应用统计II 第5讲 主成分分析 (Principal component analysis)

Instructor: 郝壮

haozhuang@buaa.edu.cn
School of Economics and Management
Beihang University

May 24, 2022

主成分分析(Principal Component Analysis)

- 工作目的：高维数据的**降维**
- 基本原理：平移变换+旋转变换
- 计算方法
- 分析技巧：分析精度, 主成分命名等

PCA的主要功能

在信息损失最小的前提下, 对高维空间进行降维处理.

数据： n样本点(观察)$\times$ p 维变量
原始数据

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

经过PCA变换后($m << p$):

$$\begin{bmatrix} y_{11} & \cdots & y_{1m} \\ y_{21} & \cdots & y_{2m} \\ \vdots & \cdots & \vdots \\ y_{n1} & \cdots & y_{nm} \end{bmatrix}_{n\times m}$$

案例：在一个低维空间辩识系统要比在高维空间容易得多

- 例1. Stone [美] 1947年关于国民经济的研究. 利用美国1929-1939年各年数据, 得到17个反映国民收入与支出的变量.

  原变量17个, 经过PCA变换后(PCA精度97.4%), 降维成3个变量

  - F_1(总收入 I);
  - F_2 (总收入变化率 ΔI);
  - F_3 (经济发展或衰退趋势 t).
- 例2. 用一个企业在所有领域/场景中的低频率发生的违约事件(如城市信用, 税务, 债务, 服务)评估其信用风险-高微稀疏矩阵降维

(精度: 可解释的变异)

怎样对数据进行降维处理? (平移+旋转)

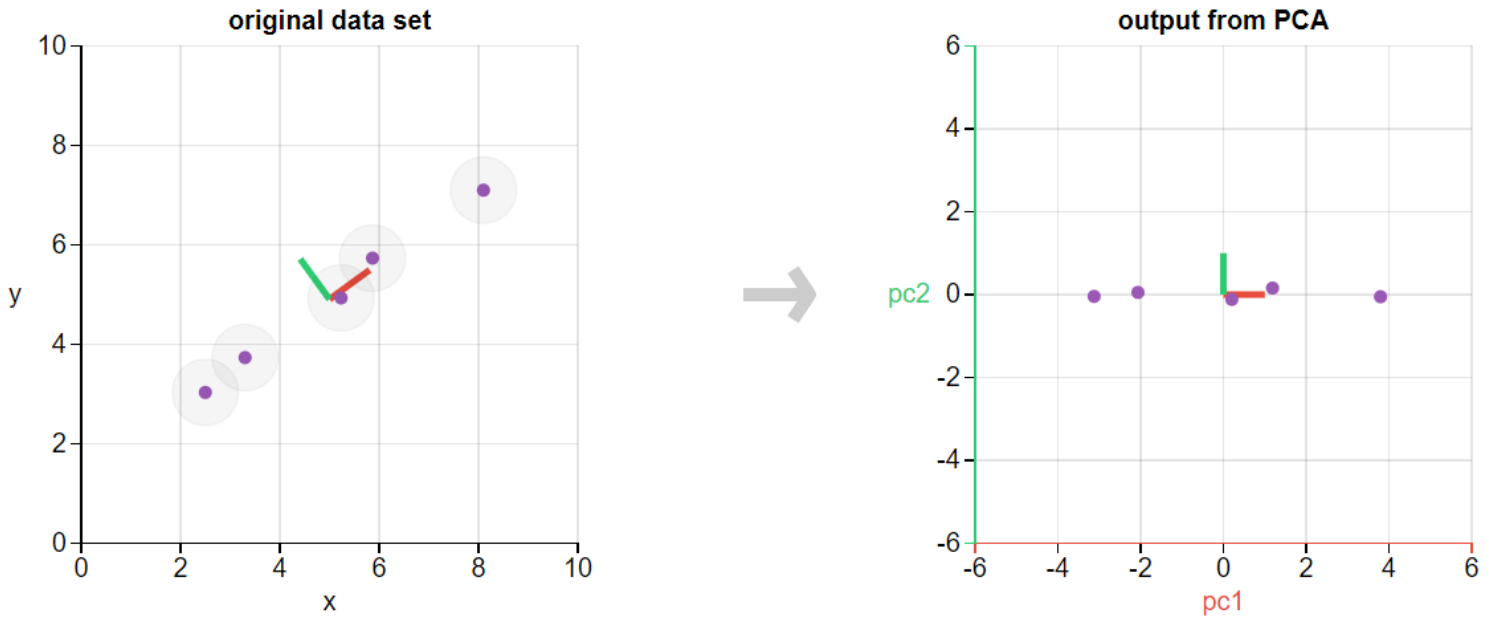
$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}_{n\times p} \Rightarrow \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots &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

1. 通过平移+旋转将原始数据X阵变成Y阵, 所有Y的列向量互相独立, 且 $Var(y_1) \geq Var(y_2) \geq \ldots$.
2. 降维之后的每个维度(主成分)都是原数据维度的线性组合： $\mathbf{y}_h = \sum_{j=1}^{p} \alpha_{hj}\mathbf{x}_j, (h = 1, \cdots, m, m << p)$,
3. 提取前m维主成分 (保留前m列).

数学上主要工作：依次寻找使y方差最大的载荷向量 α

怎样对数据进行降维处理? (平移+旋转)

通过平移+旋转省却数据变异(variation)不大方向的信息 (变异体现了信息携带量)



方差变异最大的方向是红色方向, 可以旋转平移使数据在**pc1**轴上的投影的**方差最大**, **pc2与pc1正交, 数据投影方差变小**.
降至m维：使前m维方差依次降低.

降维的两个特殊应用

- 1. 将一个高维变量系统有效的降至2维 (高维不可见空间到直观平面图示)
- 2. 将一个高维变量系统有效的降至1维 (综合指数)

增加决策知识, 提高对数据的洞察能力.

引例：管理期刊遴选研究 (4维降至2维)

- X_1 — 载文量(papers)
- X_2 — 标注“国家自然科学基金项目”(nsfc)
- X_3 —被引次数(cited_by)
- X_4 — 引证期刊数(references)

- 一张35×4 维的数据表
- 你能立刻看见这些期刊有什么特点吗?

	journal	cited_by	papers	refere~s	nsfc
1.	管理世界	33	175	54	1
2.	系统理实	47	285	64	59
3.	系工学报	21	60	35	24
4.	中国软科	20	293	117	0
5.	数量经济	37	212	22	10
6.	中国管科	1	41	26	3
7.	管理工程	8	44	40	13
8.	企业管理	1	252	0	0
9.	运筹学报	9	22	9	14
10.	经济理管	3	97	73	0
11.	管理现代	6	130	5	1
12.	中国工经	42	194	25	1
13.	金融研究	15	202	12	0
14.	经济科学	4	82	28	2
15.	科学学研	20	72	54	11

16.	科研管理	29	83	49	10
17.	宏观经济	3	270	0	0
18.	会计研究	11	123	34	0
19.	预测	9	122	47	6
20.	统计管理	8	84	34	4
21.	系统科数	13	54	34	34
22.	系统工程	59	86	75	16
23.	国际金研	2	200	0	0
24.	中外管理	7	280	0	12
25.	情报学报	5	74	70	3
26.	科学技管	39	191	41	8
27.	改革	46	121	26	0
28.	科技论坛	13	132	0	2
29.	自动化	29	156	42	86
30.	中国技经	1	52	0	0
31.	控制决策	17	171	43	60
32.	财政研究	4	162	13	0
33.	经济研究	125	119	65	0
34.	国际金融	0	179	0	0
35.	研究发展	11	105	30	8

降至二维

降维: 将4维原始数据降至2维主成分

$x_1 \quad x_2 \quad x_3 \quad x_4 \quad \Rightarrow \quad y_1 \quad y_2$

主成分中的值称为**载荷** (loadings), 也是标准化后原始数据阵的两个特征向量.

Principal components (eigenvectors)

Variable	Comp1	Comp2
cited_by	0.6337	-0.0347
papers	0.0827	0.9357
nsfc	0.4145	0.2507
references	0.6479	-0.2459

stata command: pca cited_by papers nsfc references, components(2)

主成分命名

主成分Comp1 和 Comp2 有什么现实含义呢?

Principal components (eigenvectors)

Variable	Comp1	Comp2
cited_by	0.6337	-0.0347
papers	0.0827	0.9357
nsfc	0.4145	0.2507
references	0.6479	-0.2459

主成分命名

主成分Comp1 和 Comp2 有什么现实含义呢?

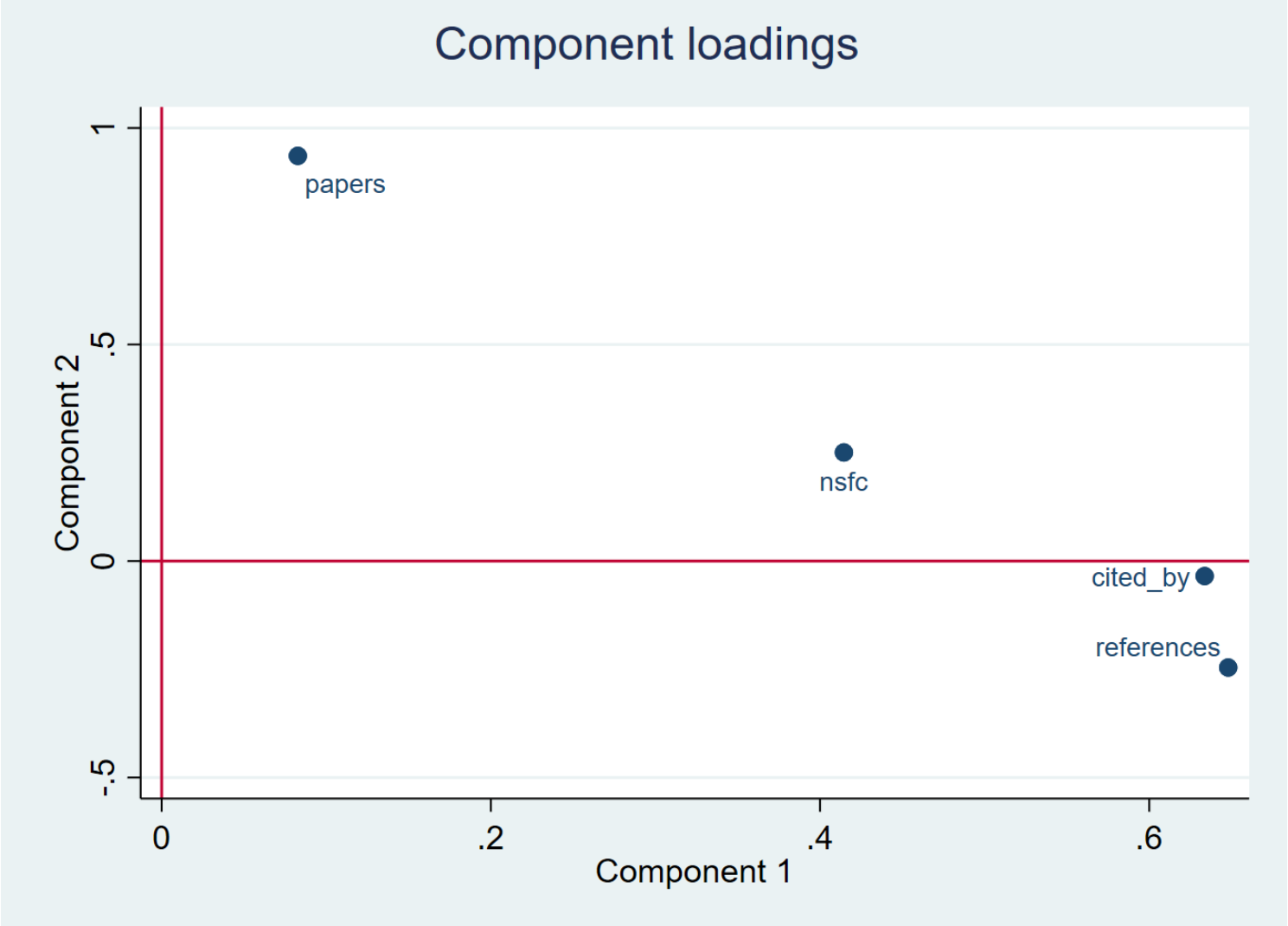
Principal components (eigenvectors)

Variable	Comp1	Comp2
cited_by	0.6337	-0.0347
papers	0.0827	0.9357
nsfc	0.4145	0.2507
references	0.6479	-0.2459

- 根据Comp1和被引次数和引证期刊的强相关, 可命名Comp1为"科学性/学术影响"等
- 根据Comp2和载文量的强相关, 可命名Comp2为"载文量"

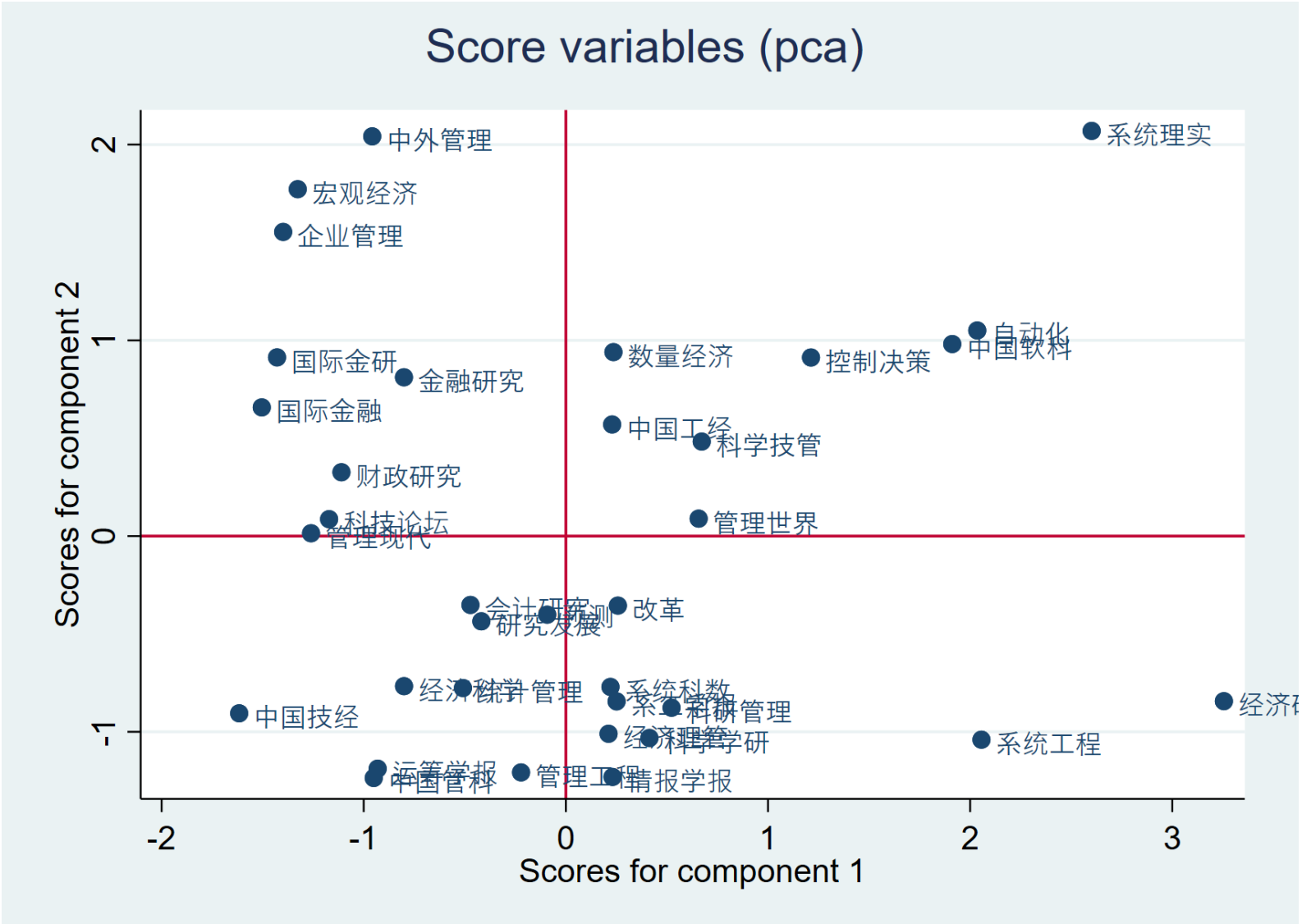
相关圆图 (component plot in rotated space)

第一和第二主成分载荷.



(loadingplot)

PCA得分散点图 (即主成分y_1, y_2取值)



(scoreplot, mlabel(journal) xline(0) yline(0))

解释的总方差

各主成分解释的总方差如下

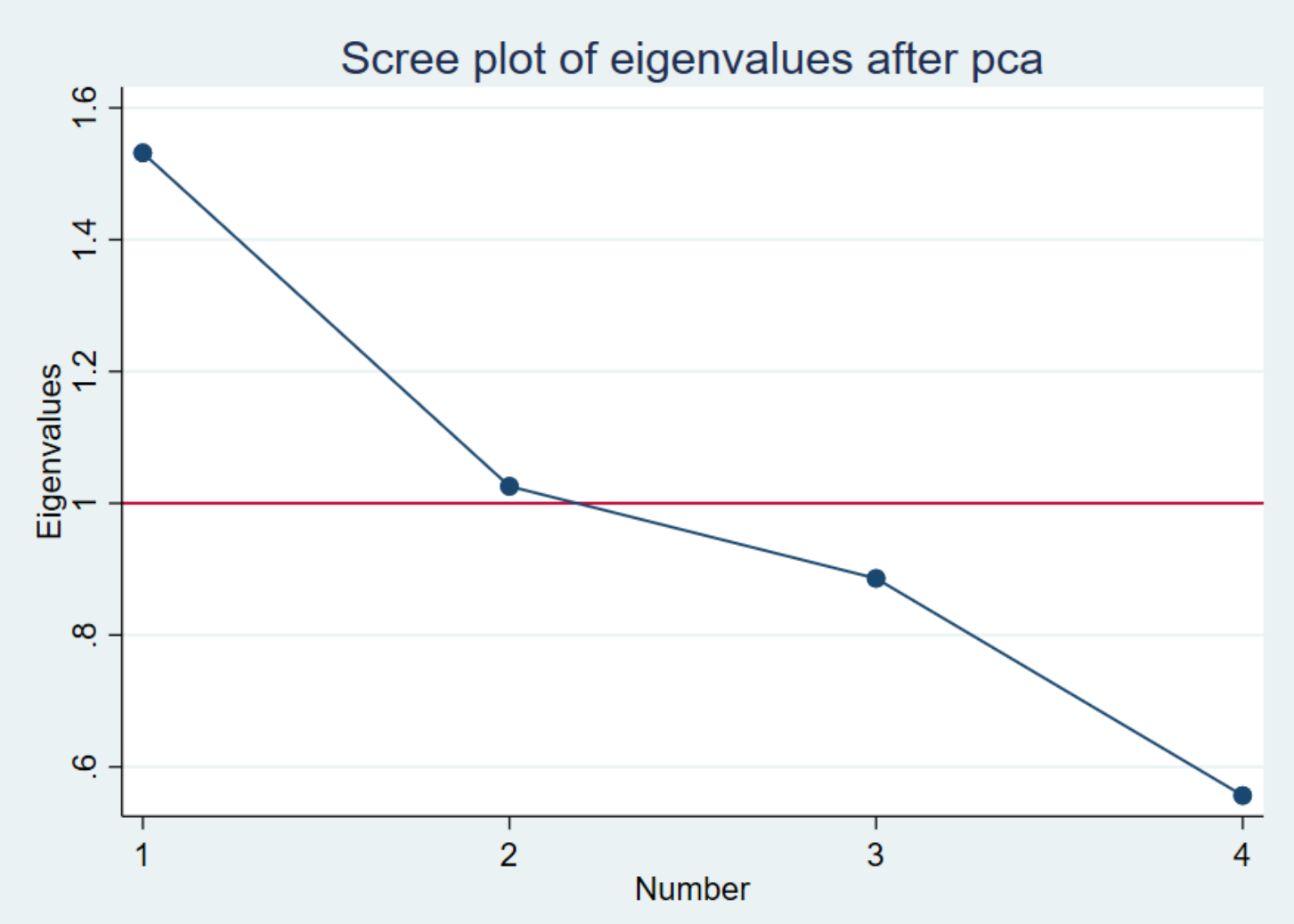
```
. pca cited_by papers nsfc references,  components(2)

Principal components/correlation                  Number of obs    =         35
                                                  Number of comp.  =          2
                                                  Trace            =          4
    Rotation: (unrotated = principal)             Rho              =     0.6394
```

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	1.5317	.505932	0.3829	0.3829
Comp2	1.02577	.139768	0.2564	0.6394
Comp3	.886001	.329472	0.2215	0.8609
Comp4	.556529	.	0.1391	1.0000

所以, 第一和第二主成分已经解释了总方差的63.94%.

碎石图 (scree plot)



引例2: 降至 1 维 (综合指数)

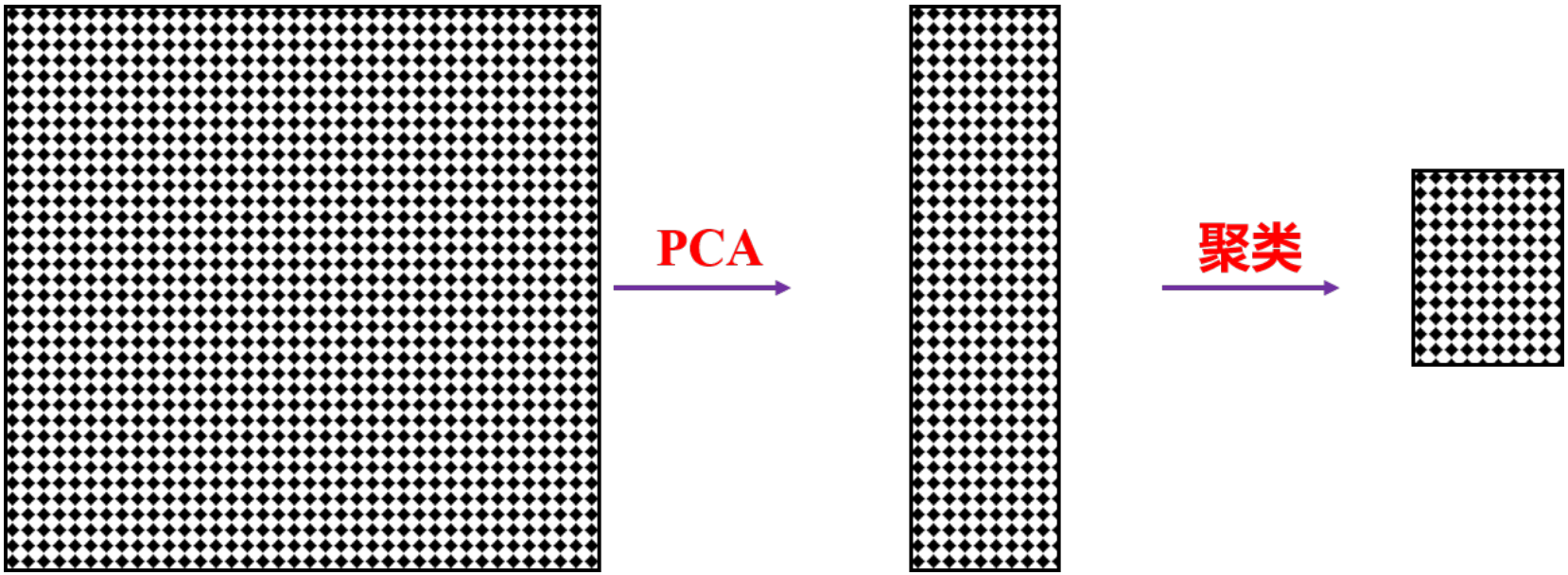
案例： Kendall (1975) 评估英国各地区农业生产水平. 48个地区, 10种农作物：小麦(x_1) , 大麦(x_2) , 燕麦(x_3) , 土豆(x_4) , 菜豆(x_5) , 马钠薯(x_6) , 萝卜(x_7), 甜菜(x_8), 牧场干草A(x_9), 牧场干草B(x_{10})

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

精度为 47.6%.
Kendall (1975) 认为该主成分可解释为生产力水平并对48个地区排序. 除个别地区以外, 所得结果与当时有关生产能力的地理分布一般知识是一致的.
Source: Kendall M. Multivariate Analysis. 1975.

PCA与聚类分析的综合应用

- 海量数据的简约处理： PCA+聚类



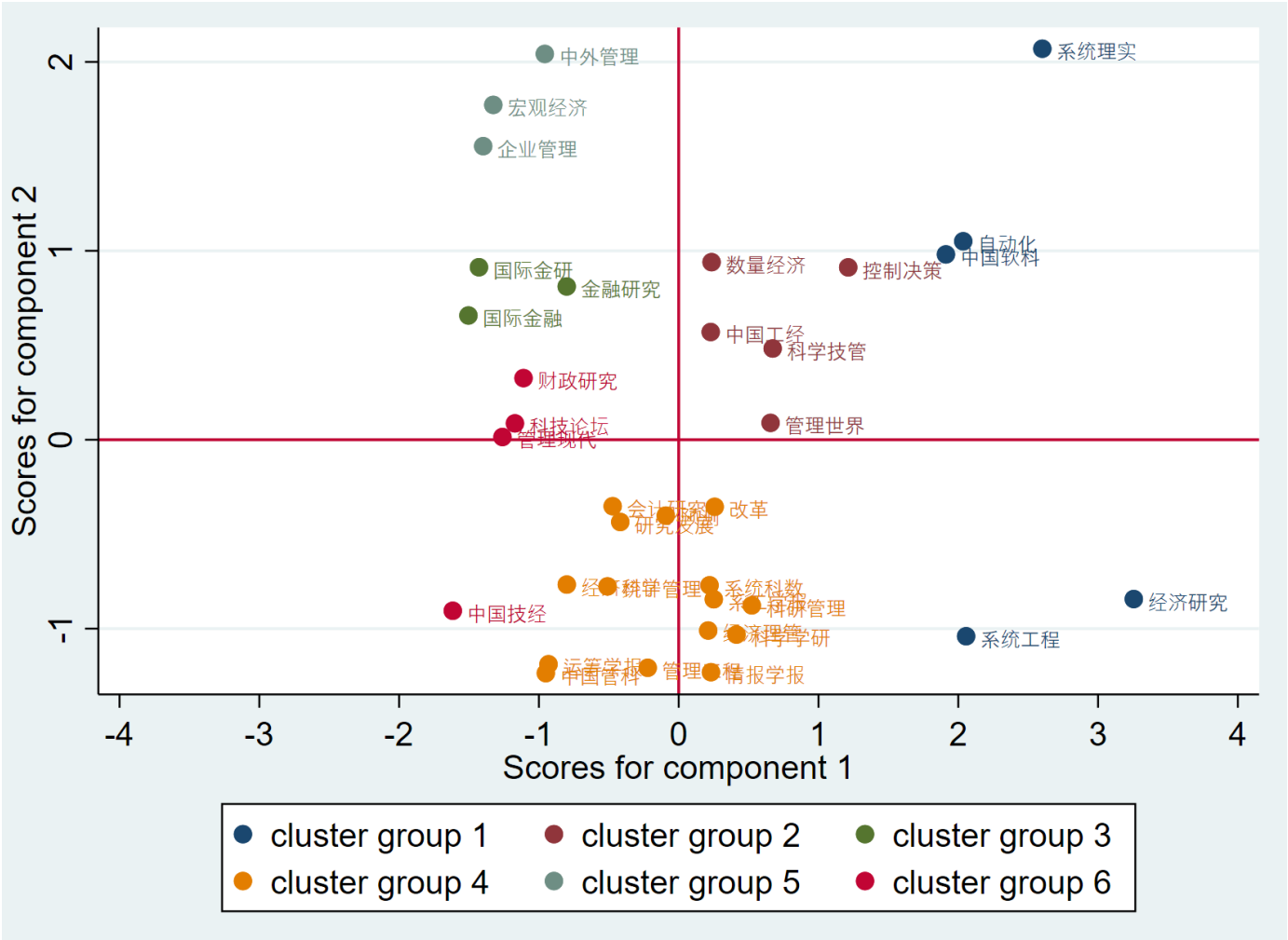
案例： PCA与聚类分析的综合应用

运用PCA和聚类分析, 对35种管理学期刊进行降维, 分类.

	mingchen	bycishu	zaiwenl	yzqikan	nsfc
1	管理世界	33	175	54	1
2	系统理实	47	285	64	59
3	系工学报	21	60	35	24
4	中国软科	20	293	117	0
5	数量经济	37	212	22	10
6	中国管科	1	41	26	3
7	管理工程	8	44	40	13
8	企业管理	1	252	0	0
9	运筹学报	9	22	9	14
10	经济理管	3	97	73	0
11	管理现代	6	130	5	1
12	中国工经	42	194	25	1
13	金融研究	15	202	12	0
14	经济科学	4	82	28	2
15	科学学研	20	72	54	11
16	科研管理	29	83	49	10
17	宏观经济	3	270	0	0
18	会计研究	11	123	34	0
19	预测	9	122	47	6

案例：主成分分析与聚类分析软件应用

1. PCA降维, 提取2维主成分
2. K-means 聚类法将35种期刊聚成6类
3. 将分组后期刊进行二维可视化表达



PCA 算法

1 数据标准化：将 $X_{(n\times p)}$ 的每一列减其均值, 除以标准差. 标准化的矩阵记为$Z_{(n\times p)}$.

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

处理后每个变量都被标准化, 量纲一致, 整个数据的重心与原点重合 (减均值为平移变换, 除以标准差为压缩变换).

2 计算 Z 的方差协方差矩阵, 即 X 各变量间的相关系数矩阵, 记该方差协方差矩阵为

$$\Sigma_{(p\times p)} = \frac{1}{n} Z'Z$$

即p个经过标准化后的变量的方差协方差阵.

思考: $\Sigma_{(p\times p)}$ 有什么性质? 对角线上元素表示什么? 上三角和下三角元素表示什么? 有什么性质?

PCA 算法

3 求 $\Sigma_{(p\times p)}$ 的所有特征值$\lambda_{(1\times 1)}$和特征向量$q_{(p\times 1)}$ (共p个特征值和p个特征向量),

将 Σ 的特征值排序： $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$,

将对应的特征向量$q_1, q_2, \ldots$进行排序, 构成 $Q_{(p\times p)}$, 要求各特征向量是标准正交的, 即:

$$q_i^T q_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

拓展知识

拓展知识: 由于Σ是对称阵, 所以正交的要求不是额外约束. 由于Σ是标准化的, 所以标准也不是额外约束.

拓展知识 (实谱定理 real spectral theorem): 对称阵的特征向量是正交的. Eigenvectors are orthogonal if the matrix is symmetric.

复习: 特征向量, 特征值

对 $n \times n$ 的矩阵$\mathbf{A}$, 如果

$$\mathbf{A}x = \lambda \mathbf{x}$$

则称

- λ 是 $\mathbf{A}$ 的一个**特征值**(eigenvalue)
- $\mathbf{x}_{(n\times 1)}$ 是$\mathbf{A}$的一个**特征向量**(eigenvector).

也可写成:

$$(\mathbf{A} - \lambda I)x = 0$$

特征值和特征向量的性质

对于$n \times n$ 的矩阵$\mathbf{A}$, 一共有n个特征值和特征向量 (λ_i, x_i) (可能有重复).

可将所有解写成矩阵形式：

$$\mathbf{A}X = X * \mathrm{diag}(L)$$

其中

$$X = (x_1, x_2, \ldots) \quad (X : n \times n)$$
$$L = (\lambda_1, \lambda_2, \ldots) \quad (L : 1 \times n)$$

复习: 特征向量, 特征值计算举例

Stata计算特征向量和特征值 (mata语言): eigensystem().

```
. mata
------------------- mata (type end to exit) -----------
: A = (1, 2 \ 9, 4)
: X=.
: L=.
: eigensystem (A,X,L)
: X
                  1               2
    +-----------------------------+
  1 |  -.316227766   -.554700196  |
  2 |  -.948683298    .832050294  |
    +-----------------------------+
: L
         1     2
    +-----------+
  1 |    7    -2  |
    +-----------+
: end
--------------------------------------------------------------
```

第一个特征值 7, 对应的特征向量是 $(-.316\backslash -.949)$. 第二个特征值 -2, 对应的特征向量是 $(-.555\backslash .832)$.

复习: 特征向量, 特征值

对称阵的特征向量正交, 如下例

```
. mata
------------------- mata (type end to exit) -----------
: A = (1, 2 \ 2, 5)
: X=.
: L=.
: eigensystem (A,X,L)
: X
[Hermitian]
                  1               2
    +-----------------------------+
  1 |  -.382683432                |
  2 |  -.923879533    .382683432  |
    +-----------------------------+
: L
                  1               2
    +-----------------------------+
  1 |  5.82842712   .171572875    |
    +-----------------------------+
: end
--------------------------------------------------------------------------------
. di -.382683432*(-.923879533)+(-.923879533)*.382683432
0
--------------------------------------------------------------------------------
```

复习: 特征向量, 特征值

方差协方差矩阵 (对称阵, 对角线元素为1, 其他元素小于1) 的特征向量标准正交

```
. mata
---------------- mata (type end to exit) -----------
: A = (1, 0.4 \ 0.4, 1)
: X=.
: L=.
: eigensystem (A,X,L)
: X
                  1               2
    +-----------------------------+
  1 |    .707106781   -.707106781  |
  2 |    .707106781    .707106781  |
    +-----------------------------+
: L
          1      2
    +-------------+
  1 |   1.4     .6  |
    +-------------+
: end
------------------------------------------------------
. di  2*.707106781*.707106781
1
```

PCA 算法

4 计算 $Y_{(n\times p)} = Z_{(n\times p)}Q_{(p\times p)}$, 取 $Y_{(n\times p)}$ 的前 m 列, 即 $X_{(n\times p)}$ 的前 m 个主成分得分. $Q_{(p\times p)}$的每一列的特征向量叫作载荷向量. 第 h 个主成分得分即:

$$\mathbf{y}_{h(n\times 1)} = \sum_{j=1}^{p} q_{hj(scalar)}\mathbf{z}_{j(n\times 1)} = \mathbf{Z}_{(n\times p)}\mathbf{q}_{h(p\times 1)}$$

拓展：为什么上述算法成立？

一、求第一主成分y_1和第一主轴q_1

- 我们的目标是想让 $\mathbf{y}_1 = \boldsymbol{Z}_{(n\times p)}\mathbf{q}_{1(p\times 1)}$ 携带最多的信息，也就是 $\mathrm{Var}\left(\mathbf{y}_1\right)$ 取到最大值.

$$\mathrm{Var}\left(\boldsymbol{y}_1\right) = \frac{1}{n}\boldsymbol{q}_1^{\mathrm{T}}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}\boldsymbol{q}_1 = \boldsymbol{q}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{q}_1$$

这里, 记 $\boldsymbol{\Sigma} = \frac{1}{n}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}$ 是 $\boldsymbol{Z}$ 的方差协方差矩阵.

拓展：为什么上述算法成立？

一、求第一主成分y_1和第一主轴q_1

把上面的问题写成数学表达式, 即求优化问题：

$$\max_{||q_1||=1} q_1^\top \Sigma q_1$$

采用拉格朗日(Lagrange)算法求解, 记 λ_1 是拉格朗日系数, 令

$$L = \boldsymbol{q}_1^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{q}_1 - \lambda_1 \left(\boldsymbol{q}_1^{\mathrm{T}} \boldsymbol{q}_1 - 1 \right)$$

拓展：为什么上述算法成立？

对 L 分别求关于 q_1 和 λ_1 的偏导, 并令其为零, 有

$$\begin{aligned}\frac{\partial L}{\partial \boldsymbol{q}_1} &= 2\boldsymbol{\Sigma}\boldsymbol{q}_1 - 2\lambda_1\boldsymbol{q}_1 = 0\\ \frac{\partial L}{\partial \lambda_1} &= -\left(\boldsymbol{q}_1^{\mathrm{T}}\boldsymbol{q}_1 - 1\right) = 0\end{aligned}$$

得

$$\boldsymbol{\Sigma}\boldsymbol{q}_1 = \lambda_1\boldsymbol{q}_1$$

由此可知, $\boldsymbol{q}_1$ 是 $\boldsymbol{\Sigma}$ 的特征向量, 它的模长等于 1, 它所对应的特征值是 λ_1.

拓展：为什么上述算法成立？

所以

$$\mathrm{Var}\left(\boldsymbol{y}_1\right)=\boldsymbol{q}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{q}_1=\boldsymbol{q}_1^{\mathrm{T}}\left(\lambda_1\boldsymbol{q}_1\right)=\lambda_1\boldsymbol{q}_1^{\mathrm{T}}\boldsymbol{q}_1=\lambda_1$$

即 $\boldsymbol{q}_1$ 所对应的特征值 λ_1 应该取到最大值.

换句话说, 所要求的 $\boldsymbol{q}_1$ 是矩阵 $\boldsymbol{\Sigma}$ 的最大特征值 λ_1 所对应的特征向量, 并且模长等于1.

$\boldsymbol{q}_1$ 称为第 1 主轴, $\boldsymbol{y}_1=X\boldsymbol{q}_1$ 称为第 1 主成分.

拓展：为什么上述算法成立？

二、求第二主成分y_2和第二主轴q_2.

- 我们的目标是想让 $\mathbf{y}_2 = \boldsymbol{Z}_{(n\times p)}\mathbf{q}_{2(p\times 1)}$ 携带第二多的信息, 也就是 $\mathrm{Var}\left(\mathbf{y}_2\right)$ 取到除$\mathrm{Var}\left(\mathbf{y}_1\right)$的最大值.

$$\mathrm{Var}\left(\boldsymbol{y}_2\right) = \frac{1}{n}\boldsymbol{q}_2^{\mathrm{T}}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}\boldsymbol{q}_2 = \boldsymbol{q}_2^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{q}_2$$

- 把上面的问题写成数学表达式, 即求优化问题：

$$\max_{||q_2||=1} q_2^{\top}\Sigma q_2$$

同时, 还需要增加约束条件:

$$\boldsymbol{q}_2^{\mathrm{T}}\boldsymbol{q}_1 = 0$$

- 采用拉格朗日(Lagrange)算法求解, 记 λ_2 是拉格朗日系数, 令

$$L = \boldsymbol{q}_2^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{q}_2 - \lambda_2\left(\boldsymbol{q}_2^{\mathrm{T}}\boldsymbol{q}_2 - 1\right)$$

拓展：为什么上述算法成立？

- 求 L 关于 q_2 与 λ_2 的偏导, 并令之为零, 得到

$$\begin{aligned} \boldsymbol{\Sigma q}_2 &= \lambda_2 \boldsymbol{q}_2 \\ \boldsymbol{q}_2^{\mathrm{T}} \boldsymbol{q}_2 &= 1 \end{aligned}$$

q_2是矩阵 Σ 的特征向量, 它的模长等于1, 它所对应的特征值是 λ_2, 而

$$\lambda_2 = \boldsymbol{q}_2^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{q}_2 = \mathrm{Var}\left(\boldsymbol{y}_2\right)$$

由于要求$\mathrm{Var}\left(\boldsymbol{y}_2\right)$第二大, 所以 λ_2 只能是矩阵 Σ 的第 2 大特征值; 同时, $\boldsymbol{q}_2$ 是对应于 λ_2 的、模长等于 1 的特征向量.

拓展：为什么上述算法成立？

依此类推, 可求得 $\boldsymbol{X}$ 数据表的第 h 主轴 $\boldsymbol{q}_h$, 它是协方差矩阵 Σ 的第 h 个特征值 λ_{h} 对应的、模长等于 1 的特征向量. 而第 h 主成分 $\boldsymbol{y}_h$ 为

$$\boldsymbol{y}_h = \boldsymbol{Z}\boldsymbol{q}_h$$

例题

背景: 采集12个地区的5个社会经济指标：人口数 (x_1), 教育程度 (x_2) , 就业数 (x_3) , 服务业人数 (x_4) , 房价(x_5). 假设所有数据已经过标准化. 通过主成分分析来评价这些地区的社会经济总体状况.

PCA结果： 主成分分析得到的最大特征值为： $\lambda_1 = 2.873,$ 该特征值对应的特征向量如表所示：

特征向量	x_1	x_2	x_3	x_4	x_5
$\mathbf{u}_1$	-0.091	0.392	-0.039	0.299	0.403

问题： 请写出第一主成分关于5个原始变量的函数依达式 (假设原数据已经经过标准化):

例题

背景: 采集12个地区的5个社会经济指标：人口数 (x_1), 教育程度 (x_2) , 就业数 (x_3) , 服务业人数 (x_4) , 房价(x_5). 假设所有数据已经过标准化. 通过主成分分析来评价这些地区的社会经济总体状况.

PCA结果： 主成分分析得到的最大特征值为： $\lambda_1 = 2.873,$ 该特征值对应的特征向量如表所示：

特征向量	x_1	x_2	x_3	x_4	x_5
$\mathbf{u}_1$	-0.091	0.392	-0.039	0.299	0.403

问题： 请写出第一主成分关于5个原始变量的函数依达式(假设原数据已经经过标准化):

$$y_1 = -0.091x_1 + 0.392x_2 - 0.039x_3 + 0.299x_5 + 0.403x_5$$

主成分的统计特征

第 h 主成分　$\mathbf{y}_h = (y_{1h}, y_{2h}, \cdots, y_{nh})' \in \mathbf{R}^n$

(1) y_h 的均值为0

$$\frac{1}{n}\sum_{i=1}^{n} y_{ih} = 0$$

(2) y_h 的方差等于特征值 λ_h

$$\mathrm{Var}\,(\boldsymbol{y}_h) = \frac{1}{n}\sum_{i=1}^{n} (y_{ih} - 0)^2 = \lambda_h$$

(3) y_j 与 y_k 的协方差等于0

$$\mathrm{Cov}\,(\boldsymbol{y}_j, \boldsymbol{y}_k) = 0, \quad \forall j \neq k$$

(证明见教材p315)

主成分y_h与原始数据x_j的相关系数$r(x_j, y_h)$

可以证明 (参考何晓群多元统计分析第四版 pp120 页) 如果

- 数据是标准化的
- x_j 是原变量, $j = 1, 2, \cdots, p$
- y_h 是主成分, $h = 1, 2, \cdots, m$
- λ_h和$\boldsymbol{q}_h$是第h特征值和特征向量

则有

$$r(\boldsymbol{x}_j, \boldsymbol{y}_h) = \sqrt{\lambda_h} q_{hj}$$

式中, q_{hj}是特征向量$\boldsymbol{q}_h$(第h主轴)的第j个分量.

主成分y_h与原始数据x_j的相关系数$r\left(x_j, y_h\right)$

教材错误： pp317 公式 (11-18)下的表达"任意 $\boldsymbol{x}_j$ 与 $\boldsymbol{F}_h$ 的相关系数恰等于 a_{hj} (特征向量) "和公式(11-19)下的表达"组合系数 a_{hj} 等于相关系数 $r\left(\boldsymbol{x}_j, \boldsymbol{F}_h\right)$"均不准确. (注意：公式(11-18)本身并无错误)

更正： 无论标准化前后的原变量和主成分的相关系数均需要用特征向量乘以根号下特征值, 区别在于是否需要除以原变量标准差.

主成分y_h与原始数据x_j的相关系数$r(x_j, y_h)$

注意

- 1. 特征向量本身并不是相关系数, 无论标准化前还是标准化后的数据.
- 2. 在原数据是标准化数据的前提下, “对应特征向量值乘以根号下特征值” 等于原变量和主成分的相关系数.

主成分y_h与原始数据x_j的相关系数$r(x_j, y_h)$

以35本管理学期刊数据为例, 我们用Stata手动计算主成分和原始数据相关系数, 并与"对应特征向量值乘以根号下特征值"做对比.

```
. // 计算各观察在第一和第二主成分上的得分
. predict pc1 pc2, score
. //输出第一主成分和原始变量引用量cited_by相关系数
. corr(pc1 cited_by)

             |      pc1 cited_by
-------------+------------------
         pc1 |   1.0000
    cited_by |   0.7842   1.0000

. // 手动计算根号下特征值乘以特征向量对应值并于相关系数做对比:  相等
. di sqrt(1.5317) * 0.6337
.78427896
```

总结：经过主成分分析

n个样本点, p个变量经过PCA后变为n个样本点, m个变量

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

均值 $g = (\bar{X}_1, \bar{X}_2, \cdots, \bar{X}_p) \rightarrow g = (0, \cdots, 0)$

方差 $s_1^2, s_2^2, \cdots, s_p^2 \rightarrow \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$

$\mathrm{Cov}(x_j, x_k) \neq 0 \quad (j \neq k) \rightarrow \mathrm{Cov}(y_j, y_k) = 0 \quad (j \neq k)$

PCA的辅助分析技术: 怎样选取精度合适的主超平面?

1. m 维主超平面(hyperspace)的精度测量
   - 主成分分析前, $\boldsymbol{Z}_{n\times p}$数据中的全部变异信息:

$$\sum_{j=1}^{p} \mathrm{Var}\left(x_j\right) = \sum_{j=1}^{p} s_j^2 = p$$

   最后等式在原数据已是标准化数据(方差为1)时成立. (见教材p316)
   - 主成分分析后保留的数据变差:

$$\mathrm{Var}\left(\boldsymbol{y}_1\right) = \lambda_1, \mathrm{Var}\left(\boldsymbol{y}_2\right) = \lambda_2, \ldots, \mathrm{Var}\left(\boldsymbol{y}_m\right) = \lambda_m$$
$$\sum_{k=1}^{m} \mathrm{Var}\left(\boldsymbol{y}_h\right) = \sum_{h=1}^{m} \lambda_h$$

形象地看: $[\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_p] \xrightarrow{PCA} [\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_m]$

方差:

$$s_1^2, s_2^2, \cdots, s_p^2 \xrightarrow{PCA} \lambda_1, \lambda_2, \cdots, \lambda_m$$

注意:

$$\sum_{h=1}^{p} \lambda_h = \sum_{h=1}^{p} \mathrm{Var}\left(\boldsymbol{y}_h\right) = \sum_{j=1}^{p} s_j^2$$

所以, 定义 "累计贡献率"=前h主成分变异/总变异

$$Q_m = \frac{\sum_{h=1}^{m} \lambda_h}{\sum_{j=1}^{p} s_j^2} = \frac{1}{p} \sum_{h=1}^{m} \lambda_h$$

最后等式为标准化后方差为1结果

2. 如何选取主成分个数

根据累计贡献率可以确定所要选取的成分的个数.

例.管理期刊评价

```
. pca bycishu zaiwenl yzqikan nsfc

Principal components/correlation                  Number of obs    =         35
                                                  Number of comp.  =          4
                                                  Trace            =          4
    Rotation: (unrotated = principal)             Rho              =     1.0000
```

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	1.5317	.505932	0.3829	0.3829
Comp2	1.02577	.139768	0.2564	0.6394
Comp3	.886001	.329472	0.2215	0.8609
Comp4	.556529	.	0.1391	1.0000

(2)若希望Q_m在80%左右, 应选取**(3)** 个主成分.

拓展： 一些科技问题的累计贡献率要求在90%以上. 但对复杂的社会科学, 行为科学或经济学中的数据, 能达到60%也可以考虑.

课堂练习

某学校收集了 100 个学生的数学, 物理, 化学, 语文, 历史, 英语的成绩. 令 $X_1, X_2, X_3, X_4, X_5, X_6$ 分别表示数学, 物理, 化学, 语文, 历史, 英语成绩的**经过标准化处理后的**取值向量. 经过主成分分析的计算, 得到这些变量协方差矩阵的特征值分别为:

$$\lambda_1 = 3.735, \lambda_2 = 1.133, \lambda_3 = 0.457, \lambda_4 = 0.323, \lambda_5 = 0.199, \lambda_6 = 0.153$$

对应 λ_1 和 λ_2 的特征向量如表 1 所示:

表 1: 对应于 λ_1 到 λ_2 的特征向量

特征向量	X_1	X_2	X_3	X_4	X_5	X_6
$\boldsymbol{u}_1$	−0.4170	−0.3488	−0.3491	0.4618	0.4268	0.4325
$\boldsymbol{u}_2$	0.3313	0.4985	0.4818	0.2877	0.4090	0.3995

(1) 请写出第一主成分与第二主成分的数学表达式

课堂练习

某学校收集了 100 个学生的数学, 物理, 化学, 语文, 历史, 英语的成绩. 令 $X_1, X_2, X_3, X_4, X_5, X_6$ 分别表示数学, 物理, 化学, 语文, 历史, 英语成绩的**经过标准化处理后的**取值向量. 经过主成分分析的计算, 得到这些变量协方差矩阵的特征值分别为:

$$\lambda_1 = 3.735, \lambda_2 = 1.133, \lambda_3 = 0.457, \lambda_4 = 0.323, \lambda_5 = 0.199, \lambda_6 = 0.153$$

对应 λ_1 和 λ_2 的特征向量如表 1 所示:

表 1: 对应于 λ_1 到 λ_2 的特征向量

特征向量	X_1	X_2	X_3	X_4	X_5	X_6
$\boldsymbol{u}_1$	−0.4170	−0.3488	−0.3491	0.4618	0.4268	0.4325
$\boldsymbol{u}_2$	0.3313	0.4985	0.4818	0.2877	0.4090	0.3995

(1) 请写出第一主成分与第二主成分的数学表达式 $y_i = Xu_i$

(2)计算主平面(m=2)累计贡献率

课堂练习

某学校收集了 100 个学生的数学, 物理, 化学, 语文, 历史, 英语的成绩. 令 $X_1, X_2, X_3, X_4, X_5, X_6$ 分别表示数学, 物理, 化学, 语文, 历史, 英语成绩的**经过标准化处理后的**取值向量. 经过主成分分析的计算, 得到这些变量协方差矩阵的特征值分别为:

$$\lambda_1 = 3.735, \lambda_2 = 1.133, \lambda_3 = 0.457, \lambda_4 = 0.323, \lambda_5 = 0.199, \lambda_6 = 0.153$$

对应 λ_1 和 λ_2 的特征向量如表 1 所示:

表 1: 对应于 λ_1 到 λ_2 的特征向量

特征向量	X_1	X_2	X_3	X_4	X_5	X_6
$\boldsymbol{u}_1$	-0.4170	-0.3488	-0.3491	0.4618	0.4268	0.4325
$\boldsymbol{u}_2$	0.3313	0.4985	0.4818	0.2877	0.4090	0.3995

(1) 请写出第一主成分与第二主成分的数学表达式 $y_i = Xu_i$

(2)计算主平面(m=2)累计贡献率 $(3.735 + 1.133)/6 = 0.811$

课堂练习

某学校收集了 100 个学生的数学、物理、化学、语文、历史、英语的成绩. 令 $X_1, X_2, X_3, X_4, X_5, X_6$ 分别表示数学、物理、化学、语文、历史、英语成绩的**经过标准化处理后的**取值向量. 经过主成分分析的计算, 得到这些变量协方差矩阵的特征值分别为:

$$\lambda_1 = 3.735, \lambda_2 = 1.133, \lambda_3 = 0.457, \lambda_4 = 0.323, \lambda_5 = 0.199, \lambda_6 = 0.153$$

对应 λ_1 和 λ_2 的特征向量如表 1 所示:

表 1: 对应于 λ_1 到 λ_2 的特征向量

特征向量	X_1	X_2	X_3	X_4	X_5	X_6
$\boldsymbol{u}_1$	−0.4170	−0.3488	−0.3491	0.4618	0.4268	0.4325
$\boldsymbol{u}_2$	0.3313	0.4985	0.4818	0.2877	0.4090	0.3995

(3) 计算第一主成分和原始变量X_1的相关系数

课堂练习

某学校收集了 100 个学生的数学、物理、化学、语文、历史、英语的成绩. 令 $X_1, X_2, X_3, X_4, X_5, X_6$ 分别表示数学、物理、化学、语文、历史、英语成绩的**经过标准化处理后的**取值向量. 经过主成分分析的计算, 得到这些变量协方差矩阵的特征值分别为:

$$\lambda_1 = 3.735, \lambda_2 = 1.133, \lambda_3 = 0.457, \lambda_4 = 0.323, \lambda_5 = 0.199, \lambda_6 = 0.153$$

对应 λ_1 和 λ_2 的特征向量如表 1 所示:

表 1: 对应于 λ_1 到 λ_2 的特征向量

特征向量	X_1	X_2	X_3	X_4	X_5	X_6
$\boldsymbol{u}_1$	−0.4170	−0.3488	−0.3491	0.4618	0.4268	0.4325
$\boldsymbol{u}_2$	0.3313	0.4985	0.4818	0.2877	0.4090	0.3995

(3) 计算第一主成分和原始变量X_1的相关系数

答: $r(X_1, y_1) = \sqrt{3.735} * (-0.4170) = -.80590038$

(4) 证明第h 主成分的方差等于λ_h, 即$Var(y_h) = \lambda_h$ 答: 略

主成分的命名

主成分$y_1, \ldots, y_m$ 是原变量$x_1, \ldots, x_p$ 的线性组合. 原变量都有明确的现实含义, 但主成分$y_1, \ldots, y_m$ 的现实含意是什么?

没有唯一的答案, 但可给出一些参考的指导意见.

主成分的命名

命名的参考指导意见:

1. 作用：指出影响系统结构的主要因素和主要特征.

- 例：分析各阶层人员生活状态.
  发展中国家：y_1——食品, y_2——穿着;
  发达国家： y_1——住宅, y_2——旅游;
  以此可以划分不同社会阶层的生活档次 (在这个方向, 人们的生活水平差距最大).

2.方法：专业知识 + 数学手段

- 数学手段：研究 y_h 与$x_1, \ldots, x_p$的相关关系; 计算主成分与原始变量的相关系数; 根据factor loading 图给定名称

PCA的问题

1. 如何进行主成分命名？对主成分的命名与解释很困难, 具有主观性, 随机性.
2. “主成分”是否等同于“主要因素”？不是.

- 例如：利用主成分分析构造评估函数 (1)样本点：n 个有关专家 (2)变量：p个评估指标
- **问题：**用第一主成分构造的评估指标完全不符合人们对实际情况的认识.
- **原因：** 第一主成分对应数据方差最大的方向, 这是专家意见分歧最大的方向!

Bottom line

PCA 是一个降维工具. 具有其应有的功能, 同时也有其局限, 使用中切勿做方法和数据无法支撑的推断或结论.

因子分析(factor analysis)方法 (略)

与PCA分析很相似. 可参考教材.

https://www.theanalysisfactor.com/
the-fundamental-difference-between-principal-component-analysis-and-f

应用统计学II作业5暨上机实验11

1. 对“污染数据 (2014)”做PCA分析和聚类分析
2. 对房地产数据做PCA分析和聚类分析(数据见houses.csv).