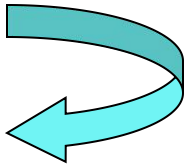
第八章 地理系统要素关系的主成分分析

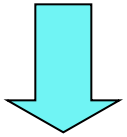
1 主成分分析的原理

- 多个指标的问题：

❖1、指标与指标可能存在相关关系



信息重叠，分析偏误



❖2、指标太多，增加问题的复杂性和分析难度

如何避免？

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

一项十分著名的工作是美国的统计学家斯通(stone)在1947年关于国民经济的研究。他曾利用美国1929一1938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，用三个新变量就取代了原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和经济发展或衰退的趋势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ΔI以及时间t因素做相关分析，得到下表：

	F1	F2	F3	I	ΔI	t
F1	1					
F2	0	1				
F3	0	0	1			
i	0.995	-0.041	0.057	1		
Δi	-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：将原来具有相关关系的多个指标简化为少数几个新的综合指标的多元统计方法。

主成分：由原始指标综合形成的几个新指标。依据主成分所含信息量的大小成为第一主成分，第二主成分等等。

主成分与原始变量之间的关系：

（1）主成分保留了原始变量绝大多数信息。

（2）主成分的个数大大少于原始变量的数目。

（3）各个主成分之间互不相关。

（4）每个主成分都是原始变量的线性组合。

主成分分析的基本思想

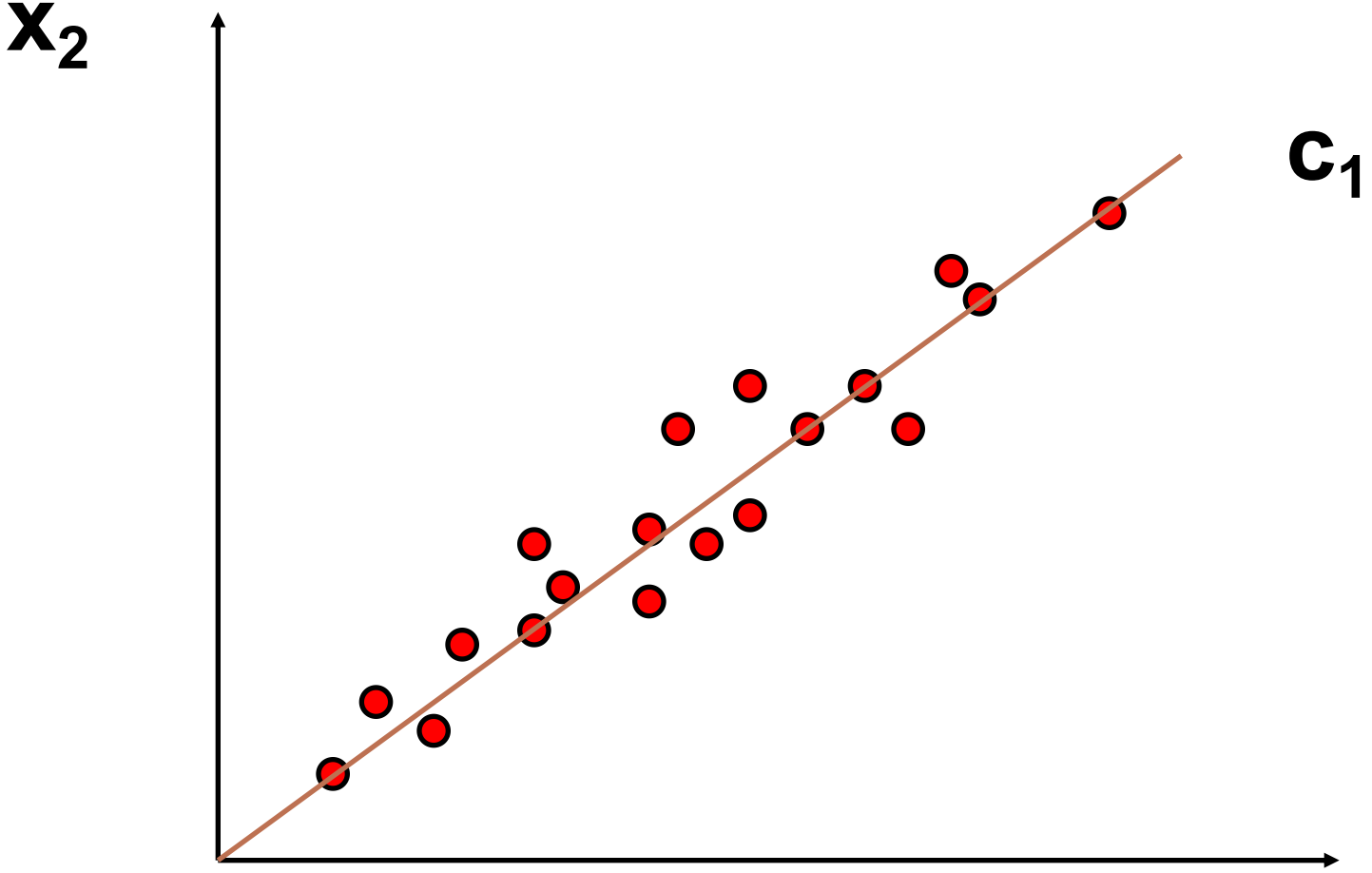
主成分分析就是把原有的多个指标转化成少数几个代表性较好的综合指标，这少数几个指标能够反映原来指标大部分的信息（85%以上），并且各个指标之间保持独立，避免出现重叠信息。主成分分析主要起着降维和简化数据结构的作用。

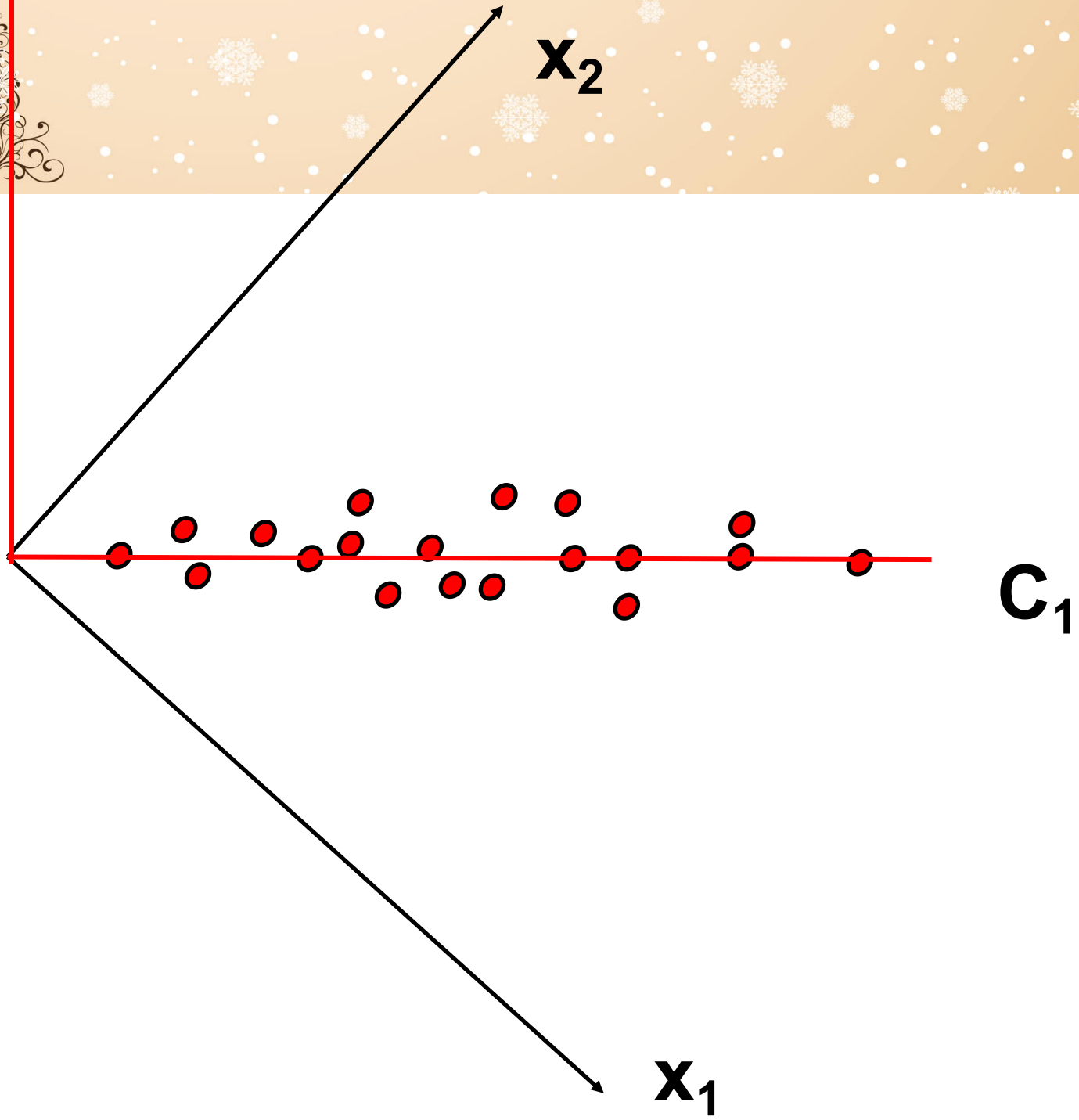
- 假设我们所讨论的实际问题中，有p个指标，我们把这p个指标看作p个随机变量，记为X_1，X_2，…，X_p，主成分分析就是要把这p个指标的问题，转变为讨论p个指标的线性组合的问题，而这些新的指标F_1，F_2，…，F_k(k≤p），按照保留主要信息量的原则充分反映原指标的信息，并且相互独立。

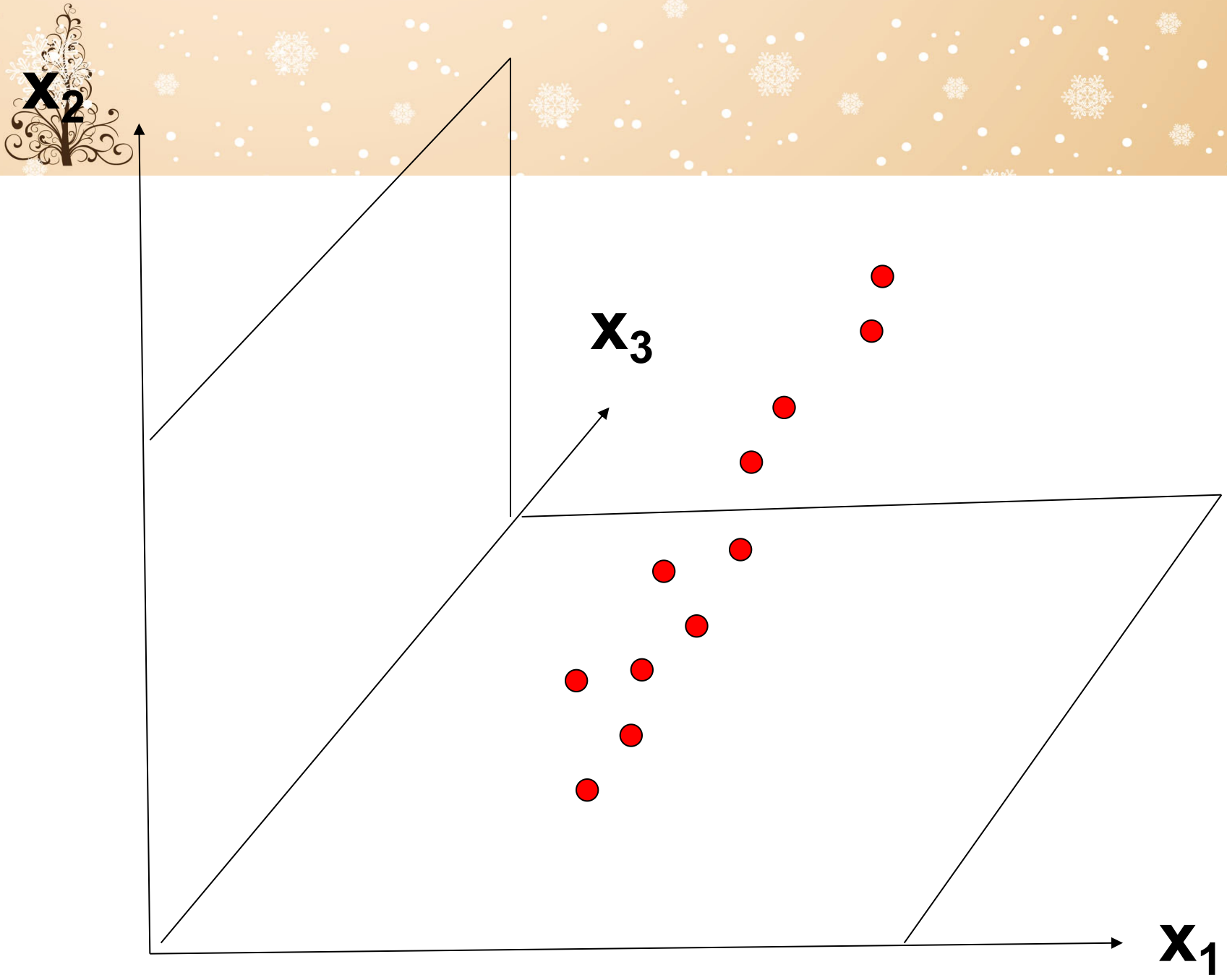
- 这种由讨论多个指标降为少数几个综合指标的过程在数学上就叫做**降维**。主成分分析通常的做法是，寻求原指标的线性组合F_i。

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
c_1


x2
C1
x1


x2
x3
x1

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。

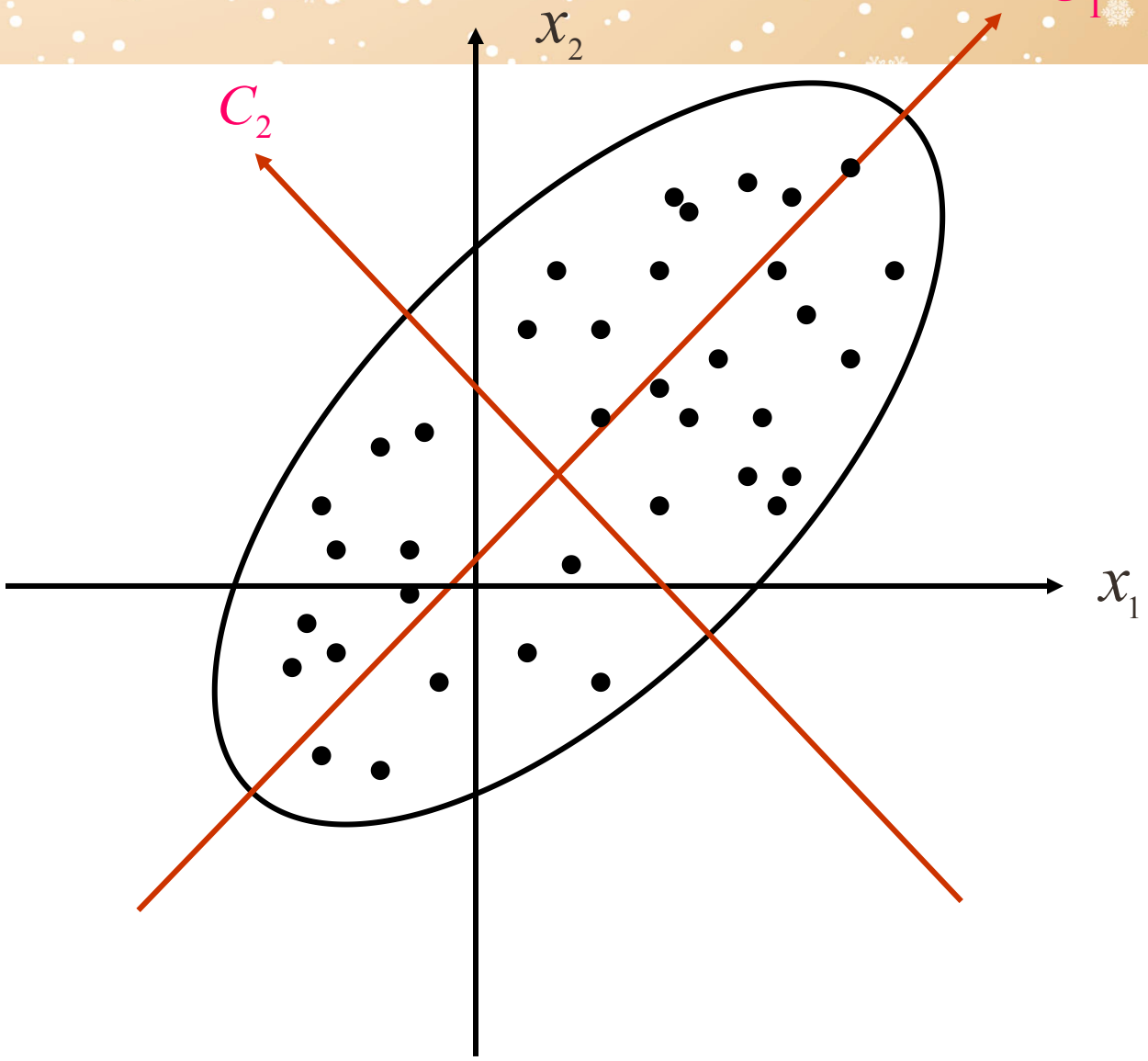
这些综合指标就称为主成分。问题是：选择几个成分合适？

- 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。

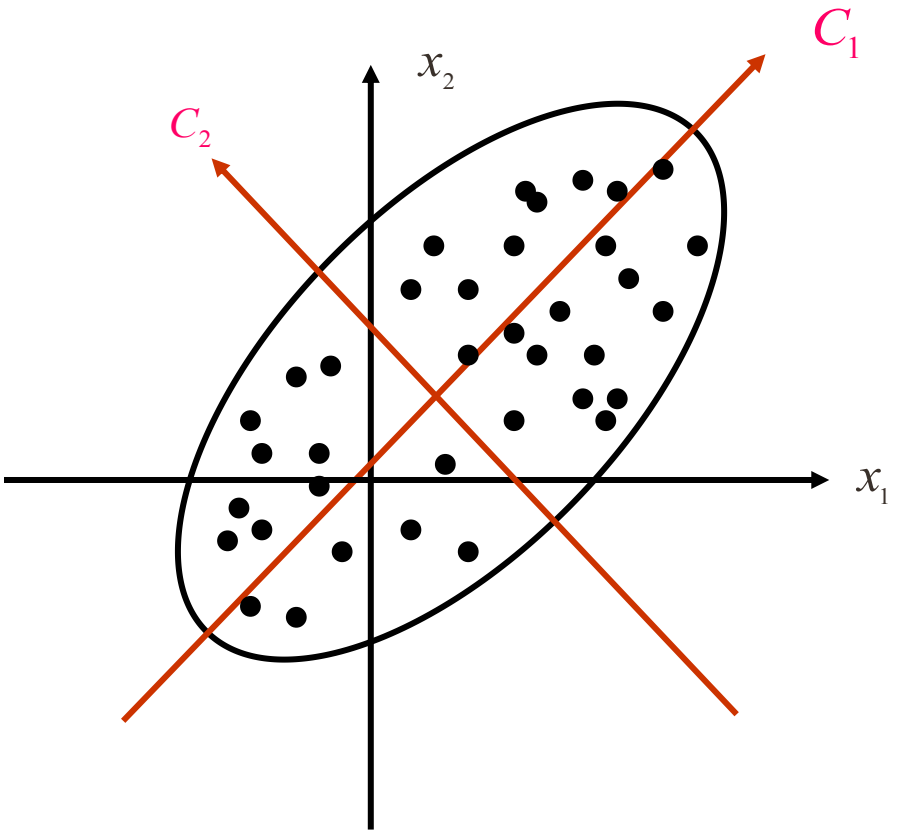
在力求数据信息丢失最少的原则下，对高维的变量空间降维，即研究指标体系的少数几个线性组合，并且这几个线性组合所构成的综合指标将尽可能多地保留原来指标变异方面的信息。这些综合指标就称为主成分。

平移、旋转坐标轴

主成分分析的几何解释



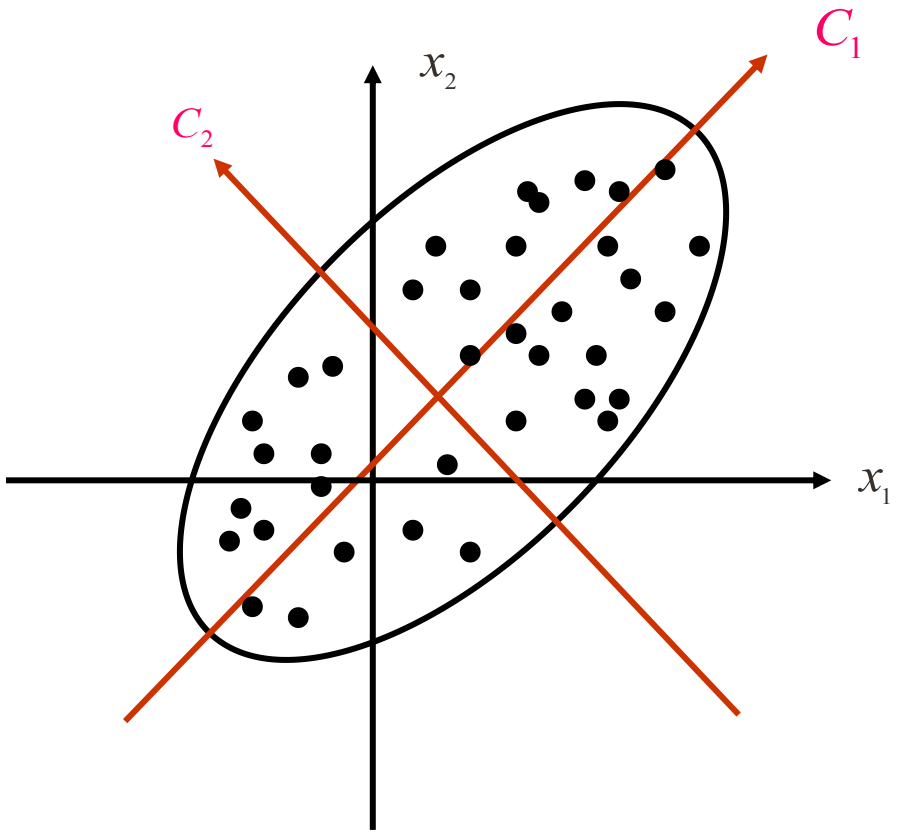
- 如果我们将x_1 轴和x_2轴先平移，再同时按逆时针方向旋转θ角度，得到新坐标轴C_1和C_2。C_1和C_2是两个新变量。



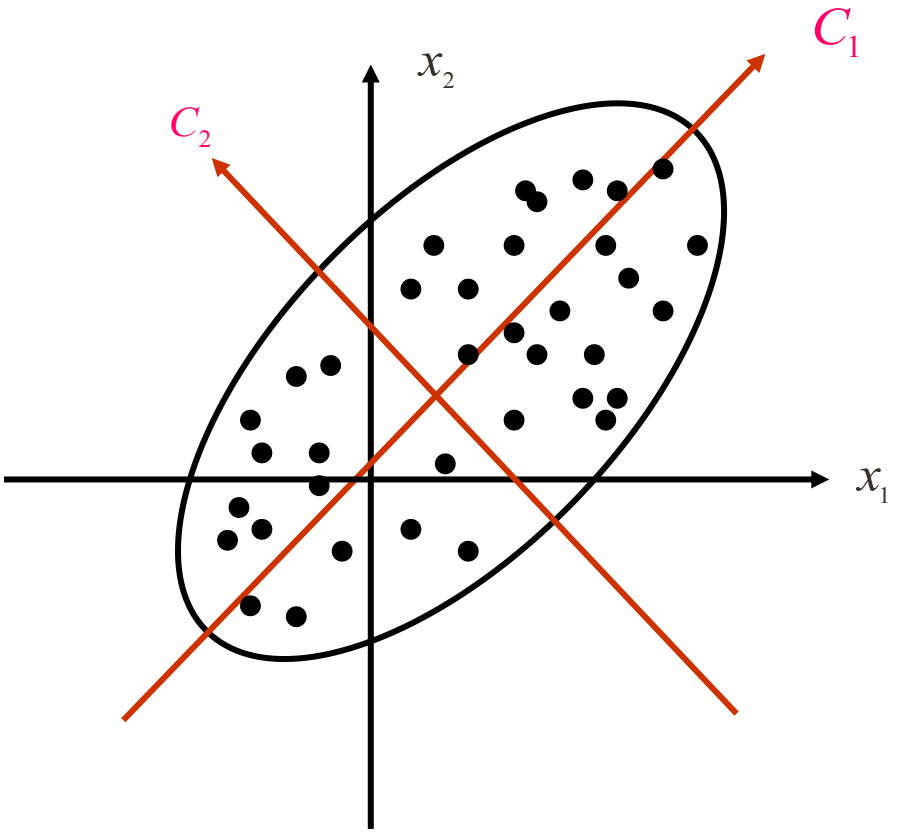
根据旋转变换的公式：

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

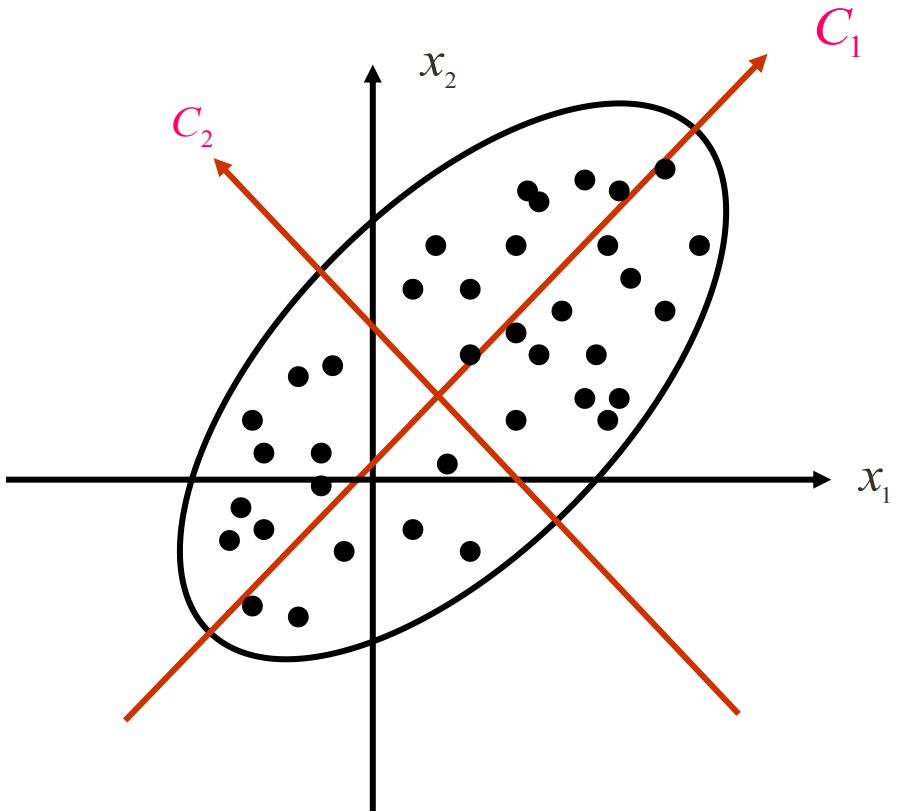
$$\begin{pmatrix} C_1 \\ C_2 \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$



旋转变换的目的是为了使得n个样品点在C_l轴方向上的离 散程度最大，即C_l的方差最大。变量C_l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_l轴上，对数据中包含的信息起到了浓缩作用。

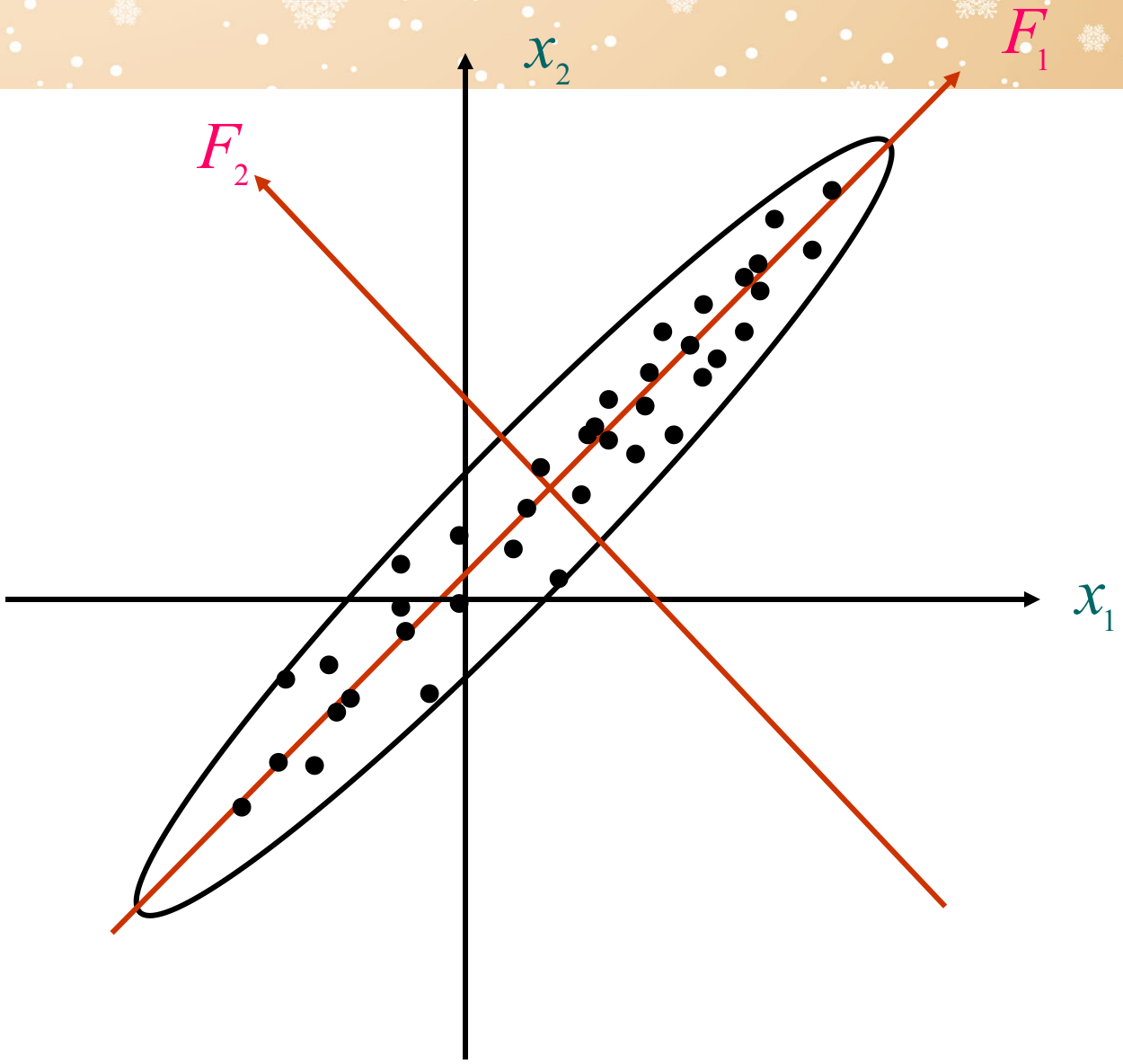


C_1，C_2除了可以对包含在X_1，X_2中的信息起着浓缩作用之外，还具有不相关的性质，这就使得在研究复杂的问题时避免了信息重叠所带来的虚假性。二维平面上的个点的方差大部分都归结在C_1轴上，而C_2轴上的方差很小。C_1和C_2称为原始变量x_1和x_2的综合变量。C简化了系统结构。



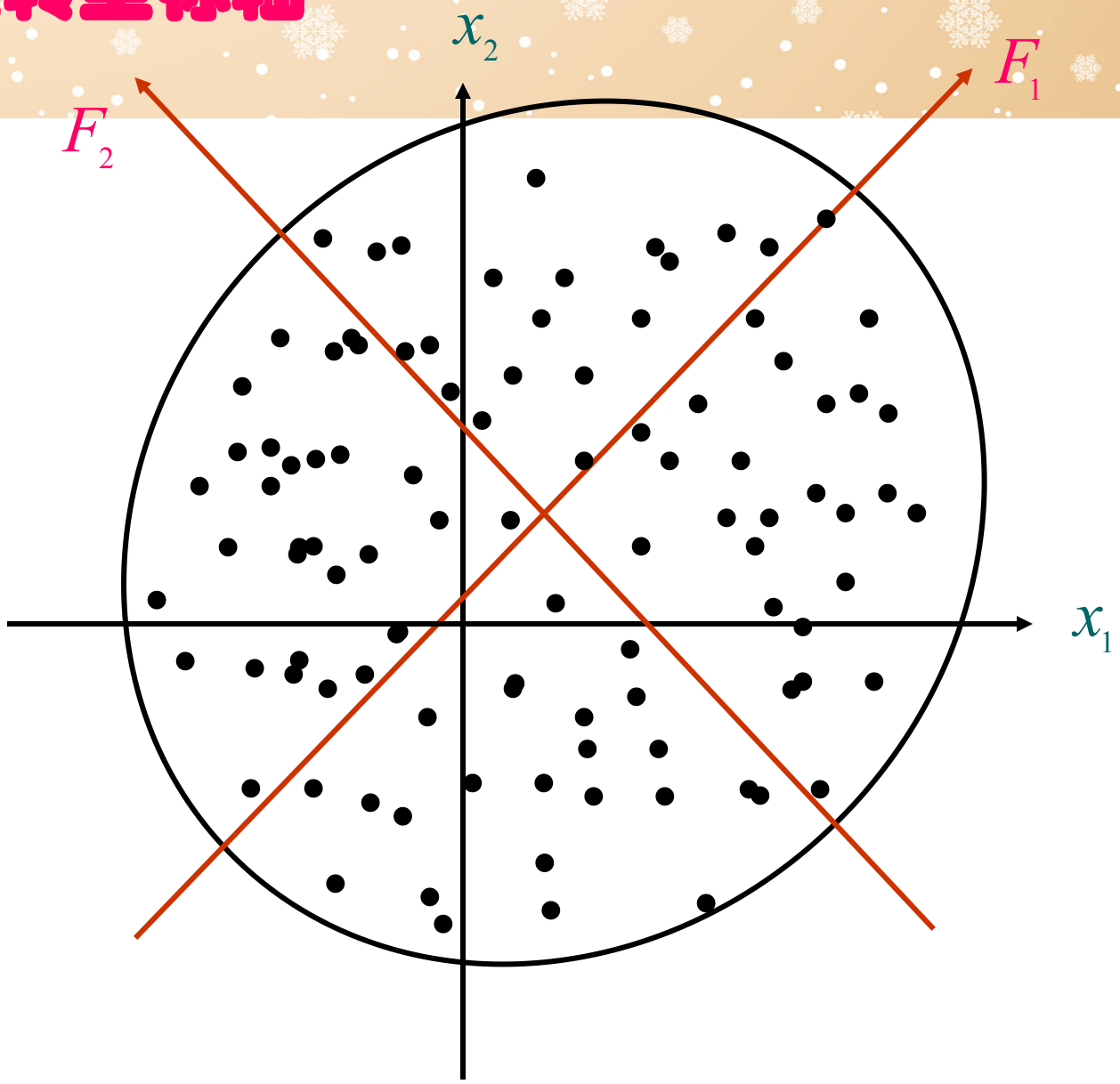
平移、旋转坐标轴

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

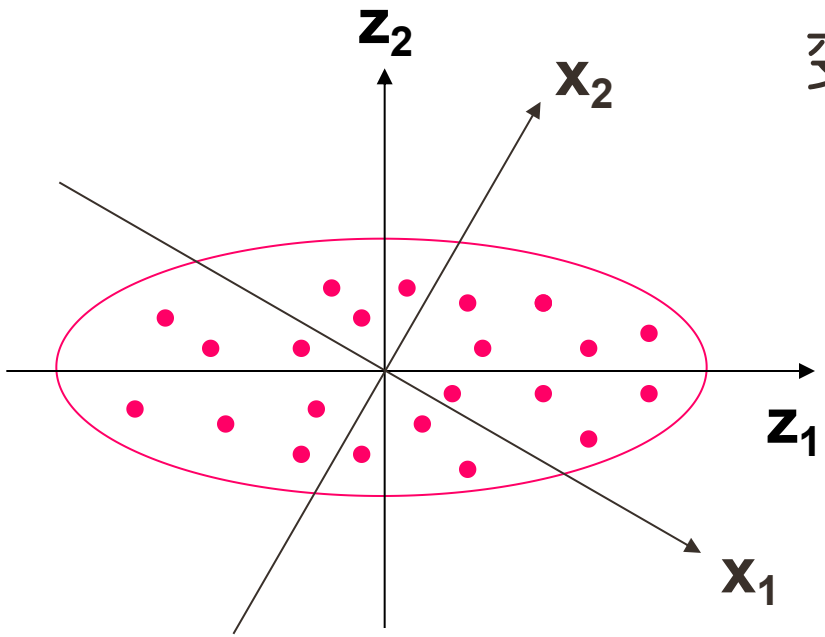
- 主成分分析的数学模型

  原始数据矩阵

$$X = \begin{array}{c} \\ \text{地区} \end{array} \begin{array}{c} \begin{array}{cccc} \text{1} & \text{2} & \cdots & \text{p} \end{array} \text{指标} \\ \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \end{array} \begin{array}{l} \\ \text{1} \\ \text{2} \\ \vdots \\ \text{n} \end{array}$$

§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：

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

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，…，x_p，综合成m个指标z_1，z_2，…，z_m($m \le p$)，可表示为：

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

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

$$Cov（Z_i，Z_j）= 0，\ i \neq j，\ i，j = 1，\ 2，\ \cdots，\ p$$

主成分的方差依次递减，重要性依次递减，即

$$Var（Z_1）\geq Var(Z_2) \geq \cdots \geq Var(Z_p)$$

§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~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 $\mathbf{u}_1,\cdots,\mathbf{u}_p$

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

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=λx（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

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

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

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

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$\lambda_1, \lambda_2, \ldots, \lambda_m$对应的特征向量$x_1, x_2, \ldots, x_m$线性无关。

补：随机变量的数字特征

- 协方差

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

§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

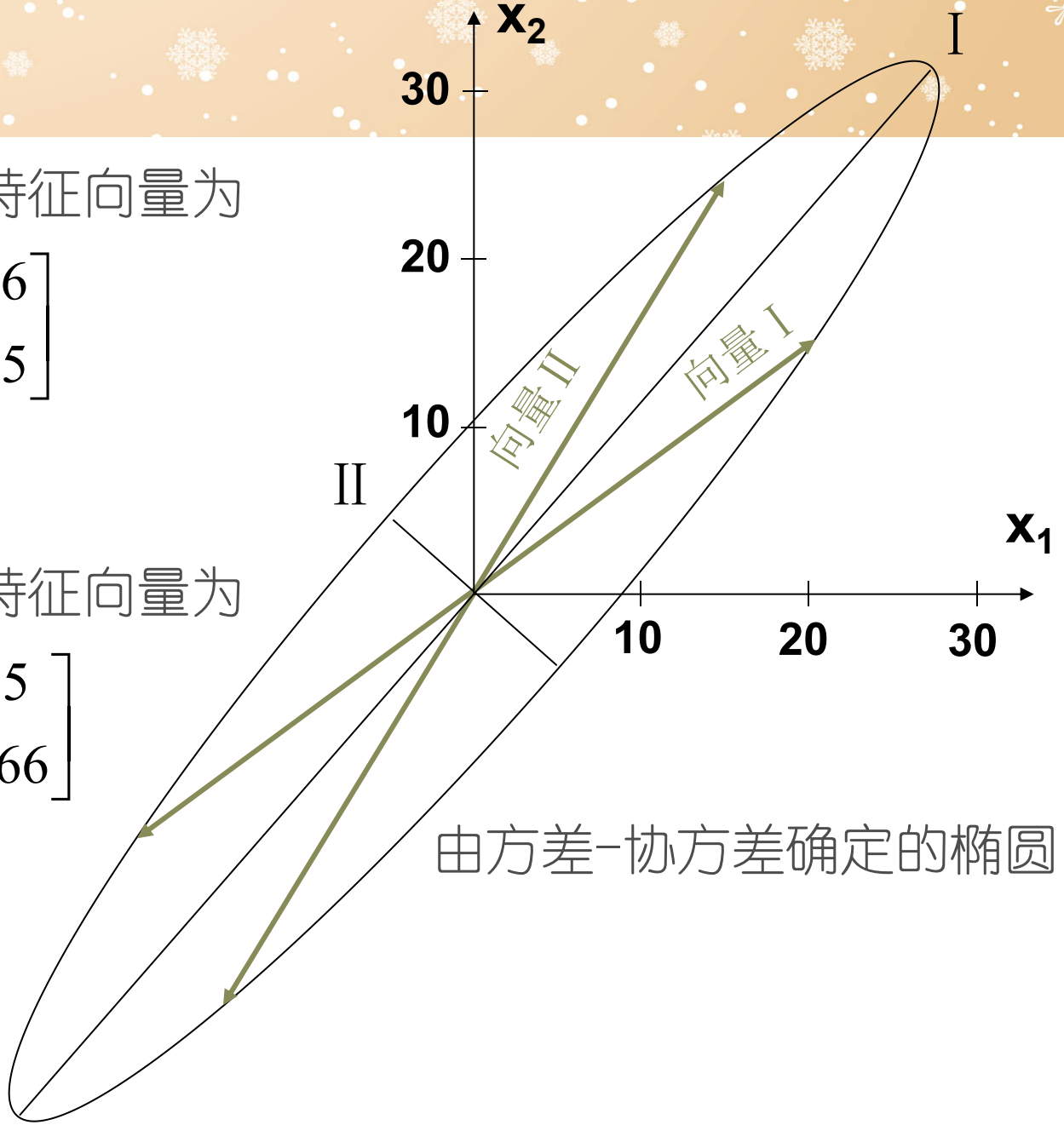
$$\mathrm{I} = \begin{bmatrix} 0.66 \\ 0.75 \end{bmatrix}$$

$\lambda_{\mathrm{I}} = 37.9$

第二主成分：特征向量为

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

$\lambda_{\mathrm{II}} = 6.5$



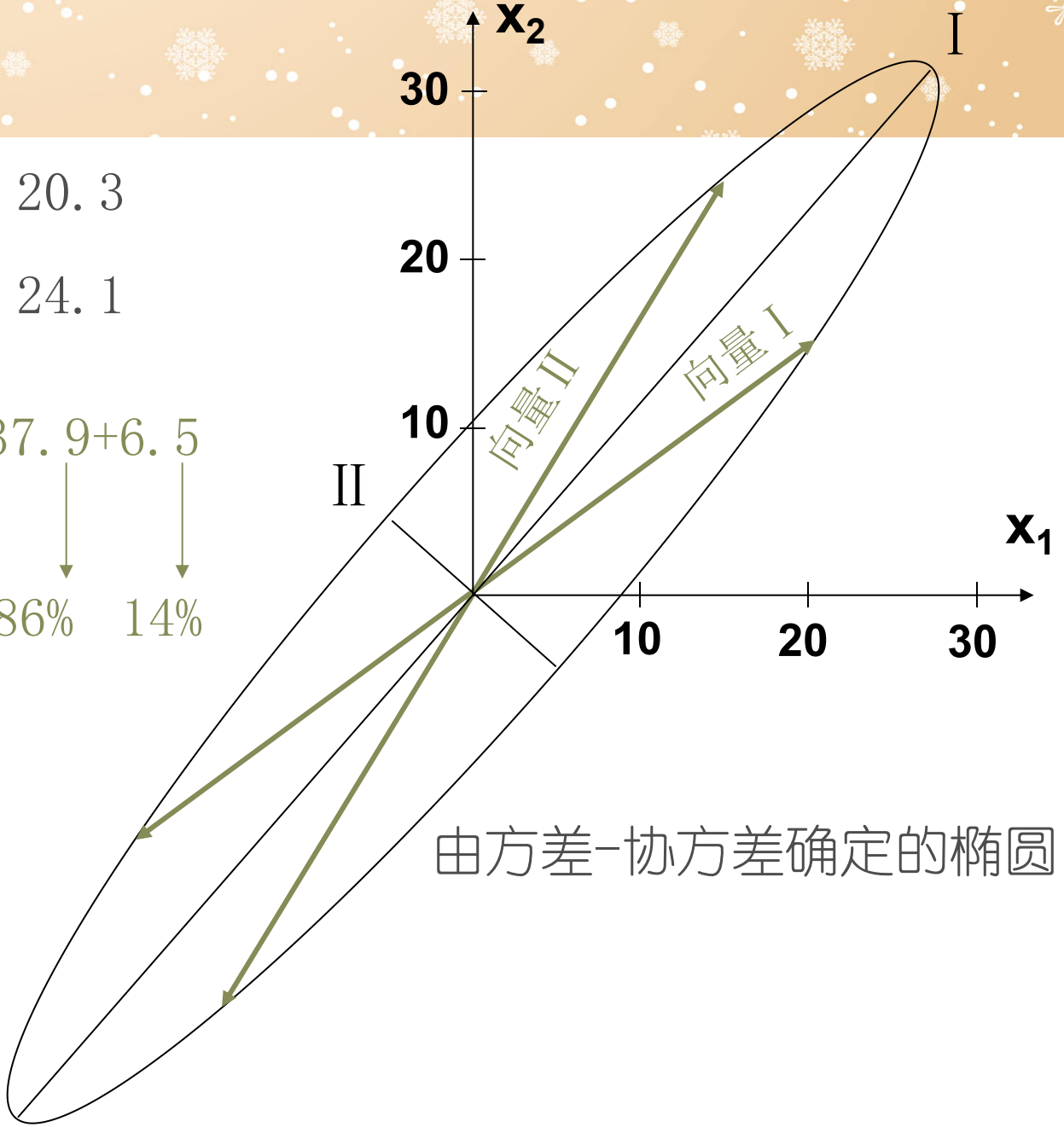
由方差-协方差确定的椭圆

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%



由方差-协方差确定的椭圆

§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，$x_2 \ldots x_p$为p个原指标变量，经过标准化后得到标准化变量X_1，$X_2 \ldots X_p$：

$$X_i = \frac{x_i - \bar{x}_i}{\sigma_i} \qquad \text{i=1，2，…p}$$

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

§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

$$\left|R-\lambda I\right|=0 \qquad \lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p\geq 0$$

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

§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

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

一般取累计贡献率达85%~95%的特征值对应的主成分即可。

$$\lambda_1, \lambda_2, \cdots, \lambda_m (m \le p)$$

我们作如下定义：

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

且使 $Var(C_1)$最大，则称C_1为第一主成分；

但系数w若无限制可使$Var(C_1)$无限大，故加约束条件：

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

组合系数（ $w_{11}, w_{12}, \ldots\ w_{1p}$）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

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

6.计算主成分得分

$$\begin{aligned} z_1 &= l_{11}x_1^* + l_{12}x_2^* + \cdots + l_{1p}x_p^* \\ z_2 &= l_{21}x_1^* + l_{22}x_2^* + \cdots + l_{2p}x_p^* \\ &\cdots \\ z_m &= l_{m1}x_1^* + l_{m2}x_2^* + \cdots + l_{mp}x_p^* \end{aligned}$$

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

主成分的性质：

主成分C_1，C_2，…，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$Corr(C_i, C_j)=0$　　$i \neq j$

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

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

(3) 各主成分的方差是依次递减的， 即

$Var(C_1) \geq Var(C_2) \geq \ldots \geq Var(C_p)$

(4) 总方差不增不减， 即

$$Var(C_1)+Var(C_2)+\cdots+Var(C_p)$$
$$=Var(x_1)+Var(x_2)+\cdots+Var(x_p)$$
$$=p$$

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，…，X_p的相关矩阵为R，(w_{i1}，w_{i2}，…，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*

$Var(C_i)= \lambda_i$

其中λ_i为相关矩阵R的第i个特征值*(eigenvalue)*

$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$

（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

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

- 2.解特征方程|R-λI |=0，求出相关阵R的特征根（eigenvalue）λ_i，且按从大到小顺序排列：
- $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ ，
- 3.求矩阵R关于λ_i的满足正规条件的特征向量（eigenvector):
- $L_i = (l_{i1}, l_{i2}, \ldots, l_{ip})$
- 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=4.6610，=2.0890，
=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。

了解了主成分分析的基本思想、数学和几何意义后，问题的关键：

- 1、如何进行主成分分析？（主成分分析的方法）

基于相关系数矩阵还是基于协方差矩阵做主成分分析。当分析中所选择的经济变量具有不同的量纲，变量水平差异很大，应该选择基于相关系数矩阵的主成分分析。

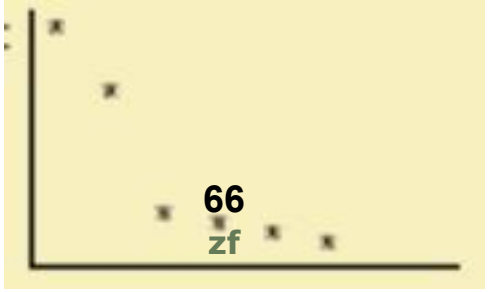
- 2、如何确定主成分个数？

主成分分析的目的是简化变量，一般情况下主成分的个数应该小于原始变量的个数。关于保留几个主成分，应该权衡主成分个数和保留的信息。

- 3、如何解释主成分所包含的地理意义？

主成分个数的确定

- **累积方差贡献率（Cumulative variance explained by components）**：通常要求累积方差贡献率达到85%以上来确定主成分个数。
- **特征根（eigenvalue）**：根据特征根来确定 $\lambda_i > \bar{\lambda}$；数据标准化情况下：$\lambda_i > \bar{\lambda} = \frac{1}{p}\sum_{i=1}^{p}\lambda_i = 1$
- **碎石图（Scree plot）**：依据特征值的变化来确定，即特征值变化趋势图由陡坡变为平坦的转折点即为主成分选择的最佳个数。



Q & A