

# 主成分分析

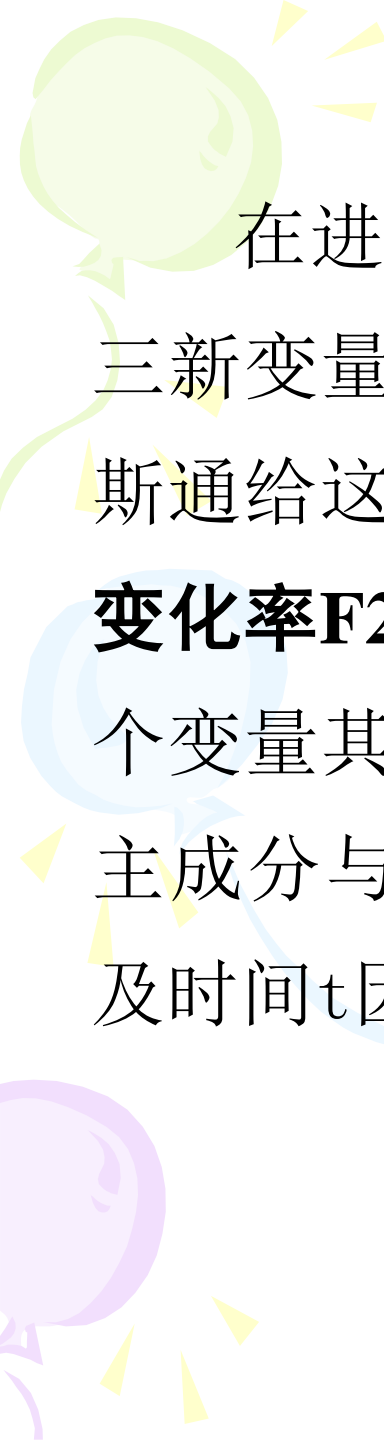
**Principal component analysis**

- 
- 主成分分析的基本思想
  - 主成分的计算
  - 主成分的性质
  - 主成分分析的应用
  - 主成分回归

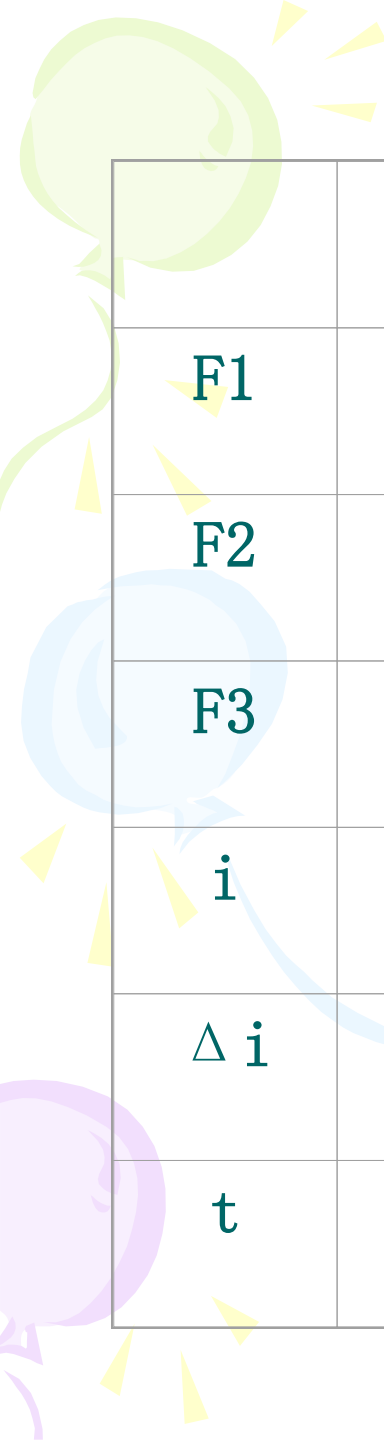


## § 1 基本思想

一项十分著名的工作是美国统计学家斯通(stone)在1947年关于国民经济的研究。他曾利用美国1929-1938年各年的数据，得到了17个反映国民收入与支出的变量要素，例如雇主补贴、消费资料和生产资料、纯公共支出、净增库存、股息、利息外贸平衡等等。



在进行主成分分析后，竟以97.4%的精度，用三新变量就取代了原17个变量。根据经济学知识，斯通给这三个新变量分别命名为**总收入F1**、**总收入变化率F2**和**经济发展趋势F3**。更有意思的是，这三个变量其实都是可以直接测量的。斯通将他得到的主成分与实际测量的总收入 $I$ 、总收入变化率 $\Delta I$ 以及时间 $t$ 因素做相关分析，得到下表：



	F1	F2	F3	i	i	t
F1	1					
F2	0	1				
F3	0	0	1			
i	0.995	-0.041	0.057	1		
$\Delta$ i	-0.056	0.948	-0.124	-0.102	1	
t	-0.369	-0.282	-0.836	-0.414	-0.112	1



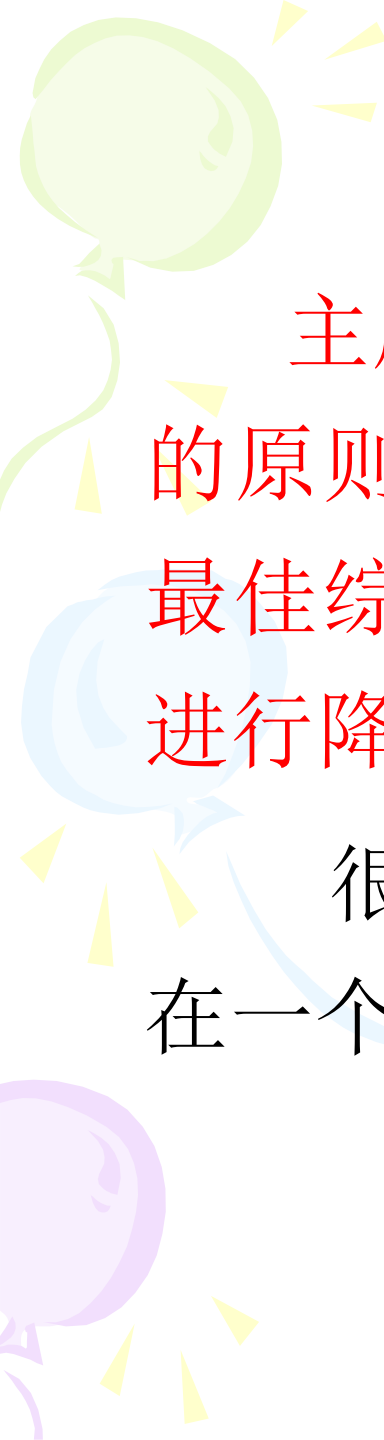
## 主成分分析的基本思想

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。



主成分分析是把各变量之间互相关联的复杂关系进行简化分析的方法。

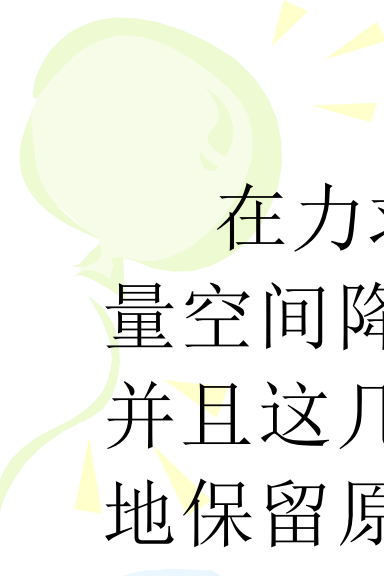
在社会经济的研究中，为了全面系统的分析和研究问题，必须考虑许多经济指标，这些指标能从不同的侧面反映我们所研究的对象的特征，但在某种程度上存在信息的重叠，具有一定的相关性。

A decorative graphic on the left side of the slide featuring a green balloon at the top, a blue balloon in the middle, and a purple balloon at the bottom, all connected by a green streamer. Yellow triangular streamers are scattered around the balloons.


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。

很显然，识辨系统在一个低维空间要比在一个高维空间容易得多。





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？
- 

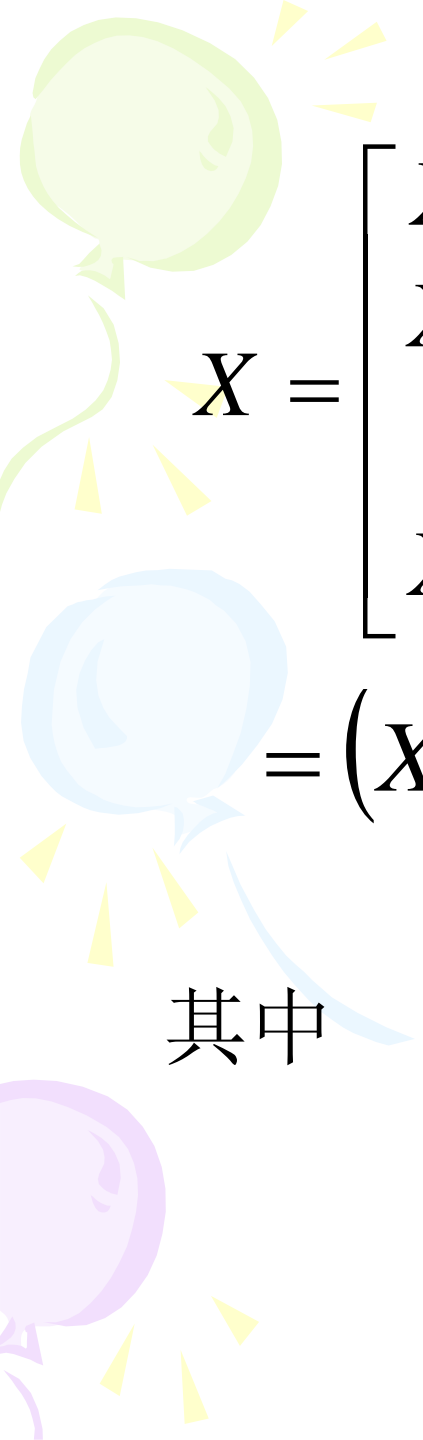
## § 2 数学模型与几何解释

假设我们所讨论的实际问题中，有 $p$ 个指标，记为  $X_1, X_2, \dots, X_p$ ,

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：

$$F_1, F_2, \dots, F_m \quad (m < p)$$

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。


$$X = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}$$

$$= (X_1 \quad X_2 \quad \cdots \quad X_p)$$

其中

$$X_i = \begin{pmatrix} X_{1i} \\ X_{2i} \\ \vdots \\ X_{ni} \end{pmatrix}$$

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 $F_i$ 。

信息量

$$F_1 = a_{11}X_1 + a_{21}X_2 + \cdots + a_{p1}X_p$$

$$F_2 = a_{12}X_1 + a_{22}X_2 + \cdots + a_{p2}X_p$$

.....

$$F_p = a_{1p}X_1 + a_{2p}X_2 + \cdots + a_{pp}X_p$$



满足如下的条件：

每个主成分的系数平方和为1。即

$$a_{1i}^2 + a_{2i}^2 + \cdots + a_{pi}^2 = 1$$

主成分之间相互独立，即无重叠的信息。即

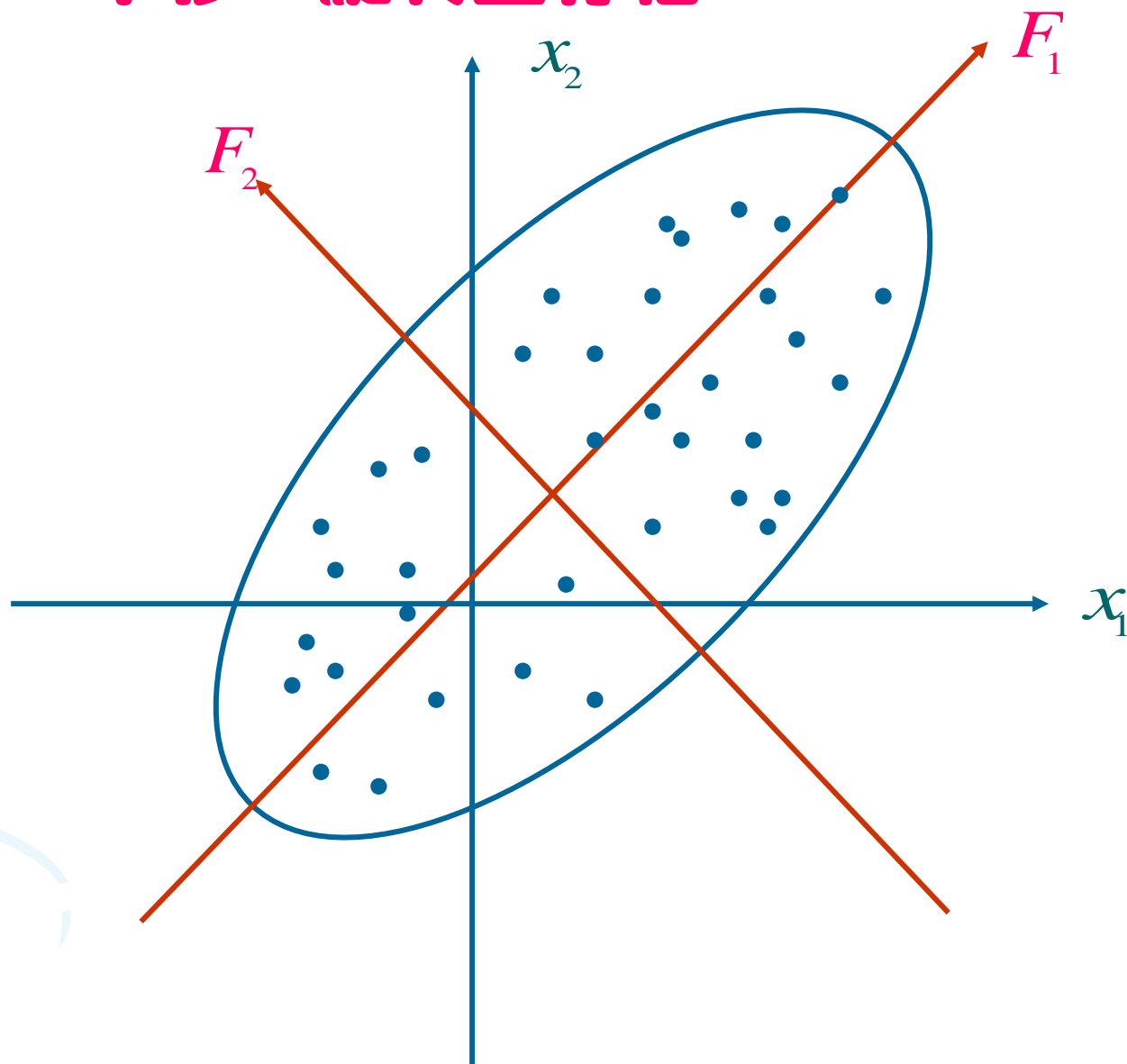
$$Cov(F_i, F_j) = 0, \quad i \neq j, \quad i, j = 1, 2, \dots, p$$

主成分的方差依次递减，重要性依次递减，即

$$Var(F_1) \geq Var(F_2) \geq \cdots \geq Var(F_p)$$

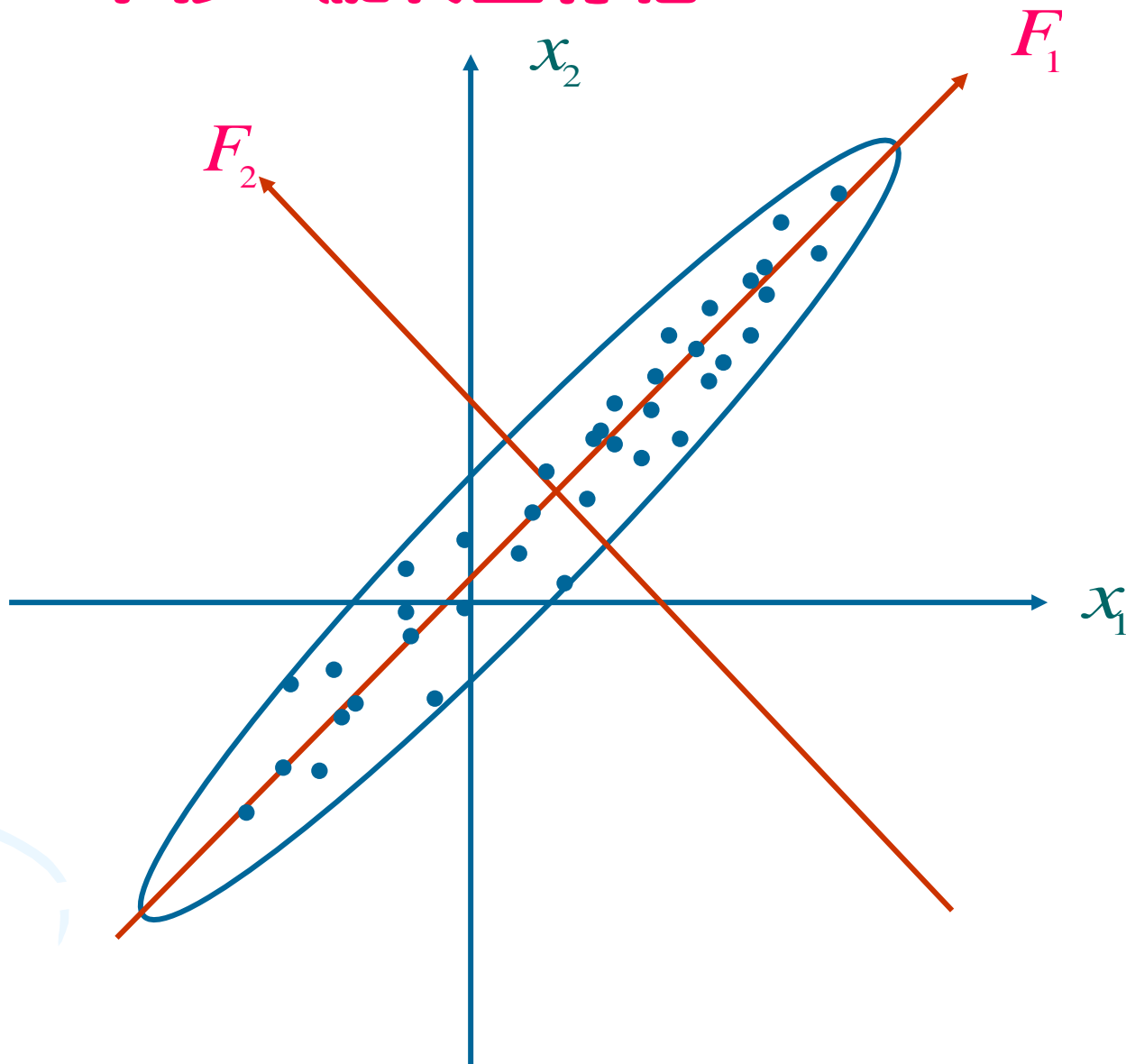
# 平移、旋转坐标轴

## 主成分分析的几何解释



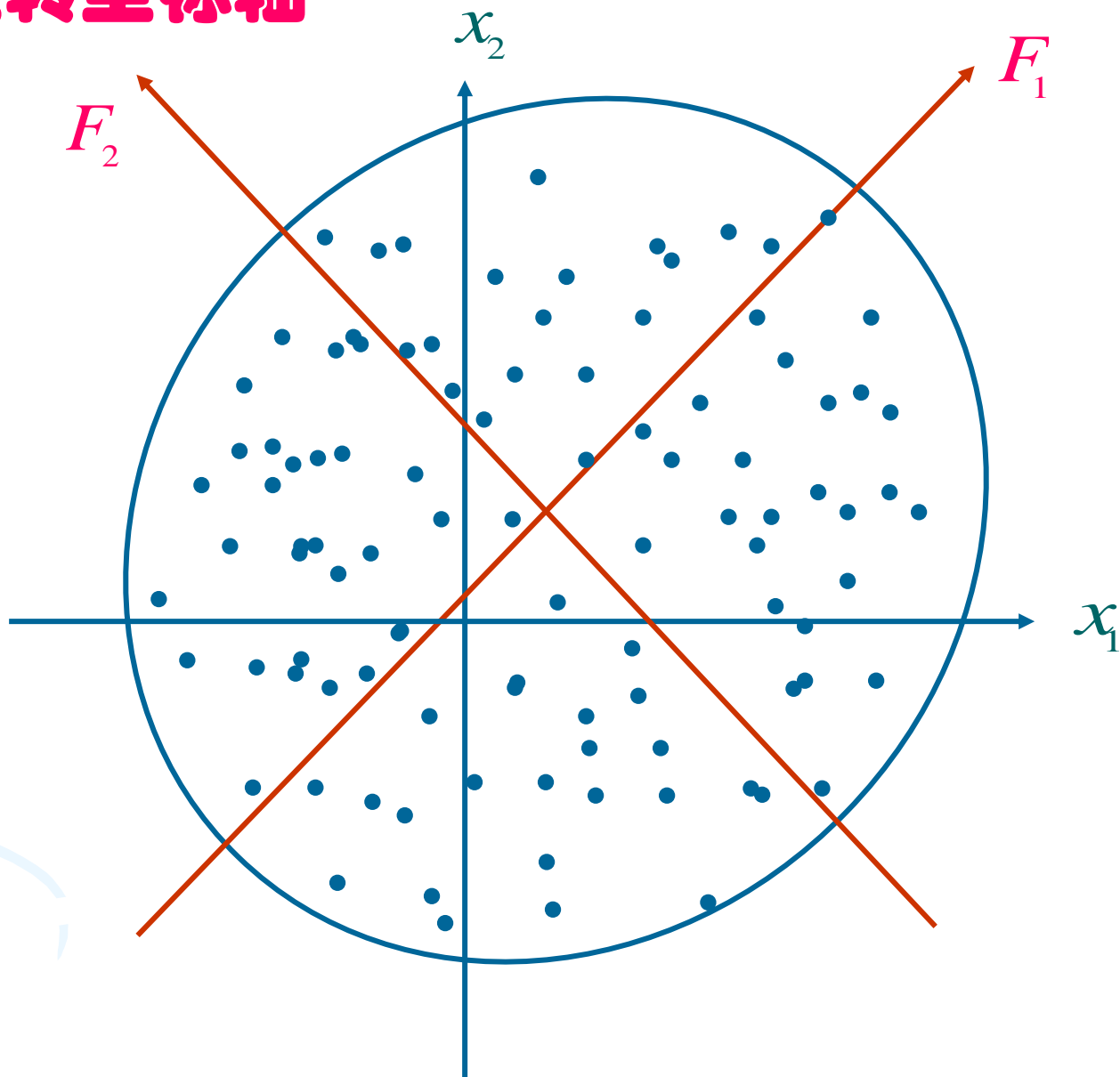
# 平移、旋转坐标轴

## 主成分分析的几何解释



# 平移、旋转坐标轴

## 主成分分析的几何解释





## § 3 主成分的推导

### 一、两个线性代数的结论

1、若 $\mathbf{A}$ 是 $p$ 阶实对称阵，则一定可以找到正交阵 $\mathbf{U}$ ，使

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

其中 $\lambda_i, i = 1.2.\cdots p$  是 $\mathbf{A}$ 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  $\mathbf{u}_1, \dots, \mathbf{u}_p$

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

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

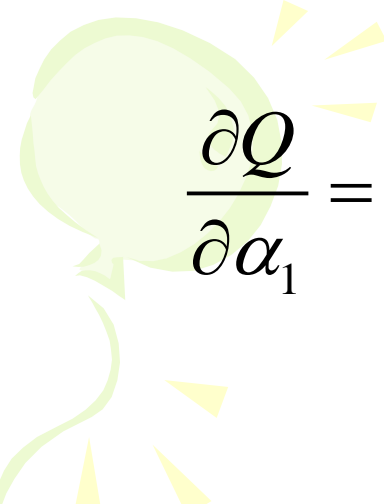
$$F_1 = a_{11}x_1 + \cdots + a_{p1}x_p = \alpha_1'X$$

$$\text{Var}(F_1) = \text{Var}(\alpha_1'X) = \alpha_1'\Sigma\alpha_1$$

寻找合适的单位向量  $\alpha_1$ ，使  $F_1$  的方差最大。

$$Q = \alpha_1'\Sigma\alpha_1 - \lambda(\alpha_1'\alpha_1 - 1) \rightarrow \max$$

$$\frac{\partial Q}{\partial \alpha_1} = 2\Sigma\alpha_1 - 2\lambda\alpha_1 = 0$$


$$\frac{\partial Q}{\partial \alpha_1} = 2\Sigma\alpha_1 - 2\lambda\alpha_1 = 0$$

$$\Sigma\alpha_1 = \lambda\alpha_1$$



表明： $\lambda$  应为  $\Sigma$  的特征值，而  $\alpha_1$  为与  $\lambda$  对应的单位特征向量。


$$Var(F_1) = \alpha_1' \Sigma \alpha_1 = \alpha_1' \lambda \alpha_1 = \lambda$$



可见  $\lambda$  应取  $\Sigma$  的最大特征根。

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

$$F_2 = a_{12}x_1 + \cdots + a_{p2}x_p = \alpha_2'X$$

寻找合适的单位向量  $\alpha_2$ ，使  $F_2$  的方差最大。

$$\text{Var}(F_2) = \text{Var}(\alpha_2'X) = \alpha_2'\Sigma\alpha_2 \rightarrow \max$$

$$\text{cov}(F_1, F_2) = \alpha_1'\Sigma\alpha_2 = \underline{0} = \alpha_2'\Sigma\alpha_1 = \alpha_2'\lambda\alpha_1 = \lambda\underline{\alpha_2'\alpha_1}$$

$$Q = \alpha_2'\Sigma\alpha_2 - \lambda(\alpha_2'\alpha_2 - 1) - 2\rho\alpha_2'\alpha_1$$

$$\frac{\partial Q}{\partial \alpha_2} = 2\Sigma\alpha_2 - 2\lambda\alpha_2 - 2\rho\alpha_1 = 0$$

$$\Sigma\alpha_2 - \lambda\alpha_2 - \rho\alpha_1 = 0$$

---

用  $\alpha_1'$  左乘上式,  $\underbrace{\alpha_1'\Sigma\alpha_2}_0 - \lambda\underbrace{\alpha_1'\alpha_2}_0 - \rho\alpha_1'\alpha_1 = 0$

因而  $\rho = 0$

$$\Sigma\alpha_2 - \lambda\alpha_2 = 0$$

表明:  $\lambda$  应为  $\Sigma$  的特征值, 而  $\alpha_2$  为与  $\lambda$  对应的单位特征向量。

$$\text{Var}(F_2) = \alpha_2'\Sigma\alpha_2 = \alpha_2'\lambda\alpha_2 = \lambda$$

这时  $\lambda$  不能再取  $\lambda_1$  了, 应取  $\lambda_2$ 。


$$F_1 = a_{11}X_1 + a_{21}X_2 + \cdots + a_{p1}X_p$$

$$F_2 = a_{12}X_1 + a_{22}X_2 + \cdots + a_{p2}X_p$$

.....

$$F_p = a_{1p}X_1 + a_{2p}X_2 + \cdots + a_{pp}X_p$$

**问题的答案**是：X的协方差矩阵S 的最大特征根  $\lambda_1$  所对应的单位特征向量即为  $(a_{11}, a_{21}, \cdots, a_{p1})$ 。并且  $\lambda_1$  就是 $F_1$ 的方差。

X的协方差矩阵S 的第二大特征根  $\lambda_2$  所对应的单位特征向量即为  $(a_{12}, a_{22}, \cdots, a_{p2})$ 。并且  $\lambda_2$  就是 $F_2$ 的方差。



写作矩阵形式：

$$\mathbf{F} = \mathbf{U}'\mathbf{X}$$

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

$$\mathbf{X} = (X_1, X_2, \dots, X_p)'$$



# 主成分的计算

先讨论二维情形

$$X = \begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \\ \vdots & \vdots \\ X_{n1} & X_{n2} \end{pmatrix} \triangleq (X_1 \quad X_2)$$

求第一主成分 $F_1$ 和 $F_2$ 。

A decorative graphic in the top-left corner of the slide features three balloons: a green one at the top, a light blue one in the middle, and a purple one at the bottom. Each balloon has a string and is surrounded by several small, yellow, triangular shapes that resemble sunbeams or confetti.

观察图，我们已经把主成分**F1**和**F2** 的坐标原点放在平均值  $(\bar{x}_1, \bar{x}_2)$  所在处，从而使得**F1**和**F2** 成为中心化的变量，即**F1**和**F2** 的样本均值都为零。

因此 $F_1$ 可以表示为

$$F_1 = a_{11}(x_1 - \bar{x}_1) + a_{21}(x_2 - \bar{x}_2)$$

**关键**是，寻找合适的单位向量  $(a_{11}, a_{21})$ ，使 $F_1$ 的方差最大。

**问题的答案**是：X的协方差矩阵S 的最大特征根  $\lambda_1$  所对应的单位特征向量即为  $(a_{11}, a_{21})$ 。并且  $\lambda_1$  就是 $F_1$ 的方差。



同样， $F_2$ 可以表示为  $F_2 = a_{12}(x_1 - \bar{x}_1) + a_{22}(x_2 - \bar{x}_2)$

寻找合适的单位向量  $(a_{12}, a_{22})$ ，使 $F_2$ 与 $F_1$ 独立，且使 $F_2$ 的方差（除 $F_1$ 之外）最大。

**问题的答案**是：X的协方差矩阵S的第二大特征根  $\lambda_2$  所对应的单位特征向量即为  $(a_{12}, a_{22})$ 。并且  $\lambda_2$  就是 $F_2$ 的方差。

# 求解主成分的步骤:

1. 求样本均值  $\bar{X} = (\bar{x}_1, \bar{x}_2)$  和样本协方差矩阵  $S$ ;

2. 求  $S$  的特征根

求解特征方程  $|S - \lambda I| = 0$  , 其中  $I$  是单位矩阵,  
解得2个特征根  $\lambda_1, \lambda_2 (\lambda_1 \geq \lambda_2)$

3. 求特征根所对应的单位特征向量

4. 写出主成分的表达式

**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

$$\bar{X} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \end{pmatrix} = \begin{pmatrix} 71.25 \\ 67.5 \end{pmatrix} \quad S = \begin{pmatrix} 323.4 & \\ 103.1 & 187.5 \end{pmatrix}$$

## 2. 求解特征方程 $|s - \lambda I| = 0$

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

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

$$(323.4 - \lambda)(187.5 - \lambda) - 103.1^2 = 0$$

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

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

### 3.求特征值所对应的单位特征向量

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

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

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

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

$$\text{解得 } (a_{11}, a_{21}) = (0.88, 0.47)$$

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

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

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

$$\text{解得: } (a_{12}, a_{22}) = (-0.47, 0.88)$$



## 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，表示该生成成绩的好坏。

第二主成分 $F_2$ 表示学生两科成绩的均衡性

## 6. 比较主成分重要性

第一主成分 $F_1$ 的方差为 $\lambda_1 = 378.9$

方差贡献率

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{Var(F_1)}{Var(F_1) + Var(F_2)} = \frac{378.9}{378.9 + 132} = 74.16\%$$

第二主成分 $F_2$ 的方差为 $\lambda_2 = 132$

$$\text{方差贡献率为} \frac{\lambda_2}{\lambda_1 + \lambda_2} = \frac{132}{378.9 + 132} = 25.84\%$$

主成分 $F_1$ 和 $F_2$ 的方差总和为 $\lambda_1 + \lambda_2 = 378.9 + 132 = 510.9$

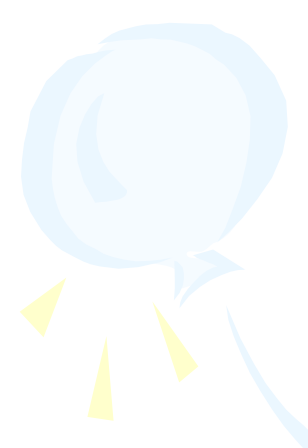
原变量 $x_1$ 和 $x_2$ 的方差总和为 $s_{11} + s_{22} = 323.4 + 187.5 = 510.9$

总方差保持不变

例2 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)。

5门课程的考试成绩

Mechanics	Vectors	Algebra	Analysis	Statistics
77	82	67	67	81
63	78	80	70	81
75	73	71	66	81
55	72	63	70	68
63	63	65	70	63



简单统计量					
	x1	x2	x3	x4	x5
均值	51.5795	57.7273	57.4205	57.8409	56.5227
StD	11.9431	10.951	7.671	8.09695	12.5542

协方差矩阵						
		x1	x2	x3	x4	x5
x1	力学(闭)	142.637	75.4702	37.5351	42.0473	52.774
x2	物理(闭)	75.4702	119.925	30.7137	26.094	35.5005
x3	代数(开)	37.5351	30.7137	58.8442	39.5964	59.9846
x4	分析(开)	42.0473	26.094	39.5964	65.5606	65.1876
x5	统计(开)	52.774	35.5005	59.9846	65.1876	157.609



经计算，得到5个主成分的表达式如下：

$$F_1 = 0.5327x_1 + 0.4208x_2 + 0.3225x_3 + 0.3422x_4 + 0.5639x_5 - 94.48$$

$$F_2 = 0.4846x_1 + 0.5405x_2 - 0.1735x_3 - 0.2210x_4 - 0.6278x_5 + 2.03$$

$$F_3 = -0.6727x_1 + 0.7197x_2 + 0.0809x_3 - 0.1005x_4 + 0.1133x_5 - 12.08$$

$$F_4 = -0.1538x_1 - 0.0373x_2 + 0.5600x_3 + 0.6216x_4 - 0.5244x_5 + 28.38$$

$$F_5 = 0.0715x_1 - 0.1073x_2 + 0.7387x_3 - 0.6615x_4 - 0.0085x_5 - 1.17$$

这5个主成分的**方差**分别为307.83、125.83、54.78、34.23和21.87。前两个主成分各自的贡献率和累积贡献率为

$$\frac{\lambda_1}{\sum_{i=1}^5 \lambda_i} = \frac{307.83}{544.58} = 56.53\%$$

$$\frac{\lambda_2}{\sum_{i=1}^5 \lambda_i} = \frac{125.83}{544.58} = 23.11\%$$

$$\frac{\lambda_1 + \lambda_2}{\sum_{i=1}^5 \lambda_i} = 56.53\% + 23.11\% = 79.63\%$$

在一般情况下，设有n个样品，每个样品观测p个指标，将原始数据排成如下矩阵：

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

1.求样本均值  $\bar{X} = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p)'$  和样本协方差矩阵  $S$ ;

2.求解特征方程  $|S - \lambda I| = 0$ , 其中  $I$  是单位矩阵

$$\begin{vmatrix} s_{11} - \lambda & s_{12} & \dots & s_{1p} \\ s_{21} & s_{22} - \lambda & \dots & s_{2p} \\ \dots & \dots & \dots & \dots \\ s_{p1} & s_{p2} & \dots & s_{pp} - \lambda \end{vmatrix} = 0$$


解得  $p$  个特征根  $\lambda_1, \lambda_2, \dots, \lambda_p$  ( $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ )

3. 求  $\lambda_k$  所对应的单位特征向量  $\alpha_k$  ( $k = 1, 2, \dots, p$ )

即需求解方程组  $(S - \lambda_k I)\alpha_k = 0$

其中  $\alpha_k = (a_{1k}, a_{2k}, \dots, a_{pk})'$





$$\begin{pmatrix} s_{11} - \lambda_k & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} - \lambda & \cdots & s_{2p} \\ \vdots & \vdots & & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} - \lambda \end{pmatrix} \begin{pmatrix} a_{1k} \\ a_{2k} \\ \vdots \\ a_{pk} \end{pmatrix} = 0$$

再加上单位向量的条件  $a_{1k}^2 + a_{2k}^2 + \dots + a_{pk}^2 = 1$

解得  $\alpha_k = (a_{1k}, a_{2k}, \dots, a_{pk})'$

#### 4. 写出主成分的表达式


$$F_k = a_{1k}(x_1 - \bar{x}_1) + a_{2k}(x_2 - \bar{x}_2) + \dots + a_{pk}(x_p - \bar{x}_p)$$

# 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:

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

且

$$\frac{\sum_{i=1}^m \lambda_i}{\sum_{i=1}^p \lambda_i} \geq 80 \sim 85\%$$

例3 设  $x = (x_1, x_2, x_3)'$  的协方差矩阵为

$$\Sigma = \begin{pmatrix} 16 & 2 & 30 \\ 2 & 1 & 4 \\ 30 & 4 & 100 \end{pmatrix}$$

经计算,  $\Sigma$  的特征值为  $\lambda_1 = 109.793, \lambda_2 = 6.469, \lambda_3 = 0.738$

相应的主成分分别为

$$y_1 = 0.305x_1 + 0.041x_2 + 0.951x_3$$

$$y_2 = 0.944x_1 + 0.1202x_2 - 0.308x_3$$

$$y_3 = -0.127x_1 + 0.992x_2 - 0.002x_3$$

第一主成分的方差贡献率为:  $\frac{\lambda_1}{\sum_{i=1}^3 \lambda_i} = \frac{109.783}{117} = 93.8\%$

The background features several large, flowing, abstract shapes in light green, light blue, and light purple. Interspersed among these are numerous small, yellow, four-pointed starburst or spark-like shapes, creating a festive and dynamic feel.

§ 4

R 型分析

# 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**。

# 主成分求解步骤

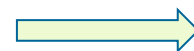
0. 对数据进行标准化变换;

1. 求样本协方差矩阵S;

2. 求特征根  $\lambda_k$

3. 求单位特征向量  $\alpha_k$

4. 写出主成分的表达式



1. 求相关  
系数矩阵R

$$F_k = a_{1k}(x_1 - \bar{x}_1) + a_{2k}(x_2 - \bar{x}_2) + \dots + a_{pk}(x_p - \bar{x}_p)$$

# § 5 主成分的性质

## 一、主成分的相关结构

1. 主成分 $F_k$ 的方差  $\lambda_k$

2. 主成分 $F_k$ 的方差贡献率为  $\frac{\lambda_k}{\sum_{k=1}^p \lambda_k}$

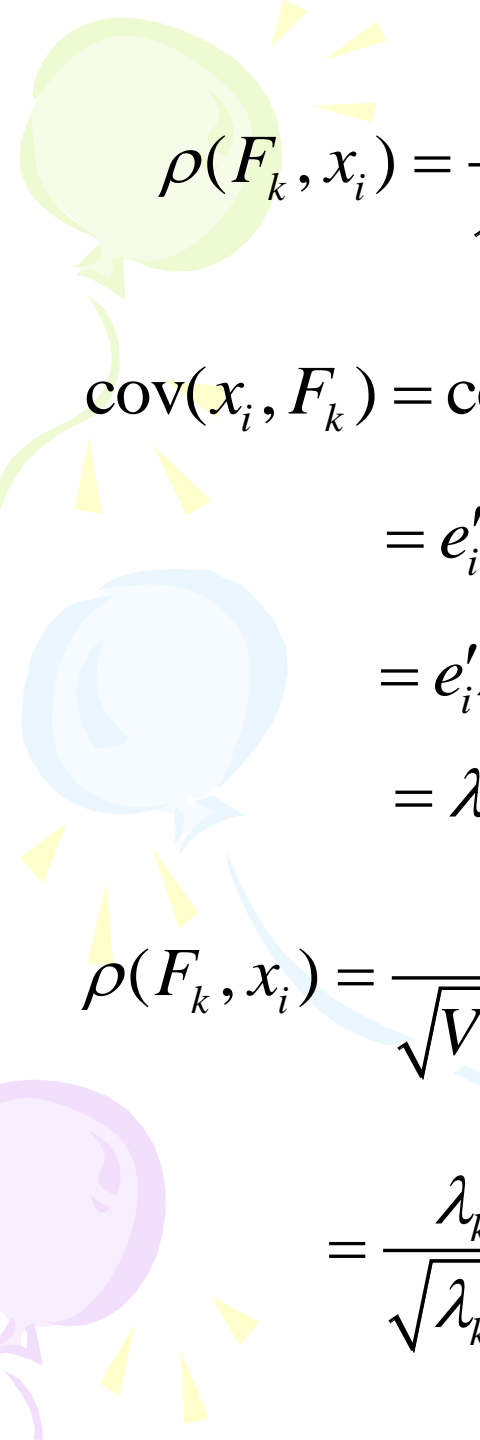
因子负荷量  
(因子载荷)

3. 主成分与每个变量之间的相关系数

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

证明

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


$$\rho(F_k, x_i) = \frac{\text{cov}(F_k, x_i)}{\sqrt{\text{Var}(F_k)\text{Var}(x_i)}}$$

$$\text{cov}(x_i, F_k) = \text{cov}(e_i' X, \alpha_k' X)$$

$$= e_i' \Sigma \alpha_k$$

$$= e_i' \lambda_k \alpha_k$$

$$= \lambda_k a_{ik}$$

$$\rho(F_k, x_i) = \frac{\text{cov}(F_k, x_i)}{\sqrt{\text{Var}(F_k)\text{Var}(x_i)}}$$

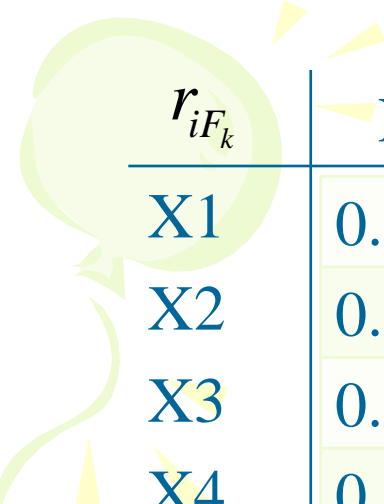
$$= \frac{\lambda_k a_{ik}}{\sqrt{\lambda_k} \sqrt{s_{ii}}} = a_{ik} \sqrt{\frac{\lambda_k}{s_{ii}}}$$

$$e_i' = (0, 0, \dots, 1, \dots, 0)$$

第  $i$  个分量为1,  
其余为0







$r_{iF_k}$	F1	F2	F3	F4	F5
X1	0.7827	0.4551	-0.4169	-0.0754	0.0280
X2	0.6741	0.5536	0.4864	-0.0200	-0.0458
X3	0.7376	-0.2537	0.0781	0.4274	0.4503
X4	0.7416	-0.3061	-0.0919	0.4494	-0.3820
X5	0.7881	-0.5610	0.0668	-0.2445	-0.0032



$r_{iF_k}^2$	F1	F2	F3	F4	F5
X1	0.6126	0.2072	0.1738	0.0057	0.0008
X2	0.4544	0.3065	0.2366	0.0004	0.0021
X3	0.5441	0.0643	0.0061	0.1827	0.2028
X4	0.5499	0.0937	0.0084	0.2020	0.1460
X5	0.6211	0.3147	0.0045	0.0598	0.0000

$r_{iF_k}^2$	F1	F2	F3	F4	F5
X1	0.6126	0.2072	0.1738	0.0057	0.0008
X2	0.4544	0.3065	0.2366	0.0004	0.0021
X3	0.5441	0.0643	0.0061	0.1827	0.2028
X4	0.5499	0.0937	0.0084	0.2020	0.1460
X5	0.6211	0.3147	0.0045	0.0598	0.0000

横行之和为1，从横行看，有  $\sum_{k=1}^p r_{iF_k}^2 = 1$

从纵向看

$$61.26\% \cdot s_{11} + 45.44\% \cdot s_{22} + 54.41\% \cdot s_{33} + 54.99\% \cdot s_{44} + 62.11\% \cdot s_{55} = \lambda_1$$

因此从纵向看，有：  $\sum_{i=1}^p r_{iF_k}^2 s_{ii} = \lambda_k$

## 二、主成分的性质

### 1. 主成分的协差阵为对角阵

$$\text{Var}(F) = \Lambda = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p \end{bmatrix}$$

## 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. $F_k$ 与 $x_i$ 的相关系数

$$\rho(F_k, x_i) = r_{iF_k} = a_{ik} \sqrt{\frac{\lambda_k}{s_{ii}}}$$

若进行R型分析，则  $\rho(F_k, x_i) = r_{iF_k} = a_{ik} \sqrt{\lambda_k}$

#### 4. $F_k$ 对 $x_i$ 的方差贡献率为 $\rho^2(F_k, x_i)$

从横行看有

$$\sum_{k=1}^p \rho^2(F_k, x_i) = \sum_{k=1}^p r_{iF_k}^2 = 1$$

从纵向看有

$$\sum_{i=1}^p \rho^2(F_k, x_i) \cdot s_{ii} = \sum_{k=1}^p r_{iF_k}^2 \cdot s_{ii} = \lambda_k$$

若进行R型分析，则

$$\sum_{i=1}^p \rho^2(F_k, x_i) = \sum_{k=1}^p r_{iF_k}^2 = \lambda_k$$



## § 6 用主成分图解样品和变量

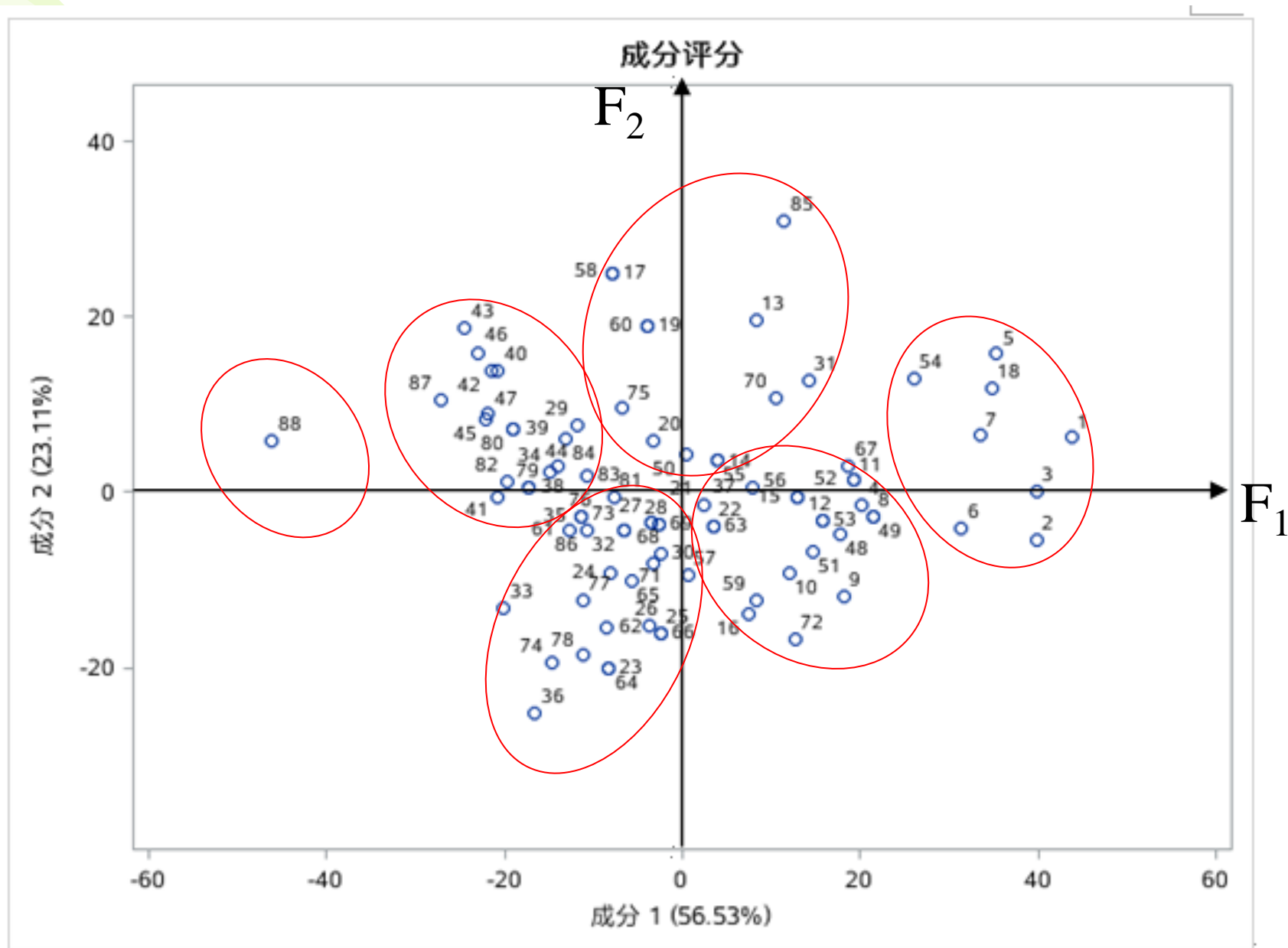
## 一、图解样品（对样品分类）

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。

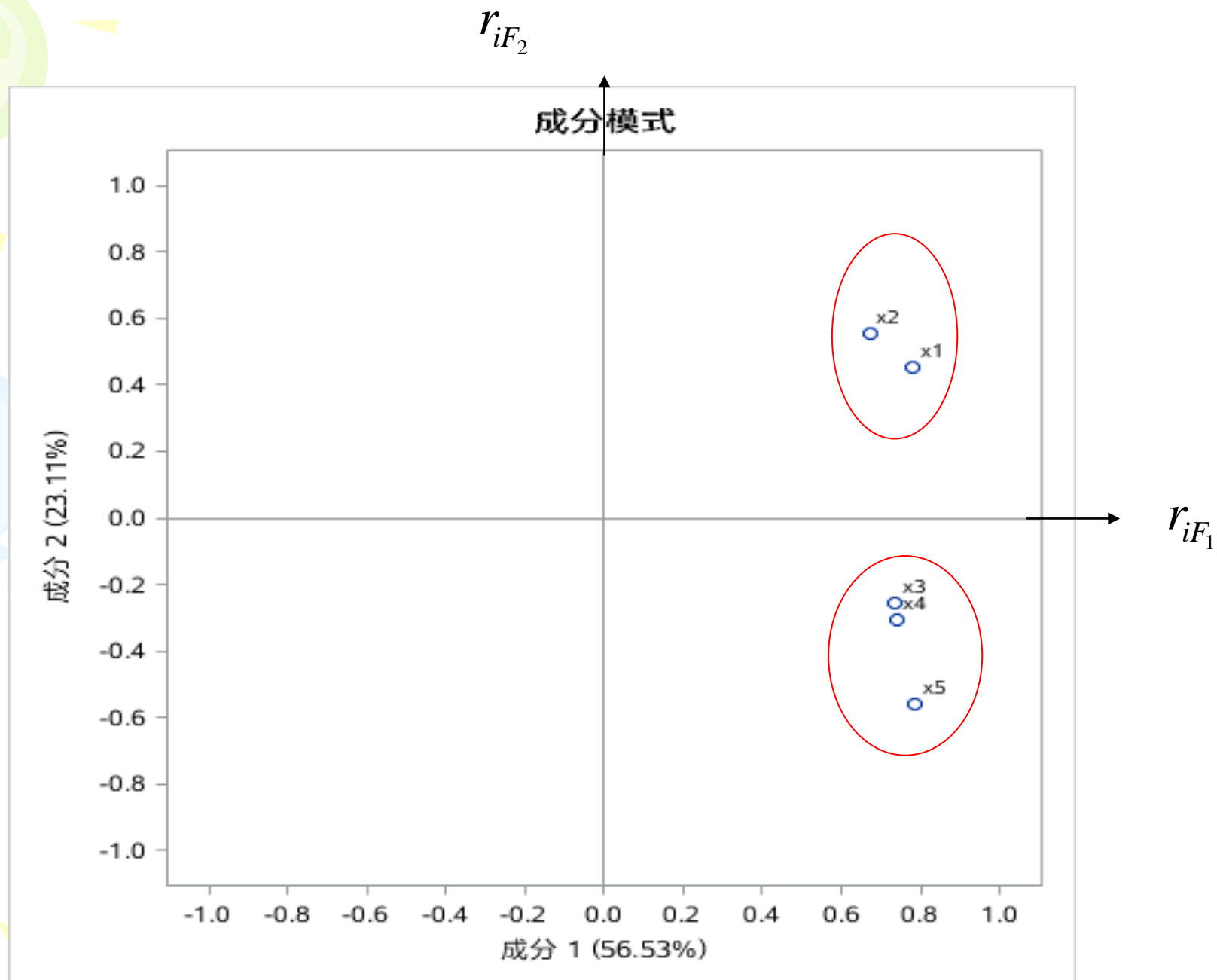
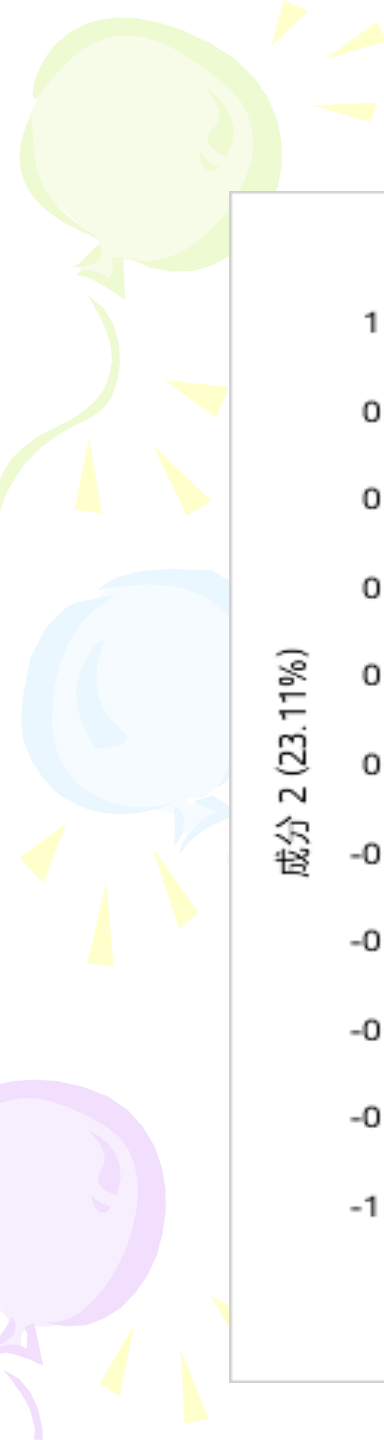




## 二、图解变量（对变量分类）

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  $\rightarrow$ ，只剩下  $\rho(F_1, x_i)$  和  $\rho(F_2, x_i)$ 。

以  $\rho(F_1, x_i)$  为横轴， $\rho(F_2, x_i)$  为纵轴，建立直角坐标系。然后以为  $\rho(F_1, x_i)$  横坐标，以  $\rho(F_2, x_i)$  为纵坐标，在坐标系中描出各变量对应的点。





§ 7

## 主成分分析用于系统评估

第一种方法，通过主成分分析得到综合指标

$$F_1 = a_{11}x_1 + a_{21}x_2 + \dots + a_{p1}x_p$$

利用  $F_1$  作为评估指标，根据  $F_1$  得分对样本点进行排序比较。但有3个前提条件：

1.  $F_1$  与全体原变量都正相关，即  $r_{iF_1} = a_{i1}\sqrt{\lambda_1} > 0$  ( $i=1,2,\dots,p$ )。
2. 各  $x_i$  ( $i=1,2,\dots,p$ ) 在数值上的分布较为均匀。
3.  $F_1$  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  $F_1, F_2, \dots, F_m$ ，以每个主成分  $F_i$  的方差贡献率  $\alpha_i = \frac{\lambda_i}{\sum_{i=1}^p \lambda_i}$

为权，构造综合评价函数

$$F = \alpha_1 F_1 + \alpha_2 F_2 + \dots + \alpha_m F_m$$



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， $F_i (i = 2, \dots, m)$  的含义违背了综合评价的本意。

以88个学生5门课程成绩的主成分分析为例

$$F_1 = 0.5327x_1 + 0.4208x_2 + 0.3225x_3 + 0.3422x_4 + 0.5639x_5 - 94.48$$

$$F_2 = 0.4846x_1 + 0.5405x_2 - 0.1735x_3 - 0.2210x_4 - 0.6278x_5 + 2.03$$

$$F = \frac{\lambda_1}{\sum_{i=1}^5 \lambda_i} \times F_1 + \frac{\lambda_2}{\sum_{i=1}^5 \lambda_i} \times F_2 = 56.53\% \times F_1 + 23.11\% \times F_2$$

$$= 0.4132x_1 + 0.3628x_2 + 0.1422x_3 + 0.1424x_4 + 0.1737x_5 - 52.94$$

显然，对开卷考成绩好的学生不公平

## 案例分析：区域经济发展的综合分析

衡量一个地区的经济发展水平，不能只考虑**GDP**，要综合宏观经济发展状况、人民生活水平、教育科技发展水平、医疗卫生状况、信息通讯水平等方面的状况。为比较我国**30**个大陆省份经济发展水平和主要影响因素，选取以下**10**个指标进行分析。

表1 指标体系

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R&D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2 指标数据（文件名：economic）

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	每万从业人员有效发明专利数	人均R&D经费支出	互联网普及率	每千人口医生数	人均财政性教育经费支出
北京	129679.9	73.51	14911.1	57229.8	25138.87	101.65	0.68	78	4.35	4465.18
天津	119616	54.42	7253.69	37022.3	14898.55	51.91	0.35	62	2.64	2802.49
河北	45146.16	39.16	662.37	21484.1	4291.91	14.23	0.05	54	2.55	1694.23
山西	41883.53	46.6	463.8	20420	5035.71	8.39	0.04	56	2.55	1674.08
内蒙古	63615.89	49.31	549.62	26212.2	6731.47	19.04	0.06	52	2.78	2220.56
辽宁	53627.33	48.99	2276.26	27835.4	5481.5	24.66	0.09	62	2.65	1484.62
吉林	54915.68	41.98	682.37	21368.3	4449.65	9.9	0.05	52	2.6	1867.04
.....										
浙江	91591.79	46.54	6680.22	42045.7	10269.49	32.54	0.2	65	3.16	2530.31
安徽	43096.58	37.08	857.55	21863.3	4486.16	78.83	0.08	47	1.93	1618.89
福建	82300.49	38.39	4373.05	30047.7	7183.64	20.78	0.12	65	2.15	2153.82
陕西	57014.82	37.51	1046.68	20635.2	5224.53	23.26	0.11	53	2.43	2156.4
甘肃	28300.91	49.62	192.69	16011	3094.67	19.67	0.03	44	2.14	2152.38
青海	43791.45	41.96	110.37	19001	4107.49	14.55	0.02	54	2.59	3128.33
宁夏	50426.4	41.78	739	20561.7	6115.06	31.45	0.04	51	2.67	2498.94
新疆	44367.81	40.01	845.03	19975.1	5979.28	13.6	0.02	55	2.55	2946.14

# SAS程序

```
proc princomp data=economic out=prin;  
run;
```

```
proc princomp data=economic out=prin n=3  
plot=pattern(ncomp=2) plot=score(ncomp=2);  
var x1-x10;  
id region;  
run;
```

# SAS输出

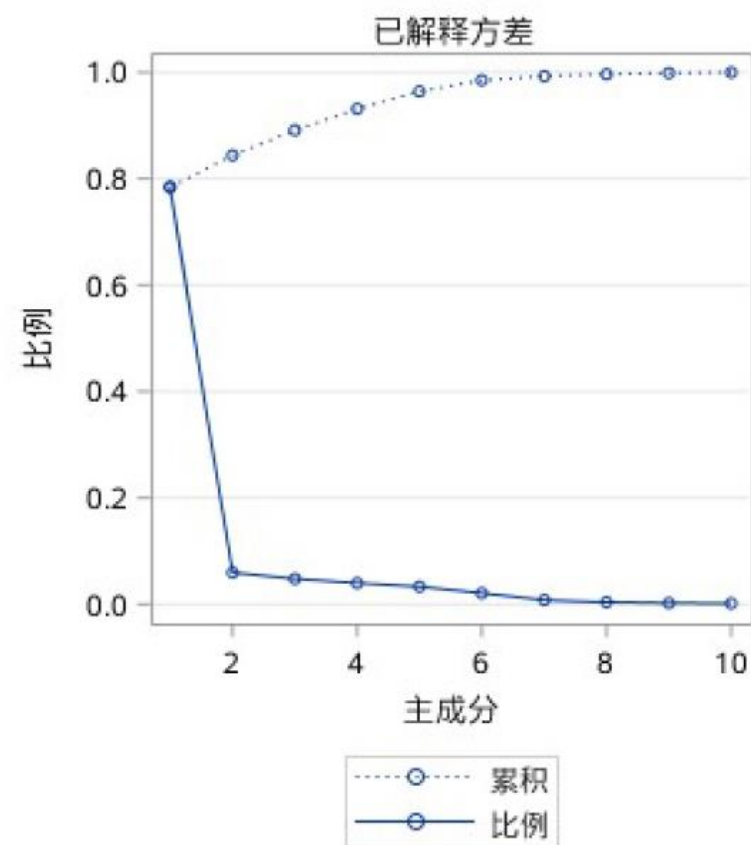
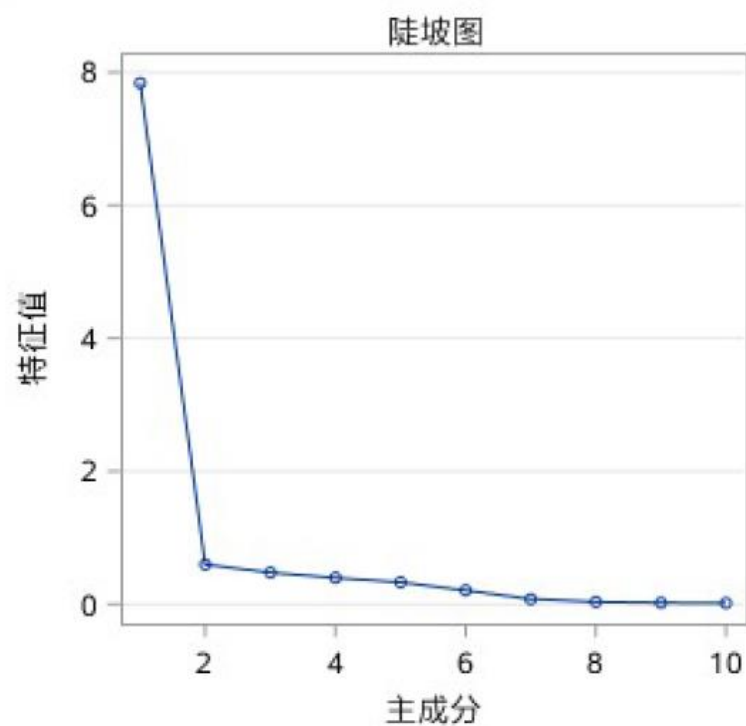
相关矩阵

		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

# SAS输出

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

# SAS输出



# SAS输出

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

$$Prin1 = 0.3301x_1 + 0.3021x_2 + 0.3375x_3 + 0.3457x_4 + 0.3460x_5 + 0.2768x_6 + 0.3465x_7 + 0.3010x_8 + 0.2790x_9 + 0.2855x_{10}$$

$$Prin2 = -0.2133x_1 + 0.3362x_2 - 0.2722x_3 - 0.1365x_4 - 0.0759x_5 - 0.4151x_6 - 0.0475x_7 - 0.0677x_8 + 0.6601x_9 + 0.3670x_{10}$$

# 结果分析

特征向量

		Prin1	Prin2
x1	人均地区生产总值	0.3301	-0.2133
x2	第三产业产值所占比重	0.3021	0.3262
x3	人均进出口额	0.3375	-0.2722
x4	人均可支配收入	0.3457	-0.1365
x5	人均地方财政收入	0.3460	-0.0759
x6	每万从业人员有效发明专利数	0.2768	-0.4151
x7	人均R&D经费支出	0.3465	-0.0475
x8	互联网普及率	0.3010	-0.0677
x9	每千人口医生数	0.2790	0.6601
x10	人均财政性教育经费支出	0.2855	0.3670

地区综合发  
展水平

$$Prin1 = 0.3301x_1 + 0.3021x_2 + 0.3375x_3 + 0.3457x_4 + 0.3460x_5 + 0.2768x_6 + 0.3465x_7 + 0.3010x_8 + 0.2790x_9 + 0.2855x_{10}$$

地区经济发展  
水平与地区发  
展潜力的比较

$$Prin2 = -0.2133x_1 + 0.3362x_2 - 0.2722x_3 - 0.1365x_4 - 0.0759x_5 - 0.4151x_6 - 0.0475x_7 - 0.0677x_8 + 0.6601x_9 + 0.3670x_{10}$$

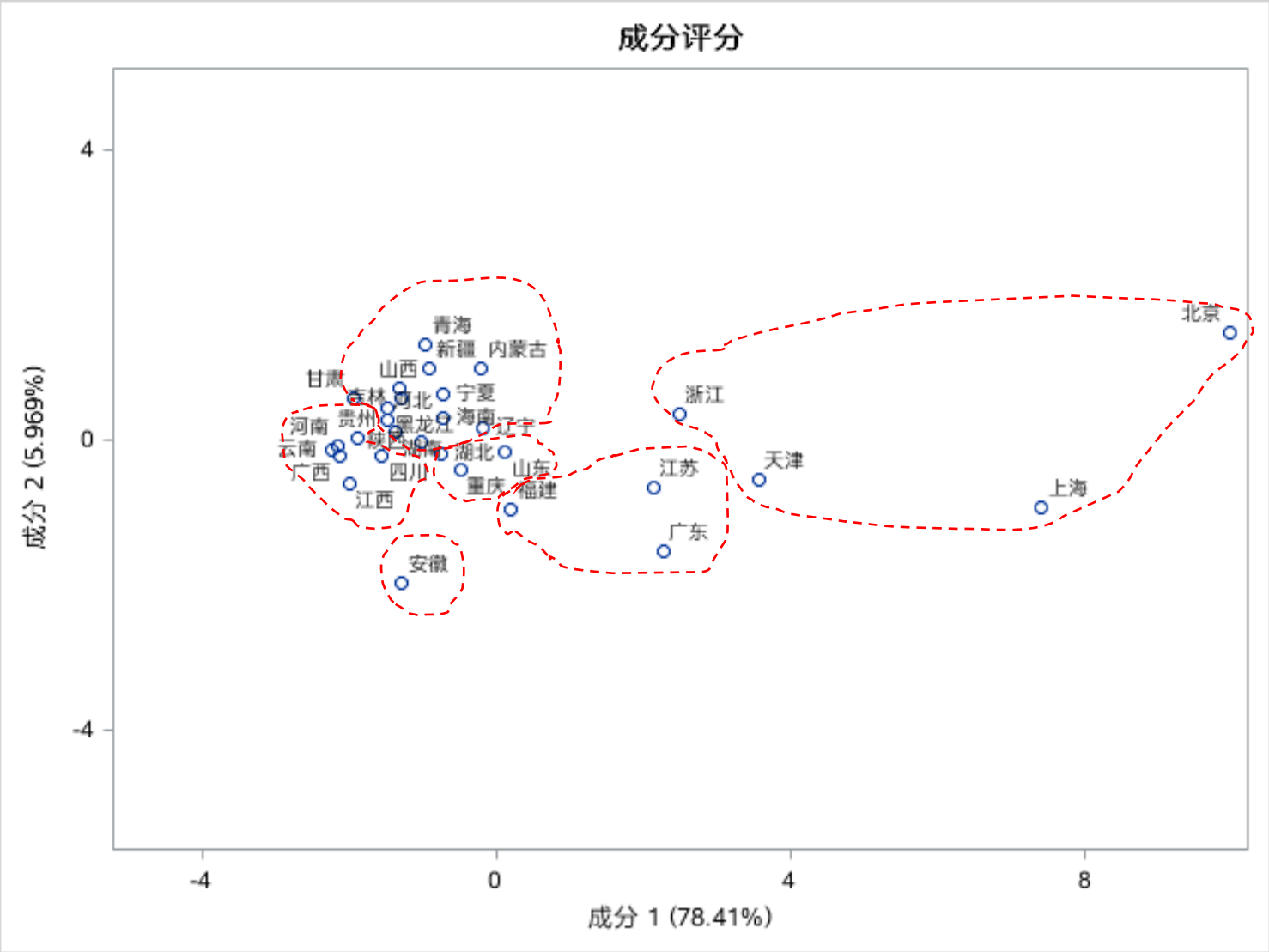
# 结果分析

表3 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	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	9.9759	1.4643	1	青海	-0.9592	1.3001	16
上海	7.4173	-0.9387	2	陕西	-1.0129	-0.0307	17
天津	3.5683	-0.5666	3	吉林	-1.2846	0.5598	18
浙江	2.4862	0.3537	4	安徽	-1.3029	-1.9792	19
广东	2.2794	-1.5272	5	山西	-1.3136	0.6912	20
江苏	2.1353	-0.6549	6	湖南	-1.3745	0.0947	21
福建	0.1970	-0.9777	7	黑龙江	-1.4791	0.4199	22
山东	0.1096	-0.1725	8	河北	-1.4965	0.2547	23
辽宁	-0.1761	0.1512	9	四川	-1.5679	-0.2133	24
内蒙古	-0.2238	0.9814	10	贵州	-1.8888	0.0253	25
重庆	-0.4978	-0.4187	11	甘肃	-1.9405	0.5632	26
宁夏	-0.7249	0.6114	12	江西	-2.0083	-0.6009	27
海南	-0.7288	0.2889	13	广西	-2.1293	-0.2279	28
湖北	-0.7548	-0.1924	14	河南	-2.1471	-0.0986	29
新疆	-0.9115	0.9889	15	云南	-2.2461	-0.1495	30



# SAS输出



# SAS输出

**类1：** 发展水平高，潜力高地区：

{北京、上海、天津、浙江}

**类2：** 发展水平较高，潜力一般地区：

{广东、江苏、福建}

**类3：** 发展水平一般，潜力一般地区：

{ 山东、湖北、重庆}

**类4：** 发展水平一般，潜力较高：

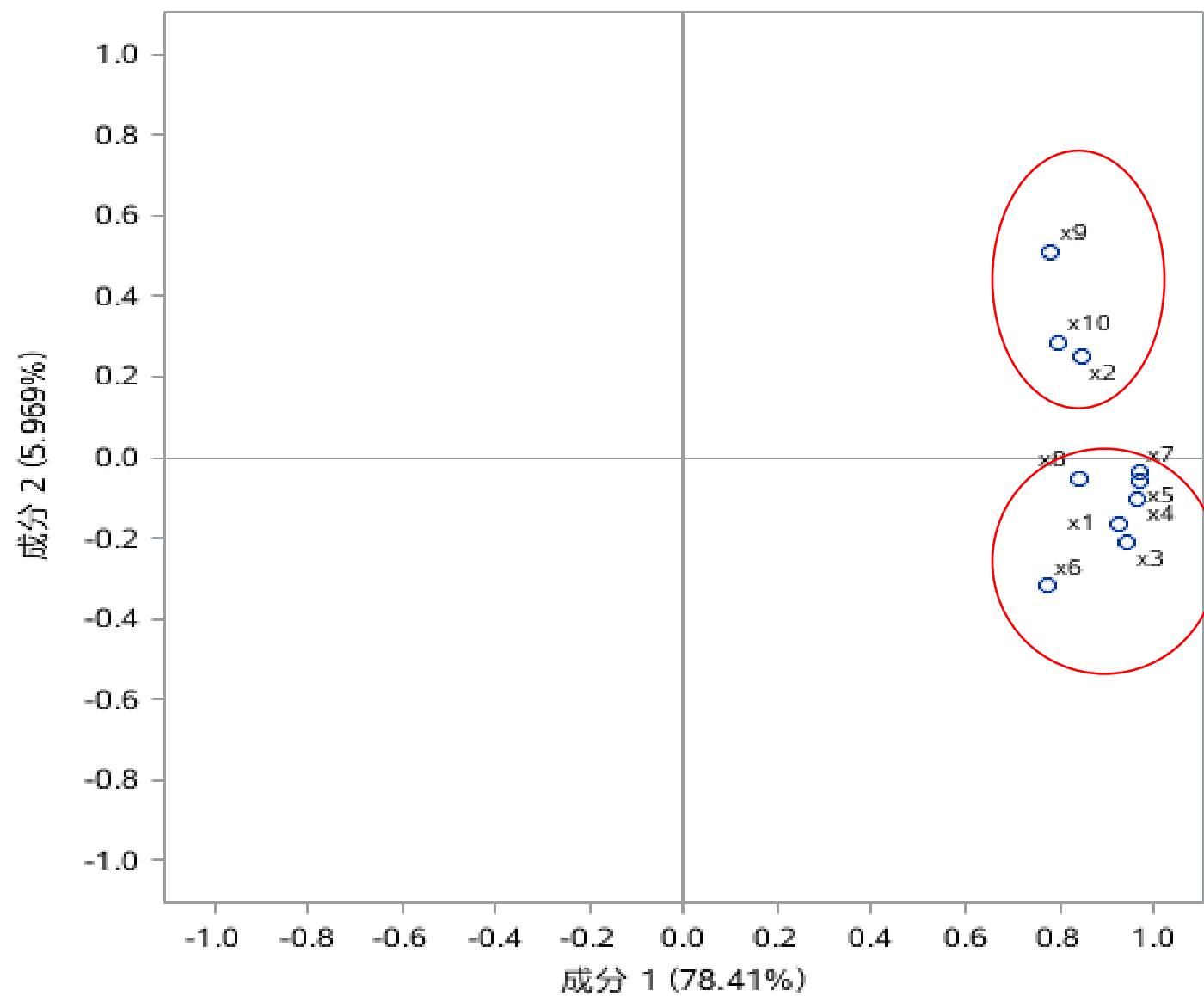
{青海、新疆、内蒙古、山西、宁夏、 吉林、黑龙江、海南、河北、辽宁、湖南、陕西}

**类5：** 发展水平较落后，潜力较低地区： {安徽}

**类6：** 发展水平落后，潜力较低地区：

{贵州、河南、云南、四川、广西、江西}

成分模式



# 主成分降维之后，

再利用新的变量（主成分），进行聚类分析、  
回归分析、综合评价等等

# 主成分回归

解决回归分析中的多重共线性问题

# 一、提出问题

## 居民消费水平的多因素分析

居民消费水平指常住住户对货物和服务的全部最终消费支出，居民消费除了直接以货币形式购买货物和服务的消费之外，还包括以其他方式获得的货币和服务的消费支出。

居民消费水平受许多因素的影响，主要有居民收入、消费观念、消费环境、国家政策等等。由于资料的可得性和代表性，选择以下变量。



$y$  : 居民消费水平（元）

$x_1$  : 农村居民家庭人均纯收入（元）

$x_2$  : 城镇居民家庭人均可支配收入（元）

$x_3$  : 国家财政支出总额（亿元）

$x_4$  : 每万人在校大学生人数（人）

$x_5$  : 每万人在校研究生人数（人）

$x_6$  : 人口自然增长率（‰）

$x_7$  : 金融机构个人人民币储蓄存款一年期存款利率（%）

数据见sasuser.vregex01

## 二、主成分回归方法

### 1. 对自变量进行主成分分析

$$F_1 = a_{11}x_1 + a_{21}x_2 + \dots + a_{p1}x_p$$

$$F_2 = a_{12}x_1 + a_{22}x_2 + \dots + a_{p2}x_p$$

.....

$$F_p = a_{1p}x_1 + a_{2p}x_2 + \dots + a_{pp}x_p$$

取m个主成分，反映原p个变量95%以上的信息



主成分

$$F_1 = a_{11}x_1 + a_{21}x_2 + \dots + a_{p1}x_p$$

$$F_2 = a_{12}x_1 + a_{22}x_2 + \dots + a_{p2}x_p$$

.....

$$F_m = a_{1m}x_1 + a_{2m}x_2 + \dots + a_{pm}x_p$$

$$Y_i = \hat{\gamma}_1 F_{i1} + \hat{\gamma}_2 F_{i2} + \dots + \hat{\gamma}_m F_{im}$$

3. 把主成分的表达式代入，得到最终的回归模型

$$Y_i = \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \dots + \hat{\beta}_p x_{ip}$$



### 三、主成分回归的实例

#### 1、经济分析数据

$X_1$  : GDP

$X_2$  : 积累总额

$X_3$  : 消费总额

$Y$  : 进口总额

求进口总额与GDP、积累总额和消费总额之间的回归方程。

数据见sasuser.vregex1

## Summary of Fit

Dependent Mean	21.89091	R-Square	0.9919
Root MSE	0.48887	Adj R-Sq	0.9884

## Parameter Estimates

Variable	DF	Estimate	Standard Error	t 值	Prob>  t
Intercept	1	-10.12799	1.21216	-8.36	0.0001
x1	1	-0.05140	0.07028	-0.73	0.4883
X2	1	0.58695	0.09462	6.20	0.0004
x3	1	0.28685	0.10221	2.81	0.0263

### Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
PCR1	1.9992	1.0010	0.6664	0.6664
PCR2	0.9982	0.9955	0.3327	0.9991
PCR3	0.0026		0.0009	1.0000

### Eigenvectors

	F1	F2	F3
x1	0.7063	-0.0357	0.7070
x2	0.0435	0.9990	0.0070
x3	0.7065	-0.0258	-0.7072

$$F1=0.7063x_1+0.0435x_2+0.7065x_3$$

$$F2=-0.0357x_1+0.9990x_2-0.0258x_3$$



Obs	x1	x2	x3	y*	F1	F2	F3
1	-1.50972	0.54571	-1.53319	-1.31852	-2.12589	0.63866	0.020722
2	-1.11305	0.48507	-1.20848	-1.20848	-1.61893	0.55554	0.071113
3	-0.76971	-0.12127	-0.80140	-0.63625	-1.11517	-0.07298	0.021730
4	-0.63637	-0.12127	-0.62209	-0.61424	-0.89430	-0.08237	-0.010813
5	-0.45970	-1.33395	-0.37008	-0.68027	-0.64421	-1.30669	-0.072582
6	-0.12970	-0.66697	-0.09869	-0.32813	-0.19035	-0.65915	-0.026553
7	0.25031	-0.72761	0.30355	0.17807	0.35962	-0.74367	-0.042781
8	0.59365	1.39458	0.69610	1.01440	0.97180	1.35406	-0.062863
9	1.05032	1.03078	1.09350	1.36654	1.55932	0.96405	-0.023574
10	1.24366	1.09141	1.19042	1.25649	1.76700	1.01522	0.044988
11	1.48033	-1.57648	1.35035	0.97038	1.93110	-1.66266	0.080613

### Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F 值	Prob>F
Model	2	9.8828	4.9414	337.2302	0.0001
Error	8	0.1172	0.0147		
Total	10	10.0000			

### Parameter Estimates

Variable	DF	Estimate	Standard Error	t 值	Prob>  t
F1	1	0.6900	0.0271	25.4859	0.0001
F2	1	0.1913	0.0383	4.9930	0.0011

标准化后的变量

$$\hat{y}^* = 0.68998F_1 + 0.19130F_2$$

$$\hat{y}^* = 0.4804x_1^* + 0.2211x_2^* + 0.4825x_3^*$$

把标准化变量还原，代入得：

$$\frac{y - 21.89}{4.5437} = 0.4805 \cdot \frac{x_1 - 194.59}{30} + 0.22 \cdot \frac{x_2 - 3.3}{1.65} + 0.4826 \cdot \frac{x_3 - 139.73}{20.63}$$

$$\hat{y} = -9.130 + 0.0727x_1 + 0.6091x_2 + 0.1062x_3$$

## 例2 国内旅游人数模型

影响人们外出旅游的因素有居民收入、交通、闲暇时间、旅游目的地治安状况、旅游目的地的环境卫生以及接待能力等等。

由于资料的可得性和代表性，选择以下变量。

$y$  国内旅游人数（百万人）

$x_1$  农村居民人均纯收入（元）

$x_2$  城镇居民人均可支配收入（元）

$x_3$  公路线路里程（万公里）

数据见sasuser.tourmx



## Summary of Fit

<b>Dependent Mean</b>	<b>558.1017</b>	<b>R-Square</b>	<b>0.9920</b>
<b>Root MSE</b>	<b>19.2003</b>	<b>Adj R-Sq</b>	<b>0.9890</b>

## Parameter Estimates

<b>Variable</b>	<b>DF</b>	<b>Estimate</b>	<b>Standard Error</b>	<b>t 值</b>	<b>Prob&gt;  t </b>
<b>Intercept</b>	<b>1</b>	<b>417.8201</b>	<b>74.0230</b>	<b>5.6445</b>	<b>0.0005</b>
<b>Incomeon</b>	<b>1</b>	<b>-0.1381</b>	<b>0.0699</b>	<b>-1.9759</b>	<b>0.0836</b>
<b>Incomeoc</b>	<b>1</b>	<b>0.1737</b>	<b>0.0302</b>	<b>5.7589</b>	<b>0.0004</b>
<b>Highway</b>	<b>1</b>	<b>-3.0009</b>	<b>0.8192</b>	<b>-3.6633</b>	<b>0.0064</b>

### Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
PCR1	2.8088	2.6238	0.9363	0.9363
PCR2	0.1850	0.1788	0.0617	0.9979
PCR3	0.0062		0.0021	1.0000

### Eigenvectors

	F1	F2	F3
x1	0.5810	-0.5167	0.6289
X2	0.5918	-0.2623	-0.7622
x3	0.5588	0.8150	0.1533

$$F1=0.5810x_1+0.5918x_2+0.5588x_3$$

$$F2=-0.5167x_1-0.2623x_2+0.8150x_3$$

### Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F 值	Prob>F
Model	2	10.7113	5.3556	166.9328	0.0001
Error	9	0.2887	0.0321		
Total	11	11.0000			

### Parameter Estimates

Variable	DF	Estimate	Standard Error	t 值	Prob>  t
F1	1	0.5767	0.0322	17.8977	0.0001
F2	1	-0.4620	0.1256	-3.6794	0.0051


$$\hat{y}^* = 0.5767 F_1 - 0.4620 F_2$$

标准化后的变量

$$\hat{y}^* = 0.5738x_1^* + 0.4625x_2^* - 0.0543x_3^*$$


把标准化变量还原，代入得：

$$\frac{y - 558.10}{182.91} = 0.5738 \cdot \frac{x_1 - 1575.63}{670.26} + 0.4625 \cdot \frac{x_2 - 4167.66}{1865.84} - 0.0543 \cdot \frac{x_3 - 121.90}{19.36}$$

$$y = 184.953 + 0.15658x_1 + 0.04534x_2 - 0.51287x_3$$


$$\hat{y}^* = 0.5767F_1$$

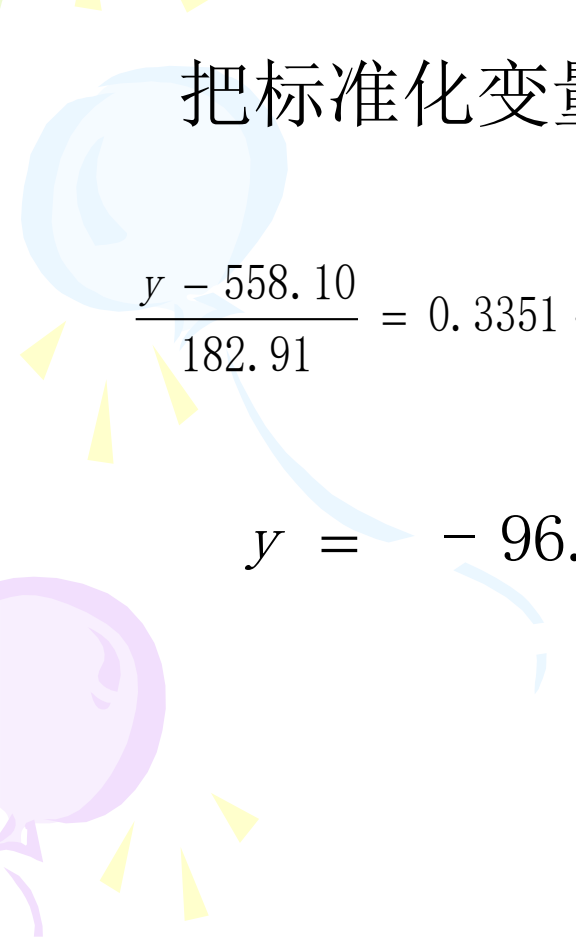
标准化后的变量



$$\hat{y}^* = 0.3351x_1^* + 0.3413x_2^* + 0.3223x_3^*$$

把标准化变量还原，代入得：

$$\frac{y - 558.10}{182.91} = 0.3351 \cdot \frac{x_1 - 1575.63}{670.26} + 0.3413 \cdot \frac{x_2 - 4167.66}{1865.84} + 0.3223 \cdot \frac{x_3 - 121.90}{19.36}$$


$$y = -96.543 + 0.0914x_1 + 0.03346x_2 + 3.0446x_3$$

# 主成分的改进

## 1、无量纲化的改进

$$\begin{aligned}\Sigma = \text{Var}(X) &= \begin{pmatrix} \text{var}(x_1) & \text{cov}(x_1, x_2) & \cdots & \text{cov}(x_1, x_p) \\ \text{cov}(x_2, x_1) & \text{var}(x_2) & \cdots & \text{cov}(x_2, x_p) \\ \vdots & \vdots & & \vdots \\ \text{cov}(x_p, x_1) & \text{cov}(x_p, x_2) & \cdots & \text{var}(x_p) \end{pmatrix} \\ &= \begin{pmatrix} 1 & \rho(x_1, x_2) & \cdots & \rho(x_1, x_p) \\ \rho(x_2, x_1) & 1 & \cdots & \rho(x_2, x_p) \\ \vdots & \vdots & & \vdots \\ \rho(x_p, x_1) & \rho(x_p, x_2) & \cdots & 1 \end{pmatrix} = R\end{aligned}$$

从标准化的数据提取的主成分，实际上只包含了各指标间相互影响这一部分信息，不能准确反映原始数据所包含的全部信息。



# 无量纲化

- 标准化
- 均值化
- 功效系数法     0-1

# 改进原始数据的无量纲化方法

## ◆ 均值化方法

$$x_{ji} \rightarrow x'_{ji} = \frac{x_{ji}}{\bar{x}_i}$$

均值化后，数据的协方差矩阵S 中的元素

$$\begin{aligned} u_{jk} &= \frac{1}{n} \sum_{i=1}^n \left( \frac{x_{ij}}{\bar{x}_j} - 1 \right) \left( \frac{x_{ik}}{\bar{x}_k} - 1 \right) \\ &= \frac{1}{n} \sum_{i=1}^n \frac{(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\bar{x}_j \bar{x}_k} = \frac{s_{jk}}{\bar{x}_j \bar{x}_k} \end{aligned}$$



## 均值化后，数据的协方差矩阵

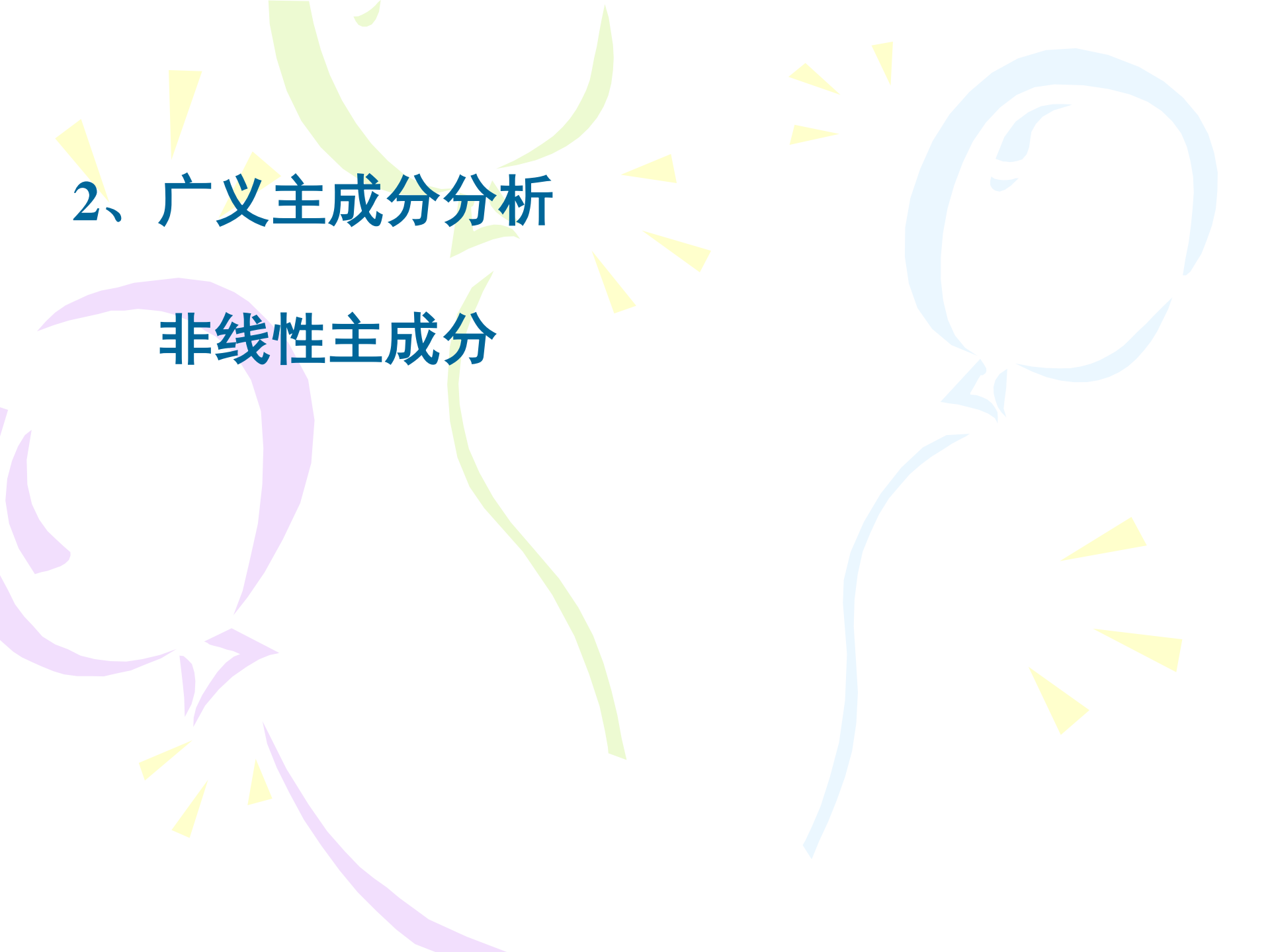
$$S = \begin{pmatrix} \frac{s_{11}}{\bar{x}_1^2} & \frac{s_{12}}{\bar{x}_1 \bar{x}_2} & \dots & \frac{s_{1p}}{\bar{x}_1 \bar{x}_p} \\ \frac{s_{21}}{\bar{x}_2 \bar{x}_1} & \frac{s_{22}}{\bar{x}_2^2} & \dots & \frac{s_{2p}}{\bar{x}_2 \bar{x}_p} \\ \vdots & \vdots & & \vdots \\ \frac{s_{p1}}{\bar{x}_p \bar{x}_1} & \frac{s_{p2}}{\bar{x}_p \bar{x}_2} & \dots & \frac{s_{pp}}{\bar{x}_p^2} \end{pmatrix}$$

对角线上是原变量标准差系数的平方，其他位置上是变量两两之间的相互关系。

均值化处理后的协方差矩阵不仅消除了指标量纲与数量级的影响，还能包含原始数据的全部信息。

## 2、广义主成分分析

### 非线性主成分



有许多实际问题，其观测数据阵并非线性结构，而呈现非线性结构。对于非线性结构的观测阵，应根据指标变量的具体的非线性结构，选用适当的曲面作坐标平面。采用原指标的非线性函数构造综合指标。

由Grandesikan（1966）和Wilkinson（1968）提出。

他们提议用原变量  $X_1, X_2, \dots, X_p$  的广义线性式

$$F = a_1 f_1(X) + a_2 f_2(X) + \dots + a_p f_p(X)$$

其中  $X = (X_1, X_2, \dots, X_p)'$

$f_1(X), \dots, f_p(X)$  为  $X$  的已知函数形式

对于给定的观测数据阵，若采用线性主成分分析效果很差（**S**或**R**的特征值取值分散，指标压缩很少或分析结果严重违反客观实际），可采用非线性主成分分析。

根据已给定的函数关系式  $Y_i = f_i(X)$   $i = 1, 2, \dots, p$

计算 $Y$ 的观测数据阵  $Y = (Y_{ij})_{n \times p}$  .

对 $Y$ 求线性主成分，求得  $k$  个线性主成分

$$F_j = a_{1j} f_1(X) + a_{2j} f_2(X) + \dots + a_{pj} f_p(X), j = 1, 2, \dots, k$$



广义主成分分析的关键在于确定非线性函数  $f_i(X)$

究竟取何种形式，应视具体情况，结合有关专业理论或实践经验给定。

# 成分向量的广义主成分分析

设随机向量  $X = (X_1, X_2, \dots, X_p)'$  满足下列条件:

$$(1) X_i > 0 \quad i = 1, 2, \dots, p \quad (2) \sum_{i=1}^p X_i = 1$$

从而每一分量可视为某一成分的含量，则称 $X$ 为成分向量。

其观测数据阵

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

称为合成数据

# “对数-线性比”主成分

**Aitchison**教授（1981年）提出用“对数-比”变换

$$Y_i = \log (X_i / g(X)) \quad i = 1, 2, \dots, p$$

$g(X)$  为成分向量 $X$ 的任一恒正函数。

一般可取  $g(X) = (X_1 X_2 \cdots X_p)^{\frac{1}{p}}$

$$Y_i = \log X_i - \frac{1}{p} \sum_{j=1}^p \log X_j$$

称之为“对数-中心化”变换

相应的 $Y$ 的观测数据阵为

$$(Y_{ij})_{n \times p} = \left( \log X_{ij} - \frac{1}{p} \sum_{j=1}^p \log X_{ij} \right)_{n \times p}$$

## 农村居民

$$PCR1 = \log \frac{X_6^{0.6623} X_7^{0.4711} X_8^{0.0174}}{X_1^{0.3258} X_2^{0.3364} X_3^{0.0043} X_4^{0.2787} X_5^{0.2056}}$$

## 城镇居民

$$PCR1 = \log \frac{X_2^{0.6490} X_7^{0.4621} X_8^{0.0027}}{X_1^{0.0934} X_3^{0.0587} X_4^{0.1479} X_5^{0.4237} X_6^{0.3894}}$$



# 主成分分析在SAS中用princomp过程:

```
proc princomp
```

```
data=duo.innovation prefix=z out=o;
```

```
run;
```

```
proc plot data=o;
```

```
plot z2*z1 $ region='*'/href=0
```

```
vref=0;
```

```
run;
```

```
proc sort data=o;
```

```
by descending z1;
```

```
run;
```

```
proc print data=o;
```

```
var region z1 z2;
```

```
run;
```

```
quit;
```

主成分分析有一个**princomp**过程就足够了。**prefix=z**表示，在输出数据集中（o中），主成分变量是z1、z2、...

**plot**过程

**href=0**表示在横坐标z1=0处画一条垂线，**vref=0**表示在纵坐标z2=0处画一条垂线。

**\$region='\*'**表示每个点在图上用\*表示，并且在\*后显示该样本点的region变量的值。

**sort**和**print**过程是很熟悉的过程了。


## 主成分分析图解样品：

```
ods graphics on;  
proc princomp data=tmp1.exec76 out=o1 plot=score(ncomp=2) ;  
id state;  
run;  
ods graphics off;
```

```
proc princomp data=tmp1.exec76 out=o1 prefix=F;  
run;  
proc plot data=o1;  
plot F2*F1 $ state='*'/href=0 vref=0;  
run;
```

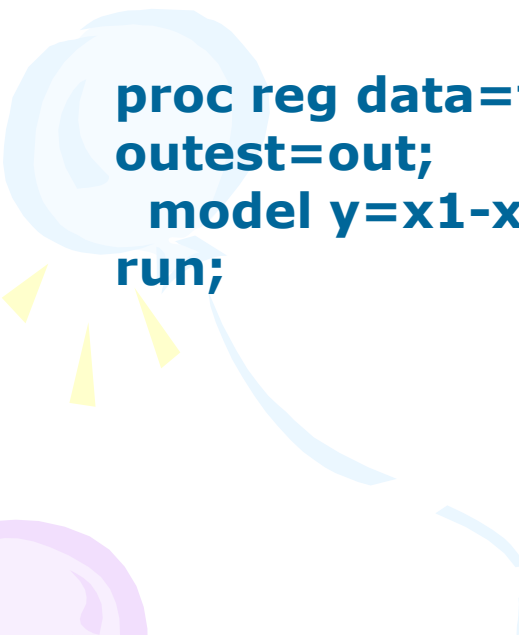
## 主成分分析图解变量：

```
ods graphics on;  
proc princomp data=tmp1.exec76 out=o1 plot=pattern(ncomp=2);  
run;  
ods graphics off;
```



```
proc corr data=tmp1.vregex1;  
var y x1-x3;  
run;
```

计算相关系数矩阵。



```
proc reg data=tmp1.vregex1  
outest=out;  
model y=x1-x3/ tol vif;  
run;
```

OLS估计回归参数，并给出共线性诊断结果。

```
proc standard data=vregex1  
out=sv mean=0 std=1;  
  var x1-x3 y;  
run;  
proc princomp data=sv  
prefix=z out=opcr ;  
  var x1 x2 x3;  
run;  
proc print data=opcr;  
  var z1 z2 y;  
run;  
proc reg data=opcr ;  
  model y=z1 z2;  
run;  
quit;  
  
proc reg data=vregex1  
outest=out;  
  model y=x1-x3/pcomit=1,2;  
run;  
quit;  
proc print data=out;  
run;
```

首先对数据标准化

然后对标准化后的数据进行主成分分析。这也就相当于是对相关系数进行的主成分分析。

作回归分析，自变量取第1、2主成分，因变量为y。

直接作主成分回归分析。

pcomit=1,2表示分别作两个回归，分别是剔除1个主成分，和剔除2个主成分，所做的主成分回归。

## 主成分回归的结果:

Obs	_MODEL_	_TYPE_	_DEPVAR_	_RIDGE_	_PCOMIT_	_RMSE_	Intercept	X1	X2	X3	Y
1	MODEL1	PARMS	Y	.	.	0.48887	-10.1280	-0.051396	0.58695	0.28685	-1
2	MODEL1	IPC	Y	.	1	0.55001	-9.1301	0.072780	0.60922	0.10626	-1
3	MODEL1	IPC	Y	.	2	1.05206	-7.7458	0.073814	0.08269	0.10735	-1

由于刚才我们分析了y与z1、z2的回归，z1、z2的回归系数均非0，且在0.01显著性水平下显著；而z1、z2又有99%的累积方差贡献率。因此我们可以认为：对y与z1、z2、z3的回归，取y对z1、z2的回归最佳。

y对z1、z2的回归，还原为x1、x2、x3的系数后就是上图中红线圈起的第2行。\_PCOMIT\_=1，表示这一行是删除了一个主成分后的 主成分回归结果。

而y与z1回归，再还原为x1、x2、x3的系数，就是上图中最后一行。

# SAS 程序

```
proc princomp data=文件名 n=? out=文件名1 outstat=文  
件名2 cov prefix=? ;  
run;
```

## 说明

data=文件名 指定用于分析的数据文件

out= 文件名自己取，保存原始数据和主成分得分

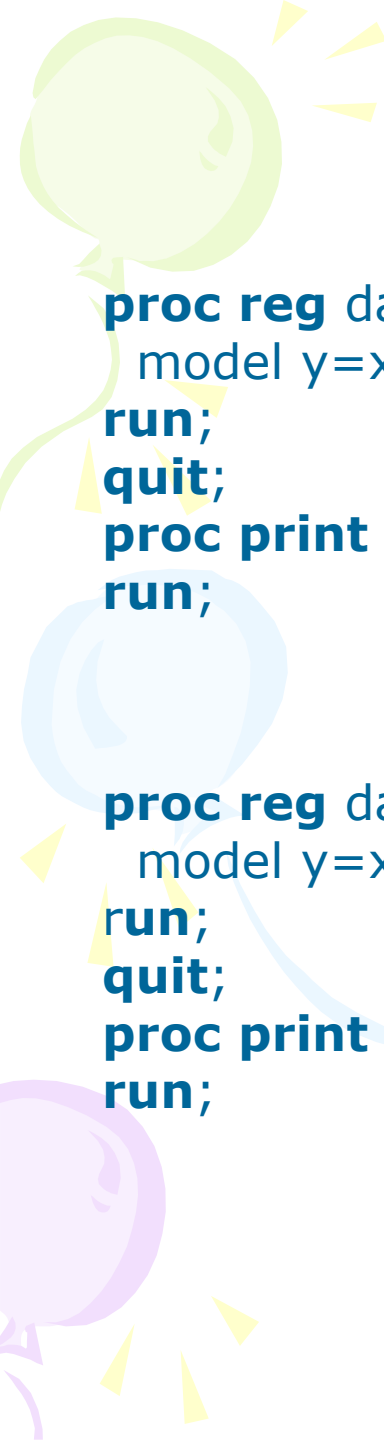
outstat= 文件名自己取，保存主成分分析过程中的统计量

cov 指定基于协差阵分析，若省略，基于相关系数矩阵

prefix= 指定主成分的代号

n= 指定主成分的个数

**特别注意，分号表示一个语句的结束，不能遗漏。**



```
proc reg data=vregex1 outest=out;  
  model y=x1-x3/ridge=0 to 2 by 0.1;  
run;  
quit;  
proc print data=out;  
run;
```

```
proc reg data=vregex1 outest=out;  
  model y=x1-x3/selection=stepwise;  
run;  
quit;  
proc print data=out;  
run;
```



浙江工商大学  
ZHEJIANG GONGSHANG UNIVERSITY

*Thank You*