

财务管理理论与实务

Financial Data Mining

唐润宇
2023-10-10



数据挖掘 Data mining

数据挖掘(Data Mining)就是从大量的数据中，提取隐藏在其中的，事先不知道的、但潜在有用的信息的过程。

常见任务 https://en.wikipedia.org/wiki/Data_mining

- 异常检测: 识别不寻常的数据记录，错误数据需要进一步调查。
- 关联规则学习: 搜索变量之间的关系。
- 聚类: 是在未知数据的结构下，发现数据的类别与结构。
- 分类: 是对新的数据推广已知的结构的任务。
- 回归: 试图找到能够以最小误差对该数据建模的函数。
- 汇总: 提供了一个更紧凑的数据集表示。

Ref: Data mining in R: <https://www.rdatamining.com/>

数据降维

Human minds are good at recognizing patterns in two dimensions and to some extent in three, but are essentially useless in higher dimensions.

常见方法:

- PCA 主成分分析
- ICA 独立分量分析
- kPCA 核主成分分析
- MDS 多维标度分析

Ref: Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.

PCA

主成分分析是一种通过降维技术把多个变量化为少数几个主成分（综合变量）的统计分析方法。

这些主成分能够反映原始变量的绝大部分信息，它们通常表示为原始变量的某种线性组合，且彼此不相关。

1901年由 Karl Pearson发明。

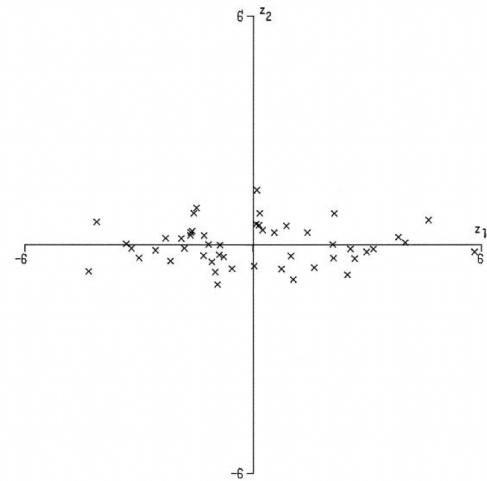
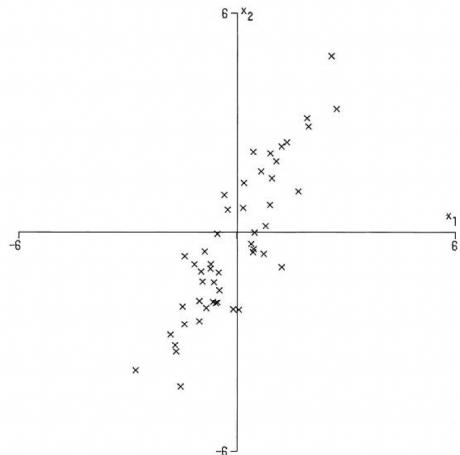
常用的科学手段：2020年Nature中有124篇文章使用了该方法。

<https://mathvoices.ams.org/featurecolumn/2021/08/01/principal-component-analysis/>

PCA

- 每一个主成分都是各原始变量的线性组合
- 主成分的数目大大少于原始变量的数目
- 主成分保留了原始变量的绝大部分信息
- 各个主成分之间互不相关。

什么是信息? Or 什么样的变量信息量更大?



PCA

- 从不同视角观测弹簧拉长变化
- 选择最优的视角?

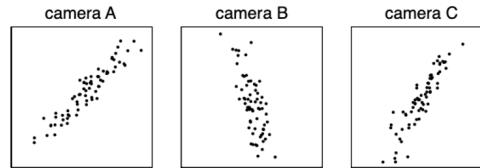
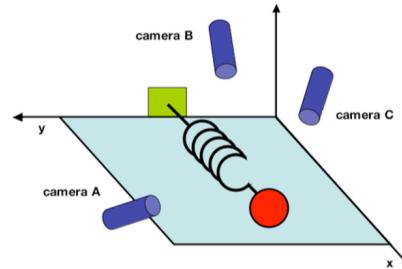


FIG. 1 A toy example. The position of a ball attached to an oscillating spring is recorded using three cameras A, B and C. The position of the ball tracked by each camera is depicted in each panel below.

PCA的数学定义

假设数据有 p 维特征 $\mathbf{x} = [x_1, x_2, \dots, x_p]'$, 假设我们知道其协方差矩阵为 Σ . 考虑线性组合

$$y_1 = \mathbf{a}_1' \mathbf{x} = a_{11}x_1 + a_{21}x_2 + \dots + a_{p1}x_p$$

$$y_2 = \mathbf{a}_2' \mathbf{x} = a_{12}x_1 + a_{22}x_2 + \dots + a_{p2}x_p$$

...

$$y_p = \mathbf{a}_p' \mathbf{x} = a_{1p}x_1 + a_{2p}x_2 + \dots + a_{pp}x_p$$

我们有 $Var(y_i) = \mathbf{a}_i' \Sigma \mathbf{a}_i$, $Cov(y_i, y_j) = \mathbf{a}_i' \Sigma \mathbf{a}_j$.

主成分: 使得 $Var(y_i)$ 尽可能大的线性无关的 y_i 其中 $(\mathbf{a}_i' \mathbf{a}_i = 1)$ 。

假设 p 维随机向量 \mathbf{x} 的协方差矩阵 Σ 的特征值和特征向量为

$(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \dots, (\lambda_p, \mathbf{e}_p)$, 其中 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$, 则第 i 个主成分为

$$\mathbf{y}_i = \mathbf{e}_i' \mathbf{x} = e_{i1}x_1 + e_{i2}x_2 + \dots + e_{ip}x_p, \quad i = 1, 2, \dots, p$$

一些基础知识回顾

- 协方差矩阵: 是一个方阵, 其 i, j 位置的元素是第 i 个与第 j 个随机变量之间的协方差。

$$\Sigma_{ij} = \text{cov}(x_i, x_j) = E[(x_i - \mu_i)(x_j - \mu_j)]$$

- 有的时候我们会给数据标准化, 此时对应着相关系数矩阵 R 。

$$R_{ij} = \text{corr}(x_i, x_j) = \frac{\text{cov}(x_i, x_j)}{\sigma_i \sigma_j}.$$

- (对称)矩阵的特征值分解(Eigendecomposition)

$$A = Q\Lambda Q^\top$$

$$S = Q\Lambda Q^T$$

对称矩阵 S 可以进行
特征值分解
特征向量组成 Q , 特征值组成 Λ

PCA 例子

CSMAR数据库 -> 财务指标分析 -> 偿债能力 时间跨度: 2022/01/01-2023/06/30

```
library(tidyverse)
library(readxl)
setwd("~/XJTU/课程/财务管理理论与实务/Codes")
data = read_xlsx("../Data/FI_T1.xlsx", skip = 1)
str(data)
```

```
## # tibble [50,254 × 35] (S3:tbl_df/tbl/data.frame)
## $ 股票代码 : chr [1:50254] "没有单位" "000001" "000001" "000001" ...
## $ 股票简称 : chr [1:50254] "没有单位" "平安银行" "平安银行" "平安银行" ...
## $ 统计截止日期 : chr [1:50254] "没有单位" "2022-03-31" "2022-03-31" "2022-06-30" ...
## $ 报表类型编码 : chr [1:50254] "没有单位" "A" "B" "A" ...
## $ 行业代码 : chr [1:50254] "没有单位" "J66" "J66" "J66" ...
## $ 行业名称 : chr [1:50254] NA "货币金融服务" "货币金融服务" "货币金融服务" ...
## $ 公告来源 : chr [1:50254] "没有单位" "0.0" "0.0" "0.0" ...
## $ 流动比率 : chr [1:50254] "没有单位" NA NA NA ...
## $ 速动比率 : chr [1:50254] "没有单位" NA NA NA ...
## $ 保守速动比率 : chr [1:50254] "没有单位" NA NA NA ...
## $ 现金比率 : chr [1:50254] "没有单位" NA NA NA ...
```

PCA 例子

整理数据:

- 选取每个公司最新的数据
- 转换数据格式
- 删掉非数字的变量
- 删掉NA的公司

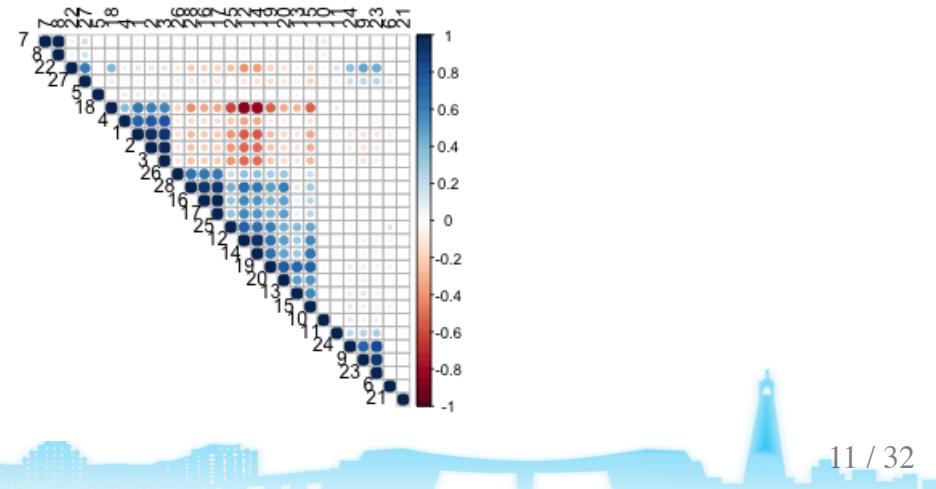
```
PCA_data <- data |>
  group_by(股票代码) |>
  slice(which.max(as.Date(统计截止日期, '%Y-%m-%d')))) |>
  ungroup() |>
  type_convert() |>
  select(-c(1:7)) |>
  drop_na()

# str(PCA_data)
```

PCA 例子

在做PCA之前，我们可以观察一下我们的数据的相关系数矩阵

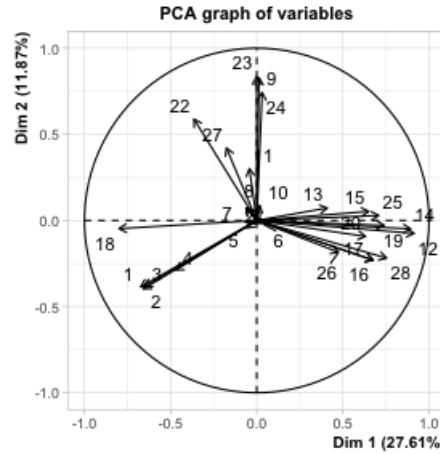
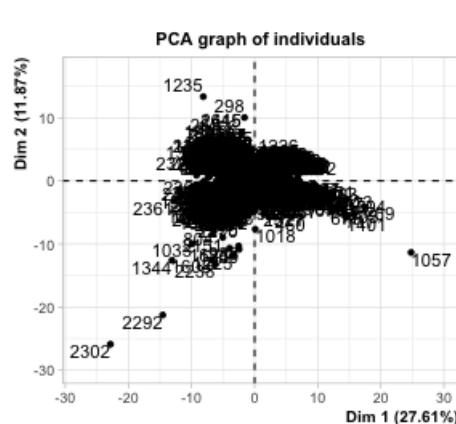
```
names(PCA_data) <- c(1:length(PCA_data))
cor.mat <- cor(PCA_data)
library("corrplot")
corrplot(cor.mat, type="upper", order="hclust",
         tl.col="black")
```



PCA 例子

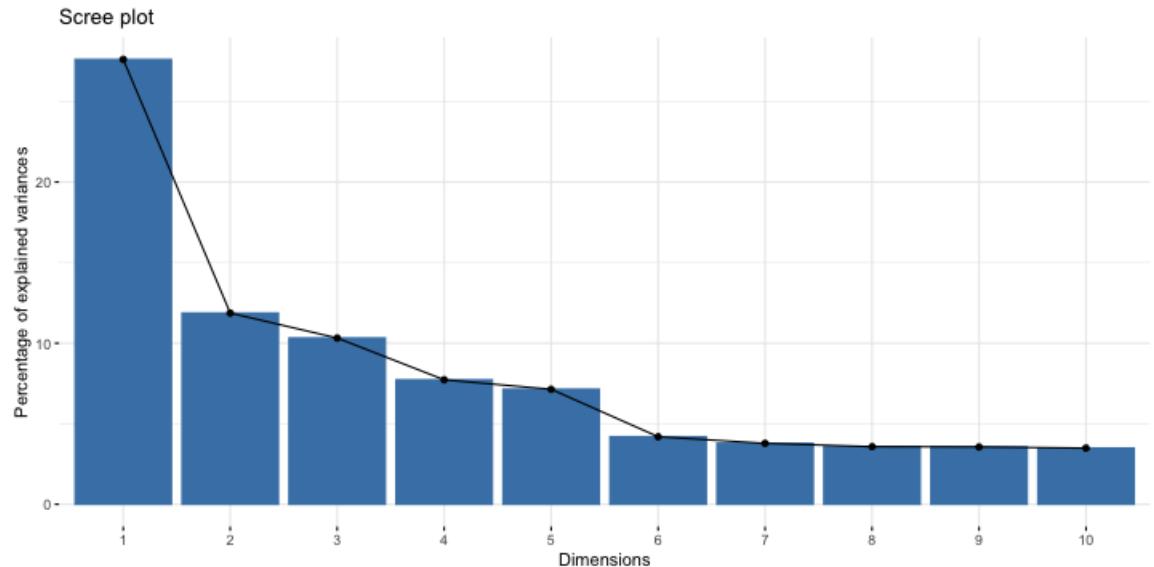
我们使用FactoMineR帮助我们做PCA

```
library(FactoMineR)  
pca.res <- PCA(PCA_data)
```



PCA 例子

```
library(factoextra)  
fviz_screenplot(pca.res)
```



PCA 例子

PCA的结果里面有很多东西，比如我们通常比较关注的特征值和特征向量。

```
pca.res <- PCA(PCA_data, ncp=5, graph=FALSE)
round(pca.res$eig, 3)
```

```
##      eigenvalue percentage of variance cumulative percentage of variance
## comp 1    7.732        27.614             27.614
## comp 2    3.324        11.870            39.484
## comp 3    2.889        10.319            49.803
## comp 4    2.163        7.726            57.529
## comp 5    1.999        7.139            64.669
## comp 6    1.174        4.193            68.862
## comp 7    1.059        3.782            72.643
## comp 8    1.001        3.576            76.220
## comp 9    0.996        3.559            79.779
## comp 10   0.976        3.487            83.265
## comp 11   0.866        3.094            86.359
## comp 12   0.834        2.978            89.337
## comp 13   0.520        1.856            91.193
## comp 14   0.434        1.548            92.742
## comp 15   0.385        1.375            94.117
## comp 16   0.329        1.174            95.291
```

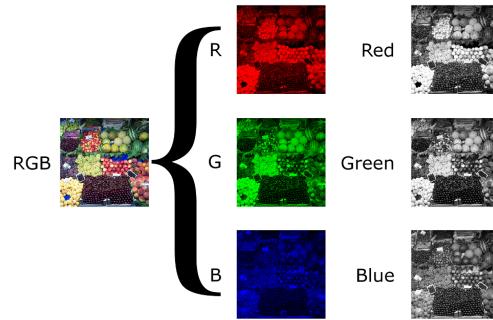
PCA的应用 --- 图片压缩

一张图片是由三原色(RGB)叠加出来的。

每个颜色下面都是一个矩阵，矩阵的维度就是我们常说的像素数。

例: 某遥遥领先品牌新机:

<https://consumer.huawei.com/cn/phones/mate60-pro/specs/>



- 后置摄像头: 5000 万像素超光变摄像头
- 后置摄像头照片分辨率: 最大可支持 8192×6144 像素

PCA的应用 --- 图片压缩



我们尝试着压缩一张钱学森图书馆的图片。

```
library(png)  
image <- readPNG("../Data/xjtu.png")  
str(image)
```

```
## num [1:535, 1:713, 1:3] 0.886 0.886 0.886 0.886 0.8
```

原图的大小为535*713.

PCA的应用 --- 图片压缩



主成分数量 = 10



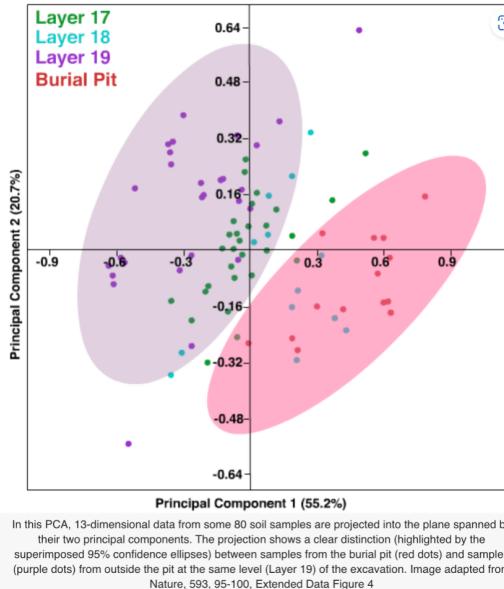
主成分数量 = 50



PCA的应用

- 78000岁的尸骨墓穴
- 如何确定是“墓穴”呢?
- 骨头周围的泥土成分和周围成分
泥土成分显著不同 !
- 80 个土壤样本
- 每个样本中13维信息 (Si, K, Ca, Ti, Mn, Fe, Zn, Ga, As, Rb, Y, Zr and Ba)

Earliest known human burial in Africa, published in *Nature* on May 5, 2021



PCA的应用

表 7

公司治理指数的载荷系数

治理指标	变量名称	变量解释	载荷系数
持股结构与股东权益	最大股东持股比例	第 1 大股东持股比例	-0.5469
	股权制衡	第 2 大到第 5 大股东持股之和除以第 1 大股东持股比例	0.4677
	股东会次数	公司年度召开的股东大会次数	0.2014
	流通股比例	公司流通股所占比例	0.2307
	国有股比例	公司国有股所占比例	-0.4910
管理层治理	两职合一	公司董事长与 CEO 是否兼任	0.1294
	管理层持股	公司管理层持股比例	0.2730
董事、监事与其他治理形式	董事会规模	公司董事会人数	-0.0964
	独立董事比例	公司董事会中独立董事所占比例	0.1013
	董事会次数	公司年度召开董事会次数	0.1551
	监事会次数	公司年度召开监事会次数	0.1113
	委员会个数	公司设立的各种委员会个数(如薪酬委员会、考核委员会、审计委员会和战略发展委员会等)	-0.0007

白重恩等. 中国上市公司治理结构的实证研究[J]. 经济研究 , 2005, 11.

张学勇,廖理.股权分置改革、自愿性信息披露与公司治理[J].经济研究, 2010, 45(04): 28-39+53.

胡楠,薛付婧,王昊楠.管理者短视主义影响企业长期投资吗? ——基于文本分析和机器学习[J].管理世界, 2021, 37(05):139-156+11+19-21

PCA 的局限性

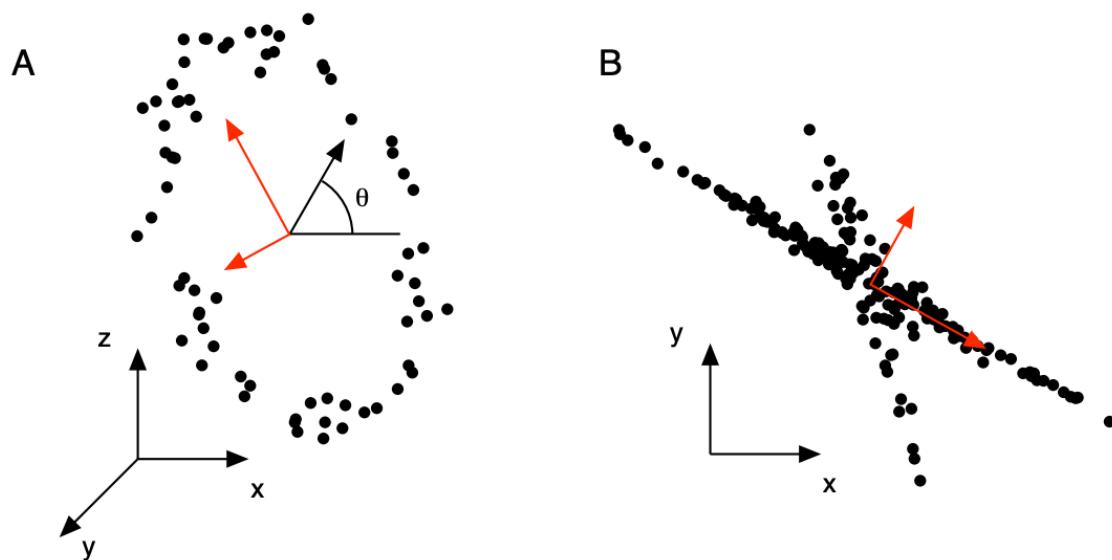


FIG. 6 Example of when PCA fails (red lines). (a) Tracking a person on a ferris wheel (black dots). All dynamics can be described by the phase of the wheel θ , a non-linear combination of the naive basis. (b) In this example data set, non-Gaussian distributed data and non-orthogonal axes causes PCA to fail. The axes with the largest

其他数据降维方法

- kPCA: 将线型函数使用核函数(kernel)变成非线性，可以描述更加复杂的关系
kernlab::kpca
- ICA: 放松PCA中的不同主成分需要线性无关的假设 ica::ica
- MDS: Multidimensional scaling

我们有n个观测值 $x_1, \dots, x_n \in \mathbb{R}^p$, 每个观测值是 p 维的, 那么MDS需要找到一组投影 $z_1, \dots, z_n \in \mathbb{R}^k, k < p$ 使得下面的stress function取得最小。

$$\min_{z_1, \dots, z_n} \sum_{i \neq j} (d_{ij} - \|z_i - z_j\|)^2.$$

实际上就是找一个低维的投影，但是能够维持原数据之间的间隔。

如果我们取 $d_{ij} = \langle x_i - \mu_i, x_j - \mu_j \rangle$, 那么此时MDS就与PCA一样了。

我们可以使用R语言中的**cmdscale**函数实现MDS.

聚类分析

- 分类: 有监督学习
- 聚类: 无监督学习

All relate to grouping or segmenting a collection of objects into subsets or "clusters", such that those within each cluster are more closely related to one another than objects assigned to different clusters.

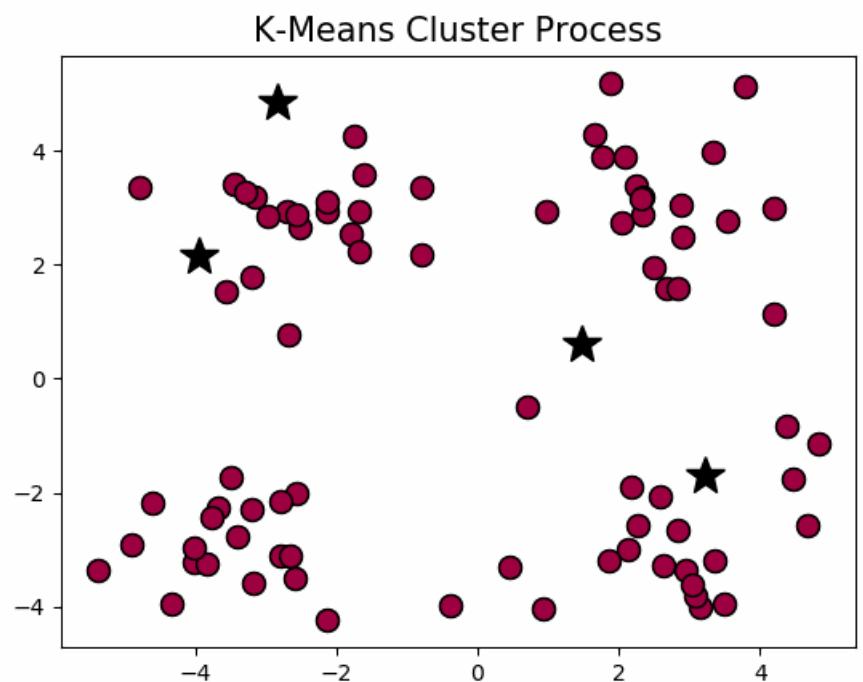
常见的聚类方法:

K-means: 基于距离的聚类

DBSCAN: 基于密度的聚类

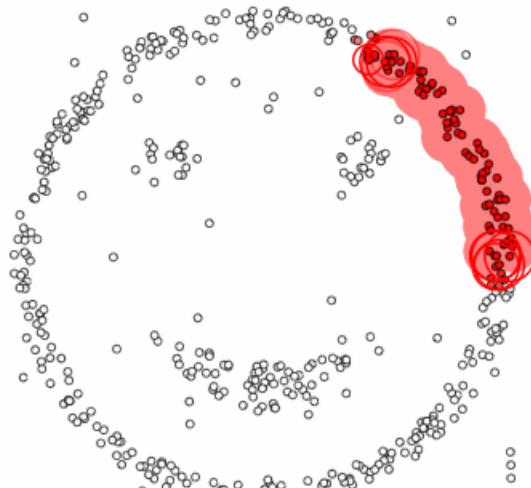
Hierarchical Clustering: 层次化聚类方法

K means



DBSCAN

epsilon = 1.00
minPoints = 4



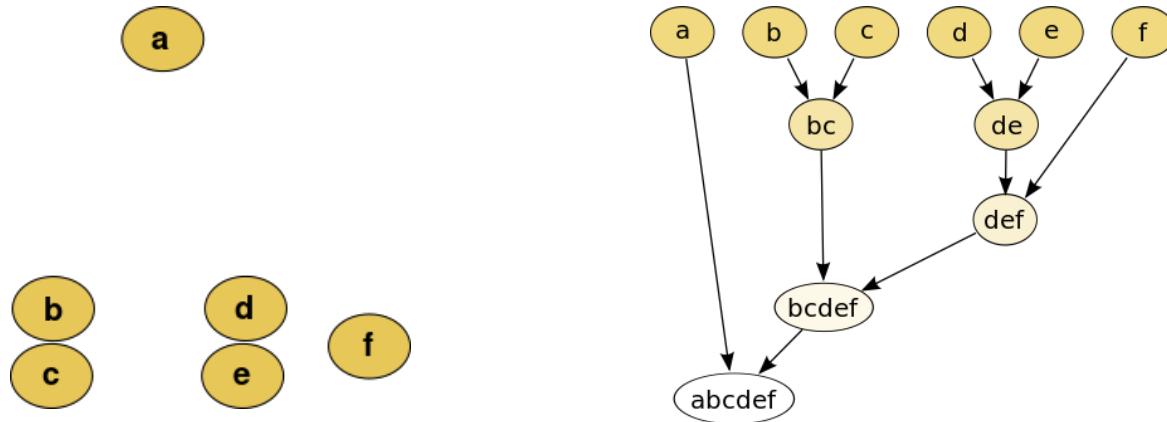
Restart



Pause



Agglomerative 层次聚类



常见聚类算法比较

上述常见的聚类算法有着不同的优缺点，我们需要针对数据的不同选择合适的方法。

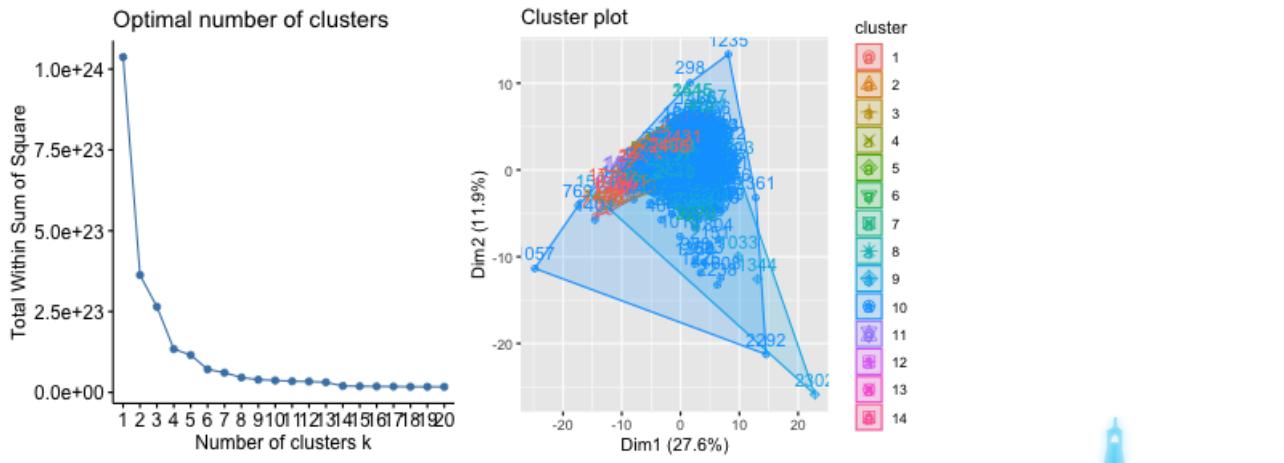
当然我们也可以都试试看~

算法	数据类型	抗噪点	聚类形状	算法效率
K-means	混合型	较差	球形	很高
DBSCAN	数值型	较好	任意形状	一般
Agglomerative	混合型	较好	任意形状	较差

例子

K-means 效果

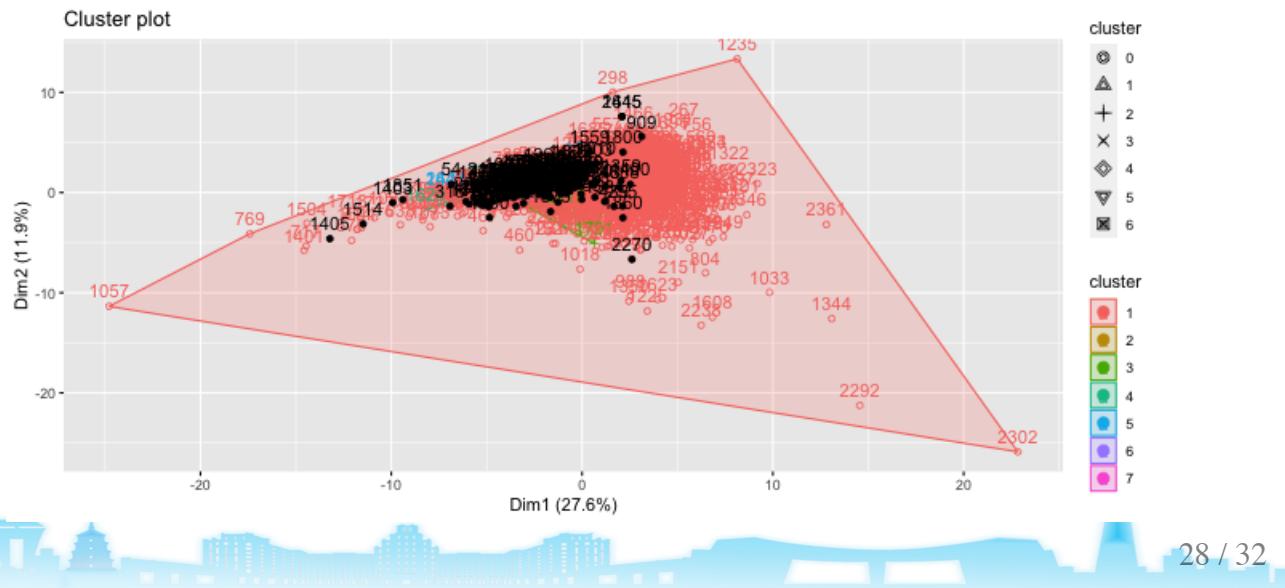
```
library(cluster)
fviz_nbclust(PCA_data, kmeans, method = "wss", k.max=20)
km <- kmeans(PCA_data, centers=14)
fviz_cluster(km, data = PCA_data)
```



例子

DBSCAN 效果

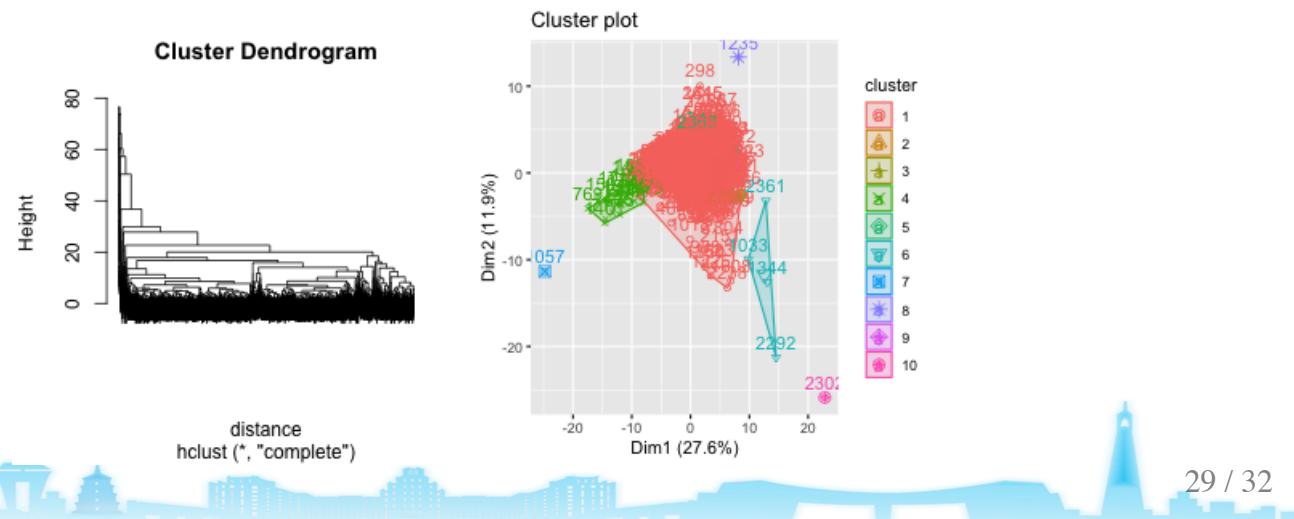
```
library(fpc)
db <- dbscan(PCA_data, eps = 2000000000)
fviz_cluster(db, data = PCA_data)
```



例子

Agglomeration 效果

```
distance <- get_dist(PCA_data, stand=TRUE)
hcl <- hclust(distance, method = "complete")
plot(hcl, labels=FALSE)
hc_cut <- hccut(distance, k=10, hc_method = "complete")
fviz_cluster(hc_cut, data = PCA_data)
```



文本挖掘

我们的财务数据里面也有很多例如公司简介的文本信息，我们可以使用R对这些文本进行分析。



```
data <- read_csv("../Data/IRR_ListedInfo.csv")
newdata <- data |>
  select(ShortName, CompanyProfile) |>
  drop_na()
```

Ref: <https://www.tidytextmining.com/>

```
test <- newdata$CompanyProfile
library(jiebaR)
library(tm)
library(tmcn)
library(wordcloud2)
cutter = worker() # we can add more options here
test_seg = segment(test, cutter)
test_df = data.frame(table(test_seg))
new_df <- test_df |>
  filter(Freq > 1)
wordcloud2(new_df)
```



Claudia Goldin 发表刊物标题词云

