多元统计与矩阵分析

张锋　*8125345@qq.com*

中国地质大学, 计算机学院, 武汉

主成分分析

Principal component analysis

内容

(1) 基本思想

(2) 数学模型及几何意义

(3) 推导与计算

(4) 结构与性质

(5) 应用

内容

(1) 基本思想

(2) 数学模型及几何意义

(3) 推导与计算

(4) 结构与性质

(5) 应用

主成分分析的基本思想

- 主成分分析就是把原有的多个指标转化成少数几个代表性较好的综合指标，这少数几个指标能够反映原来指标大部分的信息（85%以上），并且各个指标之间保持独立，避免出现重叠信息。
- 主成分分析主要起着降维和简化数据结构的作用。
- 主成分分析试图在力保数据信息丢失最少的原则下，对这种多变量的截面数据表进行最佳综合简化，也就是说，对高维变量空间进行降维处理。

例子

- 1961年英国统计学家M. Scott对157个英国城镇发展水平的调查分析，用5个综合变量表示57个原始变量，信息量保持95%。
- 1957年美国统计学家Stone关于国民经济的研究，利用1929-1938年美国经济数据，用3个新变量取代原17个变量，信息量保持97.4%。

需要解决的问题

- 在力求数据信息丢失最少的原则下，对高维的变量空间降维，即研究指标体系的少数几个线性组合，并且这几个线性组合所构成的综合指标将尽可能多地保留原来指标变异方面的信息。这些综合指标就称为主成分。要讨论的问题是：

(1)基于相关系数矩阵/协方差矩阵做主成分分析？

(2)选择几个主成分？

(3)如何解释主成分所包含的经济意义？

内容

(1) 基本思想

(2) 数学模型及几何意义

(3) 推导与计算

(4) 结构与性质

(5) 应用

数学模型与几何解释

假设我们所讨论的实际问题中，有p个指标，我们把这p个指标看作p个随机变量，记为X_1，X_2，…，X_p，主成分分析就是要把这p个指标的问题，转变为讨论 m 个新的指标F_1，F_2，…，F_m(m<p），按照保留主要信息量的原则充分反映原指标的信息，并且相互独立。

数学模型

$$\boldsymbol{X}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}'=(\boldsymbol{X}_1\ \boldsymbol{X}_2\ \ldots \boldsymbol{X}_p)'$$

其中 $\boldsymbol{X}_i=\begin{pmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{ni} \end{pmatrix}$

这种由讨论多个指标降为少数几个综合指标的过程在数学上就叫做降维。主成分分析通常的做法是，寻求原指标的线性组合Y_i，即：

$$\begin{cases} \boldsymbol{Y}_1 = a_{11}\boldsymbol{X}_1 + a_{21}\boldsymbol{X}_2 + \ldots + a_{p1}\boldsymbol{X}_p \\ \boldsymbol{Y}_2 = a_{12}\boldsymbol{X}_1 + a_{2_2}\boldsymbol{X}_2 + \ldots + a_{p2}\boldsymbol{X}_p \\ \cdots\cdots\cdots \\ \boldsymbol{Y}_p = a_{1p}\boldsymbol{X}_1 + a_{2p}\boldsymbol{X}_2 + \ldots + a_{pp}\boldsymbol{X}_p \end{cases}$$

简写为：$\boldsymbol{Y}_i = a_{1i}\boldsymbol{X}_1 + a_{2i}\boldsymbol{X}_2 + \ldots + a_{pi}\boldsymbol{X}_p = \boldsymbol{a}_i'\boldsymbol{X}$

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \\ \vdots \\ \boldsymbol{Y}_p \end{pmatrix} \qquad \boldsymbol{A} = (\boldsymbol{a}_1 \boldsymbol{a}_2 \ldots \boldsymbol{a}_p) \longrightarrow \boldsymbol{Y} = \boldsymbol{A}'\,\boldsymbol{X}$$

满足如下的条件：

每个主成分的系数平方和为**1**。即

$$a_{1i}{}^2 + a_{2i}{}^2 + \ldots + a_{pi}{}^2 = 1 \qquad \boldsymbol{a}_i' \boldsymbol{a}_i = 1$$

主成分之间相互独立，即无重叠的信息。即

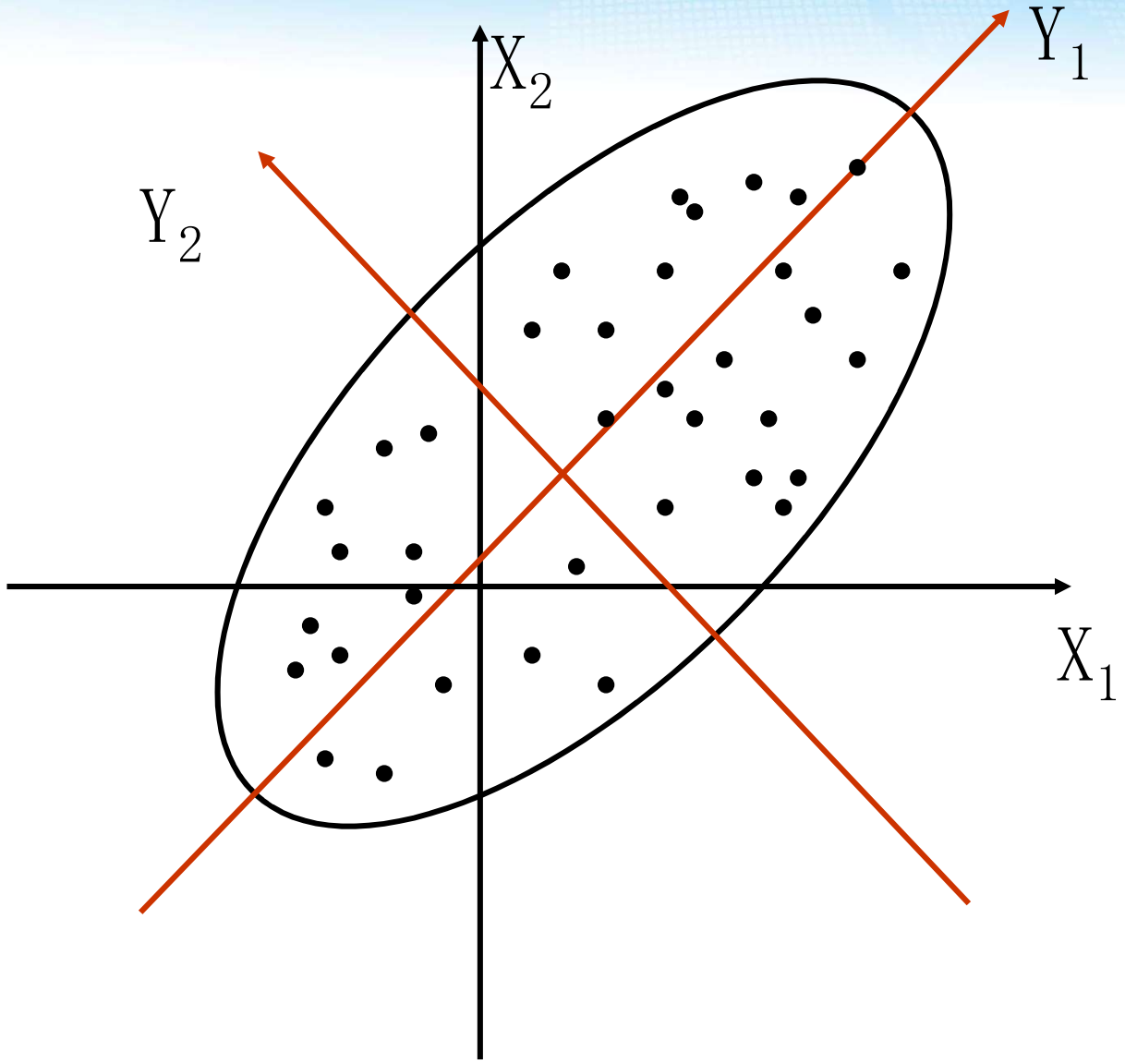
$$\mathrm{cov}(\boldsymbol{Y}_i, \boldsymbol{Y}_j) = 0$$

主成分的方差依次递减，重要性依次递减，即

$$\mathrm{var}(\boldsymbol{Y}_1) > \mathrm{var}(\boldsymbol{Y}_2) > \ldots > \mathrm{var}(\boldsymbol{Y}_p)$$

平移、旋转坐标轴

主成分分析的几何解释



几何解释

假设n个样本，每个样品都观察了两个变量(X1,X2)，散点图如下图所示，大致为一个椭圆。若在椭圆长轴方向取坐标*Y1*，短轴方向取坐标Y2，相当于在平面做坐标变换，如果用新的坐标点(Y1,Y2)表示样品点，则很显然样品点在Y1上的变化幅度最大，Y1方差最大，意味着(Y1,Y2)的信息大部分集中在Y1上。

如果上图中的椭圆形状相当扁平，使得n个点的波动大致可以归结为Y1轴的波动，如下图所示，那么我们可以认为这种波动只归结在Y1方向上而忽略Y2方向应该是合理的。这样，二维问题就降为一维，即取Y1为(X1,X2)的综合指标，也称(X1,X2)的主成分。

实践证明，原变量之间相关程度越高，主成分分析效果越好。

内容

(1) 基本思想

(2) 数学模型及几何意义

(3) 推导与计算

(4) 结构与性质

(5) 应用

主成分的推导

一、两个线性代数的结论

1、若**A**是p阶实对称阵，则一定可以找到正交阵**U**，使

$$\mathbf{U}^{-1}\mathbf{A}\mathbf{U}=\begin{bmatrix}\lambda_1 & 0 & \cdots & 0\\ 0 & \lambda_2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \lambda_p\end{bmatrix}$$

其中 $\lambda_i, i=1.2.\cdots p$ 是**A**的特征根。

2、若上述矩阵的特征根所对应的单位特征向量为 $\mathbf{u}_1,\cdots,\mathbf{u}_p$

令 $$\mathbf{U}=(\mathbf{u}_1,\cdots,\mathbf{u}_p)=\begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1p} \\ u_{21} & u_{22} & \cdots & u_{2p} \\ \vdots & \vdots & & \vdots \\ u_{p1} & u_{p2} & \cdots & u_{pp} \end{bmatrix}$$

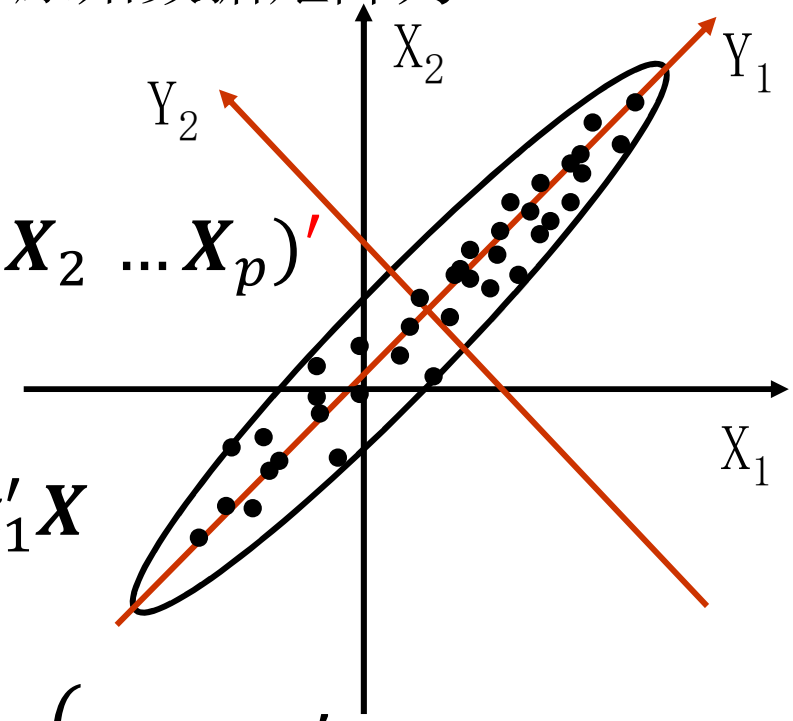
则实对称阵 $\mathbf{A}$ 属于不同特征根所对应的特征向量是正交的，即有 $\mathbf{U}'\mathbf{U}=\mathbf{U}\mathbf{U}'=\mathbf{I}$

二、主成分的推导

（一） 第一主成分

设有n个样品，每个样品观察p个指标，原始数据矩阵为：

$$\boldsymbol{X}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}' = (\boldsymbol{X}_1\ \boldsymbol{X}_2\ \ldots \boldsymbol{X}_p)'$$



$$\boldsymbol{F}_1 = a_{11}\boldsymbol{X}_1 + \cdots + a_{p1}\boldsymbol{X}_p = \boldsymbol{\alpha}_1'\boldsymbol{X}$$

$\boldsymbol{F}_1$是第一主成分，其方差必须最大

$$var(\boldsymbol{F}_1) = var(\boldsymbol{\alpha}_1'\boldsymbol{X}) = \boldsymbol{\alpha}_1'\Sigma\boldsymbol{\alpha}_1 \longrightarrow \begin{cases} \max\limits_{\alpha_1} \boldsymbol{\alpha}_1'\Sigma\boldsymbol{\alpha}_1 \\ s.t.\,\boldsymbol{\alpha}_i'\,\boldsymbol{\alpha}_i = 1 \end{cases}$$

$$\begin{cases} \max\limits_{\boldsymbol{\alpha}_1} \boldsymbol{\alpha}_1' \Sigma \boldsymbol{\alpha}_1 \\ s.t. \boldsymbol{\alpha}_i' \boldsymbol{\alpha}_i = 1 \end{cases}$$

将问题转化为以下拉格朗日函数的极大值

$$Q = \boldsymbol{\alpha}_1' \boldsymbol{\Sigma} \boldsymbol{\alpha}_1 - \lambda(\boldsymbol{\alpha}_1' \boldsymbol{\alpha}_1 - 1)$$

$$\frac{\partial Q}{\partial \boldsymbol{\alpha}_1} = 2\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 - 2\lambda\boldsymbol{\alpha}_1 = 0 \longrightarrow \boldsymbol{\Sigma}\boldsymbol{\alpha}_1 = \lambda\boldsymbol{\alpha}_1$$

$$\mathrm{var}(\boldsymbol{F}_1) = var(\boldsymbol{\alpha}_1' \boldsymbol{X}) = \boldsymbol{\alpha}_1' \Sigma \boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_1' \lambda \boldsymbol{\alpha}_1 = \lambda$$

如果第一主成分的信息不够，则需要寻找第二主成分。

（二） 第二主成分

寻找合适的单位向量 $\boldsymbol{\alpha}_2$，使$\boldsymbol{F}_2$的方差最大。

$$var(\boldsymbol{F}_2) = var(\boldsymbol{\alpha}_2'\boldsymbol{X}) = \boldsymbol{\alpha}_2'\boldsymbol{\Sigma}\boldsymbol{\alpha}_2$$

$$\max_{\boldsymbol{\alpha}_2} \boldsymbol{\alpha}_2'\boldsymbol{\Sigma}\boldsymbol{\alpha}_2$$

满足$\boldsymbol{\alpha}_2'\boldsymbol{\alpha}_2 = 1, \boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 = 0, \boldsymbol{\alpha}_2'\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 = 0$

注意到：

$$\mathrm{cov}(\boldsymbol{F}_1, \boldsymbol{F}_2) = \boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 = \boldsymbol{\alpha}_2'\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2'\lambda\boldsymbol{\alpha}_1 = \lambda\boldsymbol{\alpha}_2'\boldsymbol{\alpha}_1 = 0$$

引入拉格朗日乘数，将问题转化为求以下式子的极大值。

$$Q = \boldsymbol{\alpha}_2'\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda(\boldsymbol{\alpha}_2'\boldsymbol{\alpha}_2 - 1) - 2\rho\boldsymbol{\alpha}_2'\boldsymbol{\alpha}_1$$

$$\frac{\partial Q}{\partial \boldsymbol{\alpha}_2} = 2\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - 2\lambda\boldsymbol{\alpha}_2 - 2\rho\boldsymbol{\alpha}_1 = 0 \Longrightarrow \boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda\boldsymbol{\alpha}_2 - \rho\boldsymbol{\alpha}_1 = 0 \Longrightarrow \boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda\boldsymbol{\alpha}_2 = 0$$

主成分的计算

$$\boldsymbol{X}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}'=(\boldsymbol{X}_1\ \boldsymbol{X}_2\ \ldots \boldsymbol{X}_p)'$$

计算步骤

(1) 样本均值$\overline{\boldsymbol{X}}=(\bar{x}_1\ \bar{x}_2\ \ldots \bar{x}_p)$和样本协方差矩阵$\boldsymbol{S}=[s_{ij}]_{p\times p}$

$$s_{ij}=\frac{\sum_{k=1}^{n}(x_{ik}-\bar{x}_i)(x_{jk}-\bar{x}_j)}{n-1}\quad i,j=1,2,\ldots,p$$

(2) 求解特征方程$|\boldsymbol{S}-\lambda\boldsymbol{I}|=0$，求得$p$个特征根$\lambda_1,\lambda_2,\ldots,\lambda_p$ $(\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p)$

(3) 求λ_k对应的单位特征向量$\boldsymbol{a}_k$，即求解方程组$(\boldsymbol{S}-\lambda_k\boldsymbol{I})\boldsymbol{a}_k=0$，$\boldsymbol{a}_k=(a_{1k}\ a_{2k}\ ..a_{pk})'$

(4) 写出主成分表达式$\boldsymbol{Y}_k=a_{1k}\boldsymbol{X}_1+a_{2k}\boldsymbol{X}_2+\ldots+a_{pk}\boldsymbol{X}_p=\boldsymbol{a}_k'\boldsymbol{X}$

主成分的计算

- 计算步骤
- 主成分含义
- 主成分的重要性
- 主成分个数的选择

例1　下面是8 个学生两门课程的成绩表

语文	100	90	70	70	85	55	55	45
数学	65	85	70	90	65	45	55	65

对此进行主成分分析。

（1） 求样本均值和样本协方差矩阵

$$\bar{\boldsymbol{X}} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \end{pmatrix} = \begin{pmatrix} 71.25 \\ 67.5 \end{pmatrix}$$

$$\boldsymbol{S} = \begin{pmatrix} \frac{323.4}{7} & \\ \frac{103.1}{7} & \frac{187.5}{7} \end{pmatrix}$$

（2） 求解特征方程 $|\boldsymbol{S}-\lambda\boldsymbol{I}|=0$ $\boldsymbol{s}=\begin{pmatrix}\frac{323.4}{7} & \\ \frac{103.1}{7} & \frac{187.5}{7}\end{pmatrix}$

$$\begin{vmatrix}\frac{323.4}{7}-\lambda & \frac{103.1}{7} \\ \frac{103.1}{7} & \frac{187.5}{7}-\lambda\end{vmatrix}=0$$

$$(\frac{323.4}{7}-\lambda)(\frac{187.5}{7}-\lambda)-(\frac{103.1}{7})^2=0$$

化简得：$\lambda^2-\frac{510.9}{7}\lambda+\frac{50007.9}{7^2}=0$

解得：$\lambda_1=\frac{378.9}{7},\lambda_2=\frac{132}{7}$

（3）求特征值所对应的单位特征向量 $\boldsymbol{S}=\begin{pmatrix}\frac{323.4}{7} & \\ \frac{103.1}{7} & \frac{187.5}{7}\end{pmatrix}$

λ_1 所对应的单位特征向量 $(\boldsymbol{S}-\lambda_1\boldsymbol{I})\boldsymbol{\alpha}=0$，其中 $\boldsymbol{\alpha}=\begin{pmatrix}a_{11}\\a_{21}\end{pmatrix}$

$$\begin{cases}(\frac{323.4}{7}-\frac{378.9}{7})a_{11}+\frac{103.1}{7}a_{21}=0\\ \frac{103.1}{7}a_{11}+(\frac{187.5}{7}-\frac{378.9}{7})a_{21}=0\end{cases}$$

$$a_{11}^2+a_{21}^2=1$$

解得：$(a_{11},a_{21})=(0.88,0.47)$

λ_2 所对应的单位特征向量 $(\boldsymbol{S}-\lambda_2\boldsymbol{I})\boldsymbol{\alpha}=0$，其中 $\boldsymbol{\alpha}=\begin{pmatrix}\alpha_{12}\\\alpha_{22}\end{pmatrix}$

$$\begin{cases}(\frac{323.4}{7}-\frac{132}{7})a_{12}+\frac{103.1}{7}a_{22}=0\\ \frac{103.1}{7}a_{12}+(\frac{187.5}{7}-\frac{132}{7})a_{22}=0\end{cases}$$

$$a_{12}^2+a_{22}^2=1$$

解得：$(a_{12},a_{22})=(-0.47,0.88)$

（4） 得到主成分的表达式

第一主成分：　$y_1 = 0.88(x_1 - 71.25) + 0.47(x_2 - 67.5)$

第二主成分：　$y_2 = -0.47(x_1 - 71.25) + 0.88(x_2 - 67.5)$

（5）主成分的含义

通过分析主成分的表达式中原变量前的系数来解释各主成分的含义。

第一主成分F_1是 X_1和 X_2 的加权和，表示该生成绩的好坏。

第二主成分F_2表示学生两科成绩的均衡性

例2 下表是10位学生的身高 x_1 、胸围 x_2、体重 x_3 的数据。

身高x_1(cm)	胸围x_2(cm)	体重x_3(kg)
149.5	69.5	38.5
162.5	77.0	55.5
162.7	78.5	50.8
162.2	87.5	65.5
156.5	74.5	49.0
156.1	74.5	45.5
172.0	76.5	51.0
173.2	81.5	59.5
159.5	74.5	43.5
157.7	79.0	53.5

对此进行主成分分析。

（1） 求样本均值和样本协方差矩阵

$$\begin{pmatrix}\overline{x_1}\\ \overline{x_2}\\ \overline{x_3}\end{pmatrix}=\begin{pmatrix}161.2\\ 77.3\\ 51.2\end{pmatrix}\qquad \boldsymbol{S}=\begin{pmatrix}46.67 & & \\ 17.12 & 21.11 & \\ 30.00 & 32.58 & 55.53\end{pmatrix}$$

（2） 求解协方差矩阵的特征方程

$$\begin{vmatrix}46.67-\lambda & 17.12 & 30.00\\ 17.12 & 21.11-\lambda & 32.58\\ 30.00 & 32.58 & 55.53-\lambda\end{vmatrix}=0$$

（3）解得三个特征值和对应的单位特征向量：

$\lambda_1=98.15$ $\quad (a_{11},a_{21},a_{31})=(0.56,0.42,0.71)$

$\lambda_2=23.60$ $\quad (a_{12},a_{22},a_{32})=(0.81,-0.33,-0.48)$

$\lambda_3=1.56$ $\quad (a_{13},a_{23},a_{33})=(0.03,0.85,-0.53)$

（4） 由此我们可以写出三个主成分的表达式：

$$F_1 = 0.56(x_1 - 161.2) + 0.42(x_2 - 77.3) + 0.71(x_3 - 51.2)$$

$$F_2 = 0.81(x_1 - 161.2) - 0.33(x_2 - 77.3) - 0.48(x_3 - 51.2)$$

$$F_3 = 0.03(x_1 - 161.2) + 0.85(x_2 - 77.3) - 0.53(x_3 - 51.2)$$

（5） 主成分的含义

F_1表示学生身材大小。

F_2反映学生的体形特征

三个主成分的方差贡献率分别为：

$$\frac{\lambda_1}{\sum_{i=1}^{3}\lambda_i}=\frac{98.15}{98.15+23.60+1.56}=\frac{98.15}{123.31}=79.6\%$$

$$\frac{\lambda_2}{\sum_{i=1}^{3}\lambda_i}=\frac{23.60}{123.31}=19.1\%$$

$$\frac{\lambda_3}{\sum_{i=1}^{3}\lambda_i}=\frac{1.56}{123.31}=1.3\%$$

前两个主成分的累积方差贡献率为：

$$\frac{\lambda_1+\lambda_2}{\sum_{i=1}^{3}\lambda_i}=\frac{121.75}{123.31}=98.7\%$$

例3　对88个学生5 门不同课程的考试成绩进行分析，要求用合适的方法对这5 门课程成绩进行平均，以对88个学生的成绩进行评比。这5门课程是：Mechanics Vectors (闭)，Algebra Analysis Statistics (开)。

经计算，得到5个主成分的表达式如下：

$$F_1 = 0.51x_1 + 0.37x_2 + 0.35x_3 + 0.45x_4 + 0.53x_5 - 99.7$$

$$F_2 = 0.75x_1 + 0.21x_2 - 0.08x_3 - 0.30x_4 - 0.55x_5 + 1.5$$

$$F_3 = -0.30x_1 + 0.42x_2 + 0.15x_3 + 0.60x_4 - 0.60x_5 - 19.8$$

$$F_4 = 0.30x_1 - 0.78x_2 - 0.00x_3 + 0.52x_4 - 0.18x_5 + 11.1$$

$$F_5 = 0.08x_1 + 0.19x_2 - 0.92x_3 + 0.29x_4 + 0.15x_5 + 13.9$$

这5个主成分的方差分别为679.2，199.8，102.6， 83.7和31.8。

前两个主成分各自的贡献率和累积贡献率为

$$\frac{\lambda_1}{\sum_{i=1}^{5}\lambda_i}=\frac{679.2}{1097.1}=61.91\%$$

$$\frac{\lambda_2}{\sum_{i=1}^{5}\lambda_i}=\frac{199.8}{1097.1}=18.21\%$$

$$\frac{\lambda_1+\lambda_2}{\sum_{i=1}^{5}\lambda_i}=61.91\%+18.21\%=80.12\%$$

主成分个数的选取原则

根据累积贡献率的大小取前面m 个(m<p)主成分

选取原则：

$$\frac{\sum_{i=1}^{m-1}\lambda_i}{\sum_{i=1}^{p}\lambda_i} < 80\sim 85\% \quad \text{且} \quad \frac{\sum_{i=1}^{m}\lambda_i}{\sum_{i=i}^{p}\lambda_i} \geq 80\sim 85\%$$

R型分析的概念

为消除量纲影响，在计算之前先将原始数据标准化。标准化变量的 S=R，所以用*标准化变量*进行主成分分析相当于从原变量的*相关矩阵 R* 出发进行主成分分析。统计学上称这种分析法为R型分析，由协方差矩阵出发的主成分分析为S型分析。

S型分析和R型分析的结果是不同的。在一般情况下，若各变量的量纲不同，通常采用R型分析。

内容

(1) 基本思想

(2) 数学模型及几何意义

(3) 推导与计算

(4) 结构与性质

(5) 应用

主成分的结构

一、主成分的相关结构

1. 主成分Y_k的方差 λ_k 反映了第k个主成分Y_k所起的作用
2. 主成分Y_k的方差贡献率为$\frac{\lambda_k}{\sum_{k=1}^{p}\lambda_k}$,表示第$k$个主成分$Y_k$反映了原变量$X_1, X_2, \ldots, X_p$多少的信息
3. 主成分Y_k与每个变量X_i之间的相关系数

因子负荷量
（因子载荷）

$$r_{iY_k} = a_{ik} \cdot \sqrt{\frac{\lambda_k}{s_{ii}}}$$

如何证明？

证明：$\rho(\boldsymbol{Y}_k, \boldsymbol{X}_i) = \dfrac{\mathrm{cov}(\boldsymbol{Y}_k, \boldsymbol{X}_i)}{\sqrt{\mathrm{var}(\boldsymbol{Y}_k)\,\mathrm{var}(\boldsymbol{X}_i)}} = \dfrac{\mathrm{cov}(a_k'\boldsymbol{X}, e_i'\boldsymbol{X})}{\sqrt{\lambda_k s_{ii}}}$

其中，$\boldsymbol{a}_k' = (a_{1k}, a_{2k}, \ldots a_{pk})$; $\boldsymbol{e}_k' = (0,0,\ldots 0,1,0,\ldots,0)$ 为单位向量，第i个分量为1，其余为0.

而 $\mathrm{cov}(\boldsymbol{a}_k'\boldsymbol{X},\ \boldsymbol{e}_i'\boldsymbol{X}) = \boldsymbol{a}_k'\,\mathrm{var}(\boldsymbol{X})\boldsymbol{e}_i$

$= \boldsymbol{a}_k'\boldsymbol{\Sigma}\boldsymbol{e}_i = \ \boldsymbol{e}_i'\boldsymbol{\Sigma}\boldsymbol{a}_k$

$= \boldsymbol{e}_i'\,\lambda_k\boldsymbol{a}_k = \lambda_k\boldsymbol{a}_{\mathrm{i}k}$

所以 $\rho(\boldsymbol{Y}_k, \boldsymbol{X}_i) = r_{i\boldsymbol{Y}_k} = a_{ik}\sqrt{\dfrac{\lambda_k}{s_{ii}}}$

4. 主成分对每个原变量的方差贡献 $r_{iF_k}^2 = a_{ik}^2 \frac{\lambda_k}{s_{ii}}$

对例5.3进行计算：

r_{ik}	F1	F2	F3	F4	F5
X1	0.758	0.609	-0.175	0.156	0.026
X2	0.734	0.224	0.322	0.548	0.081
X3	0.853	-0.136	0.139	-0.003	-0.493
X4	0.796	-0.288	0.409	0.321	0.109
X5	0.812	-0.451	-0.354	-0.094	0.050

r_{ik}^2	F1	F2	F3	F4	F5
X1	0.574	0.371	0.030	0.024	0.001
X2	0.539	0.050	0.104	0.300	0.007
X3	0.727	0.018	0.019	0.000	0.243
X4	0.634	0.083	0.168	0.103	0.012
X5	0.680	0.204	0.125	0.009	0.002

二、主成分的性质

1. 主成分的协差阵为对角阵

$$D(\boldsymbol{F}) = \boldsymbol{\Lambda} = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p \end{pmatrix}$$

2. 总方差保持不变

$$\lambda_1 + \lambda_2 + \cdots + \lambda_p = s_{11} + s_{22} + \cdots + s_{pp}$$

若进行R型分析，则 $\lambda_1 + \lambda_2 + \cdots + \lambda_p = p$

3. $\boldsymbol{F}_k$与 $\boldsymbol{X}_i$的相关系数

$$\rho(\boldsymbol{F}_k, \boldsymbol{X}_i) = r_{i\boldsymbol{F}_k} = a_{ik}\sqrt{\frac{\lambda_k}{s_{ii}}}$$

若进行R型分析，则 $\rho(\boldsymbol{F}_k, \boldsymbol{X}_i) = r_{i\boldsymbol{F}_k} = a_{ik}\sqrt{\lambda_k}$

4. $\boldsymbol{F}_k$ 对 $\boldsymbol{X}_i$ 的方差贡献为

从横行看有 $$\sum_{k=1}^{p}\rho^2(\boldsymbol{F}_k,\boldsymbol{X}_i)=\sum_{k=1}^{p}r_{i\boldsymbol{F}_k}^2=1$$

从纵向看有 $$\sum_{i=1}^{p}\rho^2(\boldsymbol{F}_k,\boldsymbol{X}_i)\cdot s_{ii}=\sum_{i=1}^{p}r_{i\boldsymbol{F}_k}^2\cdot s_{ii}=\lambda_k$$

若进行**R**型分析，则

$$\sum_{i=1}^{p}\rho^2(\boldsymbol{F}_k,\boldsymbol{X}_i)=\sum_{i=1}^{p}r_{i\boldsymbol{F}_k}^2=\lambda_k$$

内容

(1) 基本思想

(2) 数学模型及几何意义

(3) 推导与计算

(4) 结构与性质

(5) 应用

图解样品（对样品分类）

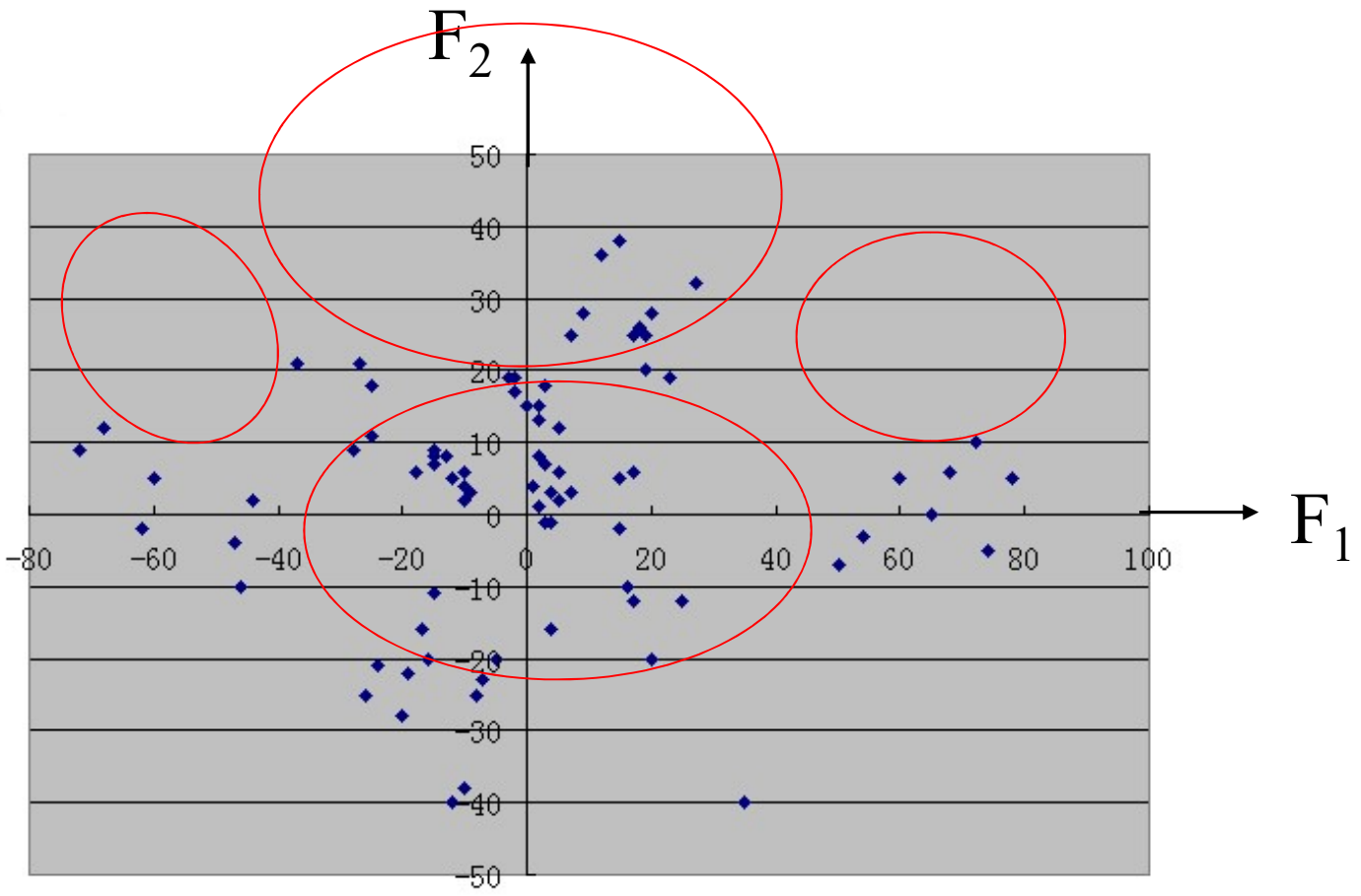
主成分分析后，若能以两个主成分代表原变量大部分的信息，则我们可以在平面上分析每一个样品点。步骤如下：

1、对每个样品分别求第一主成分F_1和第二主成分F_2的得分。

2、建立以F_1和F_2为轴的直角坐标系。以F_1为横坐标，F_2为纵坐标，在坐标系中描出各个样品点（画散点图）。

3、解释坐标系的各个象限。

图解样品（对样品分类）

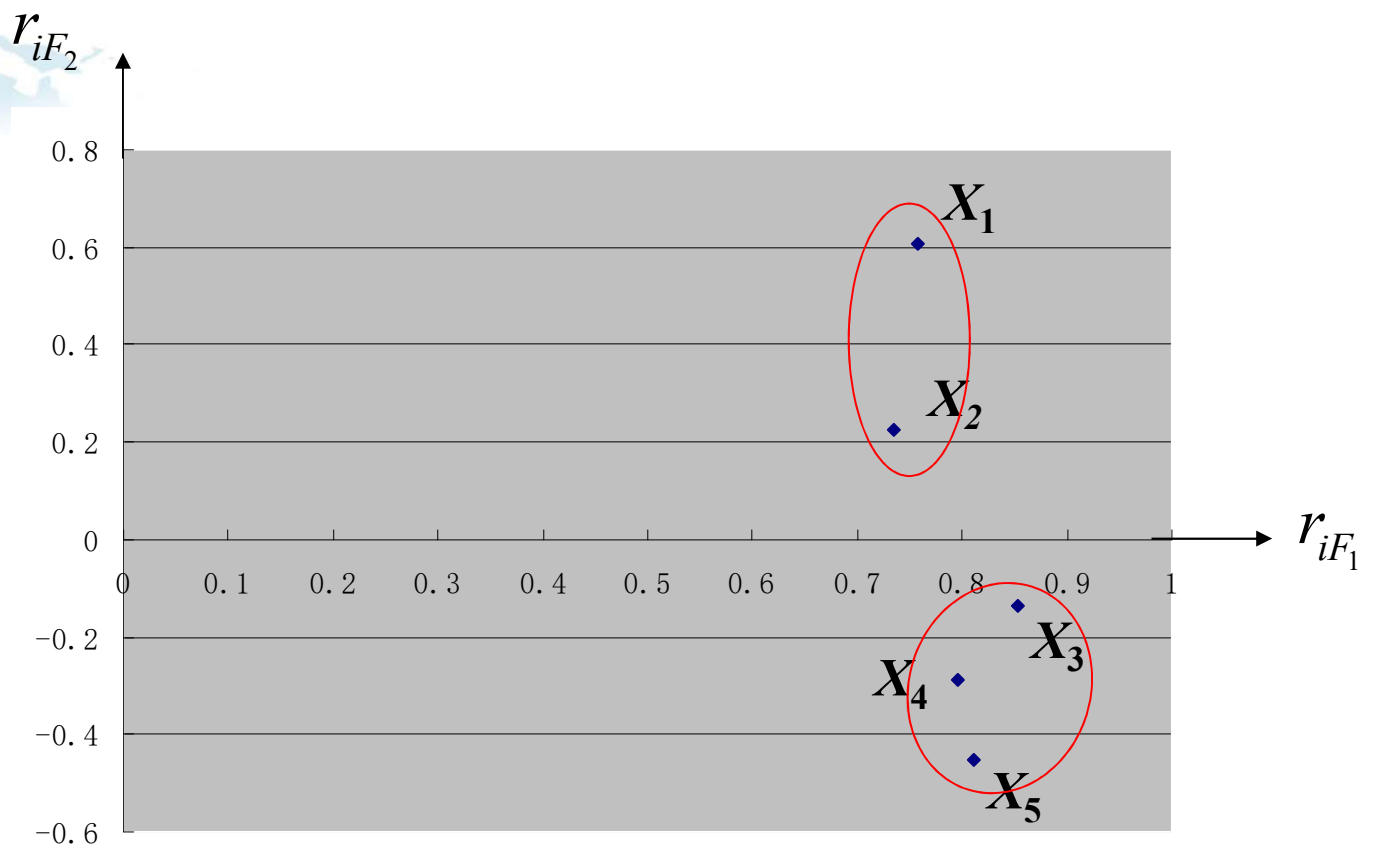


图解变量（对变量分类）

主成分分析后，若能以两个主成分代表原变量大部分的信息，则对应每个原变量 X_i ，只剩下 $\rho(\boldsymbol{F}_1, \boldsymbol{X}_i)$和$\rho(\boldsymbol{F}_2, \boldsymbol{X}_i)$。

以 $\rho(\boldsymbol{F}_1, X_i)$为横轴，$\rho(\boldsymbol{F}_2, \boldsymbol{X}_i)$为纵轴，建立直角坐标系。然后以为$\rho(\boldsymbol{F}_1, \boldsymbol{X}_i)$ 横坐标，以$\rho(\boldsymbol{F}_2, \boldsymbol{X}_i)$ 为纵坐标，在坐标系中描出各变量对应的点。

图解变量（对变量分类）


r_{iF_2}
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
r_{iF_1}
X_1
X_2
X_3
X_4
X_5

主成分分析用于综合评估

第一种方法，通过主成分分析得到综合指标

$$\boldsymbol{F}_1 = a_{11}\boldsymbol{X}_1 + a_{21}\boldsymbol{X}_2 + \cdots + a_{p1}\boldsymbol{X}_p$$

利用 $\boldsymbol{F}_1$作为评估指标，根据$\boldsymbol{F}_1$得分对样本点进行排序比较。但有**两个前提条件**：

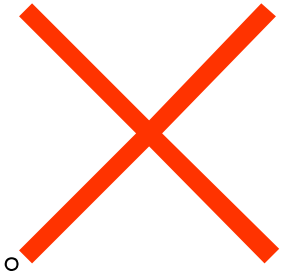
1. $\boldsymbol{F}_1$与全体原变量都正相关，即 $r_{iF_1} = a_{i1}\sqrt{\lambda_1} > 0$ (i=1,2,…,p）。

2. 各a_{i1}在各变量的分布比较均匀。

3. $\boldsymbol{F}_1$的方差贡献率较大。

第二种方法，通过主成分分析，取**前面m个**主成分 $\boldsymbol{F_1, F_2, \cdots, F_m}$ 以每个主成分 F_i 的方差贡献率 $\alpha_i = \dfrac{\lambda_i}{\sum_{i=1}^{p} \lambda_i}$ 为权，构造**综合评价函数**

$$\boldsymbol{F = \alpha_1 F_1 + \alpha_2 F_2 + \cdots \alpha_m F_m}$$



按F值的大小对样品进行排序比较或分类。

注意：实际上，这一方法不合理，$\boldsymbol{F_i}(i = 2, \cdots, m)$的含义违背了综合评价的本意。

主成分回归方法

$$\boldsymbol{F}_1 = a_{11}\boldsymbol{X}_1 + a_{21}\boldsymbol{X}_2 + \cdots + a_{p1}\boldsymbol{X}_p$$
$$\boldsymbol{F}_2 = a_{12}\boldsymbol{X}_1 + a_{22}\boldsymbol{X}_2 + \cdots + a_{p2}\boldsymbol{X}_p$$
$$\cdots\cdots$$
$$\boldsymbol{F}_p = a_{1p}\boldsymbol{X}_1 + a_{2p}\boldsymbol{X}_2 + \cdots + a_{pp}\boldsymbol{X}_p$$

主成分回归：$\boldsymbol{Y}_i = \gamma_1\boldsymbol{F}_{i1} + \gamma_2\boldsymbol{F}_{i2} + \cdots + \gamma_m\boldsymbol{F}_{im} + \varepsilon_i$

$$\hat{\boldsymbol{Y}} = \hat{\gamma}_1\boldsymbol{F}_1 + \hat{\gamma}_2\boldsymbol{F}_2 + \cdots + \hat{\gamma}_{\mathrm{m}}\boldsymbol{F}_m$$

$$\hat{\boldsymbol{Y}} = \hat{f}(\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_m)$$

由主成分分析法构造回归模型。即把各主成分作为新自变量代替原来自变量x做回归分析。

表 5-9　指标体系表

一级指标	二级指标	符号	单位
宏观经济环境	人均地区生产总值	X_1	元
	第三产业产值所占比重	X_2	%
	人均进出口总额	X_3	美元
人民生活	人均可支配收入	X_4	元
	人均地方财政收入	X_5	元
科技信息	每万从业人员有效发明专利数	X_6	项
	人均研发经费支出	X_7	万元
	互联网普及率	X_8	%
教育医疗	每千人口医生数	X_9	人
	人均财政性教育经费支出	X_{10}	元

表 5-10　指标数据表(1)

地区	人均地区生产总值	第三产业产值所占比重	人均进出口额	人均可支配收入	人均地方财政收入
北京	129 679.90	73.51	14 911.10	57 229.80	25 138.87
天津	119 616.00	54.42	7 253.69	37 022.30	14 898.55
河北	45 146.16	39.16	662.37	21 484.10	4 291.91
山西	41 883.53	46.60	463.80	20 420.00	5 035.71
内蒙古	63 615.89	49.31	549.62	26 212.20	6 731.47
辽宁	53 627.33	48.99	2 276.26	27 835.40	5 481.50
吉林	54 915.68	41.98	682.37	21 368.30	4 449.65
黑龙江	41 930.84	52.29	496.44	21 205.80	3 278.25
上海	127 315.60	64.19	19 690.65	58 988.00	27 606.29
江苏	107 044.90	45.06	7 362.31	35 024.10	10 186.60
浙江	91 591.79	46.54	6 680.22	42 045.70	10 269.49
安徽	43 096.58	37.08	857.55	21 863.30	4 486.16
福建	82 300.49	38.39	4 373.05	30 047.70	7 183.64
江西	43 196.81	38.81	962.14	22 031.40	4 851.76
山东	72 544.35	43.71	2 629.02	26 929.90	6 091.09
河南	46 467.35	37.95	811.91	20 170.00	3 553.63
湖北	60 048.37	40.45	784.65	23 757.20	5 497.94
湖南	49 316.01	43.16	525.36	23 102.70	4 011.59
广东	80 412.86	46.88	9 011.37	33 003.30	10 147.70

【输出 5-1】

相关矩阵											
		x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
x1	人均地区生产总值	1.0000	0.6797	0.8953	0.9324	0.8920	0.6742	0.9123	0.8087	0.6655	0.6587
x2	第三产业产值所占比重	0.6797	1.0000	0.7599	0.7891	0.8298	0.5998	0.8104	0.6238	0.7212	0.6668
x3	人均进出口额	0.8953	0.7599	1.0000	0.9617	0.9614	0.7121	0.9005	0.8023	0.5875	0.7139
x4	人均可支配收入	0.9324	0.7891	0.9617	1.0000	0.9536	0.7102	0.9290	0.8247	0.7128	0.6948
x5	人均地方财政收入	0.8920	0.8298	0.9614	0.9536	1.0000	0.7232	0.9287	0.7624	0.6717	0.8028
x6	每万从业人员有效发明专利数	0.6742	0.5998	0.7121	0.7102	0.7232	1.0000	0.8042	0.5813	0.5061	0.5578
x7	人均R&D经费支出	0.9123	0.8104	0.9005	0.9290	0.9287	0.8042	1.0000	0.7640	0.7663	0.7401
x8	互联网普及率	0.8087	0.6238	0.8023	0.8247	0.7624	0.5813	0.7640	1.0000	0.6688	0.6030
x9	每千人口医生数	0.6655	0.7212	0.5875	0.7128	0.6717	0.5061	0.7663	0.6688	1.0000	0.6555
x10	人均财政性教育经费支出	0.6587	0.6668	0.7139	0.6948	0.8028	0.5578	0.7401	0.6030	0.6555	1.0000

【输出 5-2】

相关矩阵的特征值				
	特征值	差分	比例	累积
1	7.84106585	7.24417795	0.7841	0.7841
2	0.59688790	0.12197055	0.0597	0.8438
3	0.47491735	0.07670241	0.0475	0.8913
4	0.39821494	0.06627548	0.0398	0.9311
5	0.33193947	0.12459535	0.0332	0.9643
6	0.20734412	0.12809405	0.0207	0.9850
7	0.07925007	0.04317261	0.0079	0.9930
8	0.03607746	0.01653377	0.0036	0.9966
9	0.01954368	0.00478451	0.0020	0.9985
10	0.01475917		0.0015	1.0000

【输出 5-3】

特征向量				
		Prin1	Prin2	Prin3
x1	人均地区生产总值	0.330116	-.213293	-.279443
x2	第三产业产值所占比重	0.302123	0.326245	0.235686
x3	人均进出口额	0.337460	-.272227	-.088823
x4	人均可支配收入	0.345657	-.136500	-.171382
x5	人均地方财政收入	0.346039	-.075903	0.070883
x6	每万从业人员有效发明专利数	0.276840	-.415135	0.600272
x7	人均R&D经费支出	0.346458	-.047502	0.110696
x8	互联网普及率	0.301035	-.067747	-.581773
x9	每千人口医生数	0.279008	0.660091	-.119614
x10	人均财政性教育经费支出	0.285487	0.366969	0.314214

【输出 5-4】

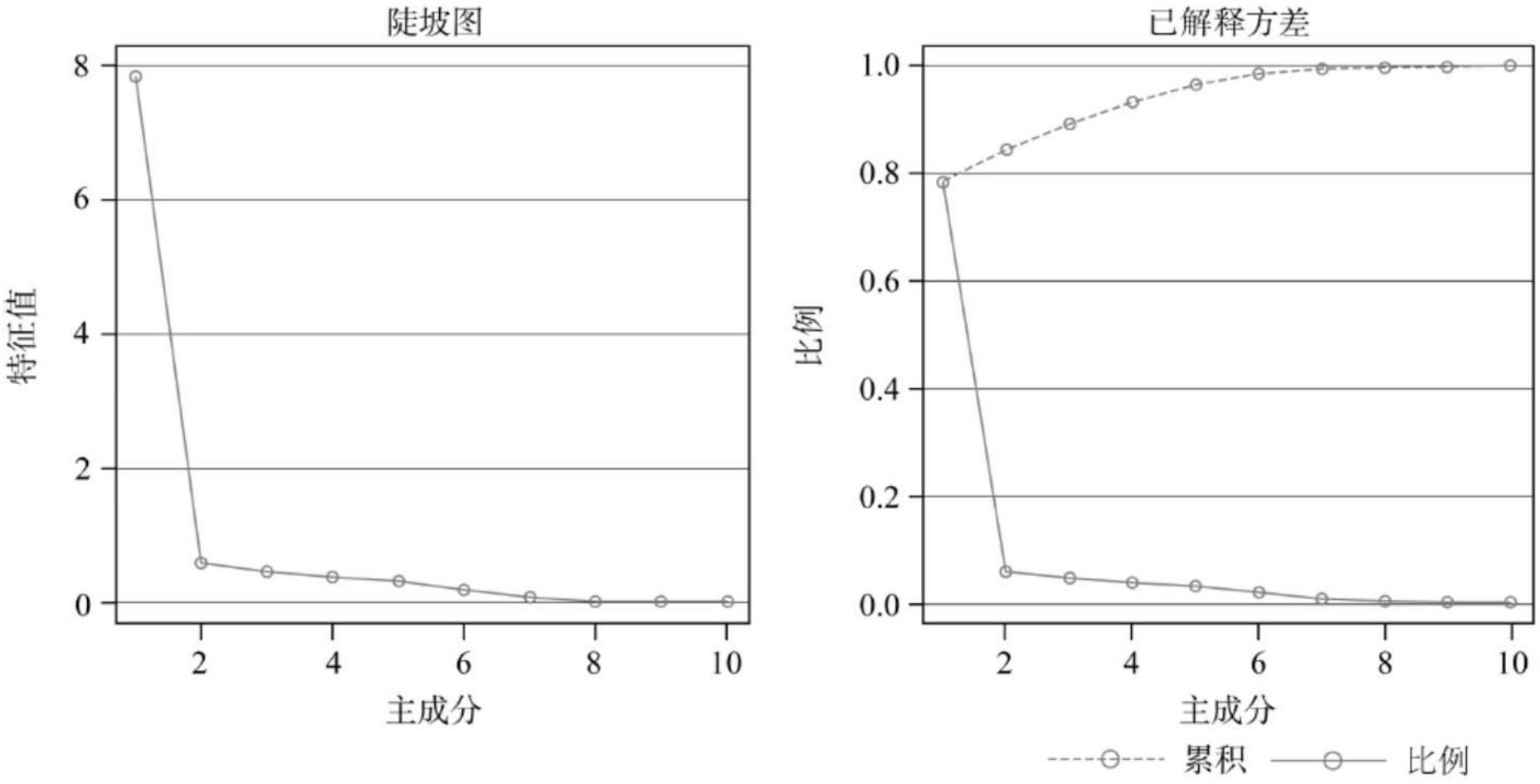
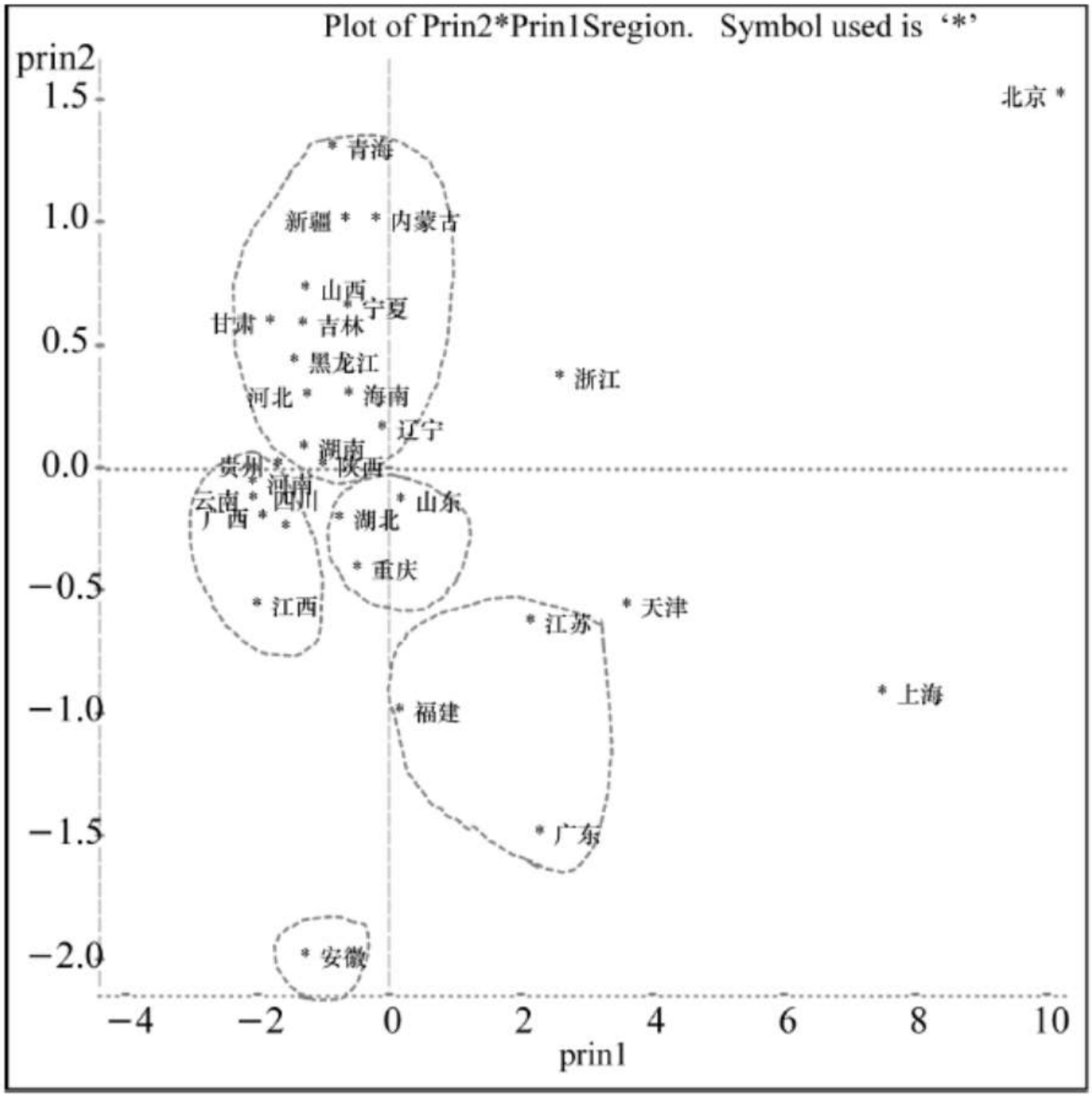
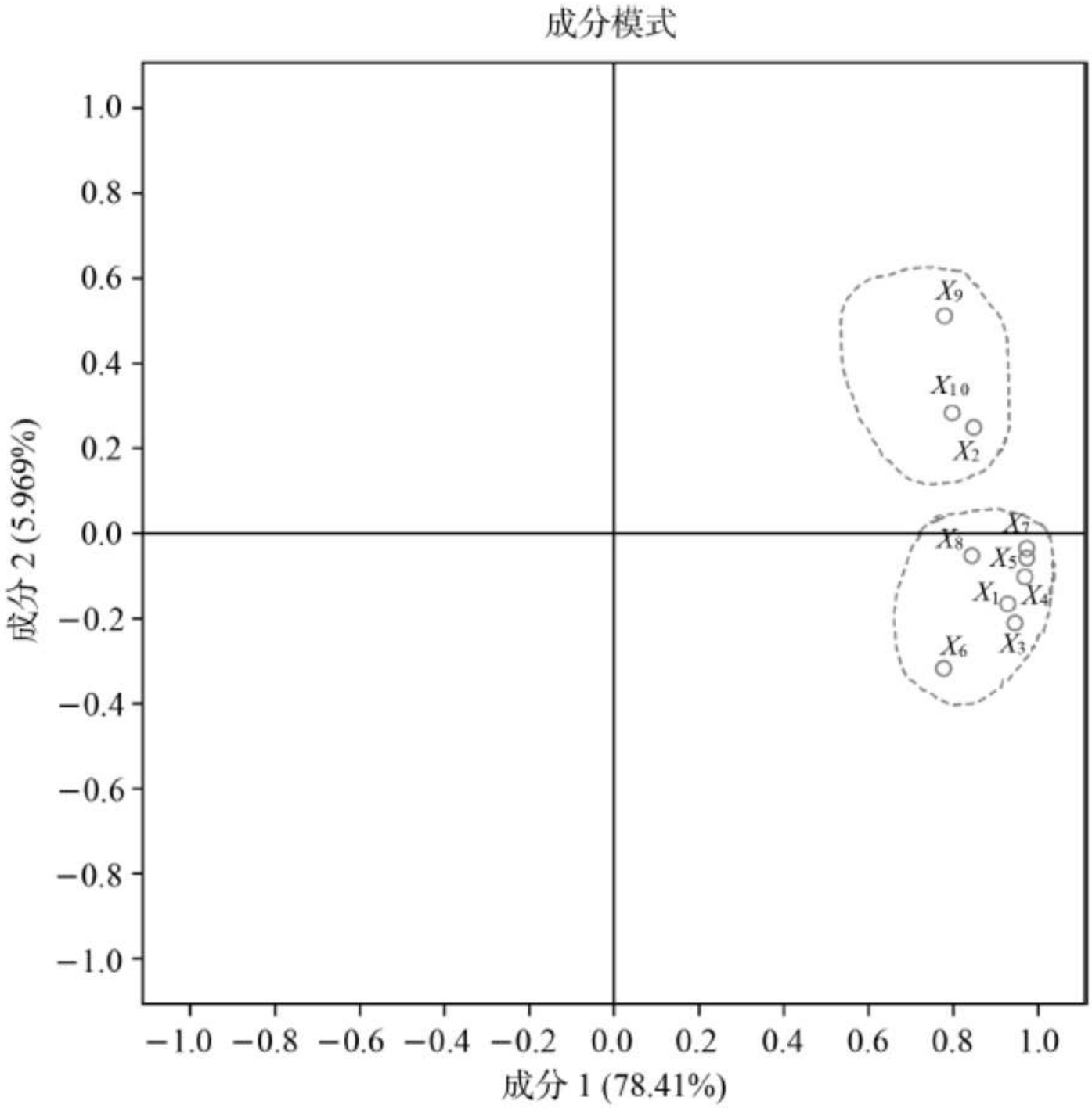

陡坡图
特征值
8
6
4
2
0
2
4
6
8
10
主成分
已解释方差
比例
1.0
0.8
0.6
0.4
0.2
0.0
2
4
6
8
10
主成分
累积
比例

表 5-12　主成分得分

地区	第一主成分得分	第二主成分得分	第一主成分得分排名
北京	9.975 9	1.464 3	1
上海	7.417 3	−0.938 7	2
天津	3.568 3	−0.566 6	3
浙江	2.486 2	0.353 7	4
广东	2.279 4	−1.527 2	5
江苏	2.135 3	−0.654 9	6
福建	0.197 0	−0.977 7	7
山东	0.109 6	−0.172 5	8
辽宁	−0.176 1	0.151 2	9


Plot of Prin2*Prin1Sregion. Symbol used is '*'
prin2
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−4
−2
0
2
4
6
8
10
prin1
北京
青海
新疆
内蒙古
山西
宁夏
甘肃
吉林
黑龙江
浙江
河北
海南
辽宁
湖南
贵州
陕西
河南
云南
四川
山东
广西
湖北
重庆
江西
天津
江苏
上海
福建
广东
安徽


成分模式
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
成分 2 (5.969%)
−1.0 −0.8 −0.6 −0.4 −0.2 0.0 0.2 0.4 0.6 0.8 1.0
成分 1 (78.41%)
X9
X10
X2
X7
X8
X5
X1
X4
X6
X3

python程序例题：

这个例子的数据很简单，就是一张照片，图片数据由3个3264*2448的矩阵组成，这三个矩阵分别代表红色绿色和蓝色，即，照片中有3264*2448=7990272个点，当图片像素很大时，传输和存储都不太方便，这时需要降低像素，主成分分析的降维功能可以用来做这件事情，希望在降维之后尽量保持图片的可识别性。

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import scipy
...
import imageio
#定义图像压缩主成分分析函数
def comp_2d(image_2d, numpc=100):
    cov_mat = (image_2d.T - np.mean(image_2d, axis=1)).T
    eig_val, eig_vec = np.linalg.eigh(np.cov(cov_mat))
    p = np.size(eig_vec, axis=1)
    idx = np.argsort(eig_val)
    idx = idx[::-1]
    eig_vec = eig_vec[:, idx]
    eig_val = eig_val[idx]
    if numpc < p or numpc > 0:
        eig_vec = eig_vec[:, range(numpc)]
    score = np.dot(eig_vec.T, cov_mat)
    recon = (np.dot(eig_vec, score).T + np.mean(image_2d,
axis=1)).T
    recon_img_mat = np.uint8(np.absolute(recon))
    return recon_img_mat
```

```
#定义使用上述函数和画图函数
def pltPCA(a, numpc):
    a_np = np.array(a)
    a_r = a_np[:, :, 0]
    a_g = a_np[:, :, 1]
    a_b = a_np[:, :, 2]
    a_r_recon, a_g_recon, a_b_recon = comp_2d(a_r, numpc), \
                                comp_2d(a_g, numpc), comp_2d(a_b, numpc)
    recon_color_img = np.dstack((a_r_recon, a_g_recon, a_b_recon))
    recon_color_img = Image.fromarray(recon_color_img)
    fig = plt.figure(figsize=(4, 3), dpi=72)
    ax = fig.add_axes([0.0, 0.0, 1.0, 1.0], frameon=False, aspect=1)
    ax.set_xticks([])
    ax.set_yticks([])
    imshow(recon_color_img)

#分别取10个和100个主成分
a = imageio.imread("1.jpeg")
pltPCA(a, 10)
pltPCA(a, 100)
plt.show()
```