DOI:10.16506/j.1009-6639.2011.12.003

【论著】

定性数据的主成分分析及其 SAS 实现

徐俊芳，李石柱，贾铁武，刘琴，周晓农

【摘要】 **目的** 探讨主成分分析在定性数据分析中的应用及其 SAS 实现。**方法** 采用分层整群抽样方法，随机抽取某血吸虫病流行区 1 247 户家庭进行问卷调查，并应用主成分分析法对居民家庭经济状况进行综合评价。**结果** 21 个家庭经济状况的变量主要受前 6 个主成分的影响，6 个主成分依次主要综合了一般的日常生活用品、家庭收支情况、家庭固有资产或不动产、高消费家庭生活用品、家庭农用机械及耕牛的信息。**结论** 进行主成分分析时，应根据数据类型选用合适的主成分分析方法，可以获得更为客观的分析结论。

【关键词】 定性数据；主成分分析；评价

中图分类号：R195.1 文献标识码：A 文章编号：1009-6639（2011）12-0991-04

Application of principal component analysis for the qualitative data related to family economic conditions

XU Jun-fang, LI Shi-zhu, JIA Tie-wu, LIU Qin, ZHOU Xiao-nong. National Institute of Parasitic Diseases, Chinese Center for Disease Control and Prevention; Key Laboratory of Parasite and Vector Biology, MOH; WHO Collaborating Center for Malaria, Schistosomiasis and Filariasis, Shanghai 200025, *China*

Corresponding author: ZHOU Xiao-nong, E-mail: xiaonongzhou1962@gmail.com

【**Abstract**】 **Objective** To explore the application of principal component analysis for qualitative data related to family conditions. **Methods** The method of stratified cluster sampling was used to select 1 247 households from a schistosomiasis endemic area. An informed consent survey was carried out for collecting the data related to family conditions. The factors associated with family economic conditions were evaluated by principal component analysis (PCA) with the SAS statistical package version 9.1 (SAS Institute, Cary, NC). **Results** A total of 21 variables associated with family economic conditions were collected for the study. The method of PCA found six principal family condition associated components, which included daily necessaries, household income and expenditure, family fixed assets, expensive life supplies, farm machinery tools and cattle. **Conclusion** The method of principal component analysis is appropriate for evaluating the major influencing factors of family economic conditions.

【**Key words**】 Qualitative data; Principle component analysis evaluation; Economic conditions

主成分分析则是从原始变量之间相互关系入手，寻找少数综合变量以概括原始变量信息，从这些复杂的观测变量中提取全面、可靠的信息，从而对研究事物或对象作出客观、正确的分析和评价。本文主要通过应用主成分分析法综合评价血吸虫病流行区村民家庭经济状况，探讨定性数据的主成分分析及其 SAS 实现。

基金项目：国家重大专项项目（2008ZX10004－011）；国家“十一五”科技支撑计划（2007BAC03A02）；上海市优秀学术带头人计划（11XD1405400）

作者单位：中国疾病预防控制中心寄生虫病预防控制所，世界卫生组织疟疾、血吸虫病和丝病虫合作中心，卫生部寄生虫病原与媒介生物学重点实验室，上海 200025

作者简介：徐俊芳，博士，主管医师，主要从事疾病控制工作

通讯作者：周晓农，E-mail：xiaonongzhou1962@gmail.com

1 主成分分析方法及 SAS 程序介绍

1.1 主成分分析法的基本原理及主要作用

主成分分析（principle component analysis）又称主分量分析，它是一种降维技术的多元统计方法[1]，它借助于一个正交变换，将其相关的原始变量转化成相互间不相关的随机主成分，其在代数上表现为将原随机变量的协方差变换成对角形矩阵，在几何上表现为坐标旋转的过程，即将原坐标系变换为新的正交坐标系，使之指向样本点散布最开的 n 个正交方向，然后对多维变量进行降维处理，把多个原始变量转化为少数几个综合变量（即主成分）。这些主成分是原始变量的线性组合，能够反映原始变量的大部分信息，为使这些主成分所包含的信息互不重叠，要求各主成分之间互不相关。主成分分析法在实际应用中的主要作用为[2-3]：（1）数据的预处理：包括对原始

变量降维和变量筛选；（2）多维数据的图形表示方法；（3）探索变量间的关系。

1.2 主成分分析的数学表达式及其性质

设有 n 个变量 x_1，x_2，…，x_n，欲寻找可以概括这 n 个变量主要信息的综合指标 Z_1，Z_2，…，Z_n。从数学上讲，就是寻找一组常数 ai_1，ai_2，…，a_{in}（$i=n$），使这 n 个变量的线性组合能够概括 n 个原始变量 x_1，x_2，…，x_n，的主要信息（式 1）：

$$\begin{cases} Z_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ Z_2 = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ \quad\vdots \qquad\qquad \vdots \\ Z_n = a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{mn}x_n \end{cases} \quad （式 1）$$

主成分 Z_1，Z_2，…，Z_n 有以下几个性质[4]：

（1）主成分间互不相关，即对任意 i 和 j，Z_i 和 Z_j（$i=n$，$j=n$）的相关系数：

$$Corr(Z_i, Z_j) = 0,\ i \neq j,\ i \neq j \quad （式 2）$$

（2）线性组合常数（a_{i1}，a_{i2}，…，a_{in}）构成的向量为单位向量，即

$$a_{i1}^2 + a_{i2}^2 + \cdots + a_{in}^2 = 1 \quad （式 3）$$

（3）各主成分的方差依次递减（或不增）（式 4），总方差不增不减（式 5），主成分是对原始变量信息的一种改组，主成分不增加总信息量，也不减少总信息量。

$$Var(Z_1) \geqslant Var(Z_2) \geqslant \cdots \geqslant Var(Z_n) \quad （式 4）$$

$$\begin{aligned} &Var(Z_1) + Var(Z_2) + \cdots + Var(Z_n) \\ &= Var(x_1) + Var(x_2) + \cdots + Var(x_n) \\ &= n \end{aligned} \quad （式 5）$$

1.3 定性数据的主成分分析及 SAS 程序

1.3.1 定性数据的主成分分析步骤

（1）原始数据的变量变换。经变量变换后，使得变换变量的协方差或相关矩阵的特征最优。在 SAS（statistics analysis system）中，定性数据的主成分分析过程提供了 3 种变换变量的方法[5]：方差总量最大化法（maximum total variance，MTV）、广义方差最小化法（minimum generalized variance，MGV）、平均相关系数最大化法（maximum average correlation，MAC）。MTV 法以主成分分析模型为基础，使协方差矩阵的前几个特征值的总和最大化，该算法带有最优化标度的经典主成分分析。MGV 法使用了多重回归迭代算法，使变换变量的协方差矩阵行列式最小化，当研究由非线性变换变量组成的数据矩阵的秩及其线性与非线性相关性时，可用 MGV 法。MAC 法使用了有约束的多重回归迭代算法，使相关矩阵元素的平均值最大化，当所有变量都正相关，或在任何变换中都无单调性约束时，可用 MAC 法；当一些最佳变换有单调递增限制时，则不能对负相关的变量使用 MAC 法。MAC 法可用作 MTV 法和 MGV 法的初始化算法。

（2）估计主成分，确定主成分个数。计算变换变量的标准化变量、变量间的相关系数、相关系数矩阵及特征向量。确定主成分个数的准则[6]：一是根据主成分的累计贡献率来确定，即累计贡献率为 70%～85%；另一个是根据特征根来确定，即特征根≥1。通常来说，根据累计贡献率确定的主成分个数较多，而根据特征根确定的主成分个数则较少。在实际应用中，应根据确定主成分个数的准则和实际意义来确定主成分个数。

（3）解释主成分实际意义，计算主成分得分。主成分是标准化指标变量的一个线性组合，其得分系数描述了各变量对主成分的影响作用，主成分的实际意义根据得分系数的绝对值、专业知识进行解释。主成分得分可用于进一步的统计分析中，计算公式如下：

$$z_i = \frac{a_{i1}}{S_1}X'_1 + \frac{a_{i2}}{S_2}X'_2 + \cdots + \frac{a_{in}}{S_1}X'_n - \left(\frac{a_{i1}}{S_1}\overline{X}'_1 + \frac{a_{i2}}{S_2}\overline{X}'_2 + \cdots + \frac{a_{in}}{S_n}\overline{X}'_n\right) \quad （式 6）$$

式 6 中 a_{in} 是得分系数，X'_n 是原始变量 x 的变换变量，S'_n 是变换变量 X'_n 的标准差，$\overline{X}'_n$ 是变换变量 X'_n 的均数。

1.3.2 定性数据主成分分析的 SAS 程序

在 SAS 中，实现主成分分析的过程有 proc prinqual 和 proc princomp 两个过程，前者主要应用于定性数据的主成分分析，可进行线性和非线性的变量变换，使用交替最小二乘法使变换变量的协方差或相关矩阵的特征最优[3]，该过程产生少量输出结果，只产生一个输出数据集（包含分析结果的输出数据集）。后者主要用来拟合主成分模型，可产生两个输出数据集（一个是包含原始数据和主成分得分的输出数据集，另一个是包含均值、标准差、观测个数、相关矩阵或协方差矩阵，特征值和特征向量的输出数据集）。

2 实例分析

2.1 资料来源

以湖北省江陵县血吸虫病流行区的居民为研究对象，以不同疫情程度为层、村为群，采用分层整群抽样的方法，抽取该县 6 个村，共 1 247 户 2 339 名村民进行问卷调查。调查内容包括与血吸虫病相关的个人高危行为、接受血吸虫病防治健康教育的情况、家庭卫生经济状况等项目。本文主要应用主成分分析法对家庭经济状况（21 个原始变量）进行综合评价，详见表 1。

2.2 分析方法及 SAS 程序

应用软件 SAS9.1.3 实现主成分分析，调用 proc

表 1 血吸虫病流行区村民家庭经济状况赋值

变量	变量意义	变量赋值	变量	变量意义	变量赋值
code1	家庭编码		aircondition	空调	0=无，1=有
cattle	耕牛	0=无，1=有	bicycle	自行车	0=无，1=有
house	楼房	0=无，1=有	motor	摩托车	0=无，1=有
phone	电话	0=无，1=有	tractor	拖拉机	0=无，1=有
tv	彩电	0=无，1=有	thresher	打谷机	0=无，1=有
vcd	影碟机	0=无，1=有	till	耕整机	0=无，1=有
fan	电风扇	0=无，1=有	machine	大型机械	0=无，1=有
pan	电饭锅	0=无，1=有	field	农田	0=无，1=旱田，2=水田，3=水旱
washing	洗衣机	0=无，1=有	earning	主要收入	1=农作物，2=务工，3=其他
shrink	甩干机	0=无，1=有	in_ex	收支情况	1=负债，2=基本平衡，3=有积蓄
fridge	电冰箱	0=无，1=有	expenses	主要开支	1=生产开支，2=生活开支，3=医疗费

prinqual 过程对原始变量进行变量变换，proc princomp 完成主成分分析。其调用命令格式如下[5]：

proc prinqual data=sch. family /指出被分析的数据集*/out=sch. family_trans replace; /* 给包含分析结果的数据集命名*/transform ops（cattle house phone tv vcd fan pan washing shrink fridge air condition bicycle motor tractor thresher till machine field earning in_ex expenses)；

/*transform 对括号内列举的变量进行变量变换，ops 要求对每个变量寻找出最佳得分*/id code1；/*在输出数集中用于识别的观测变量*/run；

proc princomp data=sch. family_trans /*指出被分析的数据集，即 proc prinqual 中的输出数据集*/out=sch. pca; /* 给包含 sch. family_trans 数据中的变量及主成分得分的数据集命名*/var cattle house phone tv vcd fan pan washing shrink fridge air-condition bicycle motor tractor thresher till machine field earning in_ex expenses; /*列出要分析的变量*/run。

2.3 结果

proc prinqual 过程将原始变量进行变量变换后，输出到新的数据集，以供 proc princomp 过程使用。proc princomp 过程运行结果显示：从相关矩阵的特征值看，前 6 个主成分的特征值>1；从主成分累计贡献率看，前 6 个主成分的累计贡献率达 70.99%，即，包含了 21 个变量的 70.99%的信息量。根据主成分个数的选取准则可选取前 6 个主成分。由相关矩阵特征向量可知：第 1 主成分主要综合了 phone、tv、vcd、fan、pan、bicycle、motor 等变量的信息，第 2 主成分主要综合了 earning、in_ex、expenses 3 个变量的信息，第 3 主成分主要综合了 house、field 2 个变量的信息，第 4 主成分主要综合了 washing 、shrink、fridge、aircondition 4 个变量的信息，第 5 主成分主要综合了 tractor、thresher、till、machin 4 个变量的信息，第 6 主成分主要综合了 cattle 1 个变量的信息，详见表 2、表 3。从主成分分析结果可知，该血吸虫病流行区村民的家庭经济状况第 1 主成分主要综合了一般家庭一般的日常生活用品信息；第 2 主成分主要综合了家庭收支情况的信息，第 3 主成分主要综合了固定资产或不动产的信息；第 4 主成分主要综合了高消费（或非必备）的家庭生活用品信息，第 5 主成分主要综合了家庭生产工具信息，第 6 主成分主要综合了家庭养牛情况的信息。

表 2 主成分分析的相关矩阵特征值

序码	Eigenvalue	Difference	Proportion	Cumulative
1	7.152 046 41	3.778 093 20	0.340 6	0.340 6
2	3.373 953 21	1.861 523 07	0.160 7	0.501 2
3	1.512 430 14	0.514 729 69	0.072 0	0.573 3
4	0.997 700 45	0.045 734 99	0.047 5	0.620 8
5	0.951 965 46	0.031 837 44	0.045 3	0.666 1
6	0.920 128 02	0.103 253 91	0.043 8	0.709 9
7	0.816 874 11	0.141 955 66	0.038 9	0.748 8
8	0.674 918 45	0.037 041 89	0.032 1	0.781 0
9	0.637 876 56	0.022 913 55	0.030 4	0.811 3
10	0.614 963 01	0.060 839 34	0.029 3	0.840 6
11	0.554 123 67	0.024 861 16	0.026 4	0.867 0
12	0.529 262 51	0.038 366 86	0.025 2	0.892 2
13	0.490 895 65	0.060 263 64	0.023 4	0.915 6
14	0.430 632 01	0.120 871 67	0.020 5	0.936 1
15	0.309 760 34	0.017 282 53	0.014 8	0.950 8
16	0.292 477 81	0.029 610 54	0.013 9	0.964 8
17	0.262 867 27	0.048 129 16	0.012 5	0.977 3
18	0.214 738 11	0.055 459 67	0.010 2	0.987 5
19	0.159 278 44	0.057 843 81	0.007 6	0.995 1
20	0.101 434 63	0.100 247 88	0.004 8	0.999 9
21	0.001 186 75		0.000 1	1.000 0

3 讨 论

一般来说，变量间除了存在线性关系外，还存在大量的非线性关系。经典主成分分析是基于线性函数对有线性关系的定量数据进行分析[7]。而对于定性数据来说，经典主成分分析的线性函数结果则较难解释，有研究者[8-11]针对定性数据的非线性问题，提出了最小二乘非线性主成分分析、logistic非线性主成

表 3　主成分分析的相关矩阵特征向量

变量	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6
phone	0.372 352	−0.115 289	0.027 352	0.097 167	0.098 741	−0.171 354
tv	0.302 705	0.106 832	0.078 247	0.101 359	0.054 507	0.121 195
vcd	0.290 153	0.056 432	−0.159 253	−0.071 829	−0.034 173	0.104 116
fan	0.361 355	0.080 474	0.076 926	−0.149 559	0.021 268	−0.196 902
pan	0.322 682	0.116 373	0.033 278	0.106 038	−0.039 532	−0.137 243
bicycle	0.290 653	0.172 734	−0.215 301	−0.026 177	−0.175 145	0.087 878
motor	0.343 911	0.083 497	−0.176 951	−0.126 155	0.169 688	0.049 857
earning	0.129 196	0.510 564	0.134 159	0.059 612	0.149 249	0.041 746
in _ex	0.109 864	0.436 241	−0.073 751	−0.110 273	−0.227 933	−0.100 039
expenses	0.134 709	0.384 855	0.152 368	−0.132 535	−0.162 384	0.209 676
house	0.014 641	0.050 794	0.627 687	−0.105 221	−0.156 014	0.131 087
field	−0.092 970	0.201 631	0.536 101	−0.091 662	−0.074 485	−0.118 787
washing	0.176 572	−0.059 002	−0.029 462	0.400 391	−0.184 359	−0.247 811
shrink	−0.262 861	0.109 661	0.145 522	0.317 583	0.102 617	0.107 245
fridge	−0.141 215	−0.228 619	−0.108 955	0.431 681	−0.140 582	0.091 314
aircondition	0.208 318	−0.160 038	−0.041 179	0.556 756	−0.096 622	−0.051 417
tractor	0.119 538	−0.230 201	−0.081 318	0.103 816	0.522 408	0.194 468
thresher	0.074 471	0.032 051	0.231 694	−0.141 524	0.338 489	0.308 242
till	0.062 653	0.035 433	−0.134 879	−0.177 592	0.400 795	−0.245 273
machine	0.013 177	−0.180 761	−0.157 037	0.103 956	0.361 236	−0.206140
cattle	−0.069 058	−0.302 684	0.127 546	−0.203 742	−0.204 719	0.686 228

分分析、最大似然估计的主成分分析等方法。因此，对定性数据进行主成分分析时，有必要进行适当的变量变换，使得变换变量的协方差矩阵或相关矩阵的特征最优。

proc prinqual 过程通过最优化一些变换变量的协方差矩阵或相关矩阵的性质，找出线性和非线性的变量变换，以便改进每个变量对某个主成分模型的拟合。通常进行变量变换时，对于名义变量选用 ops（或 opscore，即最佳得分）选项，以寻找每个变量的最佳得分；对于有序变量则选用 mon（monotone，即单调）选项，以寻找每个变量的单调变换。该过程除了变量变换外，其主要作用还有以下几点[3,5]：（1）将一般的主成分分析推广为一种适用于非量化分析数据的方法；（2）进行度量和非度量的多维选择分析；（3）在进一步数据分析之前，对多元数据的缺失值进行估计，使用调查数据时，该过程可以估计用不同次序所得变量的名义分类的最佳得分；（4）提炼混合的定量和定性数据，并找到它们之间的非线性联系；（5）为了以后进一步使用回归分析、聚类分析和其他分析而缩减变量的个数。proc princomp 过程除了可以完成主成分分析外，还可提示变量量间的共线关系。

本文通过调用 proc prinqual 过程执行原始数据的变量变换，改进了变量对主成分模型的拟合；通过调用 proc princomp 过程实现主成分分析，对家庭经济状况的 21 个变量转化为 6 个具有综合意义的指标变量，由于每个主成分互不相关，因此可将 6 个主成分得分相加，从而获得 1 个综合变量[12]，为进一步资料分析减少了问题的复杂性和难度。

参考文献

[1] Li L. Dimension reduction for high-dimensional data [J]. Methods Mol Biol，2010，620：417-34.

[2] 方积乾主编．医学统计学与电脑实验 [M]．上海：上海科学技术出版社，2006.

[3] 高惠璇．SAS/STAT 软件使用手册 [M]．北京：中国统计出版社，1997.

[4] 孙振球，徐勇勇．医学统计学 [M]．北京：人民卫生出版社，2005.

[5] SAS/STAT 9.1 User's Guide. Available from：http：//support.sas.com/documentation/onlinedoc/91pdf/index_913.html.

[6] 张家放，主编．医用多元统计方法 [M]．武汉：华中科技大学出版社，2002.

[7] Buehler DM，Versteegh MA，Matson K D，*et al*. One problem，many solutions：simple statistical approaches help unravel the complexity of the immune system in an ecological context [J]. PLoS One，2011，6 (4)：e18592.

[8] de Leeuw Jan. Principal Component Analysis of Binary Data. Applications to Roll-Call-Analysis. 2003，http：//www.escholarship.org/uc/item/7n7320n0.

[9] de Leeuw Jan. Nonlinear principal Component Analysis and Related Techniques. 2005，http：//www.escholarship.org/uc/item/8bj075gv.

[10] de Leeuw Jan. Principal component analysis of binary data by iterated singular value decomposition [J]. Computat Statis & Data Anal，2006，50 (1)：21-39.

[11] Lee S，Huang JZ，Hu J. Sparse logistic principal components analysis for binary data [J]. Ann Appl Stat，2010，4 (3)：1579-1601.

[12] 严今石．关于综合评价的多元统计分析方法的探讨 [D]．延边大学，2006：23-27.

（收稿日期：2011-09-19）

（陈继彬　编辑）