主成分分析应用于综合评价的局限性*

● 杜 晶,赵黎明

(天津大学 管理学院,天津 300072)

摘 要:主成分分析与因子分析在综合评价上均有广泛的应用。但在很多实践中二者对同一样本的分析却得到截然不同的结论,这种差异影响了两种方法在应用上的科学性。本文通过对两种方法数学原理的详细讨论,从样本几何的角度认为主成分分析在构造主成分得分函数的过程中,错误的将欧氏距离当作统计距离处理,从而对坐标的赋权产生前后矛盾,导致了该方法的不严谨。因此在综合评价中应慎用主成分分析。

关键词:主成分分析;因子分析;综合评价;局限性;辨析
中图分类号:F272.5 文献标识码:A 文章编号:1009-4458(2007)06-0125-03

主成分分析和因子分析都是基于数据总体方差-协方差结构分析的多元统计技术[1,2],由于二者精练而客观的揭示了数据内在的特性,在各个领域得到了广泛的应用。一般而言,研究者们使用这两种方法,是希望压缩数据或者找到影响变量的潜在因素;但在综合评价中,由于两种方法都是以数据内在特性对变量进行组合或拆分,避免了过多的主观评价,因此也成了综合评价中的常用方法[3]。

尽管主成分分析和因子分析在综合评价上都有各自标准的解法,并且这些解法均得到了广泛支持和大量应用,但实践中应用这两种方法处理同样数据,却常常得到不同的结论。这种不同不仅仅表现为评价结果方差的不同,更表现为结果的顺次和分类也截然不同。在这种情况下,应用上述两种方法进行综合评价时,已经引起了一些研究者的争论[4]。因此,有必要对两种方法的适用性问题进行讨论。本文旨在通过对两种方法最根本数学原理的讨论,明确其适用性。

一、问题的提出

1. 观点不同之争

林海明与张文霖在2005年第3期的《统计研究》上发表论文《主成分分析与因子分析的异同和SPSS软件》(以下简称《林文》),指出刘玉玫与张芠发表在2003年第12期《统计研究》上的论文《经济全球化程度的量化研究》(以下简称《刘文》)犯了"概念性的错误":即混淆了主成分分析与因子分析,错误地将初始因子分析当作主成分分析对中国等16个国家的经济全球化程度进行了量化评价,并且这种错误"致使相应经济分析有些偏离实际"[4]。为此,林海明等人给出了一个所谓正确的主成分分析解法,依照这种解法,果然得到了与《刘文》截然不同的分析结论。下表给出了两篇论文分析得到的各主成分/因子以及综合得分的序次:

表1 《刘文》与《林文》分析结论的差异[4,5]

国家	F1排名		F2排名		F3排名		F4排名		F综排名	
	PCA	因子分析	PCA	因子分析	PCA	因子分析	PCA	因子分析	PCA	因子分析
美国	3	3	1	1	2	2	14	14	1	1
英国	2	2	4	4	16	16	1	1	2	2
德国	4	4	3	3	5	6	8	8	3	3
日本	6	6	2	2	6	5	16	16	4	4
法国	5	5	5	5	14	14	4	4	5	5
新加坡	1	1	16	16	3	3	15	15	6	7
意大利	8	8	6	6	15	15	13	13	7	8
加拿大	7	7	12	12	11	11	7	7	8	9
中国	14	14	7	7	1	1	2	2	9	6
巴西	13	13	8	8	12	12	6	6	10	10
澳大利亚	10	10	14	14	10	10	5	5	11	11
韩国	12	12	11	11	7	7	12	12	12	13
墨西哥	11	11	13	13	4	4	9	9	13	14
新西兰	9	9	15	15	8	8	3	3	14	12
俄罗斯	15	15	10	10	9	9	10	10	15	15
印度	16	16	9	9	13	13	11	11	16	16

注:1. 原始数据省略,读者可参照参考文献[5];
2. 表中PCA指《林文》使用的方法;因子分析指《刘文》使用的方法。

由上表可以看到:两篇论文不但综合评价结果出现较大差异,在F3主成分/因子上的排序也出现不同。更进一步的,两篇论文依靠这些数据进行聚类分析,其结果也完全不同。《林文》对此差异的解释是:该差异是由于主成分分析与因子分析所构造的综合评价"函数方差不同造成的"[4]。但《林文》并未解释为何在部分主成分/因子排序上也出现差异,并且并未对问题进一步探讨,也未能说明在此问题上两种方法究竟谁适用,只是说"主成分分析与因子分析的实证结果是有差异的,不能混用"[4]。

2. 争论的本质

事实上,单从方法上看,两篇论文都是合理的。《刘文》使用的是初始因子分析对16国的数据进行分析(尽管文中声称使用的是主成分分析,正是这点引起《林文》的辩论);而《林文》使用的则是基于相关系数矩阵的主成分分析。从两篇论文各自的演

* 收稿日期:2007-05-21
作者简介:杜 晶(1982-),男,河北任丘人,天津大学管理学院研究生,研究方向:区域经济、产业规划。

算步骤来看, 均没有问题。进一步的, 笔者又依据两篇论文的解法对数据进行了验算, 排除了计算错误的可能性。因此, 两篇论文分析结论的不同是由于使用的方法不同造成的, 也就是主成分分析和因子分析本来就会产生差异。因此,《林文》与《刘文》的争论, 本质上是主成分分析和因子分析的矛盾。

二、主成分分析与因子分析辨析

直观来看, 分析结果的差异来自分析方法的不同。为了对问题有深入的探讨, 我们首先需要辨析主成分分析与因子分析的主要异同点。主成分分析和因子分析属完全不同的分析方法, 但在学界, 误会早已有之: 很多文献将主成分分析视为因子分析的一个特例, 将因子分析视为主成分分析的自然扩展。更一般的, 很多文献将主成分分析和因子分析视为同一种方法, 只不过因子分析多了一步因子旋转[6,7,8]。

尽管二者确实在某些步骤与基本思想上很相似: 如它们均从数据的总体方差—协方差矩阵出发, 力图逼近数据的协方差结构, 客观上都实现了维度压缩; 然而事实上, 这两种方法不论是在基本的数学模型上, 还是在具体的操作上, 都不尽相同。

1. 基本数学模型不同

设有 p 维变量, 每个变量有 n 个观测值, 则该数据集为 $p * n$ 矩阵, 记为 $X_{p\times n}$。

(1) 主成分分析数学模型。

记主成分分析中的主成分为向量 $F_{p\times n}$, 有主成分分析的数学模型:

$$F_{p\times 1} = A_{p\times n} X^T_{p\times n} \qquad (1)$$

其中: $A_{p\times n}$ 为变量的线性组合系数矩阵。

(2) 因子分析数学模型。

记因子分析中的各因子为向量 $Z_{k\times n}(k< p)$, 有因子分析的数学模型:

$$(X-\mu)_{p\times n} = L_{p\times k} Z_{k\times n} + \varepsilon_{p\times n} \qquad (2)$$

其中: $L_{p\times k}$ 为因子载荷矩阵。

从上述主成分分析和因子分析的数学模型我们看到: 主成分分析是将主成分视为原变量的直接线性组合, 属于数理模型; 而因子分析则是将原变量视为由某些不可观测的因子线性组合而成, 是对原变量的拆分, 属于计量模型。从这点上看, 因子内容比主成分内容单纯。

2. 标准演算步骤不同

(1) 主成分分析综合评价步骤。

① 第一, 将原始数据标准化;

② 计算数据相关系数矩阵(实际为标准化后数据的总体方差—协方差矩阵);

③ 计算相关系数矩阵的特征值与特征向量;

④ 依照某种法则确定提取的主成分个数;

⑤ 表达主成分, 其中系数矩阵 $A_{p\times n}$ 为对应的特征向量组成的矩阵;

⑥ 计算各主成分得分, 其中数据采用标准化后的数据;

⑦ 依据各主成分方差贡献率构造出综合评价函数, 计算综合得分。

(2) 因子分析综合评价步骤。

① ~ ④ 与主成分分析相同;

⑤ 计算因子载荷矩阵 $L_{p\times k}$, 即选定各因子与变量的相关系数矩阵;

⑥ 估算因子得分系数矩阵, 计算出各因子得分;

⑦ 依据各因子方差贡献率构造出综合评价函数, 计算综合得分。

从上文我们看到, 在综合评价中, 初始因子分析与主成分分析的步骤极为相似(因子分析为 R 型, 并采用主成分分析来求解初始因子), 它们最大的不同就在于如何计算主成分 / 因子得分。

3. 计算得分的数学原理不同

(1) 主成分得分的数学原理。

直接将标准化的数据带入公式(1):

$F_{p\times 1} = A_{p\times n} X^T_{p\times n}$, 则得到主成分得分。

其中系数矩阵 $A_{p\times n}$ 为对应的特征向量组成的矩阵, $X^T_{p\times n}$ 为标准化后数据集。

从中我们看到, 计算主成分得分实际上是将标准化后的原始数据投影到旋转后的坐标系中。(2) 因子得分的数学原理。

因子得分是通过公式(2):

$(X-\mu)_{p\times 1} = L_{p\times k} Z_{k\times 1} + \varepsilon_{p\times 1}$ 估计得到的。

由于 $k< p$, 也就是方程数量少于未知数个数, 因此我们无法直接从上式求解出 $Z_{k\times 1}$。但是有很多方法可以帮助我们估算出 $Z_{k\times 1}$ 的近似值 $\hat{Z}_{p\times n}$。常用的方法有加权最小二乘法、回归法等。其中回归法比较常见, 它是 1939 年出 *Thomson* 提出来的, 所以又称为汤姆森回归法。

三、对主成分分析数学原理的进一步讨论

通过上文对主成分分析与因子分析主要不同的讨论, 我们可以判断二者在综合评价上的差异来自于不同的原理和方法。其中, 鉴于二者前几个步骤基本相同, 计算得分时采用的不同的数学原理可能是造成差异的直接原因。那么, 究竟是哪个环节出了问题?为了解释差异产生的原因, 下文将从样本几何的角度对主成分分析的关键数学问题进行讨论。

1. 数据标准化的一般解释

主成分分析的第一步是对数据进行标准化处理, 标准化的公式是:

$$Z = (V^{1/2}(X-\mu) \qquad (3)$$

其中对角标准离差矩阵 $V^{-1/2}$ 由下式定义:

$$V^{1/2} = \begin{bmatrix} \sqrt{\sigma_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{\sigma_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\sigma_{pp}} \end{bmatrix} \qquad (4)$$

在这种情况下:

$$E(Z) = 0,\ Cov(Z) = (v^{1/2})^{-1} \sum (V^{1/2})^{-1} = \rho \qquad (5)$$

其中 $\sum$ 为原始样本的总体方差—协方差矩阵, ρ 为样本的相关系数矩阵。因为相关系数矩阵实际上就是标准化后的总体方差—协方差矩阵, 因此在很多文献中, 主成分分析或因子分析并不是从总体方差—协方差矩阵出发, 而是从相关系数矩阵出发。《林文》与《刘文》均是此种情况, 换言之, 他们都对数据进行了标准化处理。

关于标准化的目的, 一般的解释是: 很多情况下, 不同参数

的数据之间存在着量纲或取值范围的较大差异，这些差异影响了主成分分析的结果。比如 X_1 代表某企业的年销售额，取值范围在 100 万至 500 万之间，而 X_2 为比值（净收入/总资产），取值范围在 0.05 到 0.60 之间，那么可以预期，总方差的变化将几乎全部归因于 X_1：企业的年销售额。在这种情况下，我们将会给 X_1 赋予过高的权重，而几乎丢掉 X_2 的作用。因此，需要对数据进行标准化，使标准化后的 X_1 和 X_2 居于同一数量级上，使它们在主成分结构上的作用变的可比。在后面的分析中我们发现，标准化是导致主成分分析在综合评价中的局限性的开始。为了更好地解释这个问题，我们需要从另一个角度来理解数据标准化的意义。

1. 欧氏距离与统计距离

从样本几何的角度理解，主成分分析基于的是简单的距离概念。主成分分析将 p 个变量的评价体系视为 p 维空间，那么样本数据将产生一个 p 维散点图。在这个空间中主成分分析将构造从重心到某一点 $P=(x_1, x_2, x_3, \cdots x_p)$ 的距离的度量。然后它通过旋转坐标系至数据点间距离最大的方向，从而只用少数的新坐标来解释数据的大部分变异。最熟悉的距离构造方法是欧氏的：

$$d(\bar{x}, P)= \sqrt{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots(x_p-\bar{x})^2} \quad (6)$$

但是欧几里得距离对于大多数统计目的而言是不合适的。为了举例说明，假设有 2 个变量的 n 对观测值，称 2 个变量为 x_1，x_2 并假设它们互相独立，且 x_1 观测结果的可变性大于 x_2。数据散点图如下：

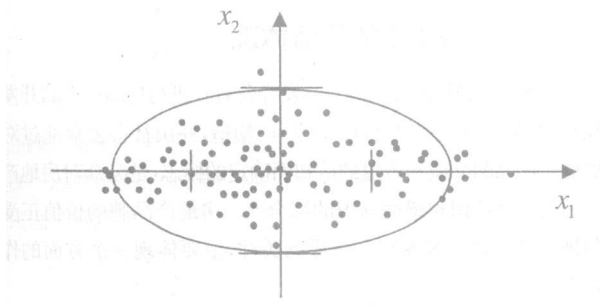

图 1 变化性不同的二维散点图

注：除非数据为多元正态分布，否则散点不一定充满于某个椭球体内。

从上图可以看到，从原点出发 x_1 方向的给定偏差并不像从原点向 x_2 方向等距离值那样的重要。这是因为 x_1 方向内在可变性本身要大于 x_2 方向，同等 x_1 坐标绝对值对系统整体的影响应当小于 x_2 坐标。消除这种差异的方法就是对距离进行加权，令：

$$x_1^* = x_1/\sqrt{S_{11}} \quad (7)$$

$$x_2^* = x_2/\sqrt{S_{22}} \quad (8)$$

更一般的，考虑重心不是原点的数据集，令：

$$x_1^* = (x_1-\bar{x})/\sqrt{S_{11}} \quad (9)$$

$$x_2^* = (x_2-\bar{x})/\sqrt{S_{22}} \quad (10)$$

重新定义的距离为：

$$d^*(\bar{x}, P)= \sqrt{\frac{(x_1-\bar{x})^2}{S_{11}}+\frac{(x_2-\bar{x})^2}{S_{22}}+\cdots\frac{(x_p-\bar{x})^2}{S_{pp}}} \quad (11)$$

由于这个距离定义考虑到数据变异的不同，因此是一个符合统计目标的距离定义，被称作统计距离。统计距离定义的空间被称作统计空间，在统计空间中，单位圆往往是一个椭圆（如图 1）；一般的，则是一个 p 维超椭球体[1]。

统计距离是多元统计分析的基础，因此也是主成分分析的基础。从统计距离的角度理解主成分分析数据标准化：标准化的目的就是将数据从欧氏空间映射到符合统计目标的统计空间。

3. 坐标赋权

我们还可以从另一个角度理解统计距离的含义：它实际上是一种坐标赋权。考虑到在不同坐标上分布的数据内在特性的不同，它们对系统变异的统计意义是不一样的。只有消除这种差异才能帮助我们分析变量的实际贡献，而消除的办法就是对坐标系的每一个轴赋予不同的权重。不论是数据标准化，还是上文所分析的统计距离，本质上是对坐标系的赋权过程（权重为 $V^{-1/2}$），从而构建出新的统计空间。而中心化处理则是对坐标系的平移，这一过程并不会改变空间的性质。

下面再回到主成分分析。主成分分析是从数据的总体方差－协方差矩阵出发，其样本坐标系是正交的。主成分分析正是将该坐标系正交旋转，获得新的坐标系，实现了对变量的拆分和重组。在其演算过程中，数据标准化对坐标系做了平移和赋权；计算相关系数矩阵的特征向量则是对坐标系的正交旋转。但是该正交旋转并未将数据标准化对原坐标系的平移和赋权动作带入到新的坐标系中。

四、主成分分析在综合评价中的误差

从样本几何的角度看，主成分分析的主要步骤对应着几个几何变换过程。如下表所示：

表 2 主成分分析的主要步骤与对应几何变换

主成分主要步骤	样本的几何变换
1. 原始数据 由于数据变异程度不同和相关性的存在，原始数据在 p 维空间的分布呈现出一定的线性特征。	
2. 数据中心化 为了方便分析，平移 p 维坐标系，使新中心为数据的期望值；但坐标系的单位球仍为正球体。	
3. 数据标准化 对坐标赋权以消除量纲影响，标准化后坐标系的单位球为 p 维椭球体。需要注意的是该椭球体的轴向并不是沿着数据集方差最大的方向。	
4. 提取主成分与计算主成分得分 从样本几何的角度来看，就是将坐标系依数据协方差矩阵的特征向量进行旋转；主成分得分就是数据点在新坐标系上的投影。	

从上表中可以看到，主成分分析的问题出在坐标旋转前后：坐标旋转前，对各坐标轴做了加权处理，这种加权表现在数据上就是标准化；坐标旋转以后，计算主成分的得分却成了直接计算数据点在旋转之后坐标系的投影——而这些数据采用的是标准化的数据，也就是默认为旋转以后的坐标系被一一对应的赋予了与旋转前相同的权重——这与事实明显不符。因为随着坐标系的旋转，数据集合在坐标轴方向上的变异特性已经发生了变化。坐标系的权重实际上在旋转前后是（转第 130 页）

表 2　　互动媒体和寄件媒体的优势和劣势[5]

	优势	劣势		优势	劣势
寄件媒体	针对性强	成本高	互动媒体	大量的受众	定向成本较高
	覆盖集中	投递问题		及时反应	收带宽限制
	便于控制	缺少内容		高度针对性	尚不属于主流媒体
	人性化冲击力			提供详细资料	安全与隐私方面的顾虑
	隐蔽性				
	反馈快而准确				
	可证实性				

表 3　　几种传统大众媒体的优势和劣势[6]

	优势	劣势		优势	劣势
电视	影响力大	成本较高	广播	低成本接触	难以记忆
	日程需要	市场传递不平均		日程需要	情景少
	快速知晓	承担义务的要求		信息量可伸缩	没有视觉影响
报纸	直接影响受众	目标受众很难精准		市场配合度高	目标受众特定、有限
	引导时间较短	定为控制能力最弱	海报广告牌	接触率高	信息没深度
	操作灵活	难以有效吸引受众		频率高	没有储存的空间
	便于储存	情景少		可立即执行	投放时间长
杂志	储存时间长	引导时间长		可地方化	信息难以搜寻
	浪费少	读者流量积累慢		进度灵活	好位置面临更多竞争
	商业可能	市场传递不平衡			
	嵌入式广告的读者	区域保险			

第一，大众媒体。现在房地产品牌传播主要采用报纸、电视。大众媒体由于费用较高，信息投放的定向性不强，不能实现定制化，而且有单向性信息发布等缺点，因此大众媒体的地位受到挑战。但是，在建立品牌知名度和形象，发布标准品牌信息时，大众媒体还是最佳选择。在选择大众媒体时，一定要研究潜在消费者的媒体，才能尽可能地实现广告信息精确投放。

第二，互动媒体。互动媒体具有双向沟通的特点，以互联网为主要代表，具有实时性、互动性、社区性、经济性、定制化等优点。在品牌传播方面，已经有许多房地产企业在公司网站上进行品牌的传播和项目的推广，如香港的新鸿基地产，在其网页上利用多媒体技术来展示项目包括整体情况到局部细节的信息，依靠这种方式来联系消费者，可以获得相关的消费者信息，建立消费者数据库。

第三，寄件媒体。寄件媒体指可以传递到个人私人区域和私人电子邮箱的媒体，如直销邮件、电子邮件、传真等。一般认为，寄件媒体在建立与顾客一对一地紧密牢固关系中是不可或缺的。

企业在进行品牌宣传之前，要以品牌的形象和自身资源为基础，根据不同传播工具的特点和成本与达到特定目标受众的能力，选择适合自己企业和产品的传播工具组合。在品牌传播过程中，围绕着品牌核心价值展开，这样就能做到让消费者从不同的传播渠道听到一个声音，使各个传播通路的行销宣传形成，整合传播的效果和累计的传播效应。

四、结语

我国房地产业竞争日趋激烈，已经由单纯的质量竞争，价格竞争演变为比质量、比价格、比信誉、比服务、比形象的品牌竞争。如今，房地产企业迅猛发展，但是真正能够做到品牌企业的却屈指可数，要在品牌竞争中获得优势，企业要通过建立品牌管理组织，制定品牌创造计划，有效的品牌定位和品牌设计、提供高质量的产品和服务以及不断的创新来创立品牌，同时要辅之以广告、媒体等品牌传播手段，来博得消费者的忠诚。□

参考文献：

[1] 金　乐．理念狂飙，房地产品牌运营论[M]．北京：中国经济出版社，2001.
[2] 董　藩．房地产营销与管理[M]．大连：东北财经大学出版社，2001.
[3] 李　健．范贻昌．关于房地产企业战略环境的 SWOT 分析[J]．内蒙古农业大学学报(社会科学版)，2006(3).
[4] 国务院发展研究中心企业研究所、清华大学房地产研究所和中国指数研究院．2006 中国房地产品牌价值研究报告[R]．2006.9.
[5] 明　阳．品牌传播学[M]．上海：上海交通大学出版社，2005.
[6] Kevin Lane Kelle. Strategic Brand Management[M]. Prentice Hall, 1998.

(接第 127 页)

不同的。这样的做法忽略了旋转后统计空间的改变，正确的做法是需要重新考虑坐标的赋权，也就是做出一个轴向与旋转后坐标轴一致的新 p 维椭球体，该椭球体代表旋转后坐标系的单位球。

五、结语

根据本文的讨论，我们得到如下结论：

第一，为了消除量纲影响，不论是主成分分析或因子分析，大部分情况下都首先对数据进行了标准化处理，但是主成分分析在计算得分的时候，只是简单的计算数据点在新坐标系的投影，没有考虑旋转前后坐标赋权的变化，因此有内在的矛盾。这个矛盾产生的误差往往易被忽略，但在某些情况下，影响了模型的准确性。

第二，因子分析在计算得分的时候，尽管采用的是最小二乘法或回归法等方法估计出来的，但保持了逻辑的正确性，因此相对于主成分分析更加精确的逼近了模型。这一点也被其他研究者认可[1]。

因此在综合评价时，应当慎用主成分分析。事实上，已经有不少研究者提出来主成分分析更多的应该作为一个中间步骤。在进一步的研究中，我们应当可以得到衡量这种差异的量化的数学表达式，并基于此分析主成分分析与因子分析的结果差异会随着数据的不同特性怎样变化，从而提出一个修正方法。□

参考文献：

[1] Richard A. Johnson, Dean W. Wichern. Applied Multivariate Statistical Analysis 4th ed.[M]. Pearson Education Company, 1998.
[2] 王伟华．基于主成分分析法的城市土地利用集约度研究[J]．内蒙古农业大学学报(社会科学版)，2005(4).
[3] 汪应洛．系统工程[M]．北京：机械工业出版社，2003.
[4] 林海明，张文霖．主成分分析与因子分析的异同和 SPSS 软件—兼与刘玉玫、卢纹岱等同志商榷[J]．统计研究，2005(3).
[5] 刘玉玫，张　芫．经济全球化程度的量化研究[J]．统计研究，2003(12).
[6] 刘　平．浅谈主成分分析与因子分析的异同[J]．辽宁师专学报，2004(3).
[7] 于秀林．多元统计分析[M]．北京：中国统计出版，1999.
[8] 丘　东．多指标综合评价方法系统分析[M]．北京：中国统计出版社，1991.