# 北京建筑大学

## 信息与计算科学专业　实验任务书指导书

**课程名称** 《数据分析》 **实验名称**《主成分分析和聚类分析》**实验地点** 腾讯会议　**指导教师** 王恒友

**学生姓名:** 李金哲　**学号:** 201707010119　**实验日期:** 2020.06.06　**成绩:**

**【实验目的】**

（1）熟悉利用主成分分析进行数据分析，能够使用 SPSS 软件完成数据的主成分分析;

（2）熟悉利用聚类分析进行数据分析，能够运用主成分分析的结果，做进一步分析，如聚类分析、回归分析等，能够使用 SPSS 软件完成该任务。

**【实验要求】**

根据各个题目的具体要求，分别运用 SPSS 软件完成实验任务。

**【实验内容】**

1、表 4.9（数据见 exercise4_5.txt）给出了 1991 年我国 30 个省市、城镇居民的月平均消费数据，所考察的八个指标如下:（单位均为元/人）

X1: 人均粮食支出;　　X2: 人均副食支出;

X3: 人均烟酒茶支出;　　X4: 人均其他副食支出;

X5:人均衣着商品支出;　　X6: 人均日用品支出;

X7: 人均燃料支出;　　X8: 人均非商品支出。

（1）求样本相关系数矩阵 R。

（2）从 R 出发做主成分分析，求出各主成分的贡献率及前两个主成分的累积贡献率;

2、(1）对题 1 中的数据，按照原有的八个指标，对 30 个省份进行聚类，给出分为 3 类的聚类结果。

（2）利用题 1 得到的前 2 个主成分指标，分别按最短距离法（最近邻居距离)、最长距离法（最远邻居距离)、类平均距离法（组间平均距离)、重心距离法；其中距离均采用欧式平方距离，对样本进行谱系聚类分析，并画出谱系聚类图；给出分为 3 类的聚类结果。并与（1）的结果进行比较。

**【实验步骤】**

## 1-1-a 实验过程

1、 选择菜单->【分析】->【降维】-> 【因子】(图 1-1)

2、 将 V1——V9 拖入变量（图 1-2）

3、 点击概述，勾选系数 和 KMO 和巴特利特球形梯度检验（图 1-3）

4、 点击提取，勾选碎石图（图 1-4）

5、 点击旋转，勾选载荷图（图 1-5）

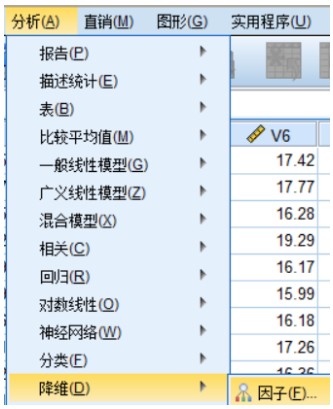


图 1-1 选择菜单


图 1-2 因子分析窗口

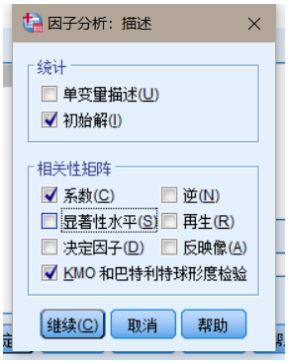


图1-3 描述窗口

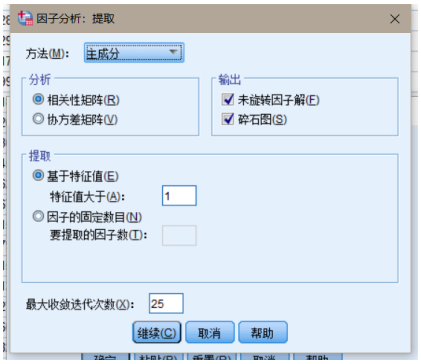


图1-4 提取窗口

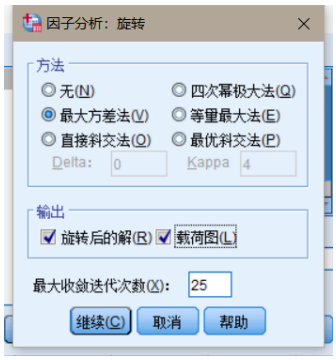


图1-6旋转窗口

# 1-1-b 实验分析

表1-1-1是原有变量的相关系数矩阵。可以看到:存在部分的相关系数都较高，各变量呈较强的线性关系,够从中提取公共因子,适合进行因子分析。由表1-1-2可知,巴特利特球度检验统计量的观测值为142.983,相应的概率P-值接近0。如果显著性水平a为0.05，由于概率P值小于显著性水平a，则应拒绝原假设，认为相关系数矩阵与单位阵有显著差异。同时，KMO值为0.569，根据 Kaiser给出的KMO度量标准可知原有变量适合进行因子分析。

表1-1-1 原有变量相关性系数矩阵

**相关性矩阵**

| | | V2 | V3 | V4 | V5 | V6 | V7 | V8 | V9 |
|---|---|---|---|---|---|---|---|---|---|
| 相关性 | V2 | 1.000 | .334 | -.055 | -.061 | -.289 | .199 | .349 | .319 |
| | V3 | .334 | 1.000 | -.023 | .399 | -.156 | .711 | .414 | .835 |
| | V4 | -.055 | -.023 | 1.000 | .533 | .497 | .033 | -.139 | -.258 |
| | V5 | -.061 | .399 | .533 | 1.000 | .698 | .468 | -.171 | .313 |
| | V6 | -.289 | -.156 | .497 | .698 | 1.000 | .280 | -.208 | -.081 |
| | V7 | .199 | .711 | .033 | .468 | .280 | 1.000 | .417 | .702 |
| | V8 | .349 | .414 | -.139 | -.171 | -.208 | .417 | 1.000 | .399 |
| | V9 | .319 | .835 | -.258 | .313 | -.081 | .702 | .399 | 1.000 |

表1-1-2 巴特利特球度检验

**KMO 和巴特利特检验**

| | | |
|---|---|---|
| KMO 取样适切性量数。 | | .569 |
| 巴特利特球形度检验 | 近似卡方 | 142.983 |
| | 自由度 | 28 |
| | 显著性 | .000 |

表1-2-1中，第一列是因子编号，以后三列组成一组，每组中数据项的含义依次是特征值(方差贡献)、方差贡献率和累计方差贡献率。

第一组数据项(第二列至第四列)描述了因子分析初始解的情况。可以看到：**第1个因子的方差贡献为3.096，解释原有8个变量总方差的38.7%**(即3.096÷8×100%)，**累计方差贡献率为38.7%**；**第2个因子的方差贡献为2.367，解释原有8个变量总方差的29.59%**(即2.367÷8×100%)，**累计方差贡献率为68.294%**[即(3.096+2.367)÷8×100%]．其余数据含义类似。在初始解中由于提取了8个因子，原有变量的总方差均被解释，累计方差贡献率为100%。

第二组数据项(第五列至第七列)描述了因子解的情况。可以看到：由于指定提取2个因子，**2个因子共解释了原有变量总方差的68.294%。总体上，原有变量的信息丢失较少，因子分析效果较理想，**

第三组数据项(第八列至第十列)描述了最终因子解的情况。可见，因子旋转后，总的累计方差贡献率没有改变，也就是没有影响原有变量的共同度，但却重新分配了各个因子解释原有变量的方差，改变了各因子的方差贡献，使因子更易于解释。

表1-2-1 因子解释原有变量总方差的情况

**总方差解释**

| 成分 | 初始特征值 | | | 提取载荷平方和 | | | 旋转载荷平方和 | | |
|---|---|---|---|---|---|---|---|---|---|
| | 总计 | 方差百分比 | 累积 % | 总计 | 方差百分比 | 累积 % | 总计 | 方差百分比 | 累积 % |
| 1 | 3.096 | 38.704 | 38.704 | 3.096 | 38.704 | 38.704 | 3.079 | 38.485 | 38.485 |
| 2 | 2.367 | 29.590 | 68.294 | 2.367 | 29.590 | 68.294 | 2.385 | 29.809 | 68.294 |
| 3 | .920 | 11.500 | 79.794 | | | | | | |
| 4 | .706 | 8.824 | 88.618 | | | | | | |
| 5 | .498 | 6.231 | 94.848 | | | | | | |
| 6 | .230 | 2.874 | 97.722 | | | | | | |
| 7 | .131 | 1.635 | 99.357 | | | | | | |
| 8 | .051 | .643 | 100.000 | | | | | | |

提取方法：主成分分析法。

# 2-1-a 实验过程

1、 选择菜单->【分析】->【分类】-> 【K-均值聚类】(图 2-1-1)

2、 将 V1——V9 拖入变量(图 2-1-2)

3、 点击选项，勾选初始聚类中心和 ANOVA 表 (图 2-1-3)


图 2-1-1选择菜单

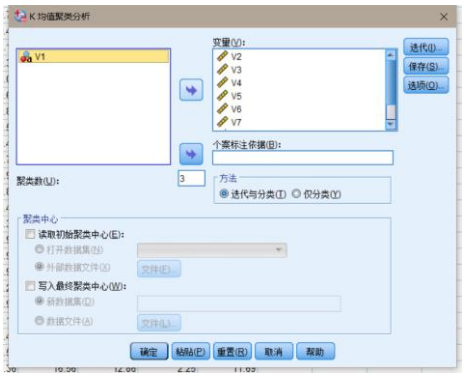


图2-1-2 K-均值聚类

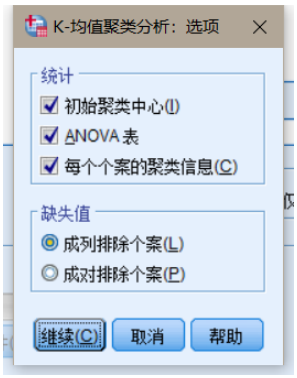


图2-1-3 选项菜单

# 2-1-b 实验分析

表2-1-1 初始聚类中心

**初始聚类中心**

| | 聚类 | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| V2 | 7.94 | 6.25 | 12.47 |
| V3 | 39.65 | 35.02 | 76.39 |
| V4 | 20.97 | 4.72 | 5.52 |
| V5 | 20.82 | 6.28 | 11.24 |
| V6 | 22.52 | 10.03 | 14.52 |
| V7 | 12.41 | 7.15 | 22.00 |
| V8 | 1.75 | 1.93 | 5.46 |
| V9 | 7.90 | 10.39 | 25.50 |

表 2-1-1 展示了 3 个类的初始类中心点的情况。3 个初始类中心点的数据分别是(7.94，39.65，20.97，20.82， 22.52，12.41，1.75，7.90)，(6.25，35.02，4.72，6.28，10.03、7.15，1.93，10.39)，(12.47，76.39，5.52，11.24，14.52，22.00，5.46，25.50)。 **可得第 3 类各指数均是最优的，第 1 类次之，第 2 类各指数最不理想。**

表2-1-2 迭代记录

**迭代历史记录[a]**

| | 聚类中心中的变动 | | |
|---|---|---|---|
| 迭代 | 1 | 2 | 3 |
| 1 | 11.748 | 8.572 | 9.960 |
| 2 | 8.583 | 3.108 | .000 |
| 3 | .729 | .670 | .000 |

| | | | |
|---|---|---|---|
| 4 | 1.053 | .517 | .000 |
| 5 | .878 | .537 | .000 |
| 6 | .000 | .000 | .000 |

a. 由于聚类中心中不存在变动或者仅有小幅变动，因此实现了收敛。任何中心的最大绝对坐标变动为 .000。当前迭代为 6。初始中心之间的最小距离为 26.262。

表2-1-2展示了3个类中心点每次送代时的偏移情况。由表2-1-2可知，第1次送代后，3个类的中心点分别偏移了 11.748，8.572，9.960，**第1类中心点偏移最大**。第2次跌的同理，直到第5次选代后，3个类的中心点的偏移均小于指定的判定标准 (0.02)，聚类分析结束

表2-1-3 最终聚类中心

| 最终聚类中心 | | | |
|---|---|---|---|
| | 聚类 | | |
| | 1 | 2 | 3 |
| V2 | 8.79 | 8.48 | 10.38 |
| V3 | 47.05 | 31.14 | 70.37 |
| V4 | 8.82 | 7.07 | 6.76 |
| V5 | 12.80 | 9.14 | 16.73 |
| V6 | 16.96 | 16.31 | 17.29 |
| V7 | 12.80 | 10.21 | 18.56 |
| V8 | 1.88 | 1.78 | 3.09 |
| V9 | 13.63 | 11.51 | 24.20 |

表2-1-3展示了3个类的最终类中心点的情况。3个最终类中心点的数据分别是(8.79. 47.05. 8.82, 12.80, 16.96, 12.80, 1.88, 13.63)，( 8.48, 31.14, 7.07, 9.14, 16.31, 10.21, 1.78, 11.51)，(10.38, 70.37, 6.76, 16.73, 17.29, 18.56, 3.09, 24.20)。仍然可见，**第3类各指数均是最优的，第1类次之，第3类各指数最。**

表2-1-4 个案统计量

| 每个聚类中的个案数目 | | |
|---|---|---|
| 聚类 | 1 | 10.000 |
| | 2 | 18.000 |
| | 3 | 2.000 |
| 有效 | | 30.000 |
| 缺失 | | .000 |

表2-1-4显示了，聚类2包含的样本最多，聚类3包含的样本最少。通过K聚类分析可以对支出类别情况产生大致的了解。我们可以**把不同地区的人均消费水平大致分成3个类：其中第2类最多，其他两类包含的较少**。具体地区所属分类见表2-1-5.

表2-1-5 聚类成员

**聚类成员**

| 个案号 | 聚类 | 距离 |
|---|---|---|
| 1 | 2 | 7.761 |
| 2 | 2 | 7.653 |
| 3 | 2 | 3.893 |
| 4 | 2 | 4.665 |
| 5 | 2 | 4.147 |
| 6 | 2 | 3.947 |
| 7 | 2 | 5.032 |
| 8 | 2 | 3.655 |
| 9 | 2 | 4.032 |
| 10 | 2 | 5.546 |
| 11 | 2 | 3.908 |
| 12 | 2 | 5.544 |
| 13 | 2 | 6.855 |
| 14 | 2 | 8.246 |
| 15 | 2 | 7.825 |
| 16 | 2 | 5.810 |
| 17 | 1 | 7.904 |
| 18 | 1 | 7.730 |
| 19 | 2 | 6.254 |
| 20 | 2 | 9.152 |
| 21 | 1 | 7.954 |
| 22 | 1 | 9.695 |
| 23 | 1 | 14.545 |
| 24 | 1 | 4.497 |
| 25 | 1 | 3.925 |
| 26 | 1 | 5.448 |
| 27 | 1 | 10.383 |
| 28 | 1 | 18.209 |
| 29 | 3 | 9.960 |
| 30 | 3 | 9.960 |

# 2-2-a 实验过程

1、 选择菜单->【分析】->【分类】-> 【系统聚类】(图 2-2-1)

2、 将 V1——V9 拖入变量（图 2-2-2）

3、 单机方法，然后根据实验方法选择 聚类方法 中的方法（图 2-2-3）


图2-2-1 选项菜单


图2-2-2 系统聚类菜单


图2-2-3 方法菜单

## 2-2-b 实验分析

根据图2-2-1 到 图2-2-4可以分别得出，当选择**最短聚类法**时，可以分**成组别1、组别2和其他**，这三类；选择**类平均距离法**时，只能分成**（组别1，组别2）和其他**，这两类；选择**最长距离法时**，可以分成**（组别1~组别5）和其他**，这两类；选择**重心距离法**时，可以分成**组别1、组别2和其他**，这三类。

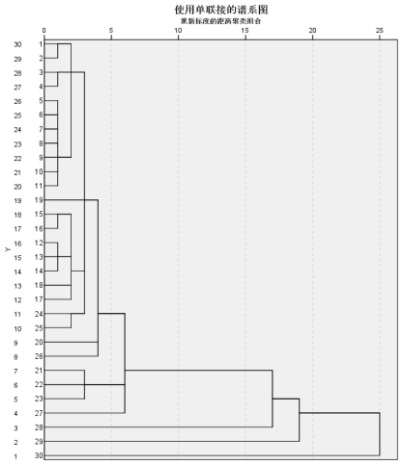


图2-2-1 最短距离法（最近邻居距离）

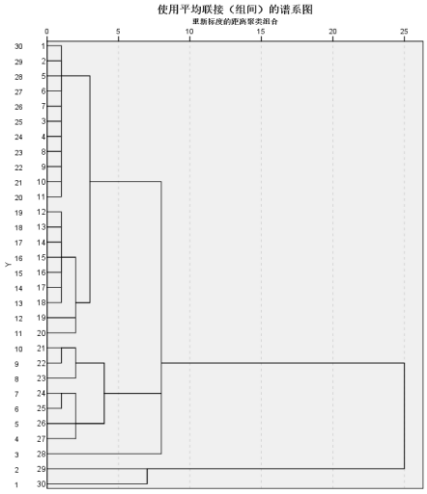


图2-2-3 类平均距离法（组间平均距离）

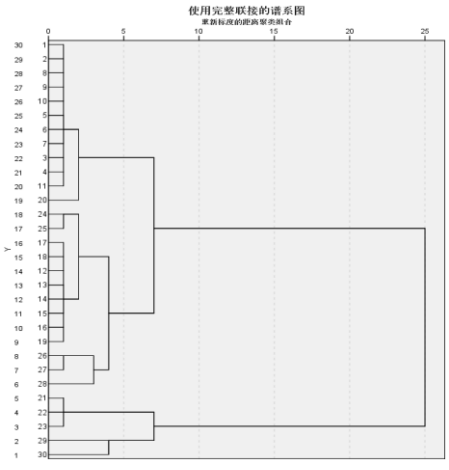


图2-2-2 最长距离法（最远邻居距离）

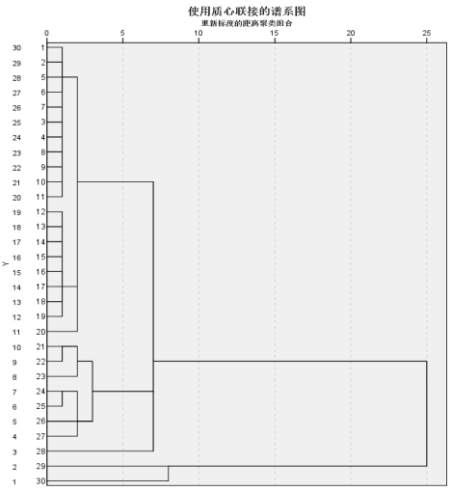


图2-2-4 重心距离法