成分数据主成分分析及其应用*

张崇甫 陈述云

(重庆建筑高等专科学校,重庆,630030) (贵州省统计局,贵阳,550001)

摘 要

本文分析了传统主成分分析在成分数据分析中的不适应性,介绍了艾奇逊的中心化对数比变换和成分数据主成分分析,并以农民消费结构为例,重点讨论了成分数据主成分分析的实施步骤。

关键词: 成分数据,主成分分析,中心化对数比变换

一、引言

成分数据是指分布在有限区域内的服从单位和约束条件的数据,如产业结构、国民经济各部门投资比例、家庭生活费用支出比例等等。在成分数据的统计分析中,定和限制常被有意或无意地忽略,为无约束条件数据计设的经典统计方法被不适当地使用,导致了不可靠的推断。英国统计学家艾奇逊(J. Aichison)根据成分分量的比值不受"定和"限制的影响和比值的对数常常服从正态分布的特点,提出用成分数据分量比值(即"对数比")作为研究成分数据的一个"变换"工具,使传统的统计分析方法也能适用于成分数据的统计分析。本文就是从分析传统主成分分析处理成分数据的不适性出发,介绍艾奇逊提出的对数比统计方法和成分数据主成分分析,并结合实例提出用这一方法解决实际问题的实施步骤。

二、对传统主成分分析方法的思考

所谓(p维)成分数据是指满足单位和约束条件

$$\sum_{i=1}^{p} x_i = 1 \qquad (x_i > 0 \quad i = 1, 2, \cdots, p)$$

的一些向量($x_1, x_2, \cdots, x_p$)所组成的数据集。

用传统主成分分析方法研究高维成分数据的变异性会因单位和约束而遇到一些困难,这些困难集中表现在协方差矩阵的解释和数据的线性性两个方面。

1. 协方差矩阵的解释受到限制

传统主成分分析的出发点是协方差或相关系数矩阵。对 p维成分数据

$$x = (x_1, x_2, \cdots, x_p)$$

* 收稿日期: 1995年 4月 8日

来讲,因$\sum_{i=1}^{p} x_i = 1$,则对成分 x_i 有

$$\sum_{j=1}^{p} \mathrm{Cov}(x_i, x_j) = 0$$

即

$$\sum_{i \neq j} \mathrm{Cov}(x_i, x_j) = -\mathrm{Var}(x_i)$$

因 $Var(x_i) > 0$(除非 x_i 为常数变量),这样对某个 x_i,在 $(p-1)$个协方差 $Cov(x_i, x_j)(i \neq j)$中至少有一个是负数。也就是说,(理论)协方差矩阵

$$V = (\mathrm{Cov}(x_i, x_j))$$

中,至少有 p个协方差必为负,这种负偏性也呈现在协方差矩阵的估计之中。从而相关系数并非象无单位和约束条件时那样,均匀地分布在 $(-1, 1)$区间内,且受协方差或相关系数矩阵的非负定性所约束,这样,用传统主成分分析方法研究高维成分数据就必然会带来解释上的困难。这个负偏性问题在 Z维成分数据中特别突出。在 Z维 (x_1, x_2)情况下

$$\mathrm{Cov}(x_1, x_2) = -\mathrm{Var}(x_1) = -\mathrm{Var}(x_2)$$

因此,x_1、x_2间的相关系数

$$d = \mathrm{Cov}(x_1, x_2)\left\{\mathrm{Var}(x_1)\mathrm{Var}(x_2)\right\}^{-\frac{1}{2}} = -1$$

这表明 z维成分间的相关系数必为定值 -1,而不是在 $(-1, 1)$内均匀取值。

以上分析表明,成分数据的协方差矩阵具有明显的负偏性,截然不同于对无约束条件下的协方差矩阵的经典解释。

2.传统主成分分析是一种"线性"降维技术。

实际上,有些成分数据(不局限于成分数据)是"非线性"的,这时如果用传统主成分分析方法对这些数据进行分析并描出按变异性大小的前 z个主成分的散点图,就会发现这些数据在散点图上,明显呈现出曲线特征。这一事实表明,传统主成分分析不适宜处理"非线性"成分数据。究其原因是,在纯几何意义上,传统主成分分析是一种"线性"降维技术,表现为其主成分是原始变量的线性组合,这对于描述非线性成分数据的变异性是不合适的。

基于上述两个原因,有必要对传统主成分分析进行改造,使其适合成分数据分析。

三、成分数据主成分分析

设有 p维成分数据随机向量

$$x = (x_1, x_2, \cdots, x_p)$$

作中心化对数比变换

$$y_i = \lg(x_i / g(x)), \qquad g(x) = \sqrt[p]{x_1 x_2 \cdots x_p}$$

此时,随机向量

$$Y = (y_1, y_2, \cdots y_p)$$

可在 p维实空间 R^p 中任意取值,因为对数比变换是空间

$$\left\{(x_1, x_2, \cdots, x_p) \mid \sum_{i=1}^{p} x_i = 1, x_i > 0\right\}$$

到 R^p 的一一对应。由此可看出,中心化对数比变换有效地摆脱了成分数据的约束而进入到 R^p

的自由度中,成分数据的协方差矩阵的负偏性也消失了。这样,对成分数据 $(x_1, x_2, \cdots, x_p)$的主成分分析,就可通过对随机向量 $(y_1, y_2, \cdots y_p)$的主成分分析来实现。

在对 $Y=(y_1, y_2, \cdots y_p)$的主成分分析中,我们的兴趣集中于寻求在 $a'a=1$的条件下,线性组合 $a'Y$的方差 $Var(a'Y)$的极大值,其中

$$a=(a_1, a_2, \cdots, a_p)'。$$

容易证明

$$\mathrm{Var}(a'y)=\mathrm{Var}(a'\lg(x/g(x)))=a'\Gamma a$$

式中

$$\Gamma=[\mathrm{Cov}(\lg(x_i/g(x)), \lg(x_j/g(x))]$$

为中心化对数比协方差矩阵。

设 Γ的 p个以大小为序的特征根分别为

$$\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p$$

相应的标准化特征向量为 $a_1, a_2, \cdots, a_p$,它们满足

$$(\Gamma-\lambda_i I)a_i=0, i=1, 2, \cdots, p$$

则

$$a'_i \lg(x/g(x))$$

为第 i个主成分。显然,$a'_i lg(x/g(x))$是 x的非线性组合。

设有 p维成分向量 $x=(x_1, x_2, \cdots, x_p)$的样本资料 $(x_{ij})_{n\times p}$,根据以上分析可得成分数据主成分分析的基本步骤:

(1)对原始数据作中心化对数比变换

$$y_{ij}=\lg x_{ij}-\frac{1}{p}\sum_{t=1}^{p}\lg x_{it}$$

(2)计算中心化对数比样本协方差矩阵

$$S=(S_{ij})_{p\times p}$$

其中

$$s_{ij}=\frac{1}{n}\sum_{t=1}^{n}(y_{ti}-\bar{y}_i)(y_{tj}-\bar{y}_j)$$

$$\bar{y}_i=\frac{1}{n}\sum_{i=1}^{n}y_{ti}, \bar{y}_j=\frac{1}{n}\sum_{t=1}^{n}y_{tj}$$

(3)从 S出发求样本主成分

设 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p$是 S的 p个特征根,$a_1, a_2, \cdots, a_p$是相应的标准化特征向量,则第 i个主成分为

$$F_i=\sum_{j=1}^{p}a_{ij}\lg x_{ij}$$

余下的处理同传统的主成分分析。

四、实例分析

以全国 30个省(市、区)的农民家庭为研究对象,将其人均生活费用支出分为食品、衣着、居住、家庭设备及服务、医疗保健、交通通讯、文教娱乐及服务、其他非商品消费及服务八个部分,它们在人均生活消费支出中所占的比重分别记为 $x_1, x_2, \cdots, x_8$,则 $x=(x_1, x_2, \cdots x_8)$就构成

成分向量,因此对农民消费结构的分析应使用成分数据主成分分析方法。为了对比,先利用1993年的资料(来源于《中国统计年鉴(1994)》)作传统的主成分分析。由样本相关系数矩阵求得依大小为序的前四个特征根为

$$\lambda_1 = 3.2272, \lambda_2 = 1.8598, \lambda_3 = 1.4147, \lambda_4 = 0.5751$$

按累积变异性贡献率不低于85%为准,应选4个主成分(前4个主成分的累积变异性贡献率为88.46%),降维效果不太明显。为此,再对此问题作成分数据主成分分析。

对原始数据作中心化对数比变换,经计算发现中心化对数比协方差矩阵的最大特征根为9.8376,远远大于第二大特征根 $\lambda_2 = 0.2881$,致使第一主成分的变异性贡献率高达94.42%,降维效果极其显著。

计算得到的第一主成分为

$$\begin{aligned} F_1 &= 0.7307\lg x_1 + 0.0677\lg x_2 + 0.2553\lg x_3 \\ &\quad - 0.0382\lg x_4 - 0.2058\lg x_5 - 0.36\lg x_6 \\ &\quad + 0.0217\lg x_7 - 0.4715\lg x_8 \\ &= \lg\left(\frac{x_1^{0.7307} x_2^{0.0677} x_3^{0.2553} x_7^{0.0217}}{x_4^{0.0382} x_5^{0.2058} x_6^{0.36} x_8^{0.4715}}\right) \end{aligned} \tag{1}$$

上式表明,影响我国农民消费结构的主要原因是食品、住房、医疗保健、交通通讯、其他非商品消费等方面,其中食品支出对消费结构的影响,明显大于其它各个方面。由于其他非商品消费包括的项目多,较复杂,本文暂不分析这一因素。

从第一成分数据主成分的表达式(1)中的系数可以看出,食品与住房支出是同步的,即随着农民生活消费支出的增加,农民用于食品、住房两方面的支出或同时下降或同时上升;同时,医疗保健、交通通讯的支出也是同步的。事实上,1993年我国农民生活消费支出中食品和住房所占比重分别是58.06%和13.88%,分别比1990年下降0.74和3.46个百分点。1993年医疗保健、交通通讯分别占我国农民生活消费支出的3.54%和2.26%,分别比1990年上升0.28和0.82个百分点。这一结论充分展示了成分数据的单位和约束的事实,因而部分成分分量上升必然导致其它分量下降。而用传统主成分分析处理这一问题时,得不出这个结论。

在四个影响农民消费结构的主要因素中,食品的影响最大,这是符合我国国情的。我国是一个低收入的发展中国家,农民消费支出主要用于食品。同时,我国各地的经济发展水平极不平衡,农民的收入水平参差不齐,形成农民在食品方面的支出相差悬殊。如1993年我国农民食品消费支出比重最高的是贵州,达到71%,最低的是上海,为46.4%,高低相差达24.6个百分点。这些事实表明,食品支出是形成我国农民消费结构差异的最重要原因。

五、结语

本文初步分析了传统主成分分析在成分数据分析中的不适应性,介绍了中心化对数比变换和成分数据主成分分析的实施步骤。并以农民消费结构问题为例,说明应慎用传统主成分分析技术。

(下转第48页)

$$\frac{0.2320 - F(\Theta^{\circ}|0.8)}{F(\Theta^{\circ}|1.0) - F(\Theta^{\circ}|0.8)} = \frac{0.95 - 0.8}{1.0 - 0.8} \tag{5}$$

现 $F(\Theta^{\circ}|0.95) = 0.2320$,节取刻度参数 0.8和 1.0的 *Von Mises* 分布表(见表 1),可以判断 $\Theta^{\circ}_1 = 120^{\circ}$, $\Theta^{\circ}_2 = 130^{\circ}$,因此,(3)(4)式可以表示为

$$\frac{F(\Theta^{\circ}|0.8) - 0.2211}{0.2588 - 0.2211} = \frac{\Theta^{\circ} - 120^{\circ}}{130^{\circ} - 120^{\circ}} \tag{6}$$

$$\frac{F(\Theta^{\circ}|1.0) - 0.1957}{0.2346 - 0.1957} = \frac{\Theta^{\circ} - 120^{\circ}}{130^{\circ} - 120^{\circ}} \tag{7}$$

(5)至(7)式联立,不难求出 $\Theta^{\circ} \approx 127^{\circ}16'$

参考文献

[1] 项静恬等,动态和静态数据处理,气象出版社,1991年,第 1093至 1099页。

A Insert Value method of Von Mises Distribution Table

Li Yuan sheng

(Beijing Graduate School China University of Mining and Technology)

Abstrast

The article gives a insert value method of Von Mises distribution Table.

Key words Von Mises distribution, directional data, insert value.

(上接第 14页)

参考文献

〔1〕 J. Aitchison,成分数据的统计分析,中国地质大学出版社,1989.周蒂译.

The Principal Component Analysis on Component Data and It′t Applications

ZHang Chongfu(CHongQing Architectural Higher College)

CHen SHuyun(Statistics Bureau of Gui ZHou Province)

Abstract

The textanalysis the unadaptability of traditional principal component analysis on composition data, gives a introduction of Aichison′s contralized logarithm ratio transformation and the principal component analysis on composition data, and set an example based on peasants′ consuming structure, focuses on discussion the appliance steps of the principal component analysis on composition data.

Key words Composition data, principal component analysis, contralized logarithm ratio transformation