

CH7. 多元统计应用

2014年6月25日

主要内容

- ❖ 多元样本及描述
- ❖ 主成分分析
- ❖ 判别分析
- ❖ 聚类分析

一元统计分析： 研究一个随机变量统计规律的学科

多元统计分析： 研究多个随机变量之间相互依赖关系以及
内在统计规律性的统计学科。

多元统计分析的应用：

经济学；
体育科学

生态学
考古学

环境保护

医学

地质学

军事科学

教育学

社会学

文学

例如:

- 根据统计数据对各省、市、自治区的经济社会情况分类

-----聚类分析

- 根据统计数据对全国各省、市、自治区经济效益做综合评价

-----主成分分析

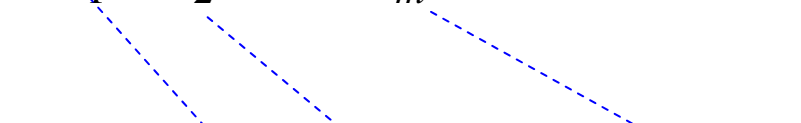
- 对红楼梦统计分析, 判断前80回与后40回作者是否为同一人?

-----判别分析

7.1 多元统计的样本及其描述

本节介绍描述多元数据性质及相互关系的几个指标和变换

把m个指标 $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_m$ 的n次观测值写成矩阵的形式


$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

X 称为样本数据阵

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} = [x_1, x_2, \dots, x_m] = \begin{bmatrix} \mathbf{X}_{(1)}^T \\ \mathbf{X}_{(2)}^T \\ \vdots \\ \mathbf{X}_{(n)}^T \end{bmatrix}$$

其中, X 的第 i 列为指标 X_i 的 n 个观测结果

X 的第 j 行 $\mathbf{X}_{(j)}^T$ 是对 m 个指标的第 j 次观测结果, 它称为一个样品.

n 次的观测结果即 n 个样品构成一个样本

说明:

1) m 个指标 $(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_m)^T$ 构成一个 m 维随机变量.

2) 样本数据阵 $X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$ 在取样观测前不

能确定其取值,它是一个随机矩阵.而一旦取样得到了观测值, X 就变成了一个数据矩阵.

3) 类似于一元统计是以正态总体为基础，多元统计分析是以多元正态总体为基础的，即假设m个指标的总体：

$$(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_m)^T \sim N_m(\mu, \Sigma)$$

其中， $\mu = E(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_m)^T$ 为总体的期望

$$\Sigma = D(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_m)^T = \left[\text{cov}(X_i, X_j) \right]_{m \times m}$$
$$= \begin{bmatrix} D(X_1) & \text{cov}(X_1, X_2) & \cdots & \text{cov}(X_1, X_m) \\ \text{cov}(X_2, X_1) & D(X_2) & \cdots & \text{cov}(X_2, X_m) \\ \vdots & \vdots & & \vdots \\ \text{cov}(X_m, X_1) & \text{cov}(X_m, X_2) \cdots & & D(X_m) \end{bmatrix}$$

为总体的协方差矩阵

问题: 如果总体的期望 μ ,总体协方差阵 Σ 未知,
如何根据样本数据阵 X 来估计 μ 和 Σ ?

μ 和 Σ 的极大似然估计

一元正态情形: $L(\hat{\mu}, \hat{\sigma}^2) = \max_{\mu, \sigma^2} L(\mu, \sigma^2)$

$$\hat{\mu} = \bar{x}, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

多元正态情形: $L(\hat{\mu}, \hat{\Sigma}) = \max_{\mu, \Sigma} L(\mu, \Sigma)$

$$\hat{\mu} = \bar{X}, \quad \hat{\Sigma} = \frac{1}{n} A$$

其中: $\bar{X} = [\bar{x}_1, \bar{x}_2, \dots, \bar{x}_m]^T$, $\bar{x}_i = \frac{1}{n} \sum_{k=1}^n x_{ki}$ $A = \sum_{i=1}^n (X_{(i)} - \bar{X})(X_{(i)} - \bar{X})^T$

$$\bar{X}$$

$$\bar{x}_i = \frac{1}{n} \sum_{k=1}^n x_{ki} \longrightarrow \text{指标 } x_i \text{ 的样本均值}$$

$$\bar{X} = [\bar{x}_1, \bar{x}_2, \dots, \bar{x}_m]^T \longrightarrow \text{n个样品的均值, n个指标观测值的"中心"}$$

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

显然, 样本均值 $\bar{X} = [\bar{x}_1, \bar{x}_2, \dots, \bar{x}_m]^T$ 为总体期望 μ 的无偏点估计

$$\hat{\Sigma} = \frac{1}{n} \mathbf{A} = \frac{1}{n} \sum_{i=1}^n (X_{(i)} - \bar{X})(X_{(i)} - \bar{X})^T \text{ 不是总体协方差阵 } \Sigma \text{ 的无偏估计}$$

可以证明,总体协方差阵 Σ 的无偏估计为 **样本协方差阵S** :

$$S = \frac{1}{n-1} \mathbf{A} = [s_{ij}]_{m \times m}$$

$$\text{其中: } s_{ij} = \frac{1}{n-1} \sum_{k=1}^n (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$$

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

易见, s_{ii} 就是样本数据阵 \mathbf{X} 的第 i 列数据的修正样本方差

样本协方差阵 **S** 对应一元正态总体的 修正样本方差

A/n 对应一元正态总体的 样本方差

样本相关阵**R**

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

$$R = [r_{ij}]_{m \times m}$$

$$\text{其中: } r_{ij} = \frac{\sum_{k=1}^n (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)}{\sqrt{\sum_{k=1}^n (x_{ki} - \bar{x}_i)^2} \sqrt{\sum_{k=1}^n (x_{kj} - \bar{x}_j)^2}} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}}$$

r_{ij} 为指标变量 X_i, X_j 的样本(即样本数据阵的第*i*列,第*j*列)相关系数

样本相关阵**R**表示了各指标变量之间线性关系的强弱.

参见,第五章回归分析中的定义

样品的距离

数学上把满足:非负性,对称性,和三角不等式的函数通称为距离(函数)

设 E 表示一个点集, 距离 d 是直积 $E \times E$ 到非负实数集的函数,满足:

$$(1) \quad d(x, y) \geq 0, \quad \forall x, y \in E;$$

$$(2) \quad d(x, y) = 0 \quad \text{当且仅当} \quad x = y;$$

$$(3) \quad d(x, y) = d(y, x) \quad \forall x, y \in E$$

$$(4) \quad d(x, y) \leq d(x, z) + d(z, y) \quad \forall x, y, z \in E$$

欧氏距离及其局限性

如图,A,B,C,D四点

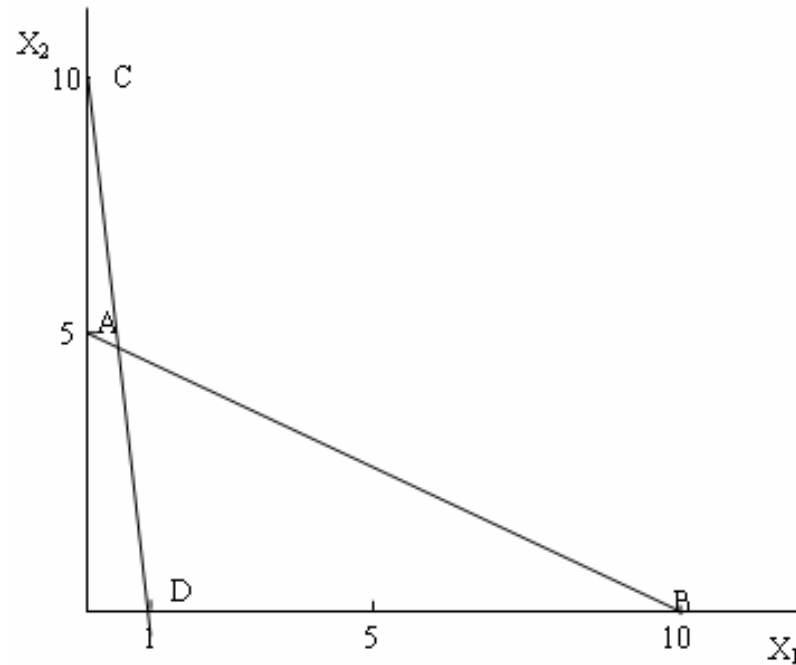
X_1, X_2 分别表示重量和长度

单位分别为 kg 和 cm

A,B,C,D四点的坐标分别为:

A (0,5) B(10,0)

C(0,10) D(1,0)



$$AB = \sqrt{5^2 + 10^2} = \sqrt{125} \quad CD = \sqrt{10^2 + 1^2} = \sqrt{101} \quad (AB > CD)$$

若把 X_2 单位换为mm,则四点坐标分别为A (0,50); B(10,0);C(0,100); D(1,0),此时:

$$AB = \sqrt{50^2 + 10^2} = \sqrt{2600} \quad CD = \sqrt{100^2 + 1^2} = \sqrt{10001} \quad (AB < CD)$$

欧式距离与量纲有关!

-向量的各分量如果单位不全相同，则欧氏距离一般没有意义

-即使单位相同，但如果各分量的变异性差异很大，则变异性大的分量在欧氏距离的平方和中起着决定性的作用，而变异性小的分量却几乎不起什么作用

-在实际应用中，为了消除单位的影响和均等地对待每一分量，我们常须先对各分量作标准化变换，然后再计算欧氏距离

— 欧氏距离经变量的标准化之后能够消除各变量的单位或方差差异的影响，但不能消除变量之间相关性的影响，以致有时用欧氏距离仍然不太合适

闵氏(MINKOWSKI)距离

样品 $X_{(i)}^T$ 与 $X_{(j)}^T$ 的闵氏距离:

$$d_{ij} = d(X_{(i)}^T, X_{(j)}^T) = \left(\sum_{k=1}^m |x_{ik} - x_{jk}|^q \right)^{\frac{1}{q}}$$

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{(1)}^T \\ \mathbf{X}_{(2)}^T \\ \vdots \\ \mathbf{X}_{(n)}^T \end{bmatrix}$$

特别地: 样品 $X_{(i)}^T$ 与 $X_{(j)}^T$ 的闵氏距离 d_{ij} :

1) 当 $q=1$ 时, d_{ij} 为绝对值距离 $d(X_{(i)}^T, X_{(j)}^T) = \left(\sum_{k=1}^m |x_{ik} - x_{jk}| \right)$

2) 当 $q=2$ 时, d_{ij} 为欧式距离 $d(X_{(i)}^T, X_{(j)}^T) = \sqrt{\sum_{k=1}^m |x_{ik} - x_{jk}|^2}$

3) 当 $q \rightarrow \infty$ 时, d_{ij} 为切比雪夫距离 $d(X_{(i)}^T, X_{(j)}^T) = \max_{1 \leq k \leq m} |x_{ik} - x_{jk}|$

马氏(Mahalanobis)距离

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{(1)}^T \\ \mathbf{X}_{(2)}^T \\ \vdots \\ \mathbf{X}_{(n)}^T \end{bmatrix}$$

样品 $X_{(i)}^T$ 与 $X_{(j)}^T$ 的马氏距离:

$$d_{ij} = d(X_{(i)}^T, X_{(j)}^T) = \sqrt{(X_{(i)} - X_{(j)})^T \mathbf{S}^{-1} (X_{(i)} - X_{(j)})}$$

其中, \mathbf{S} 为样本协方差矩阵

马氏距离消除了指标变量单位的影响,是一个无量纲的量

当总体协方差阵 Σ 已知时,马氏距离公式中用 Σ 代替样本协方差阵 \mathbf{S}

两个样品之间的距离反映了两个样品的差异程度

样本数据的标准化变换

在数据处理时，常常因各变量的单位不同而需要对每个变量作标准化变换

比如:某班级数学平均成绩 75, 标准差 6.26; 物理平均成绩80, 标准差 10

该班某同学数学成绩是80分,物理是85分.问该同学数学考得好还是物理考得好?

如果直接比较两门课的成绩: 数学80<物理85, 认为物理考得好 是显然不合理的.因为两门课程考试的难度可能不同,成绩没有可比性

但是,如果考虑到班级全体同学的考试成绩,则可以相对比较:

数学 $(80-75)/6.25=0.8$, 而物理: $(85-80)/10=0.5$,即相对说来数学比物理考得好

常用的标准化变换

1) 标准差标准化

(消除了量纲,标准化后各指标的样本均值为0,修正方差为1)

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \quad (i=1,2,\dots,n;j=1,2,\dots,m)$$

其中 s_j 为指标 X_j 的修正样本标准差,即 $s_j = \sqrt{s_{jj}}$

2) 极差标准化 (消除了量纲,标准化后的数据都在区间[0,1]内)

$$\tilde{x}_{ij} = \frac{x_{ij} - \min_{1 \leq i \leq n} x_{ij}}{R_j} \quad (i=1,2,\dots,n;j=1,2,\dots,m)$$

其中 R_j 为指标 X_j 样本数据的极差.

样本数据阵 X 经过标准化变换后变为 $\tilde{X} = [\tilde{x}_{ij}]_{n \times m}$ 称为标准化的样本数据阵

7.2 主成分分析

主成分分析（**Principal Component Analysis**）一般认为是由皮尔逊(Pearson)1901年首先引入，后被霍特林(Hotelling)1933年进一步发展的一种多元统计分析方法

在实际生活中，我们经常会遇到需要对多个变量进行统计推断的统计分析问题。变量个数多了，就不容易看清变量之间的相互关系，不容易从中得出有用的结论，会给统计分析带来很大的困难

比如: 对一个人做健康体检，会包含几十项生理指标：血压、心率、血糖、血脂、胆固醇、血小板、甲胎蛋白、.....，等等。直接根据这么多数据，判断一个人是否健康、是否有病，是很复杂的问题,能否把这些数据简化为几个主要的指标？

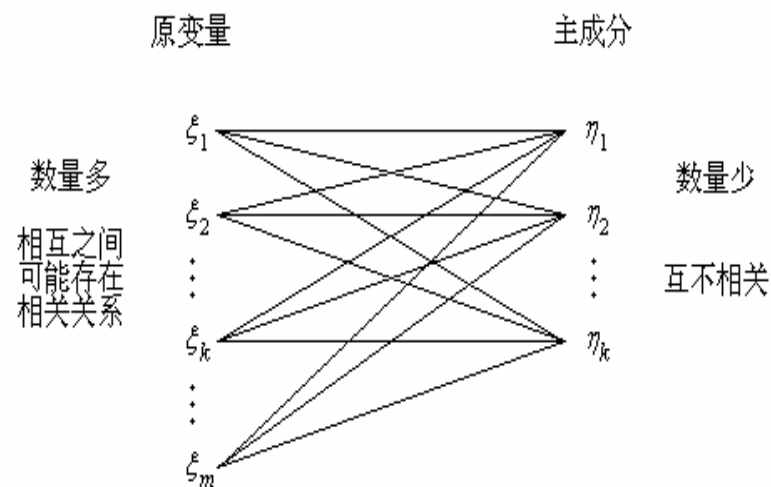
又比如我们去定做一套服装，从理论上说，需要测量身长、袖长、裤长、胸围、腰围、臀围、领口、袖口、裤口等十几种、几十种尺寸。可是实际上，裁缝师傅并没有测量这么多指标.因为这些尺寸之间往往是有一定比例关系的，只要知道几个主要的尺码，就大致上了解了原来十几种甚至更多指标中所包含的信息。

主成分分析的基本思想

对原来多个变量进行适当的组合，组合成一些综合指标，用较少的综合指标来近似代替原来的多个变量。这种由原来多个变量组合而成的综合指标，就称为**主成分**（也称**主分量**，**Principal Component**）

主成分选取的原则:

- 1) 主成分是原变量的线性组合
- 2) 各个主成分之间互不相关
- 3) 主成分的个数不超过原有变量的个数. 并且主成分应尽可能多地反映原变量信息



主成分分析的过程和推导

设对 m 个变量 $\xi_1, \xi_2, \dots, \xi_m$ 进行 n 次观测，得到观测数据矩阵 X ：

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \xrightarrow{\text{标准差标准化}} \begin{bmatrix} \tilde{x}_{11} & \tilde{x}_{12} & \cdots & \tilde{x}_{1m} \\ \tilde{x}_{21} & \tilde{x}_{22} & \cdots & \tilde{x}_{2m} \\ \vdots & \vdots & & \vdots \\ \tilde{x}_{n1} & \tilde{x}_{n2} & \cdots & \tilde{x}_{nm} \end{bmatrix} = \tilde{X}$$

$$\text{其中: } \tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \quad \bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij} \quad s_j = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}$$

易证(ex7.3): 样本相关阵 $R = \frac{1}{n-1} \tilde{X}^T \tilde{X}$

样本相关阵 $R = \frac{1}{n-1} \tilde{X}^T \tilde{X}$ 为半正定对称矩阵,故存在正交矩阵U使得:

$$R = U \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_m \end{bmatrix} U^T = U \Lambda U^T$$

其中: $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m \geq 0$ 为R的特征值

$$U = \begin{bmatrix} u_{11} & \cdots & u_{1m} \\ \vdots & & \vdots \\ u_{m1} & \cdots & u_{mm} \end{bmatrix} \text{为正交阵,}$$

称为主成分载荷阵 (**Principal Component Loading Matrix**)

记原变量标准化后的变量为 $\tilde{\xi}_1, \tilde{\xi}_2, \dots, \tilde{\xi}_m$ (即 \tilde{X} 的各列分别为其样本)

令 $[\eta_1 \cdots \eta_m] = [\tilde{\xi}_1 \cdots \tilde{\xi}_m] U$, 则 η_1, \dots, η_m 就是主成分

$$\begin{cases} \eta_1 = u_{11} \tilde{\xi}_1 + \cdots + u_{m1} \tilde{\xi}_m \\ \vdots \\ \eta_m = u_{1m} \tilde{\xi}_1 + \cdots + u_{mm} \tilde{\xi}_m \end{cases}$$

对应 标准化变元 $\tilde{\xi}_1, \tilde{\xi}_2, \dots, \tilde{\xi}_m$ 的样本数据阵为 \tilde{X}

主成分 $\eta_1, \eta_2, \dots, \eta_m$ 的样本数据阵为: $Y = \tilde{X}U$

Y 称为 **主成分得分阵 (Principal Component Score Matrix)**

性质:

1) 主成分 $\eta_1, \eta_2, \dots, \eta_m$ 的样本均值都等于0

证明: 设主成分 η_i 的样本均值为 \bar{y}_i , 则

$$[\bar{y}_1 \cdots \bar{y}_m] = \frac{1}{n} [1 \quad 1 \quad \cdots \quad 1] Y = \frac{1}{n} [1 \quad 1 \quad \cdots \quad 1] \tilde{X} U = [0 \quad 0 \quad \cdots \quad 0] U = [0 \quad 0 \quad \cdots \quad 0]$$

2) 主成分互不相关 (样本协方差阵为对角阵):

$$\begin{aligned} S &= [s_{ij}]_{m \times m} = \left[\frac{1}{n-1} \sum_{k=1}^n (y_{ki} - \bar{y}_i)(y_{kj} - \bar{y}_j) \right]_{m \times m} \\ &= \left[\frac{1}{n-1} \sum_{k=1}^n (y_{ki})(y_{kj}) \right]_{m \times m} = \frac{1}{n-1} Y^T Y \\ &= \frac{1}{n-1} (\tilde{X} U)^T \tilde{X} U = \frac{1}{n-1} U^T \tilde{X}^T \tilde{X} U = U^T R U = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_m \end{bmatrix} \end{aligned}$$

3) 主成分 $\eta_1, \eta_2, \dots, \eta_m$ 的修正样本方差分别为 $\lambda_1, \lambda_2, \dots, \lambda_m$

如果设 $Y = [y_1 \cdots y_m]$,

其中 y_i 是由主成分 η_i 的观测值 (得分) 组成的列向量

$$\begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_m \end{bmatrix} = \frac{1}{n-1} Y^T Y = \frac{1}{n-1} \begin{bmatrix} y_1^T \\ \vdots \\ y_m^T \end{bmatrix} [y_1 \cdots y_m]$$

$$\Rightarrow \frac{1}{n-1} y_i^T y_i = \lambda_i$$

\Rightarrow 若某 $\lambda_k = 0$, 则必有 $y_k = 0$ (即主成分 η_k 的得分 (观测值) 为零向量)

\Rightarrow 若 $\lambda_k = 0$, 则必有 $\lambda_{k+1} = \dots = \lambda_m = 0$, 于是 $y_k = \dots = y_m = 0$

$$Y = [y_1 \quad \cdots \quad y_m] = \tilde{X}U = \tilde{X}[u_1 \quad \cdots \quad u_m]$$

$$\Rightarrow \tilde{X} = YU^T = [y_1 \quad \cdots \quad y_m] \begin{bmatrix} u_1^T \\ \vdots \\ u_m^T \end{bmatrix} = y_1 u_1^T + \cdots + y_m u_m^T$$

若某 $\lambda_k=0$, 则 $\lambda_{k+1} = \dots = \lambda_m = 0$, 于是 $y_k = \dots = y_m = 0$

$$\Rightarrow \tilde{X} = YU^T = [y_1 \quad \cdots \quad y_m] \begin{bmatrix} u_1^T \\ \vdots \\ u_m^T \end{bmatrix} = y_1 u_1^T + \cdots + y_{k-1} u_{k-1}^T$$

即原来 m 个变量的观测数据，只用前 $k-1$ 个主成分，就可以完全精确地表示出来了

在实际问题中，不一定有特征值为0的情况，但常常会遇到后面若干个特征值都很小，近似等于0的情况(对应主成分的样本观测值近似为零向量),此时也可以忽略这些主成分.

那么，怎样才能算“特征值很小，近似等于0”呢？

$$\sum_{j=1}^m \lambda_j = \text{trace } \Lambda = \text{trace}(U^T R U) = \text{trace } R = m$$

由此可见每个特征值的平均大小等于 1。习惯上：如果前 k 个特征值大于 1，后 m-k 个特征值小于 1，或者前 k 个特征值的和/m大于 80%~85%，就可以近似认为后面 m-k 个特征值都近似等于0。

$\frac{\lambda_i}{\lambda_1 + \dots + \lambda_m}$ 反映了第 i 个主成分 η_i 对表示原变量贡献的大小(贡献率)

$\frac{\lambda_1 + \dots + \lambda_k}{\lambda_1 + \dots + \lambda_m}$ 称为前 k 个主成分的累积贡献率

主成分 $\eta_1, \eta_2, \dots, \eta_m$ 的样本均值全为0, 修正样本方差为 $\lambda_1, \lambda_2, \dots, \lambda_m$

$\tilde{\eta}_j = \frac{\eta_j}{\sqrt{\lambda_j}}$ 为主成分 η_j 的标准化, $\tilde{\eta}_j$ 称为 **因子** ($j=1, 2, \dots, m$)

由 $[\eta_1 \cdots \eta_m] = [\tilde{\xi}_1 \cdots \tilde{\xi}_m] U$, 得:

$$[\eta_1 \cdots \eta_m] \begin{bmatrix} 1/\sqrt{\lambda_1} & & \\ & \ddots & \\ & & 1/\sqrt{\lambda_m} \end{bmatrix} = [\tilde{\xi}_1 \cdots \tilde{\xi}_m] U \begin{bmatrix} 1/\sqrt{\lambda_1} & & \\ & \ddots & \\ & & 1/\sqrt{\lambda_m} \end{bmatrix}$$

$$\Rightarrow [\tilde{\eta}_1 \cdots \tilde{\eta}_m] = [\tilde{\xi}_1 \cdots \tilde{\xi}_m] U \Lambda^{-1/2} = [\tilde{\xi}_1 \cdots \tilde{\xi}_m] \tilde{U} \Lambda^{-1}$$

其中 $\tilde{U} = U \begin{bmatrix} \sqrt{\lambda_1} & & \\ & \ddots & \\ & & \sqrt{\lambda_m} \end{bmatrix}$ 称为因子载荷阵 (**Factor Loading Matrix**)

$$\begin{bmatrix} \tilde{\eta}_1 \cdots \tilde{\eta}_m \end{bmatrix} = \begin{bmatrix} \tilde{\xi}_1 \cdots \tilde{\xi}_m \end{bmatrix} U \Lambda^{-1/2}$$



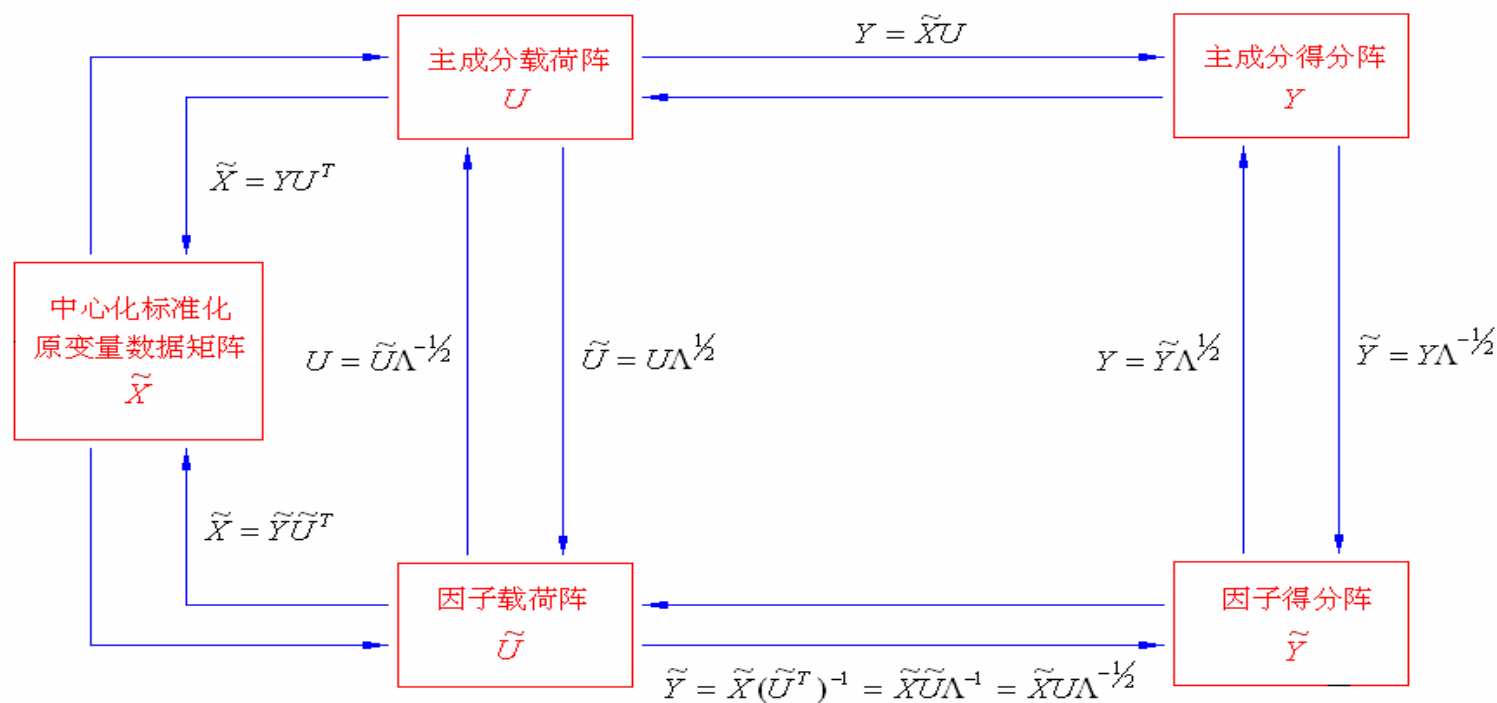
$$\begin{cases} \tilde{\eta}_1 = \frac{1}{\sqrt{\lambda_1}} (u_{11} \tilde{\xi}_1 + \cdots + u_{m1} \tilde{\xi}_m) \\ \vdots \\ \tilde{\eta}_m = \frac{1}{\sqrt{\lambda_m}} (u_{1m} \tilde{\xi}_1 + \cdots + u_{mm} \tilde{\xi}_m) \end{cases}$$



$$\begin{cases} \tilde{\xi}_1 = u_{11} \sqrt{\lambda_1} \tilde{\eta}_1 + \cdots + u_{1m} \sqrt{\lambda_m} \tilde{\eta}_m \\ \vdots \\ \tilde{\xi}_m = u_{m1} \sqrt{\lambda_1} \tilde{\eta}_1 + \cdots + u_{mm} \sqrt{\lambda_m} \tilde{\eta}_m \end{cases}$$

因子 $[\tilde{\eta}_1 \cdots \tilde{\eta}_m] = [\tilde{\xi}_1 \cdots \tilde{\xi}_m] U \Lambda^{-1/2}$ 对应的样本数矩阵记为 \tilde{Y}

$$\tilde{Y} = \tilde{X} U \Lambda^{-1/2} = \tilde{X} U \begin{bmatrix} \frac{1}{\sqrt{\lambda_1}} & & \\ & \ddots & \\ & & \frac{1}{\sqrt{\lambda_m}} \end{bmatrix} \text{ 称为因子得分阵 (Factor Score Matrix)}$$



主成分分析的应用

1) 按第1个主成分排序(第一主成分贡献率较大时适用)

例: 选取**58**个与和谐发展有关的经济增长、人文发展、社会进步、生态文明指数，根据全国全国**30**个省、直辖市、自治区（不包括西藏、台湾）的统计数据，对全国各省、直辖市、自治区排序。

对统计数据作主成分分析，得到的前**3**个主成分贡献率如下：

	第1主成分	第2主成分	第3主成分
特征值	21.39754663	7.761203640	3.901120059
贡献率	36.892322%	13.381386%	6.7260691%
累计贡献率	36.892322%	50.273707%	56.999776%

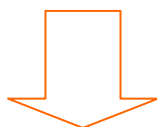
由于第1个特征值特别大，贡献率占了比较大的百分比，所以，我们可以考虑把第1个主成分的得分作为“和谐发展指数”的得分，按照得分大小，对全国30个省、直辖市、自治区进行排序，最后得到结果如下：

排名	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
地区	北京	上海	天津	浙江	广东	江苏	辽宁	福建	海南	山东	吉林	新疆	湖北	湖南	河北

排名	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
地区	黑龙江	重庆	陕西	内蒙古	山西	四川	江西	广西	河南	安徽	宁夏	青海	云南	甘肃	贵州

2) 主成分回归(用变元的主成分代替原变元求回归函数)

$$\frac{1}{n-1} \tilde{X}^T \tilde{X} = R = U \Lambda U^T = U \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_m \end{bmatrix} U^T$$



$$r(\tilde{X}) = r(\tilde{X}^T \tilde{X}) = r(R) = \text{非零特征值的个数}$$

例: 某地共有**12**个观测降水量的气象站, 积累了**10**年的降水量观测数据。考虑到各个气象站测到的降水量之间, 可能有相关关系, 有些气象站的数据, 可能是不必要的, 所以, 现在为了节省开支, 希望减少一些气象站, 问可减少几个?

12个气象站观测到的降水量，就是12个变量。我们对10年的观测数据作主成分分析

主成分序号	1	2	3	4	5	6	7	8	9	10	11	12
贡献率	27.5 %	22.5 %	16.6 %	10.7 %	9.5%	8.4%	3.3%	1.1%	0.6%	0.0%	0.0%	0.0%
累计贡献率	27.5 %	49.9 %	66.5 %	77.1 %	86.6 %	95.0 %	98.3 %	99.4 %	100 %	100 %	100 %	100 %

最后3个主成分的贡献率都是0，前9个主成分的贡献率达到100%，所以，最多只要9个气象站就足够了。

前5个主成分的累计贡献率已经达到86.6%，如果希望节省更多开支，我们可以只取前面5个主成分

$$[y_1 \quad \cdots \quad y_m] = Y = \tilde{X}U = \tilde{X}[u_1 \quad \cdots \quad u_m] = [\tilde{X}u_1 \quad \cdots \quad \tilde{X}u_m]$$

$$\lambda_j = 0 \quad \Rightarrow \quad y_j = 0 \quad \Rightarrow \quad \tilde{X}u_j = y_j = 0$$

$$u_{1j}\tilde{\xi}_1 + \cdots + u_{mj}\tilde{\xi}_m = 0$$

$$u_{1j} \frac{\xi_1 - \bar{x}_1}{s_1} + \cdots + u_{mj} \frac{\xi_m - \bar{x}_m}{s_m} = 0$$

有几个特征值等于**0**，就有几个相互独立的线性关系

例1 1966年，Malinvand收集了一组有关法国经济的数据，其中 ξ_1 为国内总产值， ξ_2 为存储量， ξ_3 为总消费量。三个变量的样本均值和样本标准差如下：

	ξ_1 国内总产值	ξ_2 存储量	ξ_3 总消费量
样本均值 \bar{x}_j	194.59	3.3	139.74
样本标准差 s_j	28.60	1.572	19.67

对数据中心化标准化后，进行主成分分析:

	主成分 η_1	主成分 η_2	主成分 η_3
特征值	1.999	0.998	0.003
贡献率	66.6%	33.3%	0.1%
累计贡献率	66.6%	99.9%	100%

主成分载荷阵

	主成分 η_1	主成分 η_2	主成分 η_3
原变量 $\tilde{\xi}_1$	0.7063	-0.0357	-0.7070
原变量 $\tilde{\xi}_2$	0.0435	0.9990	-0.0070
原变量 $\tilde{\xi}_3$	0.7065	-0.0258	0.7072

$$\lambda_3 \approx 0$$

$$-0.7070\tilde{\xi}_1 - 0.0070\tilde{\xi}_2 + 0.7072\tilde{\xi}_3 \approx 0$$

$$-\tilde{\xi}_1 + \tilde{\xi}_3 \approx 0$$

$$-\frac{\xi_1 - 194.59}{28.60} + \frac{\xi_3 - 139.74}{19.67} \approx 0$$

$$\xi_3 \approx 0.688\xi_1 + 5.9$$

主成分分析:

	主成分 η_1	主成分 η_2	主成分 η_3
特征值	1.999	0.998	0.003
贡献率	66.6%	33.3%	0.1%
累计贡献率	66.6%	99.9%	100%

主成分载荷阵:

	主成分 η_1	主成分 η_2	主成分 η_3
原变量 $\tilde{\xi}_1$	0.7063	-0.0357	-0.7070
原变量 $\tilde{\xi}_2$	0.0435	0.9990	-0.0070
原变量 $\tilde{\xi}_3$	0.7065	-0.0258	0.7072

3) 主成分的解释

$$[\eta_1 \cdots \eta_m] = [\tilde{\xi}_1 \cdots \tilde{\xi}_m] U \quad \Leftrightarrow \quad \begin{cases} \eta_1 = u_{11} \tilde{\xi}_1 + \cdots + u_{m1} \tilde{\xi}_m \\ \vdots \\ \eta_m = u_{1m} \tilde{\xi}_1 + \cdots + u_{mm} \tilde{\xi}_m \end{cases}$$

根据各个主成分用原变量表达时的系数，可以对主成分的含义进行解释

例2 某公司要招聘一名管理人员，有48人前去应聘。公司老板对这48名应聘者进行面谈后，从15个方面给他们打分数，这15个方面是：申请书形式，外貌，学术能力，讨人喜欢，自信，精明，诚实，推销能力，经验，积极性，抱负，理解能力，潜力，交际能力，适应性。评分范围从0分到10分。这样就得到了一个由15个变量的48次观测组成的数据矩阵

根据样本数据阵,进行主成分分析, 得到前6个主成分的特征值和贡献率如下

	主成分 η_1	主成分 η_2	主成分 η_3	主成分 η_4	主成分 η_5	主成分 η_6
特征值	7. 499	2. 058	1. 462	1. 207	0. 739	0. 493
贡献率	50. 0%	13. 7%	9. 7%	8. 0%	4. 9%	3. 3%
累计贡献率	50. 0%	63. 7%	73. 5%	81. 5%	86. 4%	89. 7%

主成分载荷阵中前4个主成分的载荷

原变量编号	原变量名称	主成分 η_1	主成分 η_2	主成分 η_3	主成分 η_4
1	申请书形式	0. 16	0. 43	0. 31	-0. 11
2	外貌	0. 21	-0. 03	-0. 01	0. 26
3	学术能力	0. 04	0. 24	-0. 41	0. 65
4	讨人喜欢	0. 22	-0. 13	0. 48	0. 33
5	自信程度	0. 29	-0. 25	-0. 24	-0. 16
6	精明	0. 32	-0. 13	-0. 15	-0. 06
7	诚实	0. 16	-0. 40	0. 30	0. 41
8	推销能力	0. 32	-0. 04	-0. 20	-0. 21
9	经验	0. 13	0. 55	0. 08	0. 06
10	积极性	0. 32	0. 05	-0. 08	-0. 15
11	抱负	0. 32	-0. 07	-0. 21	-0. 19
12	理解力	0. 33	-0. 02	-0. 11	0. 08
13	潜力	0. 33	0. 02	-0. 06	0. 19
14	交际能力	0. 26	-0. 08	0. 46	-0. 21
15	适应性	0. 24	0. 42	0. 09	-0. 03

第一个主成分在各个变量上的载荷都是正的，大小也差不多，在“精明”、“推销能力”、“积极性”、“抱负”、“理解力”、“潜力”上载荷较大，可以认为它代表了一般的办事能力和精明能干程度。这个主成分可以称为“办事能力”因子。

第二个主成分 在“申请书形式”、“经验”、“适应性”上有很大的正载荷，在“诚实”上有较大的负载荷，似乎代表了经验和适应能力。由于善于适应，难免有些不诚实。这个主成分可以称为“适应能力”因子

第三个主成分 在“讨人喜欢”、“交际能力”上有很大的正载荷，在“学术能力”上有较大的负载荷，似乎代表交际能力。由于忙于交际，在学术上用力就少了。这个主成分可以称为“交际能力”因子

第四个主成分 在“学术能力”、“诚实”上有很大的正载荷，在其他变量上载荷都比较小。这个主成分可以称为“学术能力”因子

原变量编号	原变量名称	主成分 η_1	主成分 η_2	主成分 η_3	主成分 η_4
1	申请书形式	0.16	0.43	0.31	-0.11
2	外貌	0.21	-0.03	-0.01	0.26
3	学术能力	0.04	0.24	-0.41	0.65
4	讨人喜欢	0.22	-0.13	0.48	0.33
5	自信程度	0.29	-0.25	-0.24	-0.16
6	精明	0.32	-0.13	-0.15	-0.06
7	诚实	0.16	-0.40	0.30	0.41
8	推销能力	0.32	-0.04	-0.20	-0.21
9	经验	0.13	0.55	0.08	0.06
10	积极性	0.32	0.05	-0.08	-0.15
11	抱负	0.32	-0.07	-0.21	-0.19
12	理解力	0.33	-0.02	-0.11	0.08
13	潜力	0.33	0.02	-0.06	0.19
14	交际能力	0.26	-0.08	0.46	-0.21
15	适应性	0.24	0.42	0.09	-0.03

本例中分析出4种能力因子，对于人才类型的划分，对于人才的招聘、培养、使用，都是很有意义的。

例如，如果招聘者注重的是办事能力，就可以优先录用那些在“办事能力”因子上主成分得分比较高的应聘者；

如果招聘者注重的是交际能力，就可以优先录用那些在“交际能力”因子上主成分得分比较高的应聘者

4) 主成分分析结果的图示

$$\begin{bmatrix} \tilde{u}_1^T \tilde{u}_1 & \cdots & \tilde{u}_1^T \tilde{u}_m \\ \vdots & & \vdots \\ \tilde{u}_m^T \tilde{u}_1 & \cdots & \tilde{u}_m^T \tilde{u}_m \end{bmatrix} = \tilde{U} \tilde{U}^T = (U \Lambda^{1/2})(U \Lambda^{1/2})^T = U \Lambda U^T = R = \begin{bmatrix} r_{11} & \cdots & r_{1m} \\ \vdots & & \vdots \\ r_{m1} & \cdots & r_{mm} \end{bmatrix}$$

$$\tilde{u}_i^T \tilde{u}_j = r_{ij}$$

$$\cos \theta_{ij} = \frac{\tilde{u}_i^T \tilde{u}_j}{\sqrt{\tilde{u}_i^T \tilde{u}_i} \sqrt{\tilde{u}_j^T \tilde{u}_j}} = \frac{r_{ij}}{\sqrt{r_{ii}} \sqrt{r_{jj}}} = r_{ij}$$

其中 θ_{ij} 为向量 \tilde{u}_i 与 \tilde{u}_j 的夹角

当 $\theta_{ij} = 0^\circ$ 时, $r_{ij} = \cos \theta_{ij} = 1$, 表示变量 ξ_i 和 ξ_j 正线性相关;

当 $\theta_{ij} = 90^\circ$ 时, $r_{ij} = \cos \theta_{ij} = 0$, 表示变量 ξ_i 和 ξ_j 不相关;

当 $\theta_{ij} = 180^\circ$ 时, $r_{ij} = \cos \theta_{ij} = -1$, 表示变量 ξ_i 和 ξ_j 负线性相关;

在直角坐标系中, 用因子载荷矩阵中的各行数据作为向量坐标, 作出 m 个向量, 每一个向量代表一个原变量. 根据因子载荷阵的作图可以判断原变量的相关关系:

向量夹角 $\theta \approx 0^\circ$

ξ_i 与 ξ_j 正相关

向量夹角 $\theta \approx 90^\circ$

ξ_i 与 ξ_j 不相关

向量夹角 $\theta \approx 180^\circ$

ξ_i 与 ξ_j 负相关

用因子得分阵 \tilde{Y} 中的各行数据 $\tilde{y}_1^T, \dots, \tilde{y}_n^T$ 作为点的坐标,
作出 n 个点, 每一个点代表一次观测

$$\begin{bmatrix} \tilde{y}_1^T \\ \vdots \\ \tilde{y}_n^T \end{bmatrix} = \tilde{Y} = Y \Lambda^{-1/2} = \tilde{X} U \Lambda^{-1/2} = \begin{bmatrix} \tilde{x}_1^T U \Lambda^{-1/2} \\ \vdots \\ \tilde{x}_n^T U \Lambda^{-1/2} \end{bmatrix}$$

$$\tilde{y}_i^T = \tilde{x}_i^T U \Lambda^{-1/2}$$

$$\begin{aligned} \sqrt{(\tilde{y}_i - \tilde{y}_j)^T (\tilde{y}_i - \tilde{y}_j)} &= \sqrt{(\tilde{x}_i - \tilde{x}_j)^T U \Lambda^{-1/2} \Lambda^{-1/2} U^T (\tilde{x}_i - \tilde{x}_j)} \\ &= \sqrt{(\tilde{x}_i - \tilde{x}_j)^T U \Lambda^{-1} U^T (\tilde{x}_i - \tilde{x}_j)} = \sqrt{(\tilde{x}_i - \tilde{x}_j)^T R^{-1} (\tilde{x}_i - \tilde{x}_j)} \end{aligned}$$

图中各点之间的几何距离的大小, 反映了各次观测值之间的马氏距离的远近

$$\begin{bmatrix} \tilde{y}_1^T \tilde{u}_1 & \cdots & \tilde{y}_1^T \tilde{u}_m \\ \vdots & & \vdots \\ \tilde{y}_n^T \tilde{u}_1 & \cdots & \tilde{y}_n^T \tilde{u}_m \end{bmatrix} = \tilde{Y} \tilde{U}^T = \tilde{X} = \begin{bmatrix} \tilde{x}_{11} & \cdots & \tilde{x}_{1m} \\ \vdots & & \vdots \\ \tilde{x}_{n1} & \cdots & \tilde{x}_{nm} \end{bmatrix}$$

$$\tilde{y}_i^T \tilde{u}_j = \tilde{x}_{ij}$$

设 \tilde{y}_i 是图中代表第*i*次观测的点，

\tilde{u}_j 是图中代表第*j*个变量 ξ_j 的向量，

则点 \tilde{y}_i 在向量 \tilde{u}_j 上的投影长度为 $\frac{\tilde{y}_i^T \tilde{u}_j}{\sqrt{\tilde{u}_j^T \tilde{u}_j}} = \frac{\tilde{x}_{ij}}{\sqrt{r_{jj}}} = \tilde{x}_{ij}$

第*i*次观测时， ξ_j 的值越大，点 \tilde{y}_i 在向量 \tilde{u}_j 上的投影越长

例 3 1973 年, Weber 对欧洲 25 个国家平均每人每天从 9 种食品来源中得到的蛋白质的数量作了统计。这 9 种食品来源是:

ξ_1 ——牛羊肉类, ξ_2 ——猪禽肉类, ξ_3 ——蛋类, ξ_4 ——乳类, ξ_5 ——鱼类,

ξ_6 ——谷类, ξ_7 ——薯类, ξ_8 ——花生豆类, ξ_9 ——果蔬类。

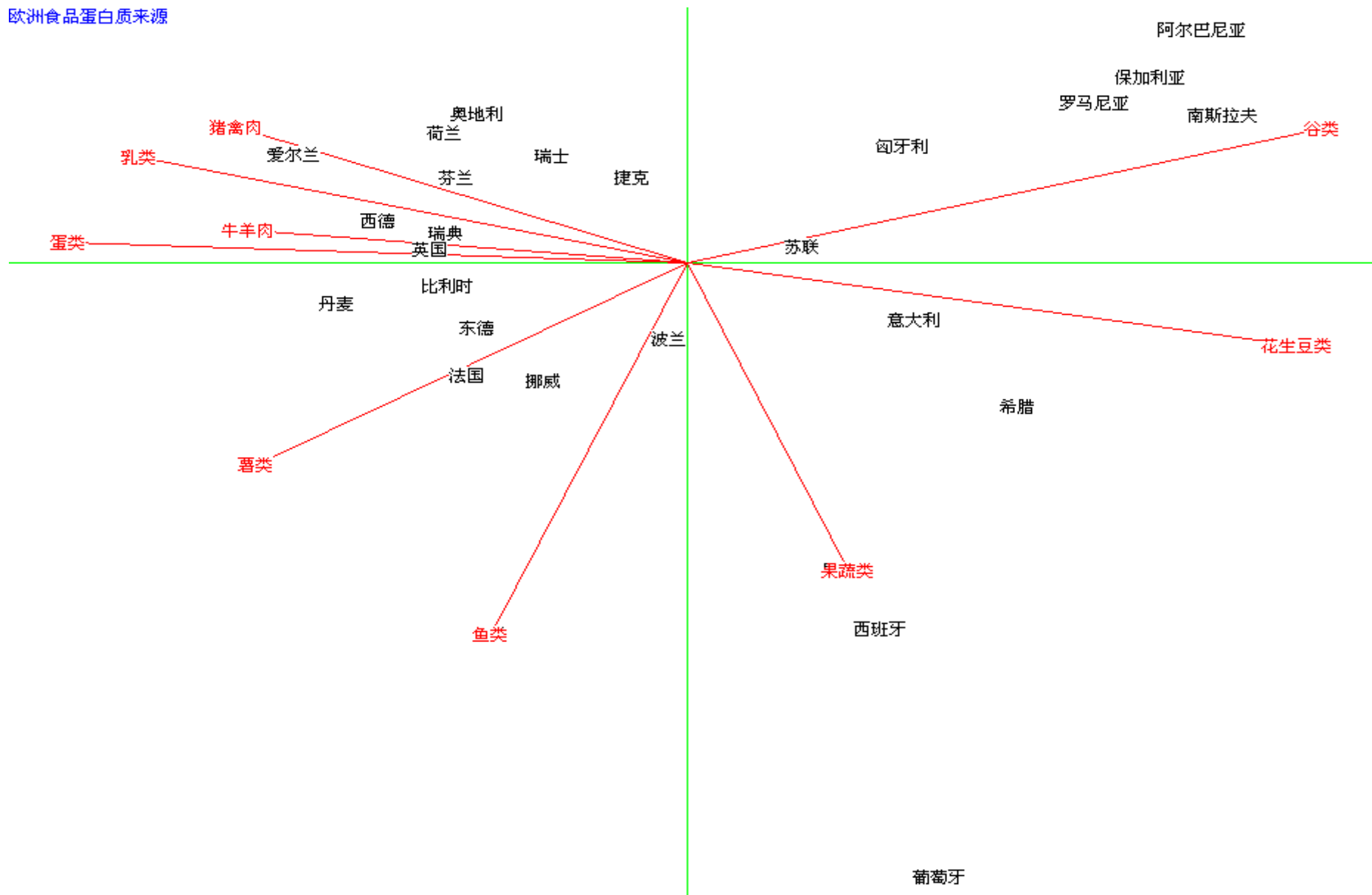
国家	牛羊肉类	猪禽肉类	蛋类	乳类	鱼类	谷类	薯类	花生豆类	果蔬类
阿尔巴尼亚	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7
奥地利	8.9	14.0	4.3	19.9	2.1	28.0	3.6	1.3	4.3
比利时	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4.0
保加利亚	7.8	6.0	1.6	8.3	1.2	56.7	1.1	3.7	4.2
捷克	9.7	11.4	2.8	12.5	2.0	34.3	5.0	1.1	4.0
丹麦	10.6	10.8	3.7	25.0	9.9	21.9	4.8	0.7	2.4
东德	8.4	11.6	3.7	11.1	5.4	24.6	6.5	0.8	3.6
芬兰	9.5	4.9	2.7	33.7	5.8	26.3	5.1	1.0	1.4
法国	18.0	9.9	3.3	19.5	5.7	28.1	4.8	2.4	6.5
希腊	10.2	3.0	2.8	17.6	5.9	41.7	2.2	7.8	6.5
匈牙利	5.3	12.4	2.9	9.7	0.3	40.1	4.0	5.4	4.2
爱尔兰	13.9	10.0	4.7	25.8	2.2	24.0	6.2	1.6	2.9
意大利	9.0	5.1	2.9	13.7	3.4	36.8	2.1	4.3	6.7
荷兰	9.5	13.6	3.6	23.4	2.5	22.4	4.2	1.8	3.7
挪威	9.4	4.7	2.7	23.3	9.7	23.0	4.6	1.6	2.7
波兰	6.9	10.2	2.7	19.3	3.0	36.1	5.9	2.0	6.6
葡萄牙	6.2	3.7	1.1	4.9	14.2	27.0	5.9	4.7	7.9
罗马尼亚	6.2	6.3	1.5	11.1	1.0	49.6	3.1	5.3	2.8
西班牙	7.1	3.4	3.1	8.6	7.0	29.2	5.7	5.9	7.2
瑞典	9.9	7.8	3.5	24.7	7.5	19.5	3.7	1.4	2.0
瑞士	13.1	10.1	3.1	23.8	2.3	25.6	2.8	2.4	4.9
英国	17.4	5.7	4.7	20.6	4.3	24.3	4.7	3.4	3.3
苏联	9.3	4.6	2.1	16.6	3.0	43.6	6.4	3.4	2.9
西德	11.4	12.5	4.1	18.8	3.4	18.6	5.2	1.5	3.8
南斯拉夫	4.4	5.0	1.2	9.5	0.6	55.9	3.0	5.7	3.2

对这些数据进行主成分分析，得到前4个主成分的特征值和贡献率如下：

	主成分 η_1	主成分 η_2	主成分 η_3	主成分 η_4
特征值	4.006	1.635	1.128	0.955
贡献率	44.5%	18.2%	12.5%	10.6%
累计贡献率	44.5%	62.7%	75.2%	85.8%

以各地区前两个主成分对应的坐标作图. 令人惊讶的是：按照食品中蛋白质来源作出的点图，竟然与按照各国地理位置作出的欧洲地图十分相似！

欧洲食品蛋白质来源



例4 1985年，有人作了一次民意调查，要求被调查者，对中国社会下列28种不同职业，按照“权力”，“声望地位”和“收入”3项指标，分别打分：

党政领导，税务干部，管理干部，法警干部，公务员，文书，清洁工，纺织工，建筑工，矿工，电工，出租司机，采购员，店员，理发师，服务员，保姆，修理工，士兵，教师，医生，工程师，科学家，教授，演员，农民，个体户，私企业主。

对这些数据进行主成分分析，得到3个主成分的特征值和贡献率如下：

	主成分 η_1	主成分 η_2	主成分 η_3
特征值	1.912	0.783	0.306
贡献率	63.7%	21.6%	10.2%
累计贡献率	63.7%	89.8%	100.0%

社会调查

