



北京航空航天大学

— 经济管理学院 —

BEIHANG UNIVERSITY
SCHOOL OF ECONOMICS AND MANAGEMENT

第 5 讲：主成分分析

康雁飞

数量经济与商务统计系

Outline

- 1 为什么主成分分析?
- 2 PCA 原理
- 3 PCA 算法
- 4 案例分析
- 5 主成分分析在 R 中的实现
- 6 作业

Outline

- 1 为什么主成分分析?
- 2 PCA 原理
- 3 PCA 算法
- 4 案例分析
- 5 主成分分析在 R 中的实现
- 6 作业

搜集了 329 个社区的 9 项指标评分：气候和地形、住房、医疗保健与环境、犯罪、交通、教育、艺术、娱乐、经济。

- 除了住房和犯罪之外，得分越高越好；对于住房和犯罪，得分越低越好。
- 目标是：如何对 329 个社区进行评级？

可能存在的问题及解决思路

- 当变量数量较多时，协方差矩阵可能过大，难以正确研究和解释
- 变量之间的两两相关性太多，无法全部考虑
- 当数据集非常庞大时，图形显示也可能不是特别有帮助

问题：如何展示一个 9 维数据？

- 在信息损失最小的前提下，对高维空间进行降维处理。
- 在一个低维空间辨识系统要比在高维空间容易得多。

- 主成分分析，简称 PCA (Principal Component Analysis)
 - 1 是一种用于降低数据维度并找出数据集中变量之间的模式和关系的统计技术
 - 2 原始变量通过线性组合转换为一组新的互相正交的变量，称为主成分
 - 3 这些主成分按照保留数据方差的方式排列，第一个主成分解释最大方差，第二个主成分解释剩余方差中的最大部分，以此类推
 - 4 有助于简化数据集、减少冗余信息、压缩数据表示，并揭示数据中的模式和结构
- 广泛的应用：数据降维与可视化、模式识别与数据挖掘（特征提取、数据压缩和分类任务等）等

主成分分析的主要目的

通过线性组合将原始变量变换成一组新的互相正交的变量（主成分），并且这些主成分按照保留原始数据方差的大小顺利排列，可以达到：

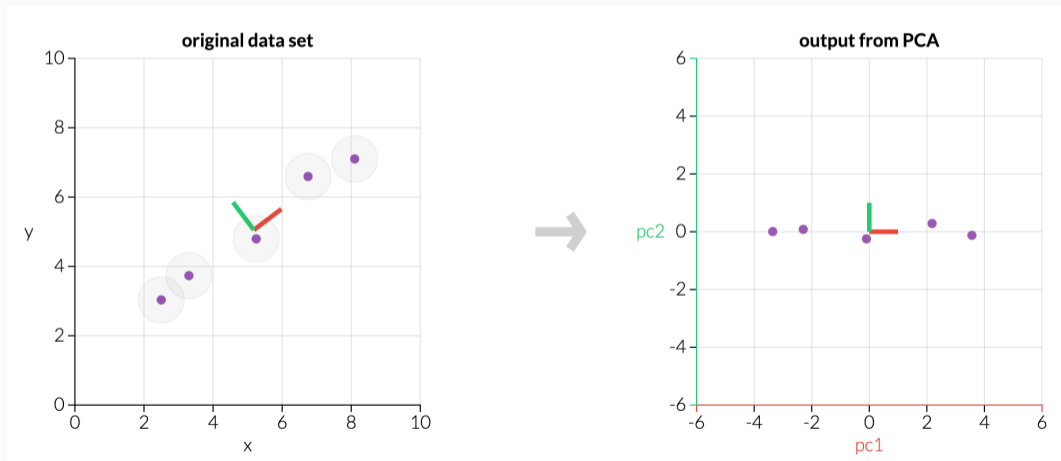
- 1 高维数据降维，发现数据中潜在的模式和结构
- 2 易于展示

Outline

- 1 为什么主成分分析?
- 2 PCA 原理
- 3 PCA 算法
- 4 案例分析
- 5 主成分分析在 R 中的实现
- 6 作业

怎样能够对数据进行降维处理？

通过平移 + 旋转省去数据变异不大方向的信息。



- 降维之后的每个维度都是原数据维度的线性组合：

$$\mathbf{y}_h = \sum_{j=1}^p \alpha_{hj} \mathbf{x}_j = X \alpha_h, \quad (h = 1, \dots, m, m \ll p).$$

$$\begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}_{n \times p} \Rightarrow \begin{bmatrix} Y_{11} & Y_{12} & \cdots & Y_{1p} \\ Y_{21} & Y_{22} & \cdots & Y_{2p} \\ \vdots & \vdots & & \vdots \\ Y_{n1} & Y_{n2} & \cdots & Y_{np} \end{bmatrix}_{n \times p} \Rightarrow \begin{bmatrix} Y_{11} & \cdots & Y_{1m} \\ Y_{21} & \cdots & Y_{2m} \\ \vdots & & \vdots \\ Y_{n1} & \cdots & Y_{nm} \end{bmatrix}_{n \times m}$$

- 线性组合的方差和协方差： $\text{var}(\mathbf{y}_h) = \alpha_h^T \Sigma \alpha_h$, $\text{cov}(\mathbf{y}_i, \mathbf{y}_j) = \alpha_i^T \Sigma \alpha_j$.

第一主成分 (PC1): y_1

- 满足在所有的线性组合中具有最大方差，即尽可能地多解释数据中的变化
- 选择 α_1 满足：

$$\max \text{var}(\mathbf{y}_1) = \alpha_1^T \Sigma \alpha_1, \text{ subject to } \alpha_1^T \alpha_1 = 1.$$

第二主成分 (PC2): y_2

- 第二个主成分是 X 变量的线性组合，尽可能多地解释剩余的变化，同时受到第一和第二主成分之间相关性为零的约束

- 选择 α_2 满足：

$$\begin{aligned} \max \text{ var}(\mathbf{y}_2) &= \alpha_2^T \Sigma \alpha_2, \text{ subject to} \\ \alpha_2^T \alpha_2 &= 1, \text{ cov}(\mathbf{y}_1, \mathbf{y}_2) = \alpha_1^T \Sigma \alpha_2 = 0 \end{aligned}$$

- 所有随后的主成分都具有相同的特性：它们是线性组合，尽可能多地解释剩余的变化，并且彼此之间不相关。

第 i 个主成分 (PC i): \mathbf{y}_i

- 选择 α_i 满足

$$\max \text{var}(\mathbf{y}_i) = \alpha_i^T \Sigma \alpha_i$$

- 1 $\alpha_i^T \alpha_i = 1$

- 2 $\text{cov}(\mathbf{y}_l, \mathbf{y}_i) = \alpha_l^T \Sigma \alpha_i = 0, l = 1, \dots, i - 1.$

问题：如何得到 α_i ?

Outline

- 1 为什么主成分分析?
- 2 PCA 原理
- 3 PCA 算法**
- 4 案例分析
- 5 主成分分析在 R 中的实现
- 6 作业

- $\lambda_1, \dots, \lambda_p$ 为协方差矩阵 Σ 的特征根，并且排序使得

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$$

$\alpha_1, \alpha_2, \dots, \alpha_p$ 是对应的特征向量。

- 主成分中线性组合的系数（载荷向量，loadings）为协方差矩阵 Σ 的单位特征向量。
- $\text{var}(\mathbf{y}_i) = \lambda_i$: 第 i 个主成分的方差为协方差矩阵 Σ 的特征根 λ_i 。

上述主成分分析基于协方差矩阵来进行，可能存在的问题：

- 1 主成分分析的结果取决于测量尺度。
- 2 样本方差最高的变量往往在前几个主成分中被强调。
- 3 只有当所有变量具有相同的计量单位时，才应考虑使用协方差函数进行主成分分析。

解决方法： \Rightarrow 基于中心化和标准化后的数据进行主成分分析，即基于相关系数矩阵。

基于相关系数矩阵的主成分分析

- 1 数据标准化：将 X 的每一列减其均值，除以标准差。标准化的矩阵记为 Z ，其中 $Z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$ 。
- 2 计算 Z 的方差协方差矩阵，即 X 的相关系数矩阵，记 $\Sigma = \frac{1}{n-1} Z^T Z$ 。
- 3 求 Σ 的特征值和特征向量，将 Σ 的特征值排序： $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ ，对应的特征向量进行排序，构成 Q ，这个过程即将 Σ 对角化： $\Lambda = Q^T \Sigma Q$ 。
- 4 计算 $Y = ZQ$ ，取 Y 的前 m 列，即 X 的前 m 个主成分得分。 Q 的每一列叫作载荷向量。第 h 个主成分得分即：

$$y_h = \sum_{j=1}^p q_{hj} z_j = Z q_h.$$

为什么上述算法成立?

我们的目标是

- 1 想让 \mathbf{y}_1 携带最多的信息，也就是 \mathbf{y}_1 的方差取到最大值，依次类推。
- 2 降维之后不相关。

$$\mathbf{q}_1 = \operatorname{argmax}_{\|\mathbf{q}_1\|=1} \{\mathbf{q}_1' \Sigma \mathbf{q}_1\}$$

$$\mathbf{q}_2 = \operatorname{argmax}_{\|\mathbf{q}_2\|=1} \{\mathbf{q}_2' \Sigma \mathbf{q}_2\} \quad \text{subject to} \quad \mathbf{q}_1' \Sigma \mathbf{q}_2 = 0$$

\vdots

$$\mathbf{q}_\ell = \operatorname{argmax}_{\|\mathbf{q}_\ell\|=1} \{\mathbf{q}_\ell' \Sigma \mathbf{q}_\ell\} \quad \text{subject to} \quad \mathbf{q}_k' \Sigma \mathbf{q}_\ell = 0 \quad \forall k < \ell$$

- 拉格朗日算法求得 q_1 满足 $\Sigma q_1 = \lambda_1 q_1$. 其中 $\lambda_1 = \text{var}(\mathbf{y}_1)$.
- $\lambda_h = \text{var}(\mathbf{y}_h)$.
- $\bar{\mathbf{y}}_h = 0$.
- $\text{cov}(\mathbf{y}_i, \mathbf{y}_j) = 0$.

PCA 的累计贡献率

- 前 m 个主成分的累计贡献率:

$$\frac{\sum_{k=1}^m \lambda_k}{\sum_{k=1}^p \lambda_k}.$$

- 注意: $\sum_{k=1}^p \lambda_k = \sum_{k=1}^p \text{var}(\mathbf{y}_k) = \sum_{k=1}^p \text{var}(\mathbf{z}_k) = p.$

主成分与原变量的相关系数

- $\text{cor}(\mathbf{y}_1, \mathbf{z}_j) = \sqrt{\lambda_1} \mathbf{q}_{1j}$.

Outline

- 1 为什么主成分分析?
- 2 PCA 原理
- 3 PCA 算法
- 4 案例分析**
- 5 主成分分析在 R 中的实现
- 6 作业

降维的两个特殊作用

- 1 将一个高维变量系统有效的降至二维（抽象思维 → 形象思维）
- 2 将一个高维变量系统有效的降至一维（综合指数）

案例一：管理期刊遴选研究

- 35 个样本点
- 4 个特征: Citation (被引次数)、PaperNo (载文量)、RefNo (引证期刊量)、NSFC (标注“国家自然科学基金项目”)
- 样本示例:

##	Citation	PaperNO	RefNO	NSFC
## 管理世界	33	175	54	1
## 系统理实	47	285	64	59
## 系工学报	21	60	35	24
## 中国软科	20	293	117	0
## 数量经济	37	212	22	10
## 中国管科	1	41	26	3

主成分和原变量之间的关系?

载荷 (loadings)

```
##
## Loadings:
##          Comp.1 Comp.2 Comp.3 Comp.4
## Citation  0.634          0.395  0.664
## PaperNO   -0.936  0.281 -0.197
## RefNO     0.648  0.246  0.131 -0.709
## NSFC      0.414 -0.251 -0.865  0.132
##
##          Comp.1 Comp.2 Comp.3 Comp.4
## SS loadings  1.00  1.00  1.00  1.00
## Proportion Var  0.25  0.25  0.25  0.25
## Cumulative Var  0.25  0.50  0.75  1.00
```

前两个主成分:

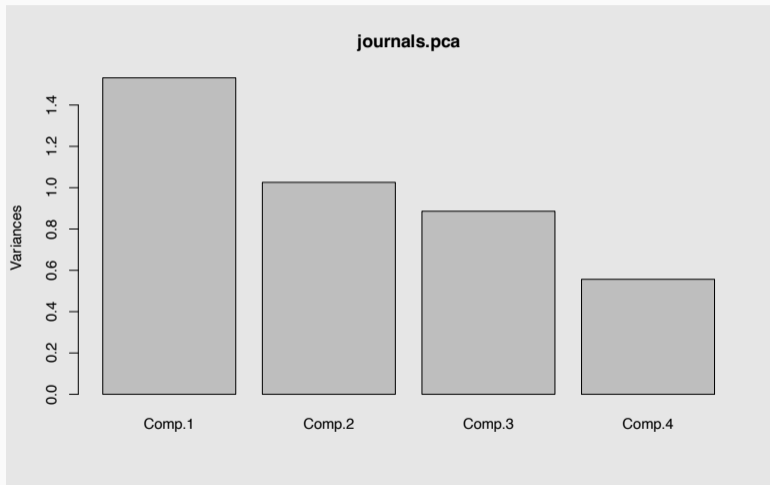
$$Y_1 = 0.634 * Citation + 0.648 * RefNo + 0.414 * NSFC$$

$$Y_2 = -0.936 * PaperNo + 0.246 * RefNo - 0.251 * NSFC$$

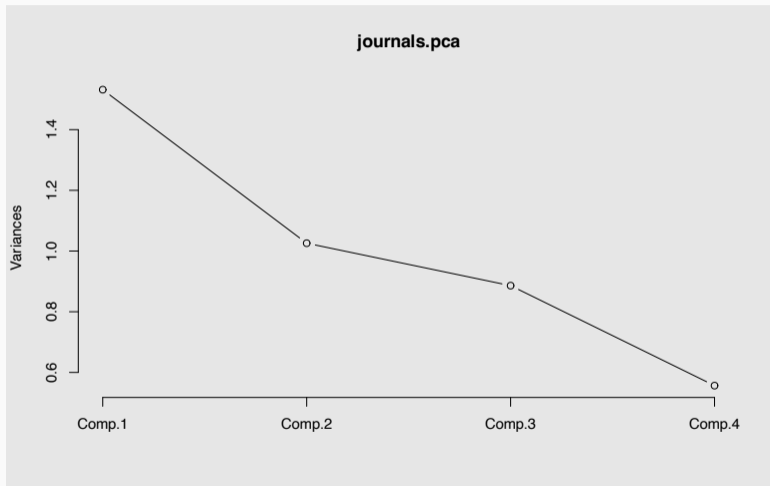
前两个主成分解释了多少信息？

```
## Importance of components:  
##                Comp.1    Comp.2    Comp.3    Comp.4  
## Standard deviation    1.2376191 1.0128027 0.9412763 0.7460085  
## Proportion of Variance 0.3829253 0.2564423 0.2215003 0.1391322  
## Cumulative Proportion 0.3829253 0.6393676 0.8608678 1.0000000
```

碎石图：条形图



碎石图：折线图



主成分得分 (scores)

##	Comp.1	Comp.2	Comp.3	Comp.4
## 管理世界	0.66652842	-0.09044110	0.89506349	-0.33498953
## 系统理实	2.63911891	-2.10015964	-0.99424599	-0.10131949
## 系工学报	0.25449784	0.85696057	-0.85180952	0.28562855
## 中国软科	1.93936103	-0.99528927	1.47169792	-2.66842290
## 数量经济	0.23870367	-0.95372324	0.54386527	0.58036927
## 中国管科	-0.96426677	1.25396608	-0.36413074	-0.13038629
## 管理工程	-0.22495473	1.22511228	-0.61375999	-0.24190476
## 企业管理	-1.41905599	-1.57657886	0.43351126	-0.02271772
## 运筹学报	-0.94468388	1.20595510	-0.87475479	0.66398914
## 经济理管	0.21410146	1.02476840	0.24094165	-1.47497153
## 管理现代	-1.27895150	-0.01450266	0.03939508	0.31321626
## 中国工经	0.23218572	-0.57833643	0.97492533	0.62785225
## 金融研究	-0.81275449	-0.82304264	0.53731028	0.18554668
## 经济科学	-0.81233253	0.77792029	-0.10640871	-0.21336411
## 科学学研	0.41931154	1.04656538	-0.15223777	-0.36049748
## 科研管理	0.53040476	0.88986910	0.05883912	-0.01311399
## 宏观经济	-1.34581534	-1.79823718	0.53419476	-0.01402184
## 会计研究	-0.47904284	0.35704298	0.28179363	-0.29612825

主成分与原始变量的相关系数

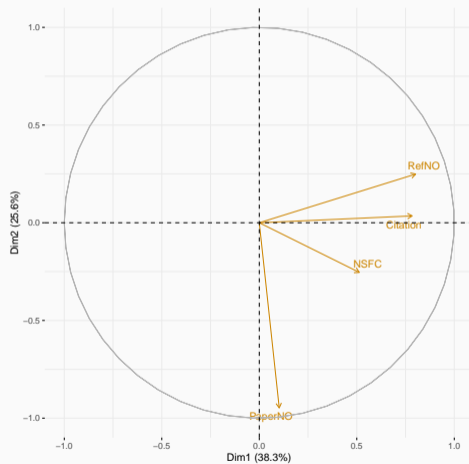
1 第一主成分与原始变量的相关系数

##	Citation	PaperNO	RefNO	NSFC
##	0.784	0.102	0.802	0.513

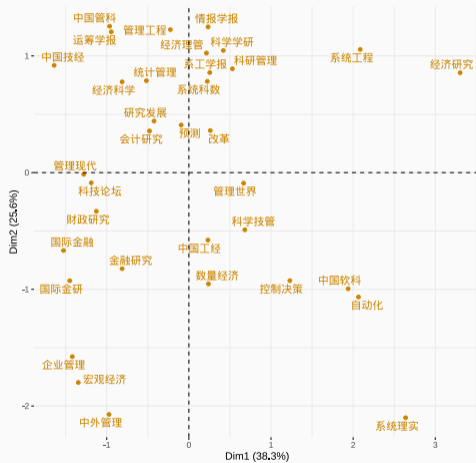
2 第二主成分与原始变量的相关系数

##	Citation	PaperNO	RefNO	NSFC
##	0.035	-0.948	0.249	-0.254

PCA 变量图（相关圆图或者载荷图）



PCA 得分散点图（高维数据的可视化）或主平面图



案例二：污染数据分析

■ 30 个省份的污染数据，含 17 个污染指标

```
## [1] "二氧化硫排放量"    "氮氧化物排放量"    "烟（粉）尘排放量" "PM2.5"  
## [5] "焦炭产量"           "烧碱产量"           "水泥产量"           "平板玻璃产量"  
## [9] "钢材产量"           "发电量"             "煤炭消费量"         "原油消费量"  
## [13] "汽车保有量"         "建成区面积"         "汽油消费量"         "柴油消费量"  
## [17] "燃料油消费量"
```

■ 将二氧化硫排放量、氮氧化物排放量和烟（粉）尘排放量三个变量降维到第一主成分：大气污染排放强度

■ 将除 PM2.5 之外的剩余 13 个变量降维到第一主成分：污染产能综合水平

(一) 大气污染排放强度：第一主成分解释 90.7%

Importance of components:

##	Comp.1	Comp.2	Comp.3
## Standard deviation	1.6496552	0.41995896	0.31980020
## Proportion of Variance	0.9071208	0.05878851	0.03409072
## Cumulative Proportion	0.9071208	0.96590928	1.00000000

##

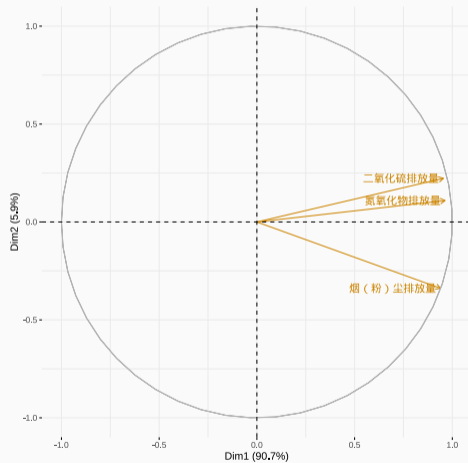
Loadings:

##	Comp.1	Comp.2	Comp.3
## 二氧化硫排放量	0.579	0.531	0.619
## 氮氧化物排放量	0.584	0.261	-0.769
## 烟（粉）尘排放量	0.570	-0.806	0.159

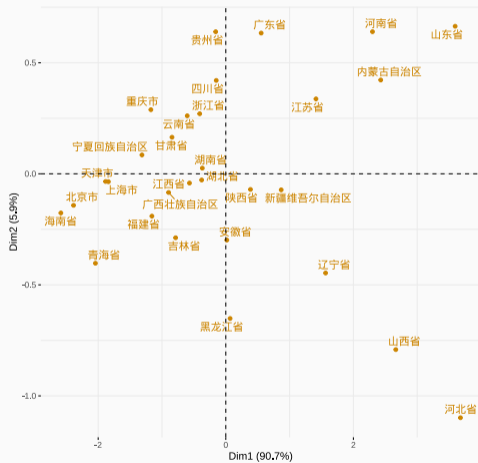
##

##	Comp.1	Comp.2	Comp.3
## SS loadings	1.000	1.000	1.000

载荷图



主平面图



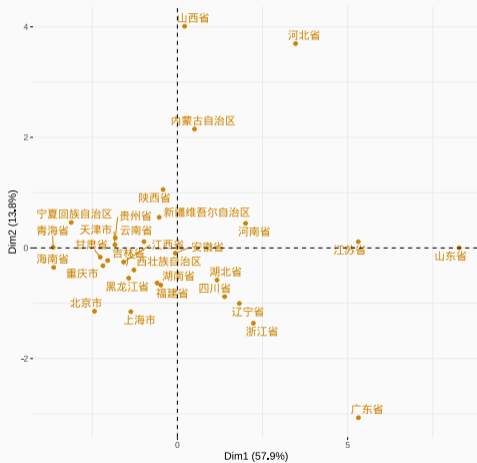
(二) 污染产能综合水平：第一主成分解释 57.9%

Importance of components:

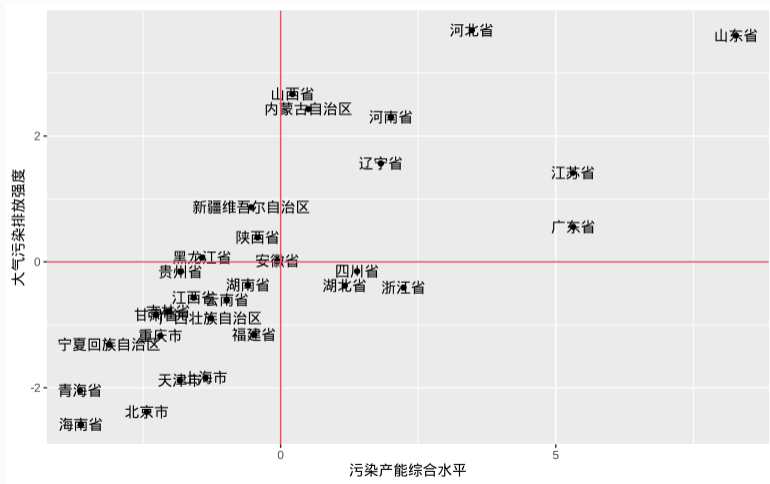
##	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
## Standard deviation	2.7438480	1.3415918	1.13333586	0.97628709	0.70084772
## Proportion of Variance	0.5791309	0.1384514	0.09880386	0.07331819	0.03778366
## Cumulative Proportion	0.5791309	0.7175823	0.81638620	0.88970439	0.92748805
##	Comp.6	Comp.7	Comp.8	Comp.9	Comp.10
## Standard deviation	0.50188127	0.45657340	0.41388159	0.343561200	0.25658558
## Proportion of Variance	0.01937575	0.01603533	0.01317677	0.009079561	0.00506432
## Cumulative Proportion	0.94686381	0.96289913	0.97607590	0.985155463	0.99021978
##	Comp.11	Comp.12	Comp.13		
## Standard deviation	0.247117886	0.199800715	0.161725850		
## Proportion of Variance	0.004697481	0.003070794	0.002011942		
## Cumulative Proportion	0.994917263	0.997988058	1.000000000		

cairo_pdf

主平面图

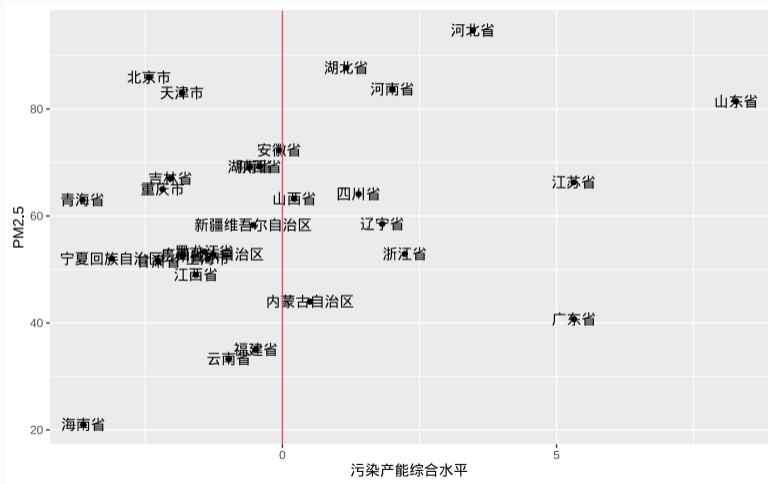


(三) 30 个地区污染产能综合水平和大气污染排放强度的综合排序



- 1 山东、河北：大气污染排放强度和污染产能综合水平高
- 2 江苏、广东：污染产能综合水平高
- 3 北京市、上海市、天津市、重庆市：大气污染排放强度荷污染产能综合水平低

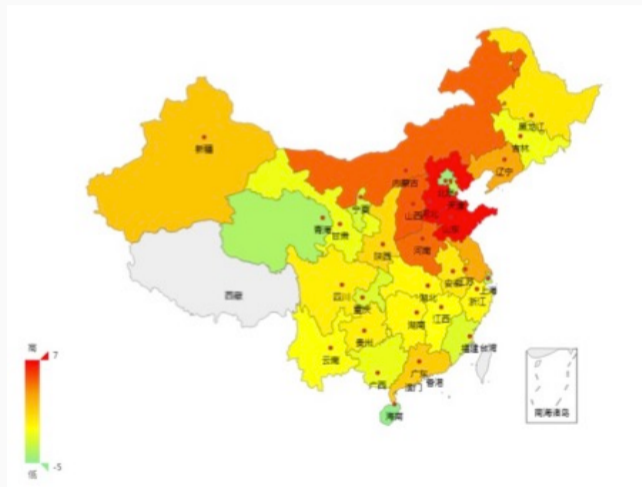
30 个地区污染产能综合水平和 PM2.5 平均浓度的散点图



重点关注

- 1 北京市、天津市：污染产能综合水平低，但是 PM2.5 高
- 2 河北省、山东省：污染产能综合水平和 PM2.5 高
- 3 江苏省、广东省：广东省相较于江苏省来说，类似的污染产能综合水平，但是 PM2.5 低

污染排放强度的地理分布图



- 1 在污染排放染色地图中，北京、天津的颜色是绿色的，基本处于全国最好水平
- 2 除河北之外，其他污染排放比较严重的地区，也都主要是北京周边的省市
- 3 分析结论：北京市的空气污染受到很强的空间传播影响

降至一维（综合指数）

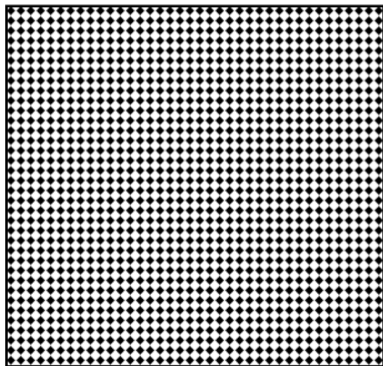
例：拟评估英国各地区农业生产水平。48 个地区，10 种农作物：小麦 (x_1)、大麦 (x_2)、燕麦 (x_3)、土豆 (x_4)、菜豆 (x_5)、马铃薯 (x_6)、萝卜 (x_7)、饲料甜菜 (x_8)、临时牧场干草 (x_9)、永久牧场干草 (x_{10})。

$$PC1 = 0.39x_1 + 0.37x_2 + 0.39x_3 + 0.27x_4 + 0.22x_5 \\ + 0.30x_6 + 0.32x_7 + 0.26x_8 + 0.24x_9 + 0.34x_{10}$$

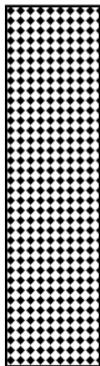
精度为 47.6%。

PCA 与聚类

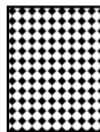
■ 海量数据的简约处理：降维 + 聚类



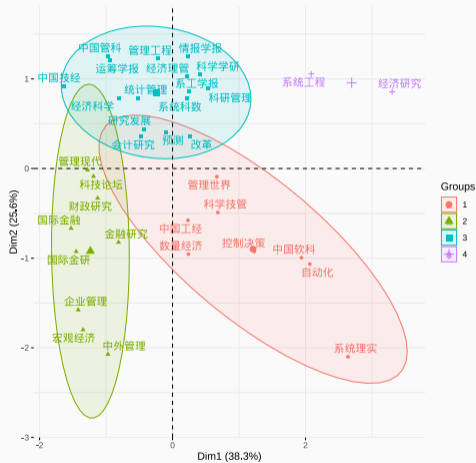
PCA
→



聚类
→



管理期刊分析



四类的特征

- 1 第一类：第二主成分得分高，载文量低（注意到第二主成分与载文量成负相关）；第一主成分得分居于平均水平
- 2 第二类：第一主成分得分（期刊综合实力或者期刊质量）高
- 3 第三类：第一主成分得分和第二主成分得分偏低，即期刊综合质量偏低、发文量高
- 4 第四类：第一主成分得分偏低，第二主成分居于平均水平

Outline

- 1 为什么主成分分析?
- 2 PCA 原理
- 3 PCA 算法
- 4 案例分析
- 5 主成分分析在 R 中的实现**
- 6 作业

PCA

```
journals.pca <- princomp(journals, cor = TRUE)
names(journals.pca)
```

```
## [1] "sdev"      "loadings" "center"   "scale"    "n.obs"    "scores"   "call"
```

```
summary(journals.pca)
```

```
## Importance of components:
```

```
##              Comp.1   Comp.2   Comp.3   Comp.4
## Standard deviation  1.2376191 1.0128027 0.9412763 0.7460085
## Proportion of Variance 0.3829253 0.2564423 0.2215003 0.1391322
## Cumulative Proportion 0.3829253 0.6393676 0.8608678 1.0000000
```

PCA 载荷

```
journals.pca$loadings
```

```
##  
## Loadings:  
##          Comp.1 Comp.2 Comp.3 Comp.4  
## Citation  0.634          0.395  0.664  
## PaperNO   -0.936  0.281 -0.197  
## RefNO     0.648  0.246  0.131 -0.709  
## NSFC      0.414 -0.251 -0.865  0.132  
##  
##          Comp.1 Comp.2 Comp.3 Comp.4  
## SS loadings  1.00  1.00  1.00  1.00  
## Proportion Var  0.25  0.25  0.25  0.25  
## Cumulative Var  0.25  0.50  0.75  1.00
```

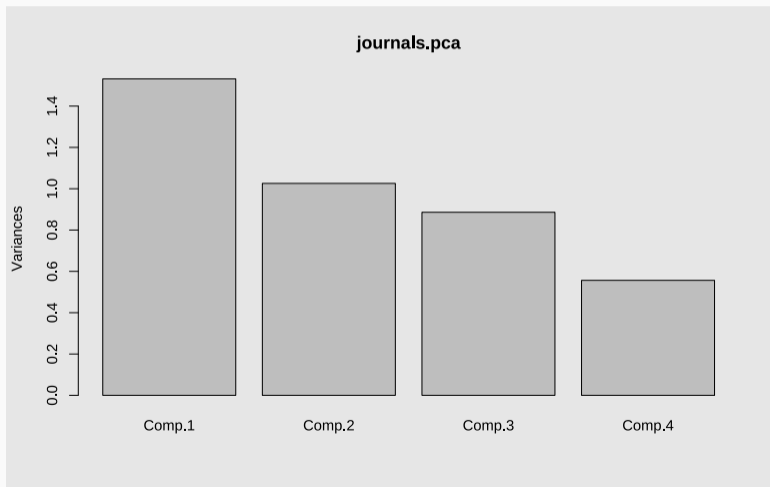
PCA 得分

```
journals.pca$scores
```

##	Comp.1	Comp.2	Comp.3	Comp.4
## 管理世界	0.66652842	-0.09044110	0.89506349	-0.33498953
## 系统理实	2.63911891	-2.10015964	-0.99424599	-0.10131949
## 系工学报	0.25449784	0.85696057	-0.85180952	0.28562855
## 中国软科	1.93936103	-0.99528927	1.47169792	-2.66842290
## 数量经济	0.23870367	-0.95372324	0.54386527	0.58036927
## 中国管科	-0.96426677	1.25396608	-0.36413074	-0.13038629
## 管理工程	-0.22495473	1.22511228	-0.61375999	-0.24190476
## 企业管理	-1.41905599	-1.57657886	0.43351126	-0.02271772
## 运筹学报	-0.94468388	1.20595510	-0.87475479	0.66398914
## 经济理管	0.21410146	1.02476840	0.24094165	-1.47497153
## 管理现代	-1.27895150	-0.01450266	0.03939508	0.31321626
## 中国工经	0.23218572	-0.57833643	0.97492533	0.62785225
## 金融研究	-0.81275449	-0.82304264	0.53731028	0.18554668
## 经济科学	-0.81233253	0.77792029	-0.10640871	-0.21336411
## 科学学研	0.41931154	1.04656538	-0.15223777	-0.36049748
## 科研管理	0.53040476	0.88986910	0.05883912	-0.01311399

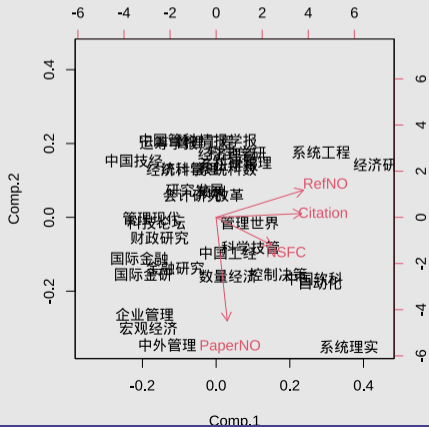
PCA 碎石图

```
plot(journals.pca)
```



主平面图 (Biplot)

```
par(family = 'SimHei')  
biplot(journals.pca)
```



R 中有关函数（小结）

- `princomp()` 函数进行主成分分析，用 `cor=` 选择是否进行标准化，`scores=` 选择是否输出主成分得分
- `summary()` 函数显示主成分分析的概况，用 `loadings=TRUE` 要求显示载荷，即主成分线性组合系数（特征向量）
- `loadings()` 单独输出载荷
- `predict()` 用来对训练数据或者新数据计算主成分得分
- `screeplot()` 显示从大到小排列的特征值的条形图或者线条，便于在特征值骤减的位置截断从而确定主成分个数
- 取前两个主成分时，用 `biplot()` 函数做主平面图（双重散点图）

其他相关学习材料

- 1 基于 factoextra 包的 PCA 可视化, 请参考

<https://www.rdocumentation.org/packages/factoextra/versions/1.0.7>.

- 2 关于 PCA 的交互可视化: <https://setosa.io/ev/principal-component-analysis/>.

主成分的命名

- 主成分 $\mathbf{y}_1, \dots, \mathbf{y}_p$ 是原变量 X_1, \dots, X_p 的线性组合，原变量有明确的物理含义。主成分 $\mathbf{y}_1, \dots, \mathbf{y}_p$ 的物理含义呢？（如何命名）
- 专业知识 + 数学手段：注意到第一主成分 \mathbf{y}_1 与原变量 X_j 之间的样本相关系数等于 $\sqrt{\lambda_1}q_{1j}$ ，即与载荷向量 q_1 成正比，因此可以通过观察 q_1 来确定 \mathbf{y}_1 的含义，其他主成分类似。

- 1 掌握利用 R 软件进行主成分分析（步骤 + 可视化）
- 2 评估需要多少个主成分（精度）
- 3 解释主成分得分，并描述一个得分高或低的主题（计算主成分与原始变量的样本相关系数）
- 4 确定何时应基于方差-协方差矩阵或相关矩阵进行主成分分析，并在进一步分析中比较主成分得分

Outline

- 1 为什么主成分分析?
- 2 PCA 原理
- 3 PCA 算法
- 4 案例分析
- 5 主成分分析在 R 中的实现
- 6 作业

作业

- 1 对“污染数据”做 PCA 分析和聚类分析（数据见污染数据.xlsx）。
- 2 对房地产数据做 PCA 分析和聚类分析（数据见 houses.csv）。
- 3 对《地方评级年鉴》数据进行 PCA 和聚类分析（数据见 places.csv）

要求：三选二进行分析，形成分析报告。在报告中注意数据结果的解释和可视化。