降维任务

陈飞宇

fchen@cqu.edu.cn

办公室：软件学院529

降维

为什么要降维？

– 去除不相关的特征（去噪、特征提取）

– 储存与计算

– 可视化

– 数据本身具有低维特点

Dimension Reduction Algorithms

线性降维方法:

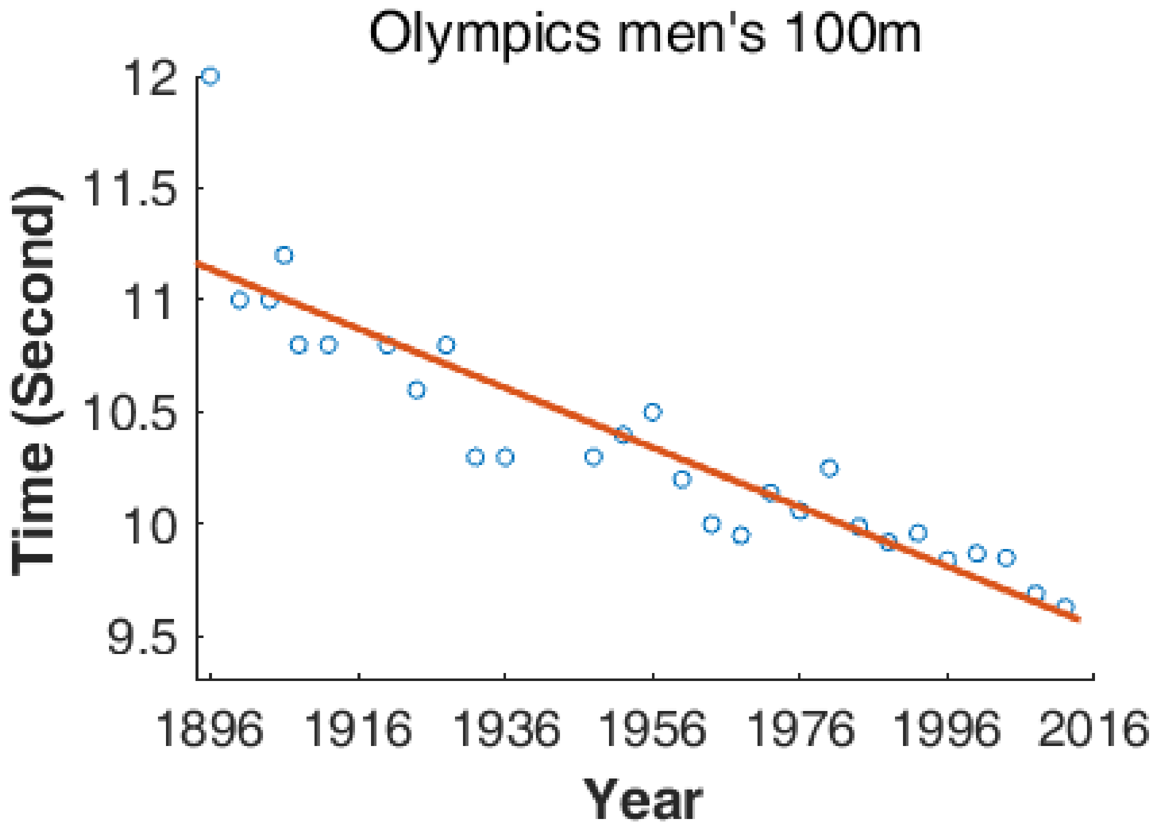
– PCA (Pearson, 1901)
– LDA (Fisher, 1936)
– SVD (Golub and Van Loan, 1983)
– NMF(Lee and Seung, 1999)

非线性降维方法:

– Kernel PCA (Scholkopf et al., 1998)
– Isomap (Tenenbaum et al., 2000)
– MDS (Cox and Cox 2001)

回归任务—直线

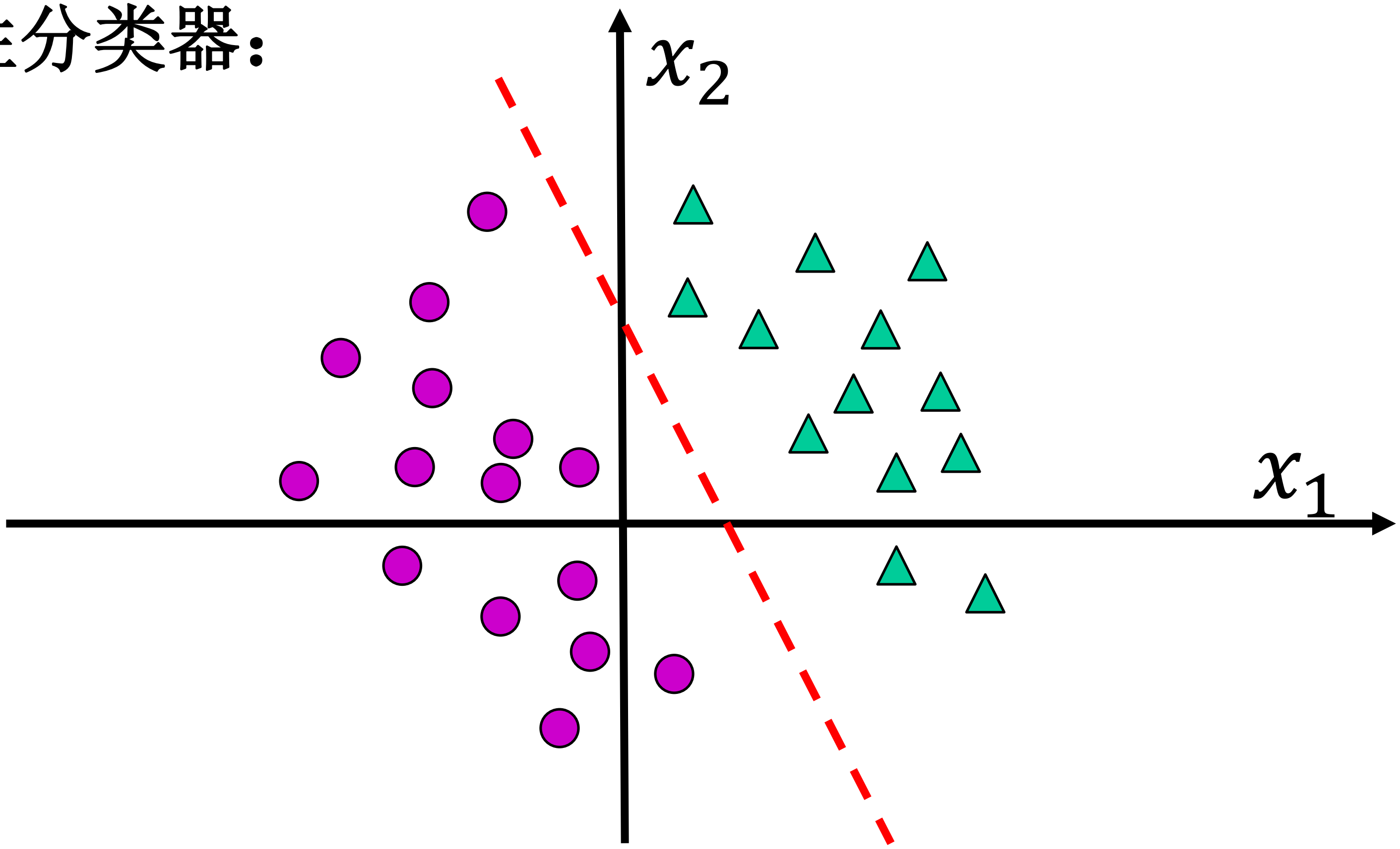
线性回归：



例子取自 A first Course in Machine Learning– S. Rogers and M. Girolami.

分类任务---直线

线性分类器：



6.1 主成分分析

陈飞宇
fchen@cqu.edu.cn
办公室：软件学院529

基础知识

方差 (variance)：

$$\mathrm{Var}(X) = E\left[(X - E[X])^2\right] = E\left[X^2\right] - E[X]^2$$

其中$E[X]$表示随机变量 X 的期望。

协方差 (covariance)：

$$\mathrm{cov}(X, Y) = E[(X - E[X])(Y - E[Y])^{\mathrm{T}}]$$

特别地，$\mathrm{cov}(X) = \mathrm{Var}(X) = \frac{1}{n-1} X X^{\mathrm{T}}$

基础知识

方差 (variance)：

$$\mathrm{Var}(X) = E\left[(X - E[X])^2\right] = E\left[X^2\right] - E[X]^2$$

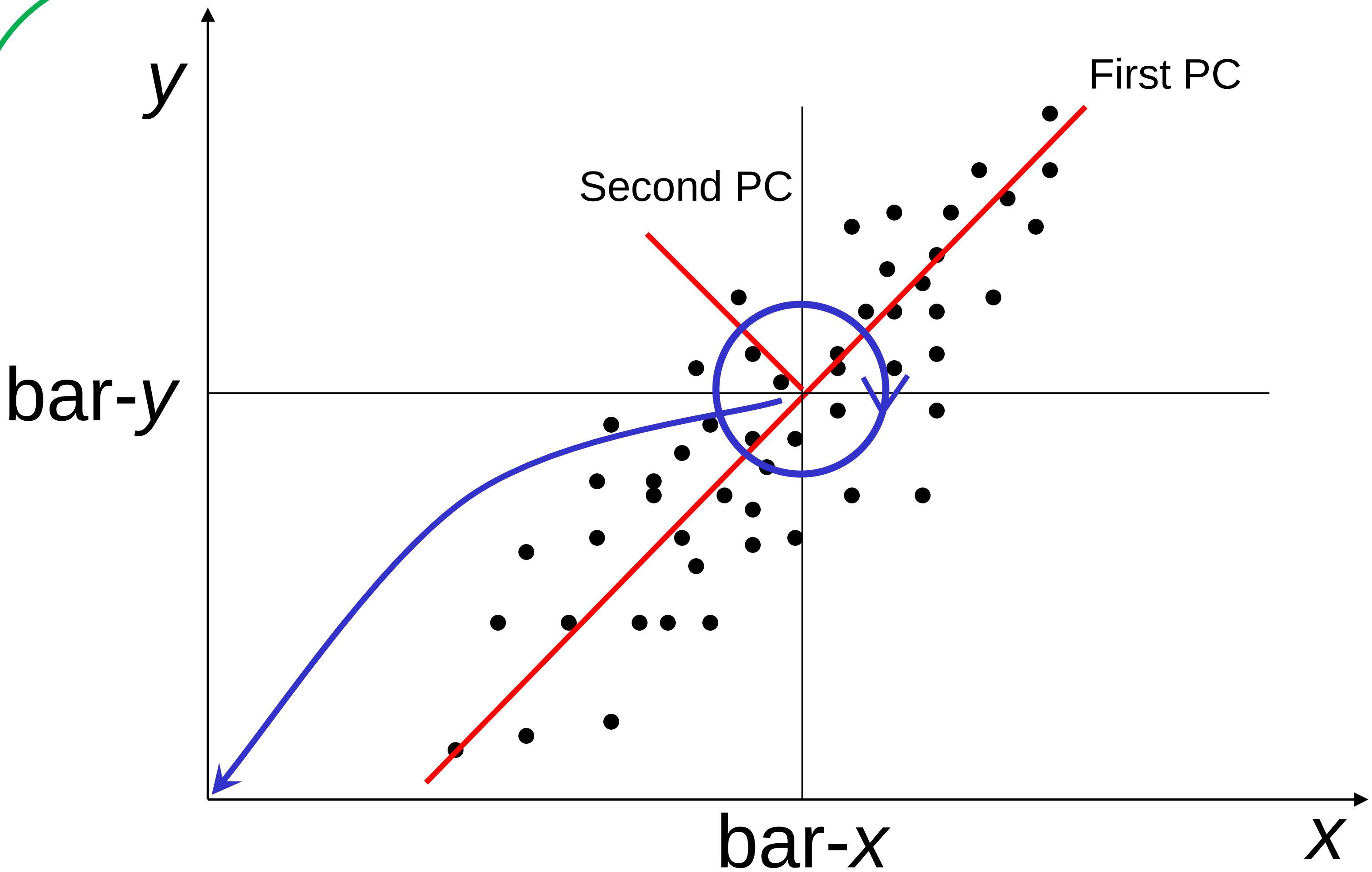
其中$E[X]$表示随机变量 X 的期望。

协方差 (covariance)：

$$\mathrm{cov}(X, Y) = E[(X - E[X])(Y - E[Y])^{\mathrm{T}}]$$

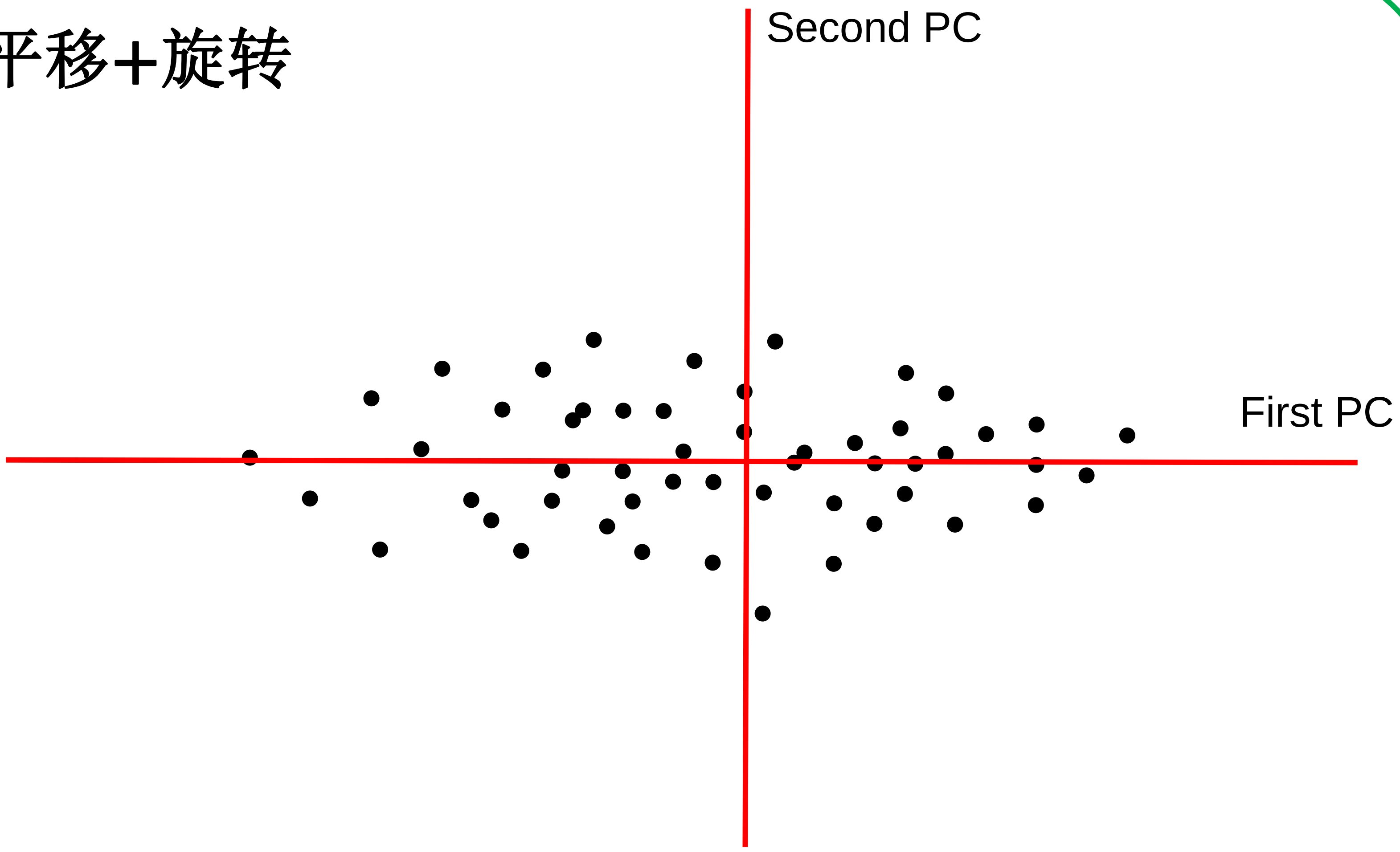
特别地， $\mathrm{cov}(X) = \mathrm{Var}(X) = \frac{1}{n-1} X X^{\mathrm{T}}$

主成分分析示例



主成分分析示例

平移+旋转



主成分分析

精确描述：

- 1^{st} pc 包含了样本方差的最大方向。
- 2^{nd} pc 与第一个主成分不相关（夹角90度），包含了剩余样本方差的最大方向。
- 前几个主成分包含了样本的绝大部分信息，以至于我们可以忽略后面的主成分。

核心思想：寻找合适的正交投影矩阵 $\boldsymbol{U}$，使得投影之后的样本 $\boldsymbol{Y}=\boldsymbol{U}^{\mathrm{T}}\boldsymbol{X}$ 方差达到最大。

$$\max \ \mathbf{Tr}(\mathbf{Var}(\boldsymbol{Y}))$$

$$\mathbf{s.t.}\ \ \boldsymbol{Y}=\boldsymbol{U}^{\mathrm{T}}\boldsymbol{X},\ \boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}=\boldsymbol{I}$$

主成分分析

假设 $\boldsymbol{Y} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{X}$是经过变换之后的数据，

$$
\begin{aligned}
&\mathrm{Var}(\boldsymbol{Y}) \\
&= \boldsymbol{E}\left((\boldsymbol{Y} - \boldsymbol{E}[\boldsymbol{Y}])(\boldsymbol{Y} - \boldsymbol{E}[\boldsymbol{Y}])^{\mathrm{T}}\right) \\
&= \boldsymbol{E}\left(\left(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{X} - \boldsymbol{E}[\boldsymbol{U}^{\mathrm{T}}\boldsymbol{X}]\right)\left(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{X} - \boldsymbol{E}[\boldsymbol{U}^{\mathrm{T}}\boldsymbol{X}]\right)^{\mathrm{T}}\right) \\
&= \boldsymbol{U}^{\mathrm{T}}\boldsymbol{E}\left((\boldsymbol{X} - \boldsymbol{E}[\boldsymbol{X}])(\boldsymbol{X} - \boldsymbol{E}[\boldsymbol{X}])^{\mathrm{T}}\right)\boldsymbol{U} \\
&= \boldsymbol{U}^{\mathrm{T}}\mathrm{Var}(\boldsymbol{X})\boldsymbol{U}
\end{aligned}
$$

主成分分析

因此，PCA模型：

$$\max \quad \mathbf{Tr}(\mathbf{Var}(\boldsymbol{Y}))$$

$$\mathbf{s.t.} \quad \boldsymbol{Y} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{X},\ \boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{I}$$

等价于

$$\max \quad \mathbf{Tr}(\boldsymbol{U}^{\mathrm{T}}\mathbf{Var}(\boldsymbol{X})\ \boldsymbol{U})$$

$$\mathbf{s.t.} \quad \boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{I}$$

此模型的最优解是矩阵Var($\boldsymbol{X}$)的前k个特征向量

拉格朗日乘子法

主成分分析模型：

$$\max \quad \mathbf{Tr}(\boldsymbol{U}^{\mathrm{T}}\mathbf{Var}(\boldsymbol{X})\,\boldsymbol{U}) \quad \mathbf{s.t.} \quad \boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}=\boldsymbol{I}$$

拉格朗日函数：

$$\mathcal{L}(\boldsymbol{U},\boldsymbol{\Lambda}) = -\mathbf{Tr}(\boldsymbol{U}^{\mathrm{T}}\mathbf{Var}(\boldsymbol{X})\,\boldsymbol{U}) + \langle \boldsymbol{\Lambda}, \boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}-\boldsymbol{I}\rangle$$

$\boldsymbol{\mathcal{L}}$ 对变量 $\boldsymbol{U}$ 求偏导：

$$\frac{\partial\mathcal{L}(\boldsymbol{U},\boldsymbol{\Lambda})}{\partial \boldsymbol{U}} = -2\mathbf{Var}(\boldsymbol{X})\boldsymbol{U} + 2\boldsymbol{\Lambda}\boldsymbol{U}$$

等价于$\mathbf{Var}(\boldsymbol{X})\,\boldsymbol{U} = \boldsymbol{\Lambda}\boldsymbol{U}$最大k个特征值对应的特征向量

PCA:最大可分性

主成分分析模型：

$$\max \quad \mathbf{Tr}(\boldsymbol{U}^{\mathrm{T}}\mathbf{Var}(\boldsymbol{X})\,\boldsymbol{U})$$

$$\mathbf{s.t.} \qquad \boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{I}$$

- 此模型的最优解是对称半正定矩阵 $\mathbf{Var}(\boldsymbol{X})$ 的前 $\boldsymbol{k}$ 个特征值所对应的特征向量。
- 此模型的目标是最大化投影后数据的方差，这个性质通常被称为PCA的 最大可分性 。

主成分分析

PCA 算法流程:

Step 1: 中心化 $\bar{X} = X - E[X]$

Step 2: 计算协方差矩阵 $C = \mathrm{cov}(\bar{X}) = \frac{1}{n-1}\bar{X}\bar{X}^{\mathrm{T}}$

Step 3: 特征值分解 $C = U\Lambda U^{\mathrm{T}}$

Step 4: 降维 $Y = U_k^{\mathrm{T}}\bar{X}$, 其中 U_k^{T} 为U^{T}的前k行。

主成分分析

中心化前：

$$\mathrm{cov}(X) = E[(X - E[X])(X - E[X])^{\mathrm{T}}] = E[\bar{X}\bar{X}^{\mathrm{T}}]$$

中心化后：

$$\mathrm{cov}(\bar{X}) = E[(\bar{X} - E[\bar{X}])(\bar{X} - E[\bar{X}])^{\mathrm{T}}] = E[\bar{X}\bar{X}^{\mathrm{T}}]$$

因此，

$$\mathrm{cov}(X) = \mathrm{cov}(\bar{X}) = \frac{1}{n-1}\bar{X}\bar{X}^{\mathrm{T}}$$

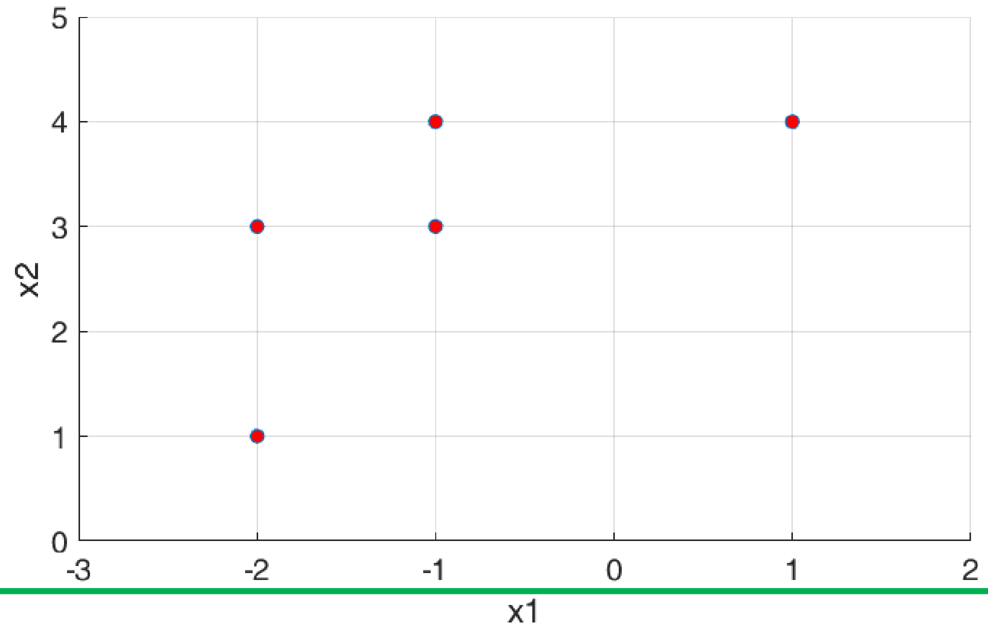
问题：中心化的意义？

例子

假设二维数据点集为

$$X=\begin{pmatrix} -2 & -2 & -1 & 1 & -1 \\ 1 & 3 & 3 & 4 & 4 \end{pmatrix}$$

试利用主成分分析算法对该数据进行降维。



例子

解：$\boldsymbol{X} = \begin{pmatrix} -2 & -2 & -1 & 1 & -1 \\ 1 & 3 & 3 & 4 & 4 \end{pmatrix}$

Step 1: 中心化

$$\overline{\boldsymbol{X}} = \boldsymbol{X} - \boldsymbol{E}[\boldsymbol{X}] = \begin{pmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{pmatrix}$$

Step 2: 计算协方差矩阵

$$\boldsymbol{C} = \mathrm{cov}(\overline{\boldsymbol{X}}) = \frac{1}{n-1}\overline{\boldsymbol{X}}\,\overline{\boldsymbol{X}}^{\mathrm{T}} = \frac{1}{4}\begin{pmatrix} 6 & 4 \\ 4 & 6 \end{pmatrix}$$

例子

Step 3: 特征值分解

$$\boldsymbol{C} = \boldsymbol{U\Lambda U}^{\mathrm{T}}$$

$$= \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 5/2 & 0 \\ 0 & 1/2 \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}$$

Step 4: 投影、降维

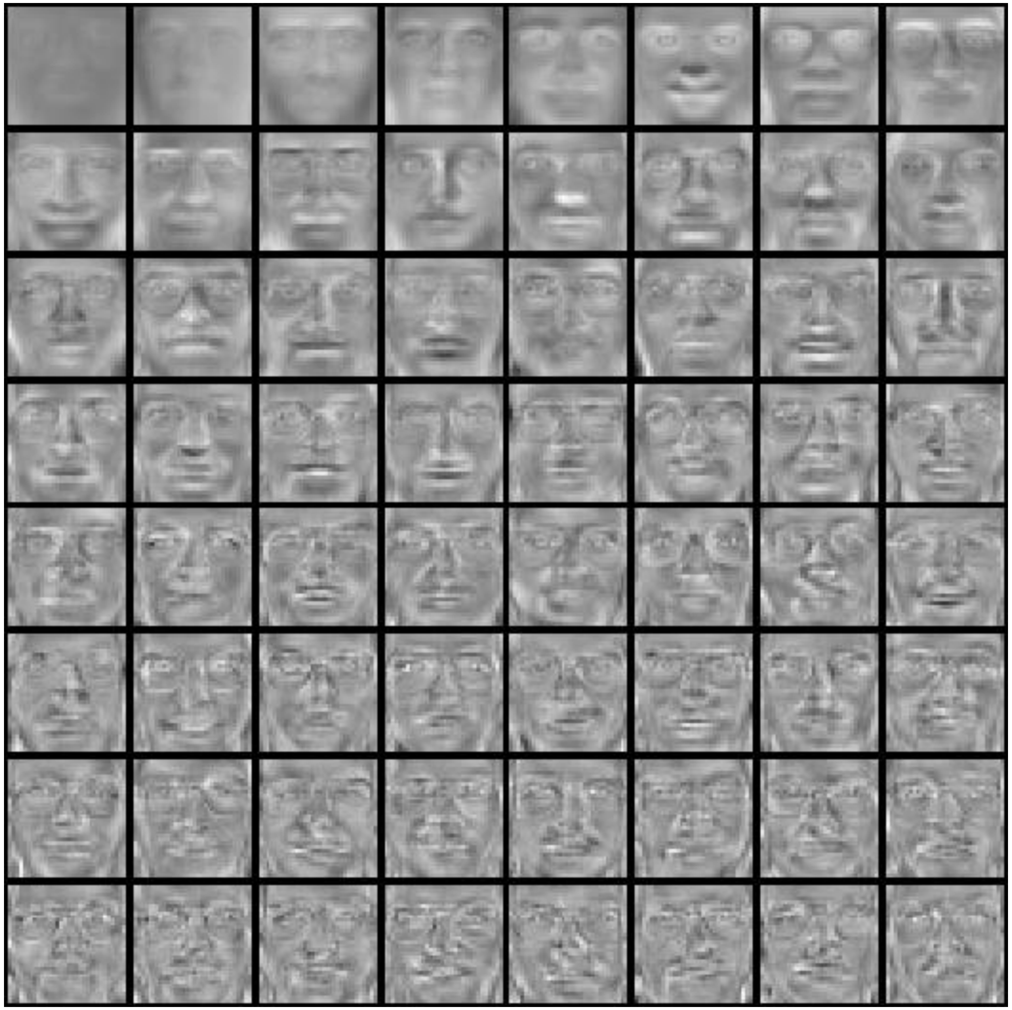
$$\boldsymbol{Y} = \boldsymbol{U}_1^{\mathrm{T}} \overline{\boldsymbol{X}}$$

$$= \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} -3/\sqrt{2} & -1/\sqrt{2} & 0 & 3/\sqrt{2} & -1/\sqrt{2} \end{pmatrix}$$

PCA应用—人脸识别

PCA应用—Eigenface



PCA局限性

What if very high dimensional data?

– e.g., HD image ($d \geq 10^4$)

Problem:

– Covariance matrix C is size (d^2)

– $d = 10^4 \rightarrow |C| = 10^8$

Speedup:

Singular Value Decomposition (SVD)

6.2 奇异值分解

陈飞宇
fchen@cqu.edu.cn
办公室：软件学院529

Singular Value Decomposition

$$\boldsymbol{A}_{\boldsymbol{m}\times\boldsymbol{n}} = \boldsymbol{U}_{\boldsymbol{m}\times\boldsymbol{r}}\boldsymbol{\Sigma}_{\boldsymbol{r}\times\boldsymbol{r}}\boldsymbol{V}_{\boldsymbol{n}\times\boldsymbol{r}}^{\mathbf{T}}$$

- **A**: **数据矩阵**

  m x *n* matrix (e.g., m documents, n terms)
- **U**: **左奇异值向量: U$^{\mathbf{T}}$U = I**

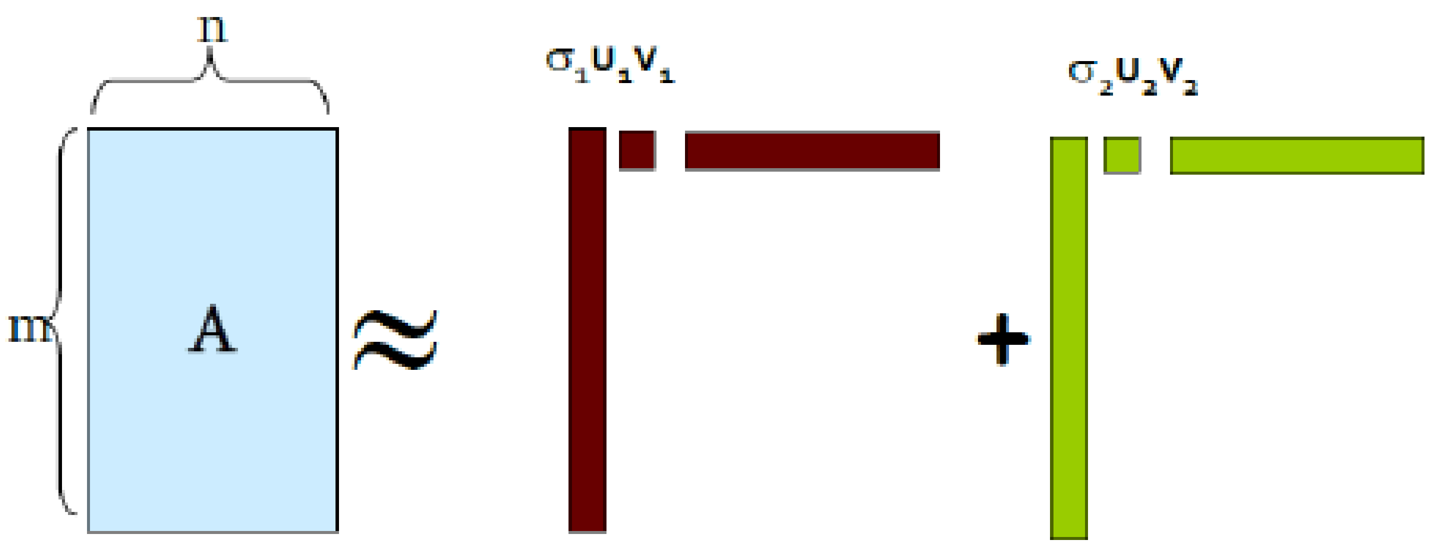
  m x *r* matrix (m documents, r concepts)
- **Σ** : **奇异值, r 为矩阵A的秩**

  r x *r* diagonal matrix (strength of ‘concept’)
- **V**:**右奇异值向量**: **V$^{\mathbf{T}}$V = I**

  n x *r* matrix (n terms, r concepts)

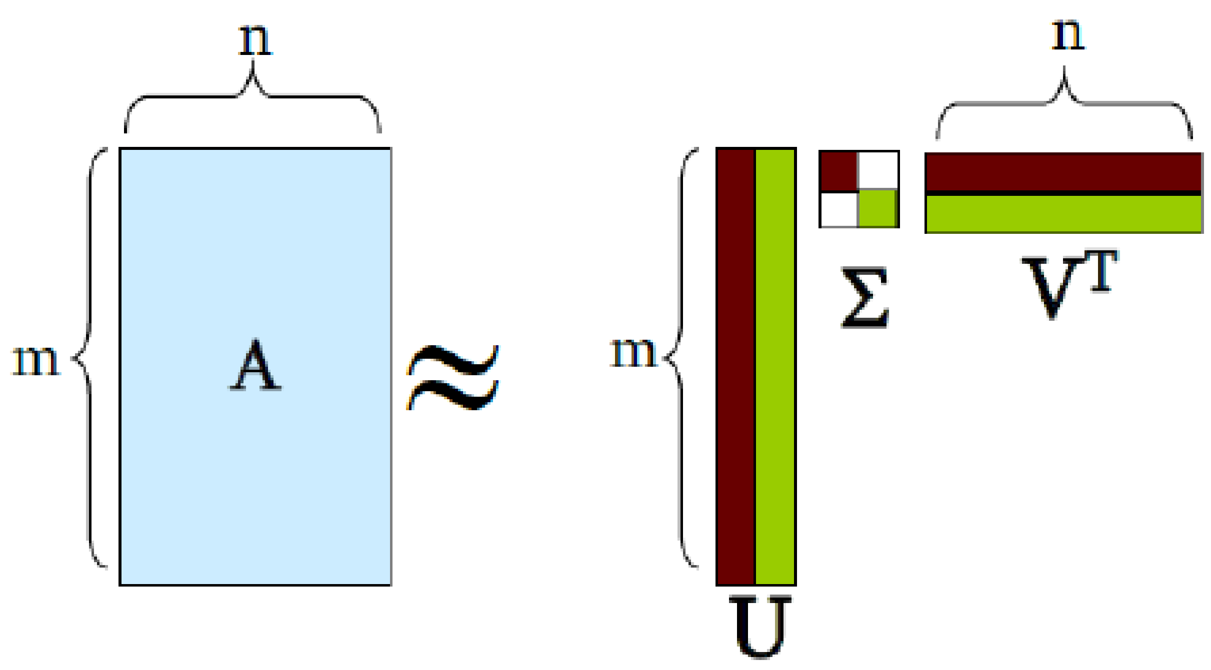
Singular Value Decomposition

$$A \approx U_{m\times k}\Sigma_{k\times k}V^{\mathrm{T}}_{n\times k} = \sum_{i=1}^{k}\sigma_i u_i v_i^T$$



Singular Value Decomposition

$$A_{m\times n} \approx U_{m\times k}\Sigma_{k\times k}V^{\mathrm{T}}_{n\times k}$$



SVD 例子

Problem:

- #1: Find concepts in text
- #2: Reduce dimensionality

term / document	data	information	retrieval	brain	lung
CS-TR1	1	1	1	0	0
CS-TR2	2	2	2	0	0
CS-TR3	1	1	1	0	0
CS-TR4	5	5	5	0	0
MED-TR1	0	0	0	2	2
MED-TR2	0	0	0	3	3
MED-TR3	0	0	0	1	1

SVD 解释

‘ documents’, ‘terms’ and ‘concepts’:

- **U**: document-to-concept similarity matrix
- **V**: term-to-concept similarity matrix
- **Σ** : its diagonal elements: ‘strength’ of each concept

SVD 例子

CS-concept

MD-concept

doc-to-concept

$$
\begin{array}{cc}
 & \begin{array}{ccccc} \text{data} & \text{information} & \text{retrieval} & \text{brain} & \text{lung} \end{array} \\
\begin{array}{c} \text{CS} \\ \\ \\ \\ \text{MD} \end{array} &
\begin{bmatrix}
1 & 1 & 1 & 0 & 0 \\
2 & 2 & 2 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 \\
5 & 5 & 5 & 0 & 0 \\
0 & 0 & 0 & 2 & 2 \\
0 & 0 & 0 & 3 & 3 \\
0 & 0 & 0 & 1 & 1
\end{bmatrix}
\end{array}
=
\boxed{\begin{bmatrix}
0.18 & 0 \\
0.36 & 0 \\
0.18 & 0 \\
0.90 & 0 \\
0 & 0.53 \\
0 & 0.80 \\
0 & 0.27
\end{bmatrix}}
\times
\begin{bmatrix}
9.64 & 0 \\
0 & 5.29
\end{bmatrix}
\times
\begin{bmatrix}
0.58 & 0.58 & 0.58 & 0 & 0 \\
0 & 0 & 0 & 0.71 & 0.71
\end{bmatrix}
$$

SVD 例子

$$
\begin{array}{cc}
 & \begin{array}{ccccc} \text{data} & \text{information} & \text{retrieval} & \text{brain} & \text{lung} \end{array} \\
\begin{array}{c} \uparrow \\ \text{CS} \\ \downarrow \\ \uparrow \\ \text{MD} \\ \downarrow \end{array} & \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}
\end{array}
=
\begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix}
\times
\boxed{\begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix}}
\times
\begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{bmatrix}
$$

‘strength’ of CS-concept

‘strength’ of MD-concept

SVD 例子

$$
\begin{array}{c c}
 & \begin{array}{ccccc} \text{data} & \text{inf.} & \text{retrieval} & \text{brain} & \text{lung} \end{array} \\
\begin{array}{c} \\ \text{CS} \\ \\ \\ \\ \text{MD} \\ \\ \end{array} &
\begin{bmatrix}
1 & 1 & 1 & 0 & 0 \\
2 & 2 & 2 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 \\
5 & 5 & 5 & 0 & 0 \\
0 & 0 & 0 & 2 & 2 \\
0 & 0 & 0 & 3 & 3 \\
0 & 0 & 0 & 1 & 1
\end{bmatrix}
\end{array}
=
\begin{bmatrix}
0.18 & 0 \\
0.36 & 0 \\
0.18 & 0 \\
0.90 & 0 \\
0 & 0.53 \\
0 & 0.80 \\
0 & 0.27
\end{bmatrix}
\mathrm{x}
\begin{bmatrix}
9.64 & 0 \\
0 & 5.29
\end{bmatrix}
\mathrm{x}
\boxed{\begin{bmatrix}
0.58 & 0.58 & 0.58 & 0 & 0 \\
0 & 0 & 0 & 0.71 & 0.71
\end{bmatrix}}
$$

MD-concept

CS-concept

term-to-concept

SVD 例子

降维：另较小的奇异值为0.

$$
\begin{bmatrix}
1 & 1 & 1 & 0 & 0 \\
2 & 2 & 2 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 \\
5 & 5 & 5 & 0 & 0 \\
0 & 0 & 0 & 2 & 2 \\
0 & 0 & 0 & 3 & 3 \\
0 & 0 & 0 & 1 & 1
\end{bmatrix}
=
\begin{bmatrix}
0.18 & \cancel{0} \\
0.36 & \cancel{0} \\
0.18 & \cancel{0} \\
0.90 & \cancel{0} \\
0 & \cancel{0.53} \\
0 & \cancel{0.80} \\
0 & \cancel{0.27}
\end{bmatrix}
\text{ x }
\begin{bmatrix}
9.64 & 0 \\
0 & \cancel{5.29}
\end{bmatrix}
\text{ x }
\begin{bmatrix}
0.58 & 0.58 & 0.58 & 0 & 0 \\
\cancel{0} & \cancel{0} & \cancel{0} & \cancel{0.71} & \cancel{0.71}
\end{bmatrix}
$$

SVD 例子

$$
\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \approx \begin{bmatrix} 0.18 \\ 0.36 \\ 0.18 \\ 0.90 \\ 0 \\ 0 \\ 0 \end{bmatrix} \text{x} \begin{bmatrix} 9.64 \end{bmatrix} \text{x} \begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \end{bmatrix}
$$

SVD 例子

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \approx \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

奇异值分解

- 对于任意矩阵来说，它的奇异值分解存在唯一，且所有奇异值都是正数。

- 利用奇异值分解进行降维处理，其意义是最小化重构误差。

奇异值分解

SVD 算法流程：

Step 1: 中心化 $\bar{\boldsymbol{X}} = \boldsymbol{X} - \boldsymbol{E}[\boldsymbol{X}]$

Step 2: 奇异值分解 $\bar{\boldsymbol{X}} = \boldsymbol{U\Sigma V}^{\mathrm{T}}$

Step 3: 降维 $\boldsymbol{Y} = \boldsymbol{U}_k^{\mathrm{T}}\bar{\boldsymbol{X}}$**, 其中** $\boldsymbol{U}_k^{\mathrm{T}}$ **为**$\boldsymbol{U}^{\mathrm{T}}$**的前k行**
。

PCA vs SVD

- **SVD** – $\bar{X} = U\Sigma V^{\mathrm{T}}$

- **PCA** – $\bar{X}\bar{X}^{\mathrm{T}} = U\Sigma V^{\mathrm{T}}(U\Sigma V^{\mathrm{T}})^{\mathrm{T}} = U\Sigma^2 U^T$

  $\bar{X}^{\mathrm{T}}\bar{X} = (U\Sigma V^{\mathrm{T}})^{\mathrm{T}} U\Sigma V^{\mathrm{T}} = V\Sigma^2 V^{\mathrm{T}}$

SVD on $\bar{X}$ $\Longleftrightarrow$ PCA on $\bar{X}$

PCA vs SVD

Computational Complexity:

- **SVD – O(mn^2) or O(nm^2) 或者更少**
  - 只需要奇异值
  - 或者只需要前几个奇异向量
  - 或者矩阵是稀疏的
- **PCA – O(n^3)**

Storage:

- **SVD – O(mn) (e.g., 数据矩阵)**
- **PCA – O(n^2) (e.g., 协方差矩阵)**

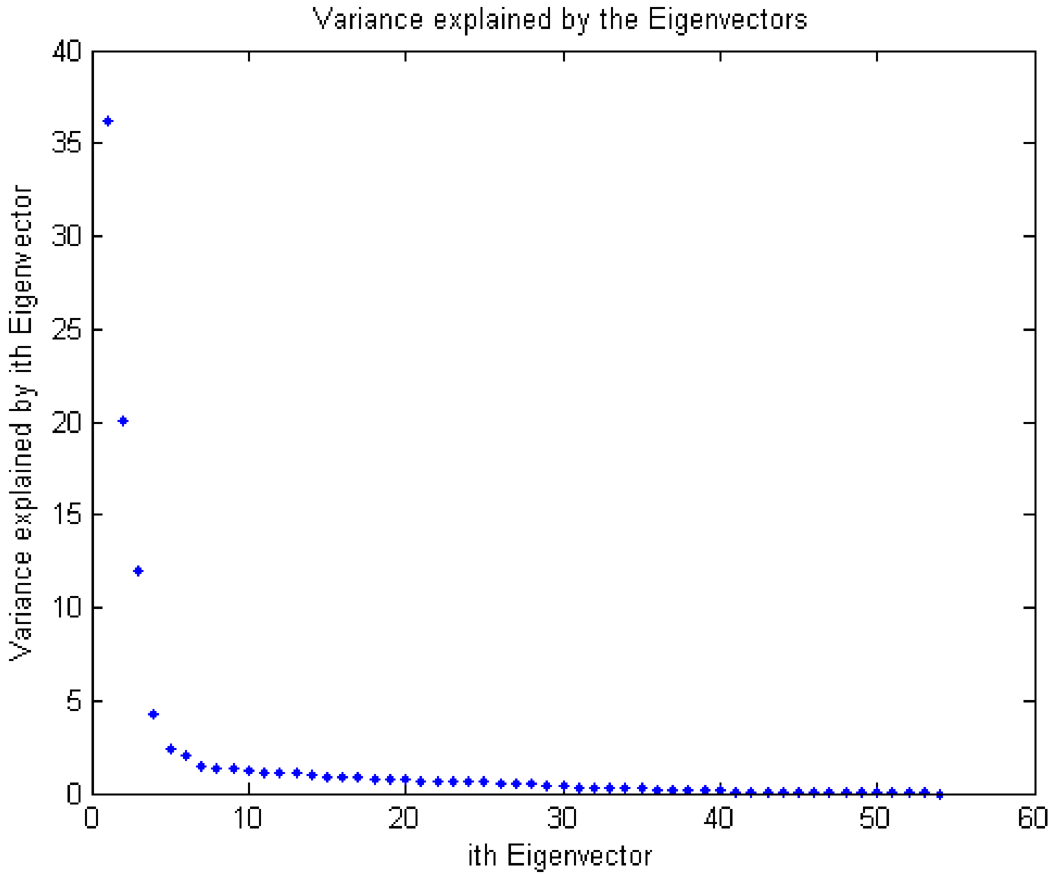
PCA vs SVD

Matlab 对比试验

- **随机数据 [m,n] = [100,4000]**
- **PCA运行时间：69.6601s**
- **SVD 运行时间：0.3414s**

主成分个数选取？

观察特征值或奇异值的分布情况来选取合适的k。



主成分分析小结

- 主成分分析（PCA）是最常用的降维技术。
- 主成分分析最大化投影后数据的方差，同时主成分分析还最小化重建误差。
- 从数值上看，奇异值分解可以大幅度地提高主成分分析算法的计算速度。

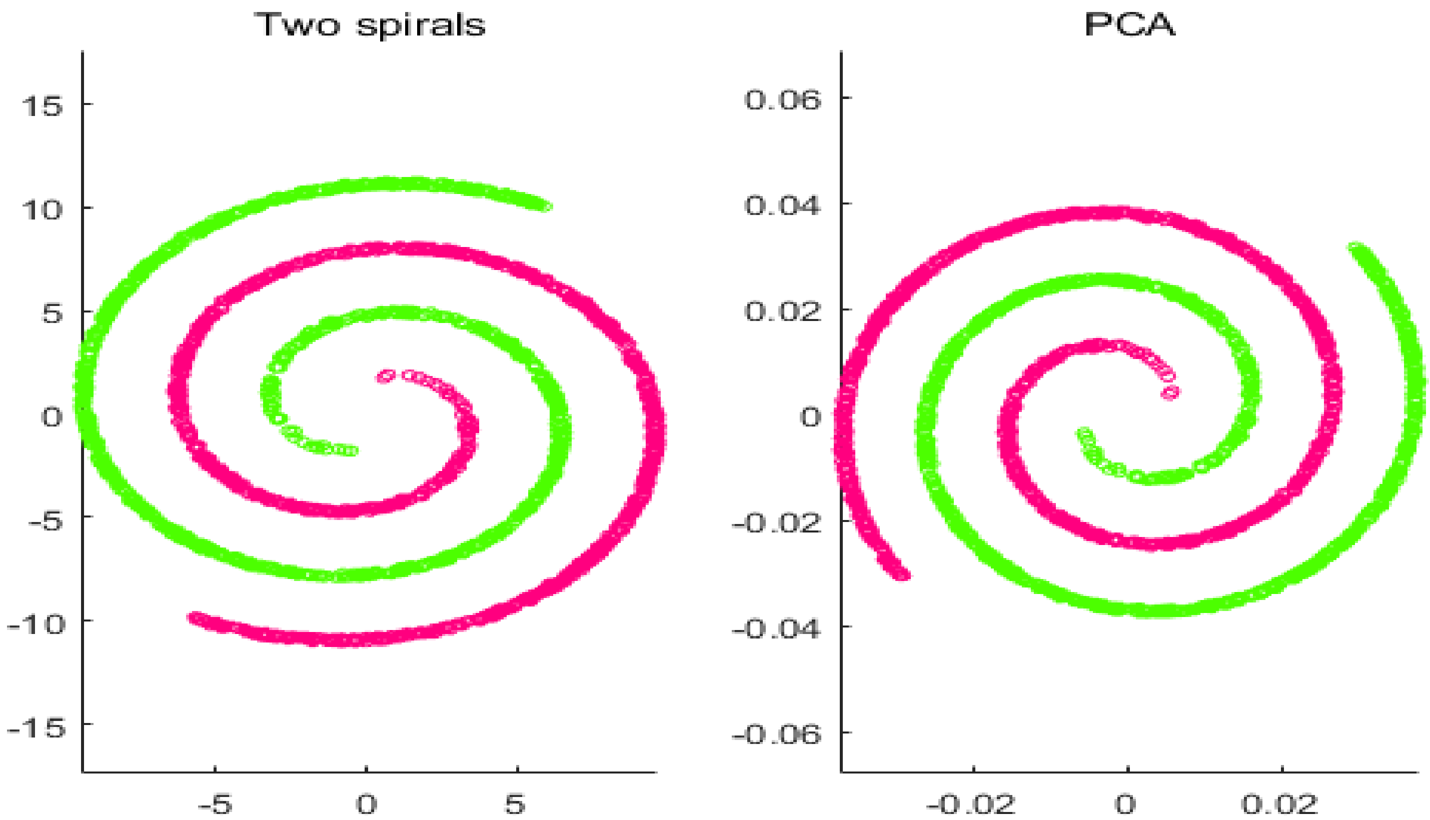
6.3 核化PCA

陈飞宇

fchen@cqu.edu.cn

办公室：软件学院529

非球状数据降维



Two spirals
15
10
5
0
-5
-10
-15
-5
0
5
PCA
0.06
0.04
0.02
0
-0.02
-0.04
-0.06
-0.02
0
0.02

解决方案

对于非线性可分的数据$\{x_i\}_{i=1}^{m}$，尝试找到一个非线性映射 ϕ，使得数据$\{\phi(x_i)\}_{i=1}^{n}$在新的空间（通常为高维空间）是线性可分的，然后再使用PCA进行降维。

椭圆方程： $w_1(x^{(1)})^2+w_2(x^{(2)})^2+b=0$

直线方程： $w_1x^{(1)}+w_2x^{(2)}+b=0$

核技巧（Kernel Trick）

然而，对于任意给定数据寻找合适的非线性映射 $\boldsymbol{\phi}$ 是不现实的。

定义核函数（**Kernel Function**）：

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{\phi}(\boldsymbol{x}_i)^{\mathrm{T}} \boldsymbol{\phi}(\boldsymbol{x}_i)$$

这样，$\boldsymbol{\phi}$ 可以隐式地由核函数表示出来，而且内积的计算也变得更加容易。

核函数

常见的核函数（核矩阵对称半正定）：

名称	表达式
线性核	$\boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j$
多项式核	$\boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j)^d$
高斯核	$\boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\frac{\Vert \boldsymbol{x}_i - \boldsymbol{x}_j \Vert^2}{2\sigma^2})$
拉普拉斯核	$\boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\frac{\Vert \boldsymbol{x}_i - \boldsymbol{x}_j \Vert}{\sigma})$
Sigmoid 核	$\boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \tanh(\beta \boldsymbol{x}_i^T \boldsymbol{x}_j + \theta)$

核化主成分分析

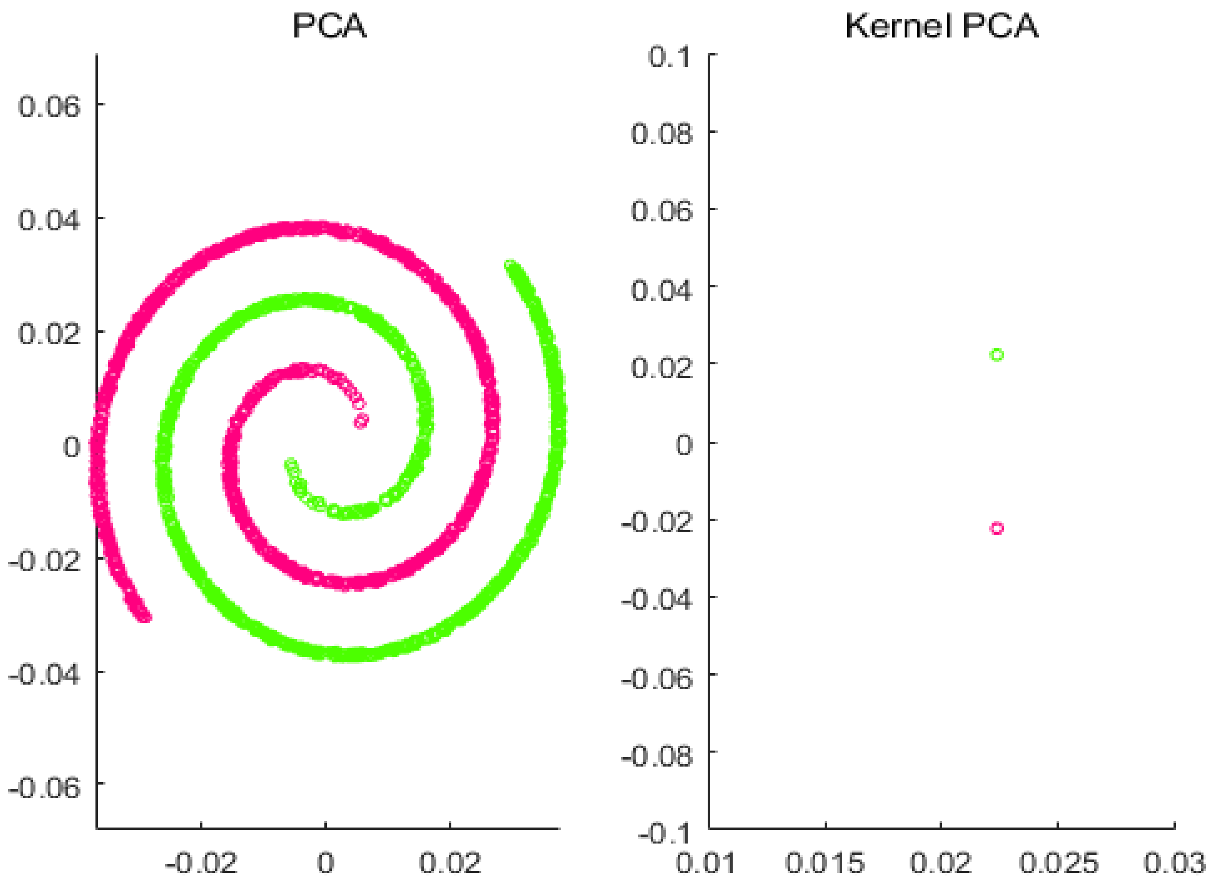
KPCA 算法流程:

Step 1: 选取合适的核函数，求出矩阵$\boldsymbol{K}$

Step 2: 特征值分解 $\boldsymbol{K} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\mathrm{T}}$

Step 3: 降维 $\boldsymbol{Y} = \boldsymbol{U}_k^{\mathrm{T}}\boldsymbol{\Lambda}^{1/2}$。

核化主成分分析



PCA
0.06
0.04
0.02
0
-0.02
-0.04
-0.06
-0.02
0
0.02
Kernel PCA
0.1
0.08
0.06
0.04
0.02
0
-0.02
-0.04
-0.06
-0.08
-0.1
0.01
0.015
0.02
0.025
0.03

核化主成分分析小结

- 核化主成分分析（**KPCA**）是非线性降维方法，是**PCA**的直接推广。
- **KPCA** 可以对非球状数据进行降维。
- 和**PCA**类似，当数据量较大时，**KPCA**的计算效率低。

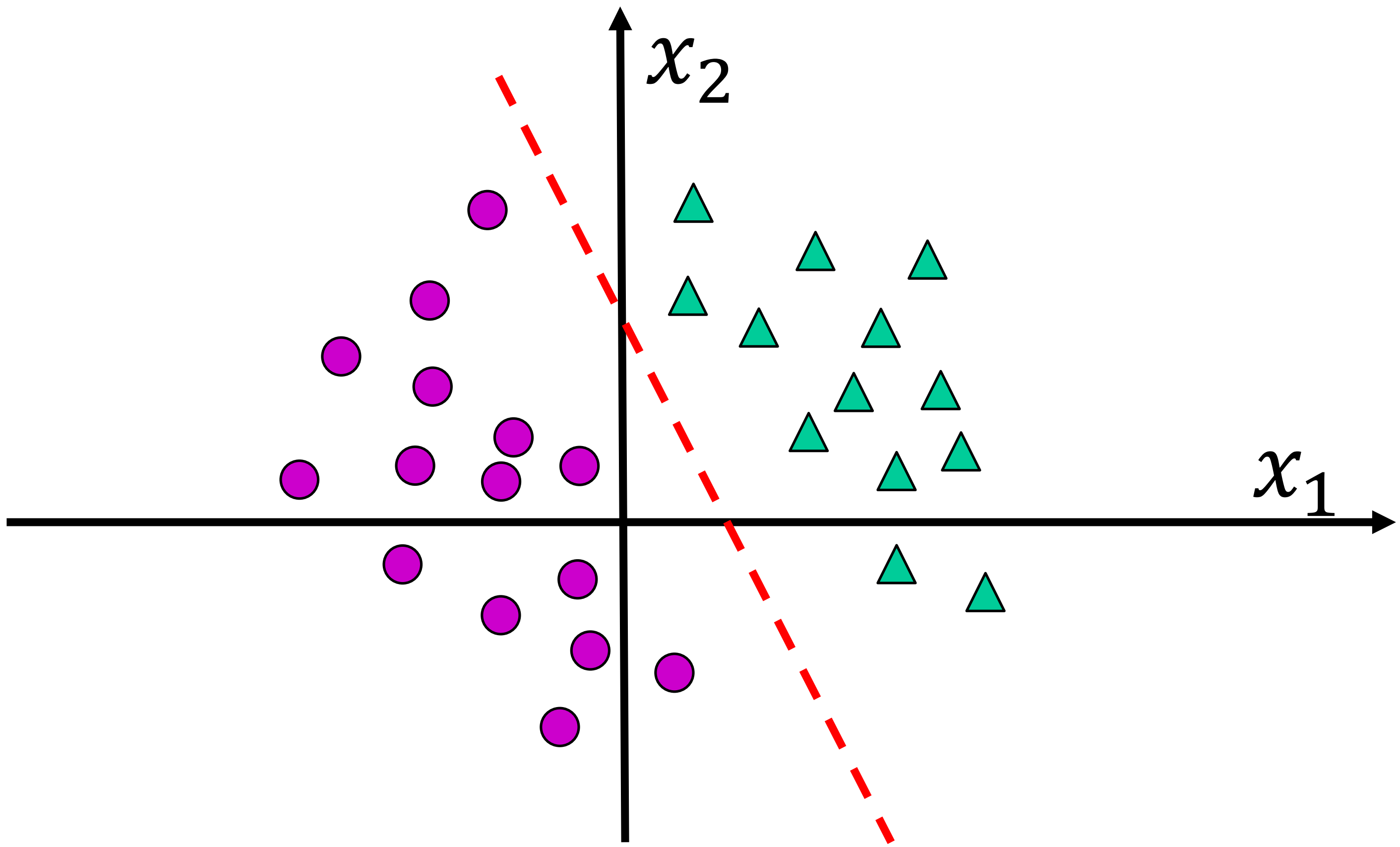
6.4 线性判别分析

陈飞宇

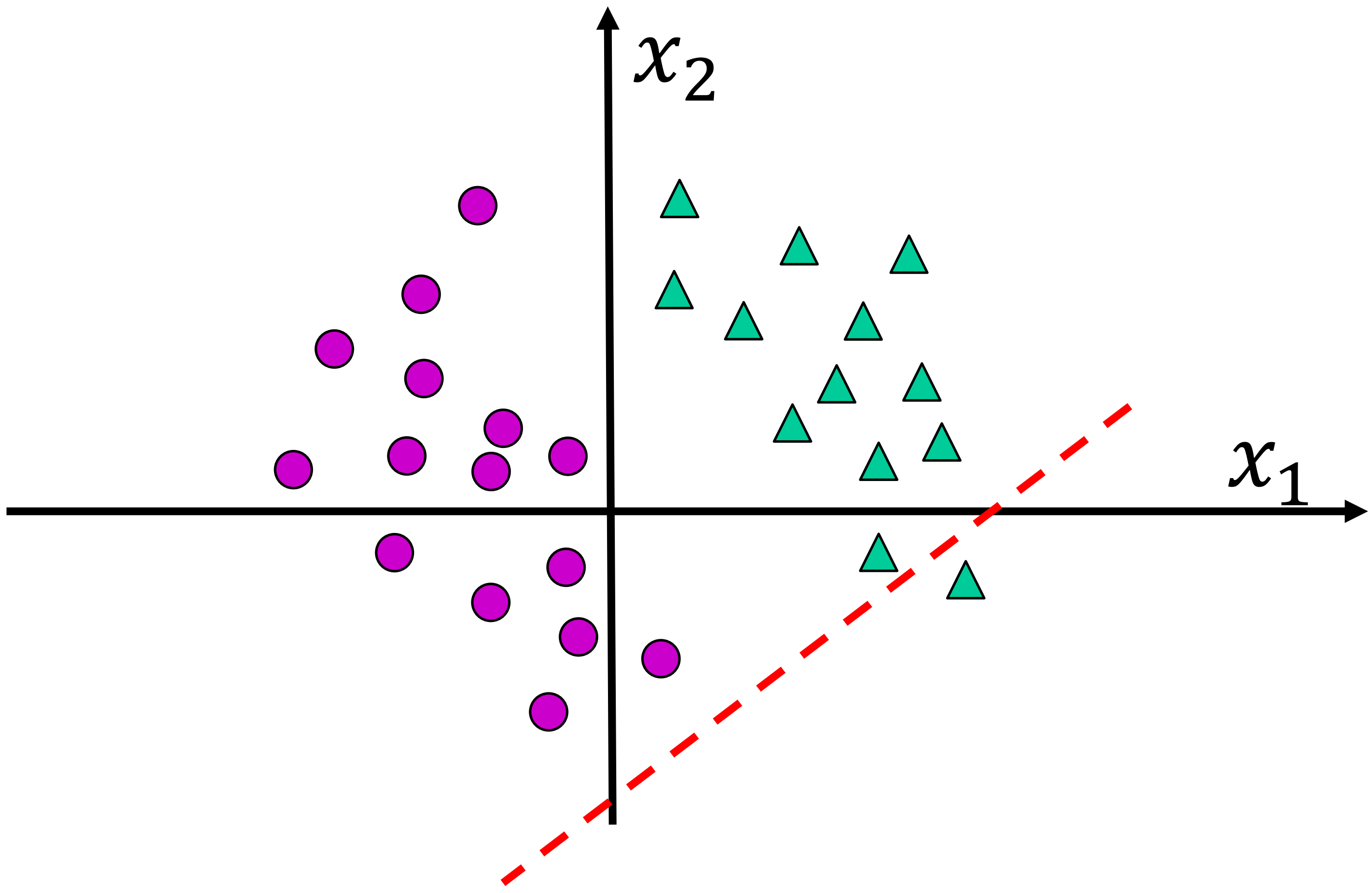
fchen@cqu.edu.cn

办公室：软件学院529

降维+分类?



降维+分类？



降维+分类？

思想：假设降维的目的是解决维度灾难，在低维空间进行分类。那么我们希望降维后的样本保持一定的可分性，即经过投影之后，同类样本的投影点尽可能接近，异类样本的投影点尽可能远离。

解决方案：让投影后的类中心之间的距离尽可能大，从而增强样本的可分性。

降维+分类?

对于二分类问题，给定训练集$\{(x_i, y_i)\}_{i=1}^{n}$，其中$y_i \in \{0,1\}$。令$D_j, n_j, \mu_j, \Sigma_j$分别表示第 $i \in \{0,1\}$类样本的集合、集合的个数、均值向量和协方差矩阵。

假设投影矩阵是w,则投影后的中心为:

$$\frac{1}{n_j}\sum_{x\in D_i} w^T x = w^T(\frac{1}{n_j}\sum_{x\in D_i} x) = w^T\mu_i$$

因此投影后的类中心之间的距离为:

$$\left\|w^T\mu_0 - w^T\mu_1\right\|_2^2 = w^T(\mu_0-\mu_1)(\mu_0-\mu_1)^T w$$

降维+分类?

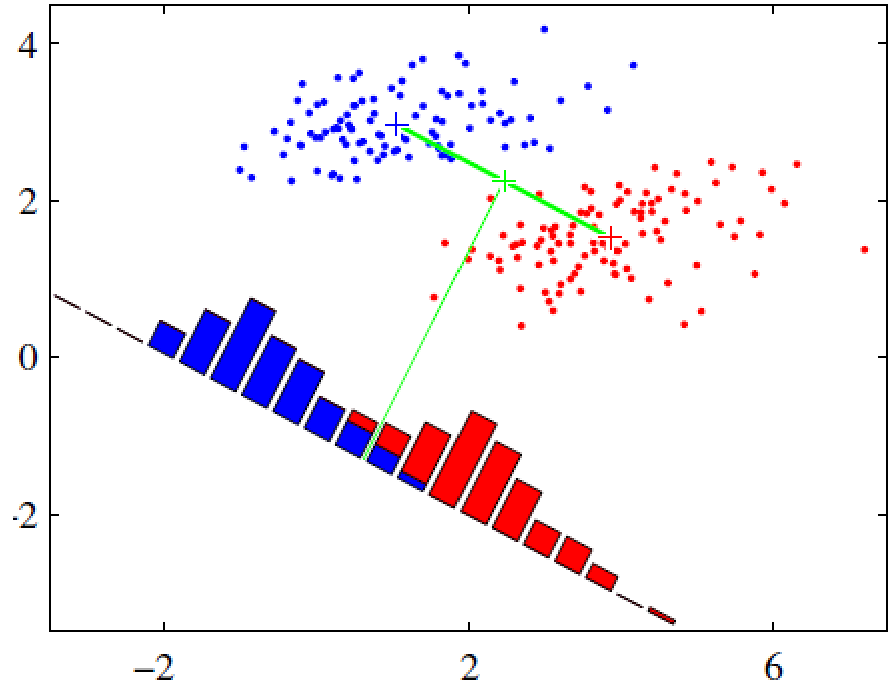
类间散度矩阵（between-class scatter matrix）：

$$S_b = (\mu_0 - \mu_1)(\mu_0 - \mu_1)^T$$

最大化类间散度?

$$\max_{w} \quad w^T S_b w$$

投影后有重合区域!



降维+分类?

投影后类内的样本不够‘紧凑’，即协方差过大，
希望投影后类内样本的协方差小。

$$\min_{\boldsymbol{w}} \quad \boldsymbol{w}^T \boldsymbol{\Sigma}_0 \boldsymbol{w} + \boldsymbol{w}^T \boldsymbol{\Sigma}_1 \boldsymbol{w}$$

类内散度矩阵（within-class scatter matrix）：

$$\boldsymbol{S}_{\boldsymbol{w}} = \boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1$$

$$= \sum_{\boldsymbol{x} \in D_0} (\boldsymbol{x} - \boldsymbol{\mu}_0)(\boldsymbol{x} - \boldsymbol{\mu}_0)^T + \sum_{\boldsymbol{x} \in D_1} (\boldsymbol{x} - \boldsymbol{\mu}_1)(\boldsymbol{x} - \boldsymbol{\mu}_1)^T$$

线性判别分析

对于二维数据来说，线性判别分析（Linear Discriminant Analysis）同时最大化类间散度矩阵，最小化类内散度矩阵。

$$\max_{w} \quad \frac{w^T S_b w}{w^T S_w w}$$

目标函数通常被成为广义瑞利商（generalized Rayleigh Quotient），最优解为广义特征值问题。

此模型的最优解有无穷多个（对最优解w进行任意缩放）。

线性判别分析

线性判别分析(LDA):

$$\max \quad \boldsymbol{w}^T \boldsymbol{S}_b \boldsymbol{w} \qquad \mathbf{s.t.} \quad \boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w} = \mathbf{1}$$

拉格朗日函数:

$$\mathcal{L}(\boldsymbol{w}, \boldsymbol{\lambda}) = -\boldsymbol{w}^T \boldsymbol{S}_b \boldsymbol{w} + \boldsymbol{\lambda}\left(\boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w} - \mathbf{1}\right)$$

$\mathcal{L}$ 对变量 $\boldsymbol{w}$ 求偏导:

$$\frac{\partial \mathcal{L}(\boldsymbol{w}, \boldsymbol{\lambda})}{\partial \boldsymbol{w}} = -2\boldsymbol{S}_b \boldsymbol{w} + 2\boldsymbol{\lambda} \boldsymbol{S}_w \boldsymbol{w}$$

极大极小问题（对偶问题）:

$$\max_{\boldsymbol{\lambda}} \ \min_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}, \boldsymbol{\lambda})$$

线性判别分析

极小问题：

$$\min_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w},\boldsymbol{\lambda}) = -\boldsymbol{w}^T\boldsymbol{S_b}\boldsymbol{w} + \boldsymbol{\lambda}\left(\boldsymbol{w}^T\boldsymbol{S_w}\boldsymbol{w} - \mathbf{1}\right)$$

$$= -\boldsymbol{w}^T\boldsymbol{\lambda}\boldsymbol{S_w}\boldsymbol{w} + \boldsymbol{\lambda}\left(\boldsymbol{w}^T\boldsymbol{S_w}\boldsymbol{w} - \mathbf{1}\right) = \boldsymbol{\lambda}$$

极大极小问题：

$$\max_{\boldsymbol{\lambda}} \boldsymbol{\lambda} \qquad \mathbf{s.t.} \quad \boldsymbol{S_b}\boldsymbol{w} = \boldsymbol{\lambda}\boldsymbol{S_w}\boldsymbol{w}$$

线性判别分析(LDA):

$$\max \ \boldsymbol{w}^T\boldsymbol{S_b}\boldsymbol{w} \qquad \mathbf{s.t.} \quad \boldsymbol{w}^T\boldsymbol{S_w}\boldsymbol{w} = \mathbf{1}$$

等价于 $\boldsymbol{S_b}\boldsymbol{w} = \boldsymbol{\lambda}\boldsymbol{S_w}\boldsymbol{w}$的最大特征值。

线性判别分析

线性判别分析(LDA):

$$\boldsymbol{S}_b\boldsymbol{w} = \lambda \boldsymbol{S}_w\boldsymbol{w}$$

常规方法是求矩阵 $\boldsymbol{S}_w^{-1}\boldsymbol{S}_b$ 的特征值问题！

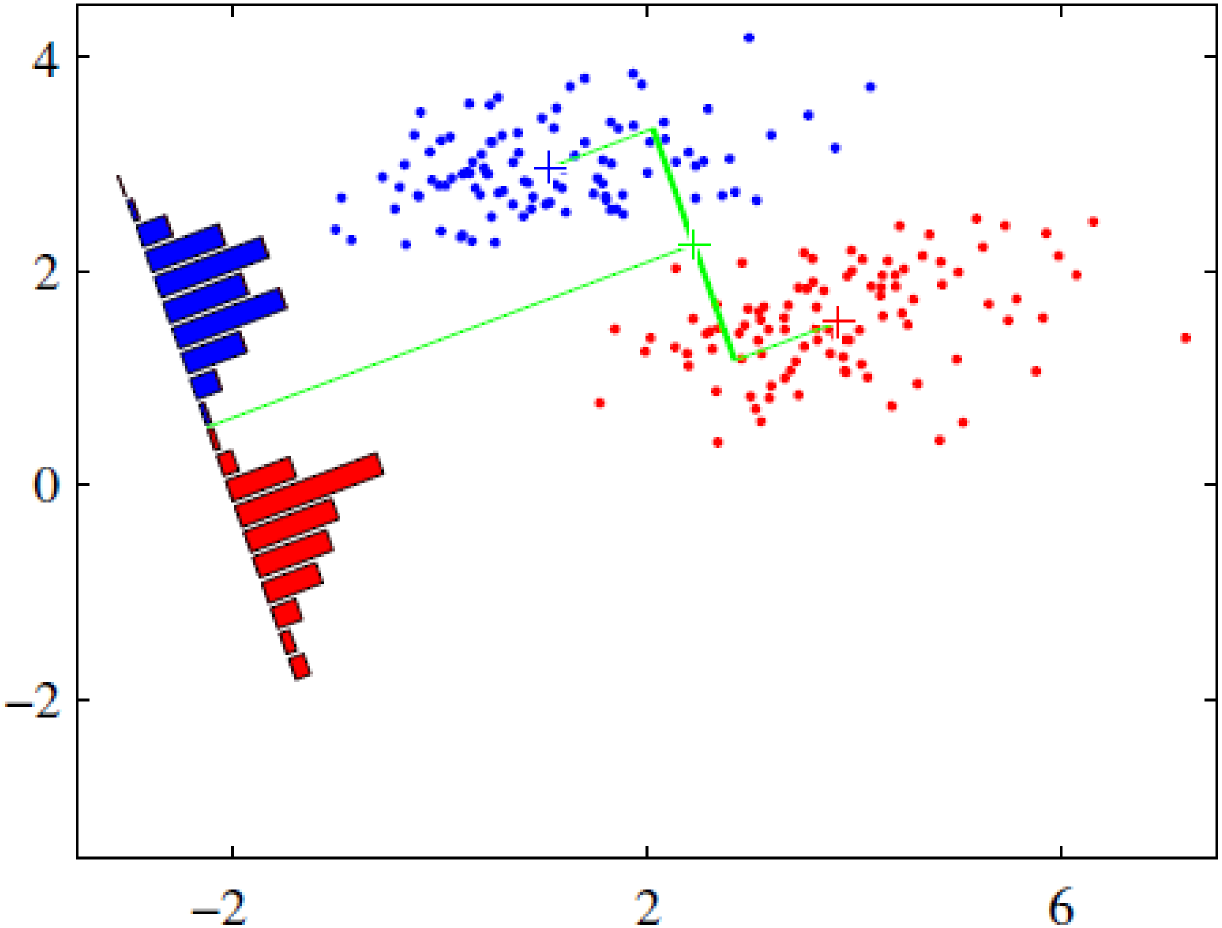
然而 $\boldsymbol{S}_b = (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T$, 我们有

$$\begin{aligned}\boldsymbol{S}_b\boldsymbol{w} &= (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T\boldsymbol{w} \\ &= ((\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T\boldsymbol{w})(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)\end{aligned}$$

原问题等价于求解:

$$\boldsymbol{S}_w\boldsymbol{w} = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$$

LDA降维示意图



LDA vs PCA

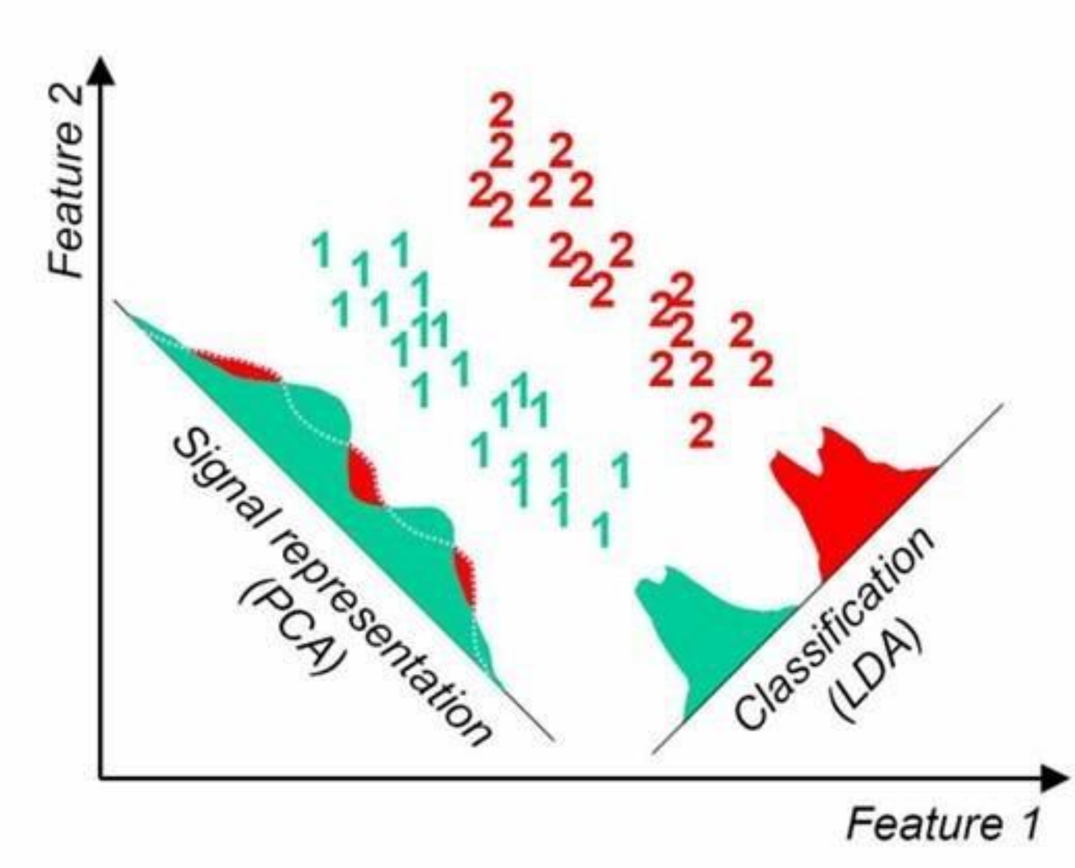
相同点：

- 都属于降维方法。
- 都转化为求解特征值问题。
- 都可以使用核化技巧。

不同点：

- PCA是非监督降维，LDA是监督降维。

LDA vs PCA


Feature 2
Feature 1
Signal representation
(PCA)
Classification
(LDA)

LDA --最小二乘法

当 $x \in X_i$ 时，令 $y = (-1)^i\, n/n_j$, 则

$$\sum_{i=1}^{n} y_i = n_0 \times n/n_0 - n_1 \times n/n_1 = 0$$

最小二乘法：

$$\operatorname*{argmin}_{w,b}\ \frac{1}{n}\sum_{i}\left[y_i - \left(w^T x_i + b\right)\right]^2$$

LDA --最小二乘法

目标函数对 w 和 b 求偏导得，

$\frac{1}{n}\sum_i(w^T x_i + b - y_i) = 0$，　　$\sum_i(w^T x_i + b - y_i)x_i = 0$

对第一个式子展开，

$$b = -\frac{1}{n}\sum_i w^T x_i = -w^T\mu = -\frac{1}{n}(n_0\mu_0 + n_1\mu_1)$$

把 b 代入第二个式子，并展开有

$$(S_w + \frac{n_0 n_1}{n} S_b)w = n(\mu_0 - \mu_1)$$

6.5 多分类LDA

陈飞宇

fchen@cqu.edu.cn

办公室：软件学院529

多分类LDA

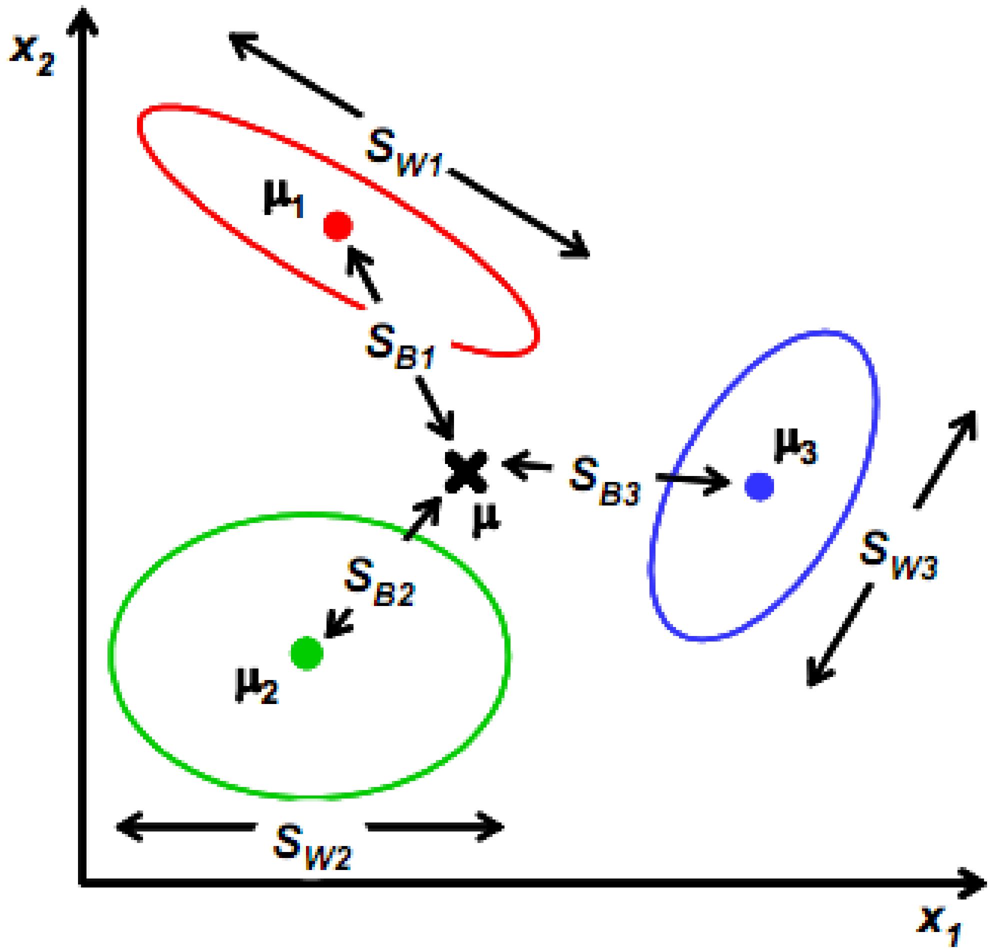
假设存在 k 个类，令$D_j, n_j, \mu_j, \Sigma_j$分别表示第 j 类样本的集合、均值向量和协方差矩阵。

类内散度矩阵（within-class scatter matrix）：

$$S_w = \sum_{j=1}^{k} \sum_{x \in D_j} (x - \mu_j)(x - \mu_j)^T$$

中心变成k个之后，如何定义类间散度矩阵？

多分类LDA



多分类LDA

类内散度矩阵（within-class scatter matrix）：

$$\boldsymbol{S}_w = \sum_{j=1}^{k} \sum_{\boldsymbol{x} \in D_j} (\boldsymbol{x} - \boldsymbol{\mu}_j)(\boldsymbol{x} - \boldsymbol{\mu}_j)^T$$

类间散度矩阵（between-class scatter matrix）：

$$\boldsymbol{S}_b = \sum_{j=1}^{k} n_j (\boldsymbol{\mu}_j - \boldsymbol{\mu})(\boldsymbol{\mu}_j - \boldsymbol{\mu})^T$$

全局散度矩阵（total scatter matrix）：

$$\boldsymbol{S}_t = \boldsymbol{S}_w + \boldsymbol{S}_b = \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T$$

全局散度矩阵 S_t

$$S_t = \sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T$$

$$\boxed{\mu = \frac{1}{n}\sum_{i=1}^{n} x_i}$$

$$= \sum_{i=1}^{n}(x_i x_i^T - x_i\mu^T - \mu x_i^T + \mu\mu^T)$$

$$= \sum_{i=1}^{n} x_i x_i^T - (\sum_{i=1}^{n} x_i)\mu^T - \mu(\sum_{i=1}^{n} x_i)^T + \sum_{i=1}^{n}\mu\mu^T$$

$$= (\sum_{i=1}^{n} x_i x_i^T) - n\mu\mu^T$$

类内散度矩阵 S_w

$$S_w=\sum_{j=1}^{k}\sum_{x\in D_j}(x-\mu_j)(x-\mu_j)^T$$

$$\boxed{\mu_j=\frac{1}{n_j}\sum_{x\in D_j}x}$$

$$=\sum_{j=1}^{k}\sum_{x\in D_j}(xx^T-x\mu_j^T-\mu_jx^T+\mu_j\mu_j^T)$$

$$=\sum_{j=1}^{k}[\sum_{x\in D_j}xx^T-\left(\sum_{x\in D_j}x\right)\mu_j^T-\mu_j\left(\sum_{x\in D_j}x\right)^T+\sum_{x\in D_j}\mu_j\mu_j^T)]$$

$$=\sum_{j=1}^{k}[\sum_{x\in D_j}xx^T-n_j\mu_j\mu_j^T]=(\sum_{i=1}^{n}x_ix_i^T)-\sum_{j=1}^{k}n_j\mu_j\mu_j^T$$

类间散度矩阵 S_b

令 $X=[x_1,x_2,\dots,x_n]$, H是标准化的示性矩阵。

$$\sum_{i=1}^{n} x_i x_i^T = XX^T$$

$$\sum_{j=1}^{k} n_j \mu_j \mu_j^T = \sum_{j=1}^{k} (\sqrt{n_j}\mu_j)(\sqrt{n_j}\mu_j)^T = \sum_{j=1}^{k} (Xh_j)(Xh_j)^T$$

$$= \sum_{j=1}^{k} Xh_j h_j^T X^T = X\left(\sum_{j=1}^{k} h_j h_j^T\right)X^T = XHH^TX^T$$

散度矩阵

$$S_t = (\sum_{i=1}^{n} x_i x_i^T) - n\mu\mu^T = XX^T - n\mu\mu^T$$

$$\boxed{\mu = \frac{1}{n}\sum_{j=1}^{k} n_j \mu_j}$$

$$S_w = (\sum_{i=1}^{n} x_i x_i^T) - \sum_{j=1}^{k} n_j \mu_j \mu_j^T = XX^T - XHH^T X^T$$

$$S_b = S_t - S_w = \sum_{j=1}^{k} n_j \mu_j \mu_j^T - n\mu\mu^T = \sum_{j=1}^{k} n_j \left(\mu_j \mu_j^T - \mu\mu^T\right)$$

$$= \sum_{j=1}^{k} n_j (\mu_j - \mu)(\mu_j - \mu)^T = XHH^T X^T - n\mu\mu^T$$

散度矩阵 $\boldsymbol{S}$

假设 $\boldsymbol{\mu} = \mathbf{0}, \boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n], \boldsymbol{H}$是标准化的示性矩阵。

全局散度矩阵: $\boldsymbol{S}_t = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T = \boldsymbol{X}\boldsymbol{X}^T$

类间散度矩阵: $\boldsymbol{S}_b = \sum_{j=1}^{k} n_j \boldsymbol{\mu}_j \boldsymbol{\mu}_j^T = \boldsymbol{X}\boldsymbol{H}\boldsymbol{H}^T\boldsymbol{X}^T$

类内散度矩阵: $\boldsymbol{S}_w = \boldsymbol{S}_t - \boldsymbol{S}_b = \boldsymbol{X}(\boldsymbol{I} - \boldsymbol{H}\boldsymbol{H}^T)\boldsymbol{X}^T$

全局散度矩阵 $\boldsymbol{S_t}$

当 $k=2$时，

$$\mu = \frac{1}{n}\sum_{j=1}^{k} n_j \mu_j$$

$$S_b = \sum_{j=1}^{k} n_j\left(\mu_j\mu_j^T - \mu\mu^T\right)$$

$$= n_1\left(\mu_1\mu_1^T - \mu\mu^T\right) + n_2\left(\mu_2\mu_2^T - \mu\mu^T\right)$$

$$= \frac{n_1 n_2}{n_1+n_2}(\mu_1-\mu_2)(\mu_1-\mu_2)^T$$

多分类LDA

多分类LDA需要求解投影矩阵$\boldsymbol{W}$，同时最大化类间散度矩阵，最小化类内散度矩阵。

$$\max_{\boldsymbol{W}} \frac{\mathrm{Tr}(\boldsymbol{W}^T \boldsymbol{S}_b \boldsymbol{W})}{\mathrm{Tr}(\boldsymbol{W}^T \boldsymbol{S}_w \boldsymbol{W})} \quad \mathrm{s.t.} \quad \boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{I}$$

其中，Tr函数表示矩阵主对角线上的元素之和，即 $\mathrm{Tr}(\boldsymbol{A}) = \sum_{i=1}^{n} \boldsymbol{A}_{ii}$，正交约束是因为每个投影方向互相垂直！

此模型的最优解为矩阵$\boldsymbol{S}_w^{-1}\boldsymbol{S}_b$的前d个最大特征值所对应的特征向量组成的矩阵。

6.6 非负矩阵分解

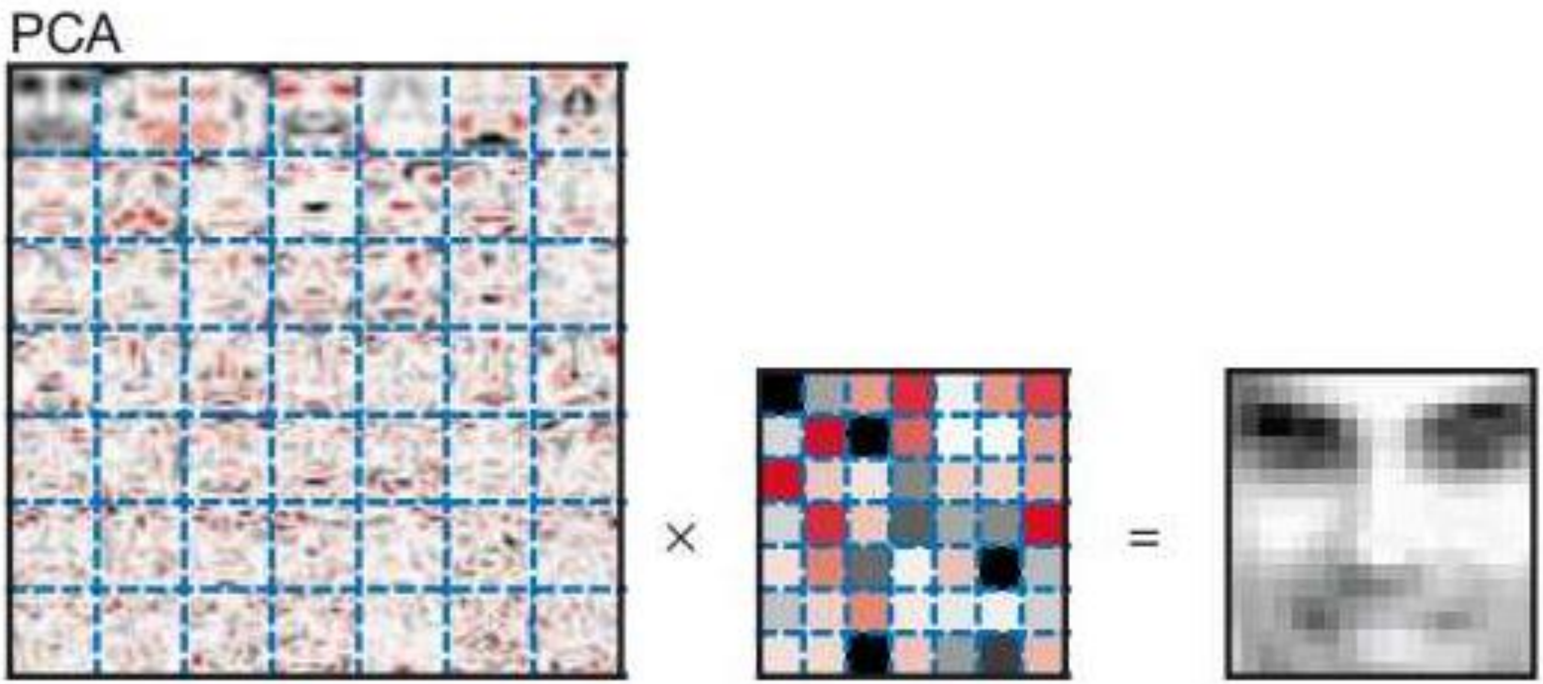
陈飞宇

fchen@cqu.edu.cn

办公室：软件学院529

PCA和LDA的局限性

PCA和LDA的投影矩阵都是正交的，投影之后结果很可能出现负数，不可解释！



非负矩阵分解

非负矩阵分解模型（**NMF**）：

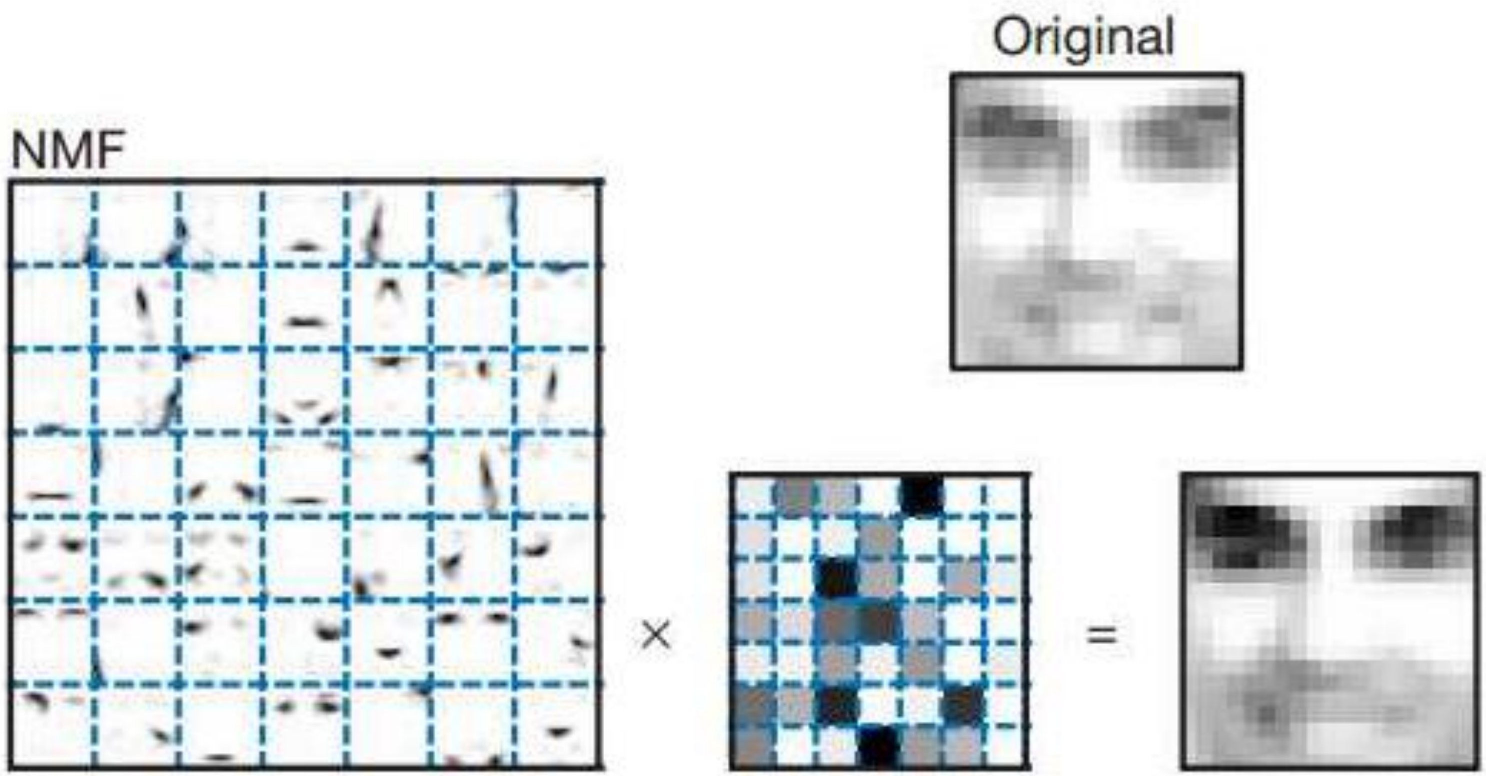
$$\min \quad \|\boldsymbol{X} - \boldsymbol{W}\boldsymbol{H}\|_F^2$$

$$\mathbf{s.t.} \quad \boldsymbol{W} \geq \mathbf{0}, \boldsymbol{H} \geq \mathbf{0}$$

- **把数据矩阵 $\boldsymbol{X}$ 分解成两个小规模矩阵的乘积，本质上说是寻找数据$\boldsymbol{X}$的低秩逼近。**
- **基矩阵和系数矩阵的非负约束使得模型更有意义。**
- **NMF是非凸的模型，目标函数是非凸的。**

非负矩阵分解

非负矩阵分解模型（NMF）：



交替最小二乘法

非负矩阵分解模型（NMF）：

$$\min \quad \|\boldsymbol{X} - \boldsymbol{W}\boldsymbol{H}\|_F^2$$

$$\text{s.t.} \quad \boldsymbol{W} \geq \mathbf{0}, \boldsymbol{H} \geq \mathbf{0}$$

迭代公式:

$$\boldsymbol{W}^+ = \max(\mathbf{0}, (\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}\boldsymbol{X})$$

$$\boldsymbol{H}^+ = \max(\mathbf{0}, \boldsymbol{X}\boldsymbol{H}^T(\boldsymbol{H}\boldsymbol{H}^T)^{-1})$$

乘积更新算法

非负矩阵分解模型（NMF）：

$$\min \quad \|\boldsymbol{X} - \boldsymbol{W}\boldsymbol{H}\|_F^2$$

$$\text{s.t.} \quad \boldsymbol{W} \geq \mathbf{0}, \boldsymbol{H} \geq \mathbf{0}$$

迭代公式:

$$\boldsymbol{W}^+ = \boldsymbol{W} \odot [\boldsymbol{X}\boldsymbol{H}^T \oslash (\boldsymbol{W}\boldsymbol{H}\boldsymbol{H}^T + \boldsymbol{\varepsilon})]$$

$$\boldsymbol{H}^+ = \boldsymbol{H} \odot [\boldsymbol{W}^T\boldsymbol{X} \oslash (\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{H} + \boldsymbol{\varepsilon})]$$