评价多个对象的水平并排序，指标间关联性很强

主成分分为主成分分析和主成分评价两个方面，分析就是单纯的分析数据是否具有主成分和主成分效果如何，评价就是根据主成分运行的结果直接评价了。

要求：指标之间相关性比较高，一般需要对数据的相关性或者主成分分析的结果进行分析后，如果效果比较好，再使用主成分分析，如果效果不好，那就不要使用主成分分析：切记

主成分简单的说法在视频中中介绍过了，说白了就是在评价的时候有很多指标，因为指标太多了，并且各个指标之间相互有影响，为了消除指标之间的影响，单纯从数据的角度寻找各个指标具有公共特征，，这些公共的特征就是主成分，也就是常说的第一主成分，第二主成分，第N主成分。具体的第一主成分第二主成分以及累计贡献率（要保证所有主成分累计对原始数据的贡献达到80%）是如何计算出的，可以不用掌握，且在比赛的时候也没必要写在论文上，只需要给出主要的结果即可。

这里我们直接给出一个例题：

例 我国各地区普通高等教育发展水平综合评价

近年来，我国普通高等教育得到了迅速发展，为国家培养了大批人才。但由于我国各地区经济发展水平不均衡，加之高等院校原有布局使各地区高等教育发展的起点不一致，因而各地区普通高等教育的发展水平存在一定的差异，不同的地区具有不同的特点。对我国各地区普通高等教育的发展状况进行聚类分析，明确各类地区普通高等教育发展状况的差异与特点，有利于管理和决策部门从宏观上把握我国普通高等教育的整体发展现状，分类制定相关政策，更好的指导和规划我国高教事业的整体健康发展。

指标的原始数据取自《中国统计年鉴，1995》和《中国教育统计年鉴，1995》除以各地区相应的人口数得到十项指标值见表 1。其中：为每百万人口高等院校数；为每十万人口高等院校毕业生数；为每十万人口高等院校招生数；为每十万人口高等院校在校生数；为每十万人口高等院校教职工数；为每十万人口高等院校专职教师数；为高级职称占专职教师的比例；为平均每所高等院校的在校生数；为国家财政预算内普通高教经费占国内生产总值的比重；为生均教育经费。

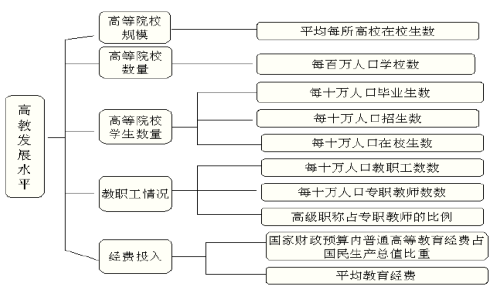


图1 高等教育的十项评价指标

表1 我国各地区普通高等教育发展状况数据

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 地区 |  |  |  |  |  |  |  |  |  |  |
| 北京 | 5.96 | 310 | 461 | 1557 | 931 | 319 | 44.36 | 2615 | 2.20 | 13631 |
| 上海 | 3.39 | 234 | 308 | 1035 | 498 | 161 | 35.02 | 3052 | 0.90 | 12665 |
| 天津 | 2.35 | 157 | 229 | 713 | 295 | 109 | 38.40 | 3031 | 0.86 | 9385 |
| 陕西 | 1.35 | 81 | 111 | 364 | 150 | 58 | 30.45 | 2699 | 1.22 | 7881 |
| 辽宁 | 1.50 | 88 | 128 | 421 | 144 | 58 | 34.30 | 2808 | 0.54 | 7733 |
| 吉林 | 1.67 | 86 | 120 | 370 | 153 | 58 | 33.53 | 2215 | 0.76 | 7480 |
| 黑龙江 | 1.17 | 63 | 93 | 296 | 117 | 44 | 35.22 | 2528 | 0.58 | 8570 |
| 湖北 | 1.05 | 67 | 92 | 297 | 115 | 43 | 32.89 | 2835 | 0.66 | 7262 |
| 江苏 | 0.95 | 64 | 94 | 287 | 102 | 39 | 31.54 | 3008 | 0.39 | 7786 |
| 广东 | 0.69 | 39 | 71 | 205 | 61 | 24 | 34.50 | 2988 | 0.37 | 11355 |
| 四川 | 0.56 | 40 | 57 | 177 | 61 | 23 | 32.62 | 3149 | 0.55 | 7693 |
| 山东 | 0.57 | 58 | 64 | 181 | 57 | 22 | 32.95 | 3202 | 0.28 | 6805 |
| 甘肃 | 0.71 | 42 | 62 | 190 | 66 | 26 | 28.13 | 2657 | 0.73 | 7282 |
| 湖南 | 0.74 | 42 | 61 | 194 | 61 | 24 | 33.06 | 2618 | 0.47 | 6477 |
| 浙江 | 0.86 | 42 | 71 | 204 | 66 | 26 | 29.94 | 2363 | 0.25 | 7704 |
| 新疆 | 1.29 | 47 | 73 | 265 | 114 | 46 | 25.93 | 2060 | 0.37 | 5719 |
| 福建 | 1.04 | 53 | 71 | 218 | 63 | 26 | 29.01 | 2099 | 0.29 | 7106 |
| 山西 | 0.85 | 53 | 65 | 218 | 76 | 30 | 25.63 | 2555 | 0.43 | 5580 |
| 河北 | 0.81 | 43 | 66 | 188 | 61 | 23 | 29.82 | 2313 | 0.31 | 5704 |
| 安徽 | 0.59 | 35 | 47 | 146 | 46 | 20 | 32.83 | 2488 | 0.33 | 5628 |
| 云南 | 0.66 | 36 | 40 | 130 | 44 | 19 | 28.55 | 1974 | 0.48 | 9106 |
| 江西 | 0.77 | 43 | 63 | 194 | 67 | 23 | 28.81 | 2515 | 0.34 | 4085 |
| 海南 | 0.70 | 33 | 51 | 165 | 47 | 18 | 27.34 | 2344 | 0.28 | 7928 |
| 内蒙古 | 0.84 | 43 | 48 | 171 | 65 | 29 | 27.65 | 2032 | 0.32 | 5581 |
| 西藏 | 1.69 | 26 | 45 | 137 | 75 | 33 | 12.10 | 810 | 1.00 | 14199 |
| 河南 | 0.55 | 32 | 46 | 130 | 44 | 17 | 28.41 | 2341 | 0.30 | 5714 |
| 广西 | 0.60 | 28 | 43 | 129 | 39 | 17 | 31.93 | 2146 | 0.24 | 5139 |
| 宁夏 | 1.39 | 48 | 62 | 208 | 77 | 34 | 22.70 | 1500 | 0.42 | 5377 |
| 贵州 | 0.64 | 23 | 32 | 93 | 37 | 16 | 28.12 | 1469 | 0.34 | 5415 |
| 青海 | 1.48 | 38 | 46 | 151 | 63 | 30 | 17.87 | 1024 | 0.38 | 7368 |

MATLAB源代码：

clc,clear

load('gj.txt'); %把原始数据保存在纯文本文件gj.txt中

sum\_0=0;%数值初始化

i=1;%数值初始化

gj=zscore(gj); %数据标准化

r=corrcoef(gj); %计算相关系数矩阵

%下面利用相关系数矩阵进行主成分分析，x的列为r的特征向量，即主成分的系数

[x,y,z]=pcacov(r); %y为r的特征值，z为各个主成分的贡献率

while ((sum\_0<98)&(z(i)>2))%根据累积贡献率确定主成分个数

n=size(z);

sum\_0=sum\_0+z(i);

i=i+1;

end

f=repmat(sign(sum(x)),size(x,1),1); %构造与x同维数的元素为±1的矩阵

x=x.\*f; %修改特征向量的正负号，每个特征向量乘以所有分量和的符号函数值

num=i-1;

df=gj\*x(:,1:num); %计算各个主成分的得分

tf=df\*z(1:num)/100; %计算综合得分

[stf,ind]=sort(tf,'descend'); %把得分按照从高到低的次序排列

stf=stf';

ind=ind';

fprintf('各成分的贡献率：\n');disp(z);

fprintf('综合得分：\n');disp(stf);

fprintf('排名：\n');disp(ind);

将上述数据保存在gj.txt文件中，运行上述MATLAB程序源代码，x为标准化后的数据，根据标准化后的数据的特征值y以及累积贡献率z来选择主成分的个数，该程序结果选择的主成分个数为4，根据所选主成分，得出了综合评价值stf，以及名次ind。具体结果见表2。

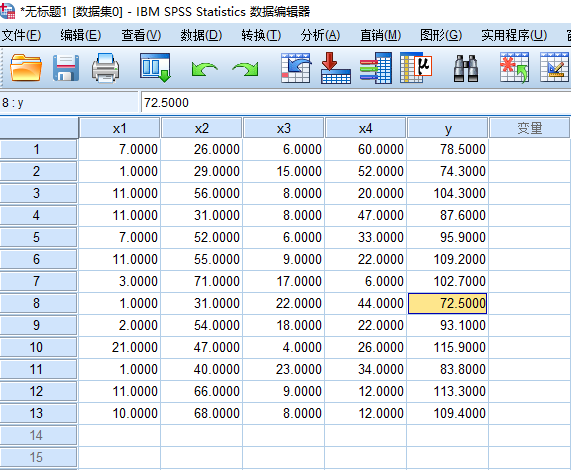
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 地区 | 北京 | 上海 | 天津 | 陕西 | 辽宁 | 吉林 | 黑龙江 | 湖北 |
| 名次 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 综合评价值 | 8.6043 | 4.4738 | 2.7881 | 0.8119 | 0.7621 | 0.5884 | 0.2971 | 0.2455 |
| 地区 | 江苏 | 广东 | 四川 | 山东 | 甘肃 | 湖南 | 浙江 | 新疆 |
| 名次 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| 综合评价值 | 0.0581 | 0.0058 | -0.268 | -0.3645 | -0.4879 | -0.5065 | -0.7016 | -0.7428 |
| 地区 | 福建 | 山西 | 河北 | 安徽 | 云南 | 江西 | 海南 | 内蒙古 |
| 名次 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
| 综合评价值 | -0.7697 | -0.7965 | -0.8895 | -0.8917 | -0.9557 | -0.9610 | -1.0147 | -1.1246 |
| 地区 | 西藏 | 河南 | 广西 | 宁夏 | 贵州 | 青海 |  |  |
| 名次 | 25 | 26 | 27 | 28 | 29 | 30 |  |  |
| 综合评价值 | -1.1470 | -1.2059 | -1.2250 | -1.2513 | -1.6514 | -1.6800 |  |  |

SPSS解法：参考下文中的多元分析中的“主成分分析”，由于用SPSS只能得到因子得分，对于排序问题，SPSS无能为力，故个人推荐使用MATLAB编程。

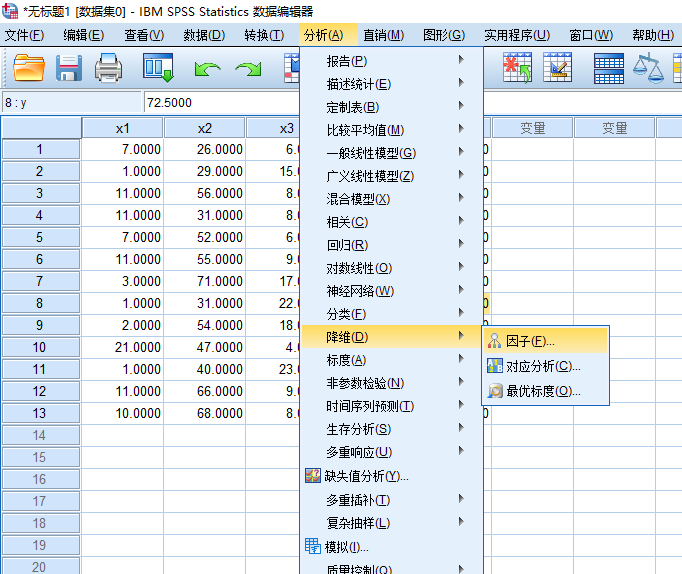
但是！！！因为使用主成分之前要做一个分析工作，那么分析工作如何使用比较方便，那就是spss，我们下面以SPSS分析为主，并不评价出结果：

SPSS解法：

1．在SPSS中输入相应数据：

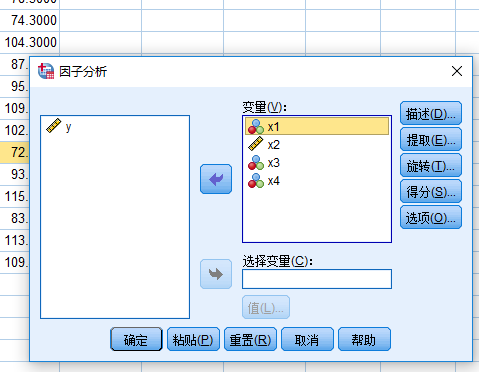


2.依次点击“分析”→“降维”→“因子”，如图：注意是因子

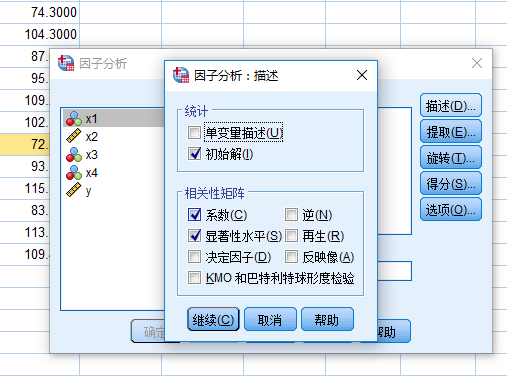


3.在“因子分析”对话框里，将x1、x2、x3、x4移到“变量”框里，如图：

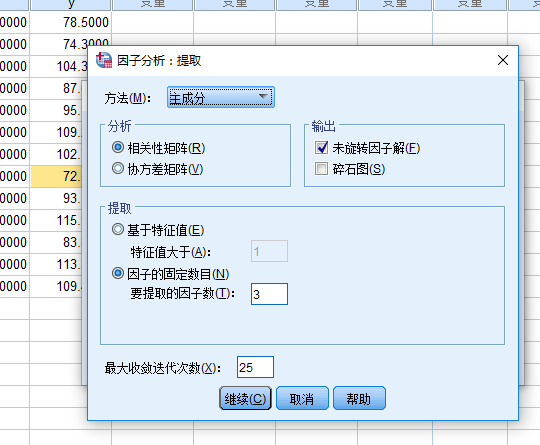




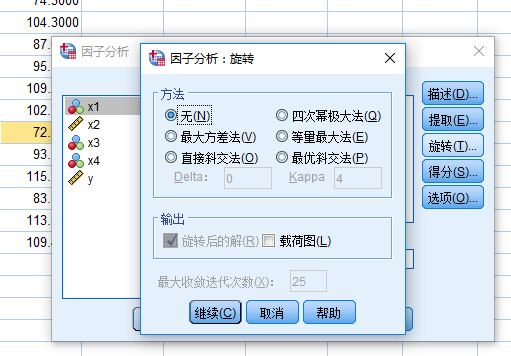
4.单击“描述”，在“描述”对话框中，“统计”中勾选“初始解”，“相关性矩阵”中勾选“系数”以及“显著性水平”，单击“继续”，如图：



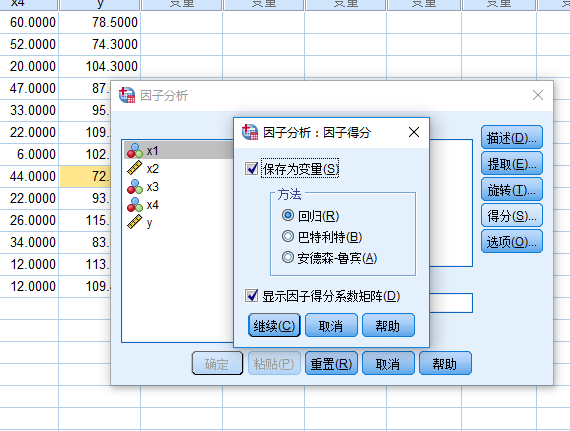
**5.单击“提取”，“方法”中选择“主成分”，**“分析”中选择“相关性矩阵”，“输出”中选择“未旋转因子解”，“提取”中选择“因袭的固定数目”，其“因子数”为3（此处为了和MATLAB解法的结果相对比，故选择了“因子的固定数目”，这一步不要忘了选择，针对这个题目是“3”针对其他题目一般不是3，会更大，故在其他数据中根据特征值大小以及累积贡献率，也可以选择“基于特征值”），如图，单击“继续”：



6.单击“旋转”，由于主成分分析不需要旋转，故此处默认即可，如图，单击“继续”：



7.单击“得分”，勾选“保存为变量”，“方法”中选择“回归”，勾选“显示因子得分系数矩阵”，如图，单击“继续”：



8.单击“选项”，“选项”对话框中默认即可，如图，单击“继续”：



结果分析：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **相关性矩阵** | | | | | |
|  | | x1 | x2 | x3 | x4 |
| 相关性 | x1 | 1.000 | .229 | -.824 | -.245 |
| x2 | .229 | 1.000 | -.139 | -.973 |
| x3 | -.824 | -.139 | 1.000 | .030 |
| x4 | -.245 | -.973 | .030 | 1.000 |
| 显著性 （单尾） | x1 |  | .226 | .000 | .209 |
| x2 | .226 |  | .325 | .000 |
| x3 | .000 | .325 |  | .462 |
| x4 | .209 | .000 | .462 |  |

上图中的“相关性”为相关性系数；

|  |  |  |
| --- | --- | --- |
| **公因子方差** | | |
|  | 初始 | 提取 |
| x1 | 1.000 | 1.000 |
| x2 | 1.000 | .999 |
| x3 | 1.000 | 1.000 |
| x4 | 1.000 | .999 |
| 提取方法：主成分分析法。 | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **总方差解释** | | | | | | |
| 成分 | 初始特征值 | | | 提取载荷平方和 | | |
| 总计 | 方差百分比 | 累积 % | 总计 | 方差百分比 | 累积 % |
| 1 | 2.236 | 55.893 | 55.893 | 2.236 | 55.893 | 55.893 |
| 2 | 1.576 | 39.402 | 95.294 | 1.576 | 39.402 | 95.294 |
| 3 | .187 | 4.665 | 99.959 | .187 | 4.665 | 99.959 |
| 4 | .002 | .041 | 100.000 |  |  |  |
| 提取方法：主成分分析法。 | | | | | | |

**上图中的“累积”为累积贡献率；【一般累积到80%就可以了，90%也行】**

|  |  |  |  |
| --- | --- | --- | --- |
| **成分矩阵a** | | | |
|  | 成分 | | |
| 1 | 2 | 3 |
| x1 | .712 | -.639 | .292 |
| x2 | .843 | .520 | -.136 |
| x3 | -.589 | .759 | .275 |
| x4 | -.819 | -.566 | -.084 |
| 提取方法：主成分分析法。 | | | |
| a. 提取了 3 个成分。就是刚才红色标记的“3” | | | |

这里的成分1、2、3为三个主成分对应的特征向量，这里的特征向量不能直接用，需要归一化后才能用，成分1为，例如：

成分1









故

其他同理可得即可，得出三个主成分表达式后，作主成分回归分析，得出回归方程，再化成标准化变量的回归方程，再恢复到原始自变量，得到主成分回归方程。

|  |  |  |  |
| --- | --- | --- | --- |
| **成分得分系数矩阵** | | | |
|  | 成分 | | |
| 1 | 2 | 3 |
| x1 | .318 | -.405 | 1.564 |
| x2 | .377 | .330 | -.728 |
| x3 | -.264 | .482 | 1.476 |
| x4 | -.366 | -.359 | -.452 |
| 提取方法：主成分分析法。  组件得分。 | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| **成分得分协方差矩阵** | | | |
| 成分 | 1 | 2 | 3 |
| 1 | 1.000 | .000 | .000 |
| 2 | .000 | 1.000 | .000 |
| 3 | .000 | .000 | 1.000 |
| 提取方法：主成分分析法。  组件得分。 | | | |