



第八章 地理系统要素关系的主成分分析



1

主成分分析的原理



问题的提出：

地理系统是多要素的复杂系统。在地理学研究中，多变量问题是经常会遇到的。变量太多，无疑会增加分析问题的难度与复杂性，而且在许多实际问题中，多个变量之间是具有一定的相关关系的。

因此，人们会很自然地想到，能否在相关分析的基础上，用较少的新变量代替原来较多的旧变量，而且使这些较少的新变量尽可能多地保留原来变量所反映的信息？



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。

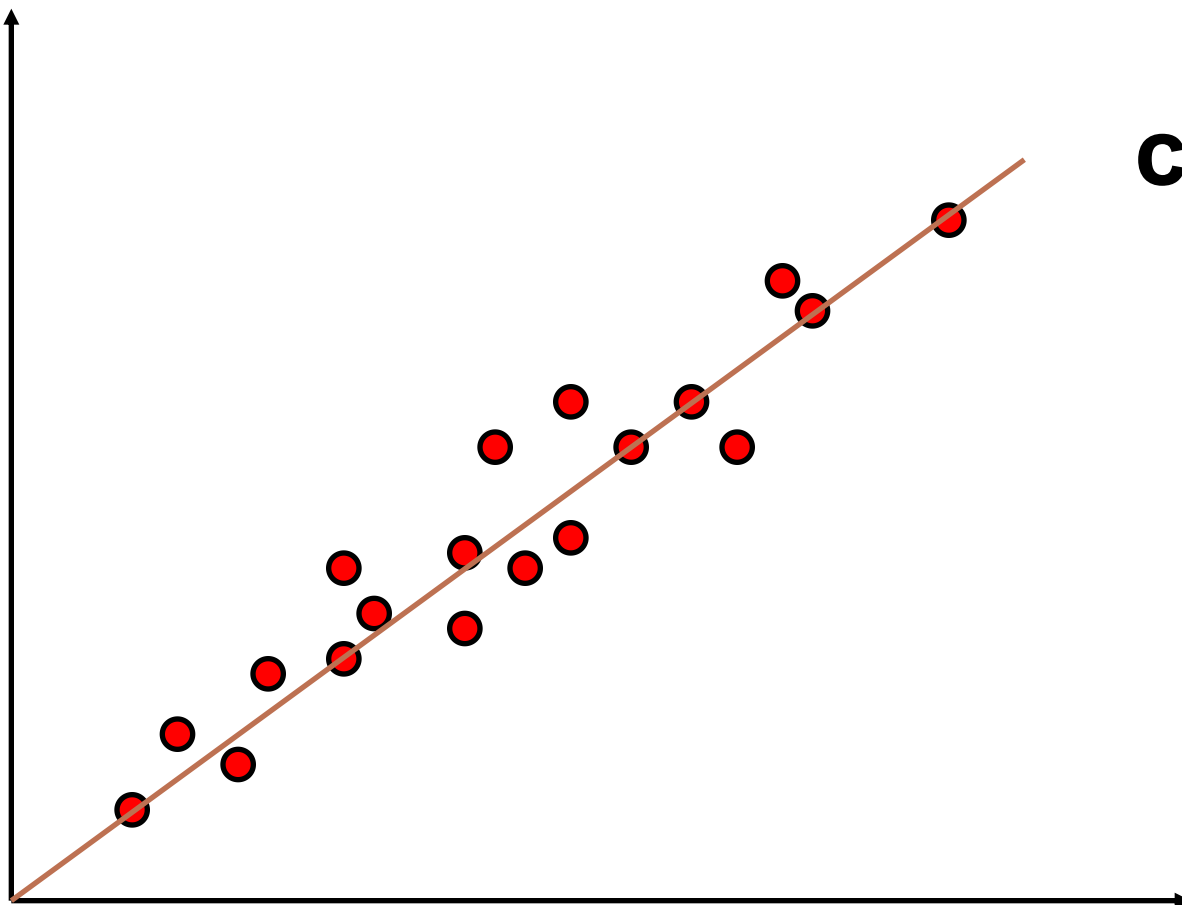


主成分分析试图在力保数据信息丢失最少的原则下，对这种多变量的截面数据表进行最佳综合简化，也就是说，对高维变量空间进行降维处理。

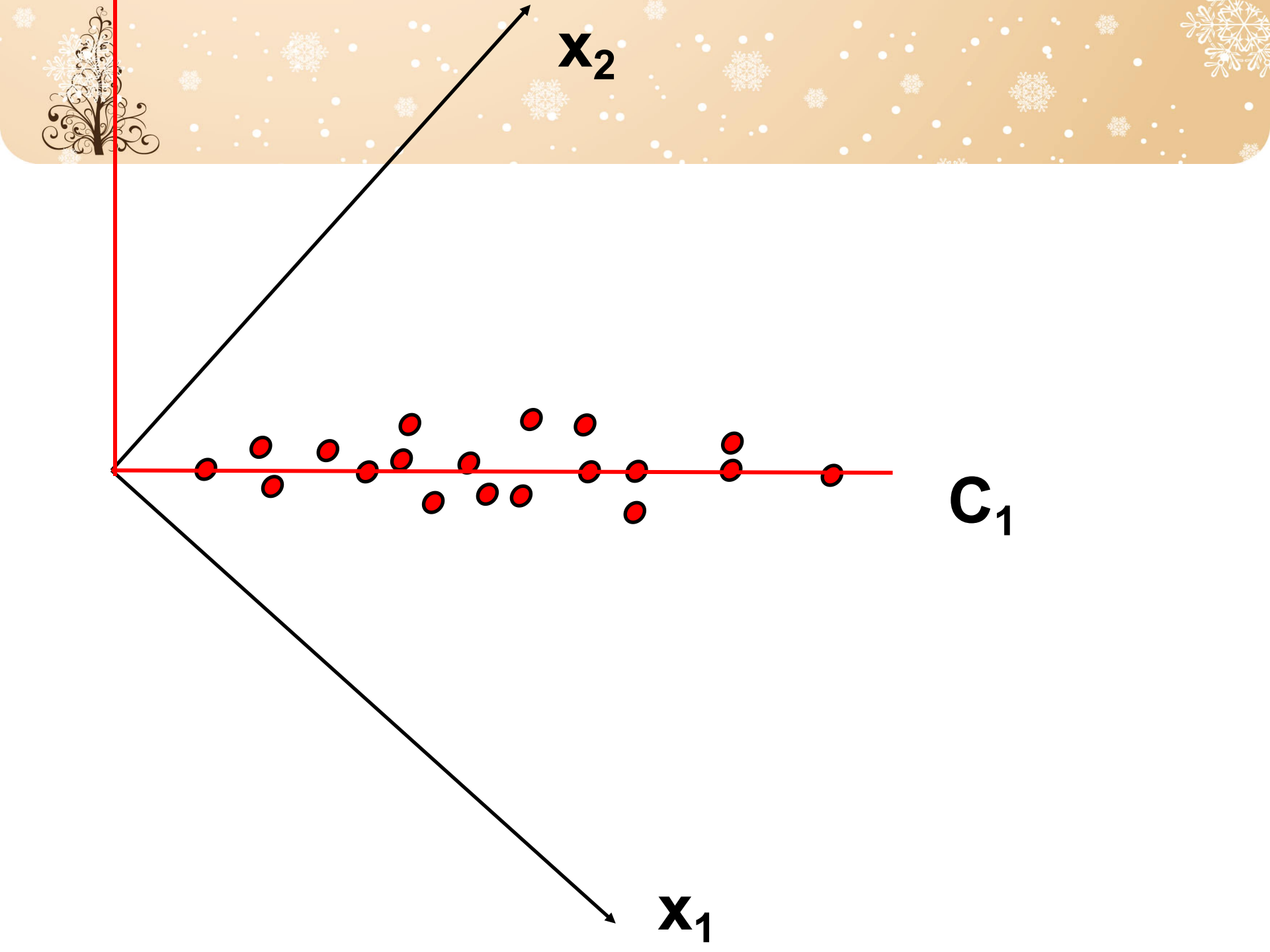
很显然，识辨系统在一个低维空间要比在一个高维空间容易得多。

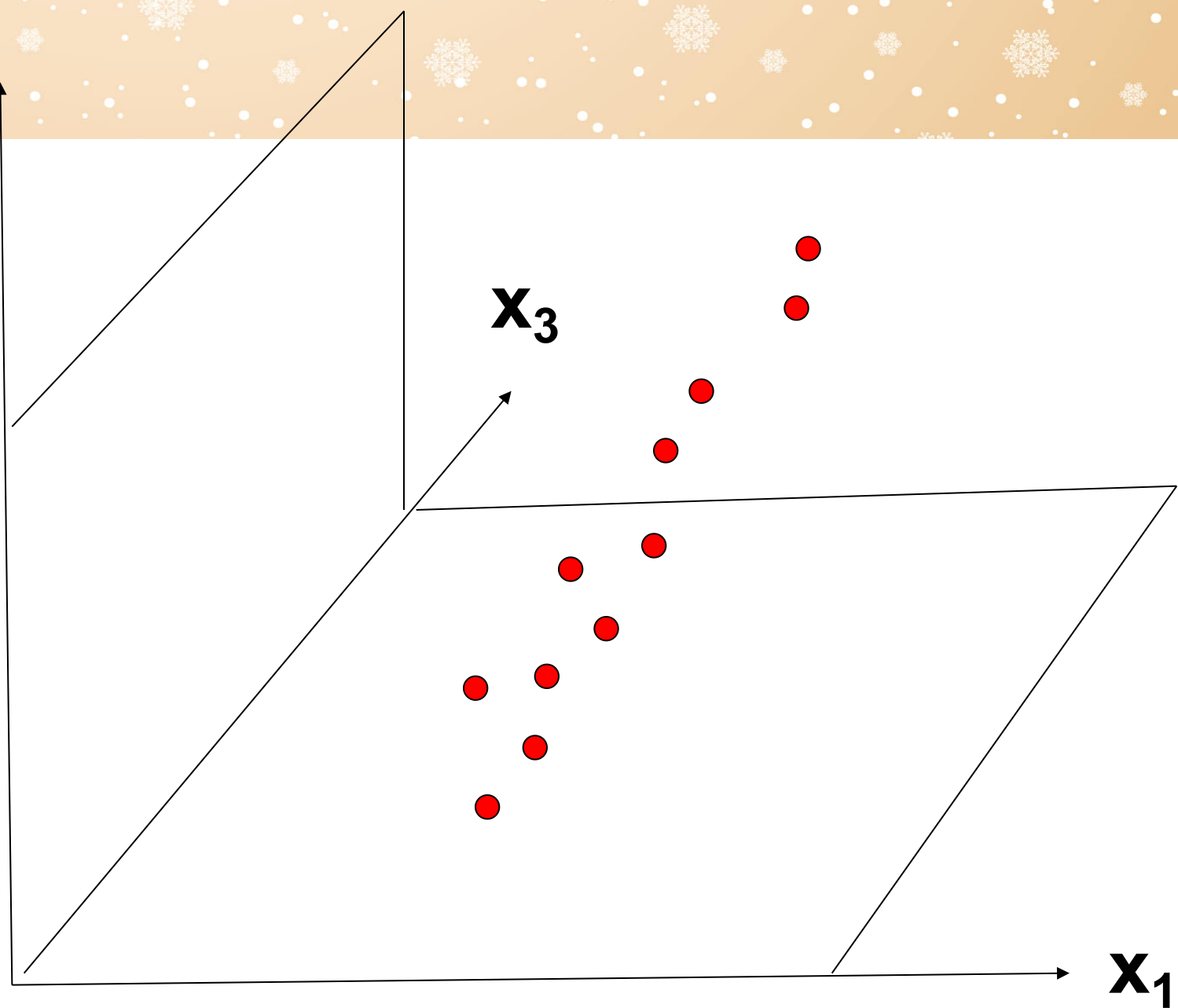


x_2



x_1







主成分分析（Principal Component Analysis, PCA）是一种数据降维技术，将多个具有较强相关性的实测变量综合成少量综合变量。

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。问题是：选择几个成分合适？



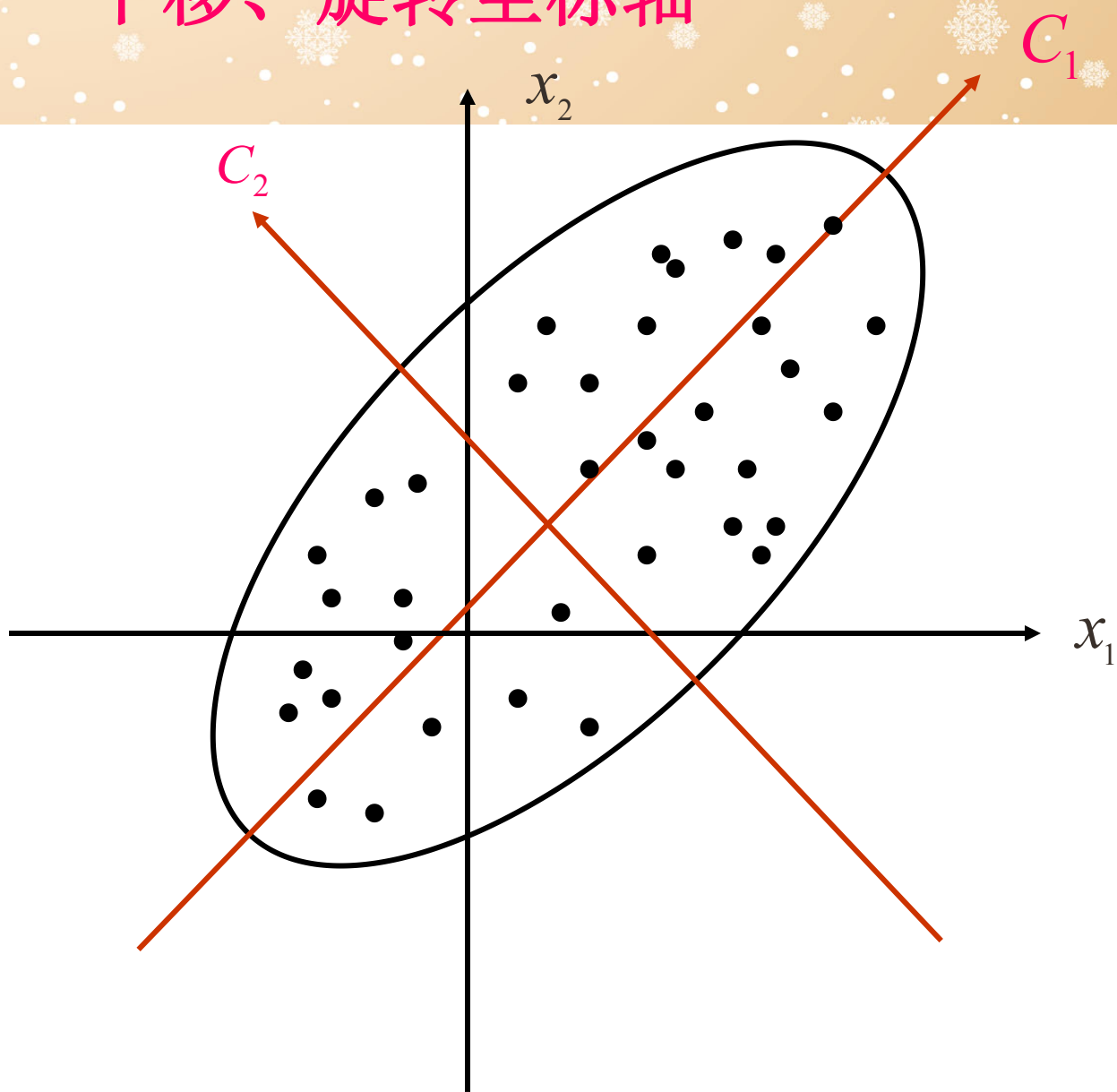
- 一个度量指标的好坏除了可靠、真实之外，还必须能充分反映个体间的变异。
- 如果有一项指标，不同个体的取值都大同小异，那么该指标不能用来区分不同的个体。
- 由这一点来看，一项指标在个体间的变异越大越好。因此我们把“**变异大**”作为“**好**”的标准来寻求综合指标。



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。

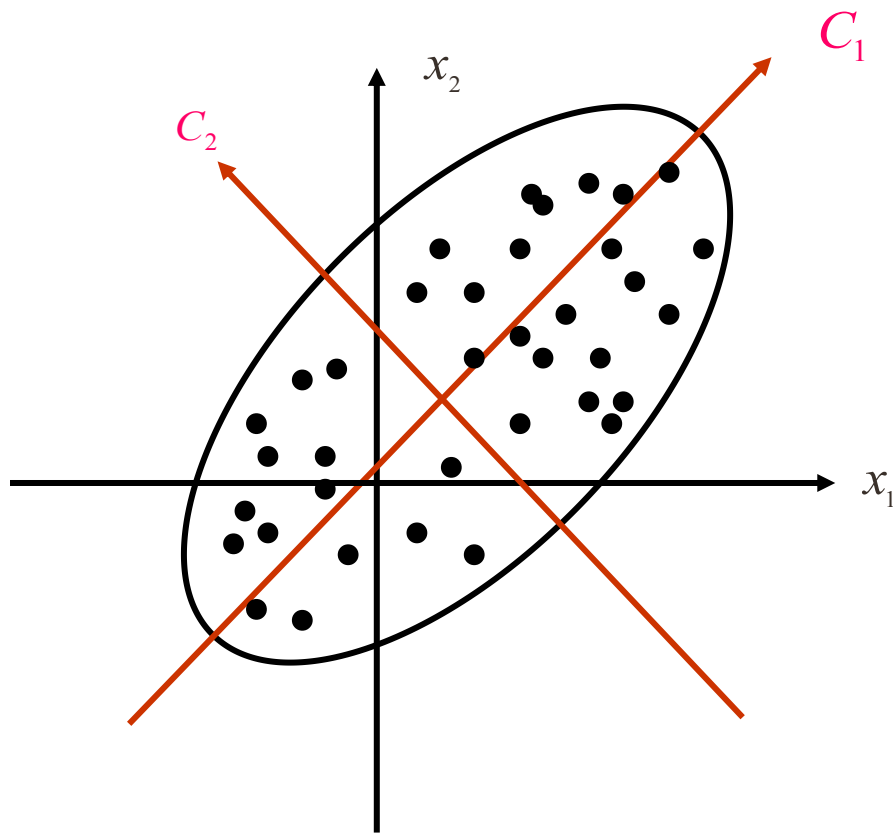
平移、旋转坐标轴

主成分分析的几何解释





- 如果我们将 x_1 轴和 x_2 轴先平移，再同时按逆时针方向旋转 θ 角度，得到新坐标轴 C_1 和 C_2 。 C_1 和 C_2 是两个新变量。

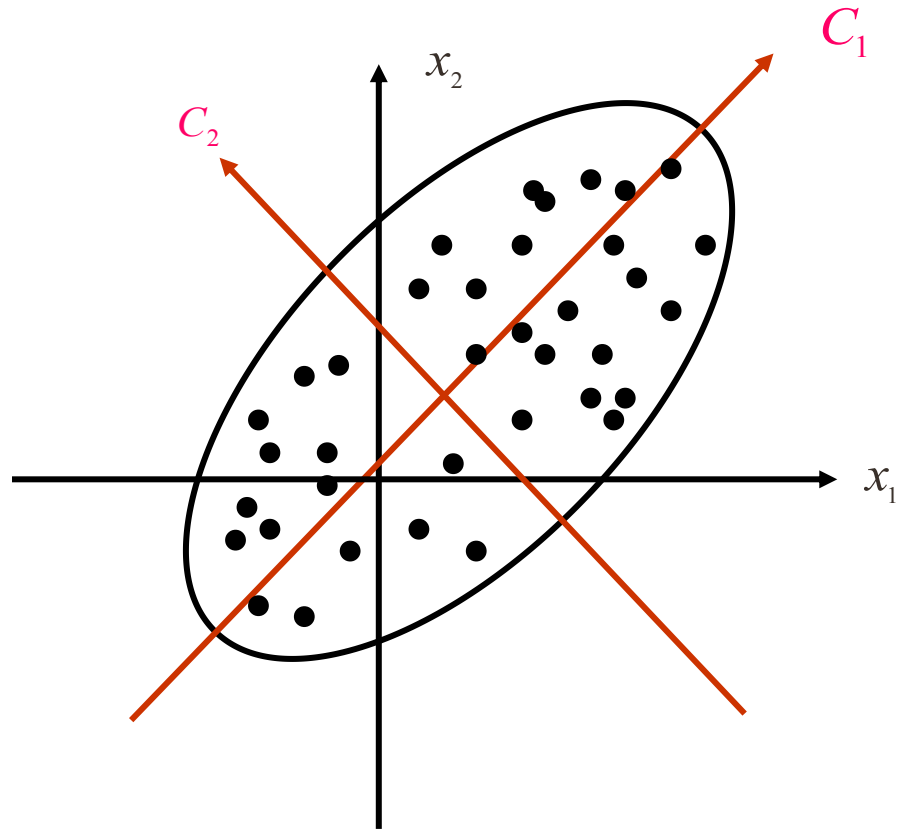




根据旋转变换的公式：

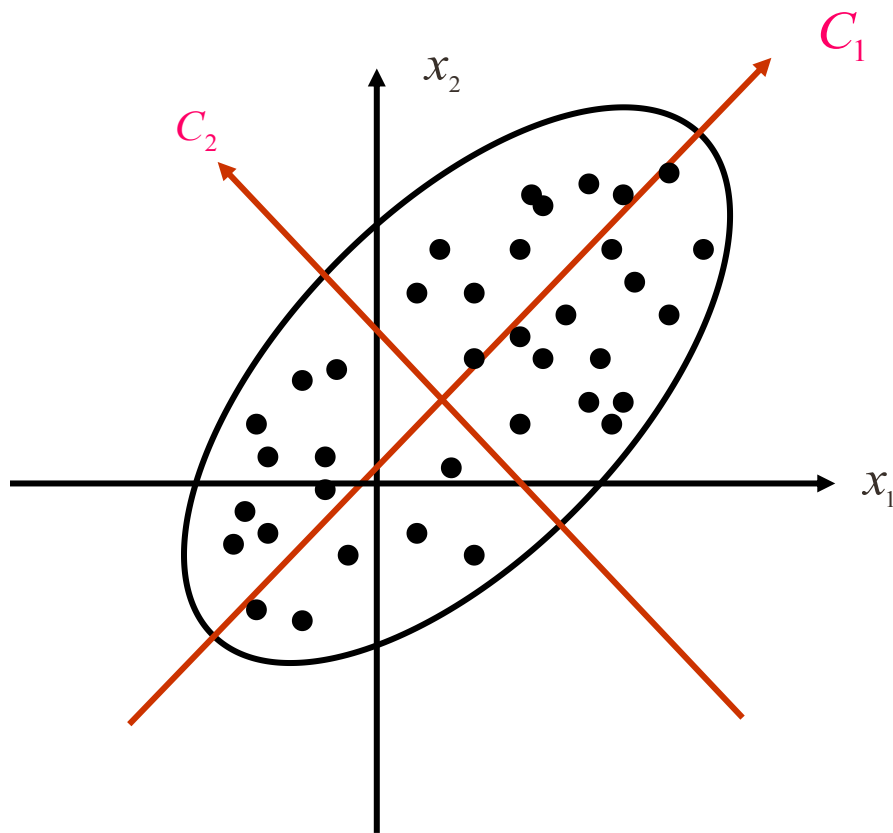
$$\begin{cases} C_1 = x_1 \cos \theta + x_2 \sin \theta \\ C_2 = -x_1 \sin \theta + x_2 \cos \theta \end{cases}$$

$$\begin{pmatrix} C_1 \\ C_2 \end{pmatrix} = \begin{pmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$



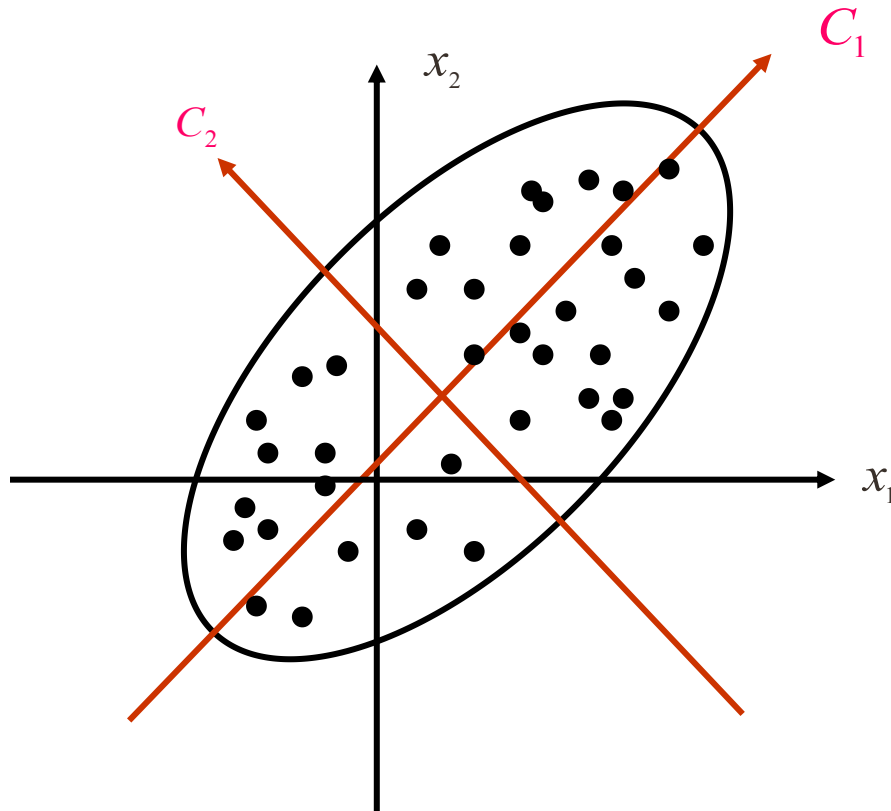


旋转变换的目的是为了使得 n 个样品点在 C_1 轴方向上的离散程度最大，即 C_1 的方差最大。变量 C_1 代表了原始数据的绝大部分信息，在研究某问题时，即使不考虑变量 C_2 也无损大局。经过上述旋转变换原始数据的大部分信息集中到 C_1 轴上，对数据中包含的信息起到了浓缩作用。



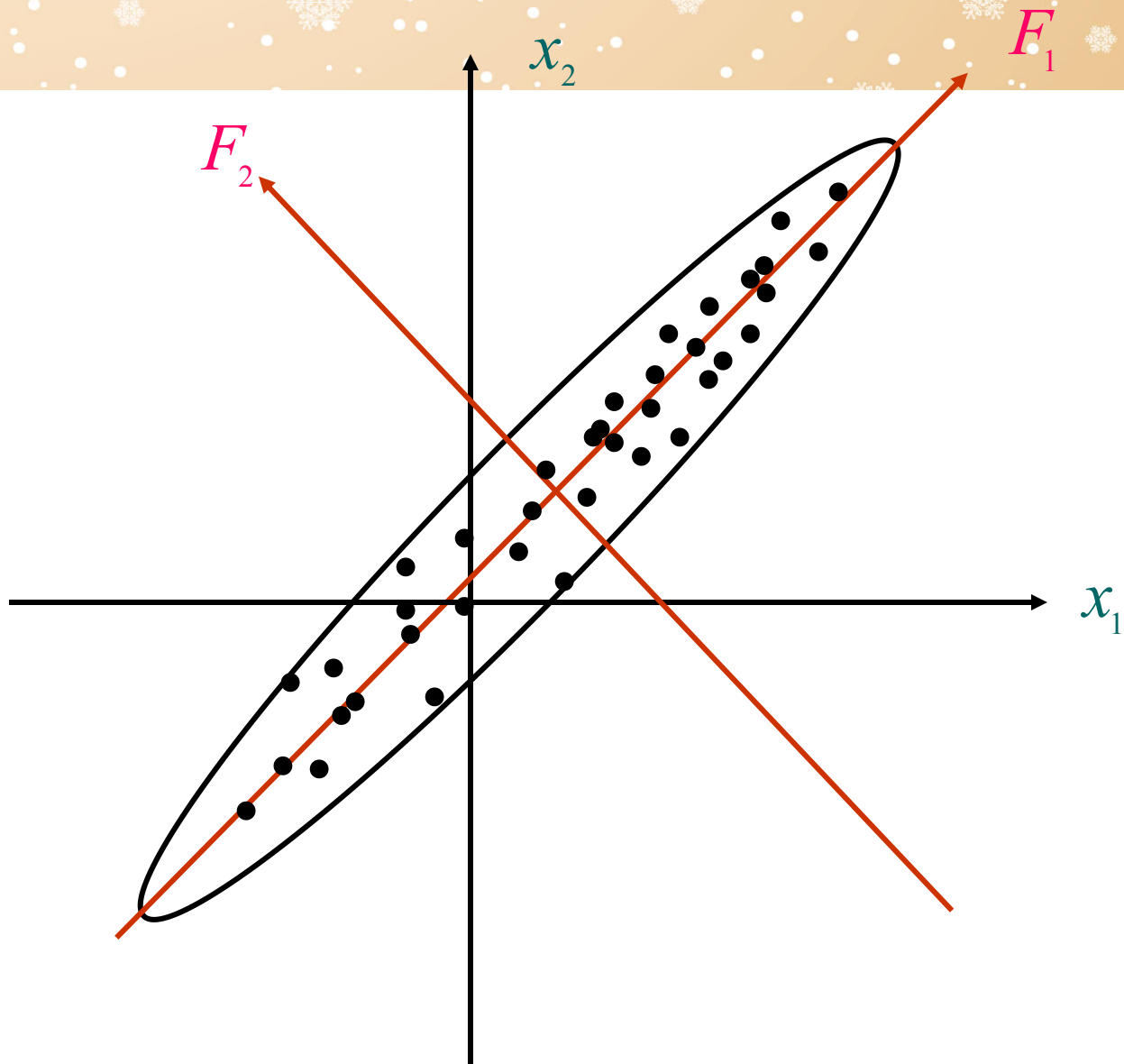


C_1 , C_2 除了可以对包含在 x_1 , x_2 中的信息起着浓缩作用之外, 还具有不相关的性质, 这就使得在研究复杂的问题时避免了信息重叠所带来的虚假性。二维平面上的个点的方差大部分都归结在 C_1 轴上, 而 C_2 轴上的方差很小。 C_1 和 C_2 称为原始变量 x_1 和 x_2 的综合变量。 C 简化了系统结构。



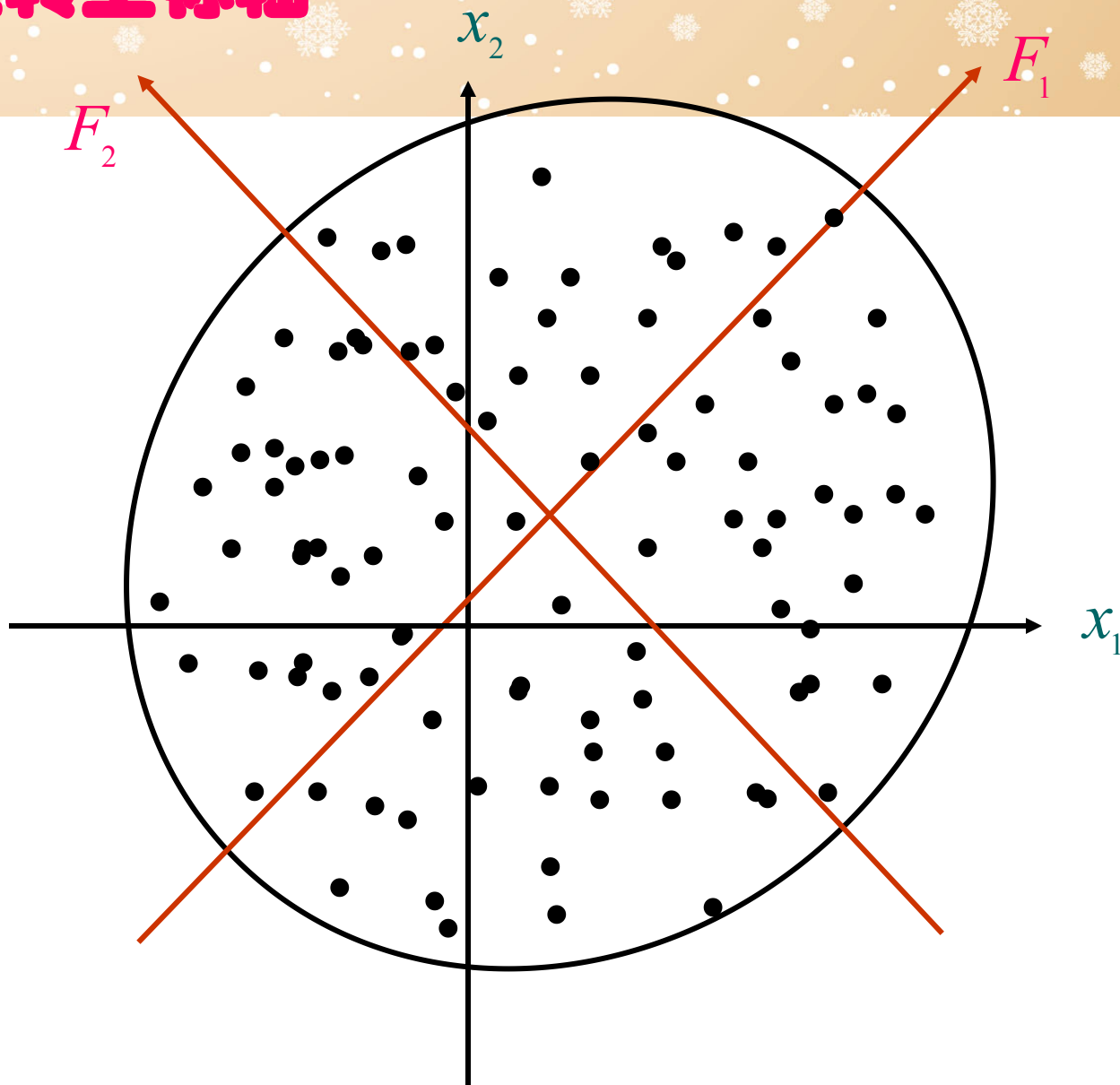
平移、旋转坐标轴

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





2

主成分分析的解法



主成分分析的基本原理

• 主成分分析的意义

假设

- n 个地理区域， p 个指标，则有 np 个观测数据。

用较少的综合指标代表原来较多的指标

- 能尽量多的反映原有信息；
- 彼此之间独立。
- 选取原则：原指标的线性组合。



主成分分析的基本原理

- 主成分分析的数学模型

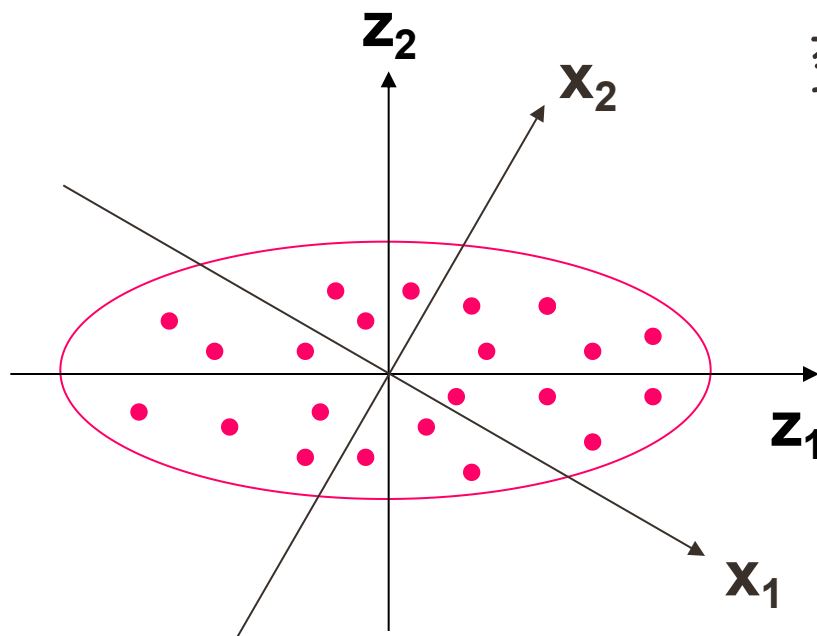
原始数据矩阵

$$X = \begin{matrix} & \begin{matrix} \mathbf{1} & \mathbf{2} & \cdots & \mathbf{p} & \text{指标} \end{matrix} \\ \begin{matrix} \mathbf{1} \\ \mathbf{2} \\ \vdots \\ \mathbf{n} \end{matrix} \begin{matrix} \text{地} \\ \text{区} \end{matrix} & \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \end{matrix}$$



§2 主成分分析的基本原理

- 主成分分析的数学模型



变换后坐标的性质：

n 个点的坐标 z_1 、 z_2 的相关几乎为0；

二维平面上 n 个点的波动大部分可归结为 z_1 轴上的波动，而 z_2 轴上的波动较小。



主成分分析的基本原理

- 主成分分析的数学模型

则称 z_1 、 z_2 是原指标 x_1 、 x_2 的主成分。

若长轴方向反映整个信息的75%，则 z_1 就是 x_1 和 x_2 的综合指标。

$$z_1 = l_{11}x_1 + l_{12}x_2$$

式中： l_{11} 、 l_{12} 为 x_1 和 x_2 对 z_1 这个综合指标的权值，
或变量 x_1 和 x_2 的回归系数。



主成分分析的基本原理

- 主成分分析的数学模型

长轴为第一主成分 z_1 ，短轴为第二主成分 z_2

数据点对于原指标和对主成分的值分别为：

$$\{x_{a1}\}, \{x_{a2}\} \quad \{z_{a1}\}, \{z_{a2}\}$$

则有：

$$\sum_{a=1}^n (x_{a1} - \bar{x}_1)^2 + \sum_{a=1}^n (x_{a2} - \bar{x}_2)^2 = \underbrace{\sum_{a=1}^n (z_{a1} - \bar{z}_1)^2}_{75\%} + \underbrace{\sum_{a=1}^n (z_{a2} - \bar{z}_2)^2}_{25\%}$$



主成分分析的基本原理

- 主成分分析的数学模型

若有 p 个指标 x_1, x_2, \dots, x_p , 综合成 m 个指标 $z_1, z_2, \dots, z_m (m \leq p)$, 可表示为:

$$\begin{cases} z_1 = l_{11}x_1 + l_{12}x_2 + \dots + l_{1p}x_p \\ z_2 = l_{21}x_1 + l_{22}x_2 + \dots + l_{2p}x_p \\ \vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \\ z_m = l_{m1}x_1 + l_{m2}x_2 + \dots + l_{mp}x_p \end{cases}$$



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

$$\text{Cov} (Z_i, Z_j) = 0, \quad i \neq j, \quad i, j = 1, 2, \dots, p$$

主成分的方差依次递减，重要性依次递减，即

$$\text{Var} (Z_1) \geq \text{Var} (Z_2) \geq \cdots \geq \text{Var} (Z_p)$$



§2 主成分分析的基本原理

- 主成分分析的数学模型

从几何上看，找主成分的问题就是找出 p 维空间中椭球体的主轴问题，就是要在 $x_1 \sim x_p$ 的相关矩阵中 m 个较大特征值所对应的特征向量。



主成分的推导

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

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

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 $\mathbf{u}_1, \dots, \mathbf{u}_p$

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

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 $\mathbf{U}'\mathbf{U} = \mathbf{U}\mathbf{U}' = \mathbf{I}$



补：矩阵的特征值与特征向量

- 一、矩阵的特征值

定义：设 A 为 n 阶矩阵， λ 是一个数，如果方程 $Ax = \lambda x$

(1) 存在非零解向量，则称 λ 为 A 的一个特征值，相应的非零解向量 x 称为与特征值 λ 对应的特征向量。

将 (1) 式改写为

$$(\lambda I - A)x = 0$$



补：矩阵的特征值与特征向量

- 一、矩阵的特征值

对应的 n 元齐次线性方程组

$$\begin{cases} (\lambda - a_{11})x_1 - a_{12}x_2 - \cdots - a_{1n}x_n = 0 \\ -a_{21}x_1 + (\lambda - a_{22})x_2 - \cdots - a_{2n}x_n = 0 \\ \vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \\ -a_{n1}x_1 - a_{n2}x_2 - \cdots + (\lambda - a_{nn})x_n = 0 \end{cases}$$

存在非零解的充要条件为 $|\lambda I - A| = 0$



补：矩阵的特征值与特征向量

- 一、矩阵的特征值

$\lambda I - A$ 为 A 的特征矩阵；

$|\lambda I - A|$ 为 λ 的 n 次多项式，称为 A 的特征多项式；

$|\lambda I - A| = 0$ 称为 A 的特征方程。



补：矩阵的特征值与特征向量

- 例：求矩阵A的特征值与特征向量

$$A = \begin{bmatrix} 3 & 1 \\ 5 & -1 \end{bmatrix}$$

特征方程为 $|\lambda I - A| = \begin{vmatrix} \lambda - 3 & -1 \\ -5 & \lambda + 1 \end{vmatrix} = 0$

化简得 $(\lambda - 4)(\lambda + 2) = 0$

故 $\lambda_1 = 4$, $\lambda_2 = -2$ 是A的两个特征值。



补：矩阵的特征值与特征向量

- 例：求矩阵A的特征值与特征向量

$$(1) \quad \lambda_1 = 4 \quad \begin{cases} x_1 - x_2 = 0 \\ -5x_1 + 5x_2 = 0 \end{cases} \text{得基础解系} \quad c \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$(1) \quad \lambda_2 = -2 \quad \begin{cases} -5x_1 - x_2 = 0 \\ -5x_1 - x_2 = 0 \end{cases} \text{得基础解系} \quad c \begin{bmatrix} 1 \\ -5 \end{bmatrix}$$



补：矩阵的特征值与特征向量

- 二、特征值与特征向量的基本性质

n 阶矩阵 A 与它的转置矩阵 A^T 有相同的特征值。

n 阶矩阵 A 互不相同的特征值 $\lambda_1, \lambda_2, \dots, \lambda_m$ 对应的特征向量 x_1, x_2, \dots, x_m 线性无关。



补：随机变量的数字特征

- 协方差

$$\text{cov}(x_1, x_2) = \frac{1}{n} \sum_{i=1}^n (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)$$



§3 主成分分析的解法

p 阶方阵的特征向量给出椭圆**主轴的方向**，
对应的特征值表示**主轴的长度**。

主成分分析的实质就是求出方差-协方差矩阵的特征值及其对应的特征向量。



第一主成分：特征向量为

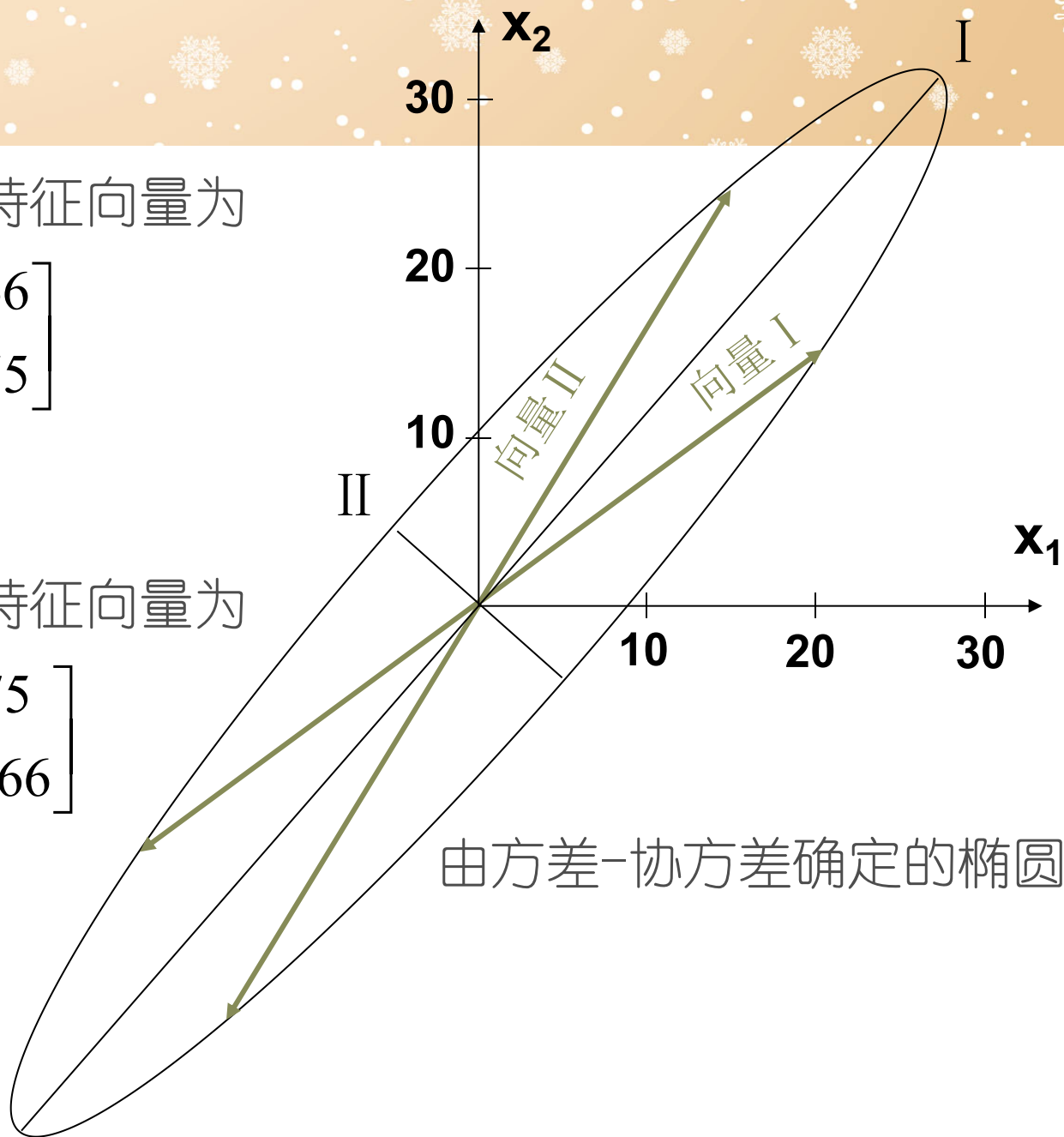
$$I = \begin{bmatrix} 0.66 \\ 0.75 \end{bmatrix}$$

$$\lambda_I = 37.9$$

第二主成分：特征向量为

$$II = \begin{bmatrix} 0.75 \\ -0.66 \end{bmatrix}$$

$$\lambda_{II} = 6.5$$





变量 x_1 的方差: 20.3

变量 x_2 的方差: 24.1

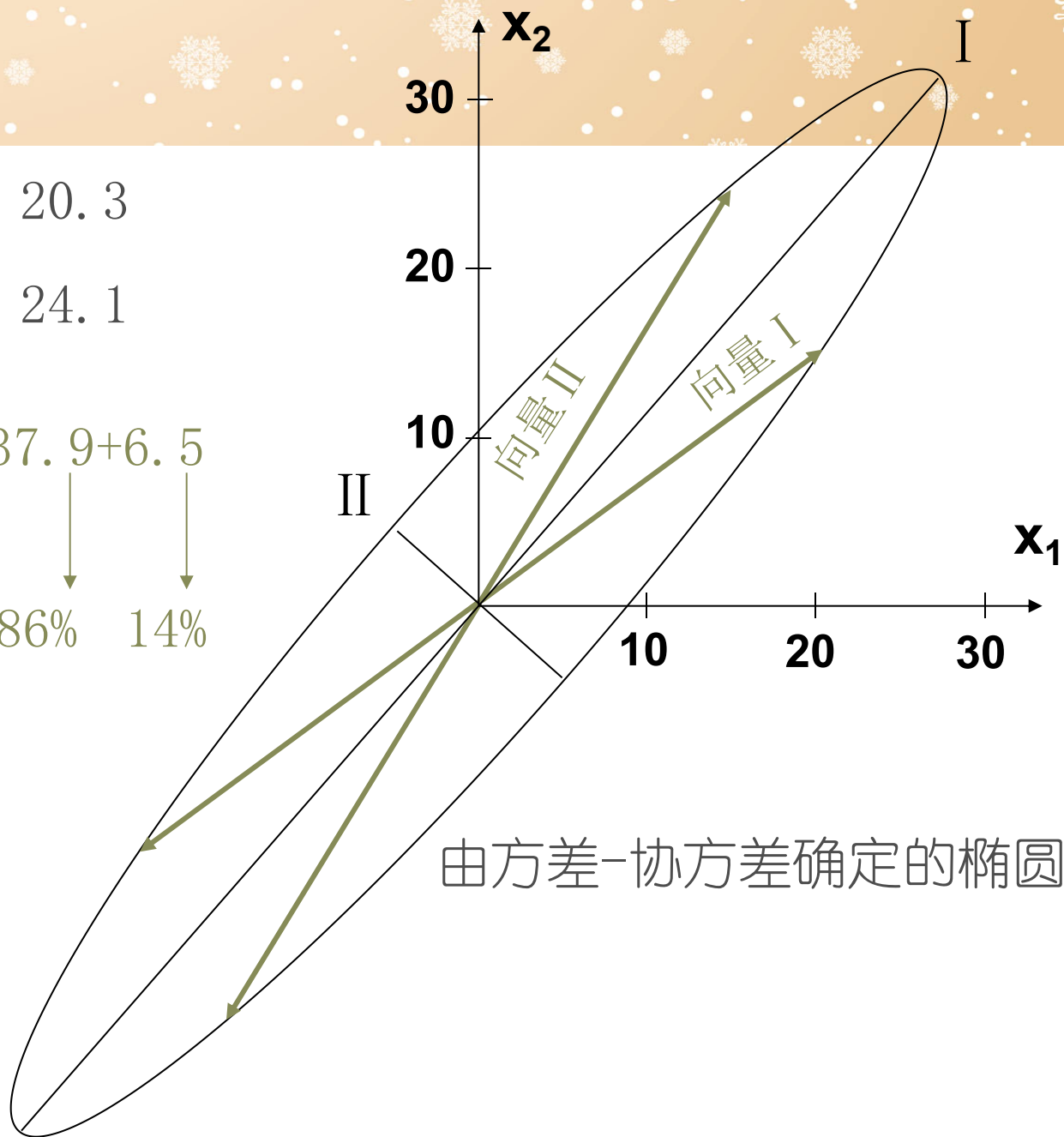
$$20.3 + 24.1 = 44.4 = 37.9 + 6.5$$

↓
46%

↓
54%

↓
86%

↓
14%





§3 主成分分析的解法

- 主成分 z_1 , z_2 的表达式

$$z_1 = 0.66x_1 + 0.75x_2$$

$$z_2 = 0.75x_1 - 0.66x_2$$

主成分得分: (P149)



主成分分析的数学模型

- 通常情况下，所分析的多个变量具有不同量纲或均数/方差相差很大，不适于用协方差矩阵做主成分分析，而采用基于相关系数矩阵的主成分分析。
- 首先将原变量标准化。设有 n 个样本， $x_1, x_2 \dots x_p$ 为 p 个原指标变量，经过标准化后得到标准化变量 $X_1, X_2 \dots X_p$ ：

$$X_i = \frac{x_i - \bar{x}_i}{\sigma_i} \quad i=1, 2, \dots, p$$

$$\bar{x}_i = \frac{1}{n} \sum_{a=1}^n x_{ai} \quad \sigma_i = \sqrt{\frac{\sum_{a=1}^n (x_{ai} - \bar{x}_i)^2}{n}}$$



§3 主成分分析的解法

- 主成分分析的步骤

- 2. 计算相关系数矩阵R

$$r_{ij} = \frac{\frac{1}{n} \sum_{a=1}^n (x_{ai} - \bar{x}_i)(x_{aj} - \bar{x}_j)}{\sigma_i \sigma_j} = \frac{1}{n} \sum_{a=1}^n x_{ai}^* x_{aj}^*$$

- 3. 计算特征值和特征向量

$$|R - \lambda I| = 0 \quad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$$

对应于 λ_k 的特征向量 $l_k = [l_{k1}, l_{k2}, \cdots, l_{kp}]^T$



§3 主成分分析的解法

- 主成分分析的步骤

4. 计算第k个特征值的贡献率和累计贡献率

$$\frac{\lambda_k}{\sum_{i=1}^p \lambda_i} \quad \sum_{j=1}^k \frac{\lambda_j}{\sum_{i=1}^p \lambda_i}$$

一般取累计贡献率达**85%~95%**的特征值对应的主成分即可。

$$\lambda_1, \lambda_2, \dots, \lambda_m (m \leq p)$$



我们作如下定义：

(1) 若 $C_1 = w_{11}X_1 + w_{12}X_2 + \dots + w_{1p}X_p$,

且使 $Var(C_1)$ 最大, 则称 C_1 为第一主成分;

但系数 w 若无限限制可使 $Var(C_1)$ 无限大, 故加约束条件:

$$w_{11}^2 + w_{12}^2 + \dots + w_{1p}^2 = 1$$

组合系数 ($w_{11}, w_{12}, \dots, w_{1p}$) 可看作一个向量, 代表 p 维空间中的一个方向, 相当于全部 n 个个体在该方向上的一个投影。要求 $Var(C_1)$ 最大就是要找一个最 “好” 的方向, 使得所有个体在该方向上的投影最为分散。



§3 主成分分析的解法

- 主成分分析的步骤

5. 计算主成分载荷

$$P(Z_k, x_i) = \sqrt{\lambda_k} l_{ki} \quad (i = 1, 2, \dots, p, k = 1, 2, \dots, m)$$

6. 计算主成分得分

$$z_1 = l_{11}x_1^* + l_{12}x_2^* + \dots + l_{1p}x_p^*$$

$$z_2 = l_{21}x_1^* + l_{22}x_2^* + \dots + l_{2p}x_p^*$$

...

$$z_m = l_{m1}x_1^* + l_{m2}x_2^* + \dots + l_{mp}x_p^*$$

$$\begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1m} \\ z_{21} & z_{22} & \cdots & z_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{mm} \end{bmatrix}$$



主成分的性质：

主成分 C_1, C_2, \dots, C_p 具有如下几个性质：

(1) 主成分间互不相关，即对任意 i 和 j ， C_i 和 C_j 的相关系数 $\text{Corr}(C_i, C_j)=0 \quad i \neq j$

(2) 组合系数 $(w_{i1}, w_{i2}, \dots, w_{ip})$ 构成的向量为单位向量

,

$$w_{i1}^2 + w_{i2}^2 + \dots + w_{ip}^2 = 1$$

(3) 各主成分的方差是依次递减的，即
 $\text{Var}(C_1) \geq \text{Var}(C_2) \geq \dots \geq \text{Var}(C_p)$



(4) 总方差不增不减， 即

$$\begin{aligned} & Var(C_1) + Var(C_2) + \cdots + Var(C_p) \\ &= Var(x_1) + Var(x_2) + \cdots + Var(x_p) \\ &= p \end{aligned}$$

这一性质说明，主成分是原变量的线性组合，是对原变量信息的一种重组，主成分不增加总信息量，也不减少总信息量。

(5) 主成分和原变量的相关系数 $Corr(C_i, x_j) = w_{ij} \sqrt{Var(C_i)}$

$$= w_{ij} \sqrt{\lambda_i}$$



(6) 令 X_1, X_2, \dots, X_p 的相关矩阵为 R , $(w_{i1}, w_{i2}, \dots, w_{ip})$ 则是相关矩阵 R 的第 i 个特征向量(eigenvector)。而且, 特征值 λ_i 就是第 i 主成分的方差, 即

$$\text{Var}(C_i) = \lambda_i$$

其中 λ_i 为相关矩阵 R 的第 i 个特征值(eigenvalue)

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$$

(7) 第 i 个主成分对所有原变量的贡献为:

$$\sum_{j=1}^p r_{C_i, x_j}^2 = \sum_{j=1}^p w_{ij}^2 \lambda_i = \lambda_i$$

(8) 所有主成分对原变量 x_j 的贡献为:

$$h_j^2 = \sum_{i=1}^p r_{C_i, x_j}^2 = \sum_{i=1}^p w_{ij}^2 \lambda_i$$



求主成分的步骤

- 1. 计算相关系数矩阵 R

$$R = \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1p} \\ r_{21} & r_{22} & \dots & r_{2p} \\ \dots & \dots & \dots & \dots \\ r_{p1} & r_{p2} & \dots & r_{pp} \end{bmatrix}$$



- 2.解特征方程 $|\mathbf{R} - \lambda \mathbf{I}| = 0$ ，求出相关阵 \mathbf{R} 的特征根 (eigenvalue) λ_i ，且按从大到小顺序排列：
- $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ ，
- 3.求矩阵 \mathbf{R} 关于 λ_i 的满足正规条件的特征向量 (eigenvector):
- $\mathbf{L}_i = (l_{i1}, l_{i2}, \dots, l_{ip})$
- 特征向量即为主成分系数。
- 当变量较多时，特征根的计算较复杂，需借助计算机软件实现。



主成分的数目的选取

如前所述， p 个随机变量，便有 p 个主成分。由于总方差不增不减， C_1 ， C_2 等前几个综合变量的方差较大，而 C_p ， C_{p-1} 等后几个综合变量的方差较小。一般来说，只有前几个综合变量才称得上主(要)成份，后几个综合变量实为“次”(要)成份。实践中总是保留前几个，忽略后几个。



- 保留多少个主成分主要考虑保留部分的累积方差在方差总和中所占百分比(即累积贡献率)，它标志着前几个主成分概括信息之多寡。实践中，一般推荐达到80%的累积方差即可。常用的判断方法有：
 1. 特征值准则：取特征值 >1 的主成分。是SPSS软件默认的方法。
 2. 累积方差比例原则：一般推荐累积方差比例达到80%以上时，即可停止选择主成分。
 3. 利用碎石图：将主成分按特征根从大到小排列，画出特征根随主成分个数变化的散点图，根据图的形状来判断保留主成分的个数。曲线开始变平的前一个点（拐点）认为是提取的最大主成分数。也就是根据特征根的变化速率来确定。



例：主成分分析在农业区划中的应用

6	68.337	2.032	76.204	1540.29	216.39	8.128	4.065	0.011	4.861
7	95.416	0.801	71.106	926.35	291.52	8.135	4.063	0.012	4.862
8	62.901	1.652	73.307	1501.24	225.25	18.352	2.645	0.034	3.201
9	86.624	0.841	68.904	897.36	196.37	16.861	5.176	0.055	6.167
10	91.394	0.812	66.502	911.24	226.51	18.279	5.643	0.076	4.477
11	76.912	0.858	50.302	103.52	217.09	19.793	4.881	0.001	6.165
12	51.274	1.041	64.609	968.33	181.38	4.005	4.066	0.015	5.402
13	68.831	0.836	62.804	957.14	194.04	9.11	4.484	0.002	5.79
14	77.301	0.623	60.102	824.37	188.09	19.409	5.721	5.055	8.413
15	76.948	1.022	68.001	1255.42	211.55	11.102	3.133	0.01	3.425
16	99.265	0.654	60.702	1251.03	220.91	4.383	4.615	0.011	5.593
17	118.505	0.661	63.304	1246.47	242.16	10.706	6.053	0.154	8.701
18	141.473	0.737	54.206	814.21	193.46	11.419	6.442	0.012	12.945
19	137.761	0.598	55.901	1124.05	228.44	9.521	7.881	0.069	12.654
20	117.612	1.245	54.503	805.67	175.23	18.106	5.789	0.048	8.461
21	122.781	0.731	49.102	1313.11	236.29	26.724	7.162	0.092	10.078



步骤如下：

(1) 将表1中的数据作标准差标准化处理，然后将它们代入公式（4）计算相关系数矩阵（见表2）。

表2 相关系数矩阵

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
x_1	1	-0.327	-0.714	-0.336	0.309	0.408	0.79	0.156	0.744
x_2	-0.33	1	-0.035	0.644	0.42	0.255	0.009	-0.078	0.094
x_3	-0.71	-0.035	1	0.07	-0.74	-0.755	-0.93	-0.109	-0.924
x_4	-0.34	0.644	0.07	1	0.383	0.069	-0.05	-0.031	0.073
x_5	0.309	0.42	-0.74	0.383	1	0.734	0.672	0.098	0.747
x_6	0.408	0.255	-0.755	0.069	0.734	1	0.658	0.222	0.707
x_7	0.79	0.009	-0.93	-0.046	0.672	0.658	1	-0.03	0.89
x_8	0.156	-0.078	-0.109	-0.031	0.098	0.222	-0.03	1	0.29
x_9	0.744	0.094	-0.924	0.073	0.747	0.707	0.89	0.29	1



(2) 由相关系数矩阵计算特征值，以及各个主成分的贡献率与累计贡献率（见表3）。由表3可知，第一，第二，第三主成分的累计贡献率已高达86.596%（大于85%），故只需要求出第一、第二、第三主成分 z_1 ， z_2 ， z_3 即可。

表3 特征值及主成分贡献率

主成分	特征值	贡献率(%)	累积贡献率(%)
z_1	4.661	51.791	51.791
z_2	2.089	23.216	75.007
z_3	1.043	11.589	86.596
z_4	0.507	5.638	92.234
z_5	0.315	3.502	95.736
z_6	0.193	2.14	97.876
z_7	0.114	1.271	99.147
z_8	0.0453	0.504	99.65
z_9	0.0315	0.35	100



(3) 对于特征值 $\lambda_1=4.6610$, $\lambda_2=2.0890$,
 $\lambda_3=1.0430$ 分别求出其特征向量 l_1 , l_2 , l_3 。

表4 主成分载荷

	Z_1	Z_2	Z_3	占方差的百分数 (%)
x_1	0.739	-0.532	-0.0061	82.918
x_2	0.123	0.887	-0.0028	80.191
x_3	-0.964	0.0096	0.0095	92.948
x_4	0.0042	0.868	0.0037	75.346
x_5	0.813	0.444	-0.0011	85.811
x_6	0.819	0.179	0.125	71.843
x_7	0.933	-0.133	-0.251	95.118
x_8	0.197	-0.1	0.97	98.971
x_9	0.964	-0.0025	0.0092	92.939

上述计算过程，可以借助于SPSS软件系统实现。



分析:

- ①第一主成分 z_1 与 x_1, x_5, x_6, x_7, x_9 呈显出较强的正相关, 与 x_3 呈显出较强的负相关, 而这几个变量则综合反映了生态经济结构状况, 因此可以认为第一主成分 z_1 是生态经济结构的代表。
- ②第二主成分 z_2 与 x_2, x_4, x_5 呈显出较强的正相关, 与 x_1 呈显出较强的负相关, 其中, 除了 x_1 为人口总数外, x_2, x_4, x_5 都反映了人均占有资源量的情况, 因此可以认为第二主成分 z_2 代表了人均资源量。



③第三主成分 z_3 ，与 x_8 呈显出的正相关程度最高，其次是 x_6 ，而与 x_7 呈负相关，因此可以认为第三主成分在一定程度上代表了农业经济结构。

显然，用三个主成分 z_1 、 z_2 、 z_3 代替原来9个变量（ x_1 ， x_2 ，...， x_9 ），描述农业生态经济系统，可以使问题更进一步简化、明了。



例：测得某地19-22岁年龄的部分城市男生身体形态指标：身高（ x_1 , cm）、坐高（ x_2 , cm）、体重（ x_3 , kg）、胸围（ x_4 , cm）、肩宽（ x_5 , cm）、骨盆宽（ x_6 , cm）。试进行主成分分析。



1:1 173.

	x1	x2
1	173.28	93.62
2	172.09	92.83
3	171.46	92.73
4	170.08	92.25
5	170.61	92.36
6	171.69	92.85
7	171.46	92.93
8	171.60	93.28
9	171.60	92.26
10	171.16	92.62
11	170.04	92.17
12	170.27	91.94
13	170.61	92.50
14	171.39	92.44
15	171.83	92.79
16	171.36	92.53
17	171.24	92.61
18	170.49	92.03
19	169.43	91.67
20	168.57	91.40
21	170.43	92.38
22	169.88	91.89

- Reports
- Descriptive Statistics
- Tables
- Compare Means
- General Linear Model
- Mixed Models
- Correlate
- Regression
- Loglinear
- Classify
- Data Reduction**
- Scale
- Nonparametric Tests
- Time Series
- Survival
- Multiple Response
- Missing Value Analysis...
- Complex Samples

	x6	var	var	var
7	27.51			
2	27.82			
3	27.46			
3	27.29			
8	27.14			
9	27.10			
9	27.18			
4	27.65			
6	26.81			
2	27.36			
50.92	85.37	38.63	26.47	
56.85	85.35	38.58	27.03	
58.39	87.09	38.23	27.04	
57.69	83.98	39.04	27.07	
57.56	87.18	38.54	27.57	
57.22	83.87	38.41	26.60	
55.96	83.02	38.74	26.97	
57.87	84.87	38.78	27.37	
56.87	86.34	38.37	27.19	
55.07	86.77	38.17	27.10	

- Factor...**
- Correspondence Analysis...
- Optimal Scaling...



1: x1

173.28

	x1	x2	x3	x4	x5	x6	var	var	var	var
1	173.28	92.61	57.69	85.87	37.61	26.67				
2										
3										
4										
5										
6										
7										
8										
9										
10										
11										
12										
13										
14										
15										
16	171.58	92.33	58.33	85.87	37.61	26.67				
17	171.24	92.61	57.69	85.87	37.61	26.67				
18	170.49	92.03	57.56	85.87	37.61	26.67				
19	169.43	91.67	57.22	85.87	37.61	26.67				
20	168.57	91.40	55.96	85.87	37.61	26.67				
21	170.43	92.38	57.87	85.87	37.61	26.67				
22	169.88	91.89	56.87	85.87	37.61	26.67				
23	167.94	90.91	55.97	85.87	37.61	26.67				
24	168.82	91.30	56.07	85.87	37.61	26.67				
25	168.02	91.26	55.28	85.63	39.66	28.07				
26	167.87	91.96	55.79	84.97	38.70	26.53				

Factor Analysis

Variables:

- # 身高 (x1) [x1]
- # 坐高 (x2) [x2]
- # 体重 (x3) [x3]
- # 胸围 (x4) [x4]
- # 肩宽 (x5) [x5]
- # 骨盆宽 (x6) [x6]

Selection:

OK Paste Reset Cancel Help

Descriptives... Extraction... Rotation...

Factor Analysis: Extraction

Method: Principal components

Analyze:

- ☒ Correlation matrix
- ☐ Covariance matrix

Display:

- ☒ Unrotated factor solution
- ☒ Scree plot

Extract:

- ☒ Eigenvalues over: 1
- ☐ Number of factors:

Maximum Iterations for Convergence: 25

Continue Cancel Help

15 : x6 27.03

	x1	x2	x3	x4	x5	x6	var	var
1	172.88	82.68	68.48	86.78	38.87	87.54		
2								
3								
4								
5								
6								
7								
8								
9								
10								
11								
12								
13								
14								
15								
16								
17								
18								
19	169.43	91.67	57.22	83.87	38.41	26.60		
20	168.57	91.40	55.96	83.02	38.74	26.97		
21	170.43	92.38	57.87	84.87	38.78	27.37		
22	168.88	91.88	56.87	86.24	38.27	27.10		

Factor Analysis

Variables:

x1
x2

OK
Paste

Factor Analysis: Extraction

Method: Principal components

Continue

Cancel

Help

Analyze

☒ Correlation matrix
☐ Covariance matrix

Display

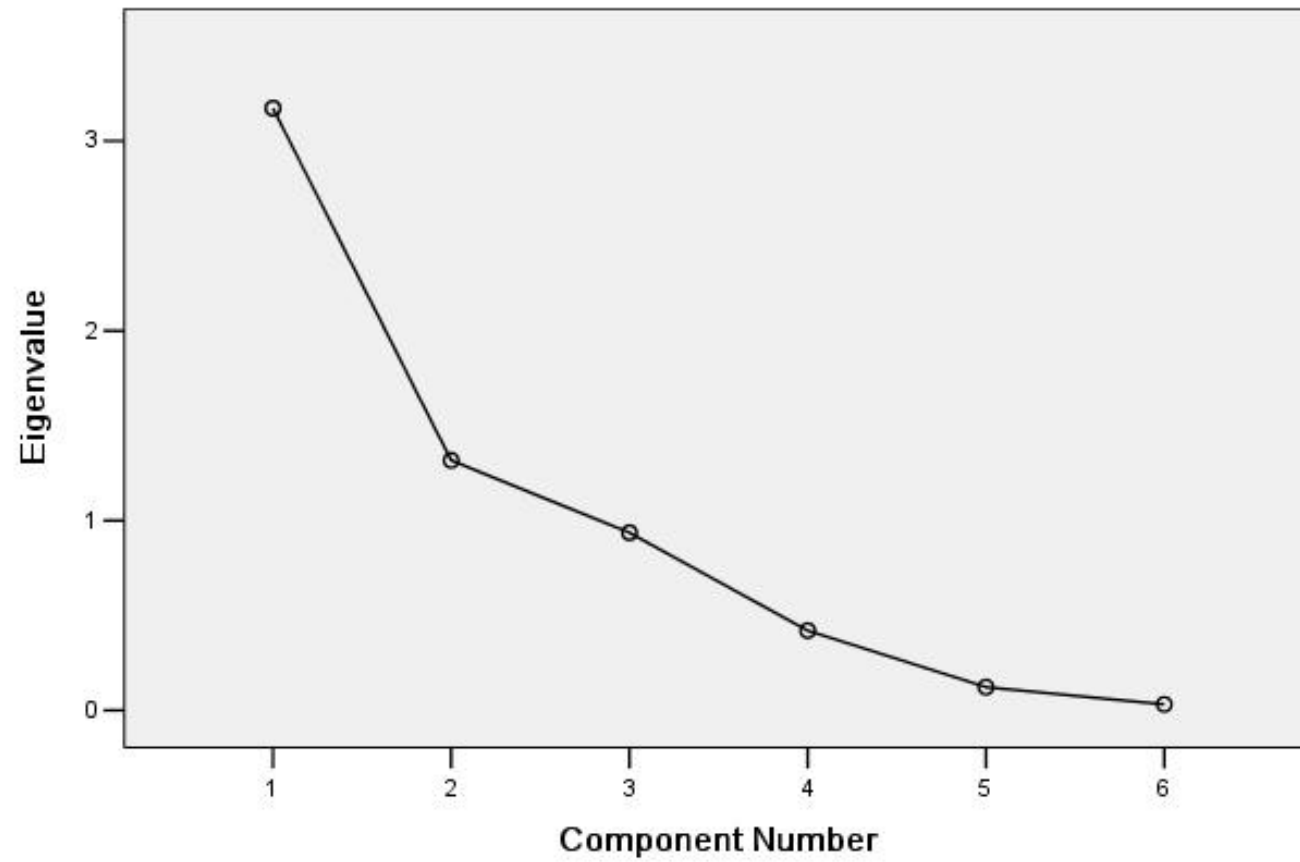
☒ Unrotated factor solution
☐ Scree plot

Extract

☐ Eigenvalues over: 1
☒ Number of factors: 6

Maximum iterations for Convergence: 25

Scree Plot



特征值、方差比例和累积贡献率

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.172	52.874	52.874	3.172	52.874	52.874
2	1.317	21.952	74.825	1.317	21.952	74.825
3	.936	15.604	90.429	.936	15.604	90.429
4	.420	7.001	97.430	.420	7.001	97.430
5	.122	2.041	99.471	.122	2.041	99.471
6	.032	.529	100.000	.032	.529	100.000

Extraction Method: Principal Component Analysis.

本例考虑保留**3**个主成分，累积贡献率可达**90%**。



- 主成分 C_i 表达式：
- **SPSS**软件不能直接给出主成分系数 w_{ij} ，经过**FACTOR** 过程产生的是因子负荷系数，但主成分分析模型需要的不是因子载荷量而是特征向量，所以还需将因子负荷系数输入数据编辑窗口，利用 “主成分相应特征根的平方根与特征向量乘积为因子负荷系数” 的性质用 **TRANSFORM——COMPUTE** 来计算特征向量，得到主成分的线性表达式。

因子负荷系数转换为主成分系数（特征向量）

Component Matrix^a

	Component					
	1	2	3	4	5	6
x1	.930	-.224	-.184	-.165	.076	.131
x2	.936	-.093	-.161	-.252	.109	-.118
x3	.910	-.208	-.101	.218	-.265	-.018
x4	.617	-.053	.717	.296	.121	-.002
x5	.336	.754	-.456	.322	.089	.003
x6	.330	.803	.379	-.301	-.110	.016

Extraction Method: Principal Component Analysis.

a. 6 components extracted.

$$\begin{aligned} C_1 &= 0.930/\sqrt{3.172}X_1 + 0.936/\sqrt{3.172}X_2 + 0.910/\sqrt{3.172}X_3 \\ &+ 0.617/\sqrt{3.172}X_4 + 0.336/\sqrt{3.172}X_5 + 0.330/\sqrt{3.172}X_6 \\ &= 0.5224X_1 + 0.5255X_2 + 0.5111X_3 + 0.3465X_4 + 0.1884X_5 + 0.1850X_6 \end{aligned}$$



特征向量:

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6
x1	0.522386	-.195138	-.190578	-.254711	0.215943	0.735666
x2	0.525457	-.081135	-.166475	-.388958	0.312044	-.664032
x3	0.511104	-.181008	-.104639	0.336278	-.756343	-.099605
x4	0.346489	-.046300	0.741022	0.456317	0.346929	-.010197
x5	0.188379	0.656718	-.471355	0.496310	0.253952	0.015620
x6	0.185038	0.699394	0.392082	-.464976	-.314787	0.087106



Q

&

A