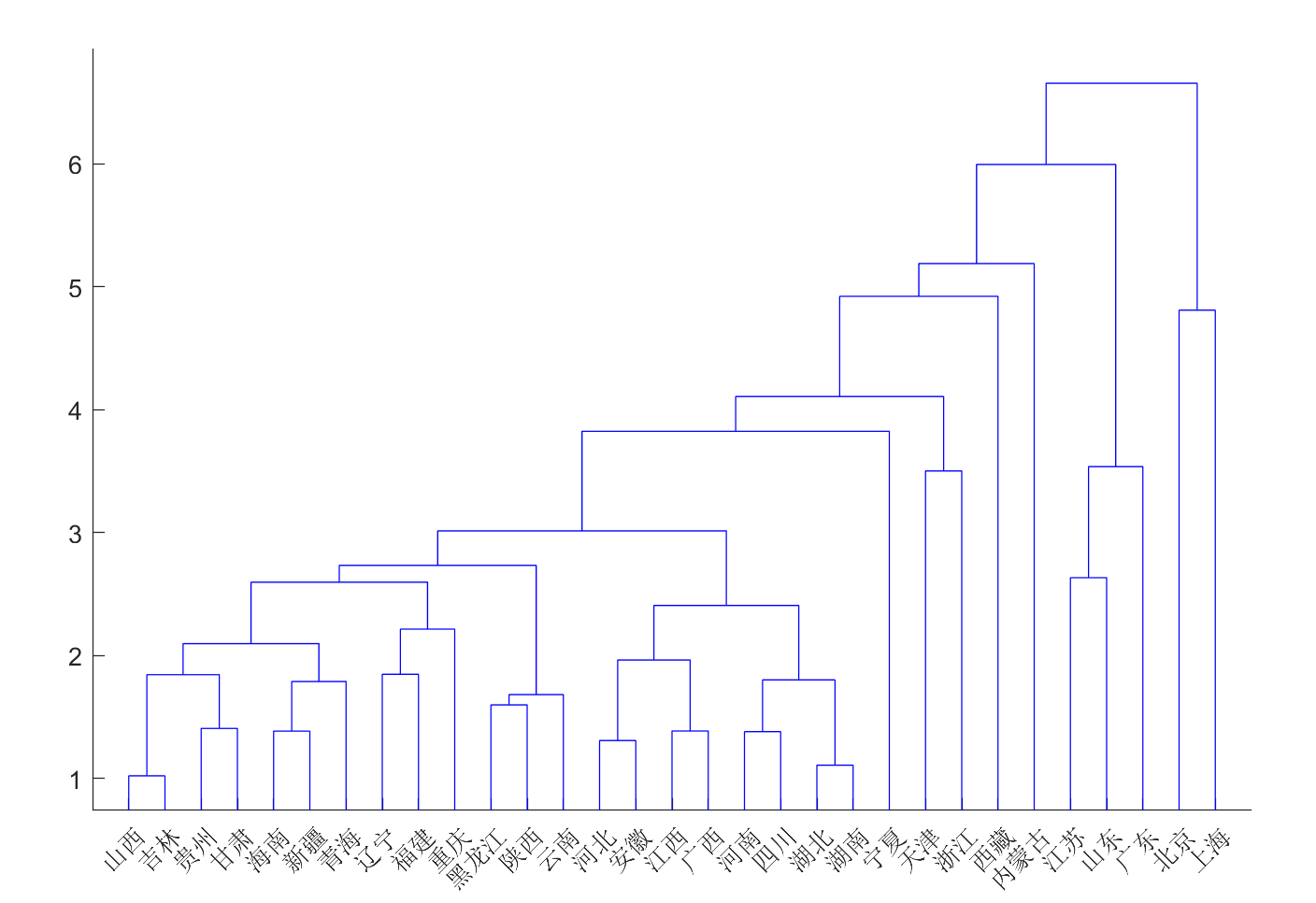
# 第四章2

##### 10193903446 汤博

1. 表格中给出了2017年中国大陆各省（直辖市、自治区）城镇化相关数据，进行聚类分析，写出具体计算步骤，最后要对结果进行解释。

* %clear;clc;  
  X = readtable('课后作业数据.xlsx','ReadRowNames',true,'ReadVariableNames',true,'VariableNamingRule','preserve');  
  % 标准差标准化数据矩阵  
  BX = zscore(table2array(X(:,2:end)),1);   
    
  % 欧氏距离  
  Y1 = pdist(BX,'euclidean');   
  %转换成距离矩阵  
  SY1 = squareform(Y1);  
  % 最短距离法定义构建具层次结构的聚类树  
  Z11 = linkage(Y1,'single');   
  % 最长距离法定义构建具层次结构的聚类树  
  Z12 = linkage(Y1,'complete');   
  % 质心距离法定义构建具层次结构的聚类树  
  Z13 = linkage(Y1,'centroid');  
  % 未加权平均距离法定义构建具层次结构的聚类树  
  Z14 = linkage(Y1,'average');  
  %检验一定算法下产生的二叉聚类树和实际情况的相符程度  
  C11 = cophenet(Z11,Y1);  
  C12 = cophenet(Z12,Y1);  
  C13 = cophenet(Z13,Y1);  
  C14 = cophenet(Z14,Y1);  
    
  % 曼哈顿距离  
  Y2 = pdist(BX, 'cityblock');  
  SY2 = squareform(Y2);  
  Z2 = linkage(Y2,'single');   
  C2 = cophenet(Z2,Y2);  
    
  %minkowski——明科夫斯基距离  
  Y3 = pdist(BX,'minkowski',2);  
  SY3 = squareform(Y3);  
  Z3 = linkage(Y3,'single');   
  C3 = cophenet(Z3,Y3);  
    
  %chebychev——切比雪夫距离  
  Y4 = pdist(BX, 'chebychev');  
  SY4 = squareform(Y4);  
  Z4 = linkage(Y4,'single');   
  C4 = cophenet(Z4,Y4);

| * **相符程度** | * **C2** | * **C3** | * **C4** |
| --- | --- | --- | --- |
|  | * 0.8264 | * 0.8386 | * 0.8269 |
| * **C11** | * **C12** | * **C13** | * **C14** |
| * 0.8386 | * 0.7883 | * 0.8808 | * 0.8843 |

* 选择欧氏距离 & 未加权平均距离法
* T = cluster(Z14,3); % 创建聚类  
  T = cluster(Z14,"criterion","distance","cutoff",3);  
  name = X.("地区");  
  [H,T,outperm] = dendrogram(Z14,31,'Labels',name'); % 画聚类图
* 

从聚类分析谱系图可以看出，在不同的聚类标准（距离）下，聚类结果不同：

当距离为1时，每个样本为单独的一类，即31个省份各自为一类；

当距离标准逐渐放大时，31个区域单元被依次聚类。样本之间距离最小的山西和吉林首先被分别聚为一类，随之，湖北和湖南, 河北和安徽也被聚为一类。

如果选取聚类标准（距离）为2，则31个区域单元被聚为17类。

如果选取聚类标准（距离）为2，则31个区域单元被聚为17类。

如果选取聚类标准(距离)为3，则31个区域单元被聚为10类。

如果进一步把聚类标准(距离)扩大到4，则31个区域单元被聚为7类；

如果进一步把聚类标准(距离)扩大到5，则31个区域单元被聚为4类；

如果进一步把聚类标准(距离)扩大到6，则31个区域单元被聚为2类，北京上海两个超大城市为一类，其他省份一类；

当聚类标准（距离）扩到7时，31个区域单元被聚为一类。

1. 根据以上数据，进行主成分分析，并对结果进行解释。

clear;clc;  
  
X = readtable('主成分数据.xlsx','ReadVariableNames',true,'VariableNamingRule','preserve');  
mX = table2array(X);

* 使用matlab自带模型

% use built-in models   
% 直接使用内置函数pca  
 % coeff 主成分系数,行包含9个原料变量的系数，其列对应于9个主成分。  
 % score 主成分分数  
 % explained 解释方差占总方差的百分比  
[coeff, score, ~, ~, explained, ~] = pca(mX);  
% 计算主成分累积贡献率  
latent = cumsum(explained);  
% 取累计贡献率达85%的特征值所对应的主成分  
i = find(latent >= 85, 1, 'first');  
% 行包含9个原料变量的系数，其列对应于累计贡献率达85%的特征值所对应的主成分。  
MainCoeff = coeff(1:i, :);

* 根据课本上的代码进行拟合

% formulas according to textbook  
% 自行编写函数  
% % 极差标准化  
% BX = (mX - min(mX)) ./ (max(mX) - min(mX));  
% 标准差标准化数据矩阵  
BX = zscore(mX,1);   
  
% 计算相关系数矩阵  
R = corrcoef(BX);   
% 计算特征值(对角矩阵 D)和特征向量(矩阵 V)  
[V, D] = eig(R);   
% 从对角矩阵D中提取特征值  
d = diag(D);   
% 对特征根进行排序  
[d, ind] = sort(d,'descend');  
% 使用 ind 对 D 的对角线元素进行重新排序。  
Ds = D(ind, ind);  
% 使用相同的索引 ind 对 V 的列进行重新排序。  
Vs = V(:, ind);  
  
% 计算主成分贡献率  
latent = d / sum(d);   
% 计算主成分累计贡献率  
Latent = cumsum(latent);   
% 取累计贡献率达 85% 的特征值所对应的主成分  
i = find(Latent>= 0.85,1,'first');  
% 主成分载荷  
B = sqrt(d(1:i)') .\* Vs(:,1:i);   
% 主成分得分  
C = mX \* B;

相关系数矩阵

|  | **x1** | **x2** | **x3** | **x4** | **x5** | **x6** | **x7** | **x8** | **x9** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **x1** | 1 | -0.3274 | -0.7144 | -0.3356 | 0.3085 | 0.4075 | 0.7902 | 0.1559 | 0.7445 |
| **x2** | -0.3274 | 1 | -0.0349 | 0.6441 | 0.4197 | 0.2549 | 0.0086 | -0.0778 | 0.0943 |
| **x3** | -0.7144 | -0.0349 | 1 | 0.0700 | -0.7397 | -0.7547 | -0.9303 | -0.1095 | -0.9238 |
| **x4** | -0.3356 | 0.6441 | 0.0700 | 1 | 0.3825 | 0.0690 | -0.0460 | -0.0310 | 0.0726 |
| **x5** | 0.3085 | 0.4197 | -0.7397 | 0.3825 | 1 | 0.7342 | 0.6721 | 0.0981 | 0.7465 |
| **x6** | 0.4075 | 0.2549 | -0.7547 | 0.0690 | 0.7342 | 1 | 0.6576 | 0.2217 | 0.7067 |
| **x7** | 0.7902 | 0.0086 | -0.9303 | -0.0460 | 0.6721 | 0.6576 | 1 | -0.0296 | 0.8904 |
| **x8** | 0.1559 | -0.0778 | -0.1095 | -0.0310 | 0.0981 | 0.2217 | -0.0296 | 1 | 0.2895 |
| **x9** | 0.7445 | 0.0943 | -0.9238 | 0.0726 | 0.7465 | 0.7067 | 0.8904 | 0.2895 | 1 |

特征值及主成分贡献率：

| **主成分** | **特征值** | **贡献率** | **累积贡献率** |
| --- | --- | --- | --- |
| **z1** | 4.6611 | 0.5179 | 0.5179 |
| **z2** | 2.0895 | 0.2322 | 0.7501 |
| **z3** | 1.0430 | 0.1159 | 0.8660 |
| **z4** | 0.5074 | 0.0564 | 0.9223 |
| **z5** | 0.3152 | 0.0350 | 0.9574 |
| **z6** | 0.1926 | 0.0214 | 0.9788 |
| **z7** | 0.1144 | 0.0127 | 0.9915 |
| **z8** | 0.0453 | 0.0050 | 0.9965 |
| **z9** | 0.0315 | 0.0035 | 1 |

主成分载荷

|  | **z1** | **z2** | **z3** |
| --- | --- | --- | --- |
| **x1** | 0.7386 | 0.5318 | -0.0611 |
| **x2** | 0.1235 | -0.8868 | -0.0282 |
| **x3** | -0.9637 | -0.0956 | 0.0949 |
| **x4** | 0.0416 | -0.8682 | 0.0370 |
| **x5** | 0.8128 | -0.4435 | -0.0109 |
| **x6** | 0.8189 | -0.1793 | 0.1248 |
| **x7** | 0.9332 | 0.1329 | -0.2514 |
| **x8** | 0.1973 | 0.1004 | 0.9700 |
| **x9** | 0.9638 | 0.0250 | 0.0917 |

从上表中可以看出：

第一主成分z1与x1, x5, x6, x7, x9呈现出较强的正相关，与x3呈现出较强的负相关，x1代表人口密度，x5代表人均粮食产量，x6代表经济作物占农作物播面比例，x7代表耕地占土地面积比，x9代表灌溉田占耕地面积之比，x3代表森林覆盖率，因此可以认为第一主成分主要体现了地区农业结构的代表，并且体现了人均资源的影响因素。

第二主成分z2与与x2, x4呈现出较强的负相关，x2代表人均耕地面积，x4代表农民人均纯收入，因此可以认为第二主成分主要体现农业经济结构特征。

第三主成分z3与x8呈现出较强的正相关，其次是和x6和x9，x7代表果园与林地面积之比，其中x2, x5分别代表经济作物占农作物播面比例和灌溉田占耕地面积之比，因此可以认为第三主成分主要体现第一产业（农林牧渔等）内部结构。

用3个成分代替原来的9个变量，可以使得问题更加简化明了。