

主成分分析与初始因子分析的异同

——兼与卢纹岱《SPSS for Windows 统计分析》商榷

林海明

(广东商学院 经济贸易与统计学院, 广州 510320)

摘要:主成分分析与初始因子分析最相近,二者的不同隐藏较深,不留意还以为同是主成分分析。事实上,这两种方法是有区别的,不能混用。本文给出了这两种方法的异同与实证比较。

关键词:主成分分析;初始因子分析;异同

中图分类号:F810.2 **文献标识码:**A **文章编号:**1002-6487(2006)04-0033-02

0 问题产生的背景

初始因子分析是指用主成分法提取初始因子载荷阵,并没有旋转过程的因子分析(对应分析用此法)^[3]。它与主成分分析最相近,二者的不同隐藏较深,不留意还以为同是主成分分析,如文[2]就出现了这种情况(经运算验明),有些教科书如[1]也没有明确区分。那么这两种方法有何异同?实证结果一致吗?

1 主成分分析与初始因子分析的异同

设 $X=(X_1, \dots, X_p)$ 为标准化随机向量($p \geq 2$), R 为相关系数矩阵, $F_m=(F_1, \dots, F_m)$ 为主成分向量, $Z_m=(Z_1, \dots, Z_m)$ 为因子向量, $m \leq p$ 。为方便,因子、因子估计、因子得分用同一记号。

不同之处:表达式、方差、标准正交性、回归过程、综合评价函数及方差见表1。

相同之处:主成分分析(原理见文献[3])与R-型初始因子分析都是对协差阵的逼近,都是打算降维解释数据集。具体为指标的正向化,指标的标准化(软件自动执行)通过相

表1 主成分分析与R-型初始因子分析的不同

区别项目	主成分分析数学模型:	R-型初始因子分析数学模型:
表达式与系数矩阵	$F_m = A_m'X$ $A_m = (a_{ij})_{p \times m} = (\alpha_1, \alpha_2, \dots, \alpha_m)$, $R\alpha_i = \lambda_i \alpha_i$, λ_i 是相应的特征值和 单位特征向量, $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m \geq 0$ 。	$X = B_m Z_m + \varepsilon$ (ε 为特殊因子), 因子载荷矩阵 $B_m = (b_{ij})$ $= (\sqrt{\lambda_1} \alpha_1, \sqrt{\lambda_2} \alpha_2, \dots, \sqrt{\lambda_m} \alpha_m)$ 为初始因子载荷矩阵*(λ_i, α_i 同左)。
方差	$\text{Var } F_i = \lambda_i$, 依次达到信息贡献最大化。	协 $\text{Var } Z_i = 1$, 没有依次达到最大化。
标准正交性	是, 要求 $A_m' A_m = I_m$ (判据之一)。	非, 因为 $B_m' B_m \neq I_m$ 。
回归过程	无。	有, 因子得分函数 $Z_m = B_m' R^{-1} X_m$ 。
综合评价函数及方差	$F_{\text{综}} = \sum_{i=1}^m (\lambda_i/k) F_i$, $\text{Var } F_{\text{综}} = (\sum_{i=1}^m \lambda_i^3)/k^2$, $k=p$ 或 $\lambda_1 + \dots + \lambda_m$ 。	$Z_{\text{综}} = \sum_{i=1}^m (\lambda_i/k) Z_i$, $\text{Var } Z_{\text{综}} = (\sum_{i=1}^m \lambda_i^2)/k^2$ (同左)。 通常 $\text{Var } Z_{\text{综}} \leq \text{Var } F_{\text{综}}$, 即 $F_{\text{综}}$ 的取值范围通常比 $Z_{\text{综}}$ 大。

表2 方差解释

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	10.837	43.348	43.348	10.837	43.348	43.348
2	5.802	23.207	66.555	5.802	23.207	66.555
3	2.060	8.240	74.795	2.060	8.240	74.795

Extraction Method: Principal Component Analysis.

关系数矩阵判断变量间的相关性,求相关系数矩阵的特征值和特征向量,主成分间、因子间线性无关,一般用累计贡献率 $\geq 85\%$ 、变量不出现丢失确定主成分、因子个数 m ,主成分与因子对 X 的贡献相同、是最大化的,命名依据都是用相同的主成分、因子与变量的相关系数。

主成分分析与初始因子分析由于方差的不同,直接导致主成分值、因子得分值的不同,故主成分分析与因子分析的定量综合评价体系不同,综合评价应该分开进行,混淆在一起是不同定量值交替错误。

以上分析说明,两者定量上不同的显著性标志是方差的不同。

2 卢纹岱《SPSS for Windows 统计分析》中轿车顾客偏好研究中正确的主成分分析结果

①卢纹岱《SPSS for Windows 统计分析》(以下简称《卢书》)中表13-16的结果是三个初始因子的载荷矩阵,不是三个主成分的矩阵(三个主成分的矩阵为本文的表4);②《卢书》图13-23为初始因子分析的变量(25个顾客偏好)散点图,不是25个顾客的主成分分数散点图;③《卢书》图13-19的结果fact-1、fact-2、fact-3为初始因子分析因子得分值结果,不是主成分分数变量;④《卢书》图13-22为初始因子分析因子得分值(17种车型)散点图,不是17

表3 初始因子载荷阵

变量	Component		
	1	2	3
v1	.274	.625	.330
v2	.956	.068	-.210
v3	.778	-.300	-.151
v4	.491	.735	.343
v5	.451	.698	-.318
v6	.238	.677	-.059
v7	.783	-.212	.170
v8	.510	-.051	.713
v9	-.513	.718	-.189
v10	.936	-.191	.050
v11	.852	.143	-.260
v12	.836	-.085	-.356
v13	.943	.000	-.149
v14	.830	.198	-.081
v15	.858	-.174	-.067
v16	-.015	.803	.077
v17	.105	.658	.235
v18	.717	.609	.096
v19	.779	.126	-.033
v20	.773	-.570	.124
v21	.071	.657	-.095
v22	.238	-.459	.753
v23	-.766	.333	.281
v24	-.162	-.753	-.209
v25	-.765	.158	-.270

表4 主成分系数矩阵

变量	A1	A2	A3
v1	.083	.259	.230
v2	.29	.028	-.146
v3	.236	-.125	-.105
v4	.149	.305	.239
v5	.137	.29	-.22
v6	.072	.281	-.041
v7	.238	-.088	.118
v8	.155	-.021	.497
v9	-.156	.298	-.132
v10	.284	-.072	.033
v11	.259	.059	-.181
v12	.254	-.035	-.248
v13	.286	.000	-.104
v14	.252	.082	-.056
v15	.261	-.072	-.047
v16	-.005	.333	.054
v17	.032	.273	.164
v18	.218	.253	.067
v19	.243	.052	-.023
v20	.235	-.237	.086
v21	.022	.273	-.066
v22	.072	-.191	.524
v23	-.233	.138	.196
v24	-.049	-.313	-.146
v25	-.232	.066	-.188

表5 主成分、综合主成分值排名

车型	F ₁	排名	F ₂	排名	F ₃	排名	F	排名
沃尔沃	6.772	1	1.03	4	1.752	2	3.32	1
大众 R	4.69	2	-0.102	8	-0.104	10	2.00	2
大众 D	4.452	4	0.644	5	-1.132	13	1.99	3
本田 A	4.562	3	0.338	6	-1.488	14	1.90	4
本田 C	3.081	5	-1.06	12	-1.491	15	0.97	5
林肯	-2.92	15	5.234	1	-1.045	12	-0.14	6
凯迪拉克	-2.84	14	4.984	2	-1.013	11	-0.16	7
庞蒂阿克	-1.67	12	2.042	3	0.559	8	-0.20	8
雪伏龙 M	-1.36	9	-0.116	9	1.468	4	-0.50	9
雪佛龙 CI	-0.18	6	-2.152	14	0.866	7	-0.51	10
福特 M	-1.32	8	-0.565	10	2.151	1	-0.53	11
普利茅斯 H	-1.18	7	-1.061	13	0.003	9	-0.76	12
普利茅斯 V	-1.66	11	-0.599	11	0.956	6	-0.78	13
福特 F	-1.40	10	-2.245	14	1.606	3	-1.00	14
普利茅斯 G	-3.08	16	0.274	7	1.207	5	-1.17	15
雪佛龙 CH	-2.62	13	-3.491	17	-2.241	17	-2.13	16
福特 P	-3.24	17	-3.157	16	-2.053	16	-2.30	17

种车型主成分散点图(17种车型主成分值见本文的表5);⑤《卢书》图13-19的结果fact-1、fact-2、fact-3(初始因子分析因子)的方差全为1(主成分F₁、F₂、F₃的方差分别为本文表2的特征值10.837、5.802、2.06)。

故《卢书》图13-19的初始因子分析因子得分值fact-1、fact-2、fact-3与本文表5主成分值F₁、F₂、F₃的取值全部不同。即主成分分析与初始因子分析的实证结果是有差异的,计量值全部不同,不能混用。

现按主成分分析法和SPSS软件应用时一对一的正确步骤^[4]给出《卢书》的主成分分析结果。

笔者根据《卢书》给出的市场研究中的轿车顾客偏好数据data13-02a用SPSS软件Analyze菜单Factor过程进行主成分分析(通过相关系数矩阵判断变量间的相关性略)得出相关系数矩阵的特征根及主成分贡献率见表2,特征向量矩阵见表3。由于前三个主成分累计方差贡献率已达到74.793%,且无变量丢失,基本包含了全部的指标所具有的信息,故取3个主成分(与《卢书》同)。

主成分命名:表3中每一个载荷量表示主成分与对应变量的相关系数,且系数符号与题意相符,第一主成分F₁与v2、v3、v7、v10、v11、v12、v13、v14、v15、v18、v19、v20十分显著正相关,与v23、v25十分显著负相关,称F₁为车的产地偏好成分^[2];第二主成分F₂与v1、v4、v5、v6、v9、v16、v17、v21十分显著正相关,与v24十分显著负相关,称F₂为车的特性(质量、动力、座位数等)偏好成分^[2];第三主成分F₃与v8、v22十分显著正相关,v8、v22背景不明,暂不命名。从这里也可以看出前三个主成分基本包含了全部的指标所具有的信息。

三个主成分的表达式还不能从输出窗口中直接得到,因

为“Component Matrix”是指**初始因子载荷矩阵**,为了得到三个主成分的表达式,以便求主成分值,还需进一步操作:将前三列因子载荷系数输入到数据编辑窗口(为变量B₁、B₂、B₃),然后利用Transform-->compute,在对话框中输入A_i=B_i/SQR(λ_i),λ_i为特征值,即可得到表4主成分系数向量A_i,i=1、2、3,于是,三个主成分表达式如下(v_i是V_i的标准化数据):

$$F_1 = 0.083v_1 + 0.29v_2 + 0.236v_3 + 0.149v_4 + \dots + 0.072v_{22} - 0.233v_{23} - 0.049v_{24} - 0.232v_{25}$$

$$F_2 = 0.259v_1 + 0.028v_2 - 0.125v_3 + 0.305v_4 + \dots - 0.191v_{22} + 0.138v_{23} - 0.313v_{24} + 0.066v_{25}$$

$$F_3 = 0.23v_1 - 0.146v_2 - 0.105v_3 + 0.239v_4 + \dots + 0.524v_{22} + 0.196v_{23} - 0.146v_{24} - 0.188v_{25}$$

应用这一线性组合计算出表5各主成分值(无回归过程)。

利用综合主成分函数(k=p=25) $F = 0.43F_1 + 0.23F_2 + 0.08F_3$ 可以求得各汽车的综合主成分(顾客偏好)值(见表5)。

轿车顾客偏好的分析:

从表5产地偏好F₁、综合偏好F值的排序看出:轿车顾客偏好欧洲车沃尔沃(第1,第1)、大众R(第2,第2)、大众D(第4,第3)和日本车本田A(第3,第4)、本田C(第5,第5)的倾向普遍高于其余美国车;从表5车的特性F₂值的排序看出:注重轿车特性(质量、动力、座位数等)的轿车顾客偏好美国车林肯(第1)、凯迪拉克(第2)。因此,高质量、豪华大型的欧洲车、日本车是新车型的开发方向。

参考文献:

- [1]张尧庭,方开泰著.多元统计分析引论[M].科学出版社,1997.
- [2]卢纹岱.SPSS for Windows 统计分析(第二版)[M].电子工业出版社,2002.
- [3]于秀林,任雪松.多元统计分析[M].北京:中国统计出版社,1999.
- [4]林海明,张文霖.主成分分析与因子分析详细的异同和SPSS软件[J].统计研究,2005,(3).

(责任编辑/李友平)