如何正确应用 SPSS 软件做主成分分析

李小胜　陈珍珍

内容提要:鉴于目前很多用 SPSS 软件分析主成分的教材中和发表的文章中有不少错误之处,本文从主成分分析与因子分析的关系出发,借用 SPSS 软件自带的数据,进行了正确的操作,并将其结果与 SAS 软件的结果进行比较,两者完全相同。

关键词:SPSS;主成分分析;因子分析
中图分类号:C812　　文献标识码:A　　文章编号:1002 - 4565(2010)08 - 0105 - 04

Correctly Using SPSS Software for Principal Components Analysis

Li Xiaosheng & Chen Zhenzhen

Abstract: In view of the errors in many books and articles about applying SPSS software for principal components analysis, this paper shows the right operations in using SPSS from the relationship between principal components analysis and factor analysis, and finds that the results are the same with that from the SAS software.

Key words: SPSS; Principal components analysis; Factor analysis

一、引言

主成分分析(principal components analysis)也称主分量分析,由霍特林(Hotelling)于 1933 年首先提出。主成分分析是利用降维的思想,在损失很少信息的前提下把多个指标转化为几个综合指标的多元统计方法。通常把转化后的综合指标称之为主成分,其中每个主成分都是原始变量的线性组合,且各个主成分之间互不相关,这就使得主成分比原始变量具有某些更优越的性能。这样在研究复杂问题时就可以只考虑少数几个主成分而不至于损失太多信息,从而更容易抓住主要矛盾,揭示事物内部变量之间的规律性,同时使问题得到简化,提高分析效率。

由于主成分分析的这些优势,在实际问题中遇到指标较多且各指标相关关系较大时,人们常考虑应用主成分分析的方法。但是目前用 SPSS 软件分析主成分的教材中和发表的文章中有很多错误和误解之处(SAS 软件中有主成分分析和因子分析的专门语句,一般不会出现这种情况):(1)如果把主成分与原始变量(或标准化后的变量)的相关系数矩阵叫做因子负荷阵,把原始变量标准化后用因子来表示的系数阵叫做因子载荷阵,那么 SPSS 软件得到的是因子载荷阵,因子载荷阵表示标准化后的主成分(或叫公因子,方差为 1)来近似标准化后原始变量的系数矩阵。(2)主成分的系数是因子载荷阵推出的,不是从因子负荷阵推出的,即从因子分析得到的载荷阵求主成分的系数时很多教材中和文章中的公式表达错误,虽然实际数据结果是对的。这时的主成分的方差不是 1,即非标准化的主成分。(3)当 SPSS 软件从相关系数求主成分时,主成分应表示为标准化后的随机变量的线性组合,有些文献中就没加区分,把主成分直接写成原始变量的线性组合。(4)为了从因子分析得到主成分的系数,在 SPSS 软件中对因子不要旋转,实际上很多人旋转了。(5)从因子得分系数矩阵得到主成分系数表达式,可以认为因子与标准化原始变量间的变换关系是可逆的,因为因子的提取采用主成分方法时,标准化后的随机变量完全由因子来表示[3]。鉴于以上错误和误解,本文从主成分分析与因子分析的关系出发,借用 SPSS 软件自带的例子,进行了正确的操作,将其结果与 SAS 软件进行比较,结果完全相同。

二、联系与区别

(一)主成分分析

设对某一事物的研究涉及到 p 个指标,记为 $X_1,X_2,\cdots X_p$,这 p 个指标构成的 p 维随机向量为 $X=(X_1,X_2,\cdots X_p)'$。对 X 进行线性变换,可以形成新的综合变量,用 Y 表示,也就是说,新的综合变量可以由原来的变量线性表示,满足下式:

$$\begin{aligned}Y_1&=b_{11}X_1+b_{12}X_2+\cdots+b_{1p}X_p\\Y_2&=b_{21}X_1+b_{22}X_2+\cdots+b_{2p}X_p\\&\cdots\cdots\\Y_p&=b_{p1}X_1+b_{p2}X_2+\cdots+b_{pp}X_p\end{aligned}\quad(1)$$

由于可以任意地对原始变量进行上述的线性变换,不同的线性变换得到的综合变量 Y 的统计特性也不尽相同。通常主成分要求满足如下的三个条件:

1. $\mathbf{b}'_i\mathbf{b}_i=1$,即 $b_{i1}^2+b_{i2}^2+\cdots+b_{ip}^2=1$,$\mathbf{b}'_i=(b_{i1},b_{i2},\cdots,b_{ip})$,$i=1,2,\cdots,p$;

2. Y_i 与 Y_j 相互无关($i\neq j;i,j=1,2,\cdots p$);

3. Y_1 是 $X_1,X_2,\cdots X_p$ 的一切满足条件 1 的线性组合中方差最大者;Y_2 是与 Y_1 不相关的 $X_1,X_2,\cdots X_p$ 的一切满足条件 1 的线性组合中方差最大者;……;Y_p 是与 $Y_1,Y_2,\cdots,Y_{p-1}$ 都不相关的 $X_1,X_2,\cdots X_p$ 的一切满足条件 1 的线性组合中方差最大者。

基于以上三个条件决定的综合变量,我们把 $Y_1,Y_2,\cdots,Y_p$ 重新记为 $G_1,G_2,\cdots,G_p$,分别称为原始变量的第一、第二、…、第 p 主成分,其系数重新记为 c_{ij},$(i,j=1,2,\cdots,p)$,根据矩阵代数的知识,每个主成分的方差($var(G_i),i=1,2,\cdots,p$),其实就是 $X_1,X_2,\cdots X_p$ 的协方差阵 $\sum$ 的非零特征值(λ_i),于是主成分与原始变量关系为:

$$\begin{aligned}G_1&=c_{11}X_1+c_{12}X_2+\cdots+c_{1p}X_p\\G_2&=c_{21}X_1+c_{22}X_2+\cdots+c_{2p}X_p\\&\cdots\cdots\\G_p&=c_{p1}X_1+c_{p2}X_2+\cdots+c_{pp}X_p\end{aligned}\quad(2)$$

记 $G=(G_1,G_2,\cdots,G_p)'$,$C=(c_{ij})_{p\times p}$,那么上式可以表示为:$G=C'X$,其中 C 是正交阵。如果数据是标准化后,即从相关系数矩阵出发,求得的特征值与对应的特征向量为主成分的系数矩阵。基于相关系数矩阵还是基于协方差矩阵做主成分分析:当分析中所选择的经济变量具有不同的量纲,变量水平差异很大,应该选择基于相关系数矩阵的主成分分析。对同度量或是取值范围在同量级的数据,还是直接从协方差矩阵求主成分。

对上述问题涉及到的 p 个指标 $X_1,X_2,\cdots X_p$,我们为了从相关系数矩阵出发,将 p 个指标标准化后记为 $ZX=(ZX_1,ZX_2,\cdots ZX_p)'$,相关系数矩阵记为 R。那么求得的主成分可以表示为:

$$\begin{aligned}F_1&=u_{11}ZX_1+u_{12}ZX_2+\cdots+u_{1p}ZX_p\\F_2&=u_{21}ZX_1+u_{22}ZX_2+\cdots+u_{2p}ZX_p\\&\cdots\cdots\\F_p&=u_{p1}ZX_1+u_{p2}ZX_2+\cdots+u_{pp}ZX_p\end{aligned}\quad(3)$$

其中:

$\mathbf{u}'_i=(u_{i1},u_{i2},\cdots,u_{ip})$,$F=(F_1,F_2,\cdots,F_p)'$,

那么上式可以表示为:$F=U'ZX$,其中 U 是正交阵。

(二)因子分析

因子分析(factor analysis)的一般模型:设对某一事物的研究涉及到 p 个指标 $X_1,X_2,\cdots X_p$,这 p 指标有着较强的相关性,为了便于研究,在指标同向化的基础上,将样本数据进行标准化。为了说明方便,将同向化和标准化后的变量向量用 ZX 表示,即 $ZX=(ZX_1,ZX_2,\cdots ZX_p)'$,其均值向量 $E(ZX)=\mathbf{0}$,协方差矩阵记为 $\mathrm{cov}(ZX)=\sum_{zx}$,其实这里的协方差矩阵 $\sum_{zx}$ 与相关系数矩阵 R_{zx} 相同,那么因子分析的一般模型为:

$$\begin{aligned}ZX_1&=a_{11}f_1+a_{12}f_2+\cdots+a_{1m}f_m+\varepsilon_1\\ZX_2&=a_{21}f_1+a_{22}f_2+\cdots+a_{2m}f_m+\varepsilon_2\\&\cdots\cdots\\ZX_p&=a_{p1}f_1+a_{p2}f_2+\cdots+a_{pm}f_m+\varepsilon_p\end{aligned}\quad(4)$$

其中 $f=(f_1,f_2,\cdots f_m)'$,$(m<p)$ 为公因子,$\varepsilon=(\varepsilon_1,\varepsilon_1,\cdots\varepsilon_p)'$为特殊因子,它们都是不可观测的随机变量,$A=(a_{ij})_{p\times m}$叫做因子载荷阵。$f$ 的均值向量 $E(f)=\mathbf{0}$,协方差 $\mathrm{cov}(f)=I$,$E(\varepsilon)=\mathbf{0}$,且 ε 与 f 相互独立,ε 的协方差矩阵是对角阵。

(三)主成分与因子分析的联系

主成分分析与因子分析都是绛维的分析方法,利用少数几个变量对数据进行解释。主成分分析是一种数据的变换,而不假定数据阵有什么样的结构形式;因子分析可以看成是一种模型分析,当模型的某些条件不满足时,因子分析可能是虚假的。主成分分析的重点放在从观测变量到主成分的变换上,因子分析重点放在从基本因子到观测变量的变换

上,主成分变换是可逆的,因子分析则不要求。当特殊因子的变差为 0 时,主成分分析和因子分析是完全等价的。那么对于一个因子分析模型怎么估计其因子载荷矩阵 A,实践中有很多方法,其中有一种就是上述的主成分分析方法,从公式 $F = U'ZX$ 我们可以得到 $ZX = UF$,具体表达式为:

$$\begin{aligned} ZX_1 &= u_{11}F_1 + u_{21}F_2 + \cdots + u_{p1}F_p \\ ZX_2 &= u_{12}F_1 + u_{22}F_2 + \cdots + u_{p2}F_p \\ &\cdots\cdots \\ ZX_p &= u_{1p}F_1 + u_{2p}F_2 + \cdots + u_{pp}F_p \end{aligned} \quad (5)$$

对上面的等式(5)只保留前 $m(m < p)$ 个主成分,而把后面的部分用 ε_i 代替,则

$$ZX_1 = Z\hat{X}_1 + \varepsilon_1 = u_{11}F_1 + u_{21}F_2 + \cdots + u_{m1}F_m + \varepsilon_1$$

$$ZX_2 = Z\hat{X}_2 + \varepsilon_2 = u_{12}F_1 + u_{22}F_2 + \cdots + u_{m2}F_m + \varepsilon_2$$

$$\cdots\cdots$$

$$ZX_p = Z\hat{X}_p + \varepsilon_p = u_{1p}F_1 + u_{2p}F_2 + \cdots + u_{mp}F_m + \varepsilon_p \quad (6)$$

其中:$\varepsilon_i = u_{m+1,i}F_{m+1} + \cdots + u_{pi}F_p$,($i = 1,2,\cdots p$)。

当主成分 $F_1,F_2,\cdots,F_p$ 是从标准化后的相关系数矩阵求出,各成分相互独立,且其方差按大到小的排序为 $\lambda_1,\lambda_2\cdots\lambda_p$,我们将式(6)做 $F_i/\sqrt{\lambda_i} \triangleq f_i$,$u_{ji}\sqrt{\lambda_j} \triangleq a_{ij}$(符号 $\triangleq$ 表示记为的意思)。通过上述变换,我们就能得到与式(4)类似的因子模型表达式。注意这里的 $\varepsilon_1,\varepsilon_2,\cdots\varepsilon_p$ 与式(4)的 $\varepsilon_1,\varepsilon_2,\cdots\varepsilon_p$ 是有区别的,式(4)要求它们相互独立而这里它们之间不独立,为了方便还是用原符号表示。

实际上对于主成分分析 SPSS 软件中没有对应的模块,但是因子分析模块中有利用主成分分析来求得因子载荷矩阵,根据上面主成分分析与因子分析的联系,我们可以从 SPSS 的因子载荷矩阵得到主成分分析的系数。由于主成分分析所得到的特殊因子 $\varepsilon_1,\varepsilon_2,\cdots\varepsilon_p$ 并不独立,因此所得的因子载荷并不完全正确。但是当共同度较大时,特殊因子所起的作用较小,那么特殊因子之间的相关性所带来的影响就几乎可以忽略不计,这时主成分分析和因子分析是完全等价的(公因子的数目与变量一样多)。这时可以利用式(4)中的 a_{ij} 反推出式(3)中的 u_{ij},它们之间的关系是 $a_{ij} = u_{ji}\sqrt{\lambda_j}$,也就是推出了从因子载荷矩阵得到主成分系数表达式。

三、主成分分析的 SPSS 实现

本文利用 SPSS 软件自带的数据集 Employee data 为例说明如何利用因子分析模块得到主成分系数。数据集 Employee data 为 Midwestern 银行在 1969 – 1971 年之间雇员情况的数据,共包括 474 条观测及如下 10 个变量:Id(观测号)、Gender(性别)、Bdate(出生日期)、Educ(受教育程度(年数))、Jobcat(工作种类)、Salary(目前年薪)、Salbegin(开始受聘时的年薪)、Jobtime(受雇时间)、Prevexp(受雇以前的工作时间)、Minority(是否少数民族)。我们将 educ、salary、salbegin、jobtime、prevexp 依次表示为 X_1,X_2,X_3,X_4,X_5。数据在同向化的基础上 SPSS 中的因子分析默认针对标准化后的数据来分析的,所以利用 Analyze → Descriptive Statistics → Descriptives…进入描述性统计对话框,依次选中变量 X_1,X_2,X_3,X_4,X_5 并点向右的箭头按钮,这五个变量便进入 variables 窗口,选中 Save standardized as variables 复选框,点击 OK 按钮,即可在数据窗口得到标准化的数据 ZX_1,ZX_2,ZX_3,ZX_4,ZX_5。接下来对标准化后的数据进行分析,点击 Analyze→Data Reduction→Factor…进入 Factor Analysis(因子分析)对话框。依次选中变量 ZX_1,ZX_2,ZX_3,ZX_4,ZX_5(用原始数据也是一样,标准化主要是在主成分表达中需要)并点向右的箭头按钮,这五个变量便进入 variables 窗口,点击右侧的 OK 按钮,即可得到表 1、表 2 和表 3。

表 1　Communalities(共同度)

	Initial	Extraction
ZX1	1.000	0.754
ZX2	1.000	0.896
ZX3	1.000	0.916
ZX4	1.000	0.999
ZX5	1.000	0.968

表 2　Total Variance Explained(总方差解释部分)

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	2.477	49.541	49.541
2	1.052	21.046	70.587
3	1.003	20.070	90.656
4	0.365	7.299	97.955
5	0.102	2.045	100.000

表 3　Component Matrix(因子载荷矩阵)

	Component		
	1	2	3
ZX1	0.846	-0.194	-0.014
ZX2	0.940	0.104	0.029
ZX3	0.917	0.264	-0.077
ZX4	0.068	-0.052	0.996
ZX5	-0.178	0.965	0.069

表 1 中的 Communalities(共同度)数据给出了该次分析从每个原始变量中提取的信息(特征根大于 1),可以看到除受教育程度(ZX_1)信息损失较大外,主成分几乎包含了各个原始变量至少 90% 的信息。表 2 中的 Total Variance Explained(总方差解释部分)则显示了各主成分解释原始变量总方差的情况,SPSS 默认保留特征根大于 1 的主成分,在本例中看到当保留 3 个主成分为宜,这 3 个主成分集中了原始 5 个变量信息的 90.66%,可见效果是比较好的。表 3 中的 Component Matrix(因子载荷矩阵)给出了标准化原始变量用公因子线性表示的近似表达式,提取三个公因子时的因子模型可以表示为:

$$ZX_1 = 0.846f_1 - 0.194f_2 - 0.014f_3 + \varepsilon_1$$
$$ZX_2 = 0.940f_1 + 0.104f_2 + 0.029f_3 + \varepsilon_2$$
$$\cdots\cdots$$
$$ZX_5 = -0.178f_1 + 0.965f_2 + 0.069f_3 + \varepsilon_p$$

根据上面的因子载荷系数 a_{ij} 与主成分系数 u_{ij} 之间的关系 $a_{ij} = u_{ji}\sqrt{\lambda_j}$,也就推出了从相关系数矩阵得到的主成分系数表达式:

$$F_1 = (0.846ZX_1 + 0.940ZX_2 + \cdots - 0.178ZX_5)/\sqrt{2.477}$$

$$F_2 = (-0.194ZX_1 + 0.104ZX_2 + \cdots + 0.965ZX_5)/\sqrt{1.052}$$

$$F_3 = (-0.014ZX_1 + 0.029ZX_2 + \cdots + 0.069ZX_5)/\sqrt{1.003}$$

实际中我们通常只选取前几个主成分,例如 F_1,F_2,F_3 来反映原 p 个变量信息。主成分系数还可以通过进入 Factor Analysis 对话框并选择好变量之后,点击对话框下部的 Scores 按钮进入 Factor Scores 对话框,选择 Display factor score coefficient matrix 选项,并按 Continue 继续,最后点击 OK 按钮运行,也可以推出主成分的系数,具体参见何晓群教授的多元统计分析。作者又应用 SAS 软件(从相关系数出发)得到的结果与上述的结果一样。

四、结论

从上面的分析可以看出,因子分析和主成分分析都依赖于原始变量,所以原始变量的选择很重要(指标的选择非常重要)。如果原始变量都本质上独立,那么降维就可能失败,这是因为很难把很多独立变量用少数综合的变量概括。数据越相关,降维效果就越好。其次,对于具体的问题指标选取之后还要对其处理,正向指标、逆向的指标和区间型指标怎样转换成可以比较的指标问题。最后,从相关系数出发建立主成分的系数矩阵还是从协方差矩阵出发建立主成分的系数还没有定论。因子分析中的特殊因子如果作用较大,不能从因子载荷阵推主成分系数。可见建立主成分模型的事前步骤和事后分析很重要,不是随便什么数据拿来用 SPSS 软件分析得出结果就行了。

参考文献

[1] 郭显光. 如何用 SPSS 软件进行主成分分析 [J]. 统计与信息论坛. 1998(2):60-64.

[2] 何晓群. 多元统计分析 [M]. 北京:中国人民大学出版社,2004.

[3] 张润楚. 多元统计分析 [M]. 北京:科学出版社,2006.

[4] 林海明、张文霖. 主成分分析与因子分析的异同和 SPSS 软件 [J]. 统计研究,2005(3):65-69.

[5] Mardia, K. V., Kent, J. T, Bibby, J. M. Multivariate Analysis [M]. London: Academic Press Inc, 1982.

[6] Landau, S., Everitt, B. A Handbook of Statistical Analysis Using SPSS [M]. Florida: Chapman & Hall, 2004.

作者简介

李小胜,男,1976 年生,安徽枞阳人,汉族,2008 年毕业于厦门大学计划统计系。博士,安徽财经大学统计与应用数学学院,讲师。研究方向:经济统计。

陈珍珍,女,1950 年生,福建厦门人,厦门大学计划统计系教授,博士生导师。研究方向:抽样技术。

(责任编辑:周　晶)