基于二次主成分分析模型解决病情确诊问题

许延鑫，熊继平

（浙江师范大学 数理与信息工程学院，浙江 金华 321004）

摘 要：通过主成分分析并结合SPSS软件得到具有高信息含量的A第一主成分和A第二主成分，并分别确定A第一主成分和A第二主成分的函数解析式。在变量基础上增加A第一主成分变量，并再次通过主成分分析得到具有高信息含量的B第一主成分和B第二主成分，并分别确定B第一主成分、B第二主成分和综合主成分的函数解析式，对三者分别进行排序，确定患病与健康的判定指标。

关键词：主成分分析；多因子综合分析；统计回归分析；SPSS技术

中图分类号：R181.2 **文献标识码**：A

Solution of condition diagnosis based on two principal components anatomic model question

XU Yan Xin，XIONG Ji Ping

（Xingzhi College College of Mathematics，Physics and Information Enginerring，ZJNU，Jinhua 321004，China）

Abstract：Through the analysis of principal components and the unification of SPSS software achieve the high information content A as to its first and the second principal component，and justify respectively their functional analysis formula. On the grounds of variable grounds，increase A a first principal component valuable，and with the second analysis on the first principal component as well as the combination of SPSS software emerges high information content B as to its first and the second principal component，and justify respectively their together with the synthesis principal components' functional analysis formulas，after integrating the original judging index about weakness and health，sequence them all.

Key words：principal components analysis；multi-factor generalized analysis；statistical regression analysis；SPSS technology

主成分分析是将多个变量通过线性变换以选出较少个数重要变量的一种多元统计分析方法。在实际课题中，为了全面分析问题，往往提出很多与此有关的变量，因为每个变量都在不同程度上反映这个课题的某些信息。但是，在用统计分析方法研究多变量的课题时，变量个数太多就会增加课题的复杂性。人们希望变量个数较少，同时得到较多的信息。变量之间存在一定的相关关系，当2个变量之间有一定相关关系时，可以解释为这2个变量反映此课题的信息有些重叠。主成分分析是对原先提出的所有变量建立尽可能少的新变量，这些新变量在反映课题的信息方面尽可能保持原有的信息[1]。

人们到医院就诊时，通常要化验指标来协助医生的诊断。诊断就诊人员是否患肾炎时通常要化验人体内各种元素含量，主要包括锌(Zn)、铜(Cu)、铁(Fe)、钙(Ca)、镁(Mg)、钾(K)及钠(Na)。表1是确诊病例的化验结果，其中

表1 确诊病例的化验结果

病例号	Zn	Cu	Fe	Ca	Mg	K	Na	病例号	Zn	Cu	Fe	Ca	Mg	K	Na
1	166	15.8	24.5	700	112	179	513	31	213	19.1	36.2	2220	249	40.0	168
2	185	15.7	31.5	701	125	184	427	32	170	13.9	29.8	1285	226	47.9	330
3	193	9.80	25.9	541	163	128	642	33	162	13.2	19.8	1521	166	36.2	133
4	159	14.2	39.7	896	99.2	239	726	34	203	13.0	90.8	1544	162	98.90	394
5	226	16.2	23.8	606	152	70.3	218	35	167	13.1	14.1	2278	212	46.3	134
6	171	9.29	9.29	307	187	45.5	257	36	164	12.9	18.6	2993	197	36.3	94.5
7	201	13.3	26.6	551	101	49.4	141	37	167	15.0	27.0	2056	260	64.6	237
8	147	14.5	30.0	659	102	154	680	38	158	14.4	37.0	1025	101	44.6	72.5
9	172	8.85	7.86	551	75.7	98.4	318	39	133	22.8	31.0	1633	401	180	899
10	156	11.5	32.5	639	107	103	552	40	156	135	322	6747	1090	228	810

1~30 号病例是已经确诊为肾炎病人的化验结果，31~60 号病例是已经确定为健康人的结果[2]。在论文中列出的数据是原始数据中 1~10 号病例及 31~40 号病例的数据，运用主成分计算时以所有数据为初始数据。

1 主成分分析模型

主成分分析数学原理[3]：设有随机变量 $X_1, X_2, \cdots, X_p$，其样本均数记为 $\bar{X}_1, \bar{X}_2, \cdots, \bar{X}_p$，样本标准差记为 $S1, S2, \cdots, Sp$。首先作标准化变换 $x_i = \frac{X_i - \bar{X}_i}{S_i}$ 有如下的定义：

(1)若 $C_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p$，$a_{11}^2 + a_{12}^2 + \cdots + a_{1p}^2 = 1$，且使 $\mathrm{Var}(C_1)$ 最大，则称 C_1 为第一主成分；

(2)若 $C_2 = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2p}x_p$，$a_{21}^2 + a_{22}^2 + \cdots + a_{2p}^2 = 1$，$(a_{21}, a_{22}, \cdots, a_{2p})$ 垂直于 $(a_{11}, a_{12}, \cdots, a_{1p})$，且使 $\mathrm{Var}(C_2)$ 最大，则称 C_2 为第二主成分；

(3)类似可有第三、四、五主成分，至多有 p 个。

2 模型应用

2.1 问题分析解决

首先将 7 种元素确定为 7 个因子 $X_1, X_2, X_3, X_4, X_5, X_6, X_7$，并将其做标准化变换，随后通过主成分分析并结合 SPSS 软件确定 A 第一主成分 C_1 和第二主成分 C_2，其特征值和特征向量分别如表 2、表 3 所示[4]。符号说明：$X_1(Y_1)$-Zn，$X_2(Y_2)$-Cu，$X_3(Y_3)$-Fe，$X_4(Y_4)$-Ca，$X_5(Y_5)$-Mg，$X_6(Y_6)$-K，$X_7(Y_7)$-Na。

表 2 特征值

	因子			特征值及其方差贡献率/%		
	所有特征值	方差贡献	累积贡献	提取特征值	方差贡献	累积贡献
1	0.201 61	72.67	72.67	0.201 61	72.67	72.67
2	0.033 62	12.12	84.79	0.033 62	12.12	84.79
3	0.020 10	7.25	92.04	0.020 10	7.25	92.04
4	0.013 77	4.96	97.00	0.013 77	4.96	97.00
5	0.006 89	2.48	99.48			
6	0.001 44	0.52	100.00			
7	-0.000 00	0.00	100.00			

表 3 特征向量

	U1	U2	U3	U4	U5	U6	U7
1	0.025 17	0.412 62	-0.140 79	-0.815 70	-0.284 86	0.002 20	0.250 54
2	-0.003 47	0.049 78	0.007 76	0.010 95	0.115 29	0.988 69	0.080 78
3	-0.007 13	0.344 49	0.925 79	0.069 68	-0.047 02	-0.030 33	0.127 28
4	-0.486 18	-0.296 07	-0.023 56	0.150 09	-0.202 66	-0.028 52	0.781 67
5	-0.086 07	0.212 78	-0.067 59	-0.116 65	0.912 26	-0.138 35	0.278 89
6	0.611 51	-0.636 02	0.238 21	-0.310 87	0.112 43	0.003 38	0.235 58
7	0.617 72	0.413 69	-0.247 31	0.443 71	-0.131 99	-0.039 96	0.412 57

因 $C_1 = [X_1\ X_2\ \cdots\ X_7] * [U_{11}\ U_{12}\ \cdots\ U_{17}]^T$，因为特征值的方差贡献率为 72.67 %，表明 C_1 包含原变量中的绝大部分信息，则在原来 7 个因子的基础上引入 C_1 作为第 8 个因子，C_1=[0.70502、0.6341、0.87415、0.80724、0.4212、0.62897、0.37992、0.85489、0.57495、0.71527、-0.74635、0.03003、-0.30047、-0.03826、-0.80605、-1.32826、-0.5588、-0.00363、0.37216、-3.19199]。再将其做标准化变化，再次通过主成分分析并结合 SPSS 软件确定 B 第一主成分 F_1、第二主成分 F_2 和综合主成分 F。根据对这 8 个因子通过 SPSS 的因子分析如表 4、表 5 所示。

表 4 特征值

	因子			特征值及其方差贡献率/%		
	所有特征值	方差贡献	累积贡献	提取特征值	方差贡献	累积贡献
1	3.921	49.011	49.011	3.921	49.011	49.011
2	2.076	25.946	74.957	2.076	25.946	74.957
3	0.724	9.045	84.002			
4	0.624	0.801	91.803			
5	0.311	3.883	95.685			
6	0.218	2.725	98.410			
7	0.127	1.590	100.000			
8	7.67E-010	9.58E-009	100.000			

表 5 特征向量

成分	Zn	Cu	Fe	Ca	Mg	K	Na	C_1
1	0.519	0.774	0.592	0.928	0.903	-0.357	-0.190	-0.928
2	-0.422	0.455	0.347	0.104	0.276	0.792	0.890	0.250

由表 5 可知 C_1 与 5 种元素有着显著的相关性[5]，可见许多变量之间直接的相关性比较强，证明它们存在信息上的重叠。

2.2 主成分表达式

主成分个数提取原则为主成分对应特征值>1 的前 m 个主成分。特征值在某种程度上可以被看成是表示主成分影响力度大小的指标，如果特征值<1，说明该主成分的解释力度还不如直接引入原变量的平均解释力度，因此一般可以用特征值>1 作为纳入标准。通过表 4 可知，提取 2 个主成分，即 m=2。从表 5 可知 C_1、Zn、Cu、Fe、Ca、Mg 在 B 第一主成分上有较高的载荷，说明 B 第一主成分基本反映了这些指标的信息，K、Na 在 B 第二主成分上有较高的载荷，说明 B 第二主成分基本反映了 K、Na 2 个指标的信息。所以提取 2 个主成分是基本反映全部指标的信息，所以决定用 2 个新的变量来代替原来的 8 个变量。通过 SPSS 将表 5 中的数据除以主成分相对应的特征值开平方根，得到两主成分中每个指标所对应的系数。将得到的特征向量与标准化后的数据相乘，然后就可以得到主成分表达式[6]：

$$F_1 = 0.26\times Y_1 + 0.39\times Y_2 + 0.30\times Y_3 + 0.47\times Y_4 + 0.46\times Y_5 - 0.18\times Y_6 - 0.10\times Y_7 - 0.47\times Y_8 \quad (1)$$

$$F_2 = -0.29\times Y_1 + 0.32\times Y_2 + 0.24\times Y_3 + 0.07\times Y_4 + 0.19\times Y_5 + 0.55\times Y_6 + 0.62\times Y_7 + 0.17\times Y_8 \quad (2)$$

由表 4 知 λ_1=3.921，λ_2=2.076，$F = \frac{\lambda_1}{\lambda_1+\lambda_2}F_1 + \frac{\lambda_2}{\lambda_1+\lambda_2}F_2$，得

$$F=0.07\times Y_1+0.36\times Y_2+0.28\times Y_3+0.33\times Y_4+0.36\times Y_5+0.07\times Y_6+0.15\times Y_7-0.25\times Y_8 \quad (3)$$

由(1)、(2)、(3)式得到B第一主成分F_1、B第二主成分F_2和综合主成分F的数据及排名，如表6所示。

由表6可以看出第一主成分中以0为临界值，0.1为修正值，即(-0.1，0.1)为不稳定状态，此状态下的就诊人员将随机被确定为患者和健康者中的1个。而当$F1>0.1$时，将此时对应的就诊人员确定为健康者；当$F1<-0.1$时，将此时的就诊人员确定为患者。经此方法判定的患者与健康者与表1中的患者与健康者基本一致，并且与用综合主成分分析得到的结果基本一致。其判定的准确性可以达到95%以上，因此具备很强的可信性与科学性。

本文创新点在于模型中连续做了2次主成分分析，即二次主成分分析，并伴有大量的数据处理和数据分析，合理的结论背后拥有强大的理论支持和数据支持，具有很强的科学性和可信性。不过，确诊病人还是需要通过医生的具体分析，以达到所需效果。

表6 综合主成分F的数据及排名

病例号	B第一主成分$F1$	排名	B第二主成分$F2$	排名	综合主成分F	排名
40	9.754 9	1	5.180 6	2	8.098 7	1
31	1.057 2	15	-0.678 4	39	0.452 6	15
39	0.623 7	18	-1.263 1	53	-0.027 8	24
36	0.622 2	19	-0.809 8	41	0.125 6	20
34	0.604 7	20	1.534 2	8	0.917 6	8
37	0.564 1	21	-1.425	9	-0.119 52	6
35	0.301 6	24	-0.411 8	30	0.059 3	21
33	-0.045	28	-0.662 4	37	-0.257 3	29
32	-0.103 3	29	0.429 8	16	-0.068 6	25
5	-0.271 1	32	-0.669 3	38	-0.406 4	32
2	-0.638 2	36	-1.412 9	58	-0.899 6	50
3	-0.708 4	37	-1.404 8	56	-0.941 6	52
38	-0.866 8	38	-0.164 8	27	-0.620 5	38
7	-0.946 4	39	-0.440 9	31	-0.768 4	46
4	-0.971 4	40	-1.002 1	47	-0.977 7	55
1	-1.067 4	41	0.046 7	23	-0.676 9	40
6	-1.239 2	46	0.818 1	11	-0.522 8	35
10	-1.263 6	47	-0.868 2	45	-1.118 4	57
8	-1.266 7	48	0.678 6	12	-0.587 8	36
9	-1.687	54	-0.654 8	36	-1.322 2	59

参考文献

[1] 主成分分析[EB/OL].http://baike.baidu.com/view/45376.htm，2009-03.

[2] 北京工业大学数学建模竞赛初赛试题B题[EB/OL].http://www.wendang.com/soft/16922.htm，2008-05.

[3] 主成分分析[EB/OL].http://ec.njue.edu.cn/tjx/wf_dytjfx/slides/chap03.2009-05.

[4] 张文霖.主成分分析在SPSS中的操作和应用[J].理论与分析，2005(12)：31-35.

[5] 王林辉.基于主成分分析的棉花品种综合评价及聚类分析[J].广东农业科学，2009(1)：29-32.

[6] 董寒青.解析SPSS对主成分分析的计算技术[J].知识丛林，2004(3)：117-118.

(收稿日期：2009-06-15)

(上接第56页)

图3所示。

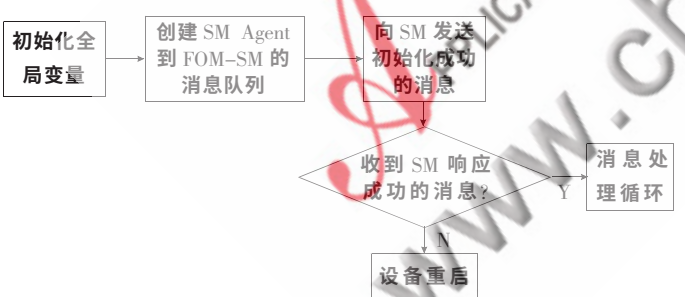

图3 初始化流程图

4.3 MIB访问

在创建代理MIB树时，访问函数已经赋予了MIB树叶节点中的访问函数指针，这样当查找到相应的叶子节点时，就会通过访问函数指针调用相应的访问函数。主要包括Get函数、Next函数、Test函数、Set函数，其中Get函数、Next函数、Set函数分别完成对Get、GetNext、Set命令的响应。

4.4 Trap设计

代理的作用之一就是检查异常事件的发生，并及时向管理站发送Trap消息。Trap设计中要考虑两方面问题：一是定义何种事件可以产生Trap；二是Trap中含有什么信息。在一些标准MIB文件规范中已经定义了产生Trap的事件，而对于企业专有MIB，要根据设备的实际情况确定产生Trap的事件。

在SM Agent的Trap设计中，既有标准MIB文件规范中预定义的Trap事件(表示代理已经开始工作的coldStart)，也有为FOM系统专门设计的Trap事件，如对出现卡的状态变化、系统各路信号的ES、SES、UAS超过设定阈值，各路信号出现信号丢失(LOS)、帧失步(LOF)等情况时给出实时告警。

在PC机上运行MIB Browser软件，通过向目标机发送Get、Set指令，制造FOM预定义的异常现象，观察Trap警告信息，均可以得到正确的信息。

参考文献

[1] 徐钊，杨福锦，郑红党.FOM光纤自愈环网[J].光网络技术，2003(8)：6-8.

[2] 徐钊，张林，李建成.FOM-ADM嵌入式网管的设计[J].光通信技术，2004(5)：32-34.

[3] 李娜，赵兵，翟文艳.FOM嵌入式网管SNMP Agent的设计与实现[J].工矿自动化，2008(2)：44-47.

(收稿日期：2009-06-12)