



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 算法
- 3 PCA 与聚类
- 4 主成分分析在 R 中的实现
- 5 作业

Outline

- 1 为什么主成分分析?
- 2 PCA 算法
- 3 PCA 与聚类
- 4 主成分分析在 R 中的实现
- 5 作业

主成分分析的目的

- 主成分分析，简称 PCA (Principal Component Analysis)
- 高维数据降维
- 易于展示

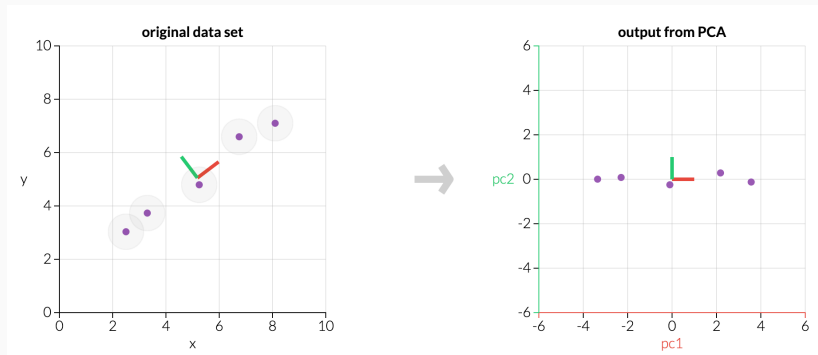
PCA 的主要功能

问题：如何展示一个 10 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。

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。



来源: <https://setosa.io/ev/principal-component-analysis/>.

Outline

- 1 为什么主成分分析?
- 2 PCA 算法
- 3 PCA 与聚类
- 4 主成分分析在 R 中的实现
- 5 作业

PCA 算法

- 降维之后的每个维度都是原数据维度的线性组合:

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

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

- 降维之后

- 均值为 0: $\bar{\mathbf{y}}_h = 0$.
- 维度之间互不相关: $\text{cov}(\mathbf{y}_i, \mathbf{y}_j) = 0$.

PCA 算法

- 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 $\Sigma = \frac{1}{n} Z^T Z$ 。
- 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：

$$\mathbf{y}_h = \sum_{j=1}^p q_{hj} \mathbf{z}_j = Z \mathbf{q}_h.$$

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 $\lambda_1 = \text{var}(\mathbf{y}_1)$.
- $\lambda_h = \text{var}(\mathbf{y}_h)$.
- $\bar{\mathbf{y}}_h = \mathbf{0}$.
- $\text{cov}(\mathbf{y}_i, \mathbf{y}_j) = 0$.

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

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

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）

例：管理期刊遴选研究

##	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
## 管理工程	8	44	40	13
## 企业管理	1	252	0	0
## 运筹学报	9	22	9	14
## 经济理管	3	97	73	0
## 管理现代	6	130	5	1
## 中国工经	42	194	25	1
## 金融研究	15	202	12	0
## 经济科学	4	82	28	2
## 科学学研	20	72	54	11
## 科研管理	29	83	49	10
## 宏观经济	3	270	0	0

主成分和原变量之间的关系？

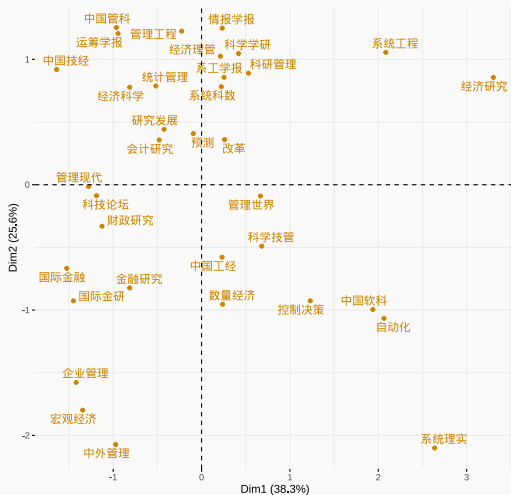
载荷 (loadings)

```
##
## 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
```

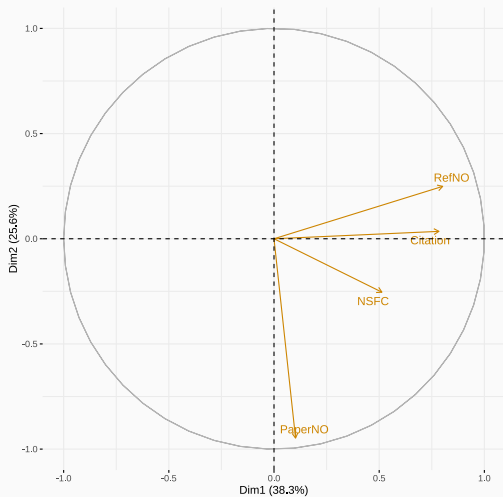

主成分得分 (scores)

##	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
## 预测	-0.09370184	0.40747852	0.04079867	-0.64988412

PCA 得分散点图 (高维数据的可视化)

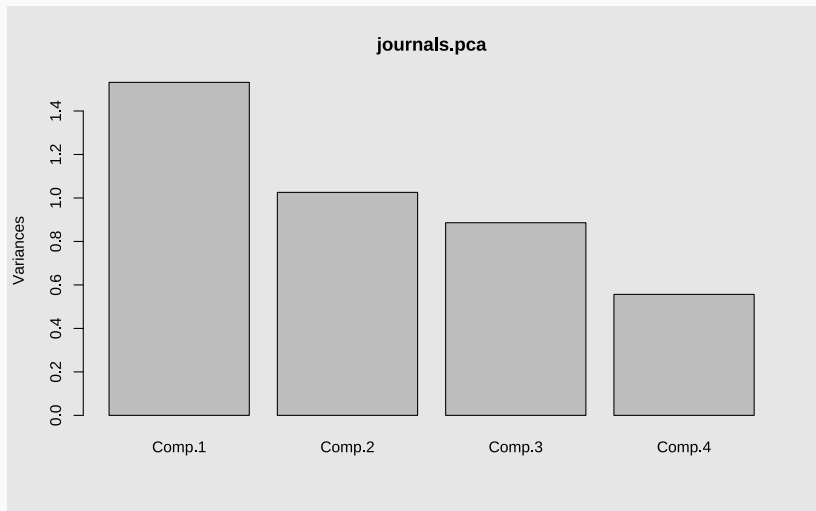


PCA 变量图 (相关圆图)



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
```



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})。

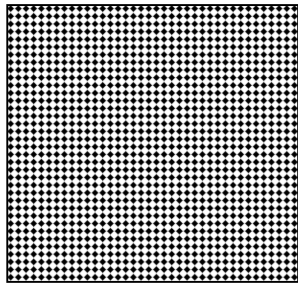
$$PC1 = 0.39x_1 + 0.37x_2 + 0.39x_3 + 0.27x_4 + 0.22x_5 \\ + 0.30x_6 + 0.32x_7 + 0.26x_8 + 0.24x_9 + 0.34x_{10}$$

精度为 47.6%。

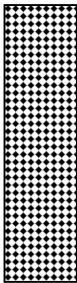
Outline

- 1 为什么主成分分析?
- 2 PCA 算法
- 3 PCA 与聚类**
- 4 主成分分析在 R 中的实现
- 5 作业

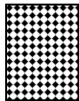
海量数据的简约处理



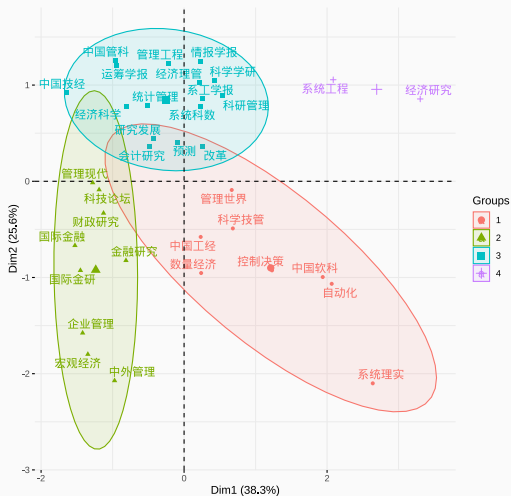
PCA
→



聚类
→



PCA + 聚类



Outline

- 1 为什么主成分分析?
- 2 PCA 算法
- 3 PCA 与聚类
- 4 主成分分析在 R 中的实现**
- 5 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
```

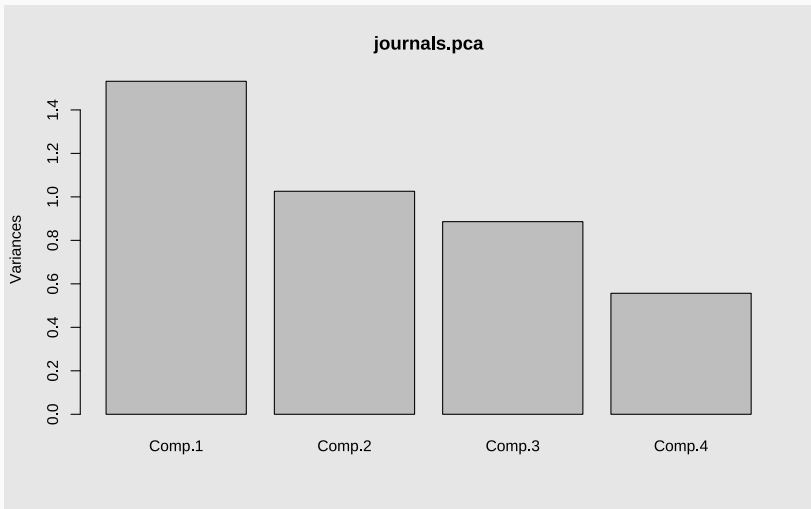
PCA 得分

```
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
##	宏观经济	-1.34581534	-1.79823718	0.53419476	-0.01402184

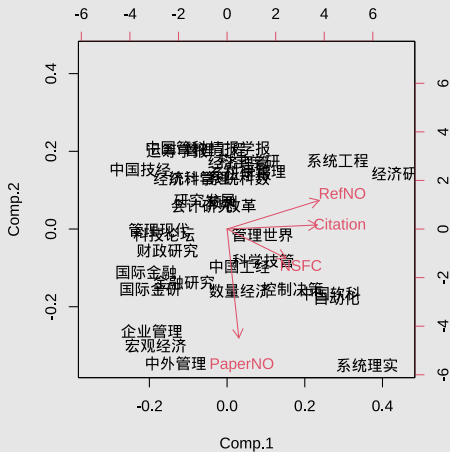
PCA 碎石图

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



Biplot

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



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/>.

Outline

- 1 为什么主成分分析?
- 2 PCA 算法
- 3 PCA 与聚类
- 4 主成分分析在 R 中的实现
- 5 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）。