

华中科技大学

# 模式识别与分类

## 案例分析

---

路志宏

[luzhihong@hust.edu.cn](mailto:luzhihong@hust.edu.cn)

华中科技大学数学与统计学院

**2023.07.04**

# 内 容

---

一、PCA（主成分分析）

二、LDA（线性分类器）

# 第一讲：主成分分析 PCA

## *Principal Component Analysis*

---

# PCA 内 容

---

- 一、问题的提出
- 二、主成分分析原理
- 三、求解步骤（算法）
- 四、实例分析
- 五、小结
- 六、参考文献

# PCA 内 容

---

- 一、问题的提出
- 二、主成分分析原理
- 三、求解步骤（算法）
- 四、实例分析
- 五、小结
- 六、参考文献

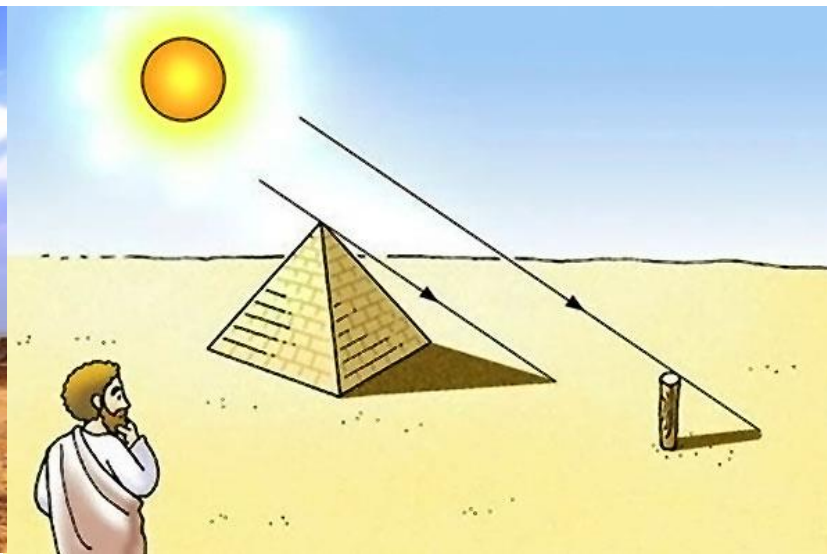
# 一、问题的提出

## ➤ 例1：测量金字塔的高度？



胡夫金字塔(公元前 2613 年)

高 **146.59 米**



泰勒斯（Thales，约前 625 ~ 前 547）：  
古希腊数学家、天文学家

原理：影长等于身長  $\Leftrightarrow$  塔影等于塔高  
相似三角形原理


# 一、问题的提出

## ➤ 例2 汇报工作

假定你是一个公司的财务经理，掌握了公司的所有数据，  
比如

固定资产、流动资金、借贷资金和期限、  
原料消耗、产值、折旧、税费、  
工资支出、利润、  
职工人数、职工分工、教育程度、

.....



原封不动  
地摆出去  
吗？

你必须要把各个方面作出高度概括，  
用少数几个指标简单明了地把情况说清楚。

# 一、问题的提出

---

## ➤ 例3 经济分析

美国经济学家库兹涅茨（1901-1985）和英国经济学家斯通（1913-1991）等人开发的国民收入和国内生产总值的核算方法被称为“20 世纪最伟大的发明之一”，并分别于1971年，1984年获诺贝尔经济学奖。

如《国民经济核算体系》(**System of National Accounts**), 简称为 **SNA**。联合国、世界银行、国际货币基金组织、经济合作和发展组织及欧洲共同体委员会等颁布和使用。



# 一、问题的提出

---

## ➤ 例3 经济分析

统计学家斯通在 1947 年关于国民经济的研究是一项十分著名的工作。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.....

17 个  
变量

# 一、问题的提出

---

## ➤ 例2 经济分析

在进行主成分分析后，竟以 97.4% 的精度，用 3 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 ；

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。

# 一、问题的提出

## ➤ 例4 学生成绩分析

**100** 个学生的数学、物理、化学、语文、历史、英语的成绩如下表（部分）。

学生编号	数学	物理	化学	语文	历史	英语
1	65	61	72	84	81	79
2	77	77	76	64	70	55
3	67	63	49	65	67	57
4	80	69	75	74	74	63
5	74	70	80	84	81	74
6	78	84	75	62	71	64
7	66	71	67	52	65	57
8	77	71	57	72	86	71
9	83	100	79	41	67	50
...	...	...				

需要少量几个综合评价指标

# 一、问题的提出

## ➤ 例4 学生成绩分析

从本例提出如下几个问题：

1. 能否把这个数据的 6 个变量  
用一两个综合变量来表示呢？
2. 这一两个综合变量包含有多少  
原来的信息呢？
3. 能否利用找到的综合变量来对  
学生排序呢？  
以及如何解释综合变量和排序呢？

学生编号	数学	物理	化学	语文	历史	英语
1	65	61	72	84	81	79
2	77	77	76	64	70	55
3	67	63	49	65	67	57
4	80	69	75	74	74	63
5	74	70	80	84	81	74
6	78	84	75	62	71	64
7	66	71	67	52	65	57
8	77	71	57	72	86	71
9	83	100	79	41	67	50
...	...	...	...	...	...	...

非常重要！

# 一、问题的提出

---

## ➤ 其它：本科生期末考试成绩单

兴趣程度	有浓厚的兴趣，会花更多的时间去学习，
学习时间	当然一定会取得优异成绩。
考试成绩	所以第二项与第一项强相关，第三项和第二项也是强相关。 <b>可以合并</b>

## ➤ 其它：去噪

信号在传输过程中，由于信道不是理想的，信道另一端收到的信号会有噪音扰动，那么如何滤去噪音？

# 内容

---

一、问题的提出

二、主成分分析原理

三、求解步骤（算法）

四、实例分析

五、小结

六、参考文献

## 二、主成分分析原理

---

### ➤ 主成分分析

- 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。
- 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，用较少的新变量代替原来较多的变量，而且使这些较少的新变量尽可能多地保留原来较多的变量所反映的信息？事实上，这种想法是可以实现的。

## 二、主成分分析原理

---

### ➤ 主成分分析的基本思想

主成分分析是研究如何以较少的信息丢失将众多原有变量浓缩成少数几个因子（主成分），使这些主成分在一定程度上复现原有变量所携带的信息的多元统计分析方法。

简而言之，**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。

**主成分分析就是综合处理这类降维问题的一种强有力方法。**



## 二、主成分分析原理

### ➤ PCA——数据分析的基础知识

设有  $n$  个样本，每个样本观测  $p$  个指标（变量）：

$X_1, X_2, \dots, X_p$ ，得到原始数据矩阵：

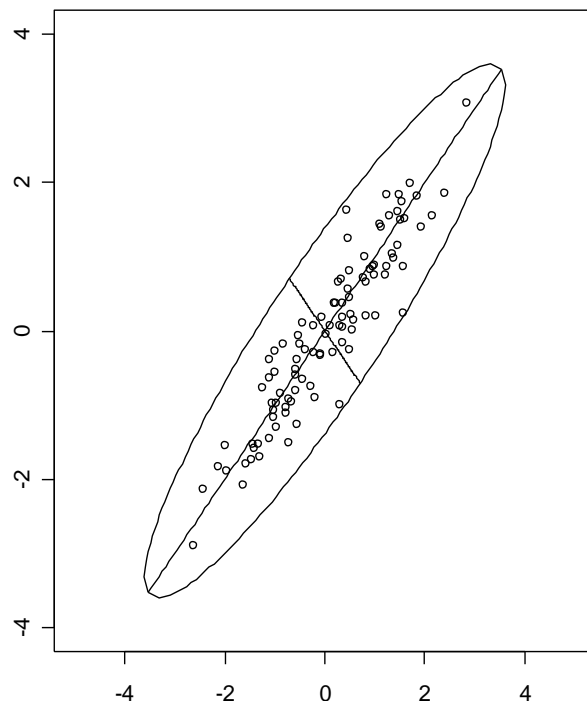
$$X_{n \times p} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}_{n \times p}$$

$\uparrow \quad \quad \uparrow \quad \quad \uparrow$

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

✓ 熟悉：均值、方差、标准差、协方差

✓ 熟悉：特征值、特征向量



## 二、主成分分析原理

✓ 基础知识：均值、方差、标准差、协方差

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

样本均值是数据散列图的中心.集中程度

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2}$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

样本方差是描述一组数据变异程度或分散程度大小的指标.

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$$

$$\text{Cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

协方差是用来度量两个随机变量关系的统计量.

## 二、主成分分析原理

---

✓ 基础知识：特征值、特征向量

**定义**  $A$  为  $n$  阶方阵， $\lambda$  为数， $X$  为  $n$  维非零向量，若

$$AX = \lambda X$$

则  $\lambda$  称为  $A$  的特征值， $X$  称为对应  $\lambda$  的  $A$  的特征向量。

称以  $\lambda$  为未知量的一元  $n$  次方程， $|\lambda I - A| = 0$  为  $A$  的**特征方程**。

## 二、主成分分析原理

例1：从一个总体中随机抽取 4 个样本作三次测量，每一个样本的观测向量为：

$$X_1 = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}, X_2 = \begin{pmatrix} 4 \\ 2 \\ 13 \end{pmatrix}, X_3 = \begin{pmatrix} 7 \\ 8 \\ 1 \end{pmatrix}, X_4 = \begin{pmatrix} 8 \\ 4 \\ 5 \end{pmatrix}$$

计算样本均值  $M$  和协方差矩阵  $S$  以及  $S$  的特征值和特征向量。

```
X = [  
    1    2    1  
    4    2   13  
    7    8    1  
    8    4    5  
];  
[n,p] = size(X);  
M = mean(X);  
Xp = X - ones(n,1) * mean(X);  
S = cov(Xp); % Y = Xp' * Xp / (n-1)  
[V,D] = eig(S);
```

$M = ( \begin{matrix} 5 & 4 & 5 \end{matrix} )$   $D =$

	1.6057	0	0
	0	13.8430	0
	0	0	34.5513

$Xp =$

-4	-2	-4
-1	-2	8
2	4	-4
3	0	0

$V =$

0.5686	0.8193	-0.0740
-0.7955	0.5247	-0.3030
-0.2094	0.2312	0.9501

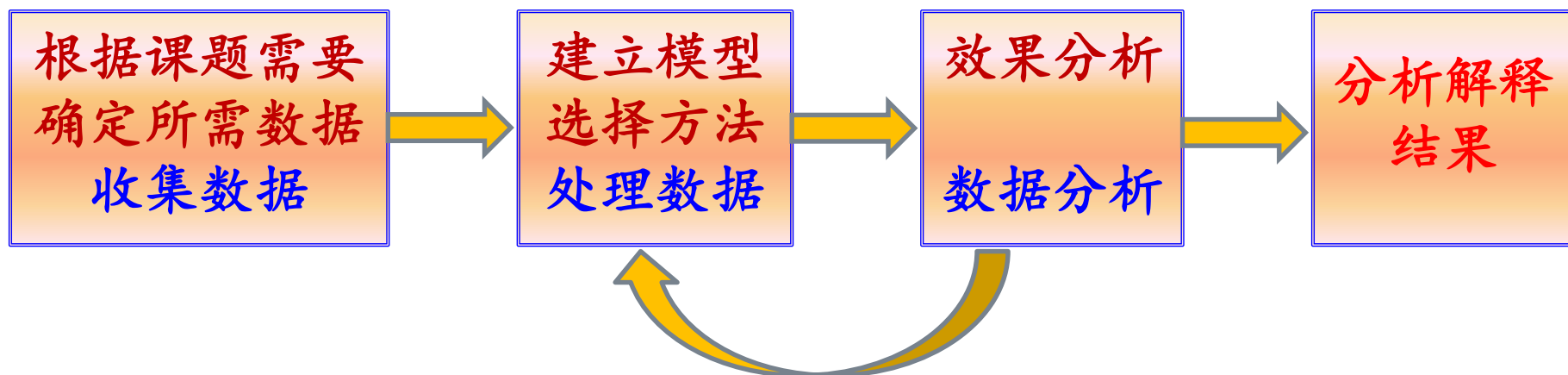
$S =$

10	6	0
6	8	-8
0	-8	32

## 二、主成分分析原理

---

### ➤ PCA---数据分析



## 二、主成分分析原理

### ➤ PCA——数据分析

学生编号	数学	物理	化学	语文	历史	英语
1	65	61	72	84	81	79
2	77	77	76	64	70	55
3	67	63	49	65	67	57
4	80	69	75	74	74	63
5	74	70	80	84	81	74
6	78	84	75	62	71	64
7	66	71	67	52	65	57
8	77	71	57	72	86	71
9	83	100	79	41	67	50
...	...	...	...	...	...	...

例中数据是六维的；也就是说，每个观测值是 6 维空间中的一个点。我们希望把 6 维空间用低维空间表示。

							平均成绩	标准差
学生编号	数学	物理	化学	语文	历史	英语		
1	65	61	72	84	81	79	73.7	
2	77	77	76	64	70	55	69.8	
3	67	63	49	65	67	57	61.3	
4	80	69	75	74	74	63	72.5	
5	74	70	80	84	81	74	77.2	
6	78	84	75	62	71	64	72.3	
7	66	71	67	52	65	57	63.0	
8	77	71	57	72	86	71	72.3	
9	83	100	79	41	67	50	70.0	
...	...	...	...	...	...	...		
单科平均成绩	74.1	74.0	70.0	66.4	73.6	63.3		
标准差	6.5659	11.9059	10.5475	14.0900	7.4517	9.6566		

## 二、主成分分析原理

---

### ➤ PCA——二维数据分析

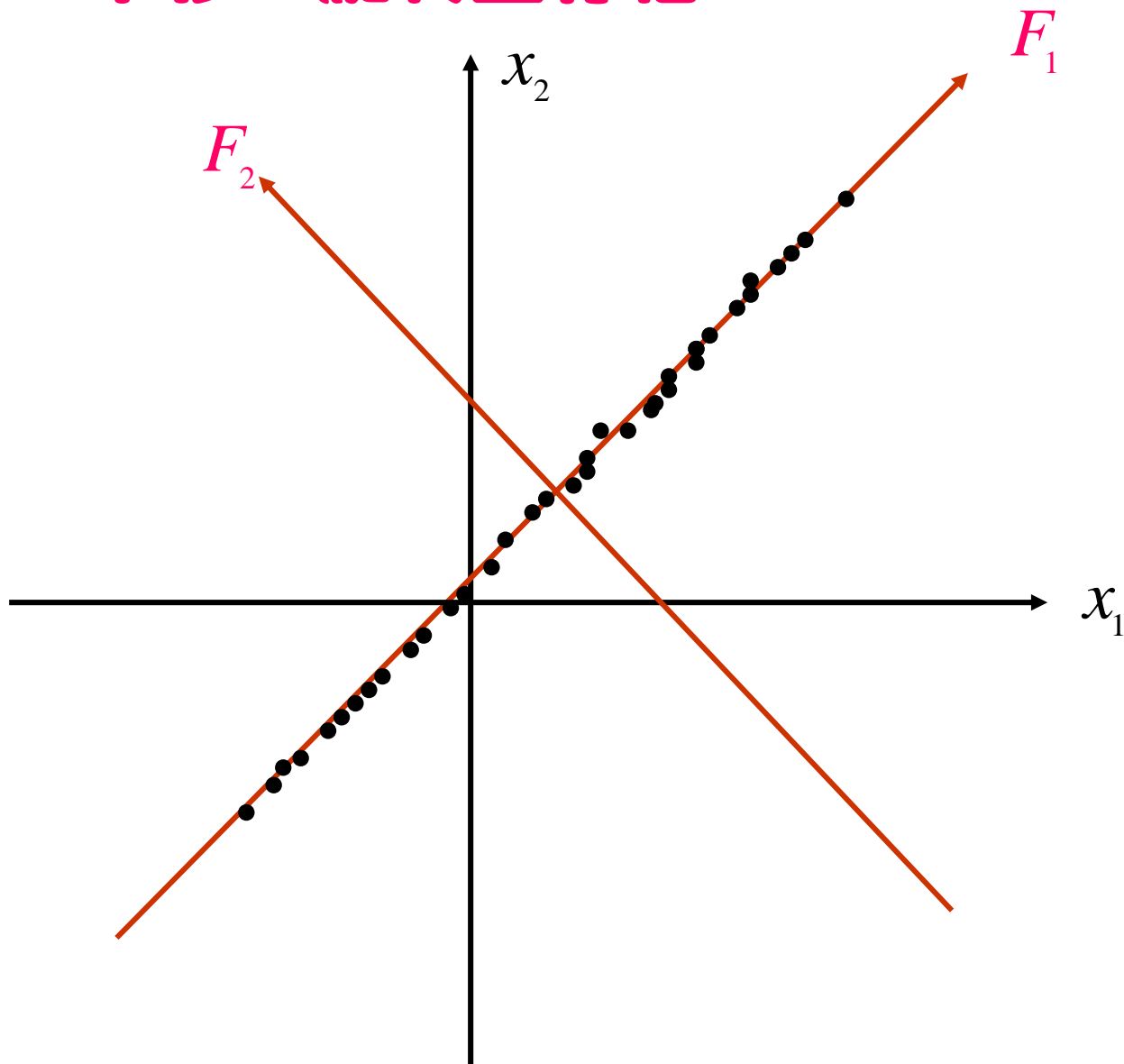
先假定数据只有二维，即只有两个变量，它们分别由横坐标和纵坐标所代表；因此每个观测值都有相应于这两个坐标轴的两个坐标值；

如果这些数据形成一个椭圆形状的点阵（这在变量的二维正态的假定下是可能的）。



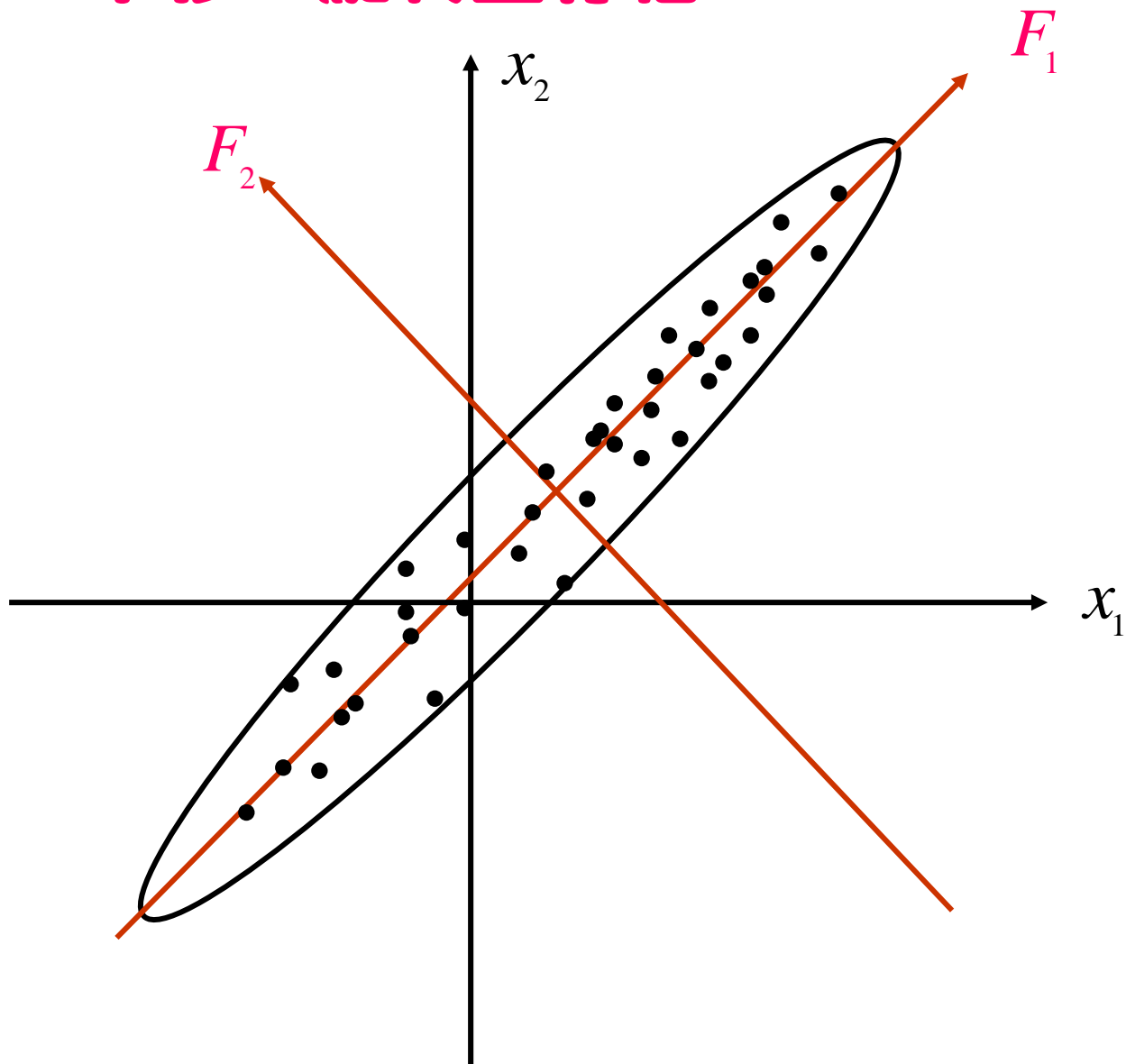
# 平移、旋转坐标轴

## 主成分分析的几何解释



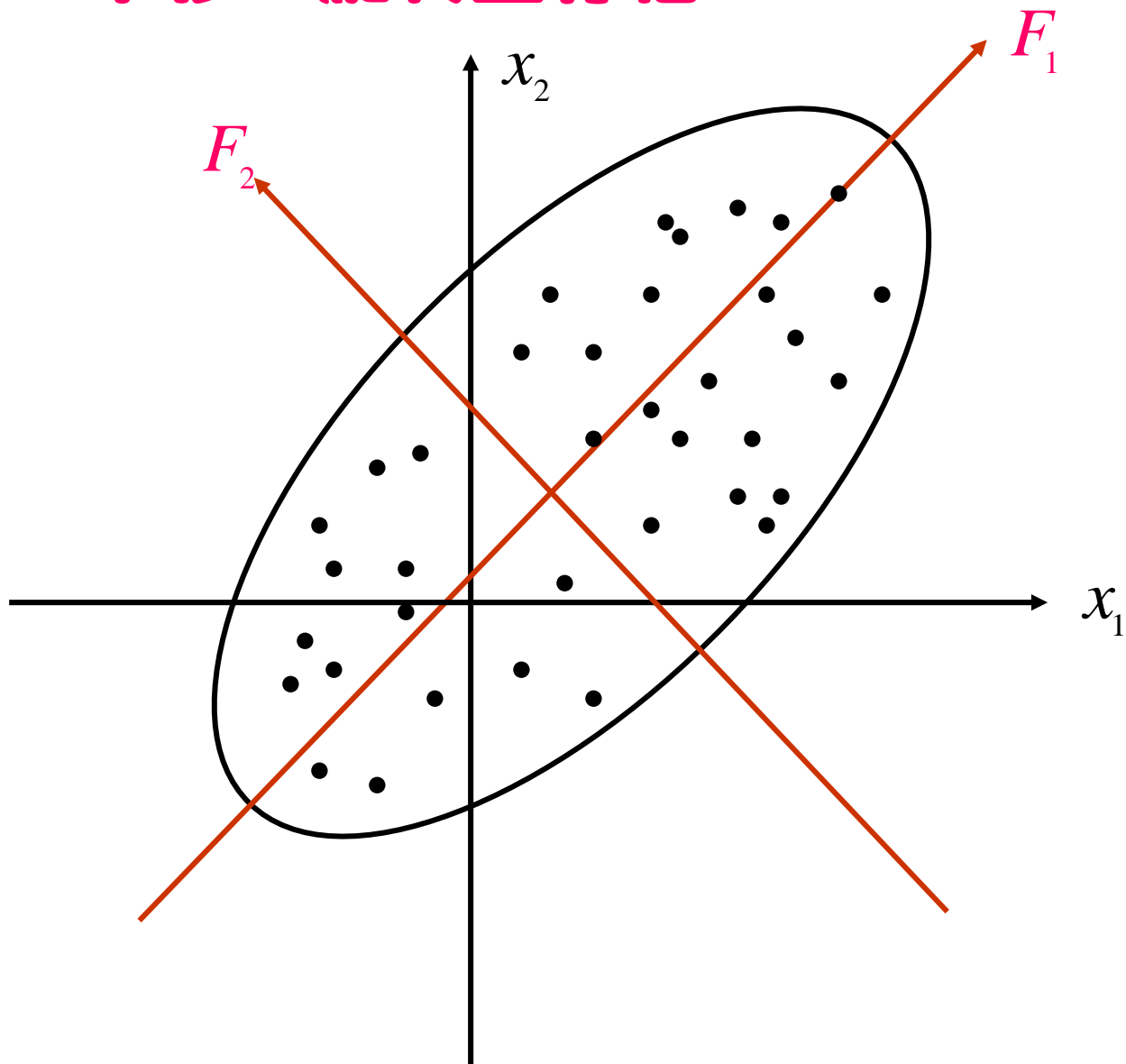
# 平移、旋转坐标轴

## 主成分分析的几何解释



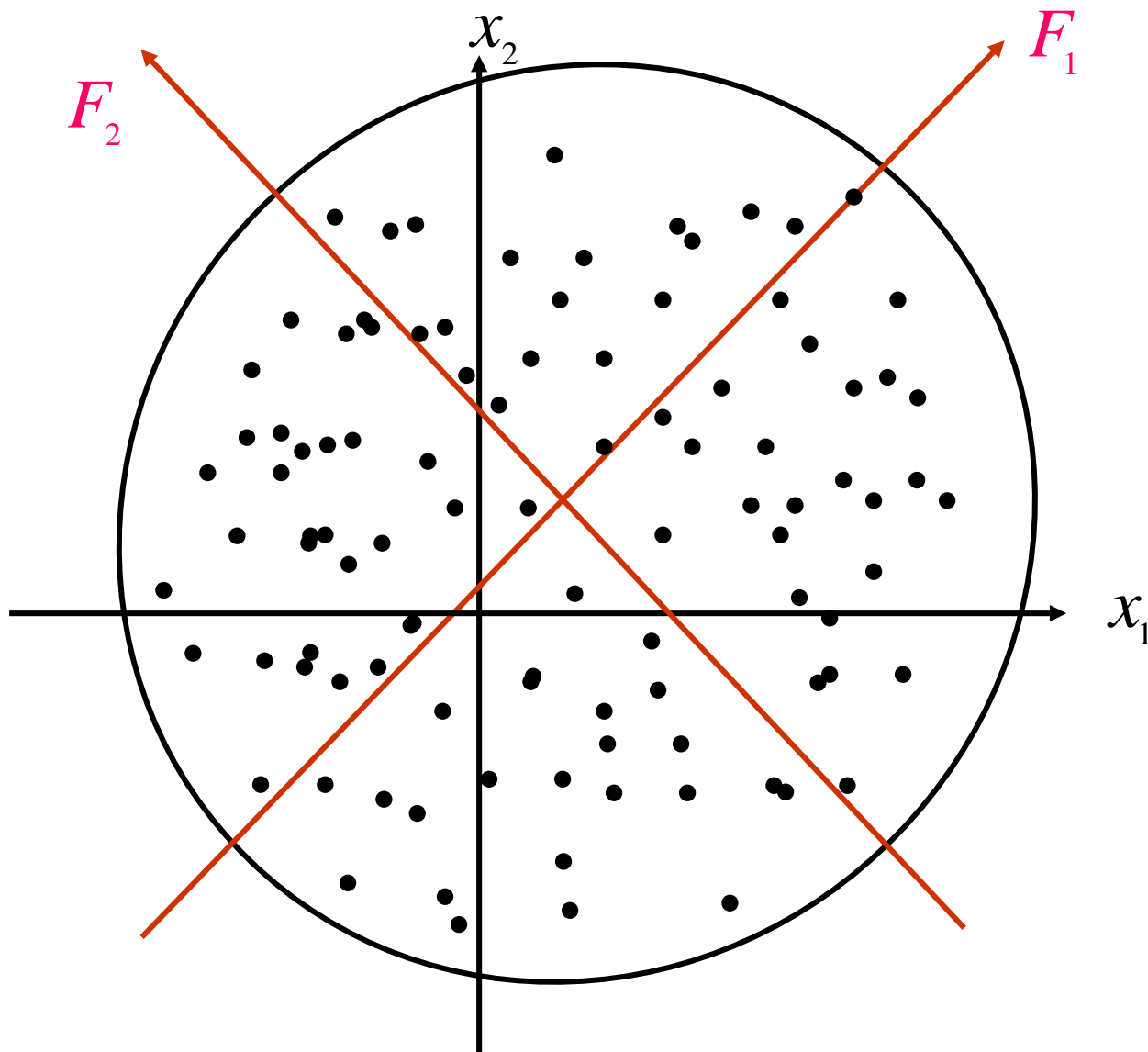
# 平移、旋转坐标轴

## 主成分分析的几何解释



# 平移、旋转坐标轴

## 主成分分析的几何解释



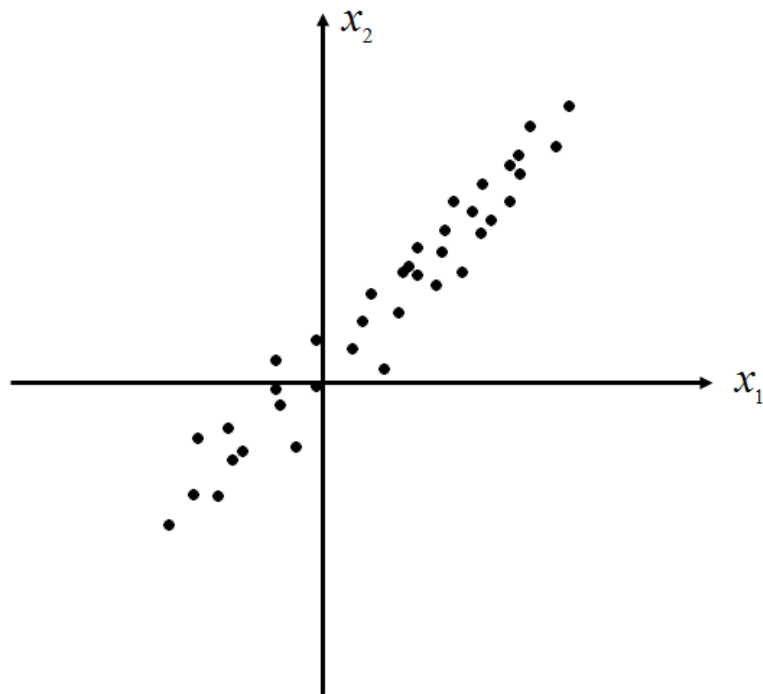
主成分分析在这种情况下是失效的！

## 二、主成分分析原理

### ➤ PCA——二维数据分析

为了方便，我们在二维空间中讨论主成分。

设有  $n$  个样本，每个样本有两个观测变量  $x_1$  和  $x_2$ ，在由变量  $x_1$  和  $x_2$  所确定的二维平面中， $n$  个样本点所散布的情况如椭圆状。

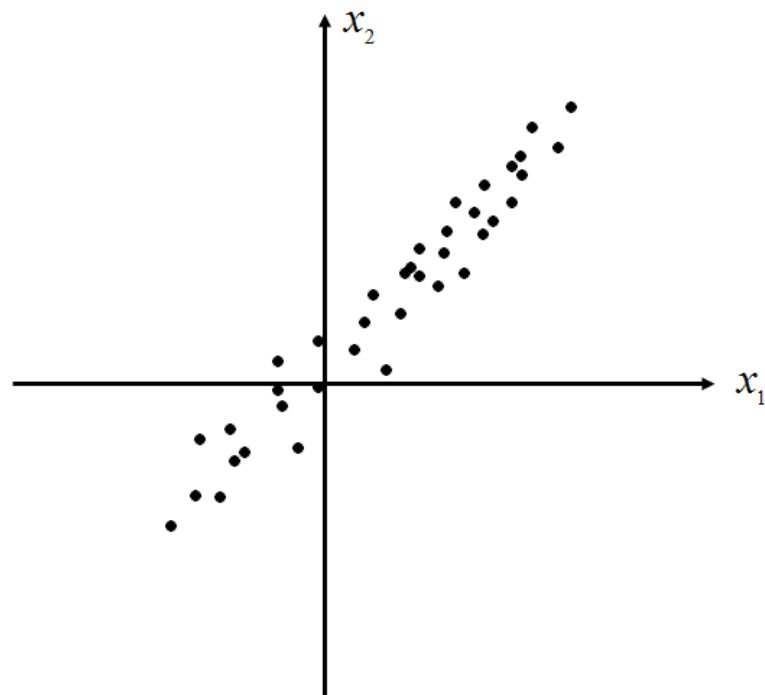


## 二、主成分分析原理

### ➤ PCA——二维数据分析

由图可以看出，这  $n$  个样本点无论是沿着  $x_1$  轴方向或  $x_2$  轴方向都具有较大的离散性，其离散的程度可以分别用观测变量  $x_1$  的方差和  $x_2$  的方差定量地表示。

显然，如果只考虑  $x_1$  和  $x_2$  中的任何一个，那么包含在原始数据中的信息将会有较大的损失。

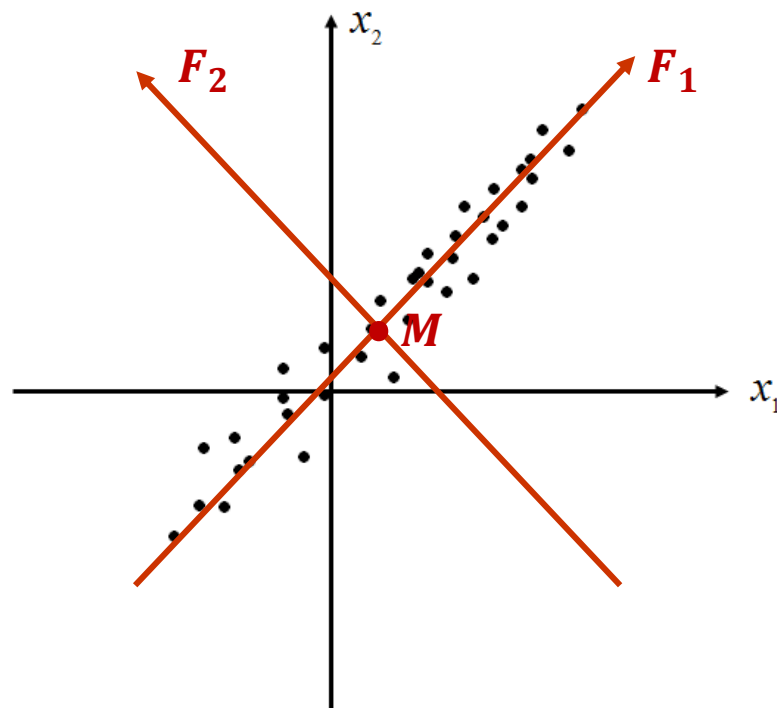


## 二、主成分分析原理

### ➤ PCA——二维数据分析

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角度，得到新坐标轴  $F_1$  和  $F_2$ 。

$F_1$  和  $F_2$  是两个新变量。



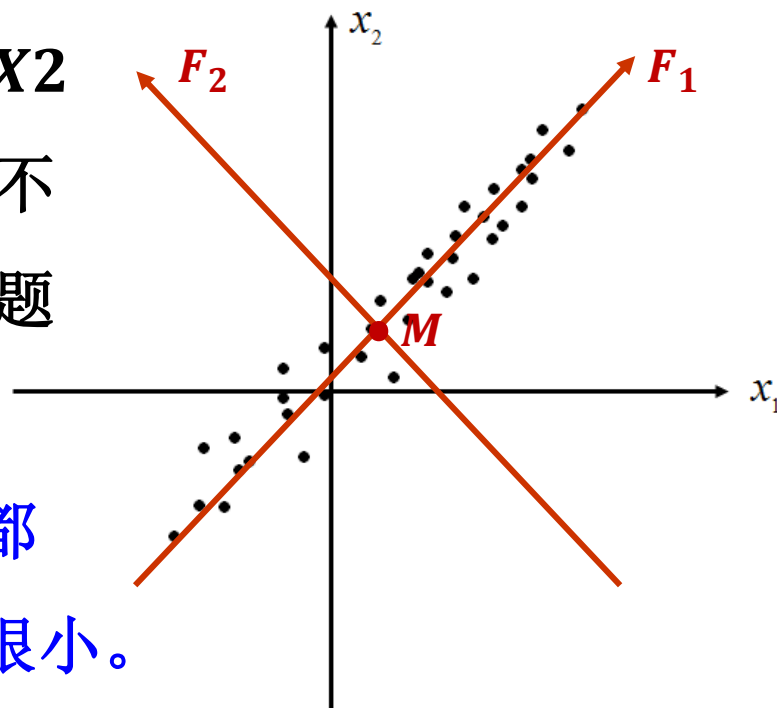
## 二、主成分分析原理

### ➤ PCA——二维数据分析

$F_1$ ,  $F_2$  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  $F_1$  轴上, 而  $F_2$  轴上的方差很小。

$F_1$  和  $F_2$  称为原始变量  $x_1$  和  $x_2$  的综合变量, 也即主成分。



$F$  简化了系统结构, 抓住了主要矛盾。



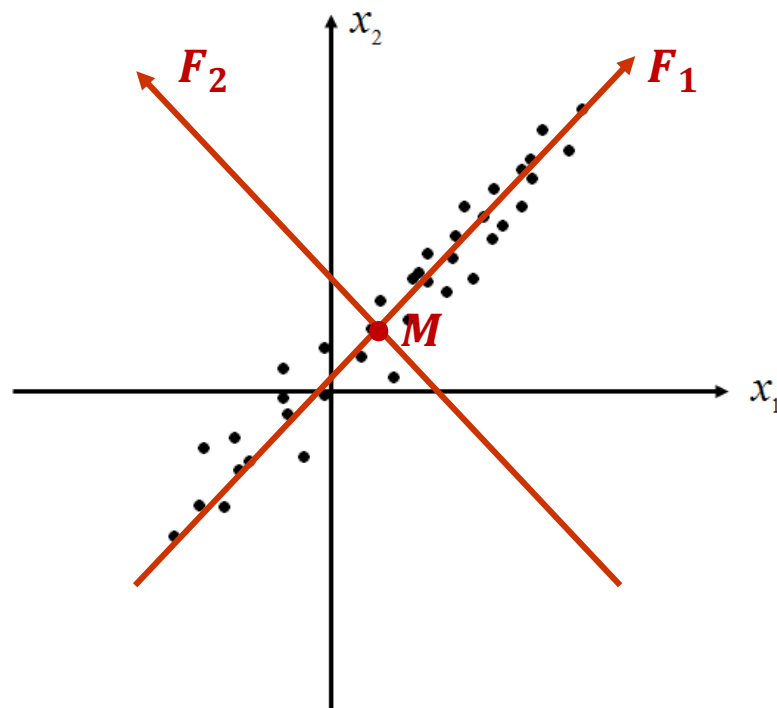
## 二、主成分分析原理

$F$  简化了系统结构，  
抓住了主要矛盾。

### ➤ PCA——二维数据分析

代表长轴的变量  $F_1$  就描述了数据的主要变化，包含数据的大部分信息；而代表短轴的变量  $F_2$  就描述了数据的次要变化（舍去），降维就完成了。

椭圆（球）的长短轴相差得越大，  
降维也越有道理。



## 二、主成分分析原理

---

### ➤ PCA——高维数据分析（推理）

对于多维变量的情况和二维类似，也有高维的椭球，只不过无法直观地看见罢了。

首先把高维椭球的主轴找出来，再用代表大多数数据信息的最长的几个轴作为新变量；这样，主成分分析就基本完成了。

注意，和二维情况类似，高维椭球的主轴也是互相垂直的。这些互相正交的新变量是原先变量的线性组合，叫做**主成分**（*principal component*）。

## 二、主成分分析原理

---

### ➤ PCA——高维数据分析（推理）

正如二维椭圆有两个主轴，三维椭球有三个主轴一样，有几个变量，就有几个主成分。

选择越少的主成分，降维就越好。

什么是标准呢？那就是这些被选的主成分所代表的主轴的长度之和占了主轴长度总和的大部分。有些文献建议，所选的主轴总长度占有所有主轴长度之和的大约 85% 即可，当然，具体选几个要看实际情况而定。

➤ 例 设  $x_1, x_2, x_3$  的协方差矩阵为  $\Sigma = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}$

解得特征根为  $\lambda_1 = 5.83, \lambda_2 = 2.00, \lambda_3 = 0.17$ .

$$u_1 = \begin{bmatrix} 0.383 \\ -0.924 \\ 0.000 \end{bmatrix}, u_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, u_3 = \begin{bmatrix} 0.924 \\ 0.383 \\ 0.000 \end{bmatrix}$$

第一个主成分的贡献率为

$$\frac{5.83}{5.83 + 2.00 + 0.17} = 72.875\%$$

尽管第一个主成分的贡献率并不小，但应该取两个主成分。

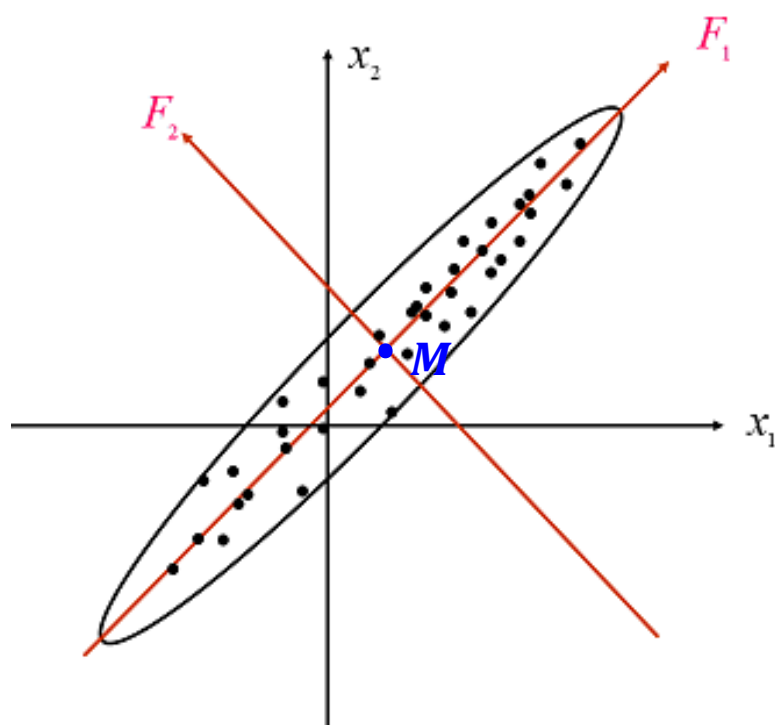
$$\frac{5.83 + 2.00}{5.83 + 2.00 + 0.17} = 97.88\%$$





## ➤ 主成分与原始变量之间的关系：（基本框架—理论分析）

1. 每个主成分都是原始变量的线性组合；
2. 各个主成分之间互不相关，是相互独立的；
3. 主成分保留了原始变量的绝大部分信息；
4. 选择的主成分个数远远少于原有变量的个数；



$$\begin{cases} F_1 = a_1^t X = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\ F_2 = a_2^t X = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\ \dots\dots\dots \\ F_p = a_p^t X = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p \end{cases}$$

$$a_{j1}^2 + a_{j2}^2 + \cdots + a_{jp}^2 = 1, j = 1, 2, \cdots, p.$$

$$\text{Cov}(F_i, F_j) = 0, i \neq j, i, j = 1, 2, \cdots, p$$

$$\text{Var}(F_1) \geq \text{Var}(F_2) \geq \cdots \geq \text{Var}(F_p)$$

# 内容

---

- 一、问题的提出
- 二、主成分分析原理
- 三、求解步骤（算法）
- 四、实例分析
- 五、小结
- 六、参考文献

### 三、PCA 求解步骤---算法

- 1) 对原始数据矩阵作中心化处理 相当于对原始变量进行坐标平移

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

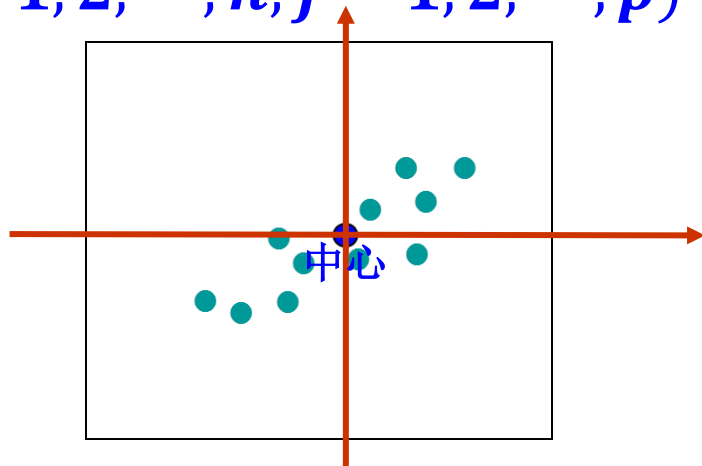
$\uparrow \quad \quad \uparrow \quad \quad \uparrow$   
 $\bar{x}_1 \quad \bar{x}_2 \quad \bar{x}_p$

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}, j = 1, 2, \dots, p$$

$$\tilde{x}_{ij} = x_{ij} - \bar{x}_j,$$

$$(i = 1, 2, \dots, n; j = 1, 2, \dots, p)$$

$\Rightarrow \tilde{X} = \begin{bmatrix} \tilde{x}_{11} & \tilde{x}_{12} & \cdots & \tilde{x}_{1p} \\ \tilde{x}_{21} & \tilde{x}_{22} & \cdots & \tilde{x}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{x}_{n1} & \tilde{x}_{n2} & \cdots & \tilde{x}_{np} \end{bmatrix}$





### 三、PCA 求解步骤---算法

#### ■ 2) 求样本协方差矩阵 $\Sigma_X$

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

于是  $\tilde{X}$  矩阵的列具有零样本均值,即零平均偏差形式。

$$\tilde{X} = (\tilde{X}_1, \tilde{X}_2, \cdots, \tilde{X}_p)$$

$$\Sigma_X = \frac{1}{n-1} \tilde{X}^T \times \tilde{X}$$

**注意:** 协方差矩阵  $\Sigma_X$   
实对称矩阵且半正定

### 三、PCA 求解步骤---算法

---

#### ■ 3) 计算样本协方差矩阵 $\Sigma_X$ 的特征值和特征向量

解特征方程  $|\lambda I - \Sigma_X| = 0$ ,

常用雅可比方法( *Jacobi* )求出特征值, 并使其按大小顺序排列

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0,$$

分别求出对应于特征值  $\lambda_i$  的特征向量

$$u_i (i = 1, 2, \cdots, p).$$

### 三、PCA 求解步骤---算法

#### ■ 4) 写出主成分, 计算主成分贡献率及累计贡献率

• 主成分:  $F_i = u_i^T \tilde{X}, i = 1, 2, \dots, p.$

• 贡献率:

$$\frac{\lambda_i}{\sum_{k=1}^p \lambda_k}, (i = 1, 2, \dots, p)$$

进行主成分分析的主要目的是希望用尽可能少的主成分

$\lambda_1, \lambda_2, \dots, \lambda_m (m < p)$   
代替原来的  $P$  个指标。

• 累计贡献率:

$$\frac{\sum_{j=1}^i \lambda_j}{\sum_{k=1}^p \lambda_k}, (i = 1, 2, \dots, p)$$

一般取累计贡献率达  
85%~95%的特征值

$\lambda_1, \lambda_2, \dots, \lambda_m$   
所对应的第1, 第2,  $\dots$ , 第  
 $m (m < p)$  个主成分。

### 三、PCA 求解步骤---算法

#### ■ 5) 对选取的主成分进行解释和分析

构造综合评价指标。如：

- 选取第 1 个主成分  $F_1$ ，作为综合评价指标。
- 计算选出的  $m$  个主成分的得分，并按得分值的大小进行排队。
- 利用主成分  $F_1, F_2, \dots, F_p$  作线性组合，并以每个主成分  $F_i$  的权重(贡献率)作为权系数构造如下的综合评价函数：

$$w = \alpha_1 F_1 + \alpha_2 F_2 + \dots + \alpha_p F_p$$

$w$  为评估指标，可根据每个样品计算出  $w$  值大小进行排序或分类划级。

# 内容

---

- 一、问题的提出
- 二、主成分分析原理
- 三、求解步骤（算法）
- 四、实例分析
- 五、小结
- 六、参考文献

## 四、PCA 案例分析

---

应收账款是指企业因对外销售产品、材料、提供劳务及其它原因，应向购货单位或接受劳务的单位收取的款项，包括应收销货款、其它应收款和应收票据等。出于扩大销售的竞争需要，企业不得不以赊销或其它优惠的方式招揽顾客，由于销售和收款的时间差，于是产生了应收款项。应收款赊销的效果好坏，不仅依赖于企业的信用，还依赖于顾客的信用。由此，评价顾客的信用等级，了解顾客的综合信用程度，做到“知己知彼，百战不殆”，对加强企业的应收账款管理大有帮助。

某企业为了了解其客户的信用程度，采用西方银行信用评估常用的 5C 方法，5C 的目的是说明顾客违约的可能性。

## 四、PCA 案例分析

---

1. 品格（用 X1 表示），指顾客的信誉，履行偿还义务的可能性。企业可以通过过去的付款记录得到此项。
2. 能力（用 X2 表示），指顾客的偿还能力。即其流动资产的数量和质量以及流动负载的比率。顾客的流动资产越多，其转化为现金支付款项的能力越强。同时，还应注意顾客流动资产的质量，看其是否会出现存货过多过时质量下降，影响其变现能力和支付能力。
3. 资本（用 X3 表示），指顾客的财务势力和财务状况，表明顾客可能偿还债务的背景。
4. 附带的担保品（用 X4 表示），指借款人以容易出售的资产做抵押。
5. 环境条件（用 X5 表示），指企业的外部因素，即指非企业本身能控制或操纵的因素。

## 四、PCA 案例分析

首先并抽取了 10 家具有可比性的同类企业作为样本，又请 8 位专家分别给 10 个企业的 5 个指标打分，然后分别计算企业 5 个指标的平均值，如下表。

企业	1	2	3	4	5	6	7	8	9	10
品格	76.5	81.5	76	75.8	71.7	85	79.2	80.3	84.4	76.5
能力	70.6	73	67.6	68.1	78.5	94	94	87.5	89.5	92
资本	90.7	87.3	91	81.5	80	84.6	66.9	68.8	64.8	66.4
担保	77.5	73.6	70.9	69.8	74.8	57.7	60.4	57.4	60.8	65
环境	85.6	68.5	70	62.2	76.5	70	69.2	71.7	64.9	68.9



原始数据

76.5000	70.6000	90.7000	77.5000	85.6000
81.5000	73.0000	87.3000	73.6000	68.5000
76.0000	67.6000	91.0000	70.9000	70.0000
75.8000	68.1000	81.5000	69.8000	62.2000
71.7000	78.5000	80.0000	74.8000	76.5000
85.0000	94.0000	84.6000	57.7000	70.0000
79.2000	94.0000	66.9000	60.4000	69.2000
80.3000	87.5000	68.8000	57.4000	71.7000
84.4000	89.5000	64.8000	60.8000	64.9000
76.5000	92.0000	66.4000	65.0000	68.9000

均值

78.6900	81.4800	78.2000	66.7900	70.7500
---------	---------	---------	---------	---------

中心化

-2.1900	-10.8800	12.5000	10.7100	14.8500
2.8100	-8.4800	9.1000	6.8100	-2.2500
-2.6900	-13.8800	12.8000	4.1100	-0.7500
-2.8900	-13.3800	3.3000	3.0100	-8.5500
-6.9900	-2.9800	1.8000	8.0100	5.7500
6.3100	12.5200	6.4000	-9.0900	-0.7500
0.5100	12.5200	-11.3000	-6.3900	-1.5500
1.6100	6.0200	-9.4000	-9.3900	0.9500
5.7100	8.0200	-13.4000	-5.9900	-5.8500
-2.1900	10.5200	-11.8000	-1.7900	-1.8500

协方差

17.4677	23.8520	-9.9489	-20.6757	-9.5172
23.8520	121.7307	-87.7656	-68.0747	-13.8078
-9.9489	-87.7656	110.3156	52.9433	26.8544
-20.6757	-68.0747	52.9433	55.7677	23.0128
-9.5172	-13.8078	26.8544	23.0128	41.5361

特征值

253.6856	43.4377	36.4894	7.3101	5.8948
----------	---------	---------	--------	--------

特征向量

-0.1343	0.0143	0.4766	-0.3259	0.8053
-0.6526	-0.4857	0.1822	0.5521	0.0155
0.5945	-0.2363	0.6916	0.2721	-0.1958
0.4172	-0.0789	-0.4457	0.5551	0.5594
0.1690	-0.8378	-0.2506	-0.4547	0.0074

贡献率

0.7315	0.1252	0.1052	0.0211	0.0170
--------	--------	--------	--------	--------

累积贡献率

0.7315	0.8567	0.9619	0.9830	1.0000
--------	--------	--------	--------	--------

原始数据

得分表

76.5000	70.6000	90.7000	77.5000	85.6000	21.8039	-10.9877	-2.8753	-2.6978	1.7211
81.5000	73.0000	87.3000	73.6000	68.5000	13.0278	3.3556	3.6161	1.6824	4.1429
76.0000	67.6000	91.0000	70.9000	70.0000	18.6173	3.9813	3.3975	-0.6803	-2.5934
75.8000	68.1000	81.5000	69.8000	62.2000	10.8926	12.6021	-0.7321	0.0115	-1.5593
71.7000	78.5000	80.0000	74.8000	76.5000	8.2669	-4.5277	-7.6398	2.9550	-1.5042
85.0000	94.0000	84.6000	57.7000	70.0000	-9.1318	-6.1566	13.9535	1.8921	-1.0688
79.2000	94.0000	66.9000	60.4000	69.2000	-17.8851	-1.5998	-2.0544	0.8283	-0.7692
80.3000	87.5000	68.8000	57.4000	71.7000	-13.4903	-0.7339	-0.6899	-5.4039	-2.0157
84.4000	89.5000	64.8000	60.8000	64.9000	-17.4549	4.7273	-0.9494	-1.7452	3.9519
76.5000	92.0000	66.4000	65.0000	68.9000	-14.6463	-0.6607	-6.0262	3.1580	-0.3054

21.8039 18.6173 13.0278 10.8926 8.2669 -9.1318 -13.4903 -14.6463 -17.4549 -17.8851

1 3 2 4 5 6 8 10 9 7

在正确评估了顾客的信用等级后，就能正确制定出对其的信用期、收帐政策等，这对于加强应收帐款的管理大有帮助。



# 课后讨论

某农业生态经济系统各区域单元的有关数据

样本 序号	x <sub>1</sub> : 人 口密度 (人 /km <sup>2</sup> )	x <sub>2</sub> : 人 均耕地 面积 (ha)	x <sub>3</sub> : 森 林覆盖 率(%)	x <sub>4</sub> : 农 民人均 纯收入 (元/人)	x <sub>5</sub> : 人 均粮食 产量 (kg/ 人)	x <sub>6</sub> : 经济 作物占农 作物播面 比例(%)	x <sub>7</sub> : 耕地 占土地面 积比率 (%)	x <sub>8</sub> : 果 园与林 地面积 之比	x <sub>9</sub> : 灌溉 田占耕地 面积之比 (%)
1	363.912	0.352	16.101	192.11	295.34	26.724	18.492	2.231	26.262
2	141.503	1.684	24.301	1752.35	452.26	32.314	14.464	1.455	27.066
3	100.695	1.067	65.601	1181.54	270.12	18.266	0.162	7.474	12.489
4	143.739	1.336	33.205	1436.12	354.26	17.486	11.805	1.892	17.534
5	131.412	1.623	16.607	1405.09	586.59	40.683	14.401	0.303	22.932
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



# 内容

---

- 一、问题的提出
- 二、主成分分析原理
- 三、求解步骤（算法）
- 四、实例分析
- 五、小结
- 六、参考文献

## 五、小结

---

根据主成分分析的定义及性质，我们已大体上能看出主成分分析的一些应用。概括起来说，主成分分析主要有以下几方面的应用。

**1. 主成分分析能降低所研究的数据空间的维数。**即用研究  $m$  维的  $F$  空间代替  $p$  维的  $X$  空间( $m < p$ )，而低维的  $Y$  空间代替高维的  $X$  空间所损失的信息很少。即：使只有一个主成分  $F1$  (即  $m = 1$ ) 时，这个  $F1$  仍是使用全部  $X$  变量( $p$  个)得到的。例如要计算  $F1$  的均值也得使用全部  $X$  的均值。在所选的前  $m$  个主成分中，如果某个  $X_i$  的系数全部近似于零的话，就可以把这个  $X_i$  删除，这也是一种删除多余变量的方法。

## 五、小结

---

2. 多维数据的一种图形表示方法。我们知道当维数大于 3 时便不能画出几何图形，多元统计研究的问题大都多于 3 个变量。要把研究的问题用图形表示出来是不可能的。然而，经过主成分分析后，我们可以选取前两个主成分或其中某两个主成分，根据主成分的得分，画出  $n$  个样品在二维平面上的分布状况，由图形可直观地看出各样本在主分量中的地位。

## 五、小结

---

3. 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。

4. 用主成分分析筛选回归变量。回归变量的选择有着重的实际意义，为了使模型本身易于做结构分析、控制和预报，好从原始变量所构成的子集合中选择最佳变量，构成最佳变量集合。用主成分分析筛选变量，可以用较少的计算量来选择量，获得选择最佳变量子集合的效果。

# 内容

---

1. 问题的提出
2. 主成分分析原理
3. 求解步骤（算法）
4. 案例分析
5. 小结
6. 参考文献



## 六、参考文献

---

1. 边肇祺, 张学工. 模式识别. 北京: 清华大学出版社, 2000.
2. I.T.Jolliffe. *Principal Component Analysis*. Springer, 2010.
3. Jonathon Shlens. *A Tutorial on Principal Component Analysis*. 2005.





华中科技大学

# 讲座2: Fisher 线性分类

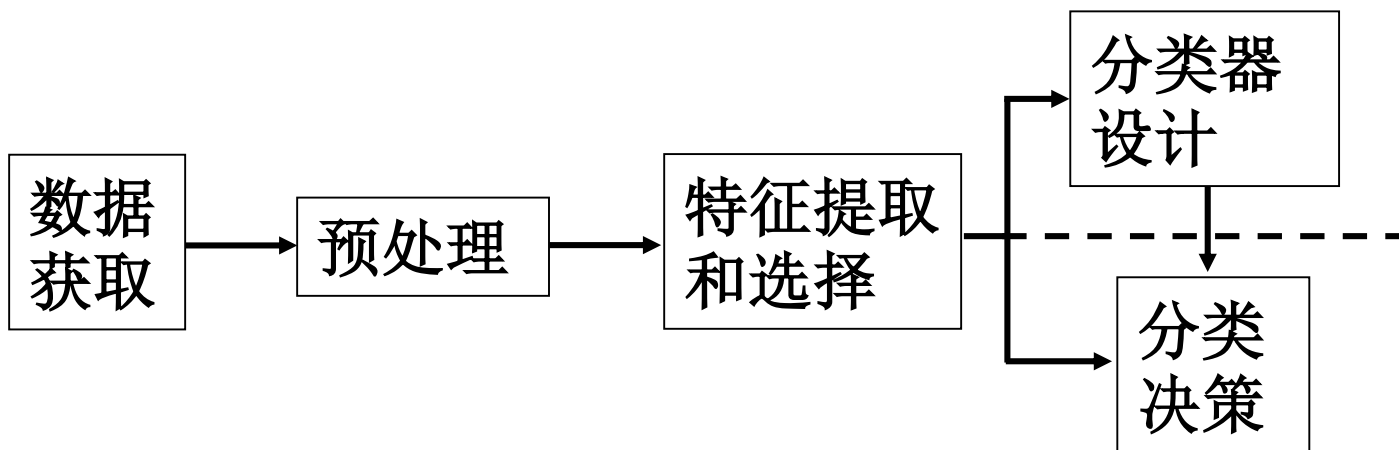
## *Fisher Linear Discriminant*

# 内容

- 一、引言：模式识别系统
- 二、判别分析之一：距离分析
- 三、判别分析之二：***Fisher*** 分析法
- 四、实例分析
- 五、小结
- 六、参考文献

# 一、引言：模式识别系统

## ➤ 模式识别系统的基本构成



- 数据采集
- 特征选取
- 模型选择 --- 训练和测试
- 计算结果和复杂度分析、反馈

# 一、引言：模式识别系统

## ➤ 实例1：模式识别

在传送带上利用光学传感器  
件对鱼按品种分类：

鲈鱼( Seabass )

鲑鱼( Salmon )



# 一、引言：模式识别系统

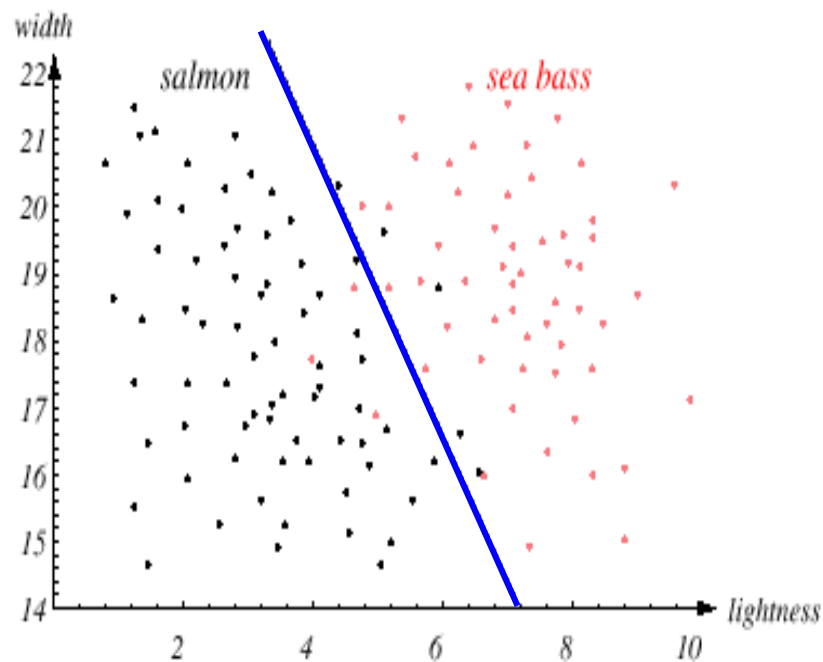
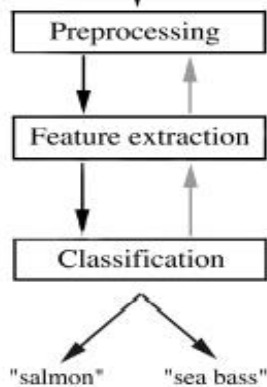
## ➤ 实例1：模式识别——识别过程

1. **数据获取：**架设一个摄像机，采集一些样本图像，获取样本数据
2. **预处理：**去噪声，用一个分割操作把鱼和鱼之间以及鱼和背景之间分开
3. **特征提取和选择：**对单个鱼的信息进行特征选择，从而通过测量某些特征来减少信息量  
长度      亮度      宽度      鱼翅的数量和形状  
嘴的位置，等等 ...
4. **分类决策：**把特征送入决策分类器



# 一、引言：模式识别系统

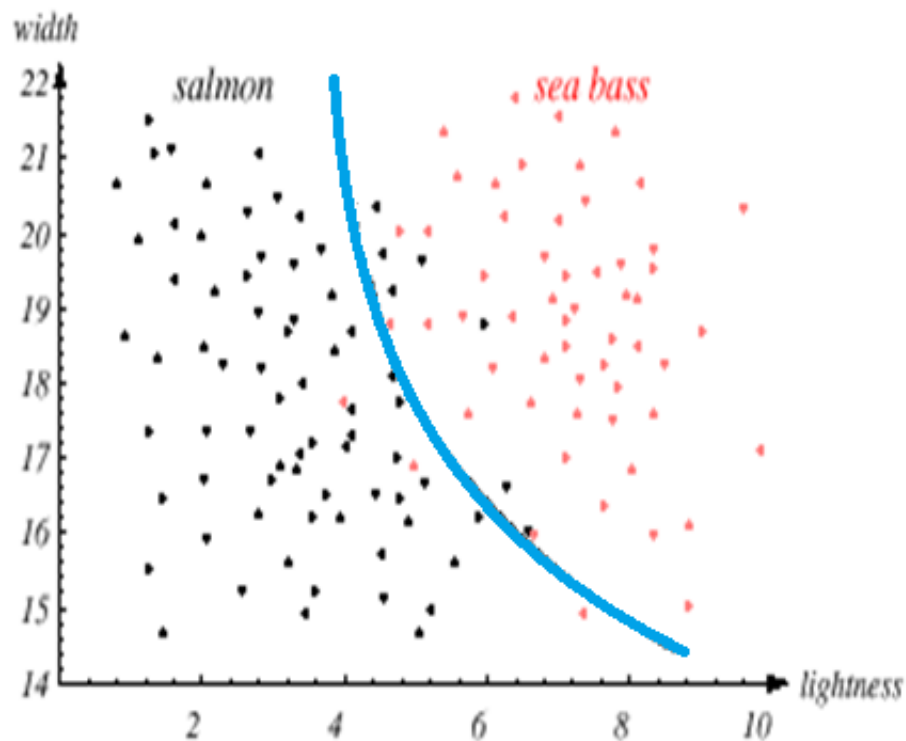
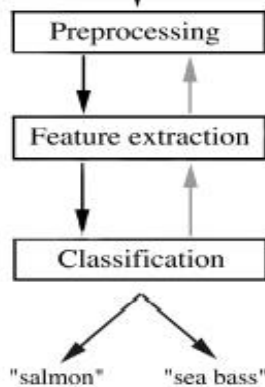
## ➤ 实例1：模式识别——识别过程





# 一、引言：模式识别系统

## ➤ 实例1：模式识别——识别过程



# 一、引言：模式识别系统

## ➤ 实例2：模式识别

19 名男女同学进行体检，测量了身高和体重，但事后发现其中有 4 人忘记填写性别，试问（在最小错误的条件下）这 4 人是男是女？体检数值如下：

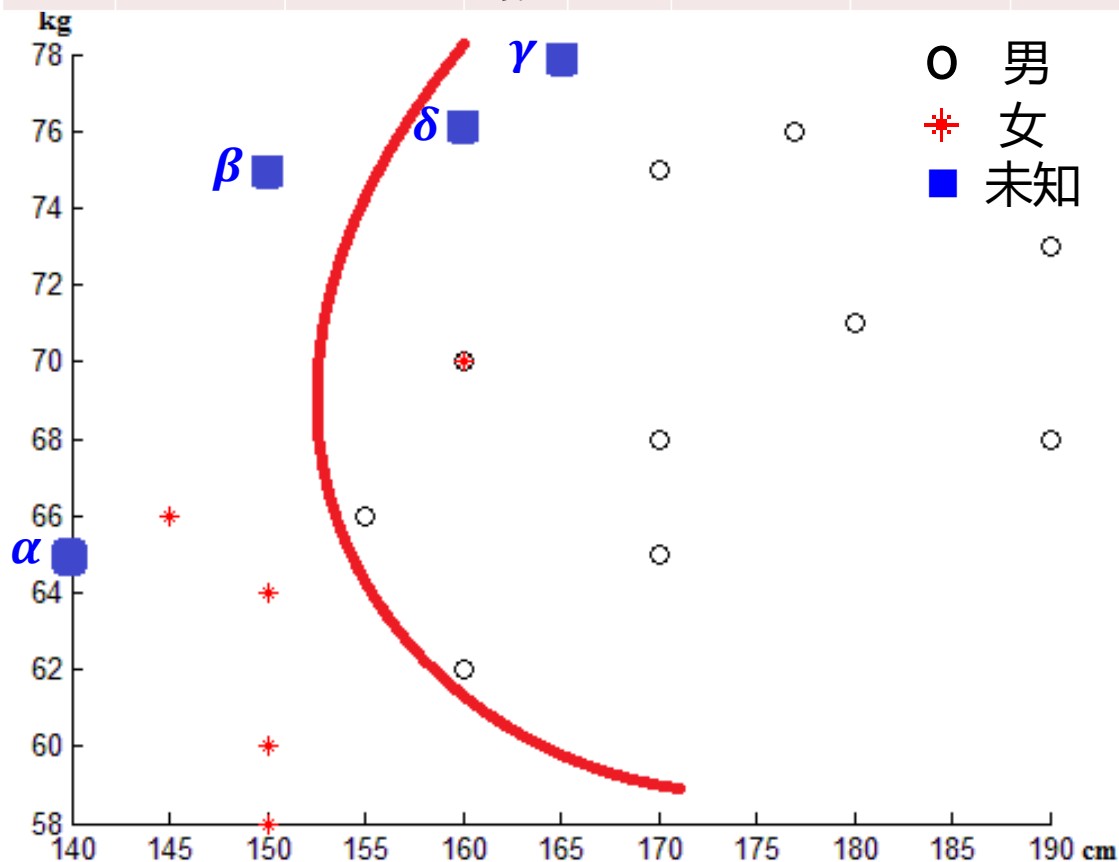
编号	身高 (cm)	体重 (kg)	性别	编号	身高 (cm)	体重 (kg)	性别
1	170	68	男	11	160	62	男
2	150	60	女	12	150	64	女
3	180	71	男	13	145	66	女
4	190	73	男	14	170	65	男
5	160	70	女	15	160	70	男
6	155	66	男	$\alpha$	140	65	?
7	190	68	男	$\beta$	150	75	?
8	177	76	男	$\gamma$	165	78	?
9	150	58	女	$\delta$	160	76	?
10	170	75	男				

# 一、引言

## ➤ 实例2：模式识别

■ 由训练样本得到的特征空间分布图

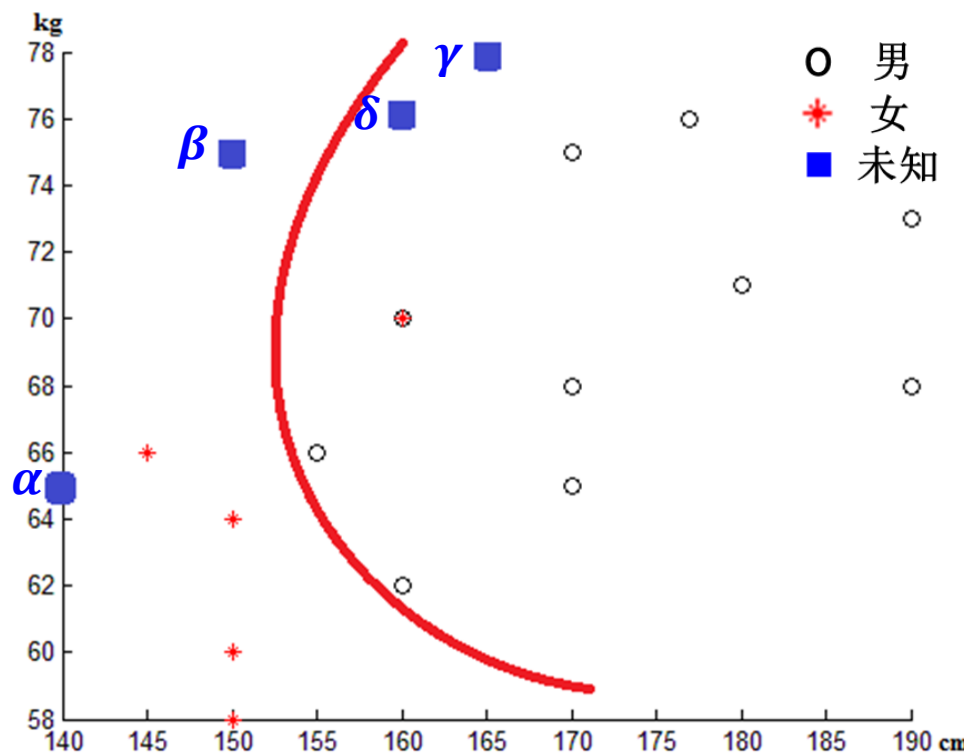
编号	身高 (cm)	体重 (kg)	性别	编号	身高 (cm)	体重 (kg)	性别
1	170	68	男	11	160	62	男
2	150	60	女	12	150	64	女
3	180	71	男	13	145	66	女
4	190	73	男	14	170	65	男
5	160	70	女	15	160	70	男
6	155	66	男	$\alpha$	140	65	?
7	190	68	男	$\beta$	150	75	?
8	177	76	男	$\gamma$	165	78	?
9	150	58	女	$\delta$	160	76	?
10	170	75	男				



# 一、引言：模式识别系统

## ➤ 实例2：模式识别

■ 由训练样本得到的特征空间分布图



1. 待识别的模式

性别（男或女）

2. 测量的特征

身高和体重

3. 训练样本

15 名已知性别的样本特征

4. 目标

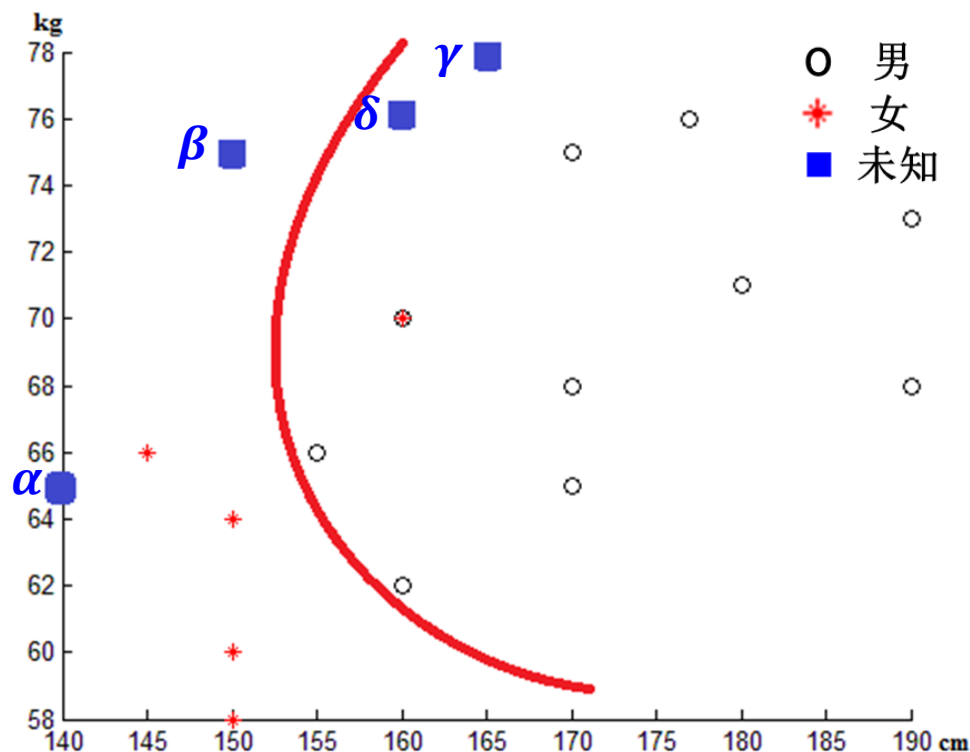
希望借助于训练样本的特征

建立判别函数

# 一、引言：模式识别系统

## ➤ 实例2：模式识别

### ■ 由训练样本得到的特征空间分布图



从图中训练样本的分布情况，找出男、女两类特征各自的聚类特点，从而求取一个判别函数（直线或曲线）。

只要给出待分类的模式特征的数值，看它在特征平面上落在判别函数的哪一侧，就可以判别是男还是女了。

# 内容

- 一、引言：模式识别系统
- 二、判别分析之一：距离分析
- 三、判别分析之二：*Fisher* 分析法
- 四、实例分析
- 五、小结
- 六、参考文献

## 二、统计方法

### ➤ 判别分析

在已知研究对象分成若干类型，并已取得各种类型的一批已知样品的观测数据，在此基础上，根据某些准则建立判别式，然后对未知类型的样品进行判别分类。

### ➤ 距离判别法

首先根据已知分类的数据，分别计算各类的重心，计算新个体到每类的距离，确定最短的距离（欧氏距离、马氏距离）。

### ➤ *Fisher* 判别法

利用已知类别个体的指标构造判别式，原则为：

同类差别较小、不同类差别较大

按照判别式的值判断新个体的类别。

## 二、统计方法

### ➤ 距离判别法

**基本思想：**首先根据已知分类的数据，分别计算各类的重心即分组（类）的均值，判别准则是：对任给的一次观测，若它与第  $i$  类的重心距离最近，就认为它来自第  $i$  类。

➤ 距离判别法，对各类或总体的分布，并无特定的要求。



## 二、统计方法（判别分析）

### ➤ 两个总体的距离判别法

设有两个总体（或称两类） $G_1$ 、 $G_2$ ，从第一个总体中抽取  $n_1$  个样品，从第二个总体中抽取  $n_2$  个样品，每个样品测量  $p$  个指标如表。

今任取一个样品，实测指标值  $X = (x_1, x_2, \dots, x_p)^T$ ，  
问  $X$  应判归为哪一类？

## 二、统计方法（判别分析）

### ➤ 两个总体的距离判别法

**G1**

变量 样品	$x_1$	$x_2$	...	$x_p$	
$x_1^{(1)}$	$x_{11}^{(1)}$	$x_{12}^{(1)}$	...	$x_{1p}^{(1)}$	
$x_2^{(1)}$	$x_{21}^{(1)}$	$x_{22}^{(1)}$	...	$x_{2p}^{(1)}$	
$\vdots$	$\vdots$	$\vdots$		$\vdots$	
$x_m^{(1)}$	$x_{m1}^{(1)}$	$x_{m2}^{(1)}$	...	$x_{mp}^{(1)}$	
均值	$\bar{x}_1^{(1)}$	$\bar{x}_2^{(1)}$	...	$\bar{x}_p^{(1)}$	

**G2**

变量 样品	$x_1$	$x_2$	...	$x_p$
$x_1^{(2)}$	$x_{11}^{(2)}$	$x_{12}^{(2)}$	...	$x_{1p}^{(2)}$
$x_2^{(2)}$	$x_{21}^{(2)}$	$x_{22}^{(2)}$		$x_{2p}^{(2)}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
$x_n^{(2)}$	$x_{n1}^{(2)}$	$x_{n2}^{(2)}$	...	$x_{np}^{(2)}$
均值	$\bar{x}_1^{(2)}$	$\bar{x}_2^{(2)}$	...	$\bar{x}_p^{(2)}$

$$\bar{x}_j^{(1)} = \frac{1}{m} \sum_{i=1}^m x_{ij}^{(1)}$$

$$\bar{x}_j^{(2)} = \frac{1}{n} \sum_{i=1}^n x_{ij}^{(2)}$$

## 二、统计方法（判别分析）

### ➤ 两个总体的距离判别法

首先计算  $x$  到  $G1$ 、 $G2$   
总体的距离，分别记为  
 $D(x, G1)$  和  $D(x, G2)$

$G1$						$G2$					
变量 样品	$x_1$	$x_2$	$\dots$	$x_p$		变量 样品	$x_1$	$x_2$	$\dots$	$x_p$	
$x_1^{(1)}$	$x_{11}^{(1)}$	$x_{12}^{(1)}$	$\dots$	$x_{1p}^{(1)}$		$x_1^{(2)}$	$x_{11}^{(2)}$	$x_{12}^{(2)}$	$\dots$	$x_{1p}^{(2)}$	
$x_2^{(1)}$	$x_{21}^{(1)}$	$x_{22}^{(1)}$	$\dots$	$x_{2p}^{(1)}$		$x_2^{(2)}$	$x_{21}^{(2)}$	$x_{22}^{(2)}$	$\dots$	$x_{2p}^{(2)}$	
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$	$\vdots$		$\vdots$	
$x_m^{(1)}$	$x_{m1}^{(1)}$	$x_{m2}^{(1)}$	$\dots$	$x_{mp}^{(1)}$		$x_n^{(2)}$	$x_{n1}^{(2)}$	$x_{n2}^{(2)}$	$\dots$	$x_{np}^{(2)}$	
均值	$\bar{x}_1^{(1)}$	$\bar{x}_2^{(1)}$	$\dots$	$\bar{x}_p^{(1)}$		均值	$\bar{x}_1^{(2)}$	$\bar{x}_2^{(2)}$	$\dots$	$\bar{x}_p^{(2)}$	

$$\begin{cases} x \in G1, \text{ if } D(x, G1) < D(x, G2); \\ x \in G2, \text{ if } D(x, G1) > D(x, G2); \\ \text{待判, if } D(x, G1) = D(x, G2) \end{cases}$$

## 二、统计方法（判别分析）

### ➤ 两个总体的距离判别法

- 如果采用欧氏距离，则可计算出

$$D(x, G1) = \sqrt{(x - \bar{x}^{(1)})'(x - \bar{x}^{(1)})} = \sqrt{\sum_{j=1}^p (x_j - \bar{x}_j^{(1)})^2}$$

$$D(x, G2) = \sqrt{(x - \bar{x}^{(2)})'(x - \bar{x}^{(2)})} = \sqrt{\sum_{j=1}^p (x_j - \bar{x}_j^{(2)})^2}$$

- 然后比较  $D(x, G1)$  和  $D(x, G2)$  大小，按距离最近准则进行判别归类。

## 二、统计方法（判别分析）

### ➤ 两个总体的距离判别法

如果采用马氏距离，即有

$$D(x, G) = \sqrt{(x - \bar{x}^{(i)})' (\Sigma^{(i)})^{-1} (x - \bar{x}^{(i)})}, i = 1, 2.$$



# 内容

- 一、引言：模式识别系统
- 二、判别分析之一：距离分析
- 三、判别分析之二： *Fisher* 分析法
- 四、实例分析
- 五、小结
- 六、参考文献

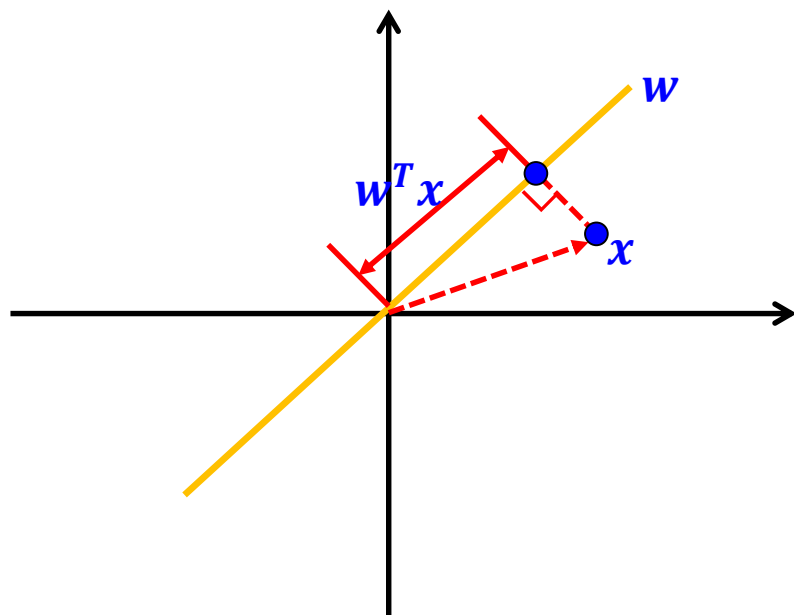
# 三、Fisher 线性判别

## ➤ 投影

称线性函数

$$y = g(x) = w^T x, \quad \|w\| = 1.$$

为向量  $x$  在向量  $w$  上的投影(或正交投影)。

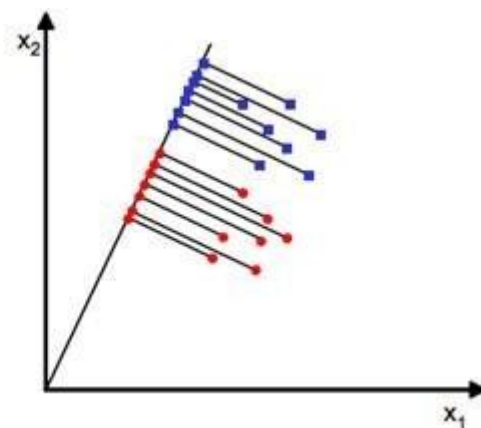
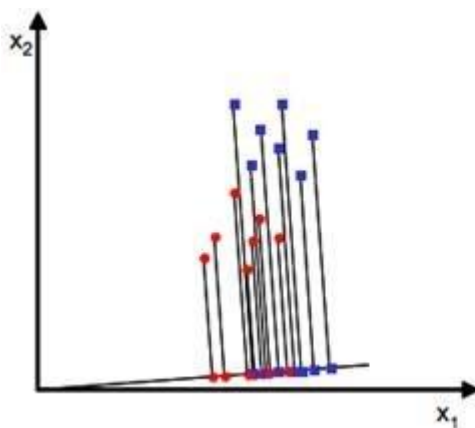
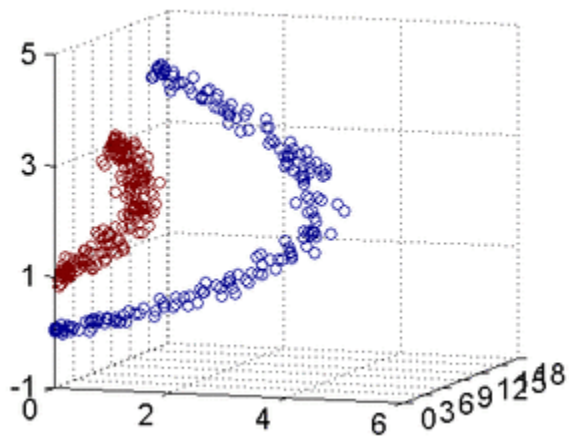


- 样本向量  $x$  与权向量  $w$  的向量点积
- 样本向量  $x$  各分量的线性加权

# 三、Fisher 线性判别

## ➤ Fisher 准则的基本原理

找到一个最合适的投影轴，使两类样本在该轴上投影之间的距离尽可能远，而每一类样本的投影尽可能紧凑，从而使两类分类效果为最佳。

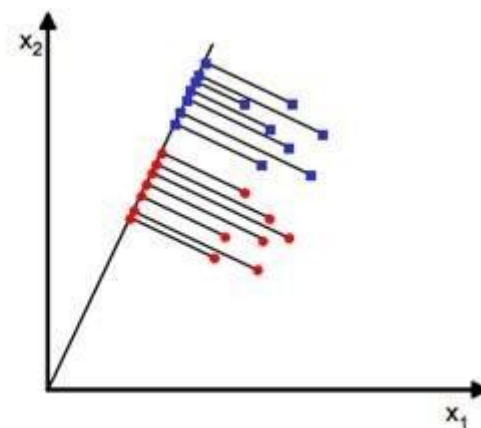
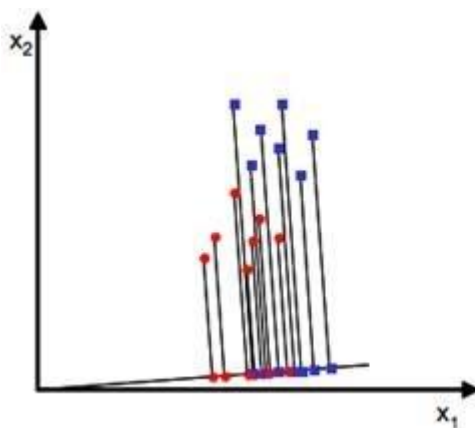
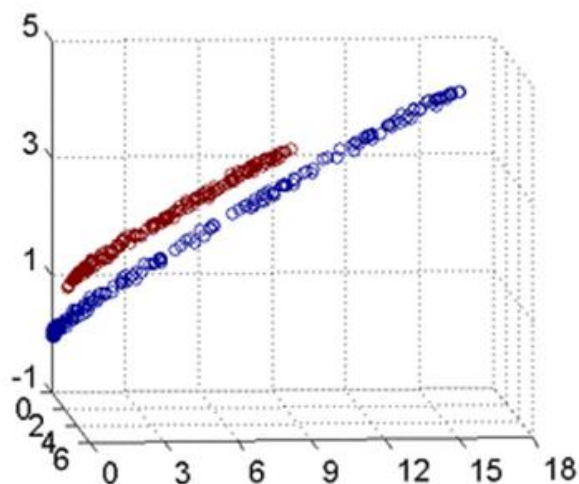




# 三、Fisher 线性判别

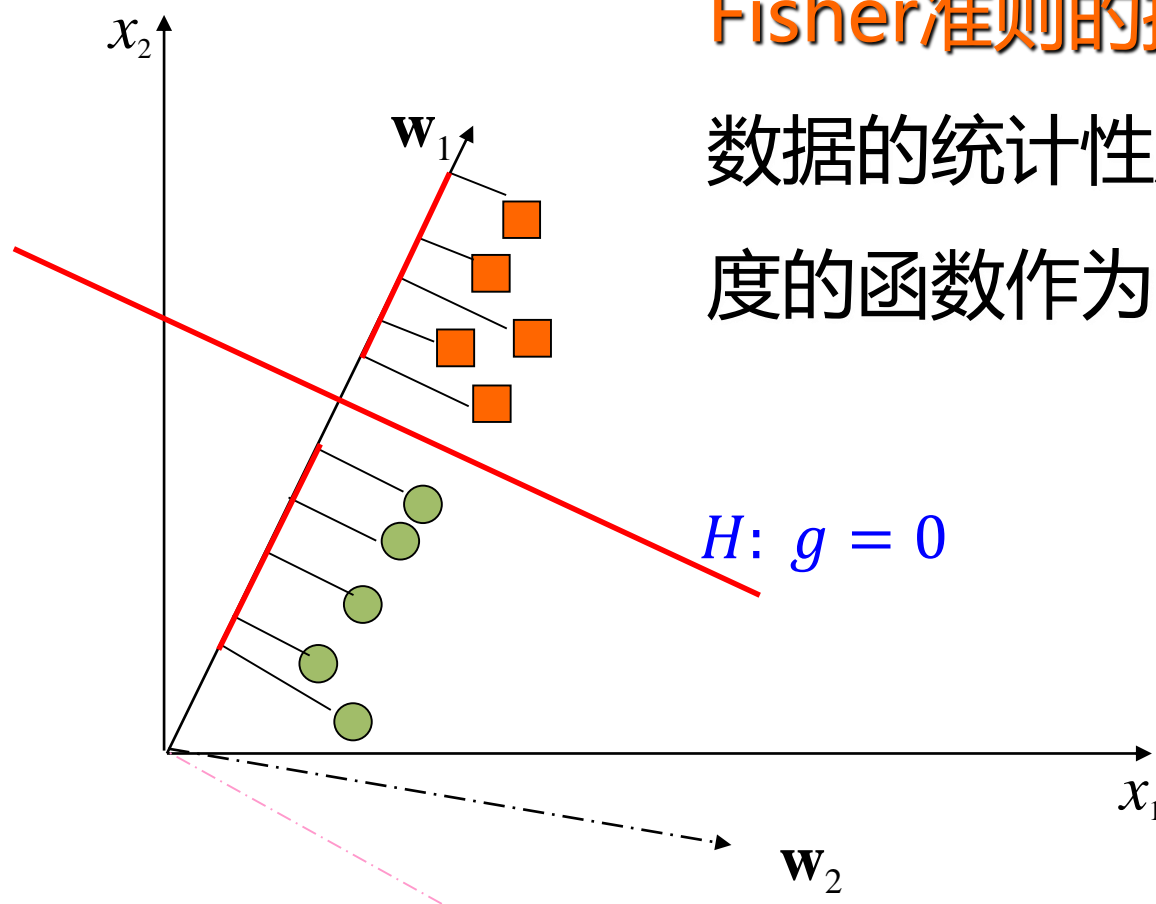
## ➤ Fisher 准则的基本原理

找到一个最合适的投影轴，使两类样本在该轴上投影之间的距离尽可能远，而每一类样本的投影尽可能紧凑，从而使两类分类效果为最佳。



# 三、Fisher 线性判别

## ➤ Fisher 线性判别图例



**Fisher准则的描述**: 用投影后数据的统计性质—均值和离散度的函数作为判别优劣的标准。

### 三、Fisher线性判别

#### ➤ $X$ 空间样本分布的统计描述量

1. 各类样本均值向量  $m_i$

$$m_i = \frac{1}{N_i} \sum_{X \in K_i} X, i = 1, 2.$$

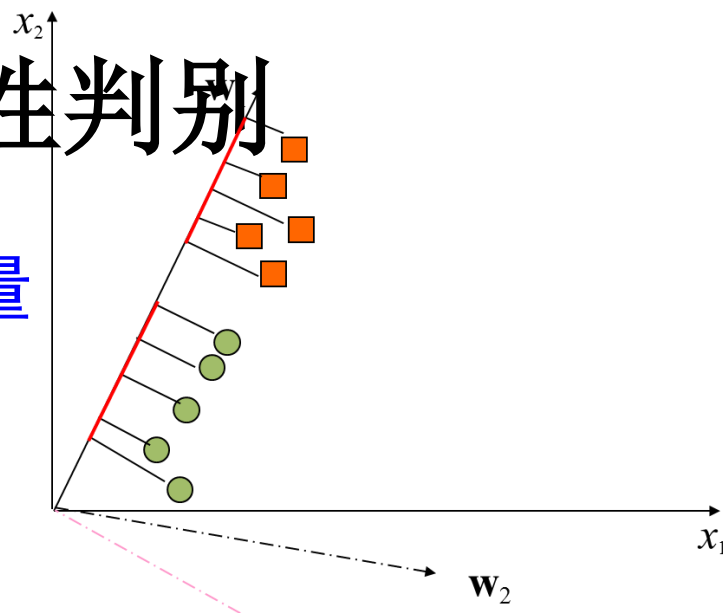
2. 类内离散度矩阵  $S_i$  与总类内离散度矩阵  $S_w$

$$S_i = \sum_{X \in K_i} (X - m_i)(X - m_i)^T, i = 1, 2.$$

$$S_w = S_1 + S_2$$

3. 样本类间离散度矩阵  $S_b$

$$S_b = (m_1 - m_2)(m_1 - m_2)^T$$



### 三、Fisher线性判别

➤  $Y$  空间样本分布的统计描述量

1. 各类样本均值向量  $\tilde{m}_i$

$$\tilde{m}_i = \frac{1}{N_i} \sum_{Y \in K_i} Y, i = 1, 2.$$

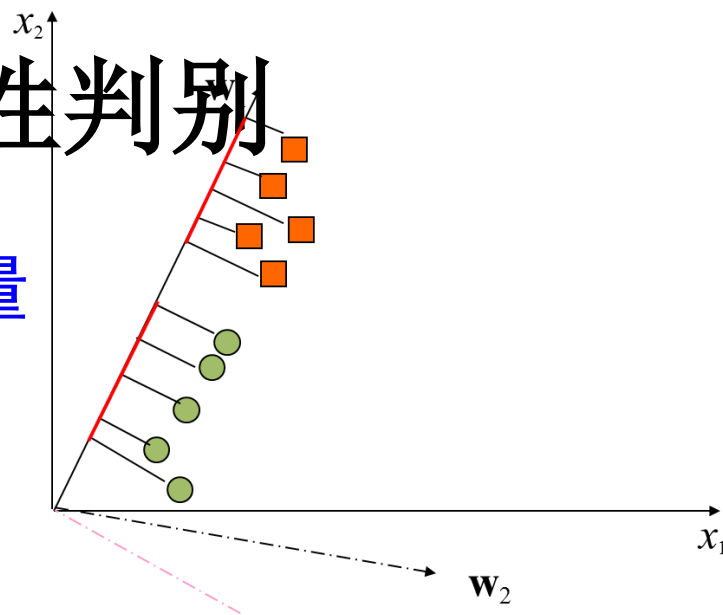
2. 类内离散度  $S_i$  与总类内离散度矩阵  $S_w$

$$\tilde{S}_i = \sum_{Y \in K_i} (Y - \tilde{m}_i)^2, i = 1, 2.$$

$$\tilde{S}_w = \tilde{S}_1 + \tilde{S}_2$$

3. 样本类间离散度矩阵  $S_b$

$$\tilde{S}_b = (\tilde{m}_1 - \tilde{m}_2)^2$$

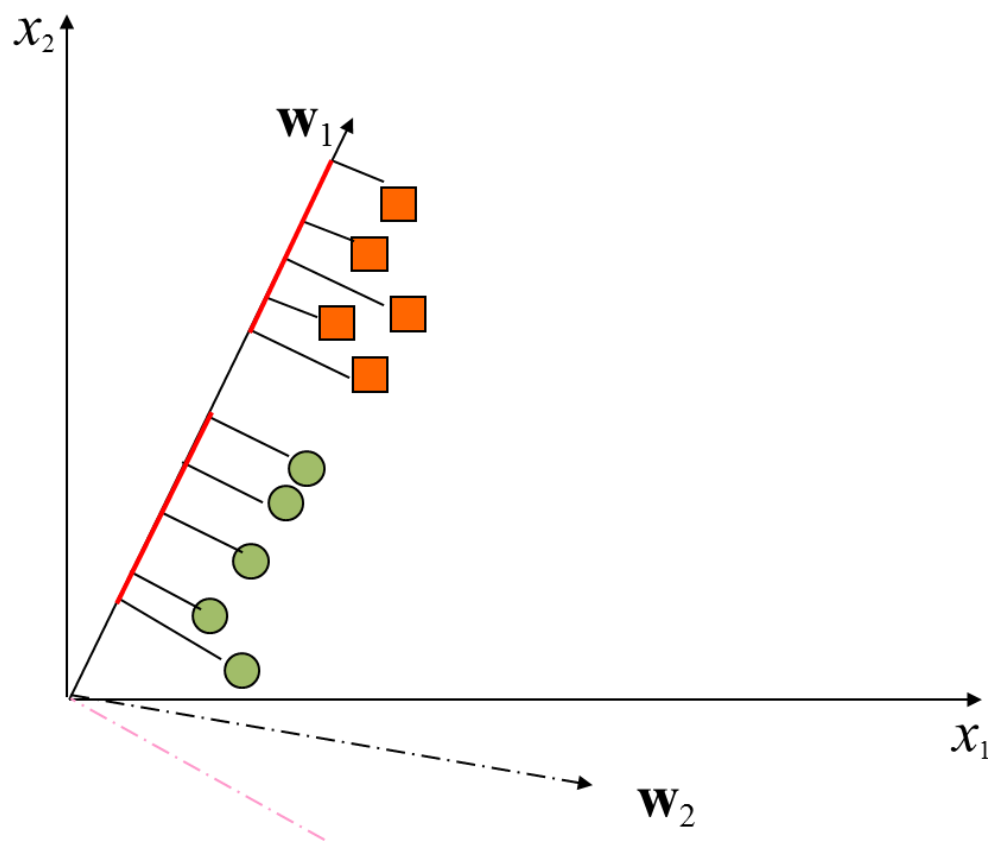


### 三、Fisher线性判别

➤ 样本  $X$  与其投影  $Y$  的统计量之间的关系

$$\begin{aligned}\tilde{m}_i &= \frac{1}{N_i} \sum_{Y \in K_i} Y \\ &= \frac{1}{N_i} \sum_{X \in K_i} w^T X = w^T m_i, i = 1, 2.\end{aligned}$$

$$\begin{aligned}\tilde{S}_b &= (\tilde{m}_1 - \tilde{m}_2)^2 \\ &= (w^T m_1 - w^T m_2)^2 \\ &= w^T (m_1 - m_2)(m_1 - m_2)^T w \\ &= w^T S_b w\end{aligned}$$

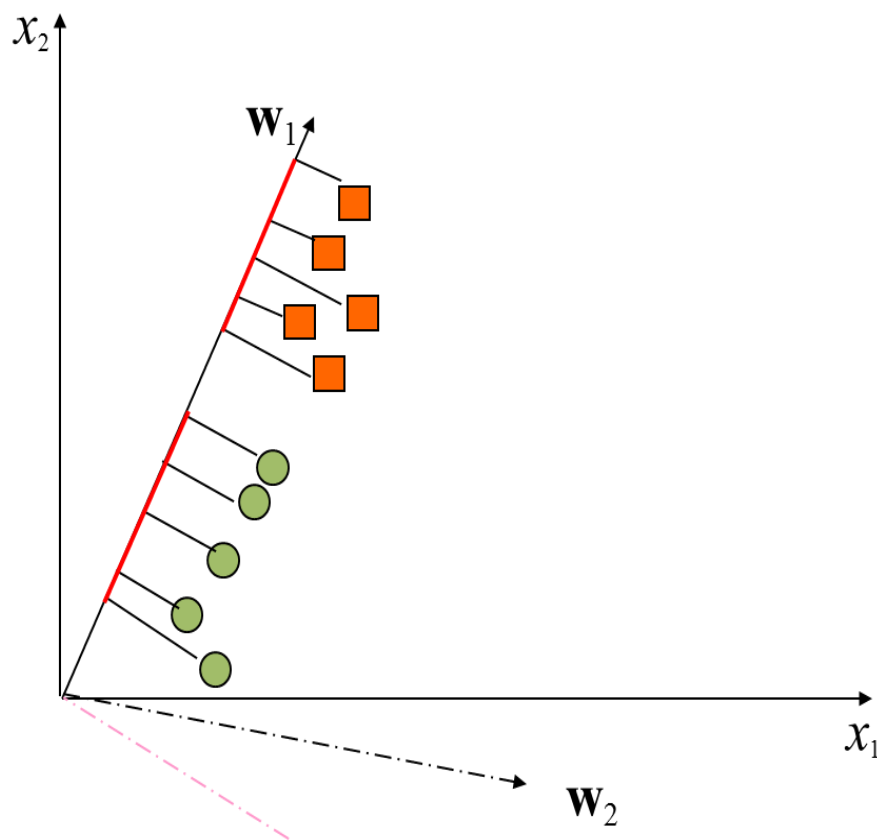


### 三、Fisher线性判别

➤ 样本  $X$  与其投影  $Y$  的统计量之间的关系

$$\begin{aligned}\tilde{S}_i &= \sum_{Y \in K_i} (Y - \tilde{m}_i)^2 \\ &= \sum_{X \in K_i} (w^T X - w^T m_i)^2 \\ &= w^T \left( \sum_{X \in K_i} (X - m_i)(X - m_i)^T \right) w \\ &= w^T S_i w\end{aligned}$$

$$\tilde{S}_1 + \tilde{S}_2 = w^T (S_1 + S_2) w = w^T S_w w$$



# 三、Fisher线性判别

## ➤ Fisher 准则函数

评价投影方向  $w$  的原则，使原样本向量在该方向上的投影能兼顾类间分布尽可能分开，类内样本投影尽可能密集的要求。

- Fisher 准则函数的定义：

$$J_F(w) = \frac{\tilde{S}_b}{\tilde{S}_w} = \frac{\tilde{S}_b}{\tilde{S}_1 + \tilde{S}_2} = \frac{w^T S_b w}{w^T S_w w}$$

- Fisher 最佳投影方向的求解：

$$w^* = \arg \max_w J_F(w)$$

# 三、Fisher线性判别

➤ *Fisher* 最佳投影方向的求解

✓ 采用拉格朗日乘子法解决

$$w^* = S_w^{-1}(m_1 - m_2)$$

$m_1 - m_2$  是一向量，对与  $m_1 - m_2$  平行的向量投影可使两均值点的距离最远。

但是如从使类间分得较开，同时又使类内密集程度较高这样一个综合指标来看，则需根据两类样本的分布离散程度对投影方向作相应的调整，这就体现在对  $m_1 - m_2$  向量按  $S_w^{-1}$  作一线性变换，从而使 *Fisher* 准则函数达到极值点。



# 三、Fisher线性判别

## ➤ Fisher线性判别分析的算法步骤

- ① 把来自两类  $G_1 / G_2$  的训练样本集  $X$  区分为  $G_1$  和  $G_2$  的两个子集  $X_1$  和  $X_2$  ;
- ② 计算各类的均值向量  $m_i, i = 1, 2$  ;
- ③ 计算各类的类内离散度矩阵  $S_i, i = 1, 2$  ;
- ④ 计算类内总离散度矩阵  $S_w = S_1 + S_2$  ;
- ⑤ 计算矩阵  $S_w$  的逆矩阵  $S_w^{-1}$  ;
- ⑥ 求解向量  $w^*$  。

$$m_i = \frac{1}{N_i} \sum_{x \in K_i} x \quad i=1,2$$

$$S_i = \sum_{x \in \mathcal{X}_i} (x - m_i)(x - m_i)^T, \quad i=1,2$$

$$S_w = S_1 + S_2$$

$$w^* = S_w^{-1}(m_1 - m_2)$$

# 三、Fisher线性判别

## ► 判别函数的确定

讨论了使 *Fisher* 准则函数极大的  $d$  维向量  $w^*$  的计算方法, 判别函数中的另一项  $w_0$  (阈值) 可采用以下几种方法确定:

$$w_0 = -\frac{\tilde{m}_1 + \tilde{m}_2}{2}$$

$$w_0 = -\frac{N_1\tilde{m}_1 + N_2\tilde{m}_2}{N_1 + N_2}$$

分类规则:

$$y = w^T x + w_0 > 0 \Rightarrow x \in G_1$$

$$y = w^T x + w_0 < 0 \Rightarrow x \in G_2$$

### 三、Fisher线性判别

➤ Fisher 公式的推导

$$\mathbf{w}^* = \arg \max_{\mathbf{w}} J_F(\mathbf{w})$$

$$J_F(\mathbf{w}) = \frac{\tilde{S}_b}{\tilde{S}_w} = \frac{\tilde{S}_b}{\tilde{S}_1 + \tilde{S}_2} = \frac{\mathbf{w}^T \mathbf{S}_b \mathbf{w}}{\mathbf{w}^T \mathbf{S}_w \mathbf{w}}$$

定义 *lagrange* 函数:

$$L(\mathbf{w}, \lambda) = \mathbf{w}^T \mathbf{S}_b \mathbf{w} - \lambda(\mathbf{w}^T \mathbf{S}_w \mathbf{w} - c)$$

$$\frac{\partial L(\mathbf{w}, \lambda)}{\partial \mathbf{w}} = \mathbf{S}_b \mathbf{w} - \lambda \mathbf{S}_w \mathbf{w} = 0 \quad \Rightarrow \quad \mathbf{S}_w^{-1} \mathbf{S}_b \mathbf{w} = \lambda \mathbf{w}$$

$$\lambda \mathbf{w} = \mathbf{S}_w^{-1} \mathbf{S}_b \mathbf{w} = \mathbf{S}_w^{-1} (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T \mathbf{w} = \mathbf{S}_w^{-1} (\mathbf{m}_1 - \mathbf{m}_2) R$$

$$\mathbf{w}^* = \frac{R}{\lambda} \mathbf{S}_w^{-1} (\mathbf{m}_1 - \mathbf{m}_2) \cong \mathbf{S}_w^{-1} (\mathbf{m}_1 - \mathbf{m}_2)$$

# 内容

一、引言：模式识别系统

二、判别分析之一：距离分析

三、判别分析之二：*Fisher* 分析法

四、实例分析

五、小结

六、参考文献

## 四、实例分析

### ➤ 例：人文发展指数的分类

衡量人文发展指数的  
指标为：

- 预期寿命
- 成人识字率
- 实际人均 *GDP*

如表有 **3** 个变量，两类  
总体各 **5** 个样本，有 **4**  
个待判样品。如何分类？

类别	序号	预期寿命 (岁)	成人识字率 (%)	人均 GDP (\$)
第一类： 高发展 水平	1	76.00	99.00	5374.00
	2	79.50	99.00	5359.00
	3	78.00	99.00	5372.00
	4	72.10	95.90	5242.00
	5	73.80	77.70	5370.00
第二类： 中等发 展水平	6	71.20	93.00	4250.00
	7	75.30	94.90	3412.00
	8	70.00	91.20	3390.00
	9	72.80	99.00	2300.00
	10	62.90	80.60	3799.00
待判 样品	11	68.50	79.30	1950.00
	12	69.90	96.90	2840.00
	13	77.60	93.80	5233.00
	14	69.30	90.30	5158.00

## 四、实例分析

### ➤ 例：人文发展指数的分类

两组线性判别的计算  
过程如下：

1. 计算样本均值
2. 计算样本协差阵
3. 建立判别函数
4. 计算判别临界值
5. 对已知类别的样品  
判别归类
6. 待判样品判别结果

类别	序号	预期寿命 (岁)	成人识字率 (%)	人均 GDP (\$)
第一类： 高发展 水平	1	76.00	99.00	5374.00
	2	79.50	99.00	5359.00
	3	78.00	99.00	5372.00
	4	72.10	95.90	5242.00
	5	73.80	77.70	5370.00
第二类： 中等发 展水平	6	71.20	93.00	4250.00
	7	75.30	94.90	3412.00
	8	70.00	91.20	3390.00
	9	72.80	99.00	2300.00
	10	62.90	80.60	3799.00
待判 样品	11	68.50	79.30	1950.00
	12	69.90	96.90	2840.00
	13	77.60	93.80	5233.00
	14	69.30	90.30	5158.00

## 四、实例分析

### ➤ 例：人文发展指数的分类

序号	判别函数 y 的值	分类
11	-12.254	2
12	-7.792	2
13	5.148	1
14	2.096	1

类别	序号	预期寿命 (岁)	成人识字率 (%)	人均 GDP (\$)
第一类： 高发展 水平	1	76.00	99.00	5374.00
	2	79.50	99.00	5359.00
	3	78.00	99.00	5372.00
	4	72.10	95.90	5242.00
	5	73.80	77.70	5370.00
第二类： 中等发 展水平	6	71.20	93.00	4250.00
	7	75.30	94.90	3412.00
	8	70.00	91.20	3390.00
	9	72.80	99.00	2300.00
	10	62.90	80.60	3799.00
待判 样品	11	68.50	79.30	1950.00
	12	69.90	96.90	2840.00
	13	77.60	93.80	5233.00
	14	69.30	90.30	5158.00



# 课后讨论

类别	序号	地区	$x_1$	$x_2$	$x_3$	$x_4$
第一组	1	辽宁	11.2	57.25	13.47	73.41
	2	河北	14.9	67.19	7.89	73.09
	3	天津	14.3	64.74	19.41	72.33
	4	北京	13.5	55.63	20.59	77.33
	5	山东	16.2	75.51	11.06	72.08
	6	上海	14.3	57.63	22.51	77.35
	7	浙江	20	83.94	15.99	89.5
	8	福建	21.8	68.03	39.42	71.9
	9	广东	19	78.31	83.03	80.75
	10	广西	16	57.11	12.57	60.91
	11	海南	11.9	49.97	30.7	69.2
第二组	12	黑龙江	8.7	30.72	15.41	60.25
	13	吉林	14.3	37.65	12.95	66.42
	14	内蒙古	10.1	34.63	7.68	62.96
	15	山西	9.1	56.33	10.3	66.01
	16	河南	13.8	65.23	4.69	64.24
	17	湖北	15.3	55.62	6.06	54.74
	18	湖南	11	55.55	8.02	67.47
	19	江西	18	62.88	6.4	58.83
	20	甘肃	10.4	30.01	4.61	60.26
	21	宁夏	8.2	29.28	6.11	50.71
	22	四川	11.4	62.88	5.31	61.49
	23	云南	11.6	28.57	9.08	68.47
	24	贵州	8.4	30.23	6.03	55.55
	25	青海	8.2	15.96	8.04	40.26
	26	新疆	10.9	24.75	8.34	46.01
	27	西藏	15.6	21.44	28.62	46.01
待判样品	28	江苏	16.5	80.05	8.81	73.04
	29	安徽	20.6	81.24	5.37	60.43
	30	陕西	8.6	42.06	8.88	56.37





# 内容

- 一、引言：模式识别系统
- 二、判别分析之一：距离分析
- 三、判别分析之二：*Fisher* 分析法
- 四、实例分析
- 五、小结
- 六、参考文献

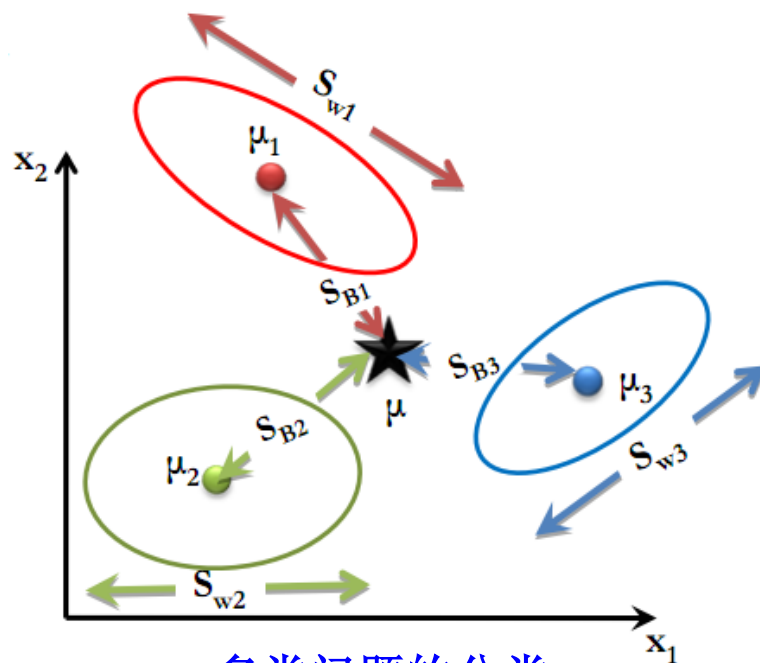
## 五、小结

线性判别，就是在特征空间内找到超平面  $H$  区分两类，或者是投影到超平面的法向量  $w$  上，来分辨两类。数据一般要服从单峰高斯分布。

多类问题是基于两类问题的。

其他分类问题常见的方法有：

线性判别方法、决策树判别、神经网络、近邻法、概率密度估计量法、支持向量机等。



多类问题的分类

# 内容

- 一、引言：模式识别系统
- 二、判别分析之一：距离分析
- 三、判别分析之二：*Fisher* 分析法
- 四、实例分析
- 五、小结
- 六、参考文献

## 六、参考文献

1. Fisher, R.A. *The Use of Multiple Measurements in Taxonomic Problems*. Annals of Eugenics, 1936,7: 179-188
2. R.O. Duda, P.E. Hart, D.H. Stork. *Pattern Classification* (2nd ed.), Wiley Interscience, 2000.
3. Friedman, J.H. *Regularized Discriminant Analysis*. Journal of the American Statistical Association, 1989
4. Martinez, A.M., Kak, A.C. *PCA versus LDA*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001,23(2):228-233.
5. Mika, S. et al. *Fisher Discriminant Analysis with Kernels*. IEEE Conference on Neural Networks for Signal Processing IX, 1999.
6. V.N. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, New York, 1995.

附：

➤ 英国统计学家和遗传学家。1890年 2月17日生于伦敦, 1962年7月29日卒于澳大利亚阿德莱德。1912年毕业于剑桥大学数学系, 后随英国数理统计学家J. 琼斯进修了一年统计力学。他担任过中学数学教师, 1918年任罗坦斯泰德农业试验站统计试验室主任。1933年, 因为在生物统计和遗传学研究方面成绩卓著而被聘为伦敦大学优生学教授。1943年任剑桥大学遗传学教授。1957年退休。1959年去澳大利亚, 在联邦科学和工业研究组织的数学统计部作研究工作。



费希尔, R. A.

Ronald Aylmer Fisher  
(1890~1962)

➤ 主要贡献有：

- ① 用亲属间的相关说明了连续变异的性状可以用孟德尔定律来解释, 从而解决了遗传学中孟德尔学派和生物统计学派的论争。
- ② 论证了方差分析的原理和方法, 并应用于试验设计, 阐明了最大似然性方法及随机化、重复性和统计控制的理论, 指出自由度作为检查K. 皮尔逊制定的统计表格的重要性。此外, 还阐明了各种相关系数的抽样分布, 进行过显著性测验研究。
- ③ 他提出的一些数学原理和方法对人类遗传学、进化论和数量遗传学的基本概念以及农业、医学方面的试验均有很大影响。例如遗传力的概念就是在他提出的可将性状分解为加性效应、非加性（显性）效应和环境效应的理论基础上建立起来的。
- ④ 主要著作有：《根据孟德尔遗传方式的亲属间的相关》、《研究者用的统计方法》、《自然选择的遗传理论》、《试验设计》、《近交的理论》及《统计方法和科学推理》等。他在进化遗传学上是一个极端的选择论者, 认为中立性状很难存在。
- ⑤ 他一生在统计生物学中的功绩是十分突出的。



祝同学们取得  
好成绩！

