

机器学习

Machine Learning

北京航空航天大学计算机学院

School of Computer Science and Engineering, Beihang University

刘庆杰 陈佳鑫

2025年春季学期
Spring 2024

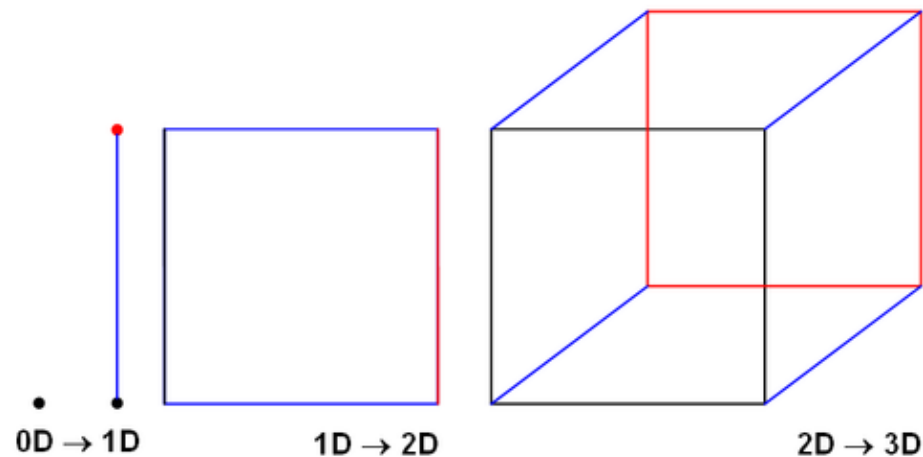
8.1 什么是降维？

- 降维的基本定义
- 为什么要降维？
- 降维的典型应用
- 代表性降维方法

降维的基本定义

● 维数 (又称维度)

- 数学中：独立参数的数目
- 物理中：独立时空坐标的数目



点是0维

直线是1维

平面是2维

体是3维



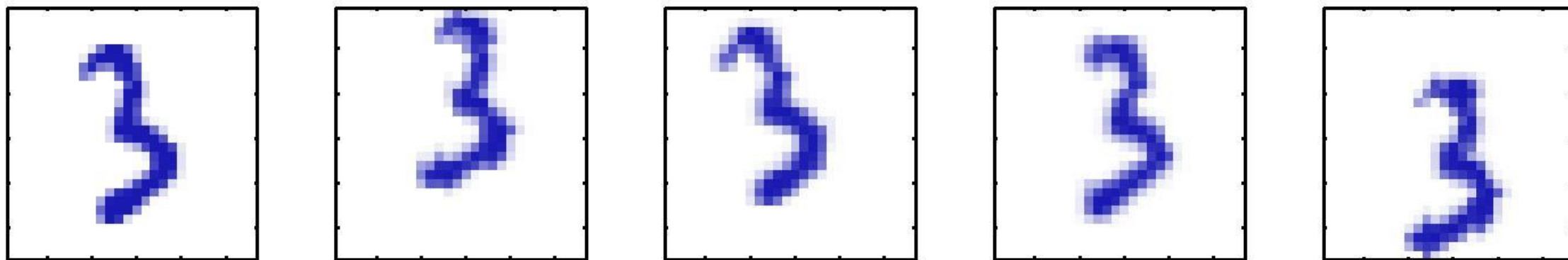
在点上定位一个点，不需要参数

在直线上定位一个点，需要1个参数

在平面上定位一个点，需要2个参数

在体上定位一个点，需要3个参数

降维的基本定义

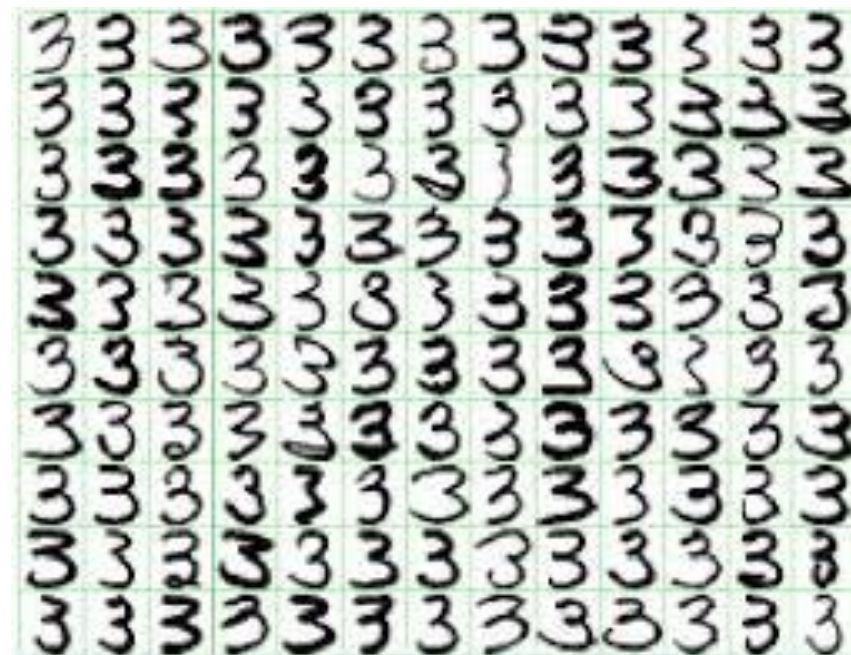


64×64像素数字放入100×100像素白板

● 维数

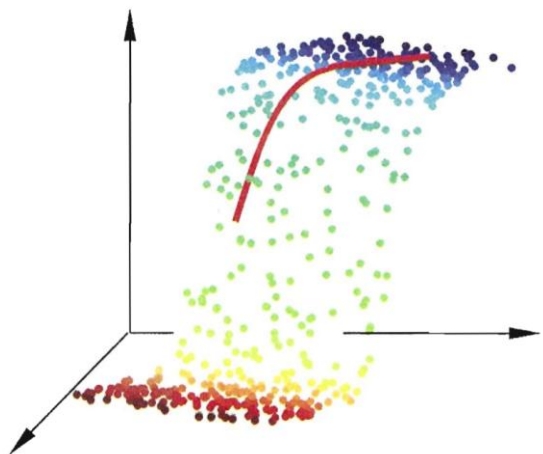
- 水平/垂直的平移变化
- 旋转变换
- 尺度变化
- 形状变化(不同人的写作习惯)
- 光照变化
- ...

隐变量
(Latent Variables)

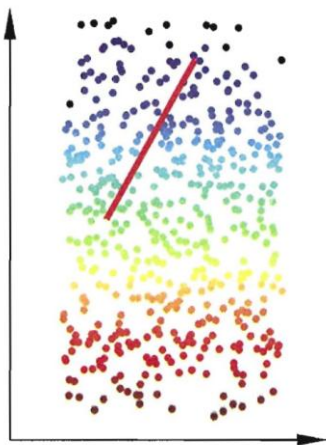


降维的基本定义

- 在高维情形下出现的数据样本稀疏，距离计算困难等问题，是所有机器学习方法共同面临的严重障碍，被称为“**维数灾难**”
- 降维 (Dimensionality Reduction)
 - 将高维数据转换为低维数据的技术，同时尽量保留原始数据的重要信息



三维空间中观察到的样本点



二维空间中的曲面

观测收集的数据样本虽然是高维的，但与学习任务密切相关的也许仅是某个低维分布，即高维空间中的一个**低维嵌入**。

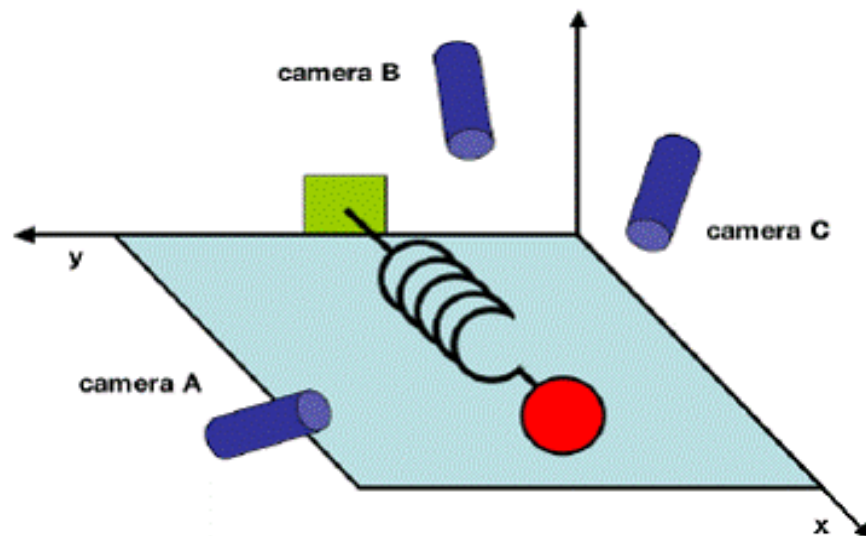
为什么要降维？

● 为什么要降维？

- 在原始的高维空间中，包含冗余信息和噪声信息，会在实际应用中引入误差，影响准确率
- 降维可以提取数据内部的本质结构，减少冗余信息和噪声信息造成的误差，提高算法效率

● 一个简单的例子

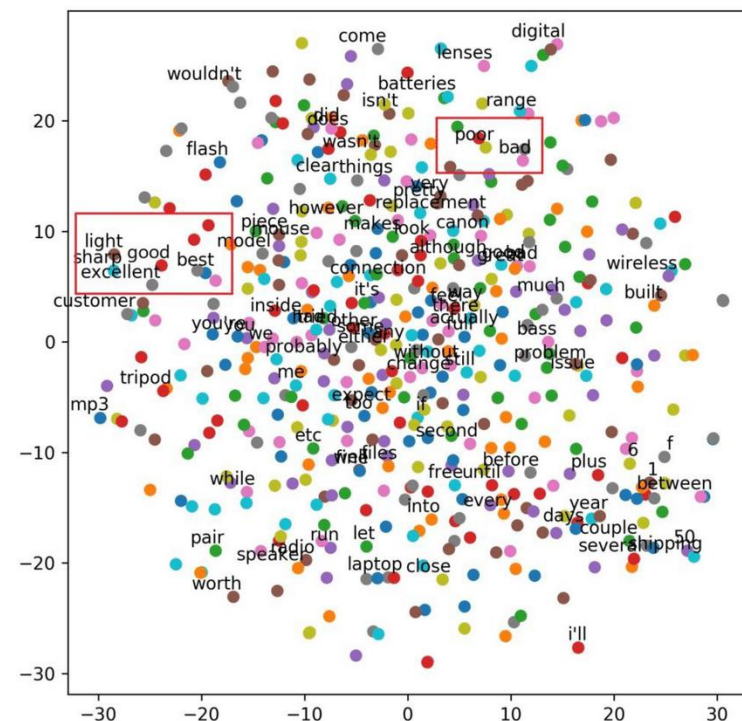
- 沿 x 轴拉小球
- 高维观测存在冗余信息



降维的典型应用

● 高维特征可视化分析

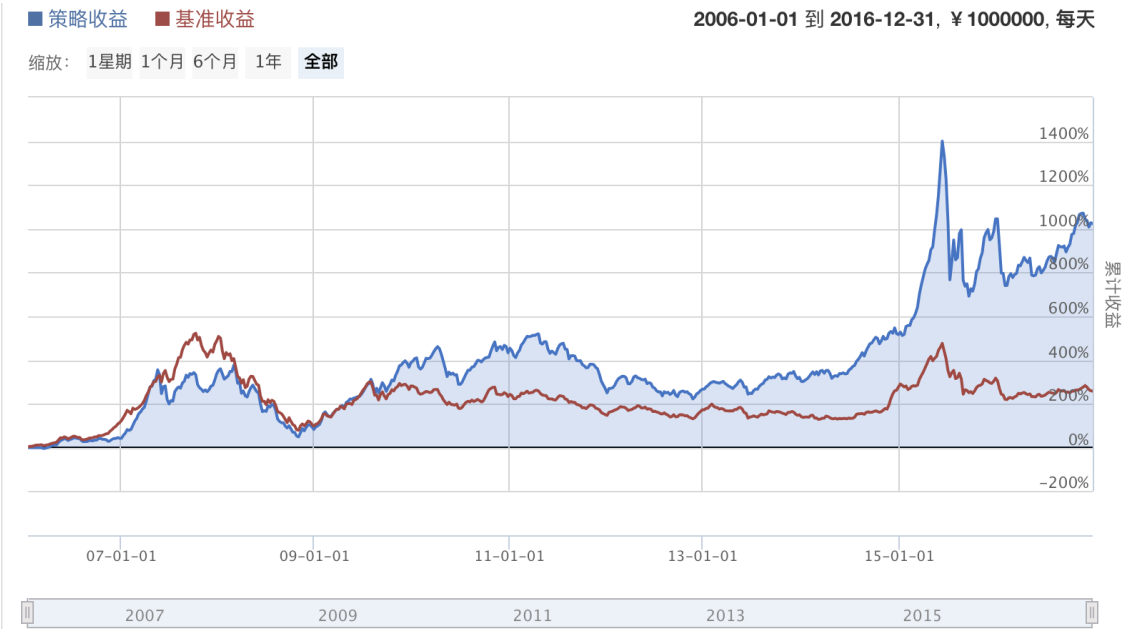
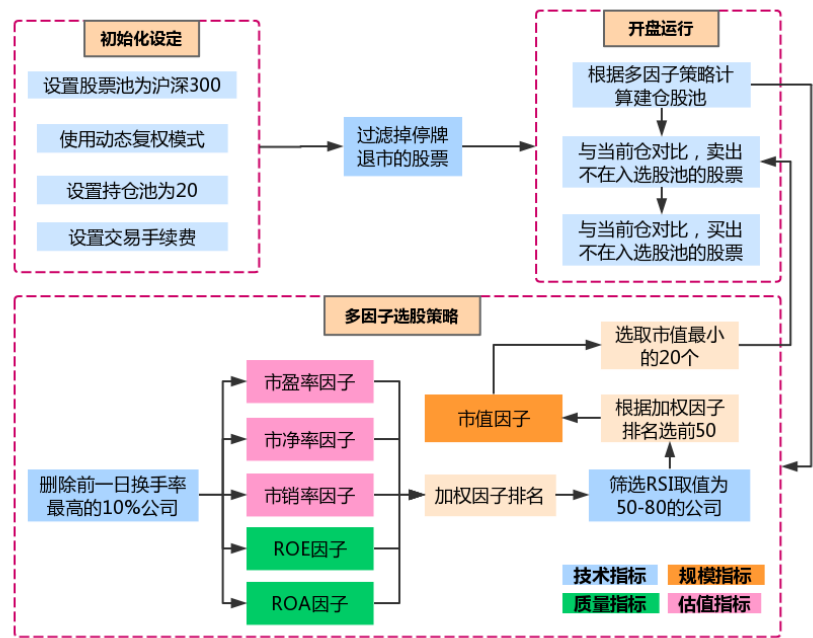
- 人类无法直观理解高维度特征嵌入，通过降维算法将高维特征投影至二维空间可视化分析，观察到相似内容的图像、文本在低维可视化空间中具备更相近的空间位置，便于高维特征的直观理解与科学分析。



降维的典型应用

● 金融数据分析

➤ 从大量的金融数据中，降维算法可以识别和提取出主要的、潜在的驱动因子，这些因子能够解释资产价格变动的大部分原因。通过该分析方法，投资者可以更准确地理解市场动态，预测未来趋势，以及优化投资策略。



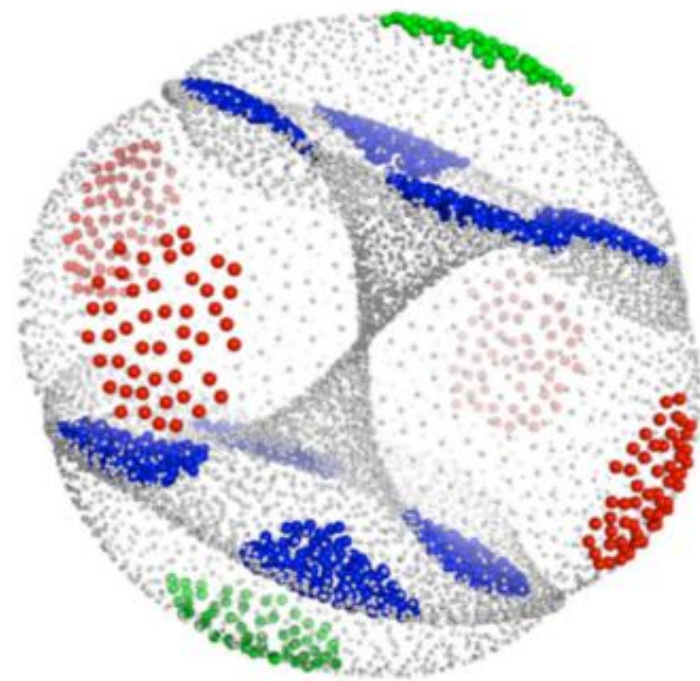
代表性降维方法

- 线性方法

- 主成分分析 (Principal Component Analysis, PCA)
- 因子分析 (Factor Analysis, FA)
- 独立成分分析 (Independent Component Analysis, ICA)

- 非线性方法

- 等距映射 (Isometric Mapping, Isomap)
- 局部线性嵌入 (Locally Linear Embedding, LLE)



8.2 主成分分析

- 主成分分析的概述
- 主成分分析的推导
- 主成分分析的应用
- 概率主成分分析
- 核主成分分析

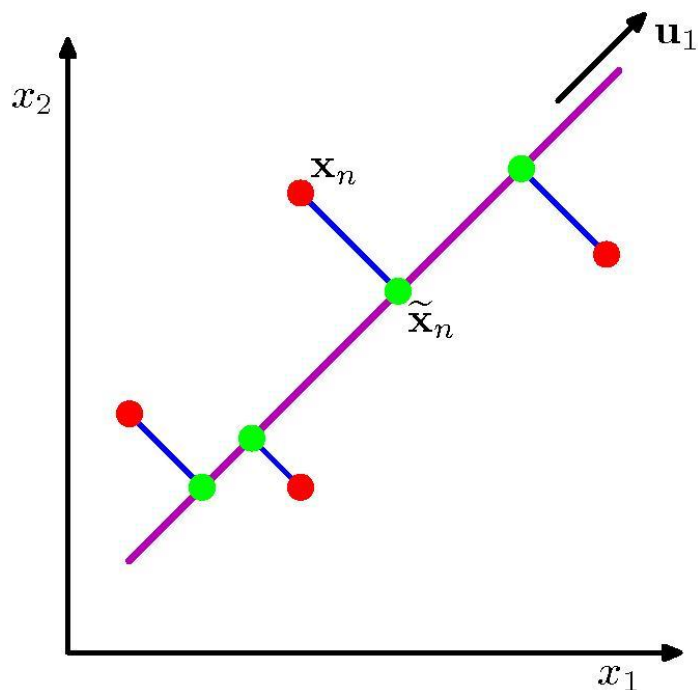
主成分分析

● 什么是主成分分析？

【1901年K. Pearson提出，针对非随机变量】

【1933年H. Hotelling提出，推广到随机向量】

➤ 主成分分析（Principal Component Analysis, PCA）是最常用的一种降维方法。PCA将原有众多具有一定相关性的指标重新组合成一组少量互相无关的综合指标。



- 原数据
- 降维后数据
- u_1 投影方向

使得降维后数据的方差尽可能大，即最大化绿色点的方差

使得降维后数据的均方误差尽可能小，即最小化蓝色线的平方和

主成分分析的推导-最大方差思想

● 最大方差思想

- 使用较少的数据维度保留住较多的原数据特性
- 将 D 维数据集 $\{x_n\}, n = 1, 2, \dots, N$ 降为 M 维, $M < D$
- 首先考虑 $M = 1$, 定义这个空间的投影方向为 D 维向量 u_1

出于方便且不失一般性, 令 $u_1^T u_1 = 1$

每个数据点 x_n 在新空间中表示为标量 $u_1^T x_n$

样本均值在新空间中表示为 $u_1^T \bar{x}$, 其中 $\bar{x} = \frac{1}{N} \sum_{n=1}^N x_n$

投影后样本方差表示为: $\frac{1}{N} \sum_{n=1}^N \{u_1^T x_n - u_1^T \bar{x}\}^2 = \boxed{u_1^T S u_1}$ 最大

其中原样本方差为: $S = \frac{1}{N} \sum_{n=1}^N (x_n - \bar{x})(x_n - \bar{x})^T$

主成分分析的推导-最大方差思想

● 最大方差思想

- 使用较少的数据维度保留住较多的原数据特性
- 目标是最大化 $u_1^T S u_1$ ，使得 $u_1^T u_1 = 1$
- 利用拉格朗日乘子法得到： $u_1^T S u_1 + \lambda_1(1 - u_1^T u_1)$
- 对 u_1 求导并置零得到： $S u_1 = \lambda_1 u_1$

u_1 是 S 的特征向量

进一步得到： $u_1^T S u_1 = \lambda_1$

u_1 是 S 最大特征值对应的特征向量时
方差取到极大值，称 u_1 为第一主成分

主成分分析的推导-最大方差思想

● 最大方差思想

- 使用较少的数据维度保留住较多的原数据特性
- 考虑更一般性的情况（ $M > 1$ ），新空间中数据方差最大的最佳投影方向由协方差矩阵 S 的 M 个特征向量 u_1, u_2, \dots, u_M 定义, 其分别对应 M 个最大的特征值 $\lambda_1, \lambda_2, \dots, \lambda_M$

首先获得方差最大的1维，生成该维的补空间；

继续在补空间中获得方差最大的1维，生成新的补空间；

依次循环下去得到 M 维的空间。

主成分分析的推导-最小均方误差思想

● 最小均方误差思想

- 使原数据与降维后的数据(在原空间中的重建)的误差最小
- 定义一组正交的 D 维基向量 $\{u_i\}, i = 1, 2, \dots, D$ ，满足

$$u_i^T u_j = \delta_{ij} \quad \text{当 } i = j \text{ 时, } \delta_{ij} = 1; \text{ 当 } i \neq j \text{ 时, } \delta_{ij} = 0。$$

由于基是完全的，每个数据点可以表示为基向量的线性组合

$$x_n = \sum_{i=1}^D \alpha_{ni} u_i$$

相当于进行了坐标变换

$$\{x_{n1}, x_{n2}, \dots, x_{nD}\} \xrightarrow{\text{变换到}} \{\alpha_{n1}, \alpha_{n2}, \dots, \alpha_{nD}\}$$

主成分分析的推导-最小均方误差思想

● 最小均方误差思想

➤ 使原数据与降维后的数据(在原空间中的重建)的误差最小

➤ 根据正交归一性得到: $\alpha_{ni} = x_n^T u_i$, 那么 $x_n = \sum_{i=1}^D (x_n^T u_i) u_i$

➤ 在 M 维变量 ($M < D$) 生成的空间中对其进行表示

$$\tilde{x}_n = \sum_{i=1}^M z_{ni} u_i + \sum_{i=M+1}^D b_i u_i$$

独特的 共享的

目标最小化失真度 $J = \frac{1}{N} \sum_{n=1}^N \|x_n - \tilde{x}_n\|^2$

求导并置零得到: $z_{nj} = x_n^T u_j, j = 1, 2, \dots, M$

$$b_j = \bar{x}^T u_j, j = M + 1, M + 2, \dots, D$$

主成分分析的推导-最小均方误差思想

● 最小均方误差思想

➤ 使原数据与降维后的数据(在原空间中的重建)的误差最小

➤ 有 $x_n - \tilde{x}_n = \sum_{i=M+1}^D \{(x_n - \bar{x})^T u_i\} u_i$

$$J = \frac{1}{N} \sum_{n=1}^N \sum_{i=M+1}^D (x_n^T u_i - \bar{x}^T u_i)^2 = \sum_{i=M+1}^D u_i^T S u_i$$

利用拉格朗日乘子法得到: $\tilde{J} = \sum_{i=M+1}^D u_i^T S u_i + \sum_{i=M+1}^D \lambda_i (1 - u_i^T u_i)$

求导并置零得到: $S u_i = \lambda_i u_i$

对应失真度为 $J = \sum_{i=M+1}^D \lambda_i$

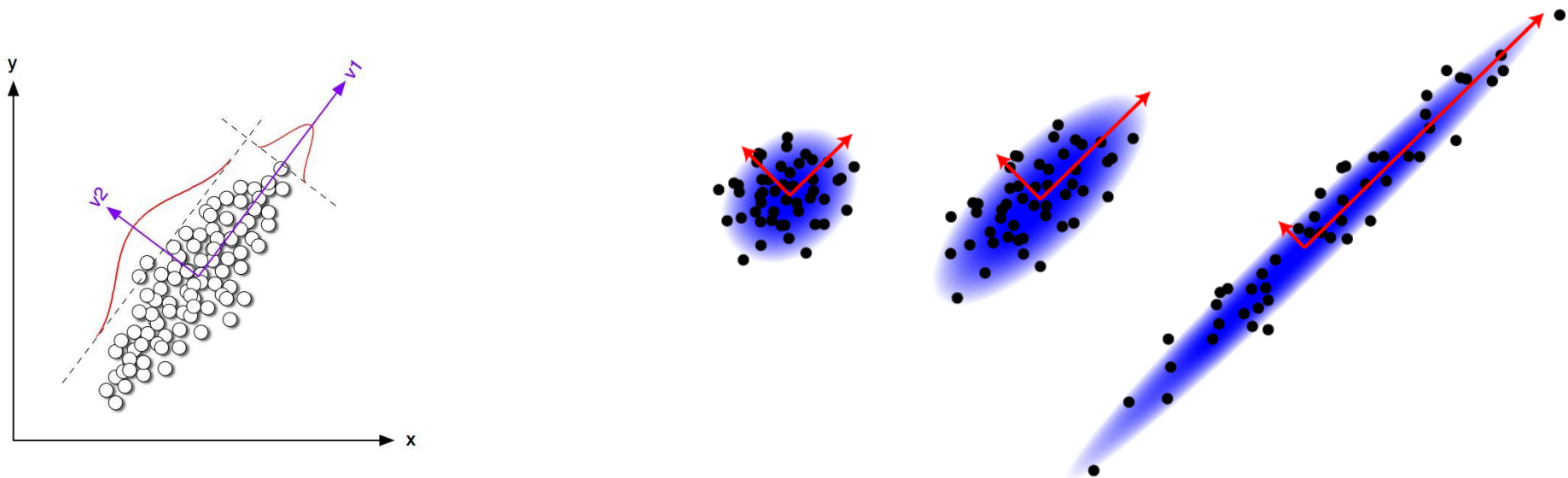
J 最小时取 $D-M$ 个最小的特征值

主子空间对应 M 个最大特征值

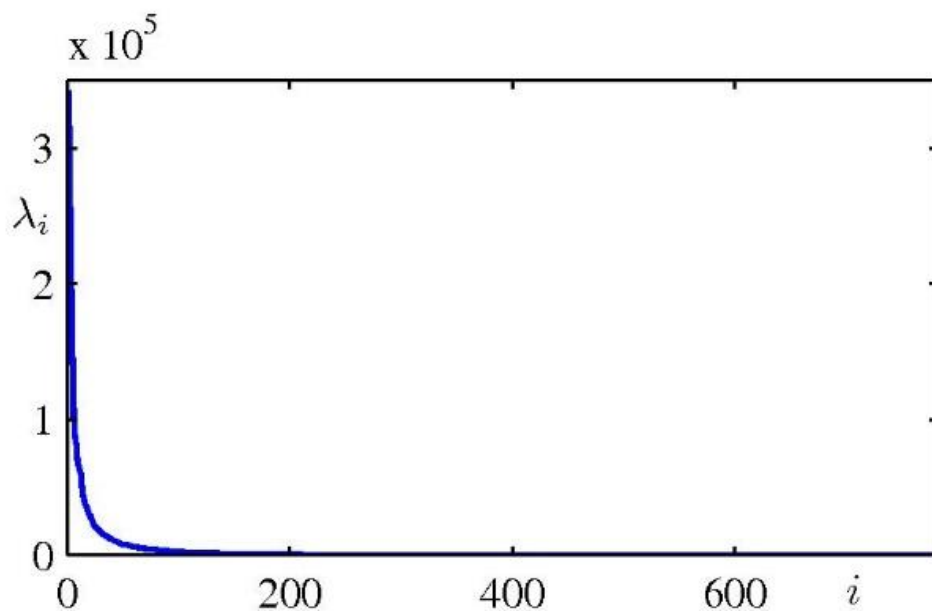
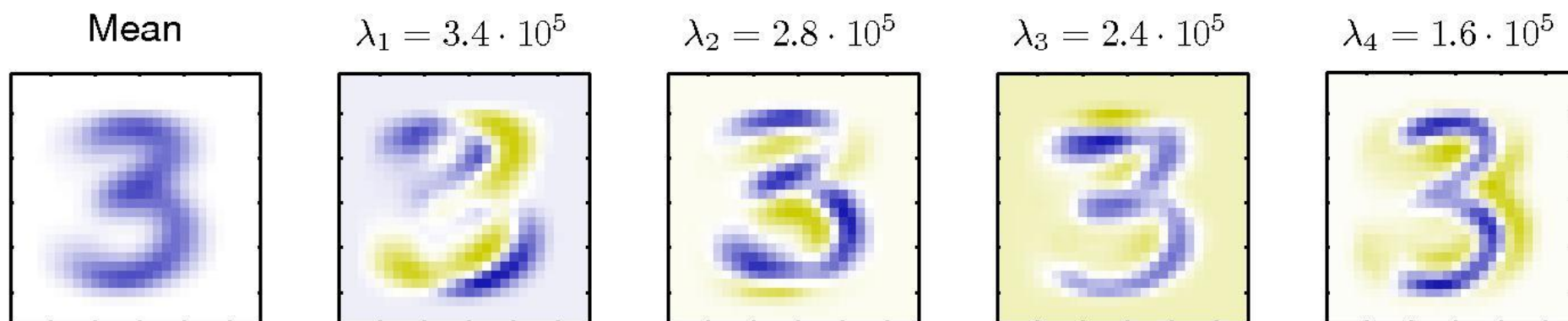
主成分分析的计算步骤

● 计算步骤

- **步骤1.** 计算给定样本 $\{x_n\}, n = 1, 2, \dots, N$ 的均值 \bar{x} 和协方差矩阵 S ;
- **步骤2.** 计算 S 的特征向量与特征值;
- **步骤3.** 将特征值从大到小排列, 前 M 个特征值 $\lambda_1, \lambda_2, \dots, \lambda_M$ 所对应的特征向量 u_1, u_2, \dots, u_M 构成投影矩阵。



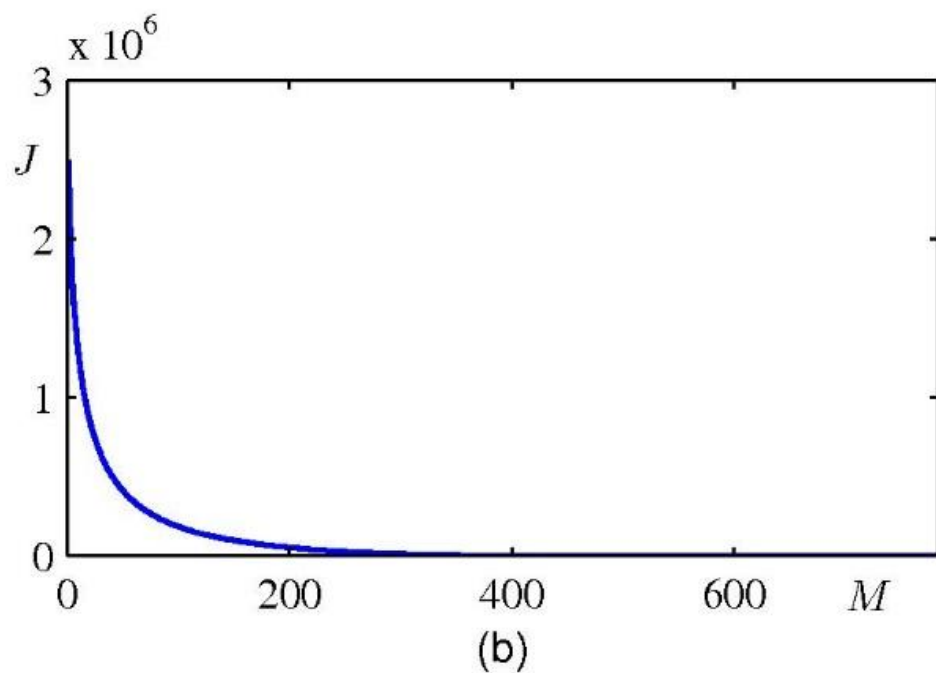
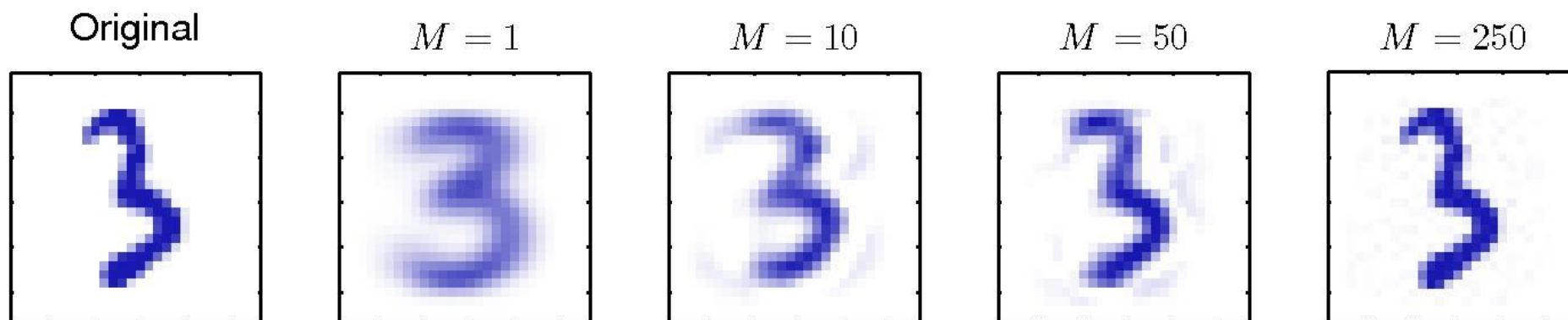
主成分分析的应用



(a)

特征值分布谱
特征值由大到小排列

主成分分析的应用



失真度分布谱
随 M 取值由小到大排列

主成分分析的应用



特征脸(Eigenfaces)#1~#8



主成分分析的应用



特征脸(Eigenfaces)#101~#108



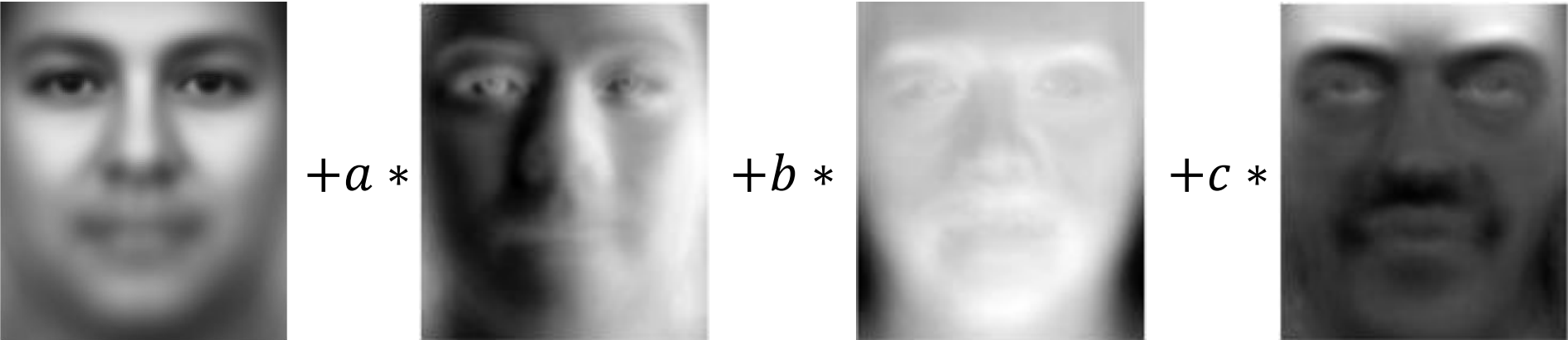
主成分分析的应用



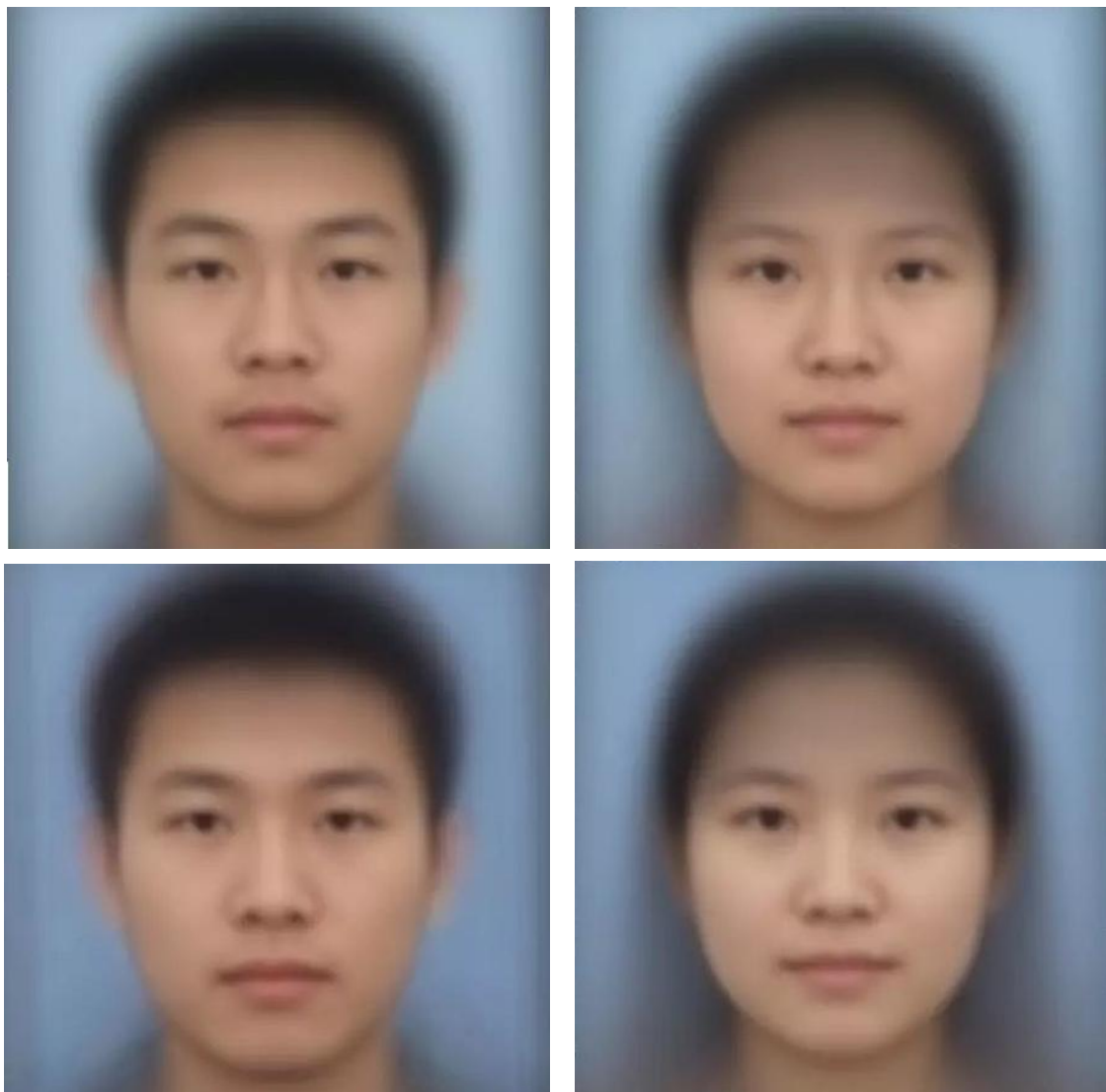
特征脸(Eigenfaces)#501~#508



主成分分析的应用

$$\begin{array}{ccccccc} & = & & +a * & & +b * & & +c * & & + \dots \\ \text{样本} & & \text{特征脸} & & \text{特征脸\#1} & & \text{特征脸\#2} & & \text{特征脸\#3} \\ & & \text{Mean} & & & & & & \end{array}$$


主成分分析的应用



主成分分析的应用

● 利用PCA处理高维数据

➤ 在实际应用中，样本维数可能很高，远大于样本的个数在人脸识别中，1000张人脸图像，每张图像 100×100 像素

➤ D 维空间， N 个样本点， $N < D$

➤ **定义：** \mathbf{X} 是 $N \times D$ 维的数据矩阵，其行向量为 $(\mathbf{X}_n - \bar{\mathbf{X}})^T$

$$\mathbf{S} = \frac{1}{N} \sum_{n=1}^N (\mathbf{X}_n - \bar{\mathbf{X}})(\mathbf{X}_n - \bar{\mathbf{X}})^T \quad \text{可以写为} \quad \mathbf{S} = N^{-1} \mathbf{X}^T \mathbf{X}$$

\mathbf{S} 的维数? $\longrightarrow D \times D$ 维 $\longrightarrow 10000 \times 10000$

$$\frac{1}{N} \mathbf{X}^T \mathbf{X} \mathbf{u}_i = \lambda_i \mathbf{u}_i \quad \longrightarrow \quad \frac{1}{N} \mathbf{X} \mathbf{X}^T \mathbf{X} \mathbf{u}_i = \lambda_i (\mathbf{X} \mathbf{u}_i)$$

令 $\mathbf{v}_i = \mathbf{X} \mathbf{u}_i$ ，得到 $\boxed{\frac{1}{N} \mathbf{X} \mathbf{X}^T} \mathbf{v}_i = \lambda_i \mathbf{v}_i$ $N \times N$ 维

主成分分析的应用

● 利用PCA处理高维数据

➤ 在实际应用中，样本维数可能很高，远大于样本的个数在人脸识别中，1000张人脸图像，每张图像 100×100 像素

➤ 对 $\frac{1}{N} \mathbf{X} \mathbf{X}^T$ 求的**特征值** λ_i 和**特征向量** v_i

$$\frac{1}{N} \mathbf{X} \mathbf{X}^T v_i = \lambda_i v_i \quad \longrightarrow \quad \frac{1}{N} \mathbf{X}^T \mathbf{X} (\mathbf{X}^T v_i) = \lambda_i (\mathbf{X}^T v_i) \quad \text{S的特征向量}$$

调整 $u_i \propto \mathbf{X}^T v_i$ 的尺度，令其满足 $\|u_i\| = 1$

$$\longrightarrow u_i = \frac{1}{N \lambda_i^{1/2}} \mathbf{X}^T v_i$$

奇异值分解(Singular Value Decomposition, SVD)

概率主成分分析

● PCA的概率表示

➤ 隐藏变量 z 以如下形式产生 D 维观测变量 x

$$X = Wz + \mu + \epsilon \longrightarrow \text{高斯噪声, 均值为0, 协方差为 } \sigma^2 I$$

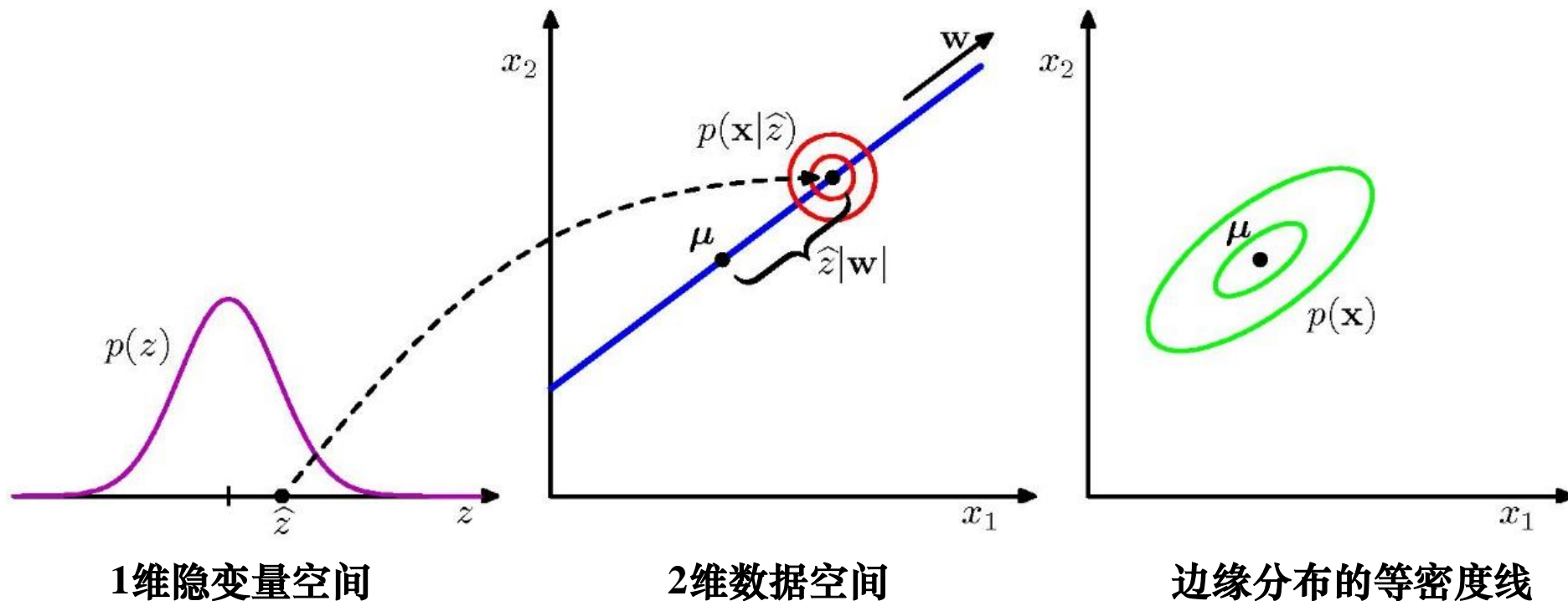
其中, z 为 M 维的隐藏变量, 且满足高斯分布

$$p(z) = \mathcal{N}(z|0, I)$$

x 以 z 为条件的分布也满足高斯

$$p(x|z) = \mathcal{N}(x|Wz + \mu, \sigma^2 I)$$

概率主成分分析



从隐空间到数据空间的映射，与PCA的传统视角相反
从数据空间到隐空间的映射，可以由贝叶斯定理得到

概率主成分分析

- PCA的概率表示

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$$

$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \sigma^2\mathbf{I})$$



$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z} = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{C})$$

$$\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}$$

$$\mathbb{E}[\mathbf{X}] = \mathbb{E}[\mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\varepsilon}] = \boldsymbol{\mu}$$

$$\text{cov}[\mathbf{X}] = \mathbb{E}[(\mathbf{W}\mathbf{z} + \boldsymbol{\varepsilon})(\mathbf{W}\mathbf{z} + \boldsymbol{\varepsilon})^T]$$

$$= \mathbb{E}[\mathbf{W}\mathbf{z}\mathbf{z}^T\mathbf{W}^T] + \mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T] = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}$$

概率主成分分析

● PCA的概率表示

最大似然估计求解

给定 $\mathbf{X} = \{x_n\}$, 求其对数似然函数

$$\begin{aligned}\ln p(\mathbf{X}|\mu, \mathbf{W}, \sigma^2) &= \sum_{n=1}^N \ln p(x_n|\mu, \mathbf{W}, \sigma^2) \\ &= -\frac{ND}{2} \ln(2\pi) - \frac{ND}{2} \ln|\mathbf{C}| - \frac{1}{2} \sum_{n=1}^N (x_n - \mu)^T \mathbf{C}^{-1} (x_n - \mu)\end{aligned}$$

对 μ 求导置零并代回



$$= -\frac{N}{2} \{D \ln(2\pi) - \ln|\mathbf{C}| - \text{Tr}(\mathbf{C}^{-1} \mathbf{S})\}$$

→ 协方差矩阵

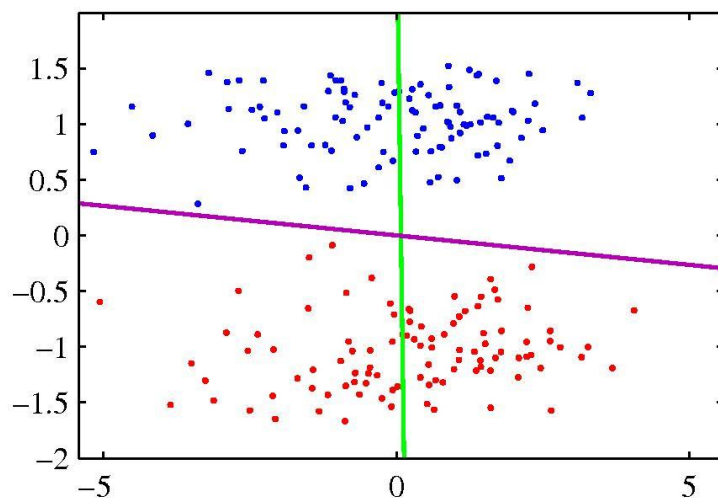
→ $\mathbf{W}_{\text{ML}} = \mathbf{U}_M (\mathbf{L}_M - \sigma^2 \mathbf{I})^{1/2} \mathbf{R}$

$$\sigma_{\text{ML}}^2 = \frac{1}{D - M} \sum_{i=M+1}^D \lambda_i$$

讨论

● PCA v.s. LDA

- **PCA**追求降维后能够最大化保持数据内在信息，并通过衡量在投影方向上的数据方差来判断其重要性。但这对数据的**区分作用并不大**，反而可能使得数据点混杂在一起
- **LDA**所追求的目标与PCA不同，不是希望保持数据最多的信息，而是希望数据在降维后能够很**容易地被区分开**



- ● 数据点 (在LDA中有标注)
- PCA的投影方向
- LDA的投影方向

讨论

● PCA的优点

- 在具有很**高普适性**，**最大程度地保持**了原有数据的**信息**
- 可**对主成分的重要性**进行**排序**，并根据需要略去部分维数，达到降维从而简化模型或对数据进行压缩的效果
- 完全**无参数限制**，在计算过程中不需要人为设定参数或是根据任何经验模型对计算进行干预，最终结果只与数据相关

● PCA的局限性

- 假设模型是**线性的**，也就决定了它能进行的主成分分析之间的关系也是线性的
- 假设数据具有**较高信噪比**，具有最高方差的一维向量被看作是主成分，而方差较小的变化被认为是噪声

核主成分分析

● Kernel PCA

➤ 将主成分分析的线性假设一般化使之适应非线性数据

➤ 传统PCA: D 维样本 $\{x_n\}, n = 1, 2, \dots, N$, $\sum_n x_n = 0$ 中心化

$$Su_i = \lambda_i u_i, \text{ 其中 } S = \frac{1}{N} \sum_{n=1}^N x_n x_n^T \quad u_i^T u_i = 1$$

核PCA: 非线性映射 $\phi(X)$, $x_n \rightarrow \phi(x_n)$, $\sum_n \phi(x_n) = 0$

$$Cu_i = \lambda_i v_i, \text{ 其中 } C = \frac{1}{N} \sum_{n=1}^N \phi(x_n) \phi(x_n)^T$$

$$\longrightarrow \frac{1}{N} \sum_{n=1}^N \phi(x_n) \{ \phi(x_n)^T v_i \} = \lambda_i v_i$$

$$\longrightarrow v_i = \sum_{n=1}^N \alpha_{in} \phi(x_n)$$

核主成分分析

● Kernel PCA

➤ 将主成分分析的线性假设一般化使之适应非线性数据

➤ 核PCA：
$$\frac{1}{N} \sum_{n=1}^N \phi(x_n) \phi(x_n)^T \sum_{m=1}^N \alpha_{im} \phi(x_m) = \lambda_i \sum_{n=1}^N \alpha_{in} \phi(x_n)$$

根据核函数 $k(x_n, x_m) = \phi(x_n)^T \phi(x_m)$

两边同乘 $\phi(x_l)^T$

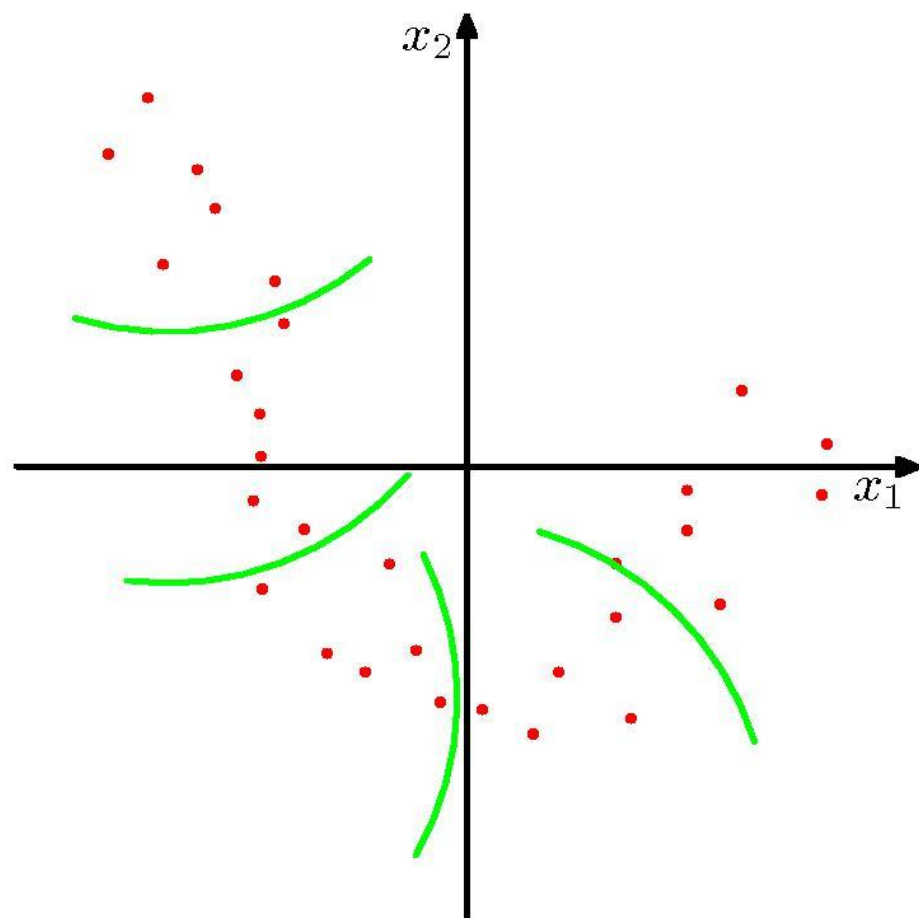
$$\longrightarrow \frac{1}{N} \sum_{n=1}^N k(x_l, x_n) \sum_{m=1}^N \alpha_{im} k(x_n, x_m) = \lambda_i \sum_{n=1}^N \alpha_{in} k(x_l, x_n)$$

$$\longrightarrow \mathbf{K}^2 \alpha_i = \lambda_i N \mathbf{K} \alpha_i$$

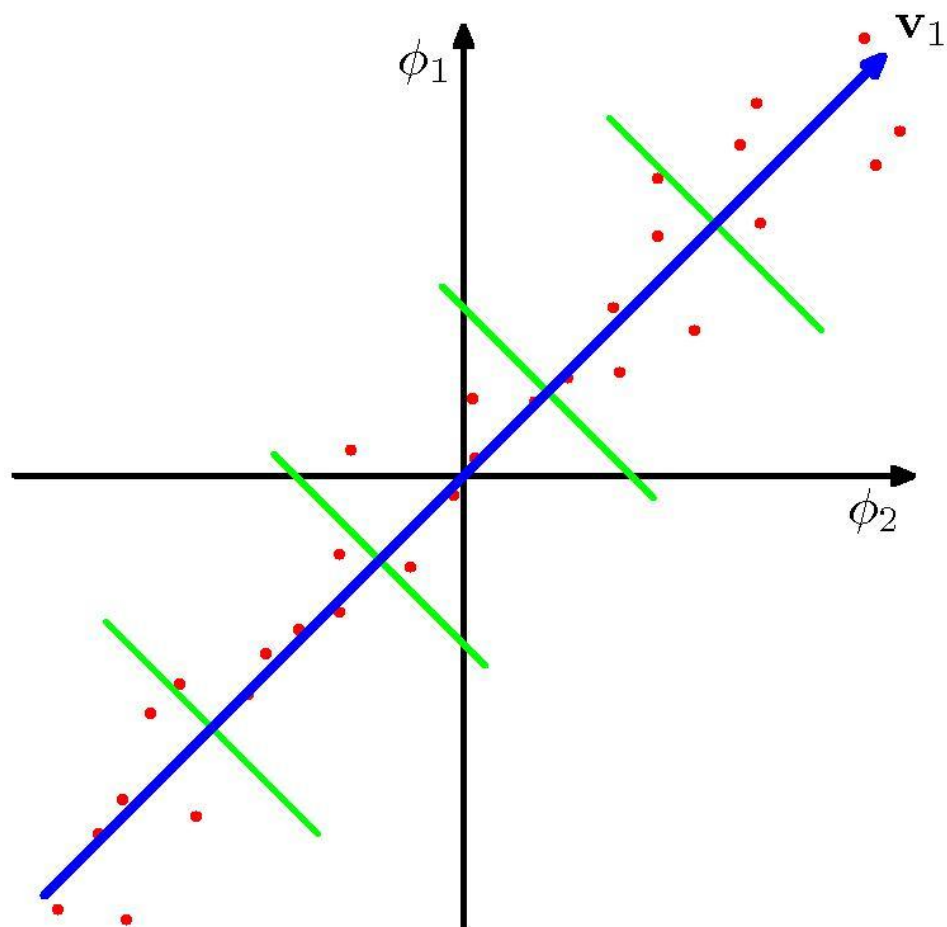
$$\longrightarrow \mathbf{K} \alpha_i = \lambda_i N \alpha_i$$

核主成分分析

● Kernel PCA



原始数据空间



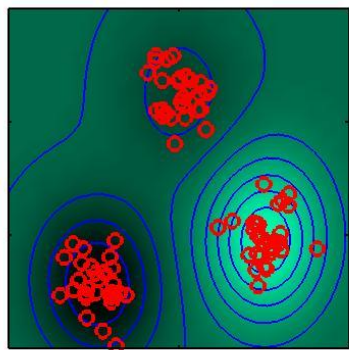
高维空间

- 数据点
- v_1 第一主成分方向
- 等值线

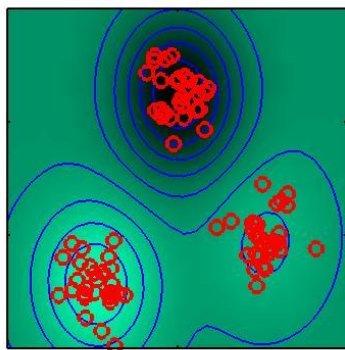
核主成分分析

● Kernel PCA

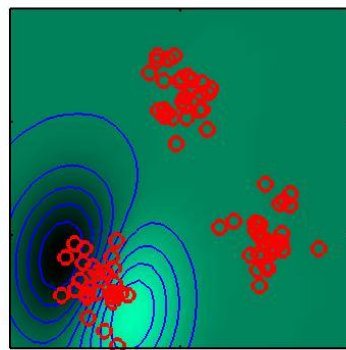
Eigenvalue=21.72



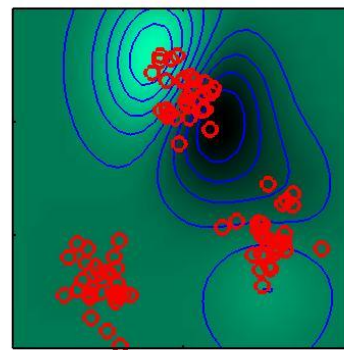
Eigenvalue=21.65



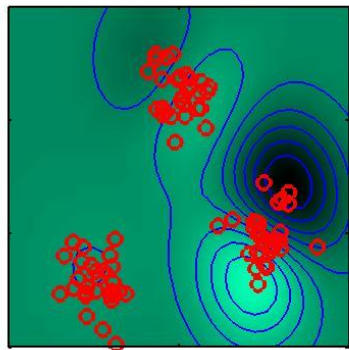
Eigenvalue=4.11



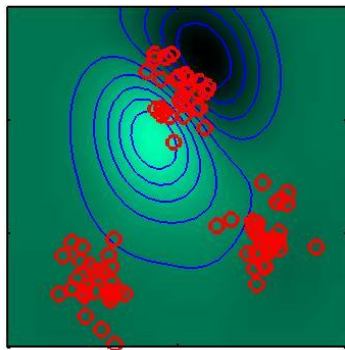
Eigenvalue=3.93



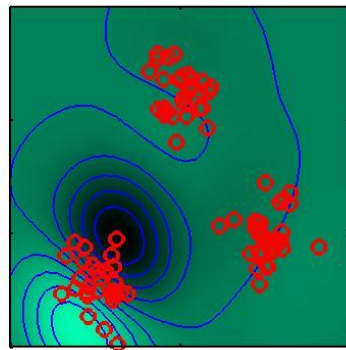
Eigenvalue=3.66



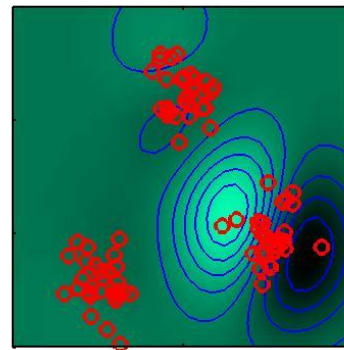
Eigenvalue=3.09



Eigenvalue=2.60



Eigenvalue=2.53



8.3 因子分析

- 因子分析的定义
- 正交因子模型
- 因子分析的应用

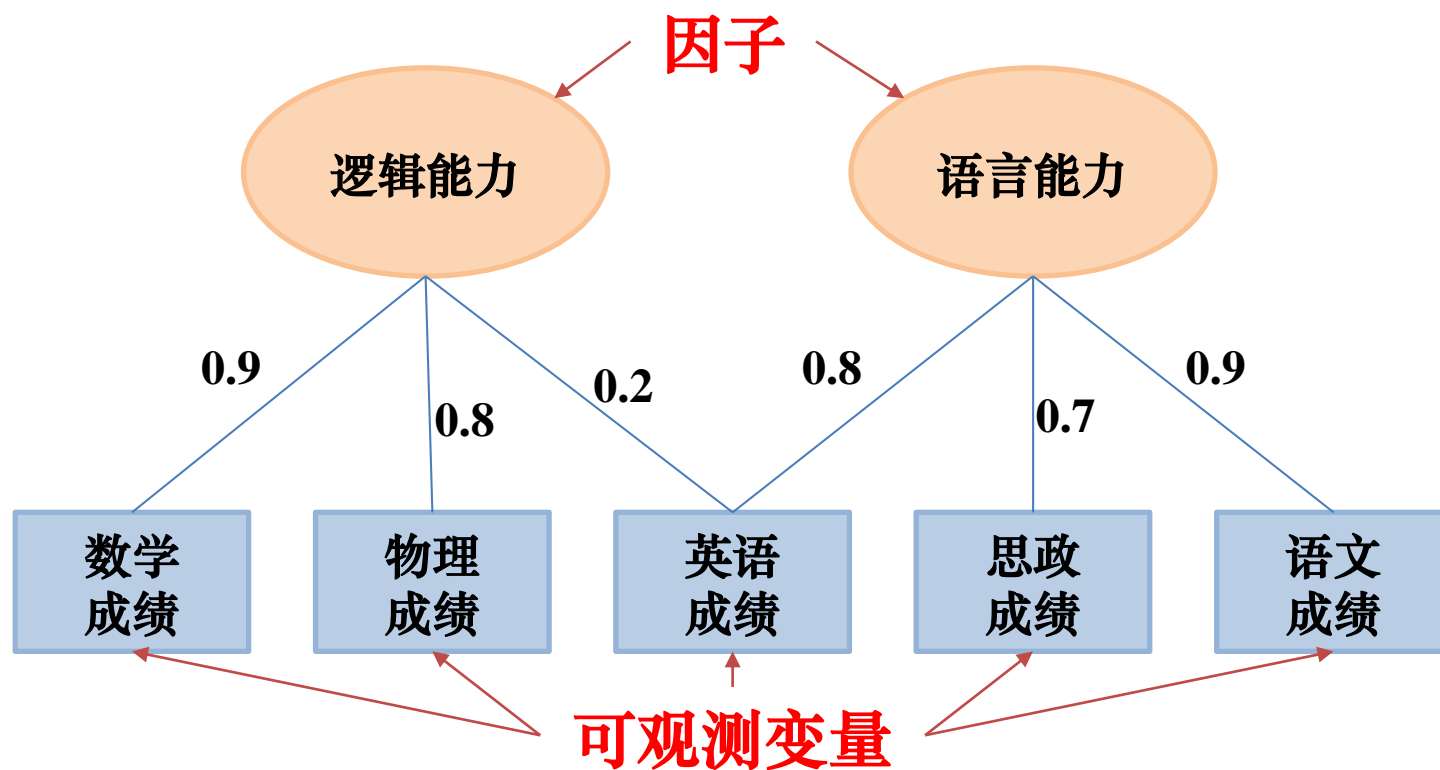
因子分析

● 因子分析 (Factor Analysis)

【1904年C. Spearman提出】

- 因子分析是指从数据内部的关联关系出发，将多个**可观测变量**归结为少数几个**因子 (隐变量)**，从而达到**降维**目的的一种多元统计分析方法

隐变量：是指不能直接观测到的变量，例如逻辑能力、语言能力等。



因子模型

- 因子模型 (Factor Model)

➤ 设有 p 维随机向量 X , $E(X) = \mu$, $Var(X) = \Sigma$

$$\begin{cases} X_1 = \mu_1 + a_{11}F_1 + a_{12}F_2 + \dots + a_{1m}F_m + \varepsilon_1 \\ X_2 = \mu_2 + a_{21}F_1 + a_{22}F_2 + \dots + a_{2m}F_m + \varepsilon_2 \\ \dots \\ X_p = \mu_p + a_{p1}F_1 + a_{p2}F_2 + \dots + a_{pm}F_m + \varepsilon_p \end{cases}$$

上式可化简为如下形式: $X = \mu + AF + \varepsilon$

其中 F 称为**公共因子**, ε 称为**特殊因子**, a_{ij} 称为第 i 个变量在第 j 个因子上的**因子载荷**, $A = (a_{ij})_{p \times m}$ 矩阵称为**载荷矩阵**

- 重要的因子模型: 正交因子模型

正交因子模型

正交因子模型假设:

$$(1) E(F) = 0, Var(F) = E(FF^T) = I \quad \text{正交性}$$

$$(2) E(\varepsilon) = 0, Var(\varepsilon) = E(\varepsilon\varepsilon^T) = \Phi = diag\{\phi_1, \dots, \phi_p\}$$

$$(3) Cov(F, \varepsilon) = E(F\varepsilon^T) = 0$$

● 在正交因子模型假设下有:

$$Cov(X, F) = Cov(AF + \varepsilon, F) = Cov(AF, F) + Cov(\varepsilon, F) = A \quad (9-3-1)$$

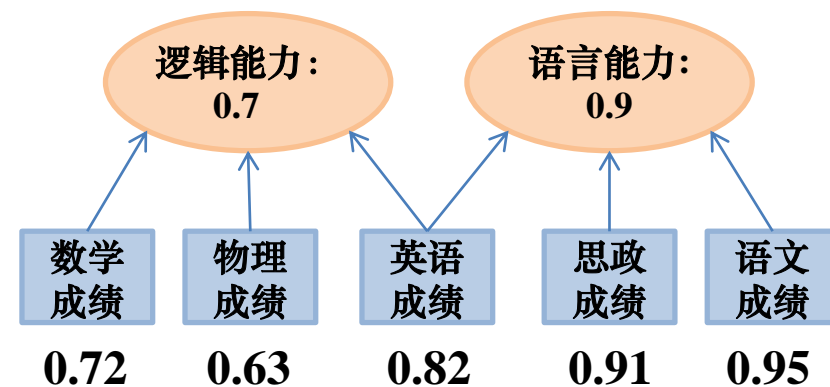
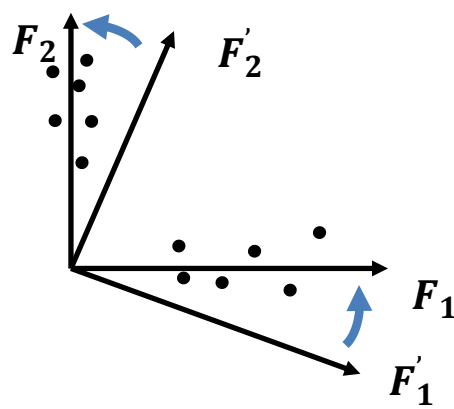
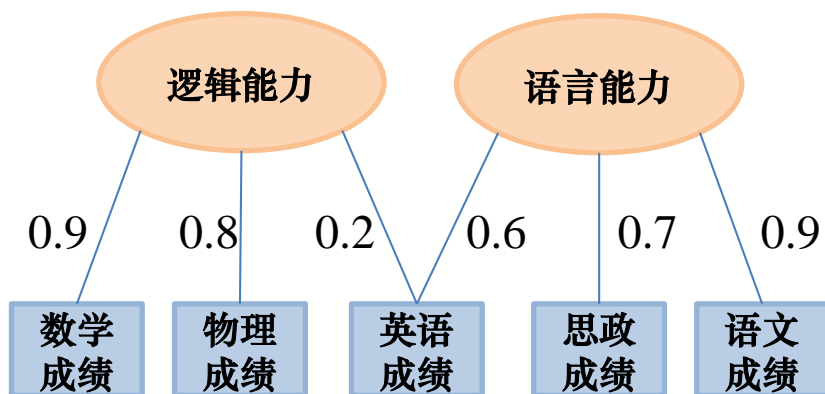
$$\Sigma = Var(X) = Var(AF + \varepsilon) = AA^T + \Phi \quad (9-3-2)$$

由式 (9-3-2) 可得, $\sigma_{ii} = Var(X_i) = h_i^2 + \phi_i$, 其中 $h_i^2 = a_{i1}^2 + a_{i2}^2 + \dots + a_{im}^2$

a_{ij}^2 表示第 j 个公共因子对 x_i 的方差贡献, ϕ_j 表示第 i 个特殊因子对 x_i 的方差贡献

模型求解步骤

- 步骤1：因子提取。求解因子模型中的载荷矩阵 A 和特殊因子 ε ，旨在将原始数据的高维数据归纳为较少的因子，以解释数据的方差
- 步骤2：因子旋转。通过旋转因子载荷矩阵 A ，对提取的因子进行变换，增强因子的解释性
- 步骤3：因子得分。基于旋转后的载荷矩阵 A ，计算每个样本在这些因子上的得分，使原本的高维数据能够在较低维度上被解释和分析



因子提取

- 因子提取旨在根据数据的协方差矩阵 Σ 求解载荷矩阵 A 和特殊因子 ε
- 因子提取的求解算法：主成分法、极大似然法、主因子法等

- 主成分法

$$X = \mu + AF + \varepsilon \Rightarrow \Sigma = AA^T + \Phi$$

- 设 Σ 的特征值为 $\lambda_1 \geq \dots \geq \lambda_p$, u_1, \dots, u_p 为对应的标准正交化特征向量, 对于所有特征值和特征向量有 $\Sigma = AA^T$, 其中 $A = (\sqrt{\lambda_1}u_1, \dots, \sqrt{\lambda_p}u_p)$
- 主成分法采用主成分分析的思想, 忽略特征值较小的 $p - m$ 个特征向量。则正交因子分析的载荷矩阵 A 可写为 $A = (\sqrt{\lambda_1}u_1, \sqrt{\lambda_2}u_2, \dots, \sqrt{\lambda_m}u_m)$
- 此时 $\Sigma \approx AA^T$, 可以用 Φ 补全 Σ 的对角线部分, 即有 $\varepsilon_j = \sqrt{\sigma_{jj} - h_j^2}$

因子旋转

- 因子旋转的前提：载荷矩阵不唯一，即载荷矩阵 A 与任意正交阵 T 矩阵相乘仍符合正交因子模型假设

设 T 为 $m \times m$ 正交阵，则有 $X = \mu + AF + \varepsilon = \mu + ATT^T F + \varepsilon$

记 $\tilde{A} = AT$ ， $\tilde{F} = T^T F$ ，则因子模型可以写为 $X = \mu + \tilde{A}\tilde{F} + \varepsilon$

其中 \tilde{F} 仍满足 $E(\tilde{F}) = 0$ ， $Var(\tilde{F}) = I$ ， $Cov(\tilde{F}, \varepsilon) = 0$

新的公共因子可以得到相同的方差结构： $\Sigma = AA^T + \Phi = \tilde{A}\tilde{A}^T + \Phi$

- 因子旋转的目标：确定一个结构简单（每个可观测变量仅在一个公共因子上有较大的载荷）的载荷矩阵 A
- 常见因子旋转算法：
最大方差旋转法 (Varimax)，最大四次方法 (Quartimax)

最大方差旋转法

- 最大方差旋转法 (Varimax) : 直观上使得可观测变量在每个因子上负荷尽可能分散, 即在每个因子上的方差最大
- 以两个因子的正交因子模型为例, 对两因子的载荷矩阵 A 进行平面正交旋转, 设旋转角度为 φ , 旋转矩阵为 Γ , 则有

$$A = \begin{bmatrix} a_{11} & a_{12} \\ \dots & \dots \\ a_{p1} & a_{p2} \end{bmatrix} \quad \Gamma = \begin{bmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{bmatrix} \quad B = \begin{bmatrix} a_{11}\cos\varphi + a_{12}\sin\varphi & -a_{11}\sin\varphi + a_{12}\cos\varphi \\ \dots & \dots \\ a_{p1}\cos\varphi + a_{p2}\sin\varphi & -a_{p1}\sin\varphi + a_{p2}\cos\varphi \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} \\ \dots & \dots \\ b_{p1} & b_{p2} \end{bmatrix}$$

最大方差旋转法要求两组数据 $(b_{11}^2, \dots, b_{p1}^2)$, $(b_{12}^2, \dots, b_{p2}^2)$ 的方差尽可能大。即

$$\max_{\varphi} V = \left[\frac{1}{p} \sum_{i=1}^p (b_{i1}^2)^2 - \left(\frac{1}{p} \sum_{i=1}^p b_{i1}^2 \right)^2 \right] + \left[\frac{1}{p} \sum_{i=1}^p (b_{i2}^2)^2 - \left(\frac{1}{p} \sum_{i=1}^p b_{i2}^2 \right)^2 \right]$$

根据求极值原理, 求得令导数 $\frac{dV}{d\varphi}$ 等于零的 φ , 即为因子旋转角度

因子得分

- 因子得分

- 因子得分的求解目标：根据 p 维可观测变量 X 和载荷矩阵 A 计算得到的每个样本在 m 个公共因子 F 上的得分值 \hat{F}
- m 维因子得分 \hat{F} 比 p 维可观测变量 X 更能体现样本的特点

- 因子得分的求解算法：最小二乘法、极大似然估计法等

- 最小二乘法：对于式 $X - \mu = AF + \varepsilon$ ，视特殊因子 ε 为随机误差，公共因子 F 为回归系数。由于 $Var(\varepsilon) = \Phi$ ，通过加权最小二乘法，得到

$$\hat{F} = (A^T \Phi^{-1} A)^{-1} A^T \Phi^{-1} (X - \mu)$$

因子分析的应用

- 示例：给定52个学生的6科成绩（数学、物理、化学、语文、历史和英语），使用正交因子模型分析科目间的相关性，将6科成绩归结为2个因子（文科因子，理科因子）
- 求解：根据6科成绩计算数据的协方差矩阵 $Var(X) = \Sigma =$

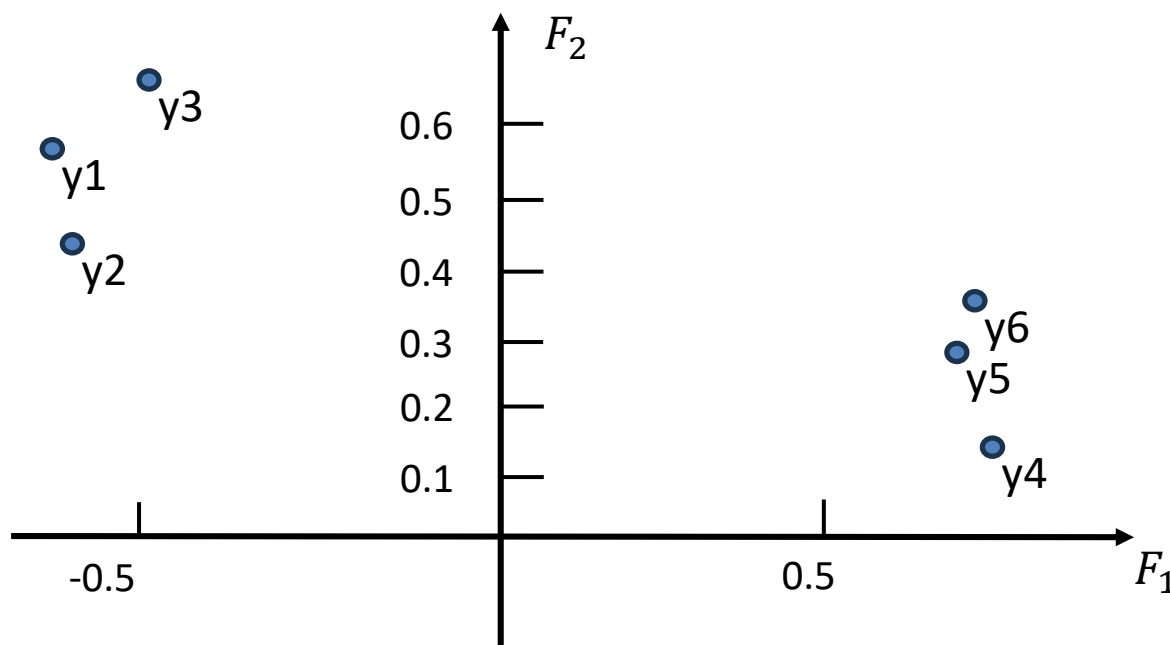
$$\begin{bmatrix} 1.000 & 0.647 & 0.696 & -0.561 & -0.456 & -0.439 \\ 0.647 & 1.000 & 0.573 & -0.503 & -0.351 & -0.458 \\ 0.696 & 0.573 & 1.000 & -0.380 & -0.274 & -0.244 \\ -0.561 & -0.503 & -0.380 & 1.000 & 0.813 & 0.835 \\ -0.456 & -0.351 & -0.274 & 0.813 & 1.000 & 0.819 \\ -0.439 & -0.458 & -0.244 & 0.835 & 0.819 & 1.000 \end{bmatrix}$$

因子分析的应用

- 因子提取：通过因子提取求得载荷矩阵 A 和特殊因子 ϵ

	数学	物理	化学	语文	历史	英语
文科因子 F_1	-0.676	-0.676	-0.676	0.917	0.917	0.883
理科因子 F_2	0.562	0.427	0.656	0.104	0.239	0.266
特殊因子 ϵ	0.228	0.459	0.333	0.148	0.210	0.150

载荷矩阵 A

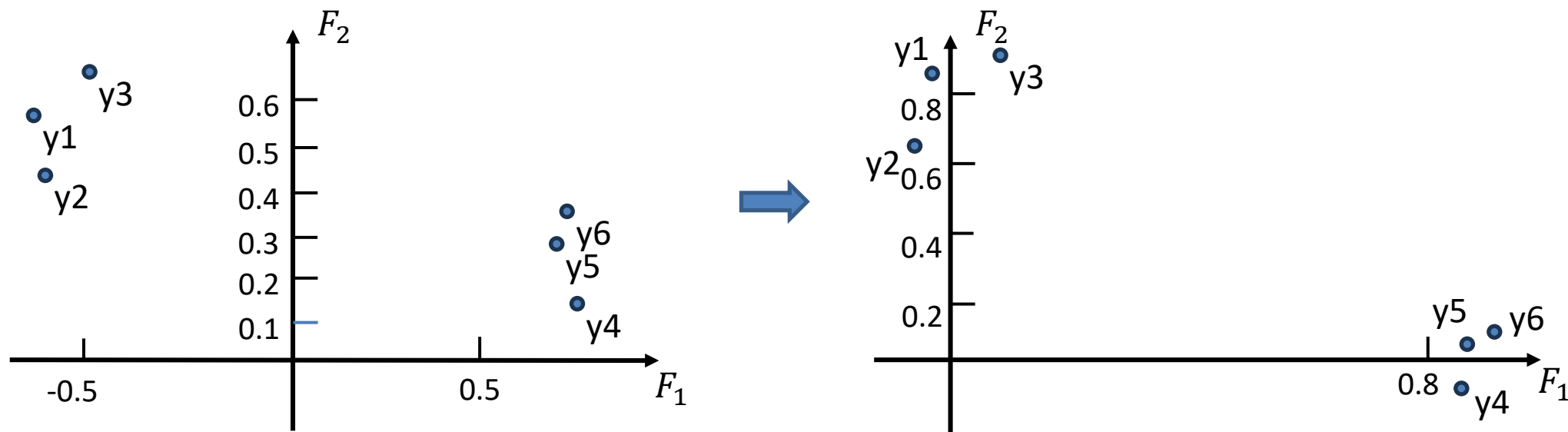


因子分析的应用

- 因子旋转：通过因子旋转得到每个可观测变量仅在一个公共因子上有较大的载荷的载荷矩阵 A

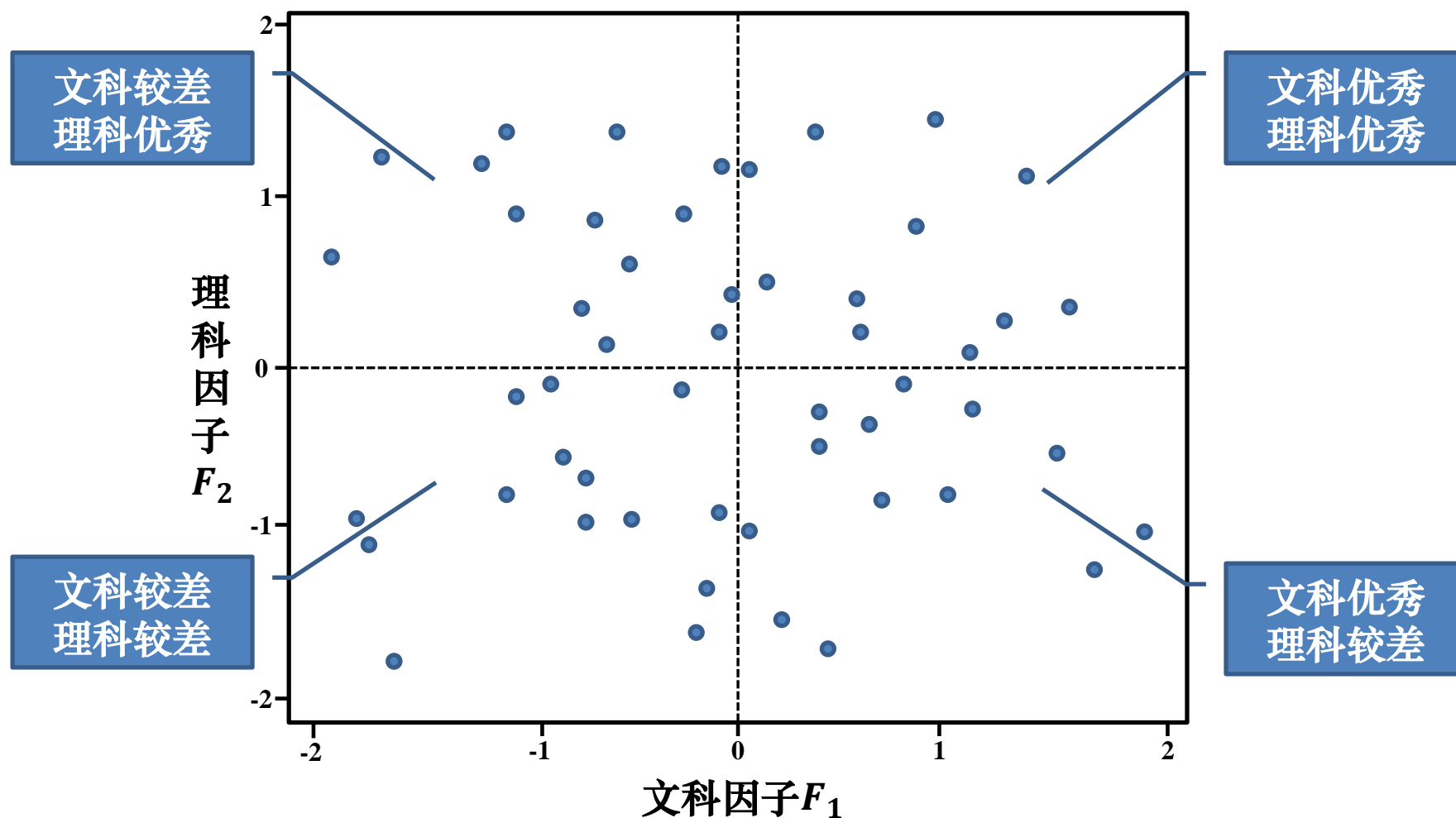
	数学	物理	化学	语文	历史	英语
文科因子 F_1	-0.089	-0.132	0.156	0.845	0.909	0.954
理科因子 F_2	0.833	0.657	0.890	-0.133	0.046	0.098
特殊因子 ϵ	0.228	0.459	0.333	0.148	0.210	0.150

载荷矩阵 A



因子分析的应用

- 因子得分：计算每个学生在文科因子和理科因子上的因子得分



因子分析 vs. 主成分分析

	因子分析	主成分分析
优化目标	提取因子解释数据中的结构关系	侧重于数据的降维和方差最大化
解释性	有助于理解变量之间的潜在联系	解释数据中的方差，本身没有明确含义
应用场景	从复杂的数据中提炼关键因素	数据的降维和可视化

8.4 独立成分分析

- 独立成分分析的定义
- 模型假设
- FastICA算法
- 独立成分分析的应用

独立成分分析的定义

- 什么是独立成分分析？

- 独立主成分分析 (Independent Component Analysis, ICA) 是一种用于信号分离的统计方法，它通过提取数据中的独立成分来识别和分离出数据内部的**本质结构**，这些独立成分不仅减少了**冗余信息**和**噪声**对数据分析的干扰，还能**保留关键信息**

- 以鸡尾酒派对为例



模型假设

- ICA模型变量定义

- x : 观测信号

- A : 混合矩阵

- s : 独立成分

$$x = As$$

- ICA模型求解目标

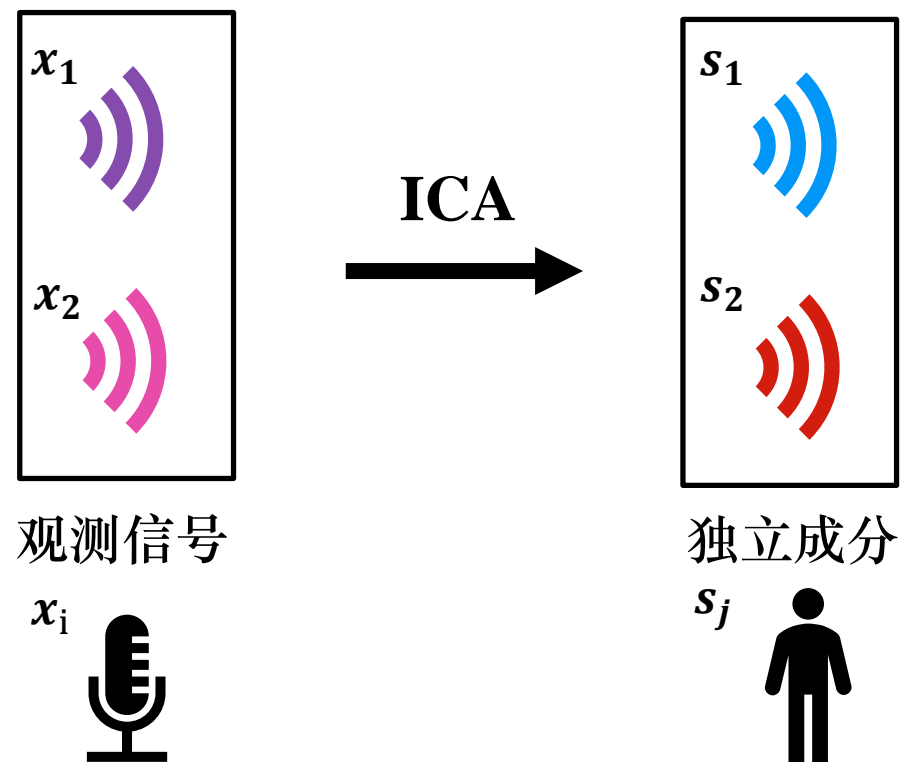
- 通过估计给权重向量 W ，使得从观测信号 x 中分离出的信号 $s = Wx$ 是相互独立的。

- 模型先验假设

- 独立成分**统计独立**

- 独立成分服从**非高斯分布**

- 典型算法：JADE、Infomax ICA、FastICA等



FastICA算法

● 原理

- 基于**非高斯性最大化**原理的独立成分分析方法，旨在从多维观测数据中提取出统计独立的信号源
- 分离过程中，可通过对分离结果的非高斯性度量来表示分离结果间的**相互独立性**，当非高斯性度量达到最大时，则表明已完成对各独立分量的分离

● 整体过程

- 步骤1：数据预处理，包含去中心化和白化
- 步骤2：寻找最大非高斯方向并计算独立成分，提取独立成分后进行正交化处理
- 步骤3：重复步骤2对其他独立成分进行提取

FastICA算法

● 步骤1：数据预处理

➤ 去中心化

对每个观测向量 x 计算均值 $x' = x - \mu$

目的：调整数据均值为零，方便后续的白化和独立成分提取

➤ 白化

1. 计算中心化后数据的协方差矩阵 $C = E[x'x'^T]$
2. 对协方差矩阵进行特征值分解 $C = EDE^T$, D 是特征值对角矩阵, E 是对应的特征向量矩阵
3. 使用特征值和特征向量对数据进行变换, 得到白化数据 $x'' = Cx'$

目的：使数据协方差为零，简化后续的独立成分提取过程

FastICA算法

● 步骤2：寻找最大非高斯方向并计算独立成分

1. 初始化：选择一个随机的单位向量 W 作为权重向量的初始值
2. 非线性变换：为了最大化信号的非高斯性，采用非线性函数 $g(\cdot)$ 对数据进行变换
3. 权重更新：根据固定点迭代法，使用输入数据 x'' 和当前的权重向量 W 进行权重更新：

$$W^+ = E[x'' g(W^T x'')] - E[g'(W^T x'')] W$$

其中， $E[x'' g(W^T x'')]$ 表示每个样本应用非线性函数后与输入数据的期望值， $E[g'(W^T x'')] W$ 是用于调整的校正项。更新后，需要对 W^+ 进行归一化

4. 计算独立成分：使用得到的权重向量 W 将白化后的数据投影到独立成分上，即

$$sx' = W^T x''$$

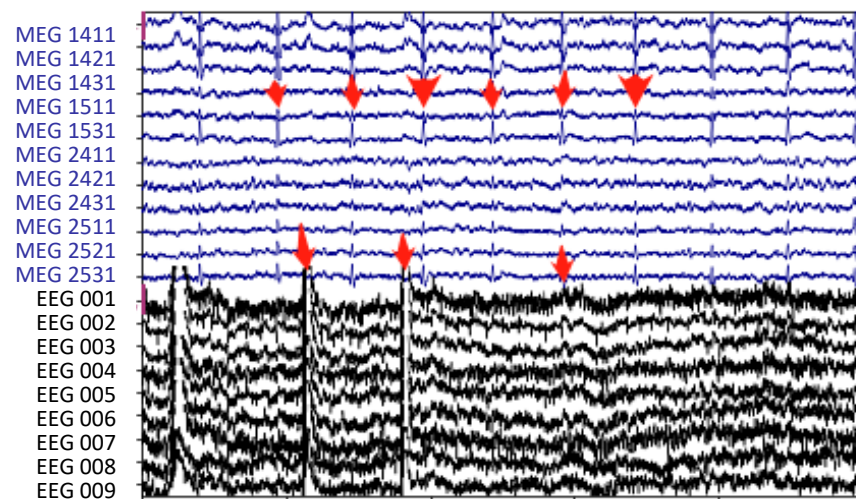
5. 正交化：如果提取多个独立成分，需要对新的权重向量进行正交化处理，以确保它们相互独立

● 步骤3：重复步骤2，对其他独立成分进行提取

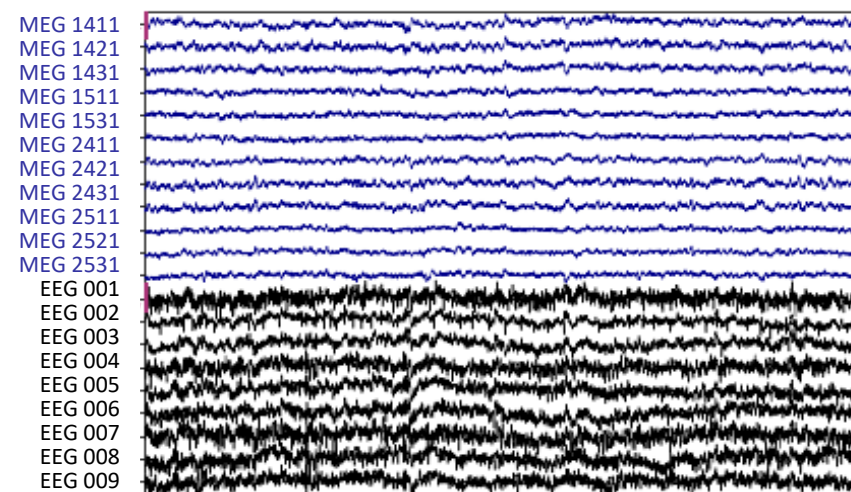
独立成分分析的应用

● 脑电图信号分析

- 脑电图 (EEG) 信号分析在神经科学研究和临床诊断中至关重要。然而，EEG 信号常常受到**伪影干扰**，例如眼动和心跳，这些**干扰信号**会**掩盖和混淆真正的脑电活动**，使得准确分析变得困难。因此，需要提取出脑电信号中的**关键成分**，**去除冗余信息和噪声**，从而揭示大脑的真实活动模式并**提高分析的精度**



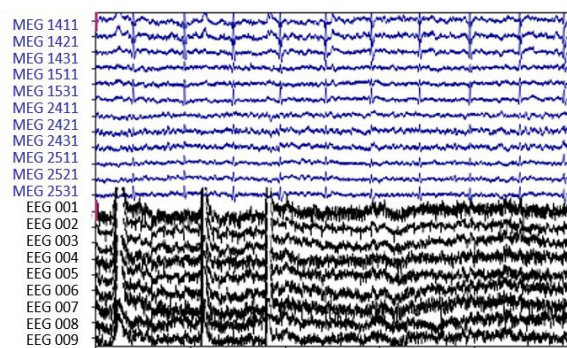
原始信号



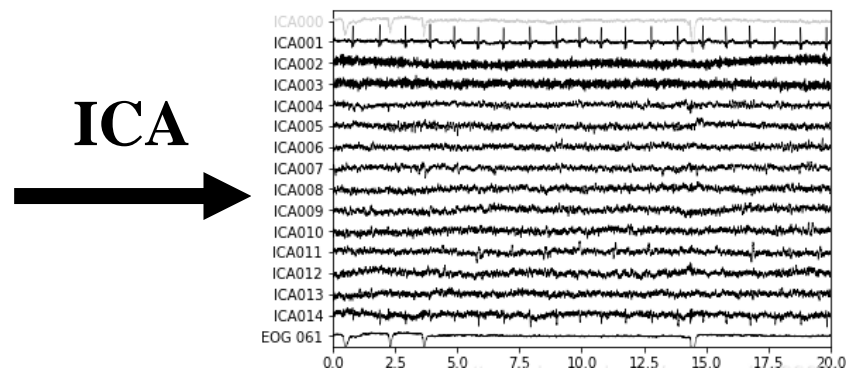
去噪后的信号

独立成分分析的应用

● 脑电图信号分析

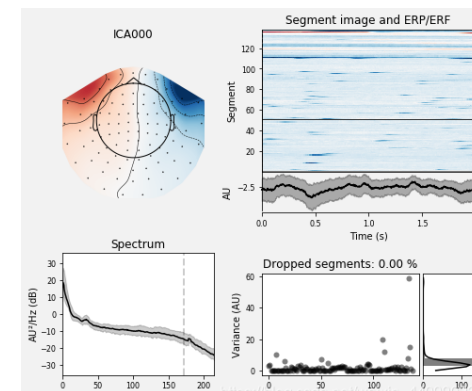


原始信号（观测信号）



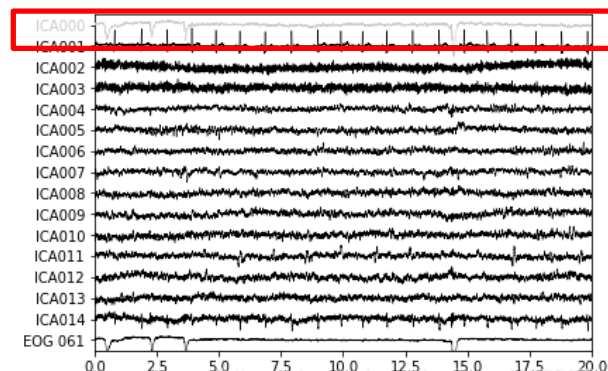
独立成分

可视化分析



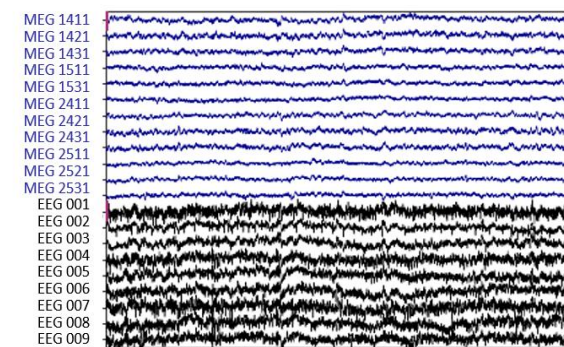
独立成分 s_j 可视化

确定噪声信号



独立成分

信号重新组合



去噪后的信号

讨论：ICA vs. PCA & 因子分析

	ICA	PCA	因子分析
优化目标	最大化信号的非高斯性，以提取独立成分	侧重于数据的降维和方差最大化	提取因子解释数据中的结构关系
解释性	有助于揭示数据中潜藏的独立源信号或独立成分	解释数据中的方差，本身没有明确含义	有助于理解变量之间的潜在关系
应用场景	常用于盲信号分离，也可用于降维	数据的降维和可视化	从复杂的数据中提炼决定性因素

小结：优点和局限性

● ICA的优点

- 灵活性：可通过不同的非线性函数来调整算法的行为，以适应不同的数据分布特征提取能力：可识别数据中重要特征，应用于分类等其他任务
- 物理解释性：提取的独立成分通常具有明确的物理或现实意义，尤其在信号处理领域

● ICA的局限性

- 假设原信号是**非高斯分布**的，也就决定了它对高斯假设的数据效果不佳
- 假设原信号是**统计独立**的
- 算法处理复杂数据集和多源信号时，可能会出现**不收敛问题**

8.5 等距映射

- 等距映射的定义
- 等距映射的流程
- 等距映射的应用

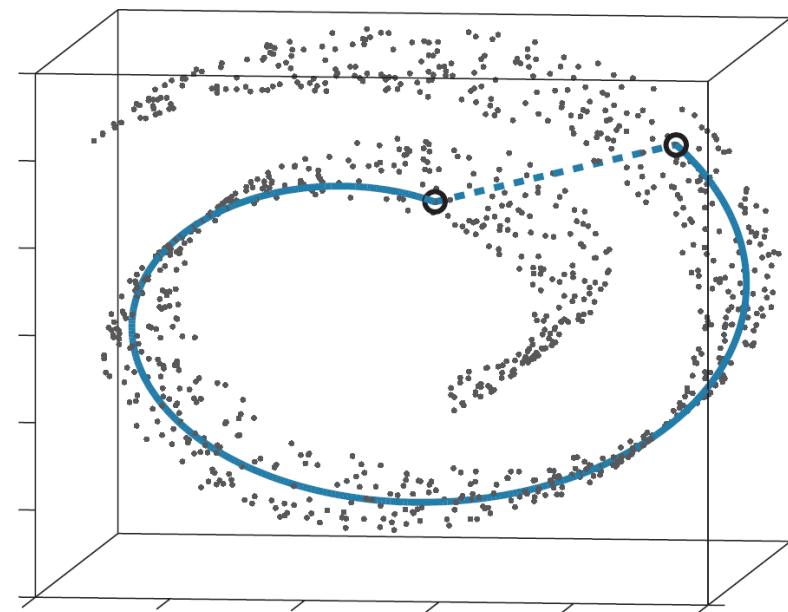
等距映射的定义

● 等距映射 (Isometric Mapping, ISOMAP)

- 一种基于流形假设的非线性数据降维方法。该方法认为高维数据集中分布在一个低维流形上，并试图在降维过程中保持样本之间的测地距离尽可能不变
- 流形 (Manifold): 一般几何对象的总称，包括各种维度的曲线或曲面等

● 非线性数据降维

- 非线性降维方法允许数据在降维过程中发生非线性变换。此类方法能够保留数据中的局部结构和非线性信息
- 代表方法：等距映射、局部线性嵌入



等距映射的流程

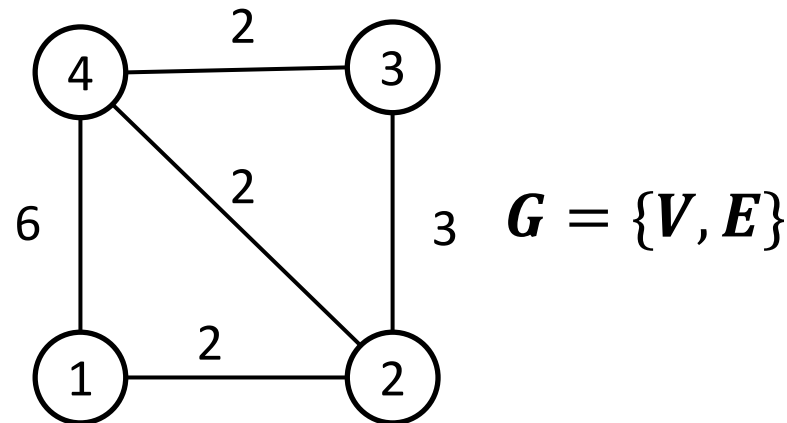
● 计算步骤

- 步骤1: **构造邻接图**。将数据中的每个样本视为一个点，将其与指定半径邻域内所有点相连(或与指定个数最近邻相连)
- 步骤2: **计算测地距离**。利用构造的临近关系图，计算图中所有点对之间的最短路径，得到距离矩阵
- 步骤3: **多维尺度分析**。将高维空间中的数据点投影到低维空间，使投影前后的距离矩阵相似度最大

等距映射的流程

● 步骤1：构建邻接图

- 对于给定数据 $X = \{x_1, x_2, \dots, x_N\}$ ，构造邻接图 $G = \{V, E\}$ ，其中 V 是顶点集合， E 是边的集合。
- 若计算样本 i 和 j 之间的距离 $d(i, j) = \text{distance}(x_i, x_j)$ ，可根据实际需求选择不同的距离度量函数
- 如果 $d(i, j)$ 小于某个阈值 ϵ ，或 j 是 i 的 K 近邻，则认为顶点 i 与 j 有边相连，且边权值设为 $d(i, j)$ 。否则认为 i 与 j 不相连， $d(i, j) = +\infty$ 。得到邻接矩阵 $D = \{d(i, j)\}$



$$D = \begin{pmatrix} 0 & 2 & +\infty & 6 \\ 2 & 0 & 3 & 2 \\ +\infty & 3 & 0 & 2 \\ 6 & 2 & 2 & 0 \end{pmatrix}$$

等距映射的流程

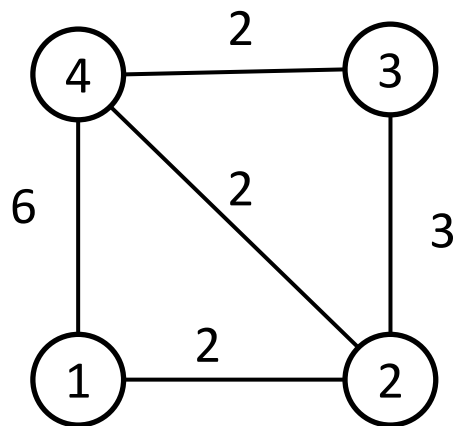
● 步骤2：计算测地距离

➤ 计算图 $G = \{V, E\}$ 中任意两点之间的最短路径，即为两点间的测地距离 $d_G(i, j)$ ，遍历得到测地距离矩阵 $D_G = \{d_G(i, j)\}$

➤ 可选方法：Dijkstra算法、Floyd-Warshall算法

➤ 以Floyd-Warshall算法为例

使用 $d(i, j)$ 初始化 $d_G(i, j)$ ，对于 $k = 1, 2, \dots, N$ ，将 $d_G(i, j)$ 更新为 $\min\{d_G(i, j), d_G(i, k) + d_G(k, j)\}$
更新完毕后的结果即为 D_G



$$D = \begin{Bmatrix} 0 & 2 & +\infty & 6 \\ 2 & 0 & 3 & 2 \\ +\infty & 3 & 0 & 2 \\ 6 & 2 & 2 & 0 \end{Bmatrix}$$

$$D_G = \begin{Bmatrix} 0 & 2 & 5 & 4 \\ 2 & 0 & 3 & 2 \\ 5 & 3 & 0 & 2 \\ 4 & 2 & 2 & 0 \end{Bmatrix}$$

等距映射的流程

● 步骤3：多维尺度分析 (Multi-dimensional Scaling)

➤ 定义平方距离矩阵 $S(i, j) = D_G^2 = \{d_G(i, j)^2\}$

➤ 构造中心矩阵(Centering Matrix) $H = I_N - \frac{1}{N} \mathbf{1}\mathbf{1}^T$

其中 I_N 表示 N 阶单位阵， $\mathbf{1}$ 表示元素全为1的长度为 N 的向量

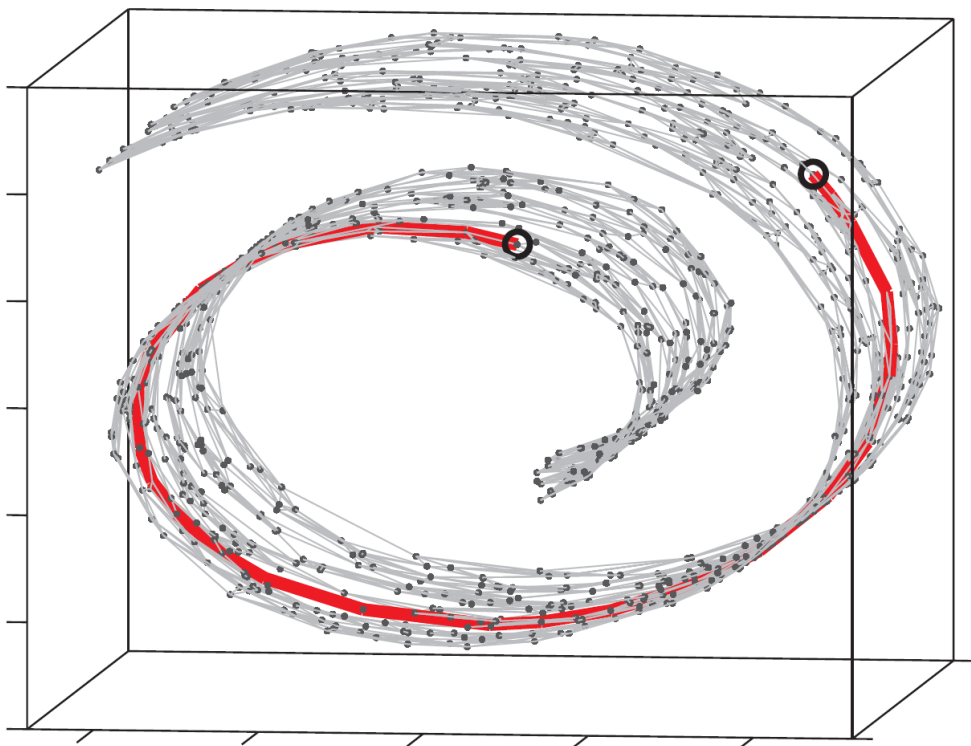
➤ 构造矩阵 $L = \frac{1}{2} HSH$ ，求解 L 的特征值与特征向量，并按特征值**降序排列**，

其中 λ_p 为第 p 个特征值， v_p 为其对应的特征向量

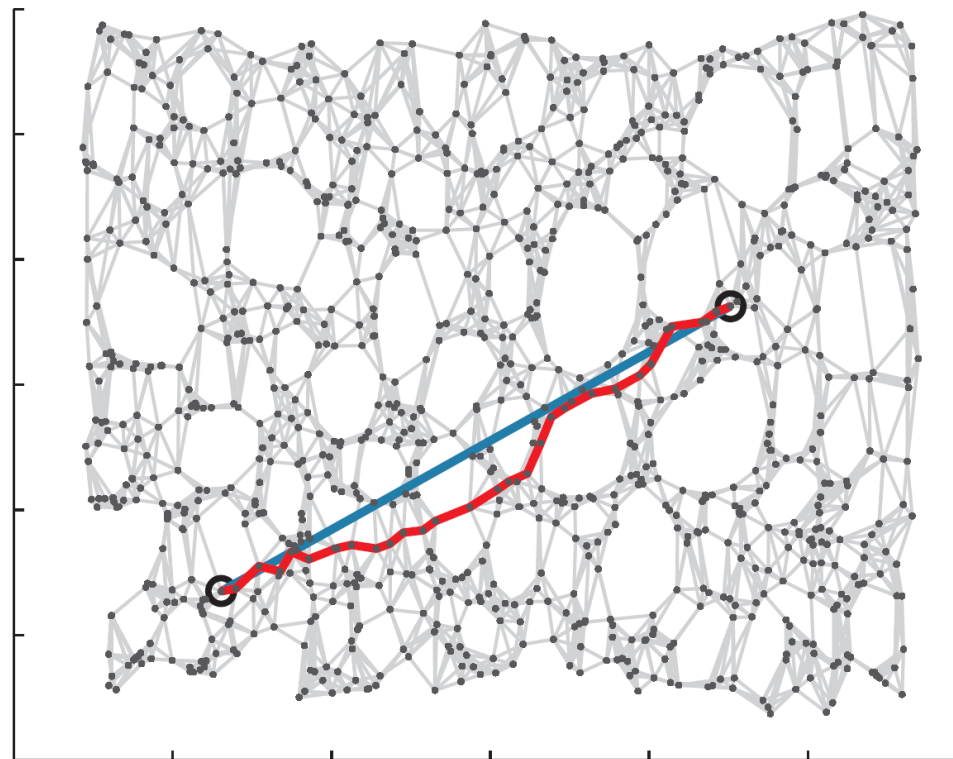
➤ $\{\sqrt{\lambda_1}v_1, \sqrt{\lambda_2}v_2, \dots, \sqrt{\lambda_d}v_d\}_{N \times d}$ 的第 i 行表示样本 x_i 降至 d 维的坐标

等距映射的效果

- 使用等距映射进行降维，能够较好的保持样本间的测地距离



降维前



降维后

等距映射优缺点

● 等距映射优点

- 非线性：可以有效捕捉数据中复杂的非线性信息
- 非迭代：无需反复更新参数，可以一次运算得到降维后的效果
- 保持全局结构：可以保持样本降维前的测地距离

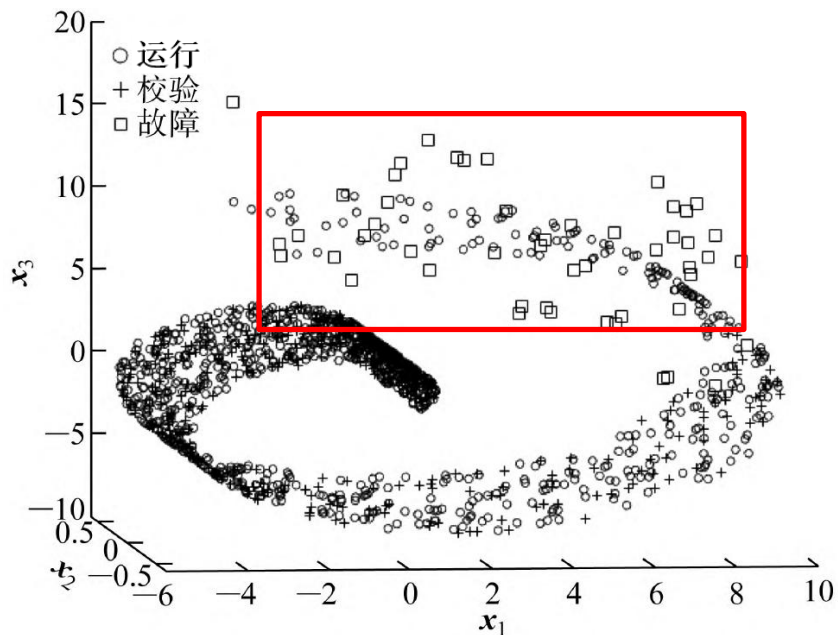
● 等距映射缺点

- 容易受噪声干扰，离群点可能导致全局距离的显著偏差
- 在大曲率区域可能存在“短路”现象，即将原本不相连的样本连接起来
- 如果数据集由多个不连通的流形组成，则可能无法正确地映射这些流形
- 寻找最短路径的计算复杂较高，不利于大规模样本数据降维

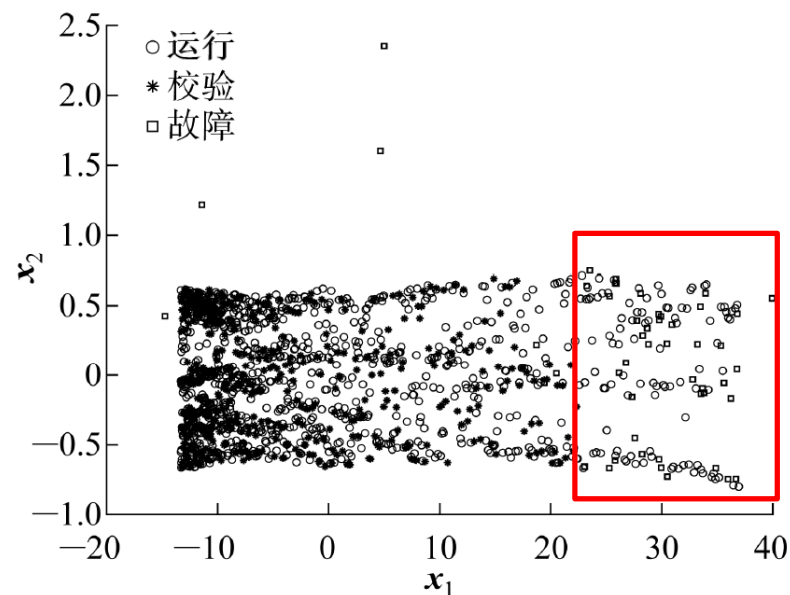
等距映射的应用

● 应用：工业设备故障诊断

- 背景：工业设备产生的数据具有高维、非线性等特点，不利于进行故障诊断。正常运行的数据点通常会形成一个紧凑的流形，而故障数据则会偏离这个流形
- 通过等距映射，可以增强数据的线性可分性，提升故障诊断的准确率



降维前



等距映射降维后

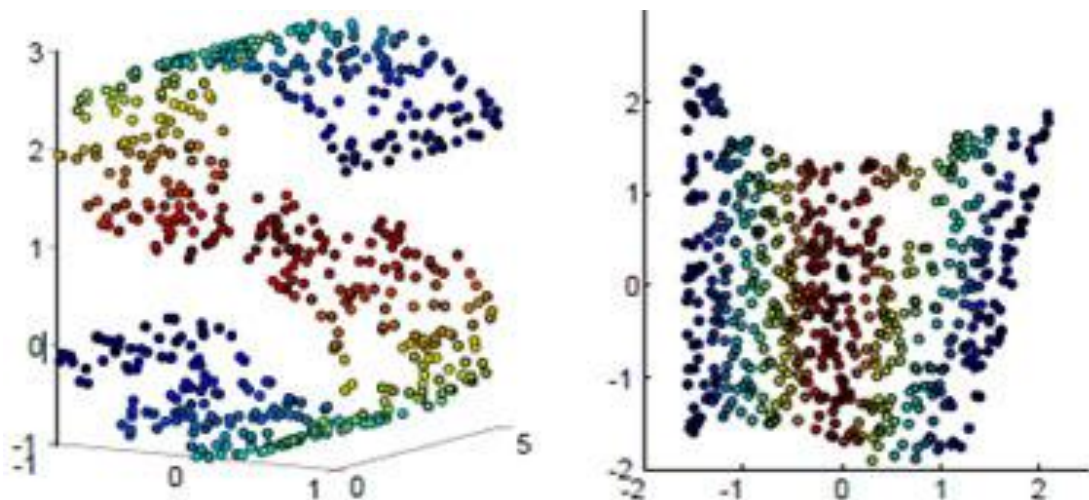
8.6 局部线性嵌入

- 局部线性嵌入的定义
- 局部线性嵌入的流程
- 局部线性嵌入的应用

局部线性嵌入的定义

● 局部线性嵌入 (Local Linear Embedding, LLE)

- 一种基于流形假设的非线性数据降维方法
- **前提假设**: 该方法假设采样数据所在的低维流形**在局部是线性的**，每个采样点可以用它的**近邻点线性**表示
- **学习目标**: 在低维空间中保持每个邻域中的权值不变，即假设嵌入映射在局部是线性的条件下，**最小化重构误差**



Nonlinear Dimensionality Reduction by Locally Linear Embedding

Sam T. Roweis¹ and Lawrence K. Saul²

SCIENCE VOL 290 22 DECEMBER 2000

局部线性嵌入的流程

● 计算步骤

➤ 步骤1: K 近邻重构

基于局部线性假设, 对于任意一个样本, 可用其距离最近的 K 个样本加权表示

➤ 步骤2: 重构系数求解

通过最小化重构误差, 计算所有样本的重构系数

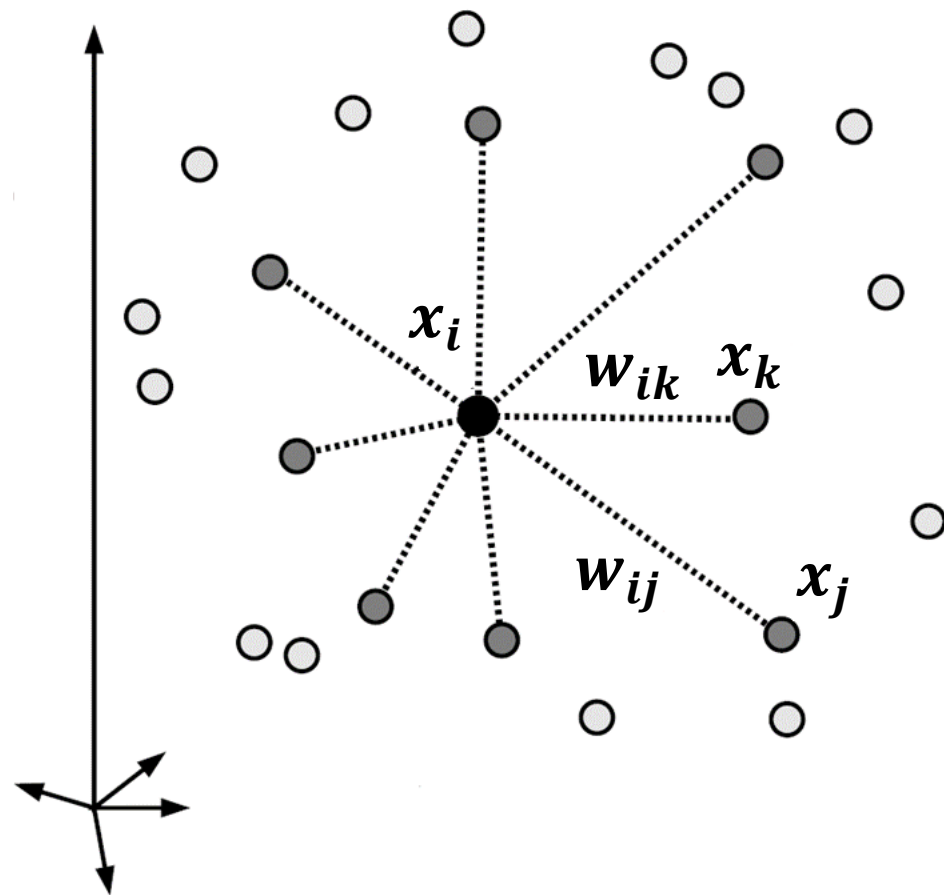
➤ 步骤3: 低维坐标求解

保持样本重构系数不变, 通过最小化降维后的重构误差, 求解低维空间坐标

局部线性嵌入的流程

● 步骤1: K 近邻重构

- 前提假设: 基于**局部线性假设**, 每个点都能用其周围的点线性表示出来
- 对于给定数据集 $X = \{x_1, x_2, \dots, x_N\}$ 中的样本 x_i , 寻找其 **K 近邻**, 表示为 $\{x_{ij}\}$, 其中 $j = 1, 2, \dots, K$
- 使用 $\{x_{ij}\}$ 对 x_i 进行重构, 即求一组权值 $\{w_{ij}\}$, 使得 $x_i = \sum_j w_{ij} x_{ij}$, 其中 $\{w_{ij}\}$ 满足 $\sum_j w_{ij} = 1$



局部线性嵌入的流程

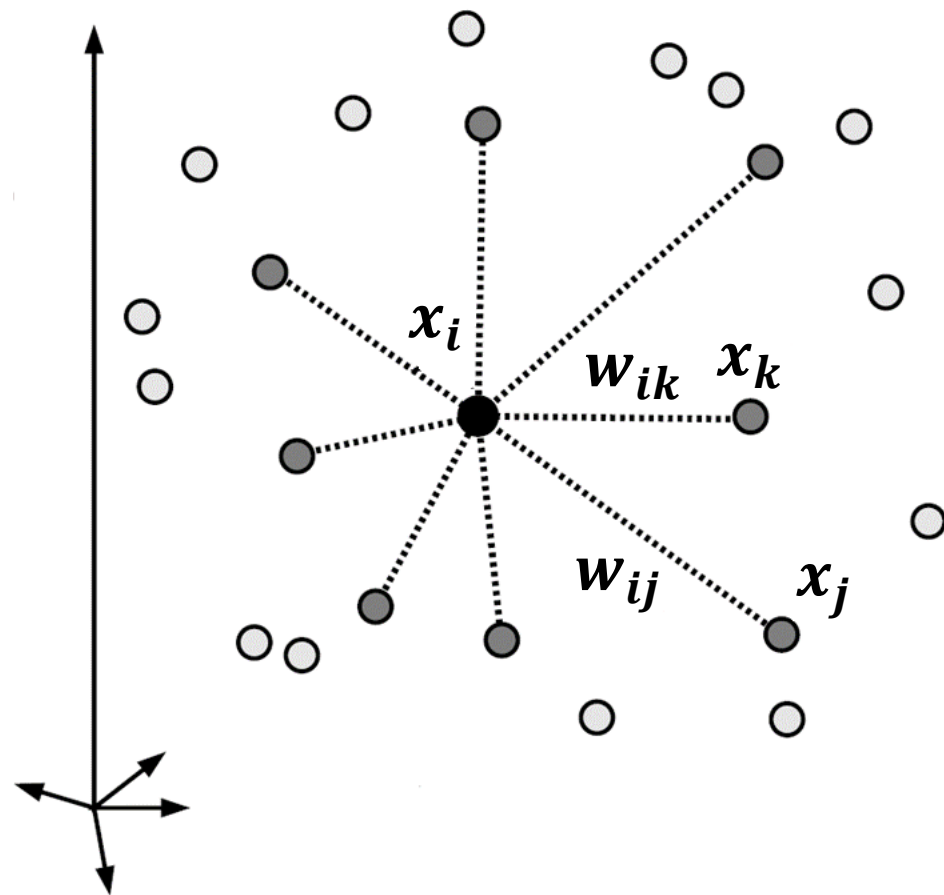
● 步骤2：重构系数求解

➤ 由于实际数据并非严格满足线性假设，故求解使重构误差最小的 $\{w_{ij}\}$

➤ 优化目标： $\min \sum_i |x_i - \sum_j w_{ij} x_{ij}|^2$
s. t. $\sum_j w_{ij} = 1$

➤ 首先构造局部协方差矩阵 $C^i = \{C_{jk}^i\}$ ，其中 $C_{jk}^i = (x_i - x_j) \cdot (x_i - x_k)$

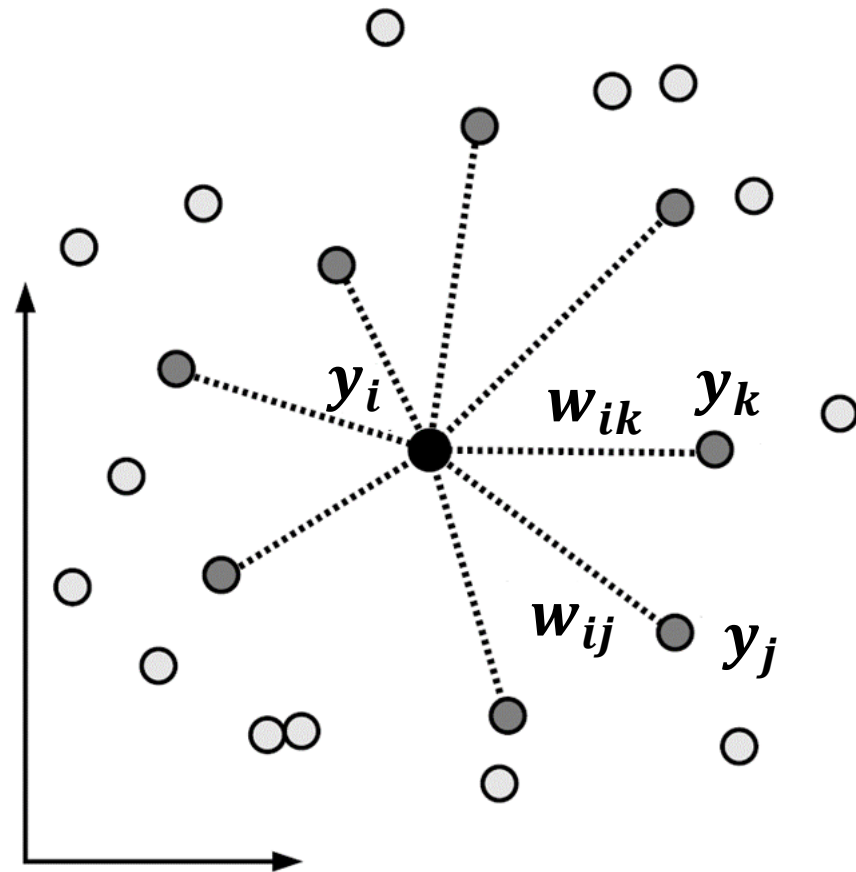
➤ 可求得 $w_{ij} = \frac{\sum_k (C^i)^{-1}_{jk}}{\sum_{jk} (C^i)^{-1}_{jk}}$



局部线性嵌入的流程

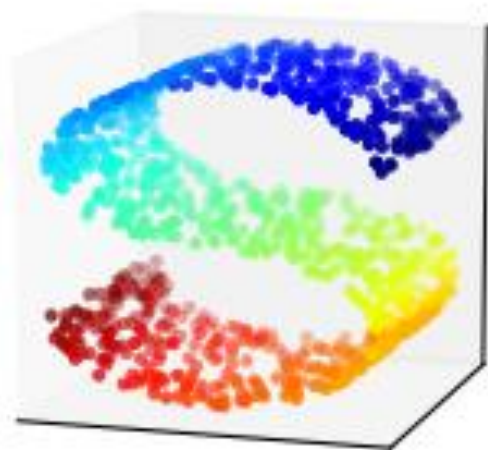
● 步骤3：低维坐标求解

- 将 $X = \{x_1, x_2, \dots, x_N\}$ 降维为 $Y = \{y_1, y_2, \dots, y_N\}$ ，使原本的重构系数 $W = \{w_{ij}\}$ 尽量不变
- 优化目标： $\min \sum_i |y_i - \sum_j w_{ij} y_j|^2$
- 可转化为 $\min \sum_{ij} M_{ij} (y_i \cdot y_j)$ ，其中 $M = (I - W)^T (I - W)$
- 通过求解可知，选择 M 的**最小 d 个非零特征值**对应的特征向量构建 $N \times d$ 的矩阵，即可得到局部线性嵌入的降维结果 Y

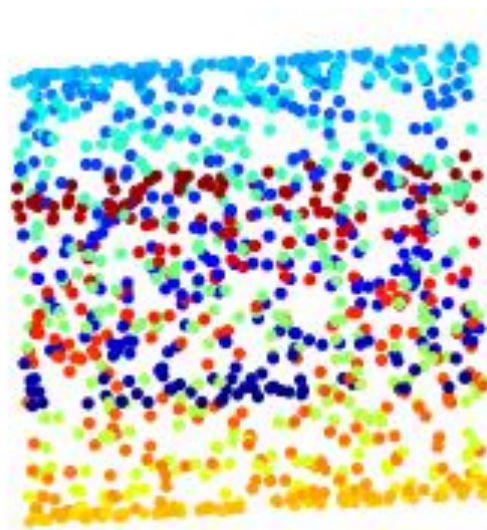


局部线性嵌入的效果

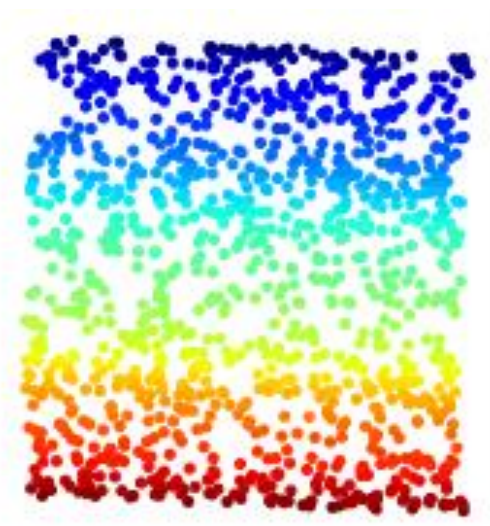
- 局部线性嵌入与主成分分析降维效果比较，有效避免了线性方法降维后局部结构改变的问题



降维前



主成分分析降维



局部线性嵌入降维

局部线性嵌入优缺点

● 局部线性嵌入优点

- 非线性：可以捕捉数据中复杂的非线性信息
- 非迭代：可以一次运算得到降维后的效果，较为高效
- 保持局部结构：在降维后，数据点之间的相对位置和距离也能得到较好的保留
- 参数较少：仅涉及调节 K 近邻的数量，使得模型的调整较为直观

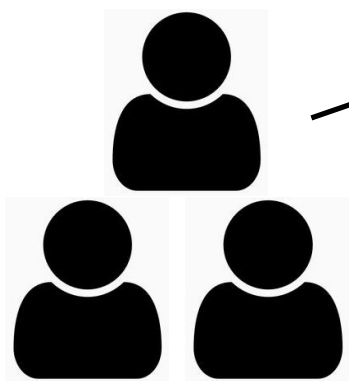
● 局部线性嵌入缺点

- 噪声点可能会破坏数据点之间的局部线性关系，从而影响降维结果的准确性
- 当数据集由多个不连通的流形组成时，可能无法正确地进行降维
- 目标函数是一个非凸优化问题，这意味着它可能陷入局部最优解

局部线性嵌入的应用

● 应用：社交网络数据可视化

- 背景：社交网络的发展逐渐催生出不同的用户类群，同一类群内的用户具有相似行为模型。通过数据降维，将有助于了解各个用户类群的特点并加以区分
- 局部线性嵌入能够在降维后保留样本的局部结构，清晰展示出用户类群内的一致性，以及不同用户类群之间的差异性



关注数 点赞数
评论数 观看数
收藏数

