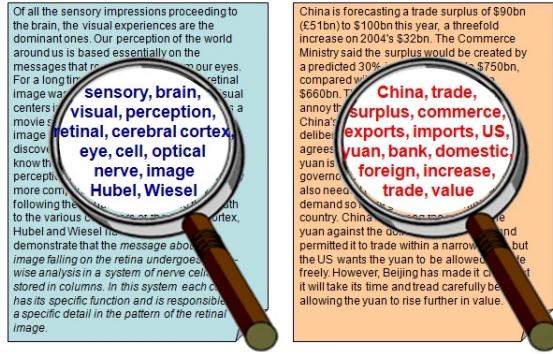


降维：主成分分析

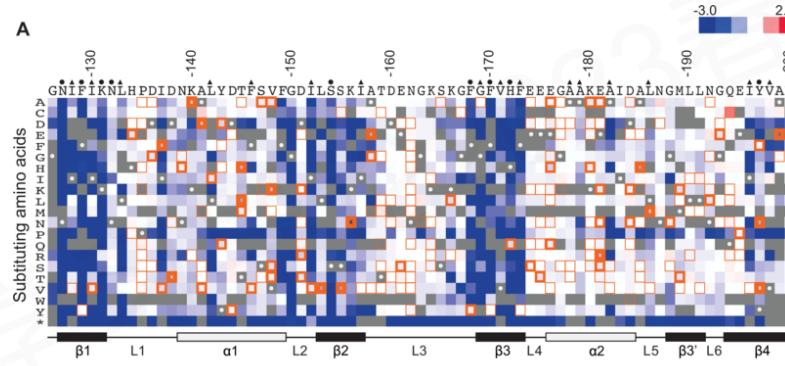
主讲教师：赵鹏

机器学习的维度问题

机器学习常常会面临高维问题

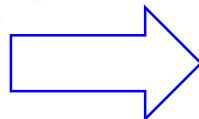


文本分析



基因组学数据分析

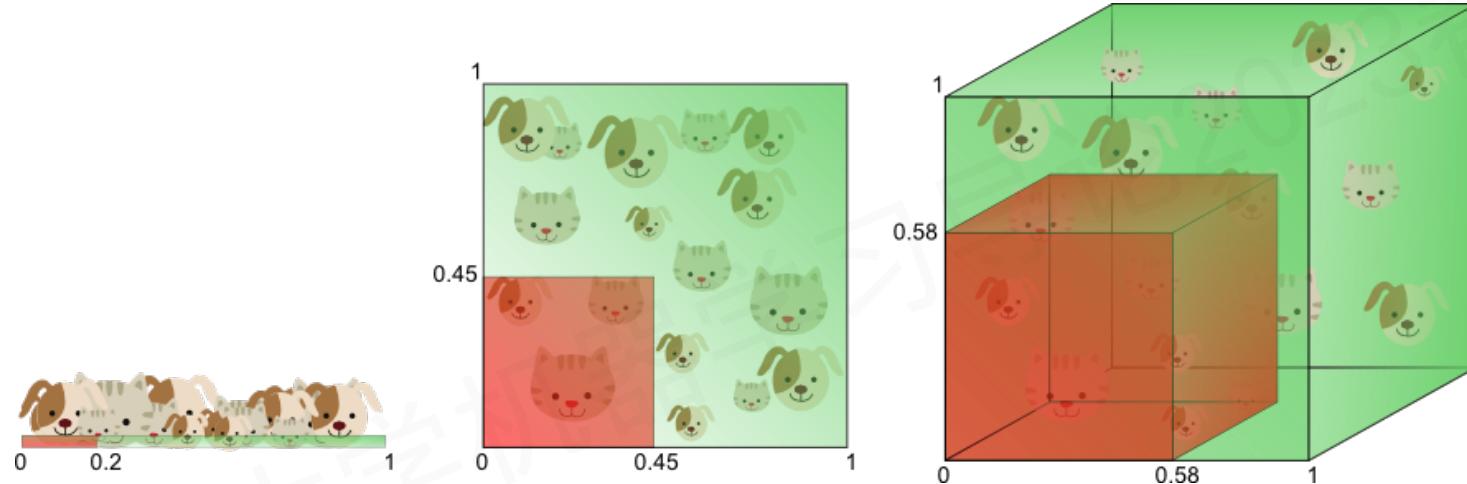
高维带来的挑战：数据样本稀疏、距离计算困难



维度灾难 (curse of dimensionality)

维度灾难 (curse of dimensionality)

高维带来的挑战：数据样本稀疏、距离计算困难



$$\sqrt[1]{0.2} = 0.2$$

$$\sqrt{0.2} \approx 0.45$$

$$\sqrt[3]{0.2} \approx 0.58$$

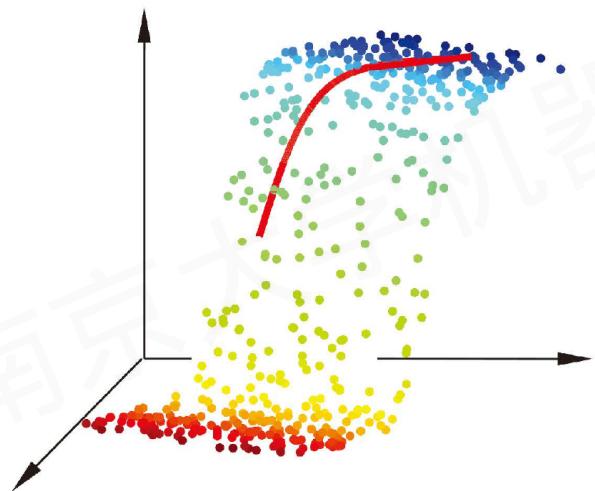
很多学习算法依赖于“密采样”假设： K 近邻算法

如何应对“维度灾难”

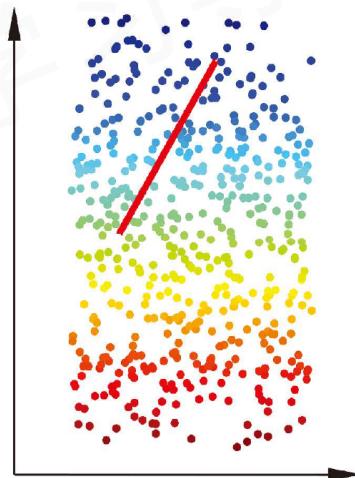
缓解“维度灾难”的重要途径——降维（维数约简）

通过某种数学变换

将原始高维属性空间，变成一个低维的“子空间”



(a) 三维空间中观察到的样本点

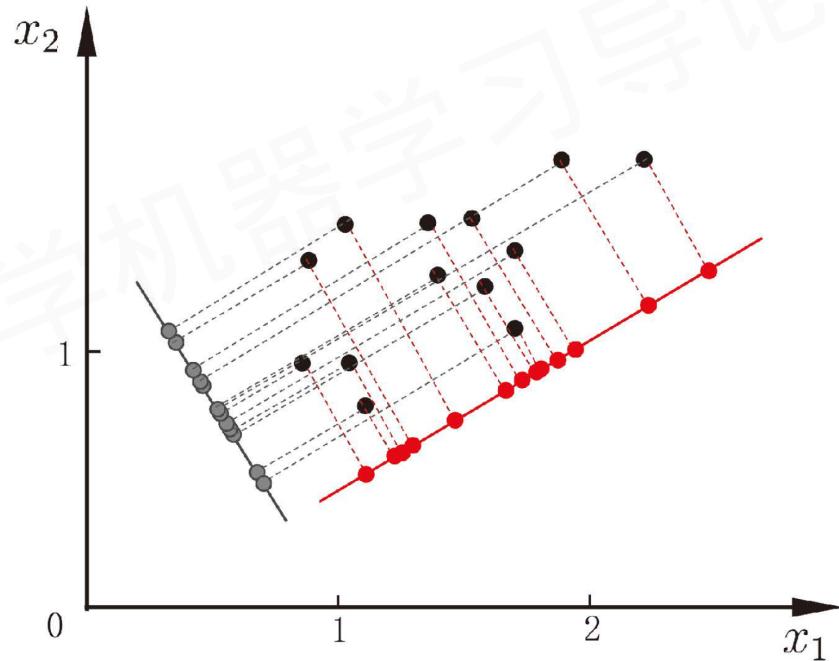


“嵌入”(embedding)

(b) 二维空间中的曲面

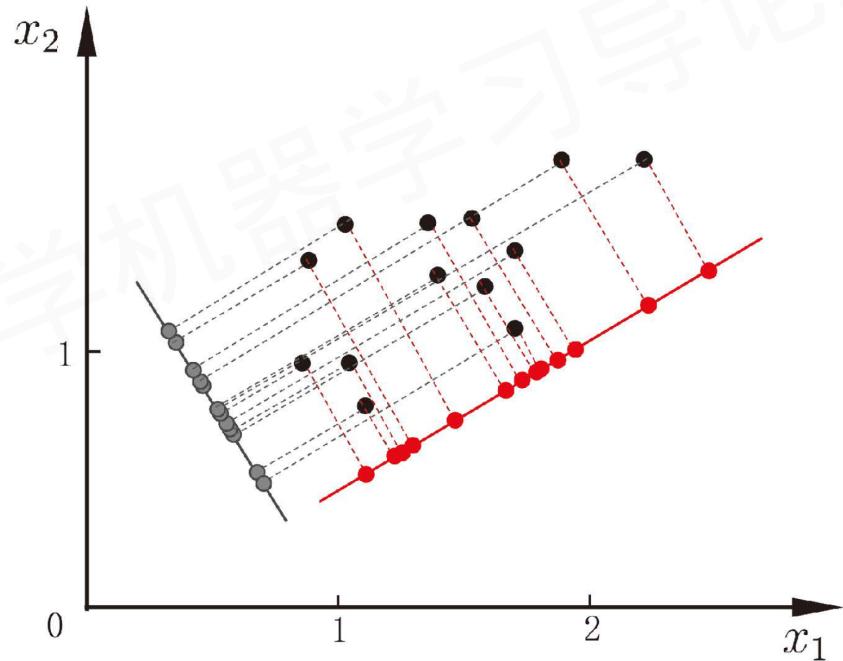
主成分分析 (Principal Component Analysis, PCA)

给定一些高维属性空间中的样本点，如何使用一个低维的超平面对所有样本进行“恰当”的表达？



主成分分析 (Principal Component Analysis, PCA)

给定一些高维属性空间中的样本点，如何使用一个低维的超平面对所有样本进行“恰当”的表达？

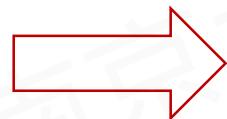


主成分分析 (Principal Component Analysis, PCA)

给定一些高维属性空间中的样本点，如何使用一个低维的超平面对所有样本进行“恰当”的表达？

若存在这样的超平面，直观上我们希望它大概具有这样的性质：

- **最大可分性**：样本点在这个超平面上的投影能尽可能分开
- **最近重构性**：样本点到这个超平面的距离都足够近



由此可得主成分分析的两种等价推导

对样本进行中心化： $\sum_i \mathbf{x}_i = \mathbf{0}$

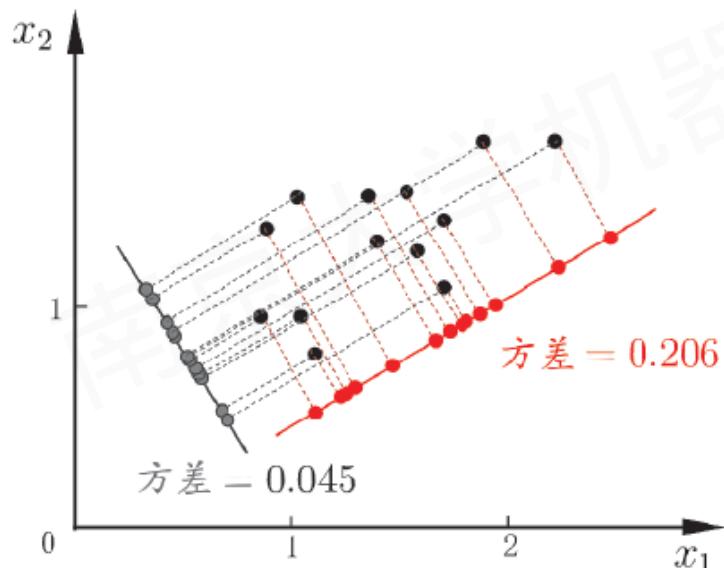
(主要为方便推导；不妨课后考虑：如未进行中心化会怎样？)

PCA - 最大可分性

记投影变换 $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_d) \in \mathbb{R}^{d \times d}$

其中每维是标准正交基向量，满足 $\|\mathbf{w}_i\|_2 = 1, \mathbf{w}_i^\top \mathbf{w}_j = 0 (\forall i \neq j)$

样本点 $\mathbf{x}_i \in \mathbb{R}^d$ 在新空间中超平面上的投影是 $\mathbf{W}[1 : d']^\top \mathbf{x}_i \in \mathbb{R}^{d'}$ ，
若所有样本点的投影尽可能分开，则应最大化投影后样本点方差



考虑投影至1维场景($d' = 1$)：

$$\text{Var}[\{\mathbf{W}^\top \mathbf{x}_i\}] = \frac{1}{m} \sum_{i=1}^m \mathbf{w}_1^\top \mathbf{x}_i \mathbf{x}_i^\top \mathbf{w}_1$$

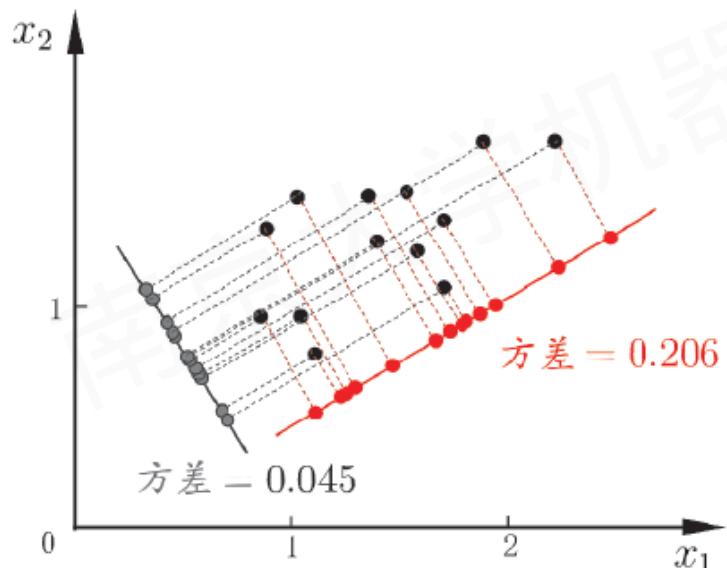
于是： $\max_{\mathbf{w}_1} \mathbf{w}_1^\top \left(\sum_i \mathbf{x}_i \mathbf{x}_i^\top \right) \mathbf{w}_1$

s. t. $\mathbf{w}_1^\top \mathbf{w}_1 = 1$

PCA - 最大可分性

样本点 x_i 在新空间中超平面上的投影是 $\mathbf{W}^T x_i$, 若所有样本点的投影能尽可能分开，则应该使得投影后样本点的方差最大化

投影后样本点的协方差矩阵是 $\sum_i \mathbf{W}^T x_i x_i^T \mathbf{W}$



于是： $\max_{\mathbf{W}} \text{tr}(\mathbf{W}^T \mathbf{X} \mathbf{X}^T \mathbf{W})$
s.t. $\mathbf{W}^T \mathbf{W} = \mathbf{I}.$

等价于：

$$\min_{\mathbf{W}} - \text{tr}(\mathbf{W}^T \mathbf{X} \mathbf{X}^T \mathbf{W})$$

s.t. $\mathbf{W}^T \mathbf{W} = \mathbf{I}.$

协方差

考虑两个实数随机变量 X 和 Y , 其期望为 $\mathbb{E}[X] = \mu$, $\mathbb{E}[Y] = \nu$.

$$\mathbf{Cov}(X, Y) = \mathbb{E}[(X - \mu)(Y - \nu)^\top] = \mathbb{E}[XY^\top] - \mu\nu$$

- 方差是协方差的一个特例, 当两个变量相同时, 定义给出方差
- 当随机变量 X 和 Y 统计独立时, 二者协方差为 0 (反之未必)

考虑 n 个随机变量 X_1, X_2, \dots, X_n , 简单起见, 不妨假设 0 期望均值可以定义协方差矩阵 $\Sigma \in \mathbb{R}^{n \times n}$, 其中

$$\Sigma_{i,j} = \mathbf{Cov}(X_i, X_j) = \mathbb{E}[X_i X_j^\top]$$

- 协方差矩阵是一个对称矩阵
- 对角元是各个随机变量的方差

PCA 求解

$$\begin{aligned} \max_{\mathbf{W}} \quad & \text{tr}(\mathbf{W}^T \mathbf{X} \mathbf{X}^T \mathbf{W}) \\ \text{s.t.} \quad & \mathbf{W}^T \mathbf{W} = \mathbf{I}. \end{aligned}$$

使用拉格朗日乘子法可得

$$\begin{aligned} L(\mathbf{W}, \Theta) &= -\text{tr}(\mathbf{W}^T \mathbf{X} \mathbf{X}^T \mathbf{W}) + \langle \Theta, \mathbf{W}^T \mathbf{W} - \mathbf{I} \rangle \\ &= -\text{tr}(\mathbf{W}^T \mathbf{X} \mathbf{X}^T \mathbf{W}) + \text{tr}(\Theta^T (\mathbf{W}^T \mathbf{W} - \mathbf{I})) \end{aligned}$$

经过推导可得，最终只需求解

$$\mathbf{X} \mathbf{X}^T \mathbf{w}_i = \lambda_i \mathbf{w}_i$$

只需对协方差矩阵 $\mathbf{X} \mathbf{X}^T$ 进行特征值分解，并将求得的特征值排序： $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$ ，再取前 d' 个特征值对应的特征向量构成 $\mathbf{W}^* = (\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_{d'})$ ，这就是主成分分析的解

关键变量：子空间方差

PCA - 最近重构性

对样本进行中心化: $\sum_i \mathbf{x}_i = \mathbf{0}$

假定投影变换后得到的新坐标系为 $\{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_d\}$,

其中每维是标准正交基向量, 满足 $\|\mathbf{w}_i\|_2 = 1, \mathbf{w}_i^\top \mathbf{w}_j = 0 (\forall i \neq j)$

若丢弃新坐标系中的部分坐标, 即将维度降低到 $d' < d$, 则样本点在低维坐标系中的投影是

$$\mathbf{z}_i = (z_{i1}; z_{i2}; \dots; z_{id'}) \quad z_{ij} = \mathbf{w}_j^\top \mathbf{x}_i$$

若基于 \mathbf{z}_i 来重构 \mathbf{x}_i , 则会得到 $\hat{\mathbf{x}}_i = \sum_{j=1}^{d'} z_{ij} \mathbf{w}_j$.

PCA - 最近重构性 (续)

原样本点 \mathbf{x}_i 与基于投影重构的样本点 $\hat{\mathbf{x}}_i$ 之间的距离为

$$\begin{aligned} \sum_{i=1}^m \left\| \sum_{j=1}^{d'} z_{ij} \mathbf{w}_j - \mathbf{x}_i \right\|_2^2 &= \sum_{i=1}^m \mathbf{z}_i^T \mathbf{z}_i - 2 \sum_{i=1}^m \mathbf{z}_i^T \mathbf{W}^T \mathbf{x}_i + \text{const} \\ &\propto -\text{tr} \left(\mathbf{W}^T \left(\sum_{i=1}^m \mathbf{x}_i \mathbf{x}_i^T \right) \mathbf{W} \right). \end{aligned}$$

\mathbf{w}_j 是正交基, $\sum_i \mathbf{x}_i \mathbf{x}_i^T$ 是协方差矩阵, 于是由最近重构性, 有:

$$\begin{aligned} \min_{\mathbf{W}} \quad & -\text{tr}(\mathbf{W}^T \mathbf{X} \mathbf{X}^T \mathbf{W}) \\ \text{s.t.} \quad & \mathbf{W}^T \mathbf{W} = \mathbf{I}. \end{aligned}$$

关键变量：重构误差

这就是主成分分析的优化目标

PCA

若存在这样的超平面，直观上我们希望它大概具有这样的性质：

- **最大可分性**：样本点在这个超平面上的投影能尽可能分开
- **最近重构性**：样本点到这个超平面的距离都足够近

由此可得主成分分析的两种等价推导

最大化子空间方差

$$\begin{aligned} \max_{\mathbf{W}} \quad & \text{tr}(\mathbf{W}^T \mathbf{X} \mathbf{X}^T \mathbf{W}) \\ \text{s.t.} \quad & \mathbf{W}^T \mathbf{W} = \mathbf{I}. \end{aligned}$$

最小化重构误差

$$\begin{aligned} \min_{\mathbf{W}} \quad & - \text{tr}(\mathbf{W}^T \mathbf{X} \mathbf{X}^T \mathbf{W}) \\ \text{s.t.} \quad & \mathbf{W}^T \mathbf{W} = \mathbf{I}. \end{aligned}$$

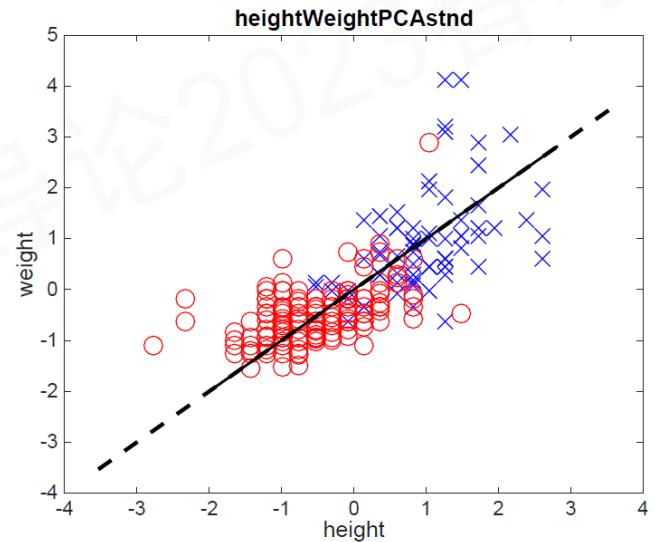
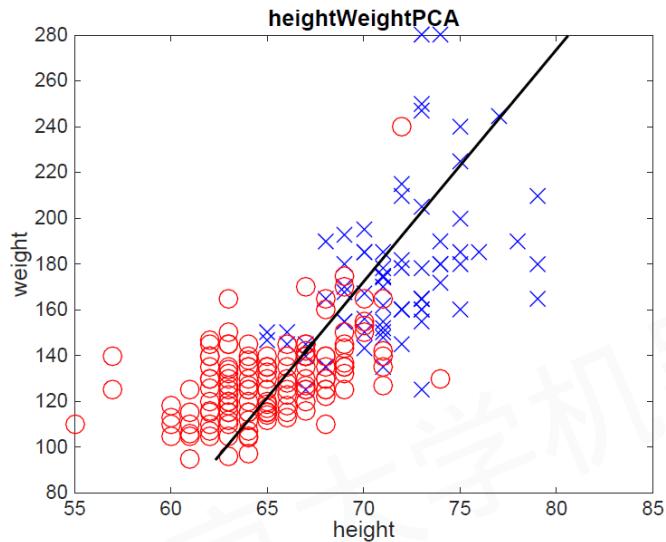
PCA

降维的子空间维度 d' 的设置：

- 用户指定
- 通过低维空间的“二次训练”进行交叉验证
- 通过重构误差判断
$$\frac{\sum_{i=1}^{d'} \lambda_i}{\sum_{i=1}^d \lambda_i} \geq t$$

PCA

协方差矩阵易受到特征尺度影响



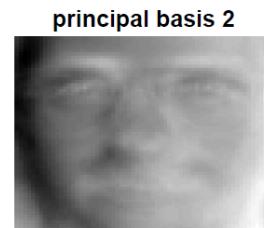
通过对数据进行标准化，使所有特征在同一尺度上

PCA

PCA 是最常用的降维方法，在不同领域有不同的称谓

例如在人脸识别中该技术被称为“特征脸”(eigenface)

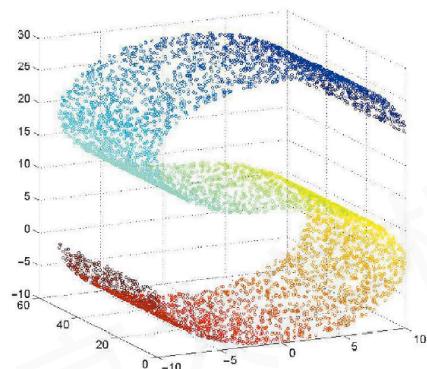
因为若将前 d' 个特征值对应的特征向量还原为图像，则得到



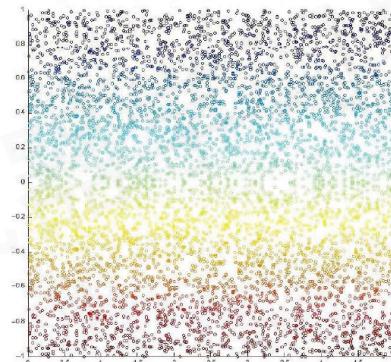
思考：降维体现在哪里？

PCA拓展：Kernelized PCA

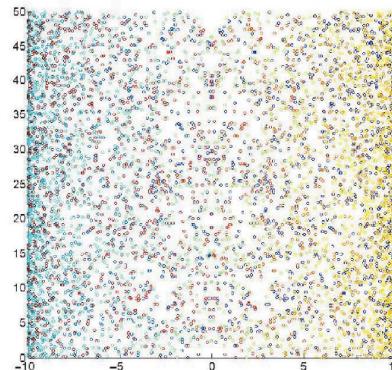
通过引入**核函数**技巧，将PCA的映射函数，从线性拓展到非线性



(a) 三维空间中的观察



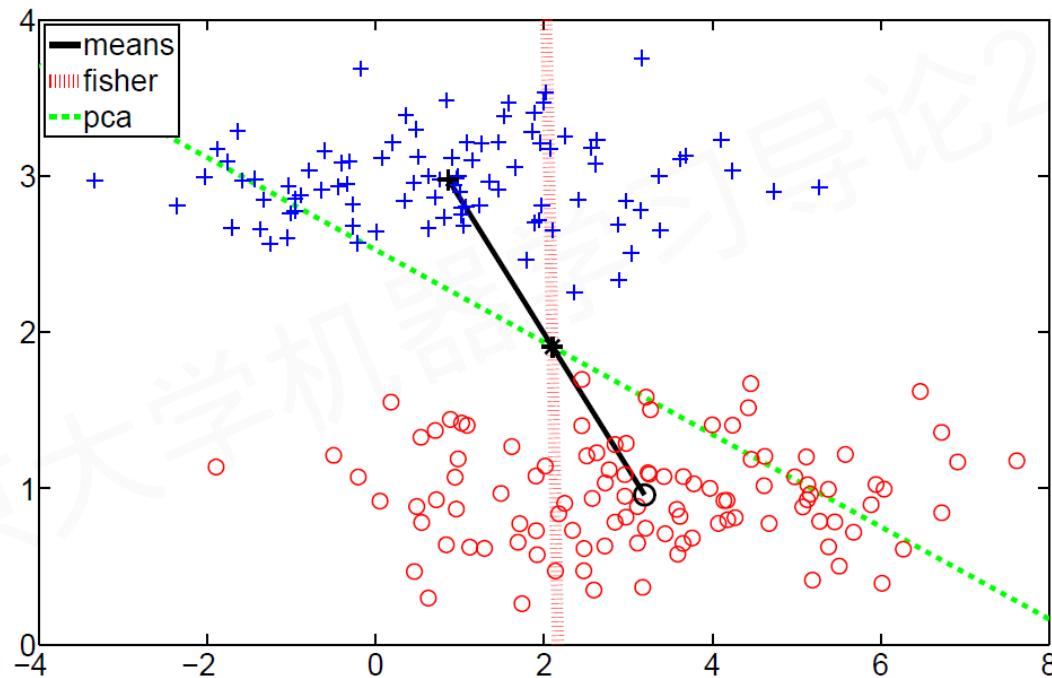
(b) 本真二维结构



(c) PCA 降维结果

PCA拓展： Fisher Discriminant Analysis (FDA)

PCA是无监督学习方法，而FDA是监督学习方法，考虑了标记的作用



PCA拓展： Robust PCA

PCA的低秩理解：

$$\min_{\text{rank}(\hat{X})=d'} \|X - \hat{X}\|_2^2$$

Robust PCA:

$$\min_{\hat{X}} \|X - \hat{X}\|_0 + \text{rank}(\hat{X})$$

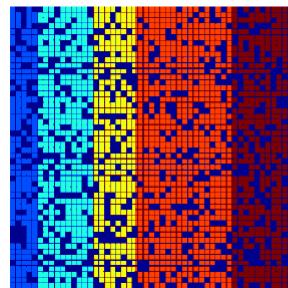


$$\min_{\hat{X}} \|X - \hat{X}\|_1 + \|\hat{X}\|_*$$

PCA拓展： Robust PCA

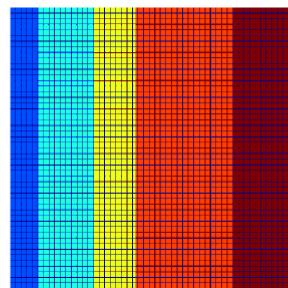
Robust PCA:

$$\min_{\hat{X}} \|X - \hat{X}\|_1 + \|\hat{X}\|_*$$



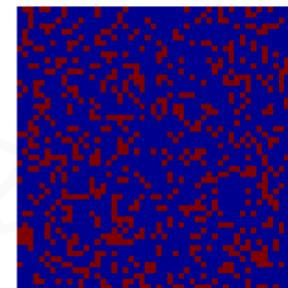
输入数据 X

=

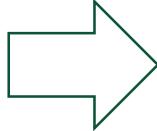


低秩 $\|\hat{X}\|_*$

+



稀疏 $\|X - \hat{X}\|_1$



前往下一站.....

