值得分享的法国专业技能
统计学与数据多维分析

François Husson
https://husson.github.io
雷恩农业高等院校统计与计算机系
继续教育培训 - 中国代表团 – 2026年8月25日至26日

目录

引言

简介

François Husson
统计学教授
任职于雷恩农业
高等院校

- **研究方向**：**数据分析、多重表格分析、缺失数据**处理 发表论文60**余篇**
- **教学**：农业与食品工程师课程、数据科学硕士课程
- **数据分析慕**课（MOOC）课程（法语和英语）
- 统计学丛书主编
- YouTube频道：https://www.youtube.com/@HussonFrancois
- **撰写**专著、开发R语言软件包

Packages: FactoMineR - missMDA - SensoMineR - Factoshiny - FactoInvestigate - RcmdrPlugin.FactoMineR

引言

雷恩农业高等院校简介

RÉPUBLIQUE FRANÇAISE
L'INSTITUT agro Rennes Angers

- 创建于1830**年**
- 农业与食品工程精英院校
- **一所在**办学领域（食品、农业、环境与景观）方面独具特色的高**等教育与科研机构**
- 农业生态与食品转型的重要参与者
- 6000**名学生**，360**名博士生**，400**名教学科研人**员
  - **生物多**样性与生态学：植物健康的生态农业管理
  - **生**态系统健康
  - **可持**续食品与理想膳食
  - **植物与养殖**业对气候变化的适应
  - **面向**农业与环境的人工智能与数据科学
  - **景**观融入城市 / 可食地景
  - 农业政策与农业生态议题
  - **植物与种子改良**
  - 动物福利
  - **公共**资源治理（生物资源、水资源与土壤管理）

雷恩农业高等院校简介

国际化战略

1. 加强校园、学生与教职工的国际化

Institut Agro 介绍：与中国的合作

教育合作

- 与苏州大学及山东农业大学开展2项机构间合作

科研合作

主要与以下合作伙伴开展科研项目及发展活动方面的合作:

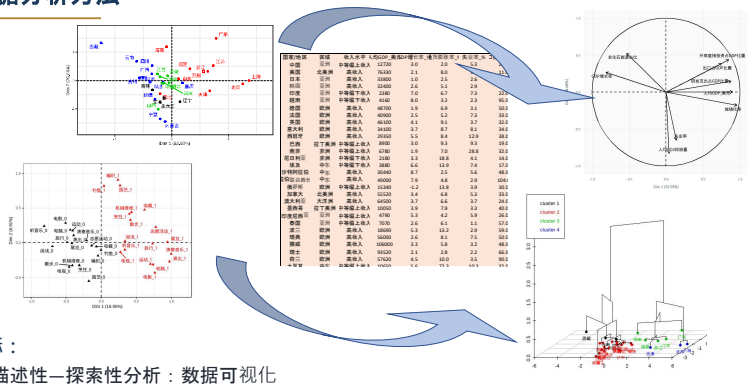
- 中国农业大学（农业生态学与绿色农业）
- 青岛农业大学
- 苏州大学（Foodprint 项目）
- 浙江大学
- 四川农业大学
- 云南大学（Platomix 项目）

11 个欧洲长期学术合作伙伴：

Wageningen University and Research – WUR
University of Hohenheim – UHOH
Czech University of Sciences Prague – CZU
L'Institut Agro, France
Estonian University of Life Sciences – EMU
Universita degli Studi della Tuscia – UNITUS
Swedish University of Agricultural Sciences – SLU
Univ. of Natural Resources and Life Sciences – BOKU
Warsaw University of Life Sciences – WULS
Norwegian University of Life Sciences – NMBU
Universidad de Lleida – UdL

1 个非欧洲长期学术合作伙伴：林肯大学——新西兰 - LU

数据分析方法



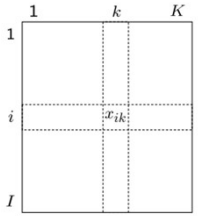
目标：

- 描述性—探索性分析：数据可视化
- 综合归纳：大型"个体×变量"表格的总结

目录

主成分分析（PCA）

主成分分析（PCA）研究的是矩形数据表，其中行代表个体，列代表数量型变量



- 经济学：地区i中指标k的数值
- 心理学：个体i对陈述k的认同程度
- 社会学：社会职业类别i中个体从事活动k所花费的时间
- 调查：商户i按1到5分的量表对以下问题k进行评分：
  - 消费者越来越多地在网上购物。
  - 租金过高。
  - 利润率变得不足。
  - 人工成本上涨过多。
  - 大型连锁店吸引了顾客。
  - 配送平台改变了市场。
  - 年轻人的消费方式有所不同。
  - 当前的经济形势减少了家庭支出。

⇒ 类似这样的数据表还有很多

经济数据

- 30个个体（行）：中国各省份
- 10个变量（列）：
  - 9个经济指标
  - 1个地理变量（地区）

变量	定义
PIB_hab_USD	人均国内生产总值，以美元（USD）计，衡量一个国家按人口平均分摊的财富产出——是衡量平均生活水平的经典指标
Croissance_PIB	GDP增长率，通常以年度变化百分比表示（常以剔除通胀影响后的实际值计算），衡量一个国家逐年的经济活力。
Inflation	通货膨胀率，以年百分比表示——反映物价总体水平的变化（通常以消费者价格指数CPI衡量）。
Chomage	失业率，占劳动人口的百分比——指没有工作且正在积极求职的劳动力比例。
Exportations_PIB	商品和服务出口占GDP的比重，以百分比表示——反映对外贸易开放程度及对外部市场依赖程度的指标。
IDE_PIB	外商直接投资占GDP的比重，以百分比表示——衡量流入该国的外资资本流量（或存量，视数据来源而定），反映其经济吸引力。
RD_PIB	研发（R&D）支出占GDP的比重，以百分比表示——反映一个国家在创新和科研投入方面的努力程度。
Energies_renouvelables	可再生能源占能源消费（或生产）总量的比重，以百分比表示——反映一个国家能源转型的指标。

经济数据

- 30个个体（行）：中国各省份
- 10个变量（列）：
  - 9个经济指标
  - 1个地理变量

省份	人均GDP_美元	GDP增长率	城镇化率	失业率	出口占GDP比重	投资占GDP	支出占GDP	化石能源占比	人均CO2排放	地区	行政类型
北京	29500	0.7	87.5	4.4	15.0	3.5	6.2	22.0	7.0	东部	直辖市
上海	26500	-0.2	89.3	4.5	38.0	5.0	4.2	20.0	8.5	东部	直辖市
天津	17000	1.0	85.0	4.0	22.0	2.0	3.4	16.0	9.8	东部	直辖市
重庆	13000	2.6	71.7	3.9	18.0	1.5	2.1	18.0	6.5	西部	直辖市
广东	14600	1.9	75.4	2.5	58.0	4.2	3.1	19.0	5.2	东部	省
江苏	21200	2.3	75.0	3.3	42.0	3.0	3.2	17.0	9.0	东部	省
浙江	18000	3.1	73.4	4.1	46.0	2.8	3.0	18.0	6.8	东部	省
山东	13200	3.9	64.5	3.6	22.0	1.5	2.7	16.0	9.5	东部	省
…	…	…	…	…	…	…	…	…	…	…	…
新疆	9700	3.6	58.4	3.9	10.0	0.5	0.8	16.0	13.5	西部	自治区
内蒙古	14800	3.9	68.8	4.1	7.0	0.5	0.8	18.0	24.0	西部	自治区
宁夏	11000	3.7	65.4	4.0	6.0	0.4	1.0	15.0	22.5	西部	自治区
西藏	7800	9.5	36.4	3.6	3.0	0.2	0.4	92.0	2.0	西部	自治区

问题与目标

数据表 = 行的集合，或列的集合

个体研究

- 根据全部变量构建彼此相似的个体分组
- 相似性总结，即个体的一种划分

变量研究

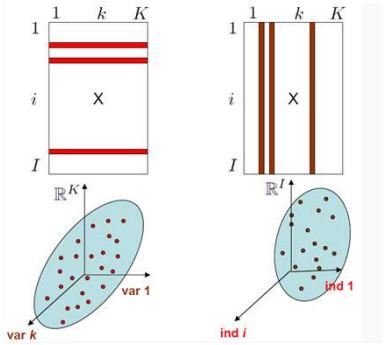
- 寻找变量之间的相似性与（线性）关联
- 关联关系总结：相关矩阵的可视化
- 寻找能够概括各变量的综合指标

两项研究之间的联系

- 通过变量刻画个体分类的特征

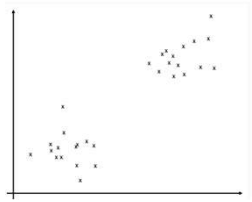
主成分分析的目标：描述性—探索性分析（可视化），以及对大型"个体×变量"表格进行归纳总结

两个点云



用两个点云来观察同一张数据表

个体点云



- 个体存在于一个K维空间中
- **研究个体点云的形状**：**将**变量中心化不会改变点云的形状⇒**因此始**终进行中心化；而将变量标准化**会改**变点云形状⇒**当各**变量单位不同时，系统性地进行标准化
- **关键问题：个体之间应采用何种距离？是否中心化—标准化？是否按人口数进行相对化处理？是否取**对数？

个体点云

1个个体 = 数据表中的1行 ⇒ K维空间中的1个点

- 若K = 1：轴向表示
- 若K = 2：点云图
- 若K = 3：三维表示较为困难
- 若K = 4：无法表示，但概念依然简单

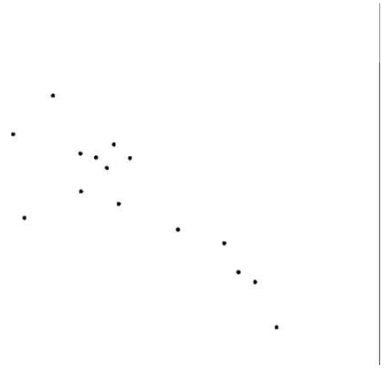
相似性的概念：个体i与i'之间的（平方）距离：

$$d^2(i,j) = \sum_{\{k=1\}}^{K} (x_{\{ik\}} - (x_{\{jk\}})^2$$ *（感谢毕达哥拉斯）*

研究个体 ⇔ 研究点云的形状

个体点云的拟合

***主成分分析旨在**为个体点云提供一个尽可能忠实的简化图像⇔**找到能最好概括数据的子空**间*

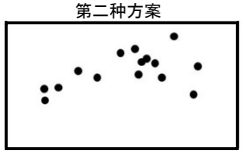


个体点云的拟合（续）

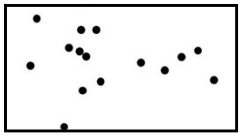
主成分分析旨在为个体点云提供一个尽可能忠实的简化图像⇔找到能最好概括数据的子空间

第一种方案

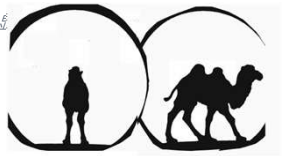
第二种方案



第三种方案

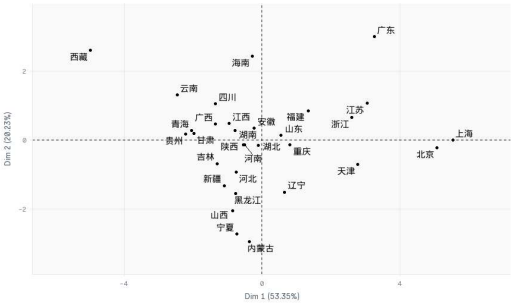


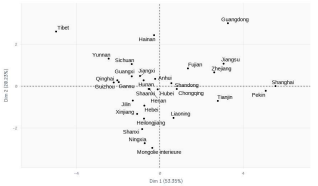
- 通过投影获得最佳近似
- 更好地呈现多样性与变异性
- *"这是什么动物？"——引自J.P. Fenelon（根据观察角度不同，插图显*



示例：个体图

在我们的示例中，哪个平面能使各城市的离散程度最大化？

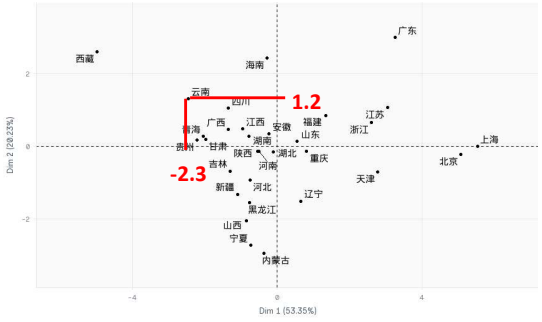


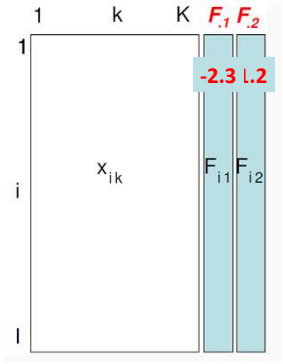


如何解释这些坐标轴？是什么因素使上海与西藏形成对立？海南与内蒙古之间又是什么？
⇒需要借助变量来解释这些变异维度

借助变量解读个体图

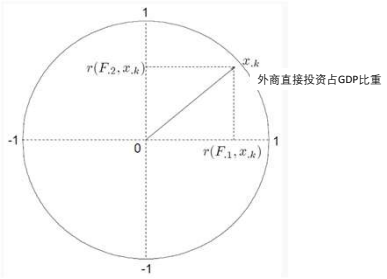
将个体在坐标轴上的坐标视为变量





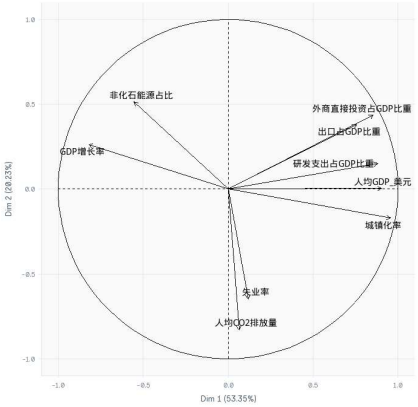
借助变量解读个体图

- 变量x.k与F.1（及F.2）之间的相关性



⇒相关圆

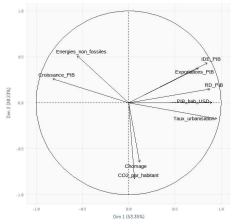
借助变量解读个体图



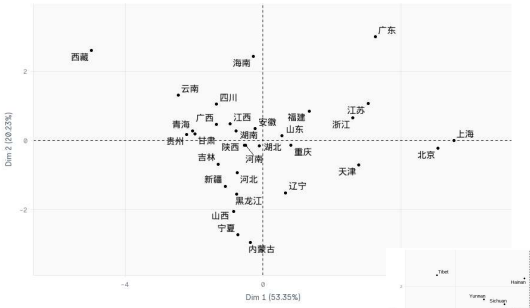
如何解释第一轴？第二轴呢？

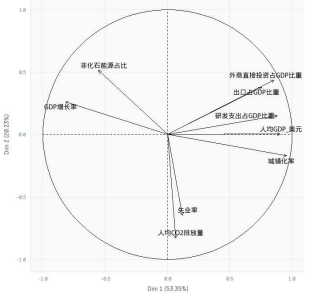
主要变异因素：

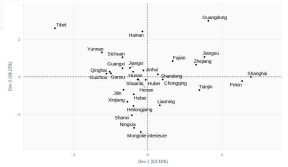
- PIB_hab_USD、RD_PIB、Taux_urbanization（城镇化率）较高的城市与之相对（位于图的右侧），左侧则是这些变量取值较低、而croissance_PIB（GDP增长率）取值较高的城市
- **在GDP水平相近的情况下**，图下方为人均CO2**排放量和失**业率较高的**城市**，图上方则为非化石能源占比较高的城市

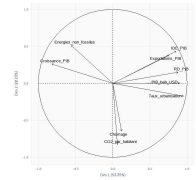


借助变量解读个体图



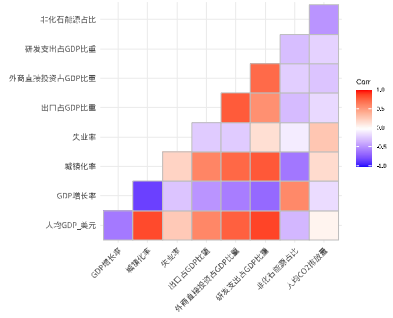






变量间关联关系的可视化

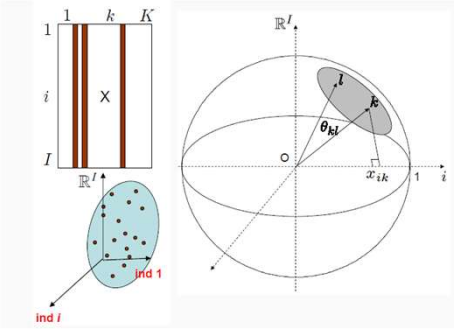
如何更好地可视化变量之间的关联（相关性）？如何可视化相关矩阵？



当变量很多时非常繁琐！我们将会做得更好！！

变量点云 N^K

1个变量 = I个数值 = I维空间中的1个点



<x.k , x.l> = ||x.k|| ||x.l|| cos(θkl)

cos(θkl) = <x.k , x.l> / (||x.k|| ||x.l||)

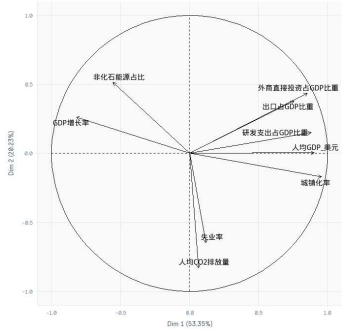
由于变量已中心化：cos(θkl) = r(x.k , x.l)

相关系数的几何解释！

由于变量已标准化 ⇒ ||x.k|| = ||x.l|| = 1 ⍰ 点位于半径为1的球面上

变量点云的拟合

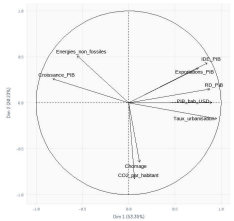
我们力求尽可能好地表示变量点云（与个体的做法相同）。我们寻找满足以下条件的正交轴：



- **第一**轴与所有变量的关联最强：max(v1) Σk r(v1, x.k)² ⇒ v1即为最能概括各变量的综合变量
- **找到与第一**轴正交、且与变量关联最强的第二轴，随后是第三轴，依此类推。

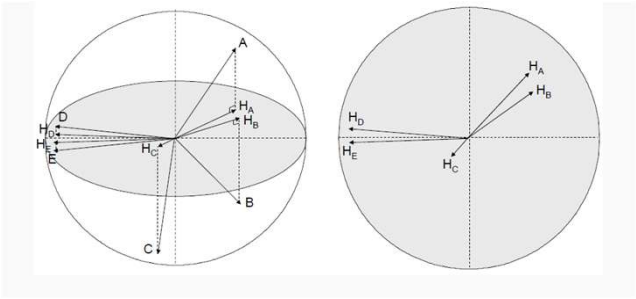
⇒ 与之前相同的表示方式！！！！

- **有助于解**读个体
- 变量点云的最优表示
- **相关矩**阵的可视化



投影……

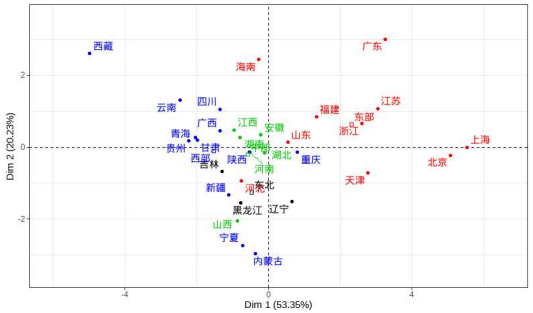
*在1维空间中，我们有：$corrélation(A,B) = \cos(\theta_{\{AB\}})$。**但我**们是在一个平面上进行可视化，因此会出现变形。*
*若变量投影良好：$corrélation(A,B) = \cos(\theta_{\{AB\}}) = \cos(\theta_{\{H_A H_B\}})$，**因此** $\cos(\theta_{\{H_A H_B\}}) = corrélation(A,B)$*

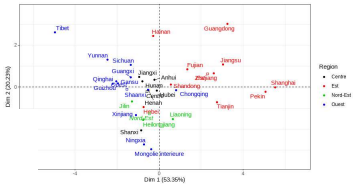


***只有投影良好的**变量才能被解释！*

补充信息

- 对于数量变量：变量的投影
- 对于分类变量的类别：投影至取该类别的个体的重心处





⇒ 补充信息不参与坐标轴的构建

维度的描述

通过数量变量：

- 计算每个变量与维度s之间的相关性
- 对（显著的）相关系数进行排序

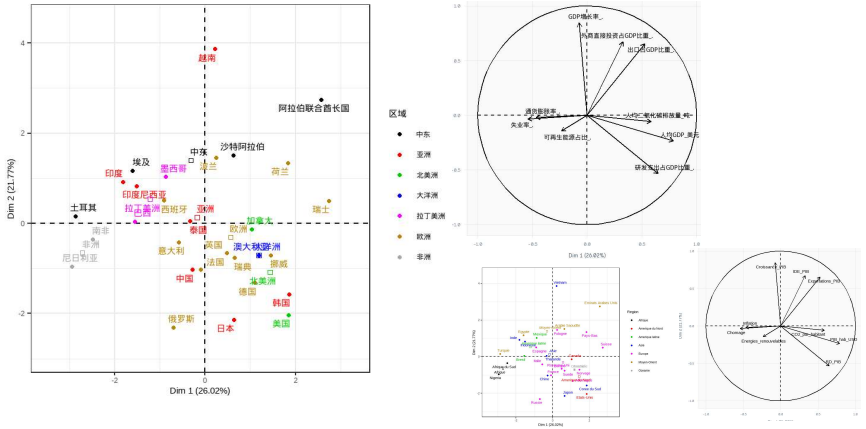
```
$Dim.1
Link between the variable and the continuous variables (R-square)
================================================================
                          correlation      p.value  n
城镇化率                    0.9549989 7.568554e-17 31
人均GDP_美元                0.9040524 3.180991e-12 31
研发支出占GDP比重           0.8809072 6.262908e-11 31
外商直接投资占GDP比重       0.8517836 1.229040e-09 31
出口占GDP比重               0.7567551 8.395875e-07 31
非化石能源占比             -0.5538826 1.226458e-03 31
GDP增长率                  -0.8201261 1.640739e-08 31

Link between variable and the categories of the categorical variables
======================================================================
        Estimate      p.value
地区=东部  2.376807 3.348483e-06
地区=西部 -1.461448 1.256628e-03

$Dim.2
Link between the variable and the continuous variables (R-square)
================================================================
                          correlation      p.value  n
非化石能源占比              0.5132278 3.150705e-03 31
外商直接投资占GDP比重       0.4350302 1.445286e-02 31
出口占GDP比重               0.3789087 3.554613e-02 31
失业率                     -0.6461464 8.621452e-05 31
人均CO2排放量              -0.8288282 8.484566e-09 31
```

```
$Dim.1
Link between the variable and the continuous
==============================================



                          correlation      p.value  n
Taux_urbanisation           0.9549989 7.568554e-17 31
PIB_hab_USD                 0.9040524 3.180991e-12 31
RD_PIB                      0.8809072 6.262908e-11 31
IDE_PIB                     0.8517836 1.229040e-09 31
Exportations_PIB            0.7567551 8.395875e-07 31
Energies_non_fossiles      -0.5538826 1.226458e-03 31
Croissance_PIB             -0.8201261 1.640739e-08 31

=================================================
              Estimate      p.value
Region=Est    2.376807 3.348483e-06
Region=Ouest -1.461448 1.256628e-03

$Dim.2
Link between the variable and the continuous
=================================================
                          correlation      p.value  n
Energies_non_fossiles       0.5132278 3.150705e-03 31
IDE_PIB                     0.4350302 1.445286e-02 31
Exportations_PIB            0.3789087 3.554613e-02 31
Chomage                    -0.6461464 8.621452e-05 31
CO2_par_habitant           -0.8288282 8.484566e-09 31
```

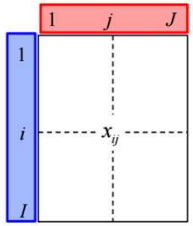
主成分分析的实践

1. 选择有效变量
2. **决定是否**对变量进行标准化
3. 执行主成分分析
4. **确定需要解**释的维度数量
5. **同**时解读个体图与变量图
6. **利用各**类指标丰富解读
7. **回**归原始数据进行解读

请你来解读：29个国家，相同的经济指标



对应分析（AFC）



xij：属于集合 I 的元素 i 且属于集合 J 的元素 j 的个体数

示例：

- 社会学：属于社会职业类别 i 且属于年龄段 j 的个体数
- 经济学：城市 i 中从事行业 j 的企业数
- 生态学：栖息地 j 中物种 i 的动物数量
- 医学：地区 i 在日期 j 的新冠死亡人数
- 文本分析：作者 i 使用词语 j 的次数

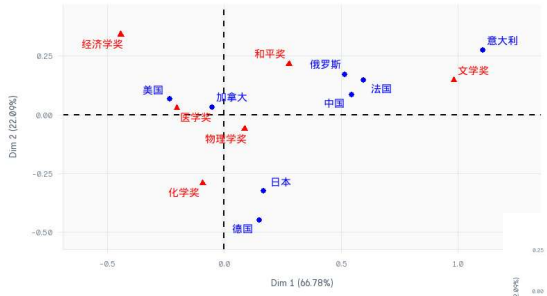
⇒ 可应用 χ^2 独立性检验的示例

诺贝尔奖数据

	化学奖	经济学奖	文学奖	医学奖	和平奖	物理学奖
法国	11	5	16	11	9	18
德国	31	1	8	17	4	26
美国	83	71	13	110	28	99
意大利	1	0	6	3	2	4
日本	7	0	2	6	2	11
俄罗斯	2	2	4	2	4	12
加拿大	4	3	1	4	2	7
中国	1	0	1	1	2	3

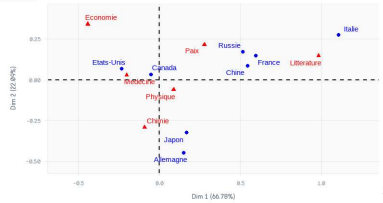
国家与奖项类别之间是否存在联系？某些国家是否具有特点？
⚠ 我们关注的是相对数据（不希望区分大国和小国）

诺贝尔奖示例



- 科学类奖项与其他奖项对立；在较小程度上，物理/化学与经济学也存在对立
- 各国的位置反映了它们在获得诺贝尔奖方面的特点

对应分析提供了一个综合的可视化图示，有助于理解表格（对于大型表格尤其如此）

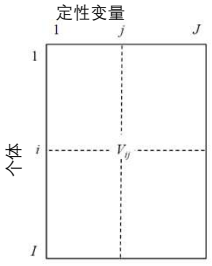


提纲

1. 引言
2. 主成分分析
3. 对应分析用于文本数据分析
4. **多重对应分析用于问卷调查分析**
5. 分类方法

多重对应分析（ACM）

用于问卷调查分析（个体 × 定性变量表）。

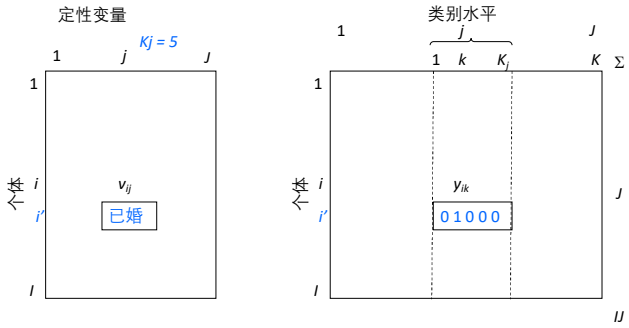


Vij 变量 j 的类别水平
由个体 i 所具有

示例：中小企业经营状况调查（为什么有些企业会倒闭？）

- 您所属的行业是？（食品 / 餐饮 / 服装 / 服务业 / 其他）
- 您的店铺经营了多长时间？（不到6个月 / 6个月至2年 / 2至5年 / 6至10年 / 超过10年）
- 您店铺目前的盈利状况是……（很差 / 差 / 一般 / 好 / 很好）
- 过去两年营业额的变化是……（大幅下降 / 下降 / 稳定 / 上升 / 大幅上升）
- 您店铺的客流量是……（大幅下降 / 下降 / 稳定 / 上升 / 大幅上升）
- 您的主要竞争对手是……（电商 / 大型商场 / 连锁店 / 其他小商户 / 无）
- 电子商务对您业务的影响是……（无 / 弱 / 中等 / 强 / 很强）
- 商铺租金水平是……（很低 / 低 / 可接受 / 高 / 很高）
- 经营成本是……（很低 / 低 / 中等 / 高 / 很高）
- 如今消费者的支出……（明显减少 / 略微减少 / 与以前相同 / 略微增加 / 明显增加）
- 您认为两年后您的店铺是否还会营业？（肯定不会 / 可能不会 / 不确定 / 可能会 / 肯定会）
- 如果您的店铺要关闭，主要原因会是什么？（客源不足 / 盈利能力低 / 租金 / 竞争 / 健康 / 退休 / 其他）
- 您主要采用什么策略来应对困难？（线上销售 / 促销 / 多元化经营 / 削减成本 / 无）

多重对应分析（ACM）



完全析取表（TDC）

多重对应分析（ACM）

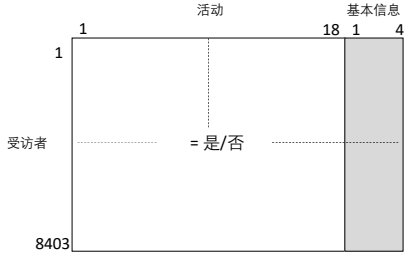
ACM 的目标：

- 个体研究
  - 一个个体 = 完全析取表中的一行 = 其所有类别水平的集合
  - 个体间的相似性；研究个体的变异性
  - 个体变异性的主要维度（与类别水平相关）
- 变量研究
  - 定性变量之间的关联（与类别水平相关）
  - 各类别水平之间关联的整体可视化
  - 综合变量（基于定性变量构建的定量指标）

➔ 与主成分分析（ACP）思路相近的问题

多重对应分析（ACM）

- 摘自法国国家统计局（Insee）2003年关于身份构建的调查，名为《生活史》
- 8403 个个体
- 两类变量：
  - 在以下休闲活动中，请指出您经常从事的活动：阅读、听音乐、看电影、看演出、参观展览、使用电脑、运动、散步、旅行、演奏音乐、收藏、志愿活动、手工修理、园艺、编织、烹饪、钓鱼，以及平均每天看电视的时长
  - 基本信息（4个问题）：性别、年龄、职业、婚姻状况



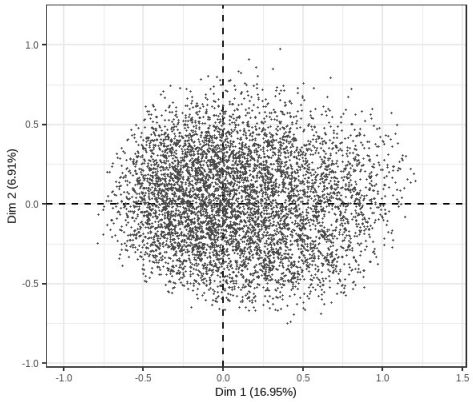
- 1 个个体 = 1 个活动特征
- 活动特征变异性的主要维度
- 这些维度与基本信息之间的关联

ACM 的原理

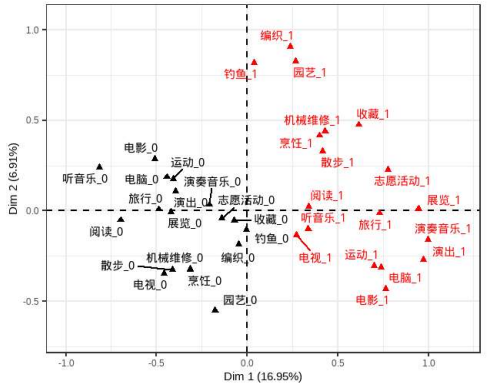
- 如果两个个体具有相同的类别水平，则它们重合
- 两个个体共有许多类别水平：距离小
- 两个个体中有一个具有罕见类别水平：为体现其特殊性，距离较大
- 两个个体共有一个罕见类别水平：为体现其共同特殊性，距离较小
- 某一类别水平位于选择该类别的所有个体的伪重心处

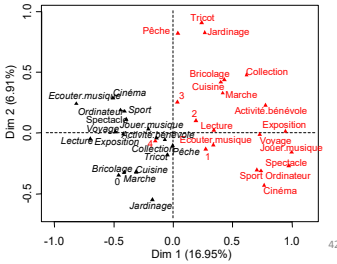
- 对个体点云的拟合
- 与所有因子分析方法一样，寻找因子维度

个体图……并不常用

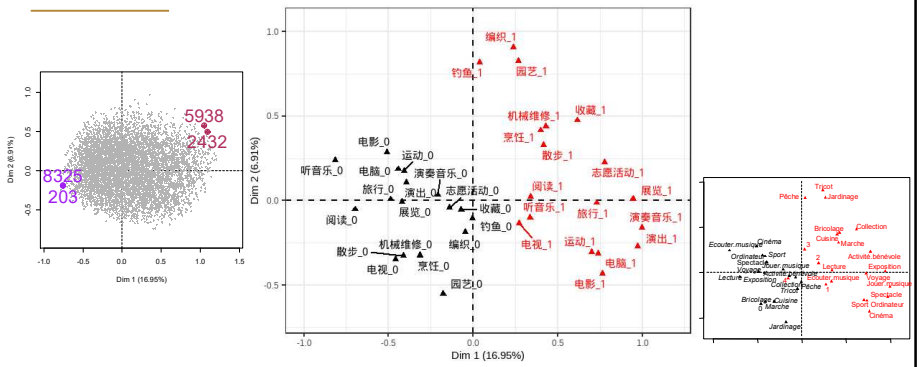


类别水平的表示图



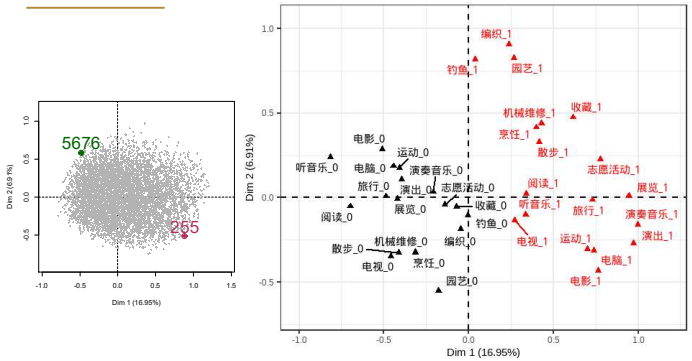


类别水平的表示图



	阅读	听音乐	电影	演出	展览	电脑	运动	散步	旅行	演奏音乐	收藏	志愿活动	机械维修	园艺	编织	烹饪	钓鱼	电视
5938	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	3
2432	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	2
8325	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
203	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4

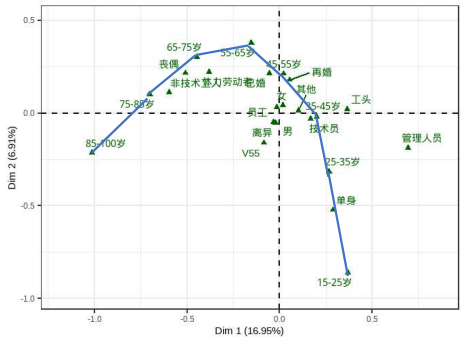
类别水平的表示图

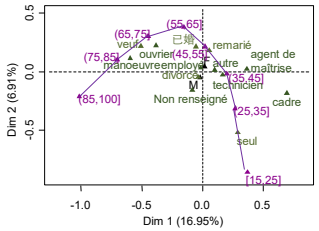


	阅读	听音乐	电影	演出	展览	电脑	运动	散步	旅行	演奏音乐	收藏	志愿活动	机械维修	园艺	编织	烹饪	钓鱼	电视
255	1	1	1	1	1	1	1	1	1	1	0	1	0	0	0	0	0	1
5676	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	4

补充类别水平的表示图

补充变量有助于理解变量与个体群体之间的联系





维度的描述

通过定量变量：

- 计算每个变量与维度 s 之间的相关性
- 对（显著的）相关系数进行排序

```
$Dim.1

Link between variable and the categories of the categorical variables
==========================================================================
                                Estimate       p.value
阅读=阅读_1                  0.23125560  0.000000e+00
听音乐=听音乐_1              0.25657041  0.000000e+00
电影=电影_1                  0.28307598  0.000000e+00
演出=演出_1                  0.30379833  0.000000e+00
展览=展览_1                  0.30388560  0.000000e+00
电脑=电脑_1                  0.26265070  0.000000e+00
运动=运动_1                  0.24685643  0.000000e+00
散步=散步_1                  0.18447699  0.000000e+00
旅行=旅行_1                  0.27033560  0.000000e+00
演奏音乐=演奏音乐_1          0.26839870  0.000000e+00
机械维修=机械维修_1          0.16540452 8.816716e-267
烹饪=烹饪_1                  0.15865273 9.423346e-247
...

演出=演出_0                 -0.30379833  0.000000e+00
展览=展览_0                 -0.30388560  0.000000e+00
电脑=电脑_0                 -0.26265070  0.000000e+00
运动=运动_0                 -0.24685643  0.000000e+00
散步=散步_0                 -0.18447699  0.000000e+00
旅行=旅行_0                 -0.27033560  0.000000e+00
演奏音乐=演奏音乐_0         -0.26839870  0.000000e+00
```

```
                                Estimate       p.value
阅读=阅读_1                  0.23125560  0.000000e+00
听音乐=听音乐_1              0.25657041  0.000000e+00
电影=电影_1                  0.28307598  0.000000e+00
演出=演出_1                  0.30379833  0.000000e+00
展览=展览_1                  0.30388560  0.000000e+00
电脑=电脑_1                  0.26265070  0.000000e+00
运动=运动_1                  0.24685643  0.000000e+00
散步=散步_1                  0.18447699  0.000000e+00
旅行=旅行_1                  0.27033560  0.000000e+00
演奏音乐=演奏音乐_1          0.26839870  0.000000e+00
机械维修=机械维修_1          0.16540452 8.816716e-267
烹饪=烹饪_1                  0.15865273 9.423346e-247
…

演出=演出_0                 -0.30379833  0.000000e+00
展览=展览_0                 -0.30388560  0.000000e+00
电脑=电脑_0                 -0.26265070  0.000000e+00
运动=运动_0                 -0.24685643  0.000000e+00
散步=散步_0                 -0.18447699  0.000000e+00
旅行=旅行_0                 -0.27033560  0.000000e+00
演奏音乐=演奏音乐_0 -0.26839870  0.000000e+00
```


ACM 小结

- ACM 是适用于“个体 × 定性变量”表格的因子分析方法
- 特征值可解释为相关比平方的平均值
- ACM 可作为一种预处理方法，将定性变量转化并归纳为定量变量：在进行分类之前作为预处理十分有用

提纲

分类方法

层次聚类分析（CAH）

- 能否将相似的个体归为一类?
- 如何描述这些类别?

分类方法

层次聚类分析（CAH）

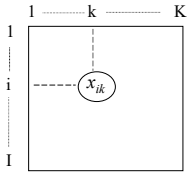
- 定义:
  - 分类：构建类别的过程
  - 类别：具有共同特征的个体（或对象）的集合（群组、类）
- 示例
  - 分类的例子：动物界、计算机硬盘、中国的地理区划、政党组织等
  - 类别的例子：社会阶层、政治阶层等
- 两种类型的分类:
  - 层次型：树状图、CAH
  - 划分法：分区

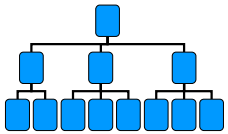
分类方法

层次聚类分析（CAH）

- 数据：个体 × 变量表

- 目标：构建一个树状结构，以便实现:
  - 揭示个体或个体群体之间的层次关系
  - 在总体中检测出“自然”的类别数



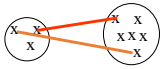


分类方法

分类——判别准则

- 个体间的相似性：欧氏距离、相似性指数等

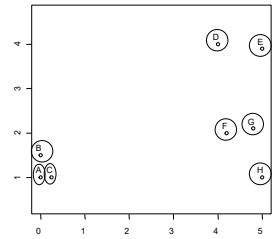
- 个体群体间的相似性：
  - 最短距离法（最小距离）
  - 最长距离法（最大距离）
  - Ward 准则



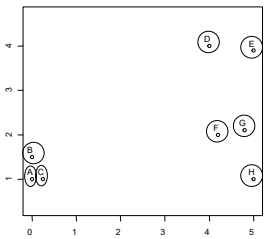
分类方法

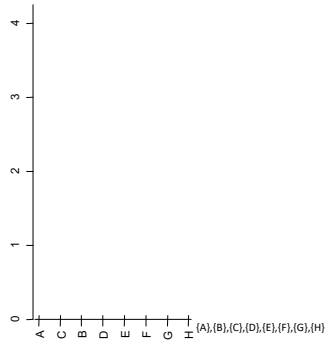
分类——算法

分类方法



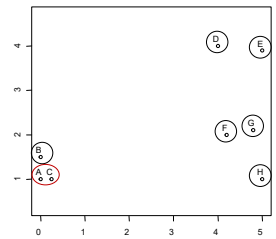
分类方法

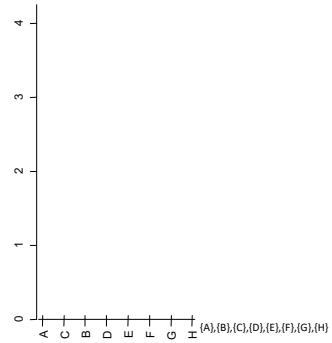




	A	B	C	D	E	F
B	0.50					
C	0.25	0.56				
D	5.00	4.72	4.80			
E	5.78	5.55	5.57	1.00		
F	4.32	4.23	4.07	2.01	2.06	
G	4.92	4.84	4.68	2.06	1.81	0.61
H	5.00	5.02	4.75	3.16	2.90	1.28	1.12

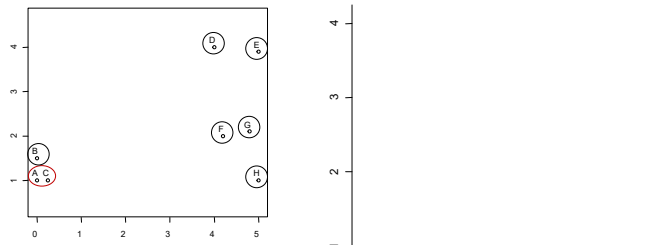
分类方法





	A	B	C	D	E	F
B	0.50					
C	0.25	0.56				
D	5.00	4.72	4.80			
E	5.78	5.55	5.57	1.00		
F	4.32	4.23	4.07	2.01	2.06	
G	4.92	4.84	4.68	2.06	1.81	0.61
H	5.00	5.02	4.75	3.16	2.90	1.28	1.12

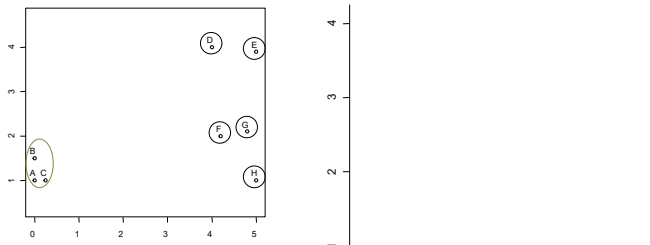
分类方法



	A	B	C	D	E	F	G
B	0.50						
C	0.25	0.56					
D	5.00	4.72	4.80				
E	5.78	5.55	5.57	1.00			
F	4.32	4.23	4.07	2.01	2.06		
G	4.92	4.84	4.68	2.06	1.81	0.61	
H	5.00	5.02	4.75	3.16	2.90	1.28	1.12

第一次合并

分类方法

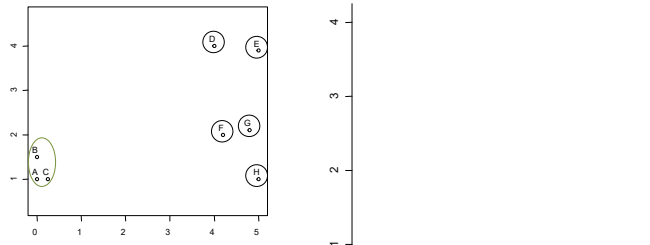


	AC	B	D	E	F	G
B	0.50					
D	4.80	4.72				
E	5.57	5.55	1.00			
F	4.07	4.23	2.01	2.06		
G	4.68	4.84	2.06	1.81	0.61	
H	4.75	5.02	3.16	2.90	1.28	1.12

	A	B	C	D	E	F	G
B	0.50						
C	0.25	0.56					
D	5.00	4.72	4.80				
E	5.78	5.55	5.57	1.00			
F	4.32	4.23	4.07	2.01	2.06		
G	4.92	4.84	4.68	2.06	1.81	0.61	
H	5.00	5.02	4.75	3.16	2.90	1.28	1.12

第一次合并

分类方法



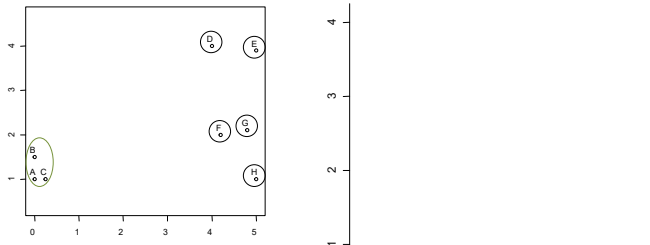
第二次合并

	AC	B	D	E	F	G
B	0.50					
D	4.80	4.72				
E	5.57	5.55	1.00			
F	4.07	4.23	2.01	2.06		
G	4.68	4.84	2.06	1.81	0.61	
H	4.75	5.02	3.16	2.90	1.28	1.12

	A	B	C	D	E	F	G
B	0.50						
C	0.25	0.56					
D	5.00	4.72	4.80				
E	5.78	5.55	5.57	1.00			
F	4.32	4.23	4.07	2.01	2.06		
G	4.92	4.84	4.68	2.06	1.81	0.61	
H	5.00	5.02	4.75	3.16	2.90	1.28	1.12

第一次合并

分类方法

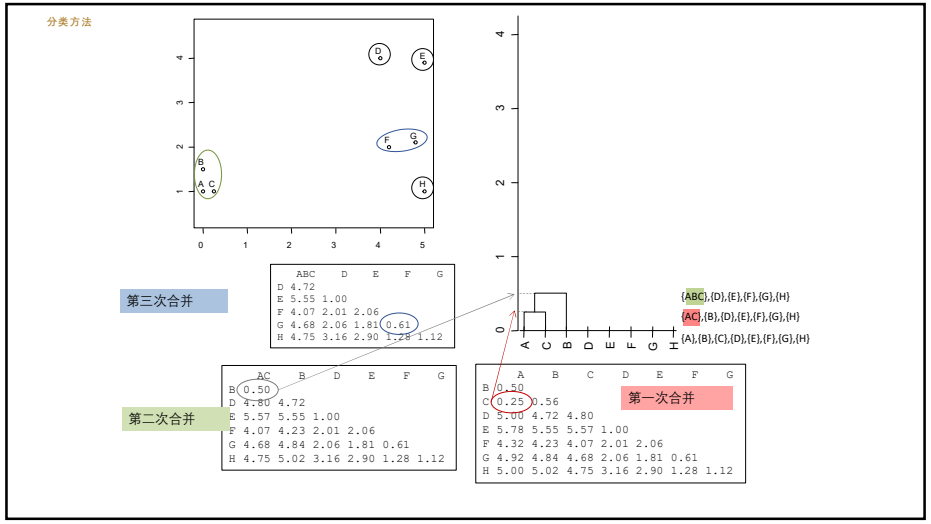


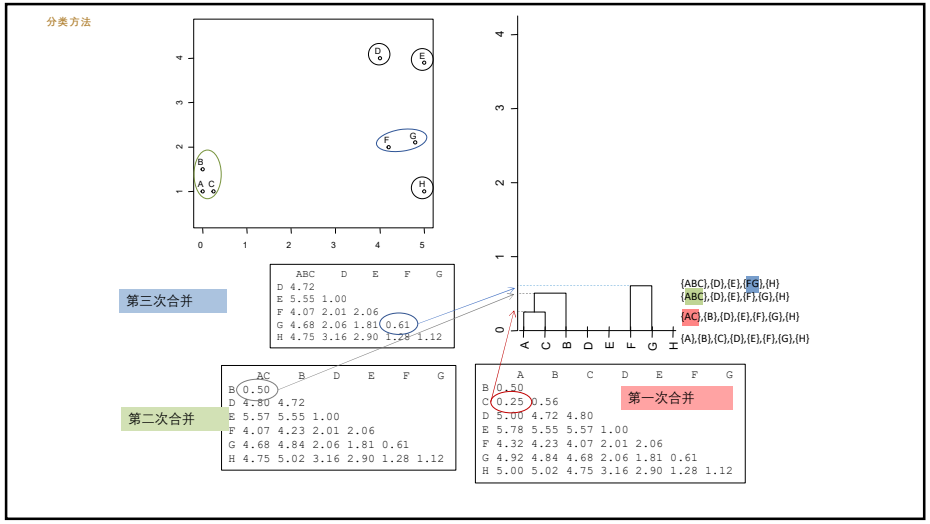
第二次合并

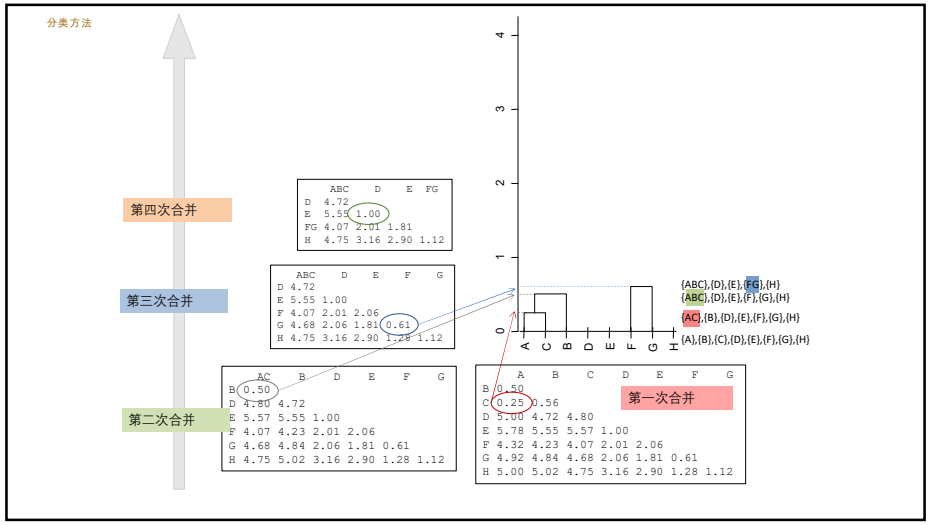
	AC	B	D	E	F	G
B	0.50					
D	4.80	4.72				
E	5.57	5.55	1.00			
F	4.07	4.23	2.01	2.06		
G	4.68	4.84	2.06	1.81	0.61	
H	4.75	5.02	3.16	2.90	1.28	1.12

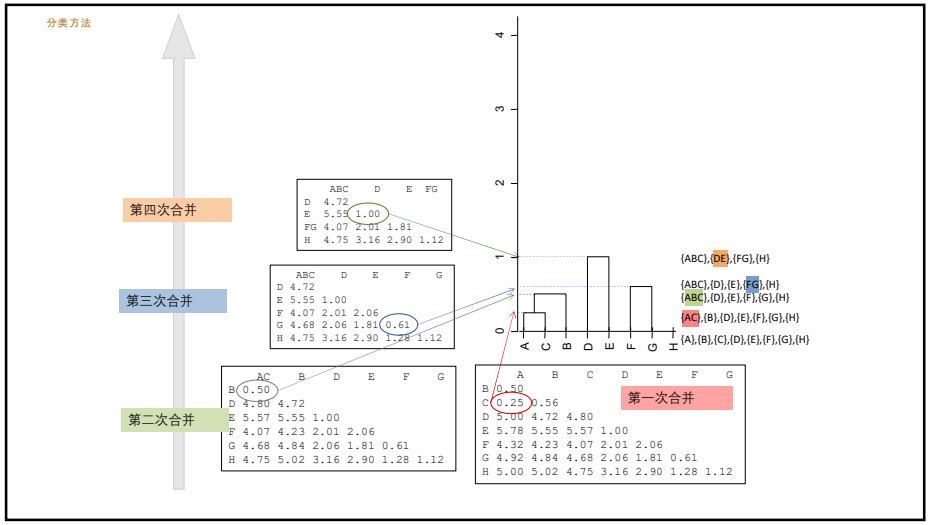
	A	B	C	D	E	F	G
B	0.50						
C	0.25	0.56					
D	5.00	4.72	4.80				
E	5.78	5.55	5.57	1.00			
F	4.32	4.23	4.07	2.01	2.06		
G	4.92	4.84	4.68	2.06	1.81	0.61	
H	5.00	5.02	4.75	3.16	2.90	1.28	1.12

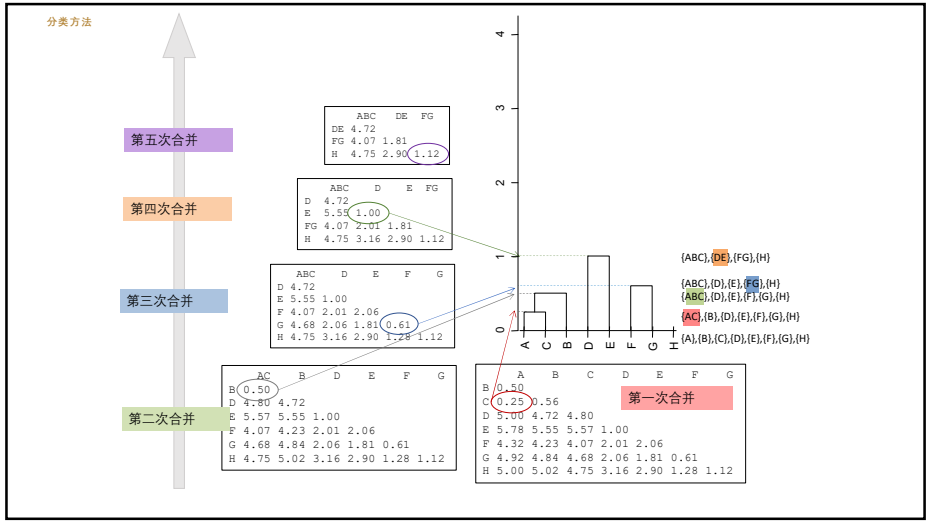
第一次合并

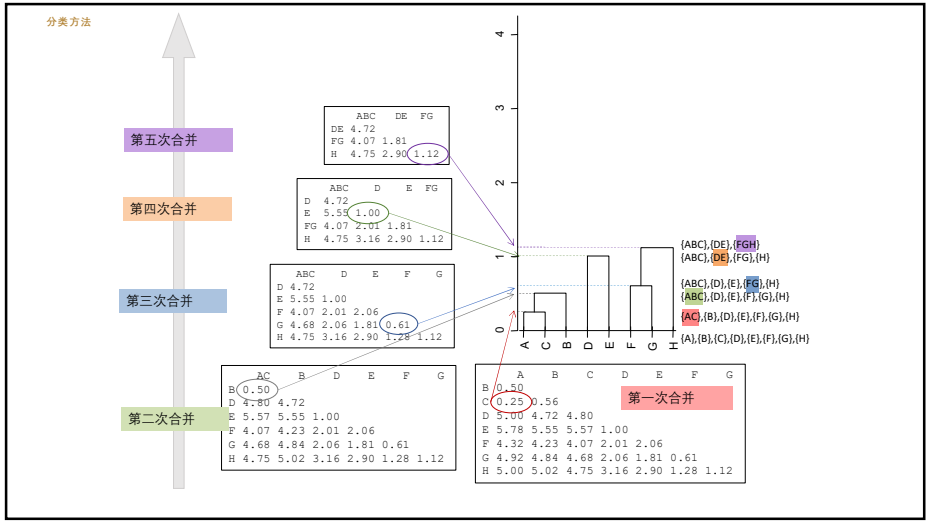

分类方法
第三次合并
第二次合并
第一次合并
{ABC},{D},{E},{F},{G},{H}
{AC},{B},{D},{E},{F},{G},{H}
{A},{B},{C},{D},{E},{F},{G},{H}

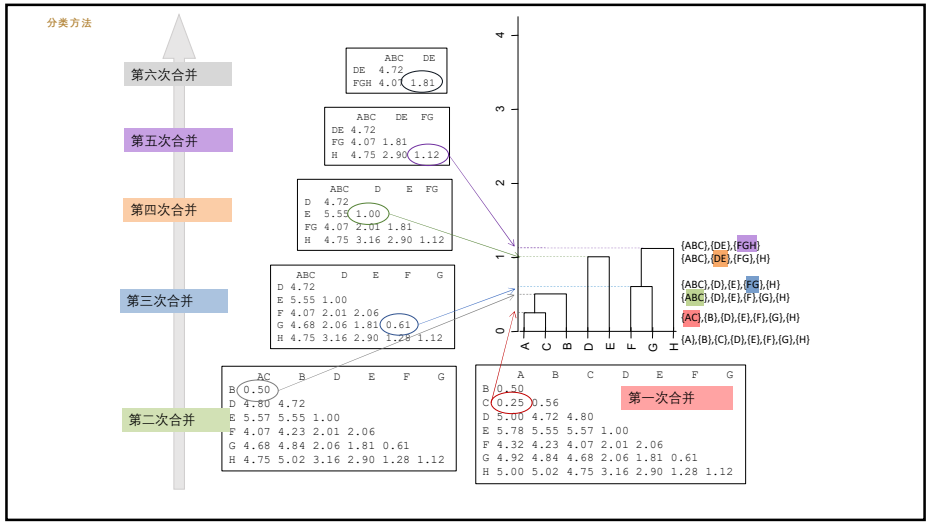

分类方法
第三次合并
第二次合并
第一次合并
{ABC},{D},{E},{FG},{H}
{ABC},{D},{E},{F},{G},{H}
{AC},{B},{D},{E},{F},{G},{H}
{A},{B},{C},{D},{E},{F},{G},{H}

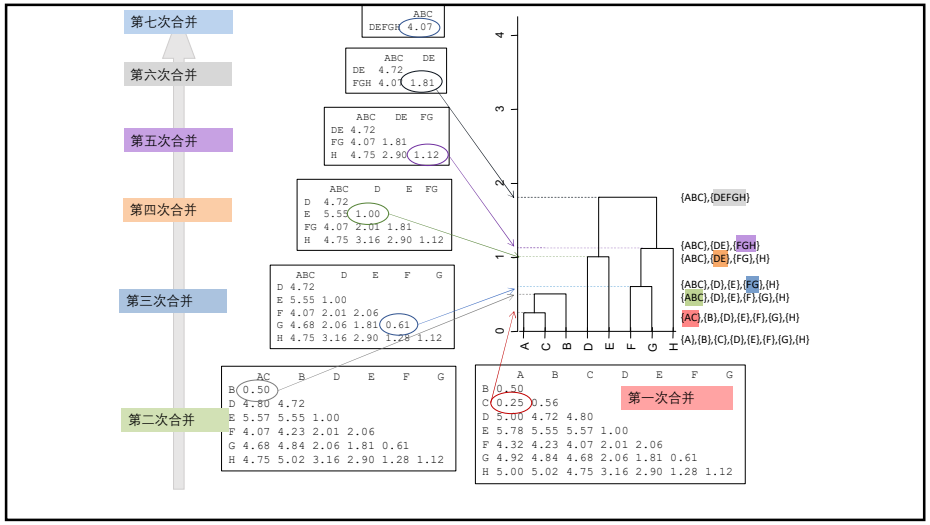

分类方法
第四次合并
第三次合并
第二次合并
第一次合并
{ABC},{D},{E},{FG},{H}
{ABC},{D},{E},{F},{G},{H}
{AC},{B},{D},{E},{F},{G},{H}
{A},{B},{C},{D},{E},{F},{G},{H}

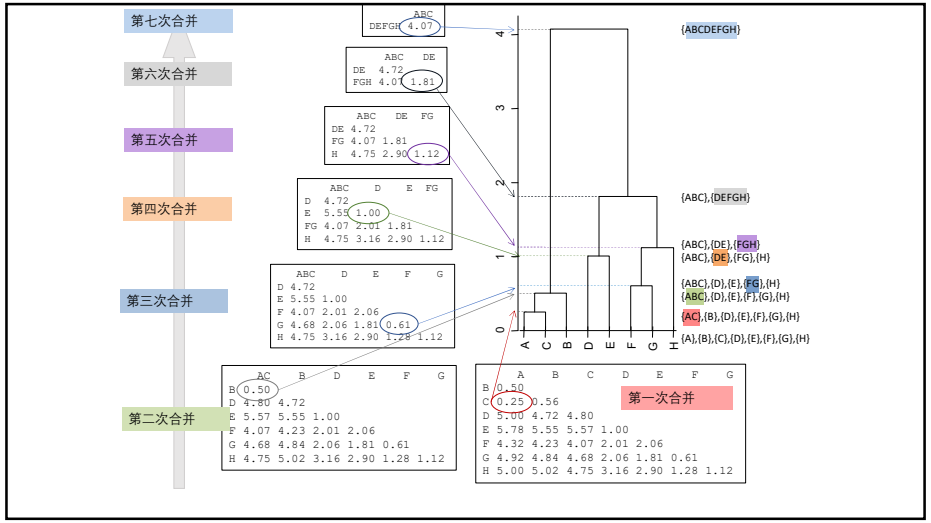

分类方法
第四次合并
第三次合并
第二次合并
第一次合并
{ABC},{DE},{FG},{H}
{ABC},{D},{E},{FG},{H}
{ABC},{D},{E},{F},{G},{H}
{AC},{B},{D},{E},{F},{G},{H}
{A},{B},{C},{D},{E},{F},{G},{H}


分类方法
第五次合并
第四次合并
第三次合并
第二次合并
第一次合并
{ABC},{DE},{FG},{H}
{ABC},{D},{E},{FG},{H}
{ABC},{D},{E},{F},{G},{H}
{AC},{B},{D},{E},{F},{G},{H}
{A},{B},{C},{D},{E},{F},{G},{H}


分类方法
第五次合并
第四次合并
第三次合并
第二次合并
第一次合并
{ABC},{DE},{FGH}
{ABC},{DE},{FG},{H}
{ABC},{D},{E},{FG},{H}
{ABC},{D},{E},{F},{G},{H}
{AC},{B},{D},{E},{F},{G},{H}
{A},{B},{C},{D},{E},{F},{G},{H}


分类方法
第六次合并
第五次合并
第四次合并
第三次合并
第二次合并
第一次合并
{ABC},{DE},{FGH}
{ABC},{DE},{FG},{H}
{ABC},{D},{E},{FG},{H}
{ABC},{D},{E},{F},{G},{H}
{AC},{B},{D},{E},{F},{G},{H}
{A},{B},{C},{D},{E},{F},{G},{H}


第七次合并
第六次合并
第五次合并
第四次合并
第三次合并
第二次合并
第一次合并
{ABC},{DEFGH}
{ABC},{DE},{FGH}
{ABC},{DE},{FG},{H}
{ABC},{D},{E},{FG},{H}
{ABC},{D},{E},{F},{G},{H}
{AC},{B},{D},{E},{F},{G},{H}
{A},{B},{C},{D},{E},{F},{G},{H}



分类方法

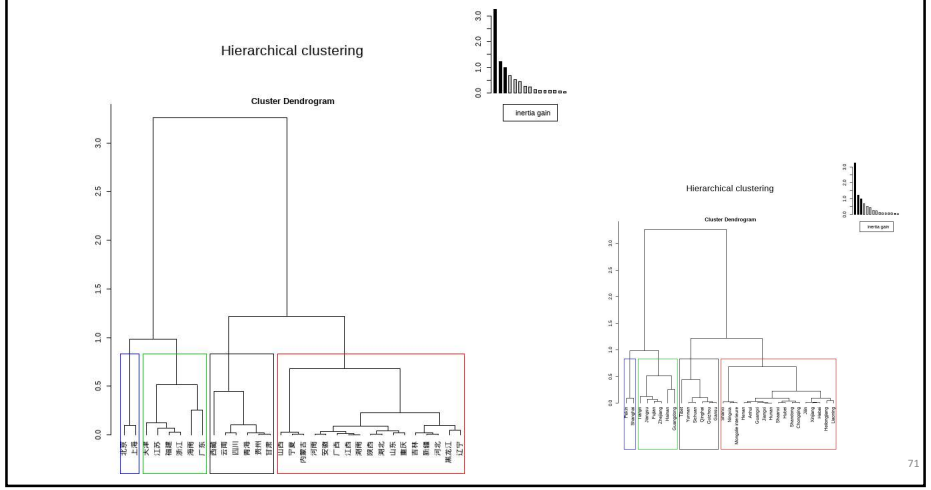
树状图与分区

所有的树状图最终都会 ……被切断！！！

通过设定一个切割水平，可以构建出一个分区

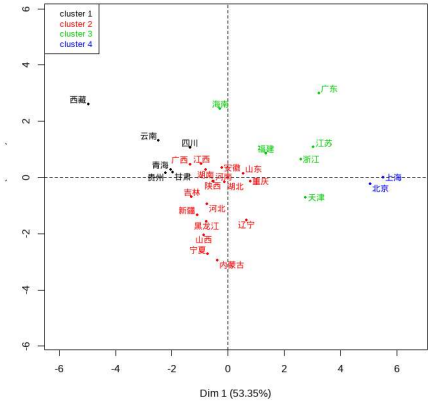
注：鉴于其构建方式，该分区并非最优，但仍具有参考价值

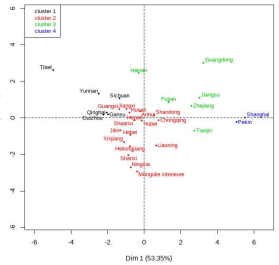
分类方法



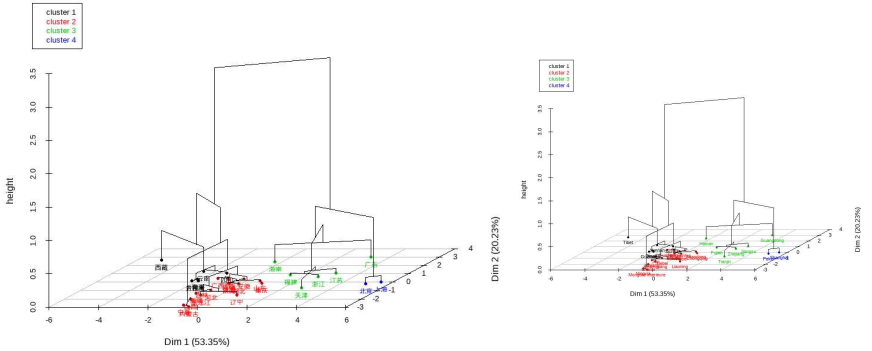
分类方法

示例



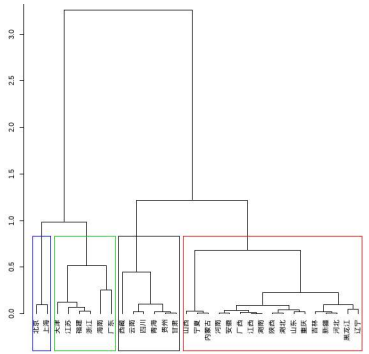


示例

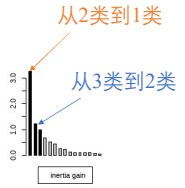


类别数的选择

- 根据树状图
- 根据层次指标直方图
- 取决于具体用途（调查等）
- 最终标准：类别的可解释性



从……过渡时损失的惯量



类别的描述

目标:

用定量变量刻画某一类别（或某个个体群体）的特征

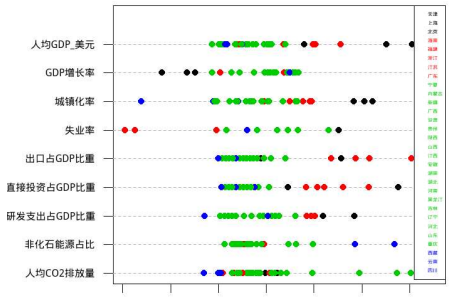
对刻画各类别特征的变量进行排序

问题:

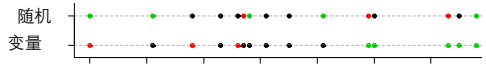
如何刻画第1类个体的特征?

哪些变量最能刻画其特征?

类别的描述

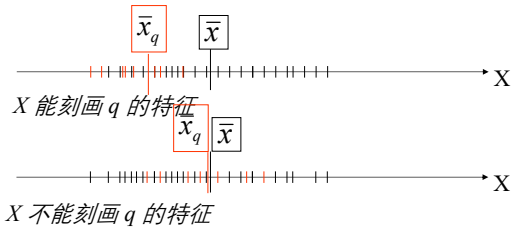


思路1：如果变量 X 在类别 q 中的取值看起来是在 X 的全部取值中随机分布的，那么 X 不能刻画类别 q 的特征



思路2：随机抽样假设越不可信，X 就越能刻画类别 q 的特征

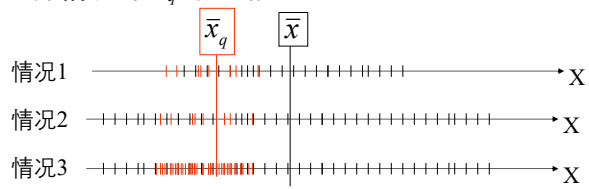
类别的描述



将类别 q 中个体（针对变量 X）的均值与总体均值进行比较

类别的描述

以下 3 种情况中 $\bar{x}_q$ 与 $\bar{x}$ 相同：



情况1和2：　情况1中取值更为极端：
在情况1中 X 对 q 的解释力更强　⇨　以 s 进行相对化处理

情况2：5个个体位于同一位置：是偶然吗？
情况3：40个个体位于同一位置：是偶然吗？
⇨　以……进行相对化处理 I_q

类别的描述

从 I 个值中随机抽取 I_q 个值的参照

它可以取哪些值 $\bar{x}_q$？（即它服从什么分布 $\bar{X}_q$？）

$$E(\bar{X}_q) = \bar{x} \qquad V(\bar{X}_q) = \frac{s^2}{I_q}\left(\frac{I - I_q}{I - 1}\right)$$

$\mathcal{L}(\bar{X}_q) = \mathcal{N}$　因为 $\bar{X}_q$ 是一个均值

类别的描述

$$\text{Valeur-test} = \frac{\bar{x}_q - \bar{x}}{\sqrt{\frac{s^2}{I_q}\left(\frac{I-I_q}{I-1}\right)}}$$ 服从标准正态分布

⇨　若 $|\text{V - test}| \geq 2$ 成立，则 X 能刻画类别 q 的特征

⇨　*V-test 值越大，X 对类别 q 的刻画就越好*

思路：按 V-test 值从大到小对变量进行排序

分类方法

类别的描述

Link between the cluster variable and the quantitative variables

	Eta2	P-value
外商直接投资占GDP比重	0.8627702	9.021128e-12
人均GDP_美元	0.8265594	2.087745e-10
非化石能源占比	0.7541122	2.227762e-08
城镇化率	0.7302165	7.678294e-08
研发支出占GDP比重	0.6916529	4.548965e-07
出口占GDP比重	0.6420226	3.298775e-06
GDP增长率	0.4985992	2.787017e-04
失业率	0.3184458	1.453253e-02

Link between the cluster variable and the quantitative variables

	Eta2	P-value
IDE_PIB	0.8627702	9.021128e-12
PIB_hab_USD	0.8265594	2.087745e-10
Energies_non_fossiles	0.7541122	2.227762e-08
Taux_urbanisation	0.7302165	7.678294e-08
RD_PIB	0.6916529	4.548965e-07
Exportations_PIB	0.6420226	3.298775e-06
Croissance_PIB	0.4985992	2.787017e-04
Chomage	0.3184458	1.453253e-02

81

分类方法

类别的描述

Description of each cluster by quantitative variables

$`1`

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value	n
非化石能源占比	4.748753	60.50	26.580645	19.163768	19.166123	2.046750e-06	6
GDP增长率	2.786335	5.35	3.596774	1.906786	1.688382	5.330777e-03	6
出口占GDP比重	-2.090804	5.70	16.103226	2.473190	13.351223	3.654568e-02	6
人均GDP_美元	-2.229094	7900.00	12403.225806	1051.982256	5420.777463	2.580765e-02	6
城镇化率	-3.077896	52.70	64.851613	7.921700	10.593664	2.084675e-03	6

$`2`

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value	n
人均CO2排放量	2.275400	10.3470588	8.480645	5.709653	4.950752	0.022881937	17
出口占GDP比重	-2.041065	11.5882353	16.103226	4.088129	13.351223	0.041244351	17
非化石能源占比	-2.776230	17.7647059	26.580645	5.620649	19.166123	0.005499335	17
外商直接投资占GDP比重	-2.849958	0.7823529	1.380645	0.360123	1.267056	0.004372498	17

$`3`

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value	n
出口占GDP比重	3.965287	35.833333	16.103226	15.2361048	13.3512228	0.0000733076	6
外商直接投资占GDP比重	3.429354	3.000000	1.380645	0.7047458	1.2670564	0.0006050195	6
城镇化率	2.304541	73.950000	64.851613	6.1578541	10.5936643	0.0211922897	6
人均GDP_美元	1.994901	16433.333333	12403.225806	3660.9045634	5420.7774630	0.0460537009	6
失业率	-2.509469	3.366667	3.793548	0.6968182	0.4564488	0.0120912953	6

$`4`

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value	n
人均GDP_美元	4.138565	28000.00	12403.225806	1500.00	5420.7774630	3.494841e-05	2
研发支出占GDP比重	3.964356	5.20	1.929032	1.00	1.1868075	7.359444e-05	2
外商直接投资占GDP比重	3.257354	4.25	1.380645	0.75	1.2670564	1.124561e-03	2
城镇化率	3.197360	88.40	64.851613	0.90	10.5936643	1.386915e-03	2
失业率	2.068651	4.45	3.793548	0.05	0.4564488	3.857881e-02	2
GDP增长率	-2.851230	0.25	3.596774	0.45	1.6883823	4.355046e-03	2

82

分类方法

类别的描述

Link between the cluster variable and the categorical variables (chi-square test)

	p.value	df
地区	0.0002062437	9

Description of each cluster by the categories

$`1`

	Cla/Mod	Mod/Cla	Global	p.value	v.test
地区=西部	50	100	38.70968	0.001254956	3.226086

$`2`

	Cla/Mod	Mod/Cla	Global	p.value	v.test
地区=中部	100	35.29412	19.35484	0.01680880	2.390864
地区=东部	20	11.76471	32.25806	0.01078815	-2.549487

$`3`

	Cla/Mod	Mod/Cla	Global	p.value	v.test
地区=东部	60	100	32.25806	0.0002852172	3.628369
地区=西部	0	0	38.70968	0.0368500613	-2.087421

$`4`
NULL

	p.value	df
Region	0.0002062437	9

Description of each cluster by the categories

$`1`

	Cla/Mod	Mod/Cla	Global	p.value	v.test
Region=Ouest	50	100	38.70968	0.001254956	3.226086

$`2`

	Cla/Mod	Mod/Cla	Global	p.value	v.test
Region=Centre	100	35.29412	19.35484	0.01680880	2.390864
Region=Est	20	11.76471	32.25806	0.01078815	-2.549487

$`3`

	Cla/Mod	Mod/Cla	Global	p.value	v.test
Region=Est	60	100	32.25806	0.0002852172	3.628369
Region=Ouest	0	0	38.70968	0.0368500613	-2.087421

$`4`
NULL

83

分类方法

类别的描述

	1	2	3	4
城镇化率	-3.08	0	2.3	3.2
人均GDP_美元	-2.23	0	1.99	4.14
出口占GDP比重	-2.09	-2.04	3.97	0
地区=东部	0	-2.55	3.63	0
地区=中部	0	2.39	0	0
人均CO2排放量	0	2.28	0	0
外商直接投资占GDP比重	0	-2.85	3.43	3.26
失业率	0	0	-2.51	2.07
研发支出占GDP比重	0	0	0	3.96
GDP增长率	2.79	0	0	-2.85
地区=西部	3.23	0	-2.09	0
非化石能源占比	4.75	-2.78	0	0

classif.html

84

分类方法

另一种划分方法——kmeans

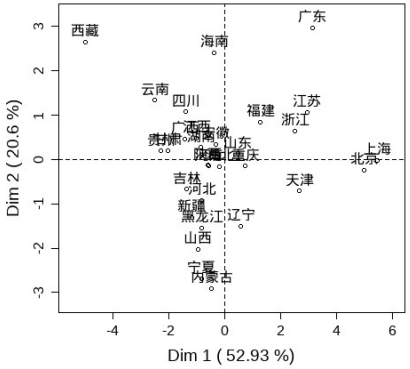
直接构建一个分区（类别数 = Q）：
动态聚类算法（围绕移动中心点）

0.1. 随机选取 Q 个点
0.2. 将其余各点分配给距离最近的类别中心
1. 计算 Q 个新的重心
2. 将各点分配给距离最近的类别中心
重复步骤 1 和 2，直至收敛

分类方法

另一种划分方法——kmeans

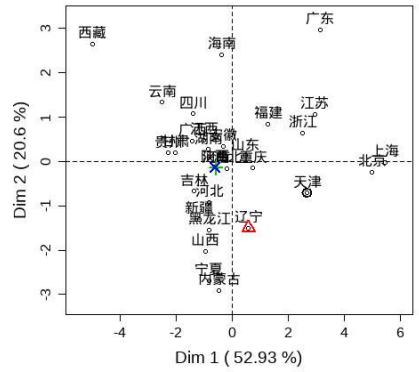
0.1. 随机选取 Q 个点
0.2. 将其余各点分配给距离最近的类别中心
1. 计算 Q 个新的重心
2. 将各点分配给距离最近的类别中心
重复步骤 1 和 2，直至收敛



分类方法

另一种划分方法——kmeans

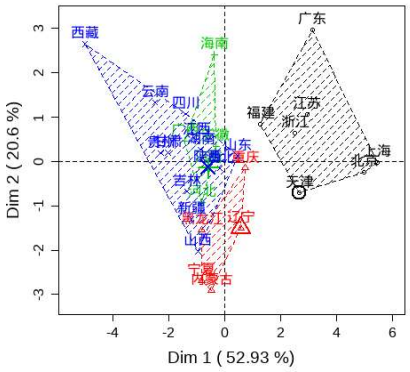
0.1. 随机选取 Q 个点
0.2. 将其余各点分配给距离最近的类别中心
1. 计算 Q 个新的重心
2. 将各点分配给距离最近的类别中心
重复步骤 1 和 2，直至收敛



分类方法

另一种划分方法——kmeans

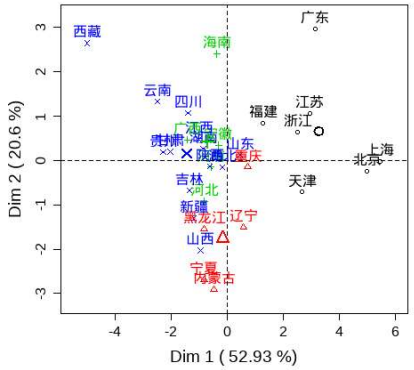
0.1. 随机选取 Q 个点
0.2. 将其余各点分配给距离最近的类别中心
1. 计算 Q 个新的重心
2. 将各点分配给距离最近的类别中心
重复步骤 1 和 2，直至收敛



分类方法

另一种划分方法——kmeans

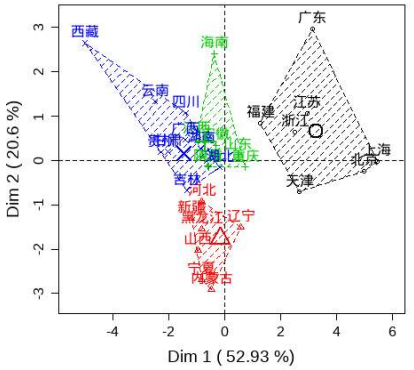
0.1. 随机选取 Q 个点
0.2. 将其余各点分配给距离最近的类别中心
1. 计算 Q 个新的重心
2. 将各点分配给距离最近的类别中心
重复步骤 1 和 2，直至收敛



分类方法

另一种划分方法——kmeans

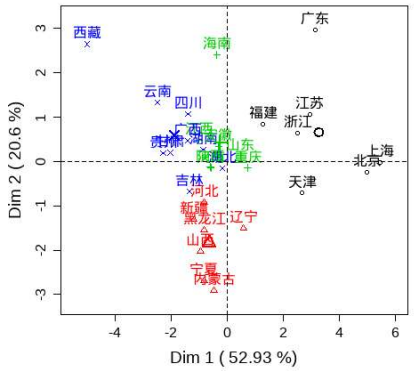
0.1. 随机选取 Q 个点
0.2. 将其余各点分配给距离最近的类别中心
1. 计算 Q 个新的重心
2. 将各点分配给距离最近的类别中心
重复步骤 1 和 2，直至收敛



分类方法

另一种划分方法——kmeans

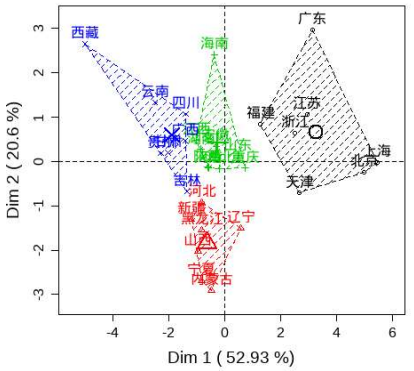
0.1. 随机选取 Q 个点
0.2. 将其余各点分配给距离最近的类别中心
1. 计算 Q 个新的重心
2. 将各点分配给距离最近的类别中心
重复步骤 1 和 2，直至收敛



分类方法

另一种划分方法——kmeans

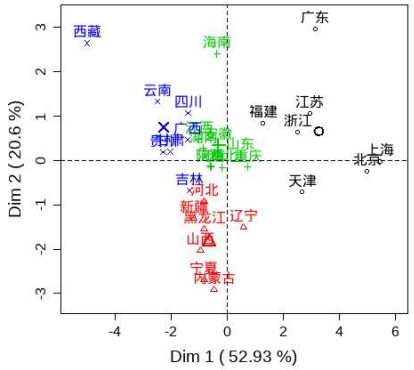
0.1. 随机选取 Q 个点
0.2. 将其余各点分配给距离最近的类别中心
1. 计算 Q 个新的重心
2. 将各点分配给距离最近的类别中心
重复步骤 1 和 2，直至收敛



分类方法

另一种划分方法——kmeans

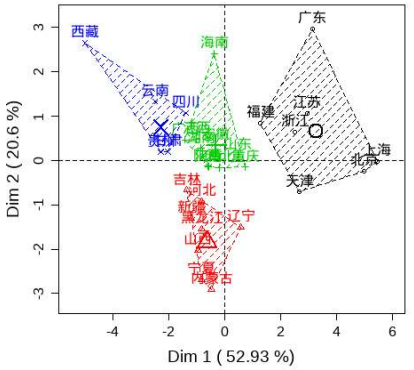
0.1. 随机选取 Q 个点
0.2. 将其余各点分配给距离最近的类别中心
1. 计算 Q 个新的重心
2. 将各点分配给距离最近的类别中心
重复步骤 1 和 2，直至收敛



分类方法

另一种划分方法——kmeans

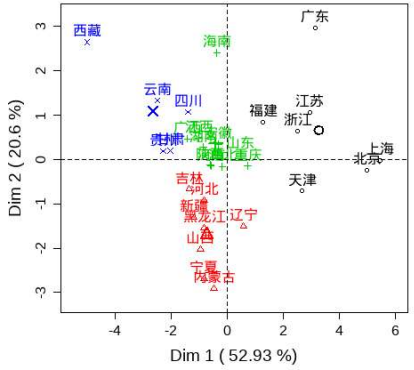
0.1. 随机选取 Q 个点
0.2. 将其余各点分配给距离最近的类别中心
1. 计算 Q 个新的重心
2. 将各点分配给距离最近的类别中心
重复步骤 1 和 2，直至收敛



分类方法

另一种划分方法——kmeans

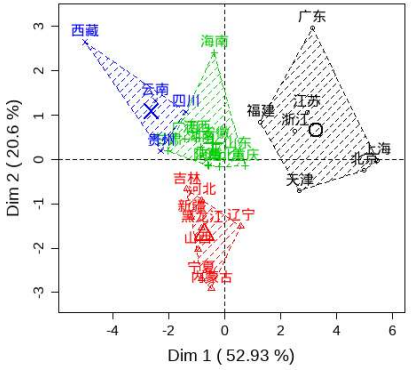
0.1. 随机选取 Q 个点
0.2. 将其余各点分配给距离最近的类别中心
1. 计算 Q 个新的重心
2. 将各点分配给距离最近的类别中心
重复步骤 1 和 2，直至收敛



分类方法

另一种划分方法——kmeans

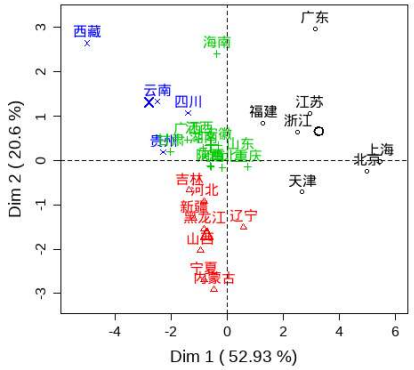
0.1. 随机选取 Q 个点
0.2. 将其余各点分配给距离最近的类别中心
1. 计算 Q 个新的重心
2. 将各点分配给距离最近的类别中心
重复步骤 1 和 2，直至收敛



分类方法

另一种划分方法——kmeans

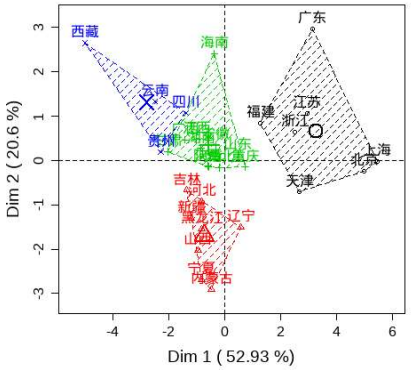
0.1. 随机选取 Q 个点
0.2. 将其余各点分配给距离最近的类别中心
1. 计算 Q 个新的重心
2. 将各点分配给距离最近的类别中心
重复步骤 1 和 2，直至收敛



分类方法

另一种划分方法——kmeans

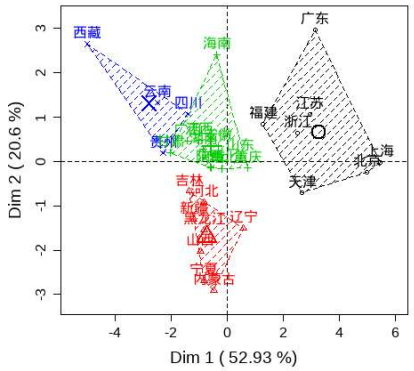
0.1. 随机选取 Q 个点
0.2. 将其余各点分配给距离最近的类别中心
1. 计算 Q 个新的重心
2. 将各点分配给距离最近的类别中心
重复步骤 1 和 2，直至收敛



分类方法

另一种划分方法——kmeans

0.1. 随机选取 Q 个点
0.2. 将其余各点分配给距离最近的类别中心
1. 计算 Q 个新的重心
2. 将各点分配给距离最近的类别中心
重复步骤 1 和 2，直至收敛



FactoMineR 与 Factoshiny：免费的实践工具

R 软件：

免费、开源的软件
众多扩展包使该软件不断发展壮大

FactoMineR 与 Factoshiny 扩展包：
全球下载量达数百万次（在中国有多少呢？）

使用帮助

参考书目：

《用R进行数据分析》（第2版）

《R语言统计与数据科学》

多维数据分析 MOOC 课程：

https://husson.github.io/MOOC_AnaDo/index.html

数据分析 YouTube 播放列表：

主成分分析 ACP / PCA
对应分析 AFC / Correspondence Analysis
多重对应分析 ACM / MCA
分类 / 聚类 Classification / Clustering
多重因子分析 AFM / MFA
缺失数据处理 / Handling missing values