·科研方法专题·

基于 SAS 与 R 软件的主成分分析

胡良平[1,2*]

(1. 军事科学院研究生院,北京 100850;
2. 世界中医药学会联合会临床科研统计学专业委员会,北京 100029
*通信作者:胡良平,E-mail:lphu812@sina.com)

【摘要】 本文目的是介绍主成分分析的概念、作用和用软件实现计算的方法。应弄清适合进行主成分分析的数据结构、如何利用 SAS 和 R 软件实现计算的具体方法,尤其是计算结果的解释和利用。值得注意的是:满足同质性的单组设计多元定量资料是适合进行主成分分析的数据结构的突出特点;主成分分析可用于下列场合:数据降维、主成分回归分析和主成分聚类分析等。

【关键词】 主成分分析;降维;相关矩阵;特征值;特征向量

中图分类号:R195.1 文献标识码:A doi:10.11886/j.issn.1007-3256.2018.02.007

Principal components analysis based on SAS and R software

Hu Liangping[1,2*]

(1. *Graduate School, Academy of Military Sciences PLA China, Beijing* 100850, *China*;
2. *Specialty Committee of Clinical Scientific Research Statistics of World Federation of Chinese Medicine Societies, Beijing* 100029, *China*
** Corresponding author: Hu Liangping, E-mail: lphu812@sina.com*)

【**Abstract**】 This paper aimed at introducing the concepts, functions and calculation based on software of the principal components analysis. It is important to find out the data structure which are suitable for the principal components analysis and the concrete calculation approach by SAS and R software specially the methods of interpretation and utilization of the calculated results. It is remarkable that the data structure with homogenization should be suitable for implementing the principal components analysis, which can be used in the following situations, such as the data dimensionality reduction, the principal components regression analysis and the principal components cluster analysis and so on.

【**Keywords**】 Principal components analysis; Dimensionality reduction; Correlation matrix; Eigenvalue; Eigenvector

1 概 述

1.1 基本概念

在科学研究中,经常需要从同一个体(或观测单位)上观测多个指标,这些指标从不同方面反映个体的性质。但指标太多,不仅会增加计算的复杂性,也会给合理分析问题和解释问题带来困难。表面上,各指标之间地位相同。实际上,各指标所包含的信息量参差不齐,且指标间往往不是相互独立的,它们所包含的信息有交叉或重叠的部分。所以,需要对众多指标进行适当的处理,以便更好地反映事物的本质特征。

1.2 何为主成分分析

主成分分析(principal components analysis)是将多个定量指标转换为少数几个综合指标的一种统计分析方法。它是将彼此相关的一组变量转化为彼此独立的一组新变量,并以其中少数的几个新变量综合反映原先多个变量所包含的主要信息,且这少数几个综合变量具有独特的专业含义。主成分变量实际上就是由原变量 $X_1 \sim X_m$ 线性组合出来的 m 个互不相关、且未丢失任何信息的新变量,也称为综合变量。

1.3 主成分分析的作用

多指标的主成分变量常被用来揭示某种事物或现象内在规律性的综合指标,研究者结合基本常识和专业知识对综合指标所蕴藏的信息予以恰当解释,就可以更深刻地揭示事物的内在规律。主要应用于以下三个方面:①降维,即利用较少的几个主成分变量就可以取代原来众多的变量所承载的信息;②基于消除多重线性回归分析中自变量间共线性关系之后的主成分变量再进行回归分析,即所谓的“主成分回归分析”;③应用于综合评价领域,就是基于综合评价指标在各个体上的“取值或得分”对

项目基金:国家高技术研究发展计划课题资助(2015AA020102)

全部个体或观测单位进行排序,还可进一步对其进行分档。这种做法和结果事实上就是将原先的“无序样品”转变成“有序样品”,此时,就相当于对“有序样品”进行聚类分析了。

1.4 适合进行主成分分析的数据结构[1]

1.4.1 问题与数据结构

【例 1】某文献计量学家收集到 23 种肿瘤类期刊的载文量(X_1)、基金论文比(X_2)、总被引频次(X_3)、影响因子(X_4)、5 年影响因子(X_5)、即年指标(X_6)、被引半衰期(X_7)和 Web 即年下载率(X_8)8 个指标的具体数据。见表 1。

1.4.2 对数据结构的分析

在表 1 中,23 种期刊都是肿瘤学方面的期刊,故可认为它们具有“同质性(简单地理解,就是具有可比性)”;$X_1 \sim X_8$ 这 8 个计量指标都是用来反映每种学术期刊的影响力、知名度、学术和社会价值等,而且,这些指标的取值都是越大越好,即所谓的“高优指标”。显然,从“性质”上来看,这些指标也是具有“同质性(简单地理解,就是具有可比性)”的。满足以上两方面(横向被称为“样品”、纵向被称为“变量”)要求的资料,称为“单组设计多元定量资料”。

表 1 23 种肿瘤类期刊的文献计量学指标及其取值

刊 名	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
中华肿瘤杂志	234	0.35	2705	1.415	1.394	0.120	6.0	38.900
癌症	316	0.49	1935	0.742	0.879	0.104	4.3	35.200
中国肿瘤临床	507	0.33	1710	0.420	0.673	0.026	5.3	22.300
中华放射肿瘤学杂志	102	0.17	942	1.011	1.290	0.029	5.3	7.500
肿瘤	191	0.40	702	0.470	0.525	0.021	4.7	25.000
中国肿瘤	243	0.13	660	0.358	0.367	0.058	3.7	15.500
中国肿瘤临床与康复	255	0.05	595	0.206	0.228	0.020	4.1	7.000
肿瘤防治研究	302	0.25	585	0.280	0.332	0.023	4.7	24.700
实用肿瘤杂志	198	0.17	566	0.326	0.332	0.035	5.3	24.100
实用癌症杂志	251	0.15	546	0.296	0.294	0.012	3.9	19.300
中国癌症杂志	188	0.14	526	0.355	0.419	0.032	3.8	25.400
肿瘤防治杂志	509	0.18	476	0.230	0.244	0.024	3.0	15.800
中国肺癌杂志	172	0.24	412	0.603	0.643	0.058	2.9	21.700
癌变·畸变·突变	120	0.45	341	0.406	0.452	0.167	5.3	31.200
实用肿瘤学杂志	233	0.09	302	0.137	0.220	0.009	5.5	8.600
临床肿瘤学杂志	325	0.06	298	0.318	0.262	0.037	2.8	15.100
中国肿瘤生物治疗杂志	82	0.70	296	0.387	0.459	0.037	4.4	11.200
肿瘤研究与临床	256	0.07	246	0.163	0.163	0.008	3.9	18.000
现代肿瘤医学	336	0.09	243	0.259	0.197	0.063	2.5	12.400
白血病·淋巴瘤	200	0.11	231	0.159	0.202	0.025	4.4	15.800
河南肿瘤学杂志	274	0.04	230	0.100	0.097	0.000	4.6	10.200
肿瘤学杂志	188	0.13	207	0.233	0.186	0.005	3.5	15.400
四川肿瘤防治	143	0.04	110	0.110	0.132	0.000	4.4	12.300

1.4.3 适合选用的统计分析方法

对于前面所呈现的“单组设计多元定量资料”而言,可以选用哪些多元统计分析方法呢?使人惊讶的是:适合分析这种数据结构的多元统计分析方法占据了全部多元统计分析方法的绝大部分。具体来说,需要按以下两种情形来划分:

(1)不提供任何附加信息

可以选择的多元统计分析方法有以下 5 种：①无序样品聚类分析法；②变量聚类分析法；③主成分分析法；④探索性因子分析法；⑤对应分析法。

（2）提供某些附加信息

可以选择的多元统计分析方法有以下 7 种：①单组设计多元方差分析（需要提供各指标的标准值）；②通径分析（需要提供通径图，即依据基本常识和专业知识绘制出变量之间相互依赖关系的图形）；③证实性因子分析［需要提供通径图，即依据基本常识和专业知识绘制出变量之间相互依赖关系的图形，变量包括“显变量（可观测其取值的变量）”与“隐变量（不可观测其取值的变量）”］；④结构方程模型分析［需要提供通径图，即依据基本常识和专业知识绘制出变量之间相互依赖关系的图形，变量包括“显变量（可观测其取值的变量）”与“隐变量（不可观测其取值的变量）”］；⑤多维尺度分析（需要提供任何两个样品之间相似度或不相似度系数，全部系数构成相似度或不相似度矩阵）；⑥典型相关分析（需要依据基本常识和专业知识将全部变量划分为两类）；⑦复相关分析（需要指出一个变量为因变量、其他变量为自变量）。

2 主成分分析的实现

2.1 基于 SAS 实现计算

2.1.1 所需要的 SAS 程序

将表 1 中的 23 行 9 列数据按文本格式存储在“F:\CCC”文件夹中，命名为“23 种肿瘤类期刊文献计量学指标资料.txt”；设所需要的 SAS 程序名为“基于肿瘤类期刊文献计量学指标进行主成分分析.SAS”：

```
data a1;
    infile 'F:\CCC\23 种肿瘤类期刊文献计量学指标资料.txt';
    input name $20. x1-x8;
run;
proc princomp data=a1 out=b1 prefix=z;
    var x1-x8;
run;
```

2.1.2 SAS 程序主要输出结果及解释

相关矩阵

	x1	x2	x3	x4	x5	x6	x7	x8
x1	1.0000	-0.1734	0.2646	-0.1541	-0.1330	-0.1191	-0.2538	0.0690
x2	-0.1734	1.0000	0.4155	0.4419	0.4776	0.5314	0.3214	0.5069
x3	0.2646	0.4155	1.0000	0.8124	0.8148	0.4726	0.4731	0.6521
x4	-0.1541	0.4419	0.8124	1.0000	0.9733	0.5647	0.3884	0.5440
x5	-0.1330	0.4776	0.8148	0.9733	1.0000	0.5058	0.4653	0.4873
x6	-0.1191	0.5314	0.4726	0.5647	0.5058	1.0000	0.2018	0.6784
x7	-0.2538	0.3214	0.4731	0.3884	0.4653	0.2018	1.0000	0.2979
x8	0.0690	0.5069	0.6521	0.5440	0.4873	0.6784	0.2979	1.0000

以上为 8 个计量变量两两之间的 Pearson 相关矩阵。

相关矩阵的特征值

	特征值	差值	比例	累积
1	4.24724987	2.98857463	0.5309	0.5309
2	1.25867524	0.29249791	0.1573	0.6882
3	0.96617733	0.28255764	0.1208	0.8090
4	0.68361969	0.19338266	0.0855	0.8945
5	0.49023703	0.21577832	0.0613	0.9557
6	0.27445871	0.20933759	0.0343	0.9901
7	0.06512112	0.05066010	0.0081	0.9982
8	0.01446102		0.0018	1.0000

以上为相关矩阵的特征值、相邻两特征值之差量、各特征值占总特征值（=8）的比例和累计百分比。

	特征向量							
	z1	z2	z3	z4	z5	z6	z7	z8
x1	-0.047547	0.848001	0.082138	0.234523	0.169319	0.270361	0.325821	0.094669
x2	0.322469	-0.174411	-0.411291	0.301167	0.774864	-0.026423	-0.012271	0.066849
x3	0.421305	0.319188	0.276649	0.047947	0.019984	-0.073226	-0.788615	-0.118797
x4	0.437866	-0.016428	0.218581	-0.432926	0.036042	-0.050005	0.172570	0.734316
x5	0.435547	-0.035764	0.297129	-0.336919	0.162679	0.032867	0.400564	-0.647475
x6	0.353989	-0.003982	-0.531820	-0.119258	-0.358822	0.664815	-0.070554	-0.043710
x7	0.267853	-0.327541	0.442339	0.694887	-0.218242	0.240178	0.165171	0.099216
x8	0.371179	0.199364	-0.364691	0.232311	-0.407886	-0.646219	0.222003	-0.047511

以上为 8 个特征值对应的特征向量。选取几个主要的主成分变量就可近似取代原先 8 个变量信息的直观判断方法见图 1。

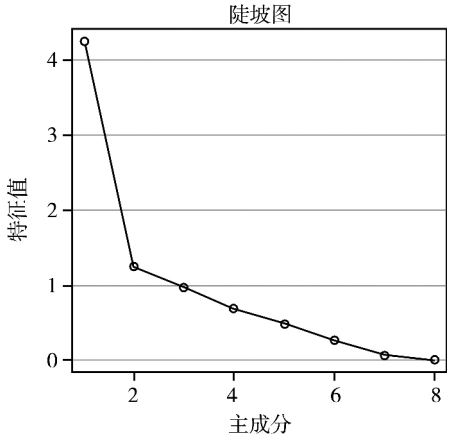

图 1 碎石图

由图 1 可知:在主成分变量为 2 个时,折线出现了明显的“拐点”,也就是说,取前两个主成分变量,就可近似反映原来的 8 个原变量所包含的信息。

各主成分变量携带的信息量占总量 8 的比例见图 2。

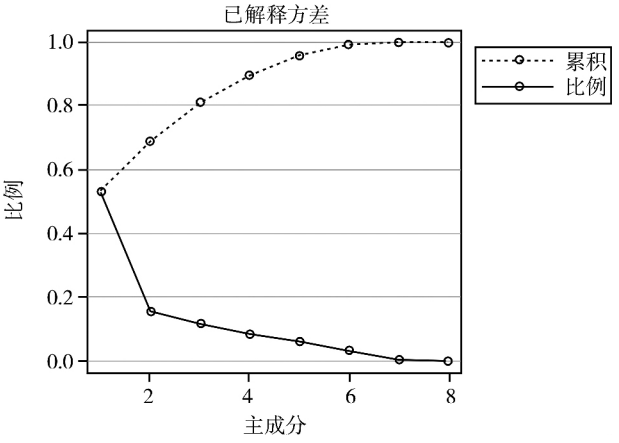

图 2 各主成分变量携带的信息量占总量 8 的比例

由图 2 可知:下面的折线代表各主成分变量携带的信息量占总量 8 的比例,上面的折线代表各主成分变量对应的特征值累积后的结果。

下面写出第一个主成分变量的线性表达式(系数来自“特征向量”第 1 列):

$$z1 = -0.047547x1 + 0.322469x2 + 0.421305x3 + 0.437866x4 + 0.435547x5 + 0.353989x6 + 0.267853x7 + 0.371179x8$$

利用“特征向量”中的系数,可以写出第 2 ~ 8 个主成分变量的表达式。

值得注意的是:“特征向量”中的各列系数都是采取了标准化变换(即每个变量减去其算术平均值除以标准差)而获得的,若希望用原变量表达出来,需要进行相反的变换,此处从略。

2.1.3 如何给主成分变量命名

(1)选取几个主成分变量

应结合特征向量各列的系数,给前几个主要的主成分变量命名。究竟应该关注前几个主成分变量呢?一般采取两种决定方法之一:第一种,选取特征值≥1 的那几个主成分变量;第二种,选取累计贡献率达到 85% 左右时所对应的那几个最大和较大特征值所对应的主成分变量。在本例中,若按前者来选取,就选两个主成分变量;若按后者来选取,就需要选 4 个主成分变量了。

(2)给选取的前两个主成分变量命名

命名的依据:根据各列特征向量的系数的绝对值大小及其左侧变量的专业含义来给各列主成分变量命名。第一主成分变量可以命名为:除“载文量”之外的其他 7 个文献计量指标的综合效应指标;而第二主成分变量可以命名为:“载文量”与“总被引频次”2 个文献计量指标的综合效应指标。

2.2 基于 R 软件实现计算[2]

2.2.1 所需要的 R 程序

将表 1 中的 23 行 9 列数据按文本格式存储在“F:\CCC”文件夹中,命名为“23 种肿瘤类期刊文献计量学指标资料含变量名.txt”;设所需要的 R 程序名为“基于肿瘤类期刊文献计量学指标进行主成分分析.txt”:

```
#设置路径为"F://CCC/"
setwd("F://CCC/")
#下面 data1 中的数据为 23 行 9 列
data1 <- read.table("23 种肿瘤类期刊文献计量学指标资料含变量名.txt",header = TRUE)
#删掉第 1 列:期刊名称
data <- data1[,-1]
attach(data)
#假定已安装 stats 子程序包
#install.packages("survival")
#加载 stats 子程序包
library(stats)
#基于 princomp()函数且相关矩阵进行主成分分析
model1 = princomp(data,cor = TRUE,scores = TRUE)
#系数保留 4 位小数
options(digits =4)
#输出模型 1 的分析结果
summary(model1,loading = TRUE)
#绘制模型 1 的碎石图
screeplot(model1,type = "line",main = "碎石图")
#基于模型 1 且前两个主成分变量绘制各指标的散布图
biplot(model1)
#计算各主成分变量在各样品上的预测值
predict(model1)
```

【R 输出结果】

Importance of components:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
Standard deviation	2.0609	1.1219	0.9829	0.82681	0.70017	0.52389	0.25519	0.120254
Proportion of Variance	0.5309	0.1573	0.1208	0.08545	0.06128	0.03431	0.00814	0.001808
Cumulative Proportion	0.5309	0.6882	0.8090	0.89447	0.95574	0.99005	0.99819	1.000000

以上为第 1 部分输出结果,其中,第 1 行“标准差”实际上就是“特征值的平方根”。

Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
x1		0.848		0.235	0.169	-0.270	-0.326	
x2	-0.322	-0.174	-0.411	0.301	0.775			
x3	-0.421	0.319	0.277				0.789	0.119
x4	-0.438		0.219	-0.433			-0.173	-0.734
x5	-0.436		0.297	-0.337	0.163		-0.401	0.647
x6	-0.354		-0.532	-0.119	-0.359	-0.665		
x7	-0.268	-0.328	0.442	0.695	-0.218	-0.240	-0.165	
x8	-0.371	0.199	-0.365	0.232	-0.408	0.646	-0.222	

以上为第 2 部分输出结果,即“特征向量”,各列中空缺处为“0”。与前面“SAS 输出的特征向量”进行比较,在第一主成分变量上“差距”非常大,很可能是“定义或算法(如:是否采取了坐标轴旋转)”不同所致。选取几个主要的主成分变量就可近似取代原先 8 个变量信息的直观判断方法见图 3。

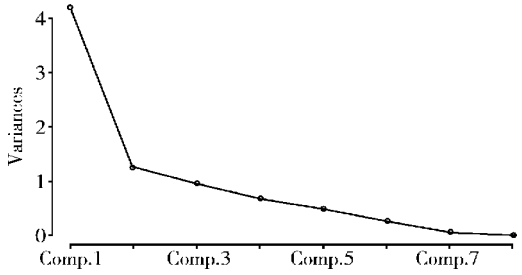

图 3 碎石图

R 软件还可以以第一主成分变量为横坐标轴、以第二主成分变量为纵坐标轴绘制出散布图(因篇幅所限,此图省略),从此图上可看出:在 8 个文献计量学指标中,唯独 x1(载文量)很特别,其他 7 个指标的性质和表现比较接近。

因篇幅所限,各主成分变量在各样品上的预测值(或得分)从略。

【说明】在医学研究中,要谨慎使用主成分分析。关键在于:应注意本文中所提及的“数据结构”。若针对文献[3]的资料,如何使用主成分分析,请读者认真思考。

参考文献

[1] 胡良平. 面向问题的统计学——(3)试验设计与多元统计分析[M]. 北京: 人民卫生出版社, 2012: 19-39.

[2] 李诗羽, 张飞, 王正林. 数据分析: R 语言实战[M]. 北京: 电子工业出版社, 2015: 211-220.

[3] 赵巍峰, 彭敏, 谢博, 等. 健康教育对精神分裂症患者病耻感影响的持续性[J]. 四川精神卫生, 2017, 30(6): 519-523.

(收稿日期:2018-04-02)

(本文编辑:陈 霞)

科研方法专题策划人——胡良平教授简介

胡良平, 男, 1955 年 8 月出生, 教授, 博士生导师, 曾任军事医学科学院研究生部医学统计学教研室主任和生物医学统计学咨询中心主任、国际一般系统论研究会中国分会概率统计系统专业理事会常务理事和北京大学口腔医学院客座教授; 现任世界中医药学会联合会临床科研统计学专业委员会会长、中国生物医学统计学会副会长,《中华医学杂志》等 10 余种杂志编委和国家食品药品监督管理局评审专家。主编统计学专著 45 部, 参编统计学专著 10 部;发表第一作者学术论文220余篇,发表合作论文 130 余篇, 获军队科技成果和省部级科技成果多项; 参加并完成三项国家标准的撰写工作; 参加三项国家科技重大专项课题研究工作。在从事统计学工作的 30 年中, 为几千名研究生、医学科研人员、临床医生和杂志编辑讲授生物医学统计学, 在全国各地作统计学学术报告 100 余场, 举办数十期全国统计学培训班, 培养多名统计学专业硕士和博士研究生。近几年来, 参加国家级新药和医疗器械项目评审数十项、参加 100 多项全军重大重点课题的统计学检查工作。归纳并提炼出有利于透过现象看本质的“八性”和“八思维”的统计学思想, 独创了逆向统计学教学法和三型理论。擅长于科研课题的研究设计、复杂科研资料的统计分析与 SAS 实现、各种层次的统计学教学培训和咨询工作。