j = 1, 2, \dots, p)$$

我们可以分别用 S 作为 Σ 的估计或者用 R 作为总体相关阵 ρ 的估计, 然后从 S 或 R 出发求主成分。

1.2 因子分析

因子是主成分分析的推广和发展, 是用少数 n 个因子去研究多个原始指标间的关系, 并尽可能合理地解释存在于原始变量间的相关性。若难以找到合理的解释, 就需要进一步进行因子旋转, 以求旋转后能得到合理的解释。具体步骤是:

对原始数据标准化, 建立相关系数矩阵 R , 求出其特征值和相应的特征向量, 确定公共因子的个数, 这些与主成分分析的做法是一致的。

计算初始因子载荷矩阵 $A = (\sqrt{\lambda_1}e_1, \sqrt{\lambda_2}e_2, \dots, \sqrt{\lambda_n}e_n)$, 解释因子的实际含义, 必要时对 A 实施方差最大正交旋转。估计各样本的因子得分进行样本的因子评价分析, 计算各样本的总因子得分估计值, 即综合评价:

$$Y = \sum_{i=1}^m \left[\frac{\lambda_i}{\sum_{j=1}^m \lambda_j} \right] \hat{Y}_i$$

进行样本的总体排序和比较分析。由此可以看出, 因子分析不仅可以对评价对象进行总体比较和排序, 更重要的是可以对评价对象在各公共因子所代表的方面进行评价。

2 主成分分析和因子分析在实例中的应用

用 SPSS 软件对 2008 年《中国统计年鉴》中“全国 2007 年城镇消费支出数据”的原始数据进行主成分分析和因子分析, 从降维的角度来比较两种方法的异同。

指标解释: x_1 —食品, x_2 —衣着, x_3 —家庭设备用品及服务, x_4 —医疗保健, x_5 —交通和通讯, x_6 —娱乐教育文化服务, x_7 —居住, x_8 —杂项商品和服务。

2.1 用主成分分析对数据进行分析

我们从表 1 可以取得每个主成分的方差, 其大小表示了对应主成分能够描述原来所有信息的多少。本文只提取前几个主成分, 由于前三个主成分的累计方差贡献率达到 89% 以上, 用三个新变量来代替原来的八个变量, 经进一步操作之后, 求得主成分表达式为:

$$Y_1 = 0.398 \times x_1 + 0.146 \times x_2 + 0.381 \times x_3 + 0.333 \times x_4 + 0.377 \times x_5 + 0.415 \times x_6 + 0.299 \times x_7 + 0.400 \times x_8$$

$$Y_2 = -0.099 \times x_1 + 0.697 \times x_2 + 0.159 \times x_3 + 0.362 \times x_4 - 0.274 \times x_5 + 0.052 \times x_6 - 0.519 \times x_7 - 0.018 \times x_8$$

$$Y_3 = 0.274 \times x_1 + 0.474 \times x_2 - 0.466 \times x_3 - 0.390 \times x_4 + 0.352 \times x_5 - 0.296 \times x_6 - 0.052 \times x_7 + 0.363 \times x_8$$

表 1 公因子方差

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5.169	64.609	64.609	5.169	64.609	64.609
2	1.398	17.474	82.083	1.398	17.474	82.083
3	0.561	7.014	89.097	0.561	7.014	89.097
4	0.326	4.076	93.174	0.326	4.076	93.174
5	0.267	3.343	96.517	0.267	3.343	96.517
6	0.149	1.861	98.377	0.149	1.861	98.377
7	0.075	0.936	99.313	0.075	0.936	99.313
8	0.055	0.687	100.000	0.055	0.687	100.000

第一主成分, 除 x_2 之外, 其余变量的系数都在 0.3 - 0.4 附近, 说明第一主成分是七个变量的综合, 而第二主成分主要反映了衣着消费支出 x_2 的信息, 第三主成分反映的是家庭设备用品及服务 (x_3) 和医疗服务 (x_4) 的信息。

最后进行排序。城镇消费支出综合排名前三是: 北京、上海和广东。

2.2 用因子分析对数据进行分析

通过 SPSS 软件可以得出, 经提取的三个公共因子与娱乐教育文化服务支出 (x_6) 的依赖程度最高, 总体来说公共因子与变量间的相关程度较强。

从表 2 可以看出, 旋转前第一个公共因子的方差贡献率为 64.609%, 第二个为 17.474%, 第三个为 7.014%, 旋转后方差贡献发生了变化, 但三个公共因子的重要性地位并未发生变化, 且总信息量也未发生改变。

表 2 公因子方差

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5.2	64.61	64.609	3.3	41.44	41.444
2	1.4	17.47	82.083	2.6	32.31	73.749
3	0.56	7.01	89.097	1.2	15.35	89.097

从表 3 可以看出, 旋转后因子间的差异更明显, 第一因子主要用来解释 x_1, x_5, x_7 和 x_8 集中反映食品、交通通讯、居住及服务的信息, 是生活必需公共因子。第二因子主要用来解释 x_3, x_4, x_6 集中反映了家庭设备用品、医疗保健及娱乐教育文化方面的信息, 是精神享受因素。第三个公共因子则主要用来解释 x_2 的信息, 是气候因素。

3 结束语

主成分分析和因子分析既有区别又有联系。前者是将原来多个具有一定相关性的指标变量组合成一组相互无关的综合指标来代替原指标; 因子分析是将原来多个具有相关性的指标分解成一组数量较少的不相关的因子, 以再现原变量与因子的关系。

表 3 旋转后的因子矩阵

	Component		
	1	2	3
score(x_1)	. 819	. 419	. 143
score(x_2)	. 039	. 255	. 922
score(x_3)	. 395	. 863	. 092
score(x_4)	. 228	. 834	. 304
score(x_5)	. 916	. 263	. 001
score(x_6)	. 571	. 783	. 071
score(x_7)	. 773	. 257	-. 422
score(x_8)	. 820	. 396	. 271

随着计算机的应用与发展,主成分分析和因子分析在经济、社会等方面将得到广泛的应用。本文主要是从理论和应用上对比分析主成分分析和因子分析这两种

方法,未能对主成分分析和因子分析算法提出改进。

参考文献:

- [1] 彭丽,冯维波.基于因子分析的重庆市县域综合实力评价[J].四川理工学院学报:自然科学版,2009,22(4):116-118
- [2] 米红,张文璋.实用现代统计分析方法与SPSS应用[M].北京:当代中国出版社,2000
- [3] 林海明.对主成分分析法运用中十个问题的解析[J].统计与决策:理论版,2007,(8):16-18
- [4] 理查德,约翰逊,迪安,等.实用多元统计分析[M].北京:清华大学出版社,2008
- [5] 魏艳华,王丙参,田玉柱.主成分分析与因子分析的比较研究[J].天水师范学院学报,2009,(3):13-15

Application and Comparison of Principal Components Analysis and Factor Analysis in Urban Consumption

RONG Wen-jing

(College of Information and Management, Chengdu University of Technology, Chengdu 610059, China)

Abstract The principal components and factor analysis are the important methods to simplify structure of data, and they have both differences and ties. This paper introduces these two methods from the perspective of data analysis and the concept of information. At the same time, according to the analysis and comparison in urban consumption expenditure by these two methods, it's convenient for us to choose a reasonable approach in practice, and help to make more effective analysis and evaluation to economic phenomena.

Key words principal components analysis; factor analysis; urban consumption; SPSS

(上接第 121 页)

Study of Continuous-time Portfolio Optimal Strategy with Liability

YUAN Min, LIU Xuan-hui, XUE Yun

(School of Science, Xi'an Polytechnic University, Xi'an 710048, China)

Abstract In this paper, we formulate mean-variance portfolio selection model with risky asset and liability in a complete market. The risky asset's price is driven by geometric Brownian motion with drift while the liability evolves according to a Brownian motion. The correlations between the risky asset and liability are considered. We employ stochastic optimal control theory to analytically solve the asset-liability management problem in a continuous-time setting. More specifically, we derive the optimal policy from a stochastic linear quadratic control framework.

Key words portfolio; liability; mean-variance model; stochastic linear-quadratic control