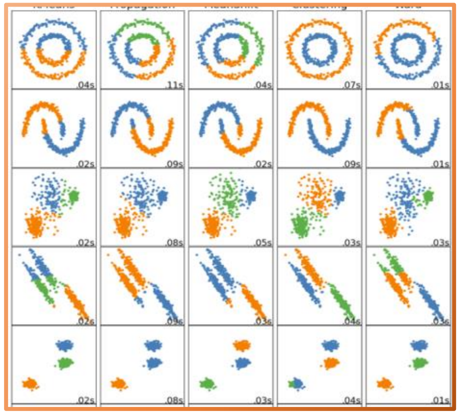


# 第十一讲　降维技术

云南大学

孙正宝

zbsun@ynu.edu.cn

# 提　纲

一　主成分分析（PCA）

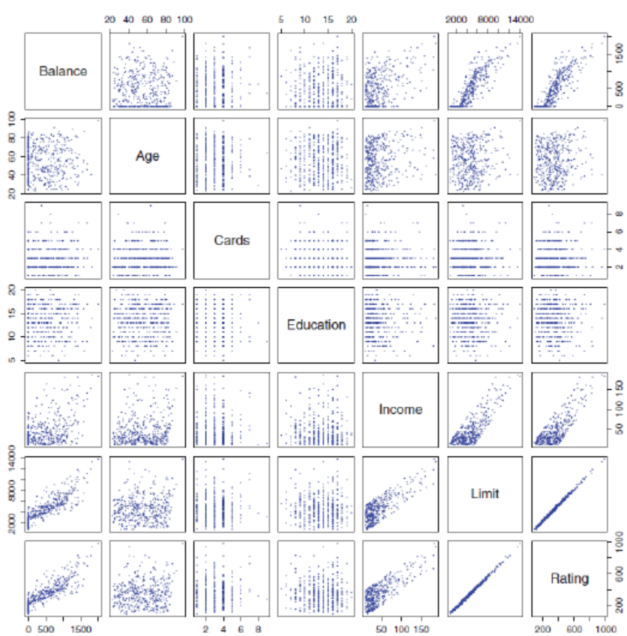
二　线性判别分析（LDA）

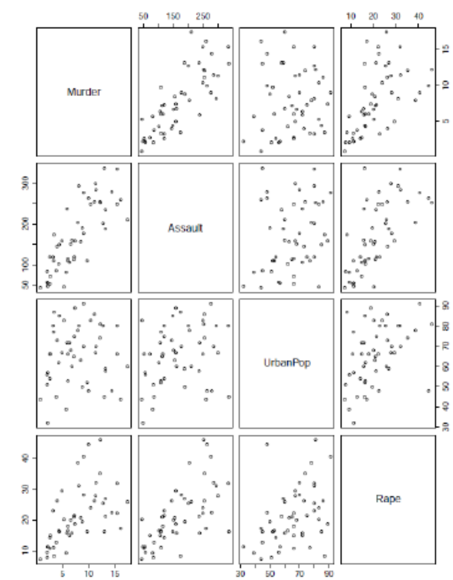
# 一、主成分分析

## 1. PCA算法概述

- 观测数据用样本矩阵***X***表示，记作

$$X = [\ \boldsymbol{x}_1 \quad \boldsymbol{x}_2 \quad \cdots \quad \boldsymbol{x}_n\ ] = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}$$



# 一、主成分分析

## 1. PCA算法概述

- 1901年，Pearson提出；1933年，Hotling发扬光大。
- 主成分分析（principal component analysis, PCA）是一种常用的无监督学习方法。
- 这一方法利用正交变换把由线性相关变量表示的观测数据转换为少数几个由线性无关变量表示的数据，线性无关的变量称为主成分。
- 主成分的个数通常小于原始变量 的个数，所以主成分分析属于降维方法。
- 主成分分析主要用于发现数据中的基本结构， 即数据中变量之间的关系。

## 2. PCA算法的基本思想

- 首先对给定数据进行规范化，使得数据每一变量的平均值为0，方差为1（标准化随机变量）。
- 然后对数据进行正交变换，原来由线性相关变量表示的数据，通过正交变换变成由若干个线性无关的新变量表示的数据。
- 新变量是可能的正交变换中信息保存量最大的，方差表示在新变量上信息的大小。
- 可以用主分成近似地表示原始数据，发现数据的基本结构，也可以把数据由少数主成分表示，对数据降维。

# 一、主成分分析

## 2. PCA算法的基本思想

如何确保正交变换以后，新的变量尽可能多的保留原始变量的信息？

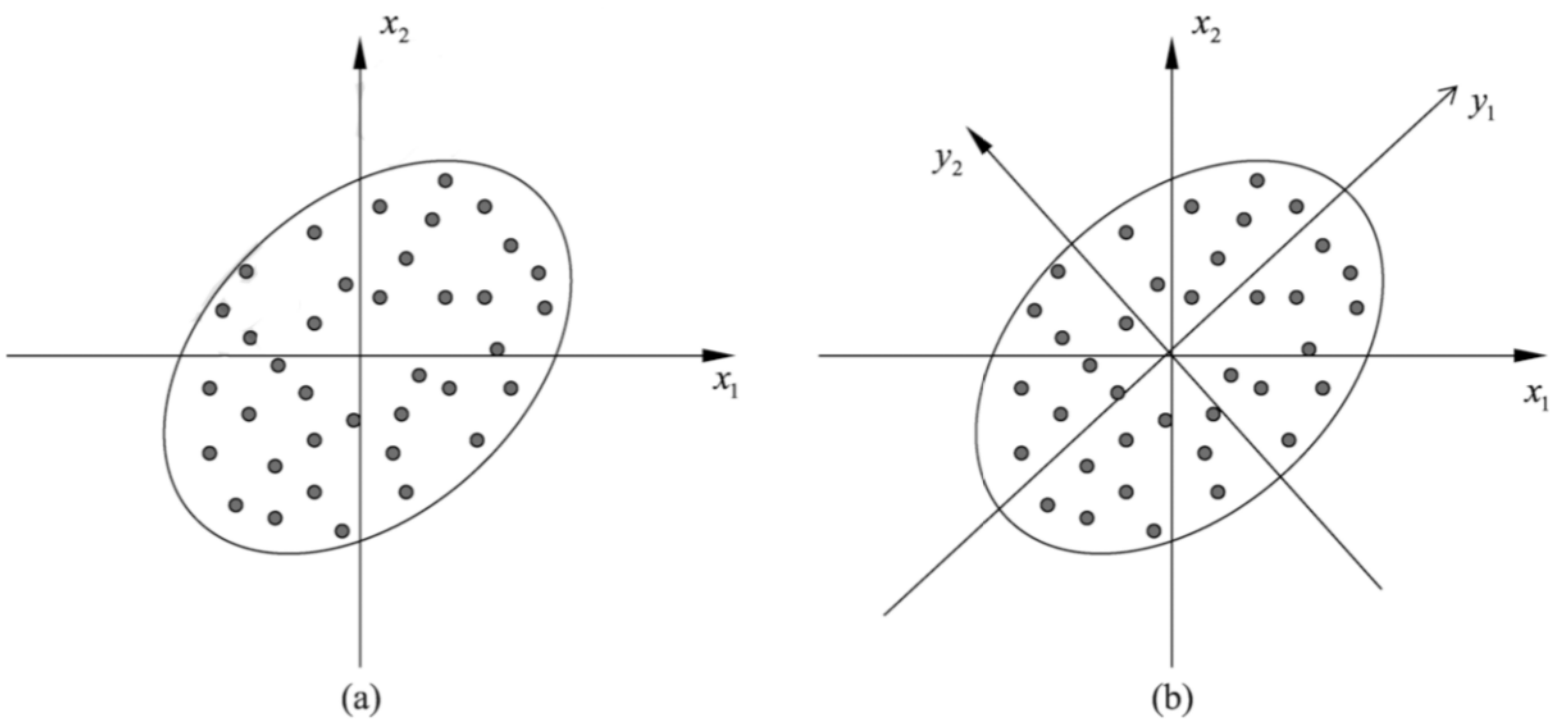
主成分分析：方差和最大的方向。

- **方差**：数据在坐标轴上的坐标值的平方表示相应变量的方差。

# 一、主成分分析

## 2. PCA算法的基本思想

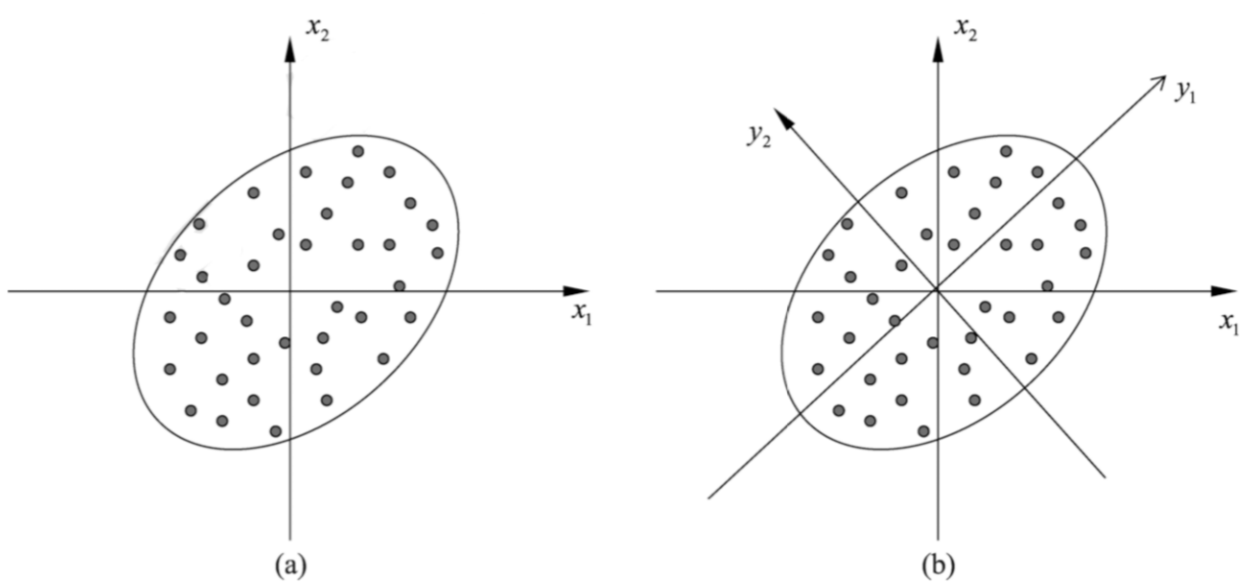
- 主成分分析对数据进行正交变换，对原坐标系进行旋转变换，并将数据在新坐标系表示。



# 一、主成分分析

## 2. PCA算法的基本思想

- 主成分分析选择方差最大的方向（PCA1）作为新坐标系的第一坐标轴，即$y_1$轴。
- 之后选择与第一坐标轴正交，且方差次之的方向（PCA2）作为新坐标系的第二坐标轴，即$y_2$轴。
- ……

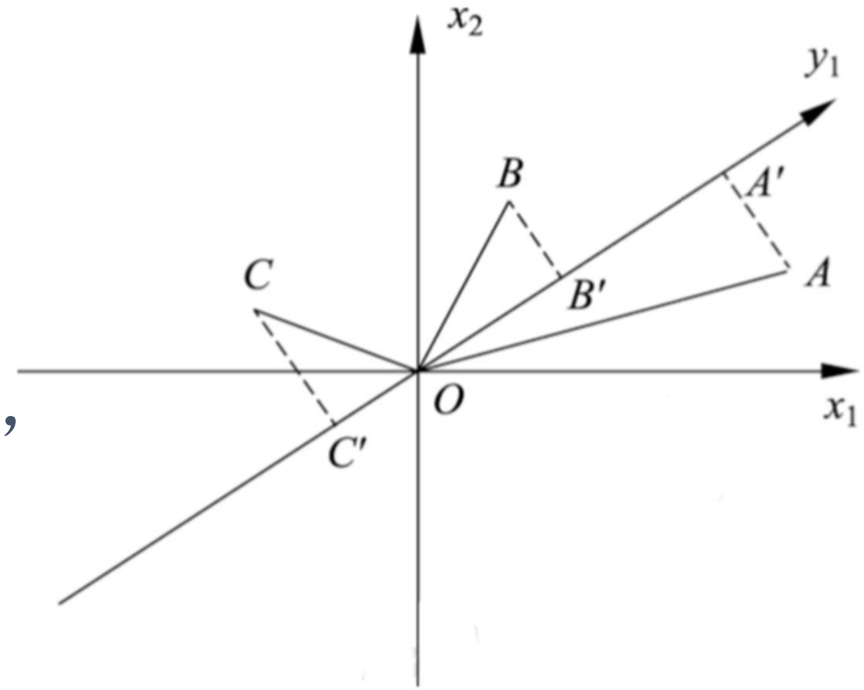


# 一、主成分分析

## 2. PCA算法的基本思想

✓示例

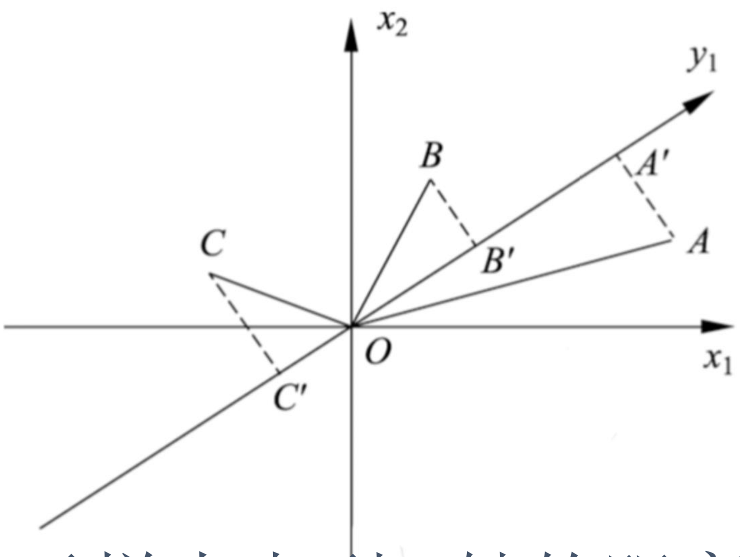
- 假设有两个变量$x_1$和$x_2$，三个样本点A、B、C，样本分布在由$x_1$和$x_2$轴组成的坐标系中。对坐标系进行旋转变换，得到新的坐标轴$y_1$，表示新的变量$y_1$，样本点A、B、C在y1轴上投影，得到$y_1$轴的坐标值A'、B'、C'。



# 一、主成分分析

## 2. PCA算法的基本思想

✓示例




- 坐标值的平方和OA’$^2$ + OB’$^2$ + OC’$^2$表示样本在变量$y_1$上的方差和；
- OA’$^2$ + OB’$^2$ + OC’$^2$最大等价于样本点到$y_1$轴的距离的平方和AA’$^2$ + BB’$^2$ + CC’$^2$最小(最小化点到直线距离平方和的回归，TLS)
- 主成分分析在旋转变换中选取离样本点的距离平方和最小的轴，作为第一主成分。第二主成分等的选取，在保证与已选坐标轴正交的条件下，以此类推。

# 一、主成分分析

## 3. PCA分析的定义

- 给定一个线性变换，如果它满足：

（1）系数向量 $\alpha_i^{\mathrm{T}}$ 是单位向量，即 $\alpha_i^{\mathrm{T}}\alpha_i = 1$，$i = 1, 2, \cdots, m$;

（2）变量 $y_i$ 与 $y_j$ 互不相关，即 $\mathrm{cov}(y_i, y_j) = 0(i \neq j)$;

（3）变量 $y_1$ 是 $\boldsymbol{x}$ 的所有线性变换中方差最大的; $y_2$ 是与 $y_1$ 不相关的 $\boldsymbol{x}$ 的所有线性变换中方差最大的; 一般地，$y_i$ 是与 $y_1, y_2, \cdots, y_{i-1}$ $(i = 1, 2, \cdots, m)$ 都不相关的 $\boldsymbol{x}$ 的所有线性变换中方差最大的; 这时分别称 $y_1, y_2, \cdots, y_m$ 为 $\boldsymbol{x}$ 的第一主成分、第二主成分、$\cdots$、第 $m$ 主成分。

## 3. PCA分析的定义

- 定义中的条件（1）表明线性变换是正交变换，$\alpha_1, \alpha_2, \cdots, \alpha_m$ 是其一组标准正交基

$$\alpha_i^{\mathrm{T}} \alpha_j = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

- 条件（2）（3）给出了一个求主成分的方法：

一般地，$x$的第$k$主成分是 $\alpha_k^{\mathrm{T}} \boldsymbol{x}$，并且 $\operatorname{var}(\alpha_k^{\mathrm{T}} \boldsymbol{x}) = \lambda_k$，这里 $\lambda_k$ 是第$k$个特征值，$\alpha_k$ 是对应的单位特征向量。

# 一、主成分分析

## 4. 如何确定主成分的个数?

- 通常取$k$使得累计方差贡献率达到规定的百分比以上来确定主成分的个数。
- 方差贡献率

$k$ 个主成分 $y_1, y_2, \cdots, y_k$ 对原有变量 $x_i$ 的贡献率定义为 $x_i$ 与 $(y_1, y_2, \cdots, y_k)$ 的相关系数的平方，记作 $\nu_i$

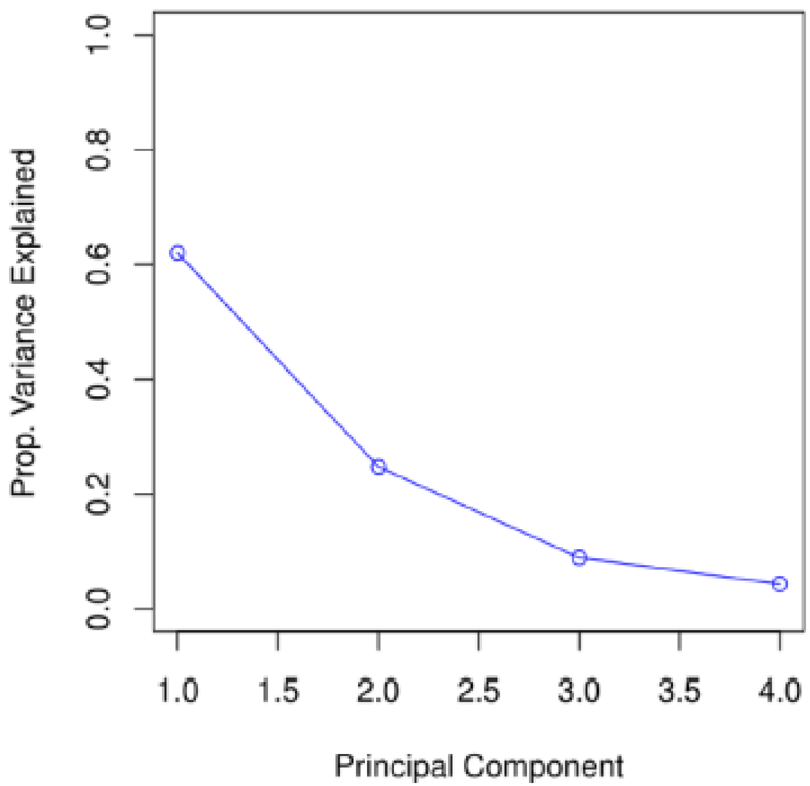
$$\nu_i = \rho^2(x_i, (y_1, y_2, \cdots, y_k))$$

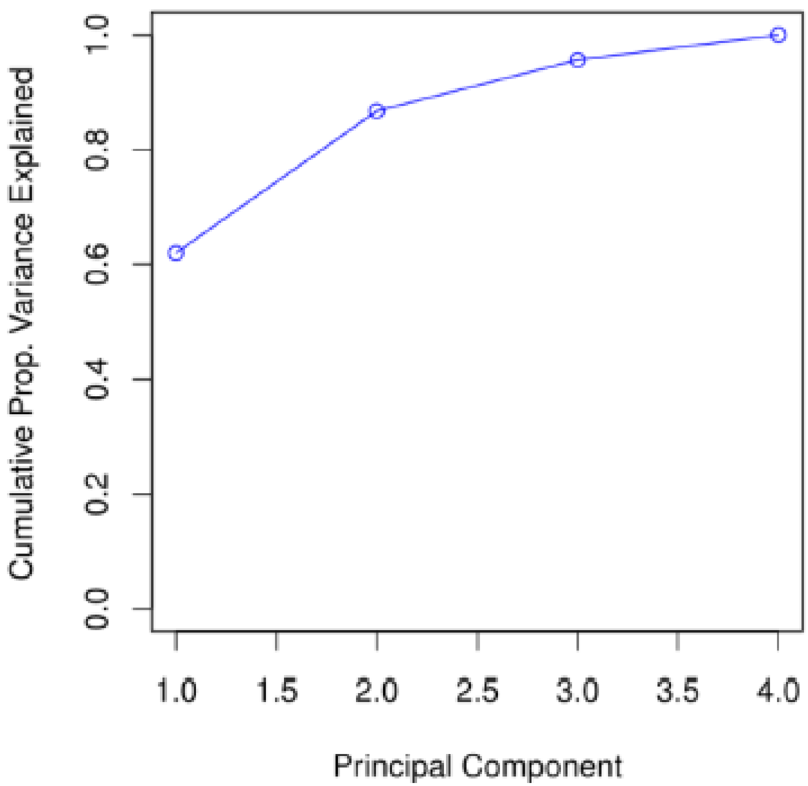
计算公式如下:

$$\nu_i = \rho^2(x_i, (y_1, y_2, \cdots, y_k)) = \sum_{j=1}^{k} \rho^2(x_i, y_j) = \sum_{j=1}^{k} \frac{\lambda_j \alpha_{ij}^2}{\sigma_{ii}}$$

# 一、主成分分析

## 4. 如何确定主成分的个数？



## 5. 例题

假设有$n$个学生参加四门课程的考试，将学生们的考试成绩看作随机 变量的取值，对考试成绩数据进行标准化处理，得到样本相关矩阵$\boldsymbol{R}$

| 课程 | 语文 | 外语 | 数学 | 物理 |
|---|---|---|---|---|
| 语文 | 1 | 0.44 | 0.29 | 0.33 |
| 外语 | 0.44 | 1 | 0.35 | 0.32 |
| 数学 | 0.29 | 0.35 | 1 | 0.60 |
| 物理 | 0.33 | 0.32 | 0.60 | 1 |

试对数据进行主成分分析。

# 一、主成分分析

## 5. 例题

- 设变量 $x_1, x_2, x_3, x_4$ 分别表示语文、外语、数学、物理的成绩。对样本相关矩阵进行特征值分解，得到相关矩阵的特征值，并按大小排序

$$\lambda_1 = 2.17, \quad \lambda_2 = 0.87, \quad \lambda_3 = 0.57, \quad \lambda_4 = 0.39$$

- 这些特征值就是各主成分的方差贡献率。假设要求主成分的累计方差贡献率大于75%，那么只需取前两个主成分即可，即*k*=2，因为

$$\frac{\lambda_1 + \lambda_2}{\sum_{i=1}^{4} \lambda_i} = 0.76$$

# 一、主成分分析

## 5. 例题

- 求出对应于特征值 $\lambda_1, \lambda_2$ 的单位特征向量

| 项目 | $x_1$ | $x_2$ | $x_3$ | $x_4$ | 方差贡献率 |
|---|---|---|---|---|---|
| $y_1$ | 0.460 | 0.476 | 0.523 | 0.537 | 0.543 |
| $y_2$ | 0.574 | 0.486 | −0.476 | −0.456 | 0.218 |

- 由 $y_i = a_i^{\mathrm{T}} \boldsymbol{x}, \quad i = 1, 2, \cdots, k$ 可得第一主成分$y_1$、第二主成分$y_2$

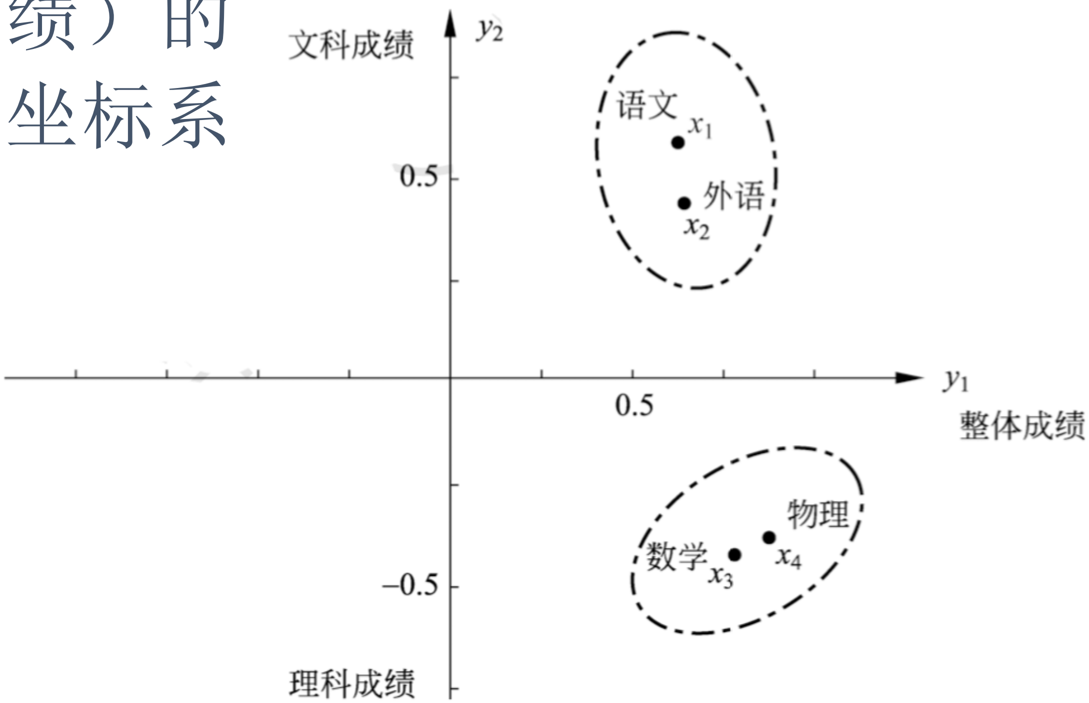
$$y_1 = 0.460x_1 + 0.476x_2 + 0.523x_3 + 0.537x_4$$

$$y_2 = 0.574x_1 + 0.486x_2 - 0.476x_3 - 0.456x_4$$

# 一、主成分分析

## 5. 例题

- 将原变量 $x_1, x_2, x_3, x_4$（语文、外语、数学、物理）和主成分 $y_1, y_2$（整体成绩、文科对理科成绩）的因子载荷量在平面坐标系中表示。



# 一、主成分分析

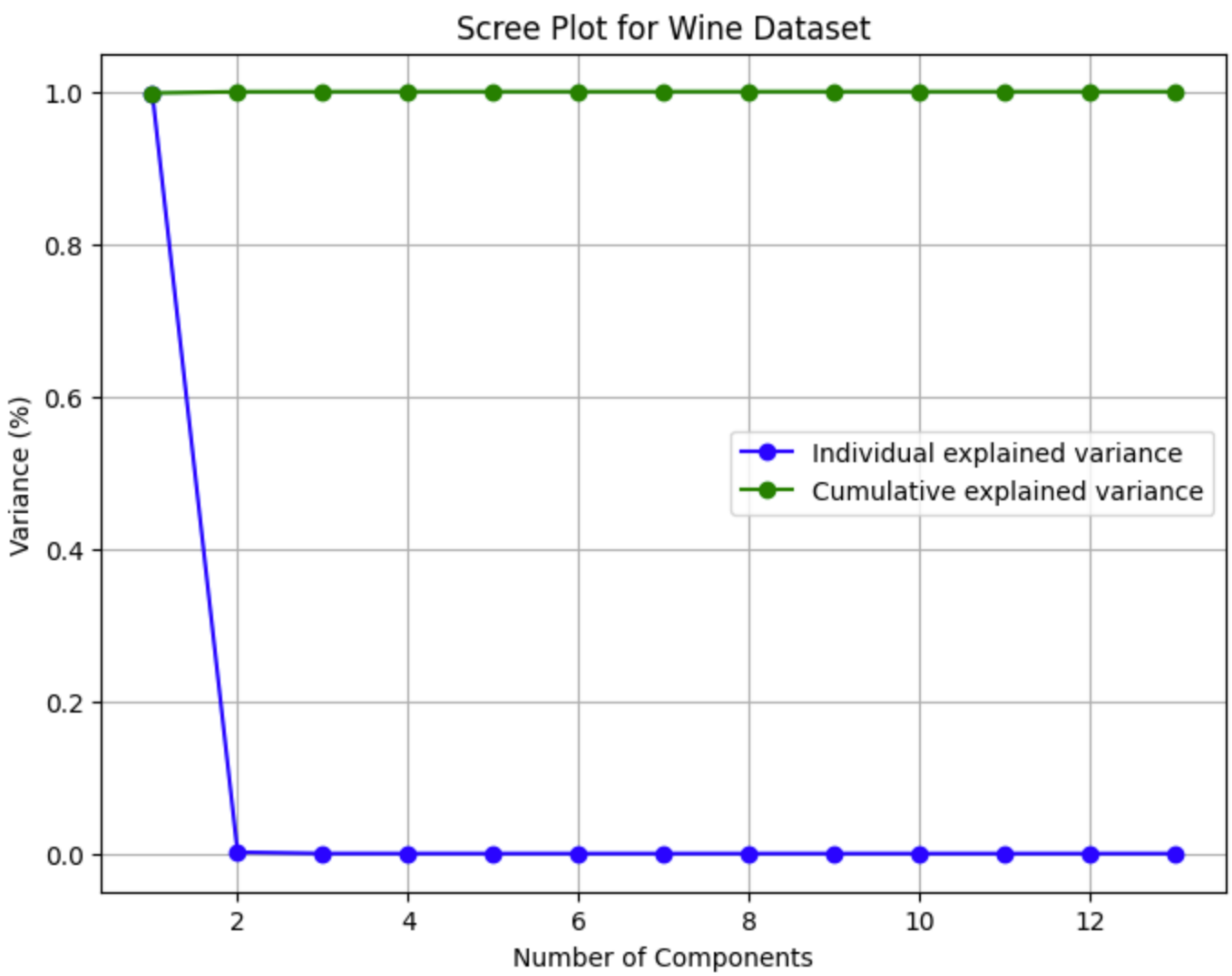
## 6. 示例：基于红酒数据集（wine）实现PCA

```
.. _wine_dataset:

Wine recognition dataset
------------------------

**Data Set Characteristics:**

:Number of Instances: 178
:Number of Attributes: 13 numeric, predictive attributes and the class
:Attribute Information:
    - Alcohol
    - Malic acid
    - Ash
    - Alcalinity of ash
    - Magnesium
    - Total phenols
    - Flavanoids
    - Nonflavanoid phenols
    - Proanthocyanins
    - Color intensity
    - Hue
    - OD280/OD315 of diluted wines
    - Proline
    - class:
        - class_0
        - class_1
        - class_2
```

```
:Summary Statistics:

============================= ==== ===== ======= =====
                                Min   Max   Mean     SD
============================= ==== ===== ======= =====
Alcohol:                      11.0  14.8    13.0   0.8
Malic Acid:                   0.74  5.80    2.34  1.12
Ash:                          1.36  3.23    2.36  0.27
Alcalinity of Ash:            10.6  30.0    19.5   3.3
Magnesium:                    70.0 162.0    99.7  14.3
Total Phenols:                0.98  3.88    2.29  0.63
Flavanoids:                   0.34  5.08    2.03  1.00
Nonflavanoid Phenols:         0.13  0.66    0.36  0.12
Proanthocyanins:              0.41  3.58    1.59  0.57
Colour Intensity:              1.3  13.0     5.1   2.3
Hue:                          0.48  1.71    0.96  0.23
OD280/OD315 of diluted wines: 1.27  4.00    2.61  0.71
...
```

# 一、主成分分析

## 6. 示例



Scree Plot for Wine Dataset
1.0
0.8
0.6
0.4
0.2
0.0
Variance (%)
2
4
6
8
10
12
Number of Components
Individual explained variance
Cumulative explained variance

# 一、主成分分析

## 6. 示例

```
Cumulative variance contributions:
Up to Component 1: 99.81%
Up to Component 2: 99.98%
Up to Component 3: 99.99%
Up to Component 4: 100.00%
```

```
Transformed Feature Matrix:
          PC1        PC2
0  318.562979  21.492131
1  303.097420  -5.364718
2  438.061133  -6.537309
3  733.240139   0.192729
4  -11.571428  18.489995
```

```
PCA Loadings:
      alcohol  malic_acid       ash  alcalinity_of_ash  magnesium  \
PC1  0.001659   -0.000681  0.000195          -0.004671   0.017868
PC2  0.001203    0.002155  0.004594           0.026450   0.999344

     total_phenols  flavanoids  nonflavanoid_phenols  proanthocyanins  \
PC1       0.000990    0.001567             -0.000123         0.000601
PC2       0.000878   -0.000052             -0.001354         0.005004

     color_intensity       hue  od280/od315_of_diluted_wines   proline
PC1         0.002327  0.000171                      0.000705  0.999823
PC2         0.015100 -0.000763                     -0.003495 -0.017774
```

# 提　纲

一　主成分分析（PCA）

二　线性判别分析（LDA）

# 二、线性判别分析

## 1. 判别分析概述

- 判别分析（Discriminant Analysis，DA）是根据观测到的某些指标对所研究的对象进行分类的一种多元统计分析方法。

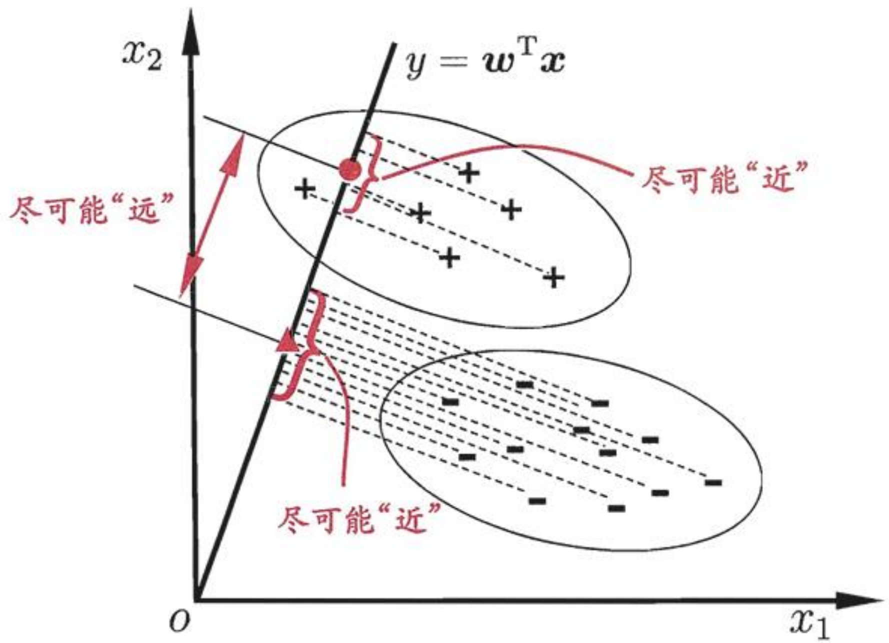
- 判别分析技术由R.A. Fisher于1936年提出。

- 线性判别分析LDA(Linear Discriminant Analysis)又称为Fisher线性判别，是一种有监督的降维技术，也就是说它的数据集的每个样本都是有类别输出的，这与PCA（无监督学习）不同。

- LDA在模式识别领域（比如人脸识别、地物识别等图形图像识别领域）中有非常广泛的应用。

## 2. LDA的算法思想

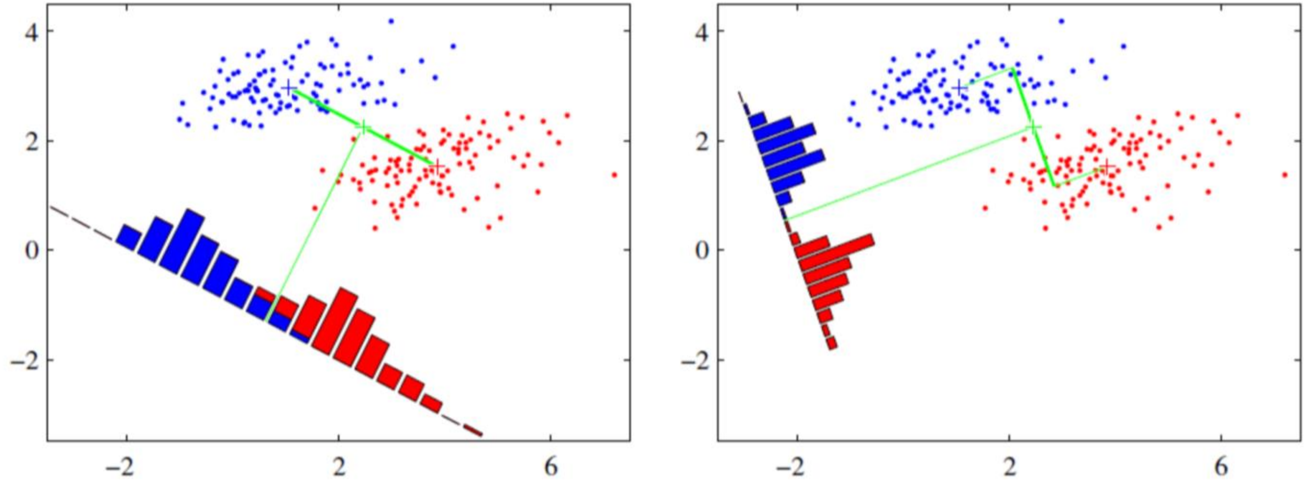
- LDA的思想是：最大化类间均值，最小化类内方差。意即将数据投影在低维度上，并且投影后同种类别数据的投影点尽可能的接近，不同类别数据的投影点的中心点尽可能的远。



# 二、线性判别分析

## 2. LDA的算法思想

- 假设我们有特征数是二维的两类数据，希望将这些数据投影到一维的一条直线，让每一类别数据的投影点尽可能的接近，而两类数据中心之间的距离尽可能的大。




- 上图的两种投影结果中哪一种更好？

## 3. 瑞利商与广义瑞利商

- 瑞利商（Rayleigh quotient）

$$R(A,x)=\frac{x^H A x}{x^H x}$$

- 其中，$x$ 是非零向量，$A$为 $n \times n$ 的厄米特矩阵。
- 厄米特矩阵（Hermitan矩阵）是满足共轭转置矩阵和自己相等的矩阵。

- 瑞利商的最大值等于矩阵 $A$ 最大的特征值，而最小值等于矩阵 $A$ 的最小的特征值。即满足

$$\lambda_{\min} \le \frac{x^H A x}{x^H x} \le \lambda_{\max}$$

当向量 $x$ 是标准正交基，即满足 $x^H x=1$ 时，瑞利商退化为 $R(A,x)=x^H A x$ 。

## 3. 瑞利商与广义瑞利商

- 广义瑞利商（Genralized Rayleigh quotient）

$$R(A,B,x)=\frac{x^H Ax}{x^H Bx}$$

• 其中，B 为正定矩阵。

- 若令 $x=B^{-1/2}x'$，则

$$x^H Ax = x'^H B^{-1/2} AB^{-1/2} x'$$

$$x^H Bx = x'^H \left(B^{-1/2}\right)^H BB^{-1/2} x' = x'^H B^{-1/2} BB^{-1/2} x' = x'^H x'$$

$$R(A,B,x') = \frac{x'^H B^{-1/2} AB^{-1/2} x'}{x'^H x'}$$

• 广义瑞利商的最大值和最小值分别是矩阵 $B^{-1/2}AB^{-1/2}$ 的最大特征值和最小特征值。

# 二、线性判别分析

## 4. 目标函数

欲使同类样例的投影点尽可能接近，可以让同类样例投影点的协方差尽可能小，即 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_0\boldsymbol{w}+\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_1\boldsymbol{w}$ 尽可能小；而欲使异类样例的投影点尽可能远离，可以让类中心之间的距离尽可能大，即 $\|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_0-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_1\|_2^2$ 尽可能大. 同时考虑二者，则可得到欲最大化的目标

$$
\begin{aligned}
J &= \frac{\|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_0-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_1\|_2^2}{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_0\boldsymbol{w}+\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_1\boldsymbol{w}} \\
&= \frac{\boldsymbol{w}^{\mathrm{T}}(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)^{\mathrm{T}}\boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}}(\boldsymbol{\Sigma}_0+\boldsymbol{\Sigma}_1)\boldsymbol{w}} .
\end{aligned}
$$

定义“类内散度矩阵”(within-class scatter matrix)

$$
\begin{aligned}
\mathbf{S}_w &= \boldsymbol{\Sigma}_0+\boldsymbol{\Sigma}_1 \\
&= \sum_{\boldsymbol{x}\in X_0}(\boldsymbol{x}-\boldsymbol{\mu}_0)(\boldsymbol{x}-\boldsymbol{\mu}_0)^{\mathrm{T}}+\sum_{\boldsymbol{x}\in X_1}(\boldsymbol{x}-\boldsymbol{\mu}_1)(\boldsymbol{x}-\boldsymbol{\mu}_1)^{\mathrm{T}}
\end{aligned}
$$

以及“类间散度矩阵”(between-class scatter matrix)

$$
\mathbf{S}_b=(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)^{\mathrm{T}} ,
$$

# 二、线性判别分析

## 4. 目标函数

- 基于广义瑞利商的定义，目标函数$J$可以重写为：

$$J = \frac{\boldsymbol{w}^{\mathrm{T}} \mathbf{S}_b \boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}} \mathbf{S}_w \boldsymbol{w}}$$

- 即类内散度与类间散度的广义瑞利商。
- LDA求其最大化目标值。

## 5. 示例：基于鸢尾花数据集实现LDA

```
**Data Set Characteristics:**

:Number of Instances: 150 (50 in each of three classes)
:Number of Attributes: 4 numeric, predictive attributes and the class
:Attribute Information:
    - sepal length in cm
    - sepal width in cm
    - petal length in cm
    - petal width in cm
    - class:
            - Iris-Setosa
            - Iris-Versicolour
            - Iris-Virginica

:Summary Statistics:

============== ==== ==== ======= ===== ====================
                Min  Max   Mean    SD   Class Correlation
============== ==== ==== ======= ===== ====================
sepal length:   4.3  7.9   5.84   0.83    0.7826
sepal width:    2.0  4.4   3.05   0.43   -0.4194
petal length:   1.0  6.9   3.76   1.76    0.9490  (high!)
petal width:    0.1  2.5   1.20   0.76    0.9565  (high!)
============== ==== ==== ======= ===== ====================

:Missing Attribute Values: None
:Class Distribution: 33.3% for each of 3 classes.
:Creator: R.A. Fisher
:Donor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)
:Date: July, 1988
```

# 二、线性判别分析

## 5. 示例

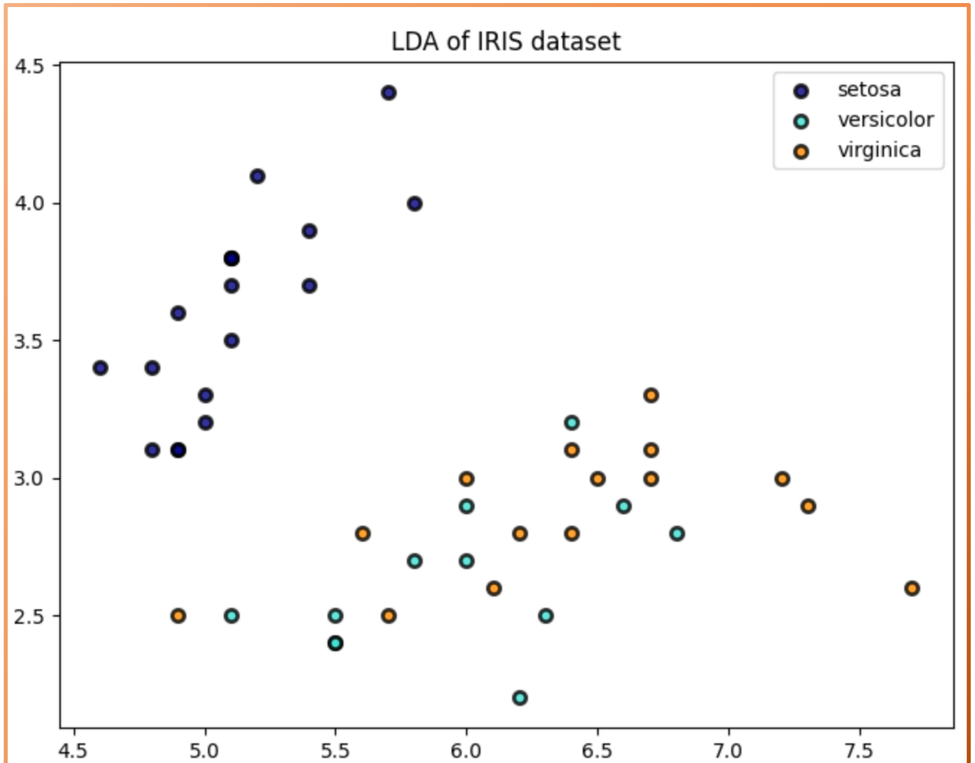
```
# 加载鸢尾花数据集
iris = load_iris()
X = iris.data
y = iris.target
target_names = iris.target_names#为作图做准备
#print(iris.DESCR)

# 划分训练集和测试集
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=2024)

# 初始化LDA分类器,n_components为降维后的维度，设置为2以便可视化
lda = LDA(n_components=2)

# 训练LDA分类器
lda.fit(X_train, y_train)

# 使用LDA分类器进行预测
y_pred = lda.predict(X_test)

# 输出分类报告和混淆矩阵
print(classification_report(y_test, y_pred))
print(confusion_matrix(y_test, y_pred))
```

```
[[18  0  0]
 [ 0 11  1]
 [ 0  0 15]]
```

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 1.00 | 1.00 | 1.00 | 18 |
| 1 | 1.00 | 0.92 | 0.96 | 12 |
| 2 | 0.94 | 1.00 | 0.97 | 15 |

# 二、线性判别分析

## 5. 示例




```
X_new = [[5.8, 4, 1.2, 0.2],
         [4.9, 3.6, 1.4, 0.1],
         [6, 2.9, 4.5, 1.5]]
y_pred_new = lda.predict(X_new)
print(y_pred_new)
```

✓ 0.0s

```
[0 0 1]
```

# PCA 与 LDA

PCA与LDA都属于常用的降维技术，且仅适用于符合高斯分布的数据集，但二者也存在显著的区别：

**相同点：**

- 两者均可用于数据降维。
- 两者在降维时均使用了矩阵特征分解的思想。
- 两者都假设数据符合正态分布。

**不同点：**

- LDA是有监督的降维方法，而PCA是无监督降维方法；LDA除了用于降维，还可以用于分类。
- LDA选择分类性能最好的投影方向，而PCA选择样本点投影具有最大方差的方向。

# LDA实验

- **数据：教材272页，习题6。**
- **实验要求：**
  - 根据给定数据集，采用线性判别法对指定用户的信用好坏进行判别；
  - 包含典型的训练和测试过程；
  - 对实验结果进行分析。

# 思考题

1. 简述PCA算法的思想及其主要步骤。
2. 简述LDA的基本思想。
3. 举例说明PCA与LDA的异同。