

The background features two white wireframe hands, one in the top right and one in the bottom left, reaching towards the center. The background is a solid blue color with faint, concentric circular patterns and small white star-like sparkles.

第9讲 特征选择与提取2

——特征提取

周文晖

杭州电子科技大学



- ✓ 特征选择与提取基本概念
基本概念, 特征形成, 特征评价准则...
- ✓ 类别可分性测度
基于距离、概率分布、熵函数的可分判据 ...
- ✓ 特征选择的最优搜索方法
穷举, 分支定界, 遗传算法 ...
- ✓ **特征提取方法**
基于类别可分离性判据的特征提取 ...
- ✓ 主成分分析
PCA基本原理 ...

特征提取方法

特征选择：从 D 个特征中选出 $d (< D)$ 个

特征提取：把 D 个/维特征变换为 $d (< D)$ 个/维新特征

这里只讨论
特征提取

特征提取的目的：

对某类模式： 压缩模式向量的维数，找到可以表征该单一类别的特征。

对多类分类： 压缩维数；找到可以区分多个类别的特征。
保留类别间的鉴别信息，突出可分性。

特征提取操作方法：

变换 $\mathbf{x}^* = \mathbf{W}(\mathbf{x})$ ，其中 $\mathbf{x} \in \mathbb{R}^D$ 是 D 个/维原始特征，变换后生成 d 个/维新特征 $\mathbf{x}^* \in \mathbb{R}^d$ 。

线性变换 $\mathbf{x}^* = \mathbf{W}^T \mathbf{x}$ $\mathbf{W}: D \times d$ 维 ---- 通常， $D > d$ ， “特征压缩”， “特征变换”

这里只讨论
线性变换

注意：维数压缩后，在新的 d 维空间里各模式类之间的分布规律应至少保持不变或更优化。

对某类模式的特征提取——基于类内散布矩阵的单类模式特征提取

对某类模式的特征提取目的：压缩模式向量的维数，找到可以有效表征该单一类别的特征。

设 ω_i 类模式的均值向量为 $\mathbf{M}$ ，类内散布矩阵(协方差矩阵)为 $\mathbf{C}$ ：

$$\mathbf{M} = E\{\mathbf{X}\} \quad \mathbf{C} = E\{(\mathbf{X} - \mathbf{M})(\mathbf{X} - \mathbf{M})^T\}$$

式中， $\mathbf{X}$ 为 D 维向量， $\mathbf{C}$ 为 $D \times D$ 的实对称矩阵。

设矩阵 $\mathbf{C}$ 的 D 个特征值分别为 $\lambda_1, \lambda_2, \dots, \lambda_D$ 。任一特征值是满足 $|\lambda \mathbf{I} - \mathbf{C}| = 0$ 的一个解。

假定 D 个特征值对应的 D 个特征向量为 $\mathbf{u}_k$, $k = 1, 2, \dots, D$ 。则 $\mathbf{u}_k$ 是满足 $\mathbf{C}\mathbf{u}_k = \lambda_k \mathbf{u}_k$ 的一个非零解。

$\mathbf{u}_k$ 是 D 维向量，可表示为 $\mathbf{u}_k = [u_{k1}, u_{k2}, \dots, u_{kD}]^T$ 。

若 $\mathbf{u}_k$ 为归一化特征向量，根据实对称矩阵的性质，有

$$\mathbf{u}_i^T \mathbf{u}_j = \begin{cases} 1, & j = i \\ 0, & j \neq i \end{cases} \quad \text{—— } D \text{ 个特征向量相互正交。}$$

对某类模式的特征提取——基于类内散布矩阵的单类模式特征提取

利用 W 对 ω_i 类的样本 X 进行变换, 得 $X^* = WX$ 。式中, X 和 X^* 都是 D 维向量。

若选 D 个归一化特征向量作为 W 的行, 则 W 为归一化正交矩阵:

$$W = \begin{bmatrix} u_1^T \\ u_2^T \\ \dots \\ u_D^T \end{bmatrix} \quad W^T = [u_1 \ u_2 \ \dots \ u_D] \quad WW^T = I$$

考察变换前后的分布规律: 均值向量 M^* 、类内距离 $\overline{D^2}$ 和协方差矩阵 C^* 的变化。

(1) 均值向量 $M^* = E\{X^*\} = E\{WX\} = WE\{X\} = WM$

(2) 变换后的类内距离

$$\begin{aligned} \overline{D^2} &= E\{\|X_i^* - X_j^*\|^2\} = E\{(X_i^* - X_j^*)^T (X_i^* - X_j^*)\} = E\{(WX_i - WX_j)^T (WX_i - WX_j)\} \\ &= E\{(X_i - X_j)^T W^T W (X_i - X_j)\} = E\{(X_i - X_j)^T (X_i - X_j)\} = E\{\|X_i - X_j\|^2\} \end{aligned} \quad \text{变换后: 类内距离保持不变。}$$

对某类模式的特征提取——基于类内散布矩阵的单类模式特征提取

$$(3) \quad C^* = E\{(X^* - M^*)(X^* - M^*)^T\} = E\{(WX - WM)(WX - WM)^T\} = WE\{(X - M)(X - M)^T\}W^T = WCW^T$$

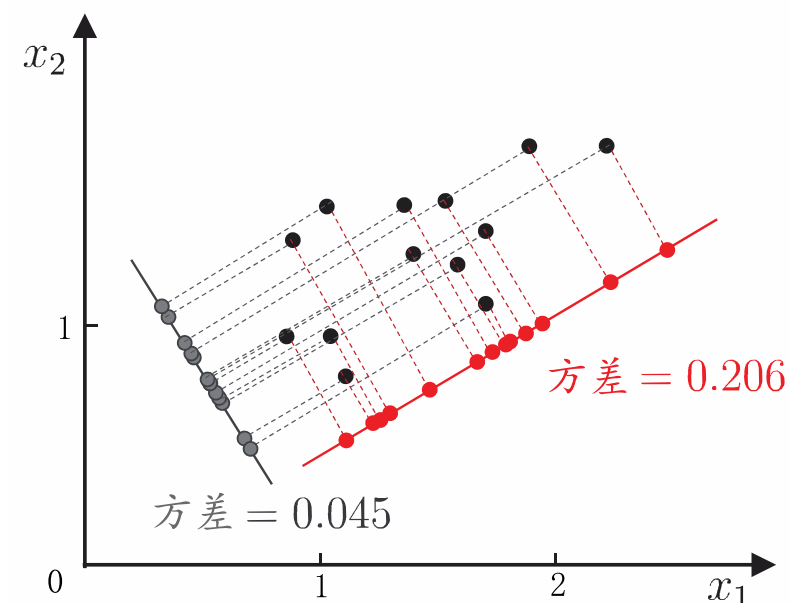
$$= \begin{bmatrix} u_1^T \\ u_2^T \\ \vdots \\ u_D^T \end{bmatrix} C \begin{bmatrix} u_1 & u_2 & \dots & u_D \end{bmatrix} = \begin{bmatrix} u_1^T \\ u_2^T \\ \vdots \\ u_D^T \end{bmatrix} \begin{bmatrix} \lambda_1 u_1 & \lambda_2 u_2 & \dots & \lambda_D u_D \end{bmatrix} = \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \dots & \\ 0 & & & \lambda_D \end{bmatrix}$$

$$Cu_k = \lambda_k u_k$$

$$u_i^T u_j = \begin{cases} 1, & j = i \\ 0, & j \neq i \end{cases}$$

变换后：协方差矩阵为对角阵，说明 X^* 的各分量不相关 —— 便于特征的取舍；

X^* 的第 k 个分量的方差等于未变换时 C 的特征值 λ_k 。



特征取舍的依据：选择方差小的特征分量，方差越小说明在该特征分量方向上投影点越紧密。

根据以上特点得到构造变换矩阵的方法： **目标：**构造一变换矩阵，可以将 D 维向量 X 变换成 d 维 ($d < D$)。

对某类模式的特征提取——基于类内散布矩阵的单类模式特征提取

思路:

将变换前的 C 的 D 个特征值 **从小到大** 排队 $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_d \leq \dots \leq \lambda_D$

不同于基于类别可分性判据的特征提取方法，特征值排列顺序相反。

↓ 前 d 个特征值

选择前 d 个特征值 $\lambda_1, \lambda_2, \dots, \lambda_d$ 所对应的特征向量 $u_1, u_2, \dots, u_d$ 作为变换矩阵 W 的行 ($D \times d$)

↓
对 X 进行 W 变换

$$\begin{bmatrix} w_{11} & \dots & w_{1D} \\ \dots & \dots & \dots \\ w_{d1} & \dots & w_{dD} \end{bmatrix} \begin{bmatrix} x_1 \\ \dots \\ x_D \end{bmatrix} \rightarrow \begin{bmatrix} x_1^* \\ \dots \\ x_d^* \end{bmatrix}$$

优点: 压缩了维数;

选择特征值 (方差) 更小的特征向量方向投影, 投影后的样本点更密集

——相当**去掉了方差大**的特征分量。

对某类模式的特征提取——基于类内散布矩阵的单类模式特征提取（例）

例 假定 ω_i 类的样本集为 $\{X\} = \{X_1, X_2, X_3\}$ ，三个样本分别为 $X_1 = [1, 1]^T$ ， $X_2 = [2, 2]^T$ ， $X_3 = [3, 1]^T$

用类内散布矩阵进行特征提取，将二维样本变换成一维样本。

解：1) 求样本均值向量和协方差矩阵。 $M = \frac{1}{3} \sum_{i=1}^3 X_i = [2, 1.3]^T$ $C = \frac{1}{3} \sum_{i=1}^3 X_i X_i^T - M M^T = \begin{bmatrix} 0.7 & 0.1 \\ 0.1 & 0.3 \end{bmatrix}$

2) 根据 $|\lambda I - C| = 0$ 求 C 的特征值，并进行选择。

由 $\begin{vmatrix} \lambda - 0.7 & -0.1 \\ -0.1 & \lambda - 0.3 \end{vmatrix} = 0$ ，得 $\lambda_1 = 0.2765$ $\lambda_2 = 0.7236$ $\because \lambda_1 < \lambda_2$ $\therefore$ 选 λ_1

3) 计算 λ_1 对应的特征向量 u_1 。由方程 $C u_1 = \lambda_1 u_1$ 得 $u_1 = [0.5, -2.1]^T$

归一化处理有 $u_1 = \frac{1}{\sqrt{0.5^2 + 2.1^2}} [0.5, -2.1]^T = \frac{1}{\sqrt{4.66}} [0.5, -2.1]^T$

由归一化特征向量 u_1 构成变换矩阵 W ： $W = \frac{1}{\sqrt{4.66}} [0.5, -2.1]$

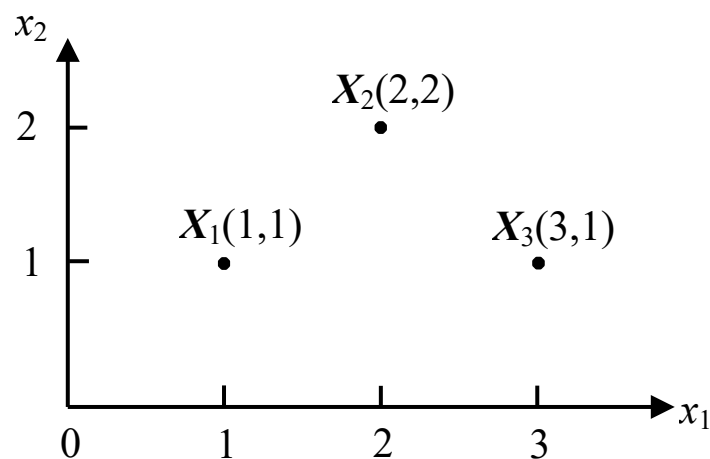
对某类模式的特征提取——基于类内散布矩阵的单类模式特征提取（例）

例 假定 ω_i 类的样本集为 $\{X\} = \{X_1, X_2, X_3\}$ ，三个样本分别为 $X_1 = [1, 1]^T$ ， $X_2 = [2, 2]^T$ ， $X_3 = [3, 1]^T$

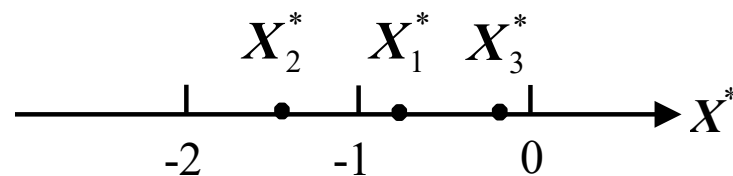
用类内散布矩阵进行特征提取，将二维样本变换成一维样本。

4) 利用 W 对 X_1, X_2, X_3 进行变换。

$$X_1^* = WX_1 = -0.74 \quad X_2^* = WX_2 = -1.48 \quad X_3^* = WX_3 = -0.28$$



(a) 变换前



(b) 变换后

对多类分类的特征提取——基于类别可分性判据的特征提取

对多类分类的特征提取思想：提取的特征不仅要能够描述每个类别的特性，还要能够区分不同的类别，突出可分性。

采用类别可分性判据作为衡量新特征的准则。

特征提取问题转换为求最优的 $\mathbf{W}^*$ ，即： $\mathbf{W}^* = \arg \max_{\mathbf{W}} J(\mathbf{W}^T \mathbf{x})$ ，其中 $J(\cdot)$ 为类别可分性判据。

回顾：特征选择中常用的基于类内类间距离的可分性判据

多类类间散布矩阵： $\mathbf{S}_b = \sum_{i=1}^c P(\omega_i)(\mathbf{M}_i - \mathbf{M}_0)(\mathbf{M}_i - \mathbf{M}_0)^T$

多类类内散布矩阵： $\mathbf{S}_w = \sum_{i=1}^c P(\omega_i)E\{(\mathbf{X} - \mathbf{M}_i)(\mathbf{X} - \mathbf{M}_i)^T\}$

基于类内类间距离的可分性判据：

$$J_1 = \text{tr}(\mathbf{S}_w + \mathbf{S}_b)$$

$$J_3 = \ln \frac{|\mathbf{S}_b|}{|\mathbf{S}_w|}$$

$$J_4 = \frac{\text{tr} \mathbf{S}_b}{\text{tr} \mathbf{S}_w}$$

$$J_5 = \frac{|\mathbf{S}_b - \mathbf{S}_w|}{|\mathbf{S}_w|}$$

$$J_2 = \text{tr}(\mathbf{S}_w^{-1} \mathbf{S}_b)$$

$$\mathbf{S}_b^* = \mathbf{W}^T \mathbf{S}_b \mathbf{W}$$

$$\mathbf{S}_w^* = \mathbf{W}^T \mathbf{S}_w \mathbf{W}$$



准则函数（变换后的可分离判据）：

$$J_1 = \text{tr}(\mathbf{W}^T (\mathbf{S}_w + \mathbf{S}_b) \mathbf{W})$$

$$J_2 = \text{tr}[(\mathbf{W}^T \mathbf{S}_w \mathbf{W})^{-1} (\mathbf{W}^T \mathbf{S}_b \mathbf{W})]$$

$$J_3 = \ln \frac{|\mathbf{W}^T \mathbf{S}_b \mathbf{W}|}{|\mathbf{W}^T \mathbf{S}_w \mathbf{W}|}$$

$$J_4 = \frac{\text{tr}(\mathbf{W}^T \mathbf{S}_b \mathbf{W})}{\text{tr}(\mathbf{W}^T \mathbf{S}_w \mathbf{W})}$$

$$J_5 = \frac{|\mathbf{W}^T \Sigma \mathbf{W}|}{|\mathbf{W}^T \mathbf{S}_w \mathbf{W}|}, \Sigma = \mathbf{S}_w + \mathbf{S}_b$$

对多类分类的特征提取——基于类别可分性判据的特征提取

思想：类别可分性判据中要求 $J(\cdot)$ **越大可分性越好**。

不同于类内散布矩阵，类内散布矩阵是越小越好。

虽然 $J_1 \sim J_5$ 这些准则函数形式不同，但得到的最优变换矩阵相同，接下来以 J_2 为例（具体推导略）。

设总体散布矩阵 $\mathbf{S}_w^{-1}\mathbf{S}_b$ 的特征值（本征值）从大到小排列为： $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d \geq \dots \geq \lambda_D$

选择前 d 个特征值 $\lambda_1, \lambda_2, \dots, \lambda_d$ 所对应的特征向量 $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_d$ 作为变换矩阵 $\mathbf{W}$ 的行 ($D \times d$)，即 $\mathbf{W} = \begin{bmatrix} \mathbf{u}_1 \\ \dots \\ \mathbf{u}_d \end{bmatrix}_{D \times d}$

对 $\mathbf{X}$ 进行 $\mathbf{W}$ 变换

$$\begin{bmatrix} w_{11} & \dots & w_{1D} \\ \dots & \dots & \dots \\ w_{d1} & \dots & w_{dD} \end{bmatrix} \begin{bmatrix} x_1 \\ \dots \\ x_D \end{bmatrix} \rightarrow \begin{bmatrix} x_1^* \\ \dots \\ x_d^* \end{bmatrix}$$



- ✓ 特征选择与提取基本概念
基本概念，特征形成，特征评价准则...
- ✓ 类别可分性测度
基于距离、概率分布、熵函数的可分判据 ...
- ✓ 特征选择的最优搜索方法
穷举，分支定界，遗传算法 ...
- ✓ 特征提取方法
基于类别可分离性判据的特征提取 ...
- ✓ **主成分分析**
PCA基本原理 ...

PCA（Principal Component Analysis）主成分分析

例：每个西瓜都有重量、颜色、形状、纹路、气味五种属性

序号	西瓜A	西瓜B	西瓜C	西瓜D	西瓜E
重量	1	3	2	8	8
颜色	1	2	2	8	9
形状	6	7	6	6	7
纹路	3	2	3	9	9
气味	1	2	1	1	1

降维



观察西瓜的五种属性，气味和形状是完全可以忽略的，因为所有西瓜的这两个属性是差不多的。

序号	西瓜A	西瓜B	西瓜C	西瓜D	西瓜E
重量	1	3	2	8	8
颜色	1	2	2	8	9
纹路	3	2	3	9	9

注：这里的**离散属性**已经做了某种数字化处理，如颜色数值越小表示越接近浅黄色，颜色数值越大越接近深绿色，形状数值越接近1代表越接近球形等。

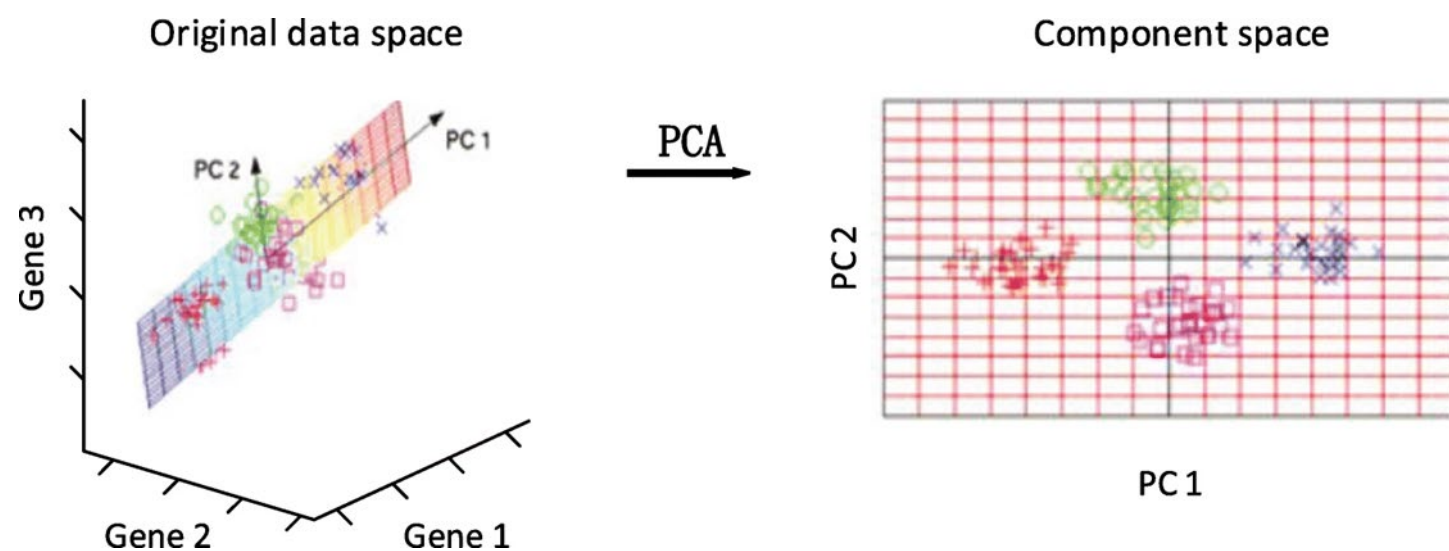
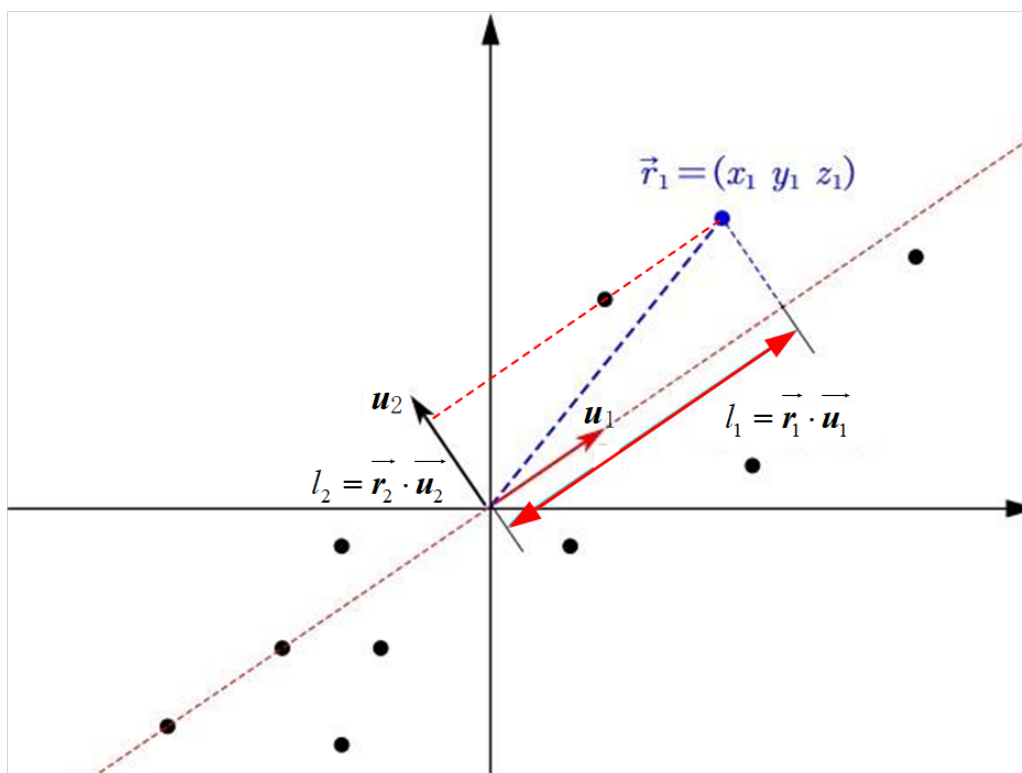
如果数据多且杂乱，或是高维数据，应如何处理？

PCA基本思想：把高维特征通过**线性变换**，投影到低维空间中，将包含信息量大的主成分（**离散度大**）保留下来，忽略对数据描述不重要的次要信息。

PCA (Principal Component Analysis) 主成分分析

PCA的基本原理：线性映射（基变换）， 将原始数据变换为一组各维度**线性无关**的表示。

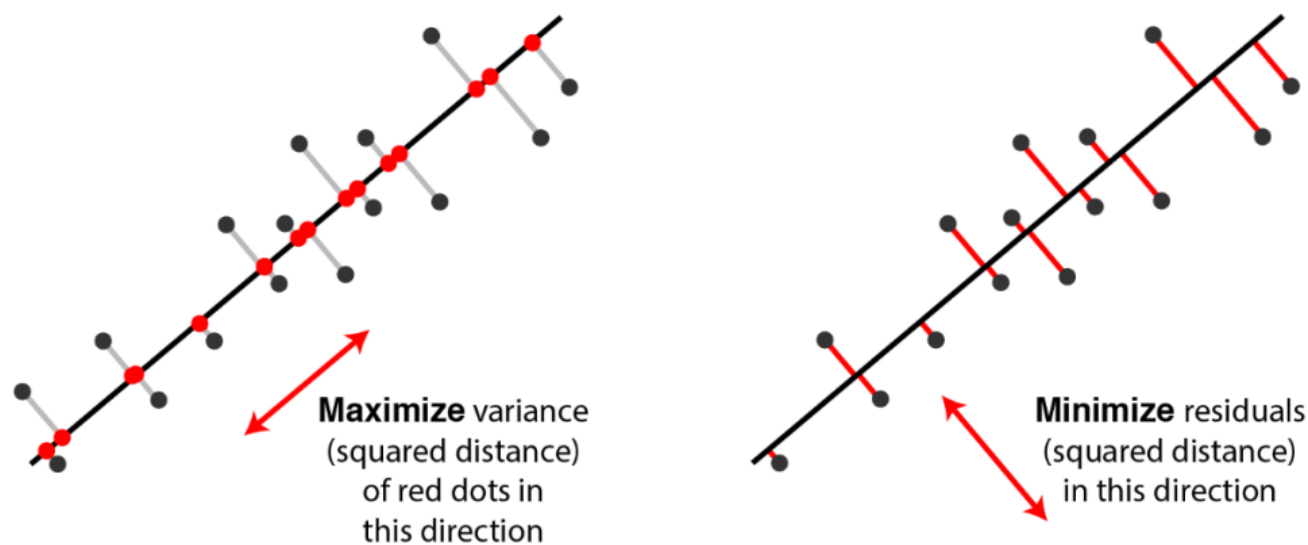
成为一组基的唯一要求是：**线性无关**（非正交基也可以，但由于正交基有较好的性质，所以通常使用正交基）



PCA (Principal Component Analysis) 主成分分析

PCA的基本原理：线性映射（基变换）。

如何用一个超平面对原始数据进行恰当的表达？



若存在这样的超平面，则应具有两个性质：

- 最大可分析性：
样本点在这个超平面的投影尽可能分开
- 最近重构性：
所有样本点到超平面的距离最近

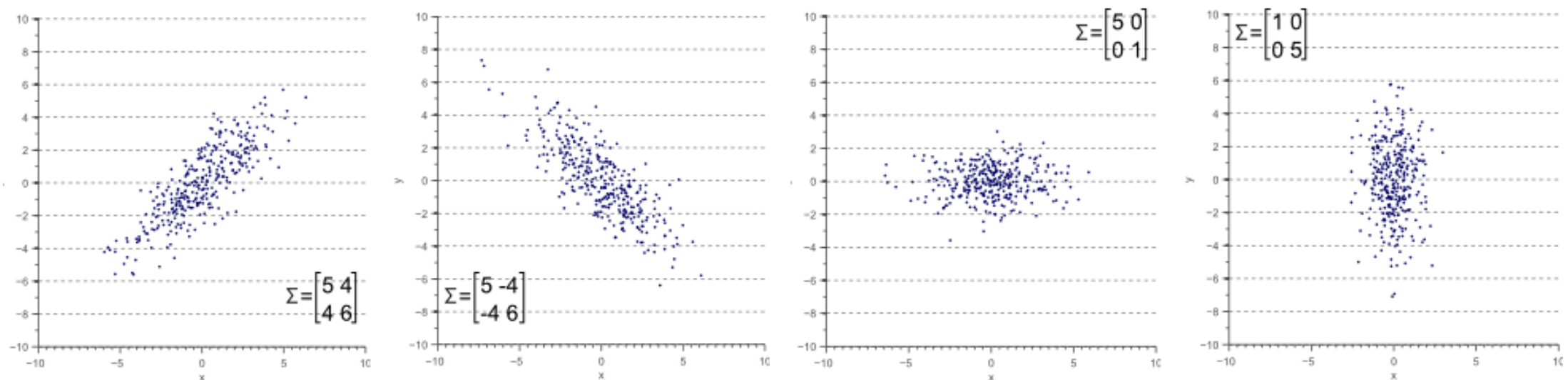
以上两种性质便是主成分分析的两种等价的推导，即**PCA最大方差理论**和**PCA最小平方误差理论**。

二维特征降到一维的示意图，若是高维特征则对应于超平面

PCA主成分分析——最大方差理论

PCA的基本手段：线性映射（基变换），将原始数据变换为一组各维度**线性无关**的表示。

协方差矩阵定义了数据的离散程度（方差）和走向（orientation）（协方差）。



协方差矩阵的**特征向量**代表了一组正交基；

对应的**特征值**代表了原始数据在各特征向量上的方差（离散度），特征值越大，方差（离散度）越大。

因此，可采用具有最大离散度的特征向量及其特征值，表示原始数据的主要成分。

PCA主成分分析

PCA最大方差理论：数学表示

原特征： $x_1, \dots, x_p$

新特征： $\xi_i, i = 1, \dots, p$

$$\xi_i = \sum_{j=1}^p \alpha_{ij} x_j = \boldsymbol{\alpha}_i^T \mathbf{x}$$

$$\boldsymbol{\xi} = \mathbf{A}^T \mathbf{x}$$

尺度： $\boldsymbol{\alpha}_i^T \boldsymbol{\alpha}_i = 1$

考虑第一个新特征 ξ_1

$$\xi_1 = \sum_{j=1}^p \alpha_{1j} x_j = \boldsymbol{\alpha}_1^T \mathbf{x}$$

$$\text{var}(\xi_1) = E[\xi_1^2] - E[\xi_1]^2 = E[\boldsymbol{\alpha}_1^T \mathbf{x} \mathbf{x}^T \boldsymbol{\alpha}_1] - E[\boldsymbol{\alpha}_1^T \mathbf{x}] E[\mathbf{x}^T \boldsymbol{\alpha}_1] = \boldsymbol{\alpha}_1^T \boldsymbol{\Sigma} \boldsymbol{\alpha}_1$$

其中， $\boldsymbol{\Sigma}$ 是 $\mathbf{x}$ 的协方差矩阵

求下列拉格朗日函数的极值：

$$f(\boldsymbol{\alpha}_1) = \boldsymbol{\alpha}_1^T \boldsymbol{\Sigma} \boldsymbol{\alpha}_1 - v \boldsymbol{\alpha}_1^T \boldsymbol{\alpha}_1$$

得：

$$\boldsymbol{\Sigma} \boldsymbol{\alpha}_1 = v \boldsymbol{\alpha}_1; \text{var}(\xi_1) = \boldsymbol{\alpha}_1^T \boldsymbol{\Sigma} \boldsymbol{\alpha}_1 = v \boldsymbol{\alpha}_1^T \boldsymbol{\alpha}_1 = v$$

最优的 $\boldsymbol{\alpha}_1$ 为 $\boldsymbol{\Sigma}$ 的最大本质值对应的本征向量， ξ_1 称作第一主成分，它在原始特征的所有线性组合里是方差最大的。

PCA主成分分析

PCA最大方差理论：数学表示

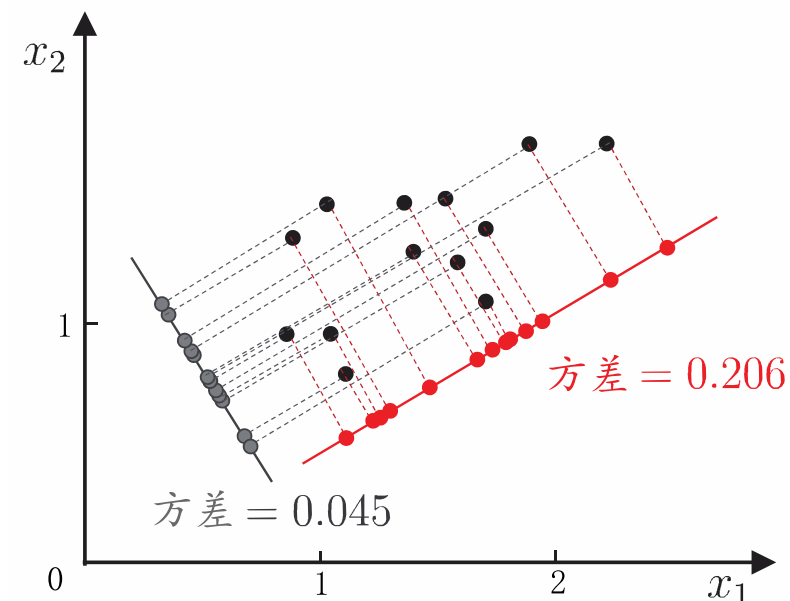
考虑第二个新特征 ξ_2 ，必须满足第一主成分不相关，即

$$E[\xi_2 \xi_1] - E[\xi_2]E[\xi_1] = 0$$

推导可得： $\alpha_2^T \Sigma \alpha_1 = 0$

在约束条件 $\alpha_2^T \alpha_1 = 0$, $\alpha_2^T \alpha_2 = 1$ 下最大化 ξ_2 的方差

α_2 为 Σ 的第二本征值对应的本征向量， ξ_2 称作第二主成分



PCA主成分分析

设 $\{X\}$ 为 ω_i 类的样本集， X 为 n 维向量。

步骤： 第一步：根据样本集求 ω_i 类的协方差矩阵（类内散布矩阵）。

$$C = \frac{1}{N} \sum_{i=1}^N (X_i - M)(X_i - M)^T \quad \text{其中, } M = \frac{1}{N} \sum_{i=1}^N X_i$$

第二步：计算 C 的特征值，对特征值从大到小进行排队，选择前 m 个。

第三步：计算前 m 个特征值对应的特征向量 $u_1, u_2, \dots, u_m$ ，并归一化处理。

将归一化后的特征向量作为矩阵 A 的行。

$$A = \begin{bmatrix} u_1^T \\ u_2^T \\ \dots \\ u_m^T \end{bmatrix}$$

第四步：利用 A 对样本集 $\{X\}$ 进行变换。

$$X^* = AX$$

则 m 维（ $m < n$ ）模式向量 X^* 就是作为分类用的模式向量。

Any Questions?

