维度约简

高　阳, 李文斌

http://cs.nju.edu.cn/rl, 2023.3.21

大纲

特征选择和降维

线性判别分析 LDA

主成分分析 PCA

独立成分分析 ICA

局部线性嵌入 LLE

ISOMAP

本讲要求的数学基础

线性代数：

特征值、特征向量、向量点乘、矩阵逆、伪逆、矩阵迹

统计学：

期望、方差、协方差、卡方检验、皮尔逊系数、熵

优化：拉格朗日乘子法

矩阵求导：

$$\frac{\partial A^{\top}A}{\partial A} = A \qquad \frac{\partial tr(AB)}{\partial A} = B^{\top}$$

大纲

特征选择和降维

线性判别分析 LDA

主成分分析 PCA

独立成分分析 ICA

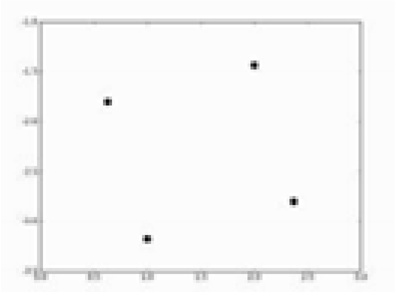
局部线性嵌入 LLE

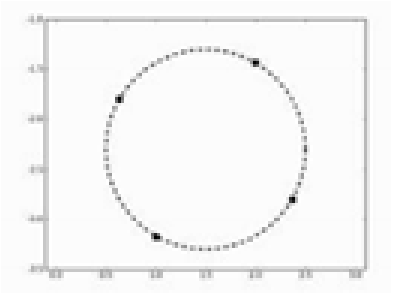
ISOMAP

特征选择和降维

- 回顾
  - ✓ 上一章中的自动编码器
- 维度约简(Dimensionality Reduction)
  - ✓ 特征选择(Feature Selection)
  - ✓ 特征诱导/变化(Feature Derivation)
- 目的
  - ✓ 降维，降低over-fitting风险
  - ✓ 增加解释性
  - ✓ 去除冗余特征

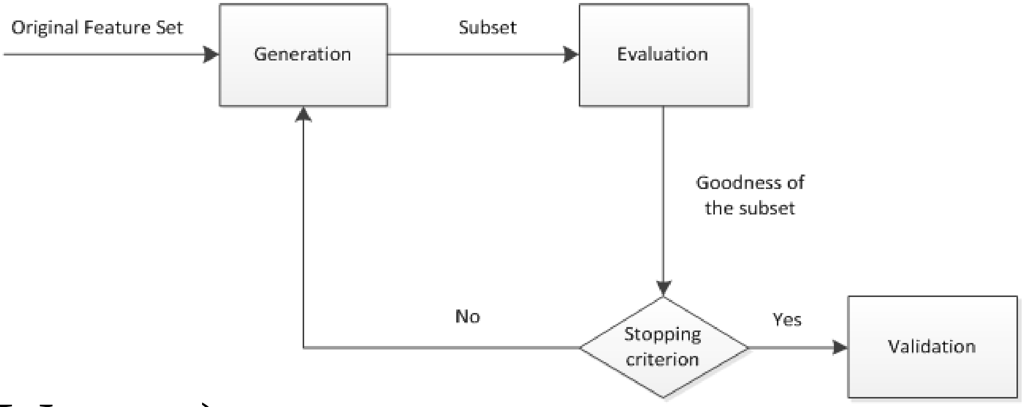
x	y
2.00	-1.43
2.37	-2.80
1.00	-3.17
0.63	-1.80





特征选择

- 搜索问题
  - ✓ N个原始特征，2^N-1非空特征空间，搜索最优的特征子集
- 搜索起点和方向
  - ✓ 前向(起点为空集)、后向(起点为全集)、双向
- 搜索策略
  - ✓ 穷举、序列、随机
- 特征评估函数
  - ✓ 过滤式(Filter) 和封装式(Wrapper)



特征选择

- 过滤式
  - ✓ 使用评价准则来<u>增强特征与类的相关性, 削减特征之间的相关性</u>
  - ✓ 距离度量(方差法)、信息度量(互信息熵)、依赖性度量(皮尔逊相关系数，卡方检验)以及一致性度量
- 封装式
  - ✓ 特征及其与下游任务目标(分类、回归、聚类等)的相关性
  - ✓ 如分类错误率
- 嵌入式
  - ✓ 模型正则化，如加上稀疏约束 (如著名的LASSO回归)

这部分内容可以专门学习，本讲不再赘述。

大纲

特征选择和降维

线性判别分析 LDA

主成分分析 PCA

独立成分分析 ICA

局部线性嵌入

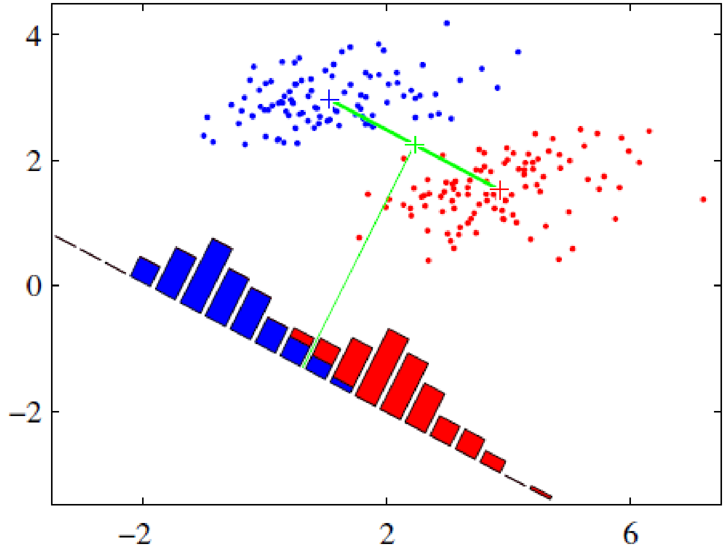
ISOMAP

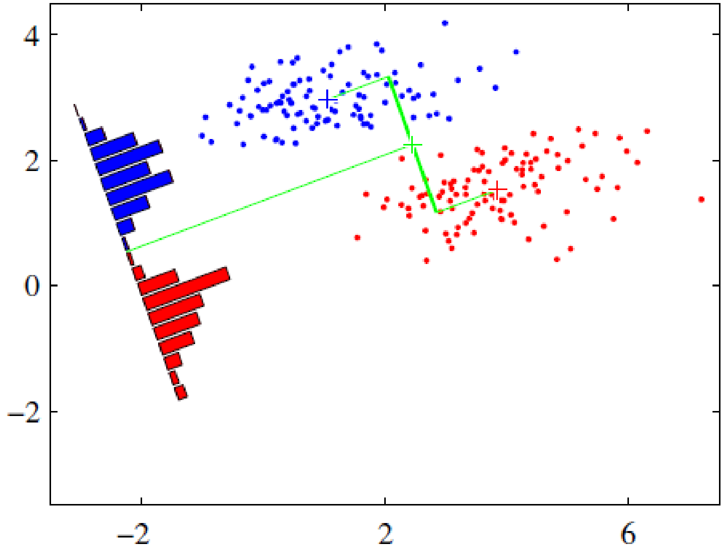
线性判别分析LDA

- 线性分类器(回忆多分类感知器)

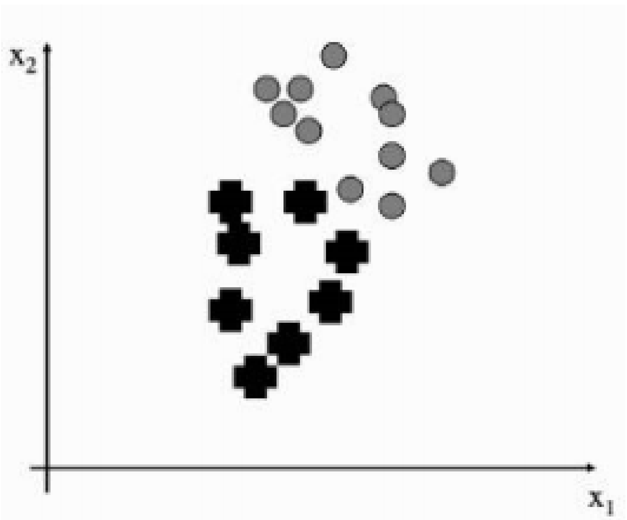
$$y_k(x) = w_k^T x + w_{k0}$$

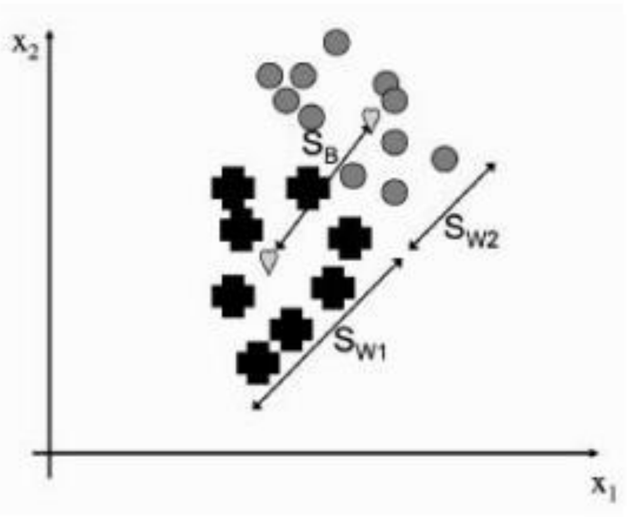
- 线性判别分析Linear Discriminant Analysis
  - ✓ 给定标注了类别的高维数据集，投影到高维的超平面，使得样本点的按类别尽最大可能区分开





线性判别分析LDA





□相关统计量

- ✓ 均值：$\boldsymbol{\mu}_1$，$\boldsymbol{\mu}_2$和$\boldsymbol{\mu}$
- ✓ 协方差：$\sum_j (\mathbf{x}_j - \boldsymbol{\mu})(\mathbf{x}_j - \boldsymbol{\mu})^T$
- ✓ 概率：p_c

有监督的特征降维方法：

Linear Discriminant Analysis

线性判别分析LDA

- (源)数据集的组内分布/散度Scatters(Within)

越小越好

$$S_W = \sum_{\text{classes } c} \sum_{j \in c} p_c (\mathbf{x}_j - \boldsymbol{\mu}_c)(\mathbf{x}_j - \boldsymbol{\mu}_c)^T.$$

- (源)数据集的组间分布(Between)

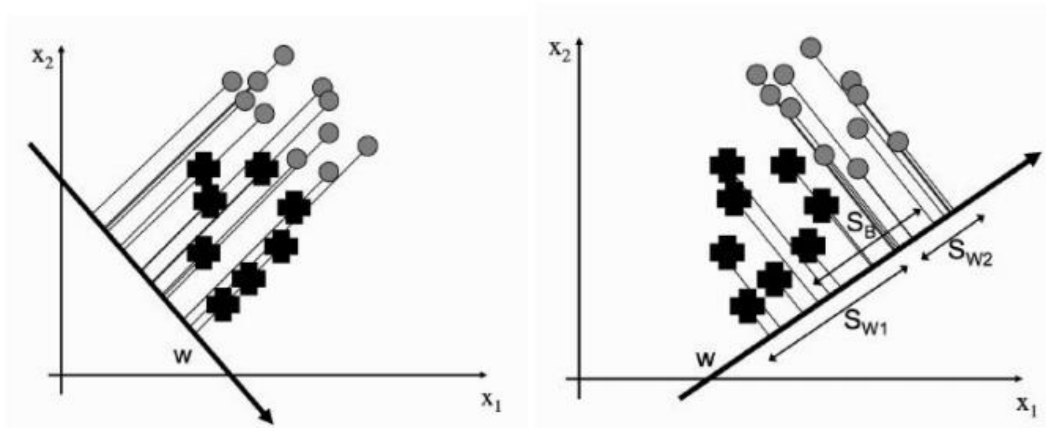
越大越好

$$S_B = \sum_{classes\ c} (\boldsymbol{\mu}_c - \boldsymbol{\mu})(\boldsymbol{\mu}_c - \boldsymbol{\mu})^T.$$

投影的目标：类别内的点距离越近越好, 类别间的点越远越好

线性判别分析LDA

投影后数据集的组内方差和组间方差/散度



$$\sum_{classes\ c}\ \sum_{j\in c} \color{red}{p_c} \left(\mathbf{w}^T(\mathbf{x}_j-\boldsymbol{\mu}_c)\right)\left(\mathbf{w}^T(\mathrm{x}_j-\boldsymbol{\mu}_c)\right)^T = \mathbf{w}^T S_W \mathbf{w}$$

$$\sum_{classes\ c} \mathbf{w}^T(\boldsymbol{\mu}_c-\boldsymbol{\mu})(\boldsymbol{\mu}_c-\boldsymbol{\mu})^T\mathbf{w} = \mathbf{w}^T S_B \mathbf{w}$$

最小化

$$\color{red}{\frac{\mathbf{w}^T S_W \mathbf{w}}{\mathbf{w}^T S_B \mathbf{w}}}$$

原文有笔误。

线性判别分析LDA

- **对w求导，并令导数为0**

$$\frac{S_B\mathbf{w}(\mathbf{w}^T S_W\mathbf{w}) - S_W\mathbf{w}(\mathbf{w}^T S_B\mathbf{w})}{(\mathbf{w}^T S_B\mathbf{w})^2} = 0$$

$$S_W\mathbf{w} = \boxed{\frac{\mathbf{w}^T S_W\mathbf{w}}{\mathbf{w}^T S_B\mathbf{w}}} S_B\mathbf{w} = \lambda S_B\mathbf{w}$$

计算$S_W^{-1}S_B$的特征值和特征向量

具体推导过程略。

线性判别分析LDA

◻算法基本流程

- ✓ 计算每个类别的均值$\boldsymbol{\mu}_i$，全局样本均值$\boldsymbol{\mu}$
- ✓ 计算类内散度矩阵S_W，类间散度矩阵S_B
- ✓ 对矩阵$S_W^{-1}S_B$做特征值分解
- ✓ 取最大的数个特征值所对应的特征向量
- ✓ 计算投影矩阵

思考散度矩阵的维度。

大纲

特征选择和降维

线性判别分析 LDA

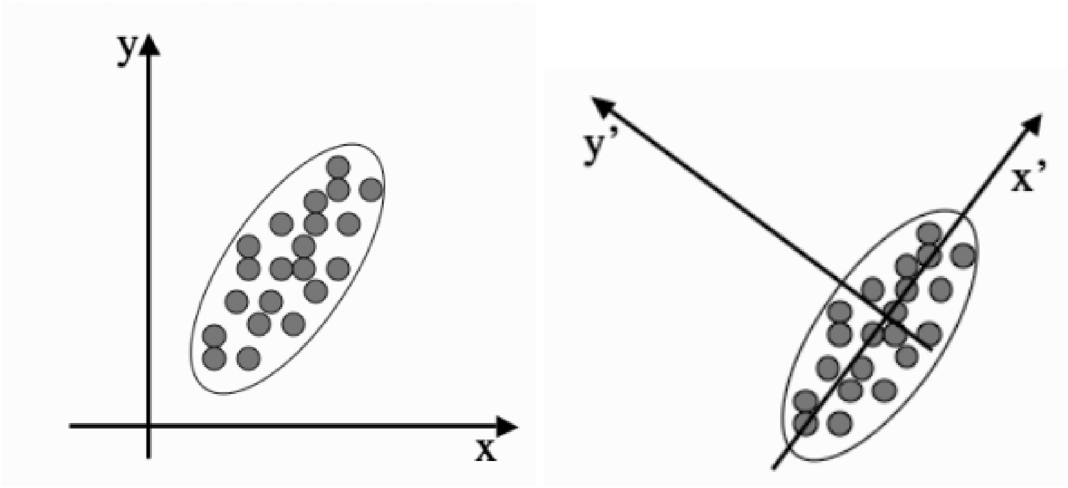
主成分分析 PCA

独立成分分析 ICA

局部线性嵌入 LLE

ISOMAP

主成分分析PCA



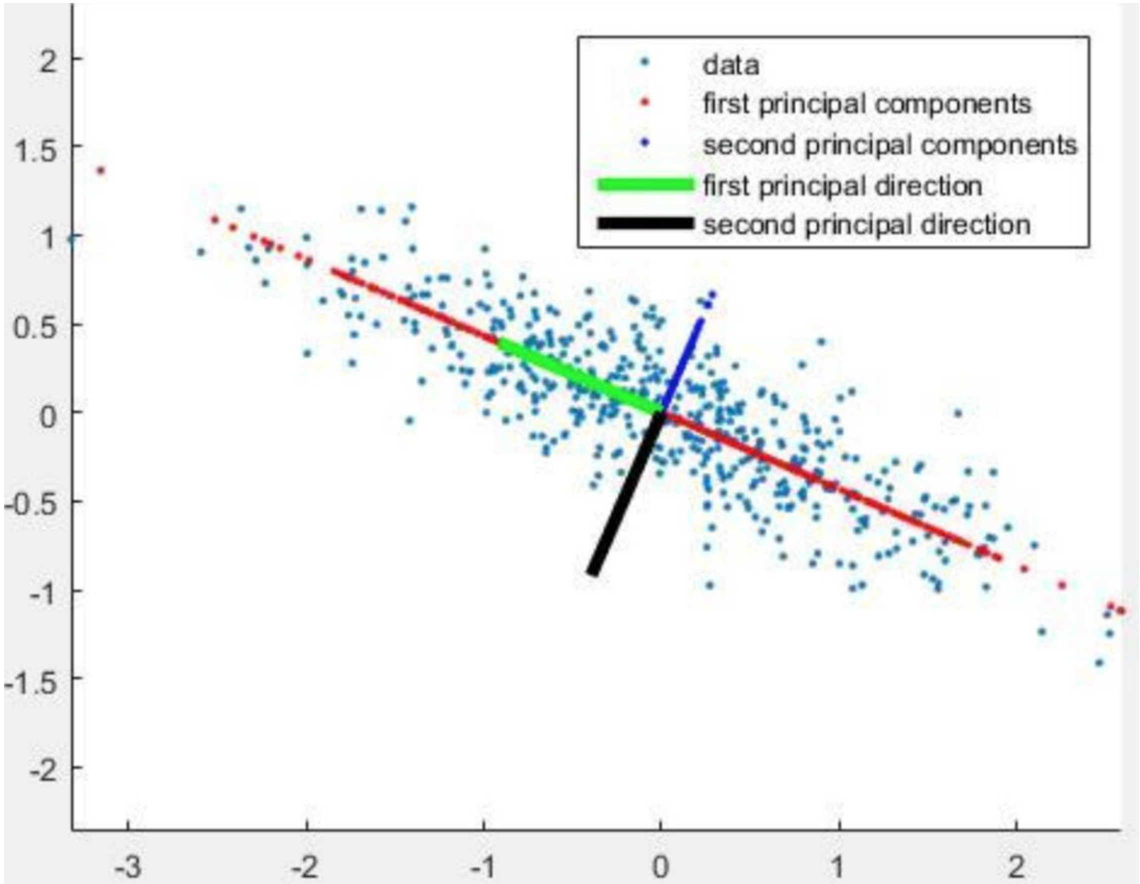
无监督的特征降维方法：

Principal Components Analysis

如何选择坐标轴呢？

主成分分析PCA

- Principal Component Analysis
  - ✓ 找到数据中的主要成分，并以之表征数据。



主成分分析PCA

◻ 主成分的特点

- ✓ 最大可分性：样本点在第一主成分上的投影其离散程度要大于其在第二主成分上投影的离散程度
- ✓ 最近可重构性：样本点到第一主成分线的平均距离要小于其到第二主成分线的距离
- ✓ 以此类推……

◻ 最大可分性理论的目标函数

最大化样本点在主成分上投影的方差

主成分分析PCA

- 给定去中心化的样本数据 $\{x_1, x_2, x_3, \ldots, x_n\}$
- 投影后的数据 $x_i^\top W$
- 投影后数据的方差

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

$$D(x) = \frac{1}{n}\sum_{i=1}^{n} W^T x_i x_i^\top W$$

样本点已被去中心化，简化了协方差矩阵的表达

$$\max\ tr(W^T C W)$$

$$s.t.\ W^T W = I$$

注意：特征值和数据方差的关系

矩阵的迹等于矩阵特征值的和。

主成分分析PCA

- 优化：拉格朗日乘子法

$$f(W) = tr(W^T CW) + \lambda(W^T W - I)$$

- 求导

$$\frac{\partial f}{\partial W} = \frac{\partial tr(W^T CW)}{\partial W} + \lambda\frac{\partial(W^T W)}{\partial W} = 0$$

- 利用矩阵迹的求导性质

$$CW = \lambda W \qquad D(x) = W^T CW = W^T \lambda W = \lambda$$

结论： 投影后的方差就是协方差矩阵的特征值，而最大方差就是协方差矩阵最大的特征值，最佳投影方向就是最大特征值所对应的特征向量

主成分分析PCA

- 算法基本流程
  - ✓ 样本去中心化
  - ✓ 计算样本的协方差矩阵
  - ✓ 对协方差矩阵做特征值分解
  - ✓ 取最大的数个特征值所对应的特征向量
  - ✓ 计算投影矩阵

主成分分析PCA

寻找一个正交变换P，使得：

$$\mathrm{cov}(\mathbf{Y}) = \mathrm{cov}(\mathbf{P}^T\mathbf{X}) = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \lambda_n \end{pmatrix}$$

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

$$= E[(\mathbf{P}^T\mathbf{X})(\mathbf{X}^T\mathbf{P})] = \mathbf{P}^T E(\mathbf{X}\mathbf{X}^T)\mathbf{P} = \mathbf{P}^T\mathrm{cov}(\mathbf{X})\mathbf{P}$$

$$\mathbf{P}\,\mathrm{cov}(\mathbf{Y}) = \mathbf{P}\mathbf{P}^T\,\mathrm{cov}(\mathbf{X})\,\mathbf{P} = \mathrm{cov}(\mathbf{X})\,\mathbf{P}$$

因为P是正交且cov(**Y**)对角阵

$$\mathbf{P}\,\mathrm{cov}(\mathbf{Y}) = [\lambda_1\boldsymbol{p}_1, \lambda_2\boldsymbol{p}_2, \cdots, \lambda_N\boldsymbol{p}_N]$$

主成分分析PCA

寻找一个正交变换P，解下式：

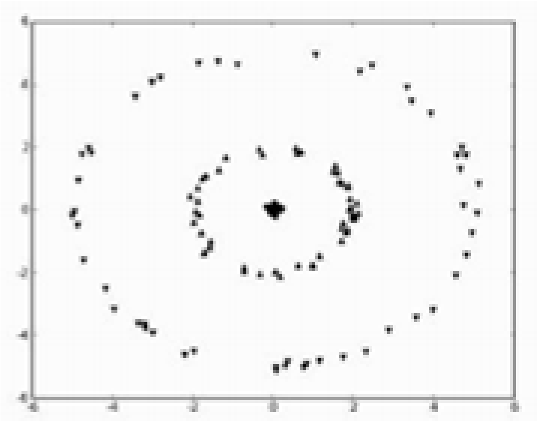
$$\boldsymbol{\lambda P} = \mathrm{cov}(\mathbf{X})\boldsymbol{P}$$

本质：求解X协方差矩阵的特征值λ和特征向量P

主成分分析算法详细流程

- ✓ 输入N个点$\mathbf{x}_i = (\mathbf{x}_{1i}, \mathbf{x}_{2i}, \cdots, \mathbf{x}_{Mi})$作为行向量
- ✓ 把这些向量构造成矩阵$\mathbf{X}$（$\mathbf{X}$将是$N \times M$阶矩阵）
- ✓ 通过减去每列的平均值来把数据中心化，并令变化后的矩阵为$\mathbf{B}$
- ✓ 计算协方差阵$\mathbf{C} = \frac{1}{N}\mathbf{B}^T\mathbf{B}$
- ✓ 计算$\mathbf{C}$的特征向量和特征值，即$\mathbf{V}^{-1}\mathbf{C}\mathbf{V} = \mathbf{D}$，其中$\mathbf{V}$由$\mathbf{C}$的特征向量组成，$\mathbf{D}$是由特征值组成的$M \times M$阶对角矩阵
- ✓ 把$\mathbf{D}$对角线上元素按降序排列，并对$\mathbf{V}$的列向量作同样排列
- ✓ 去掉那些小于η的特征值，剩下L维的特征向量，构成P输出

核PCA



Q1：能找到合适的线性变换正交坐标轴呢？

A1：核PCA。留到SVM一讲时学习。

PCA与LDA的区别

- PCA是无监督的，LDA是有监督的
- PCA目标投影后数据方差最大，LDA目标组内方差小、组间方差大
- PCA的基础是特征的协方差矩阵，投影后更难被分类
- PCA投影后的坐标是正交的，LDA无需正交
- PCA投影后维度数目与源数据相同(除非略去小的特征值所对应的特征向量)，LDA投影后维度数目与类别数目相同(基于散度矩阵)

大纲

特征选择和降维

线性判别分析 LDA

主成分分析 PCA

独立成分分析 ICA

局部线性嵌入 LLE

ISOMAP

因素分析FA

因素分析 Factor Analysis：假设数据是由多个物理源所产生(并附加测量噪声)，因素分析将数据解释为少数不相关因素的累加。

预处理：

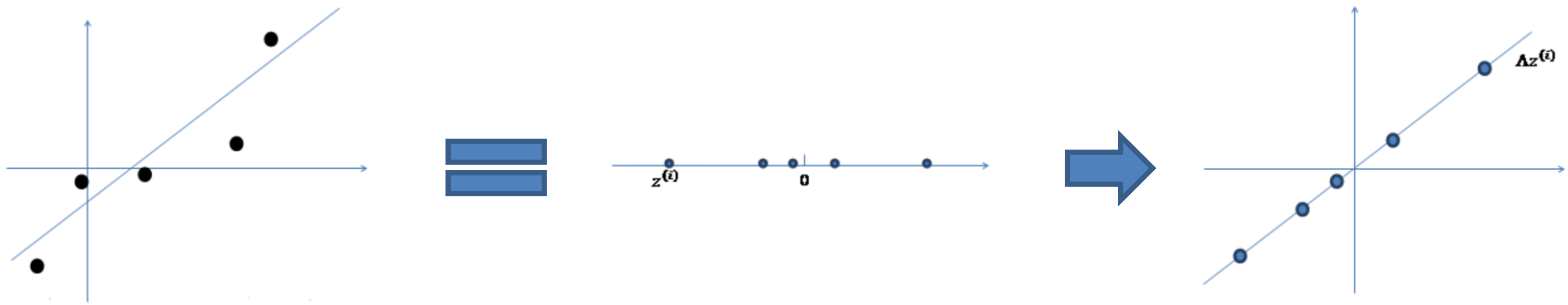
数据集 $N \times M$ 的矩阵 $\mathbf{X}$， $\mathbf{X}$的协方差矩阵为$\sum$，

经PCA预处理后， $\mathbf{b}_j = \mathbf{x}_j - \boldsymbol{\mu}_j, j = 1 \dots M$, 即$E[\mathbf{b}_i] = 0$

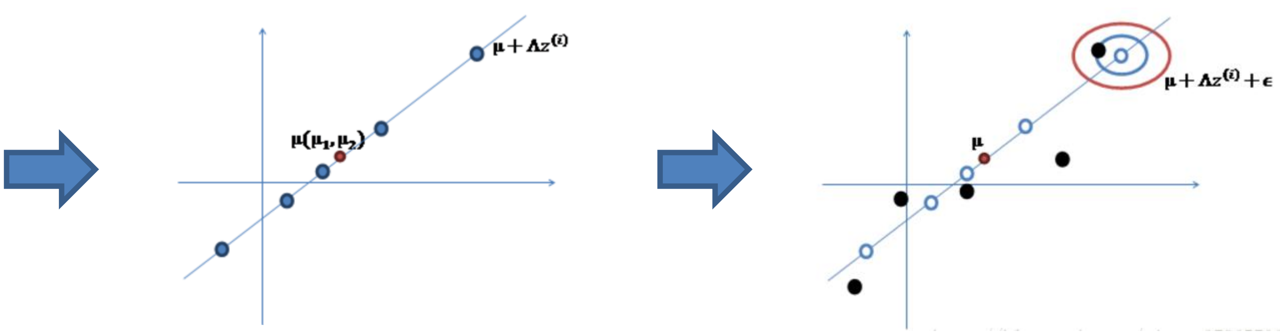
目标模型：

$$\mathbf{X} = \mathbf{W}\mathbf{Y} + \epsilon$$

因素分析FA



原始样本　　　　高斯分布中采样　　　　映射到高维



高维空间里平移　　　　加噪声扰动

因素分析FA

$$\mathbf{X} = \mathbf{W}\mathbf{Y} + \epsilon$$

- 模型的限制：
  - ✓ $i \neq j$， $\operatorname{cov}(\mathbf{b}_i, \mathbf{b}_j) = 0$
  - ✓ $\epsilon \sim (0, \Psi)$， $\Psi_i = var(\epsilon_i)$

$$\sum = \operatorname{cov}(\mathbf{X}) = \operatorname{cov}(\mathbf{W}b + \epsilon) = \mathbf{W}\mathbf{W}^T + \Psi$$

优化：用EM算法进行迭代求解

EM算法将在第五讲学习。

独立成分分析ICA

鸡尾酒会问题 cocktail party problem

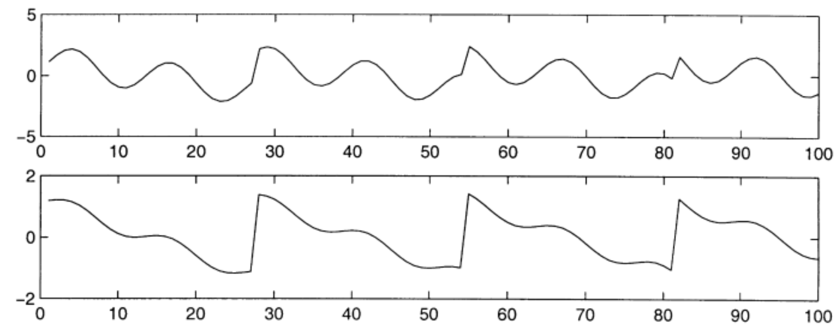
n个人在一个房间开party，房间的不同地方摆放了n个声音接收器，每个接收器在每个时刻同时采集到n个人声音的重叠声音。每个接收器与每个人的距离是不一样的，所以每个接收器接收到的声音的重叠情况也不同。

party结束后，我们得到m个声音样例，每个样例是在具体时刻t，从n个接收器接采集的一组声音数据(一个接收器得到一个数据，所以有n个数据)，如何从这m个样本集分离出n个说话者各自的声音呢？

独立成分分析ICA

盲源分离 blind source separation

假设数据实际上是来自于一些**独立**的潜在物理过程。



观测信号

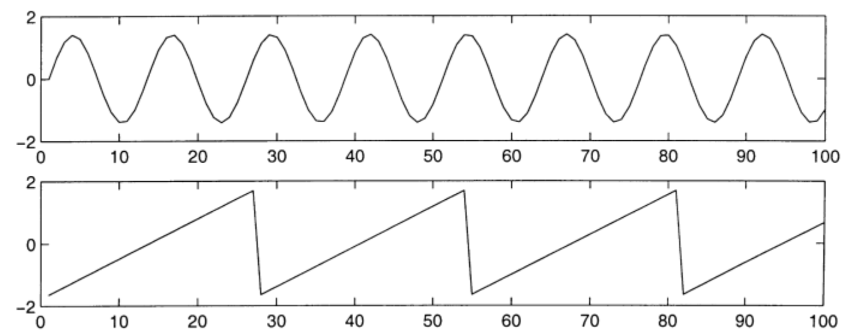
重要(两个不同的数学概念)

不相关 uncorrelated：

$\operatorname{cov}(\mathbf{b}_i, \mathbf{b}_j) = 0$，当$i \neq j$

统计独立 independent：

$E[\mathbf{b}_i, \mathbf{b}_j] = E[\mathbf{b}_i]E[\mathbf{b}_j]$



源信号

独立成分分析ICA

$$\left.\begin{aligned} x_1 &= as_1 + bs_2 \\ x_2 &= cs_1 + ds_2 \end{aligned}\right\}$$

理论上

$$\mathbf{x} = \mathbf{A}\mathbf{s} \Longrightarrow \mathbf{s} = \mathbf{A}^{-1}\mathbf{x}.$$

此处注意：和其他方法不一样的是，ICA并不要求反推出来的源数据误差最小，其约束后面解释。

◘ 基本假设

- ✓ 数据源相互独立，但混合数据不相互独立；
- ✓ 数据源必须是非高斯变量(否则不满足独立性要求)，但混合数据可以服从高斯分布(中心极限定理)；

如果A已知，上述问题很简单。但现在A未知，无法解。只能附加假设。

独立成分分析ICA

$\mathbf{s} = \mathbf{A}^{-1}\mathbf{x} = \boldsymbol{W}\mathbf{x}.$

- 令 $\boldsymbol{z} = \mathbf{A}^{T}\boldsymbol{W}$, 构造 $\boldsymbol{y} = \boldsymbol{W}^{T}\mathbf{x} = \boldsymbol{W}^{T}\boldsymbol{A}\boldsymbol{s} = \boldsymbol{z}^{T}\boldsymbol{s}.$
- $\boldsymbol{y}$是$\boldsymbol{s}$的线性组合，其非高斯性最大化等价$\boldsymbol{z}$中只有一个非0元素(因为任何独立变量相加，将增大其高斯性)
- 因此，最大化$\boldsymbol{W}^{T}\mathbf{x}$的非高斯性将求出其中的一个$\boldsymbol{s}_{i}$

独立成分分析ICA

- 独立成分分析任务
  - ✓ 已知信号为 S，经混和矩阵变换后的信号为：X=AS。
  - ✓ 对交叠信号 X，求解混矩阵W，使 Y=WX 各分量尽量相互独立。
  - ✓ 求解 W 的过程并不一定是近似 A 的逆矩阵，Y 也不是信号 S 的近似，而是为了使 Y 分量之间相互独立。
  - ✓ 目的是从仅有的观测数据 X 出发寻找一个解混合矩阵。

独立成分分析ICA

独立性的评价方式

- ✓ 熵 entropy 的定义 $\mathrm{H}(y) = -\int g(y) \log g(y) dy$
- ✓ 负熵 negentropy $J(y) = H(z) - H(y)$
- ✓ 或用近似方法

$J(y) = (E[G(y)] - E[G(z)])^2$，这里 $g(u) = \frac{1}{a} \log cosh(au)$，其导数 $g'(u) = \tanh(au)$, $1 \le a \le 2$

所有等方差的随机变量中，高斯变量的熵最大。由中心极限定理，若干个有限方差随机变量（无论其服从何种分布）的和，其逼近高斯分布。反言之，源信号比混合信号的非高斯性更强。用负熵度量其非高斯性。

独立成分分析ICA

☐ 主要算法

- ✓ InfoMax 方法(用神经网络使信息最大化)
- ✓ FastICA 方法(固定点算法，寻求 X 分量在 W 上投影$\boldsymbol{W}^T\mathbf{x}$的非高斯最大化)
  - ✓ 预处理：PCA去中心化、球化
  - ✓ 选择初始的权值向量(随机选择)
  - ✓ 计算 $\mathbf{w}^{+}=E\left\{\mathbf{x}g(\mathbf{w}^T\mathbf{x})\right\}-E\left\{g'(\mathbf{w}^T\mathbf{x})\right\}\mathbf{w}$
  - ✓ 令 $\mathbf{w}=\mathbf{w}^{+}/\ ||\mathbf{w}^{+}||$
  - ✓ 若不收敛，返回第三步

ICA与PCA的区别

- PCA 是将原始数据降维并提取出不相关的属性，而 ICA 是将原始数据降维并提取出相互独立的属性。
- PCA 目的是找到这样一组分量表示，使得重构误差最小，即最能代表原事物的特征。ICA 的目的是找到这样一组分量表示，使得每个分量最大化独立，能够发现一些隐藏因素。由此可见，ICA 的条件比 PCA 更强些。
- ICA 要求找到最大独立的方向，各个成分是独立的；PCA 要求找到最大方差的方向，各个成分是正交的。
- ICA 认为观测信号是若干个统计独立的分量的线性组合，ICA 要做的是一个解混过程。而 PCA 是一个信息提取的过程，将原始数据降维，现已成为 ICA 将数据标准化的预处理步骤。

大纲

特征选择和降维

线性判别分析 LDA

主成分分析 PCA

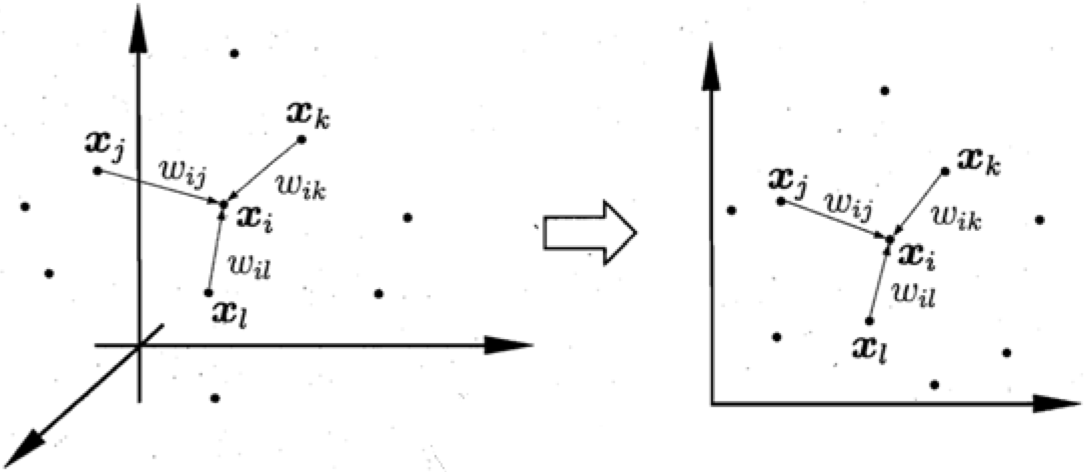
独立成分分析 ICA

局部线性嵌入 LLE

ISOMAP

局部线性嵌入

- Locally Linear Embedding, LLE[Roweis and Saul, Science2000]
  - ✓ 努力去保留相邻数据间的关系
  - ✓ 数据集中的数据用其局部近邻线性近似



$$\varepsilon = \sum_{i=1}^{N}\left(\mathbf{x}_i - \sum_{j=1}^{N}\mathbf{W}_{ij}\mathbf{x}_j\right)^2$$

局部线性嵌入

近邻点的确定

- ✓ 距离法：和当前点的距离小于预先定义的距离d的点集合
- ✓ 个数法：前k个靠得最近的点集合

权重约束

- ✓ 对任意一个点$\mathbf{x}_i$, 如果它离当前点很远，那么$\mathbf{W}_{ij} = 0$
- ✓ $\sum_j \mathbf{W}_{ij} = 1$

$$\varepsilon = \sum_{i=1}^{N} \left(\mathbf{x}_i - \sum_{j=1}^{N} \mathbf{W}_{ij} \mathbf{x}_j \right)^2$$

局部线性嵌入

- LLE的计算方法

✓ 降到L维重新构造损失函数 $= \sum_{i=1}^{N}\left(\mathbf{y}_i - \sum_{j=1}^{L}\mathbf{W}_{ij}\mathbf{y}_j\right)^2$

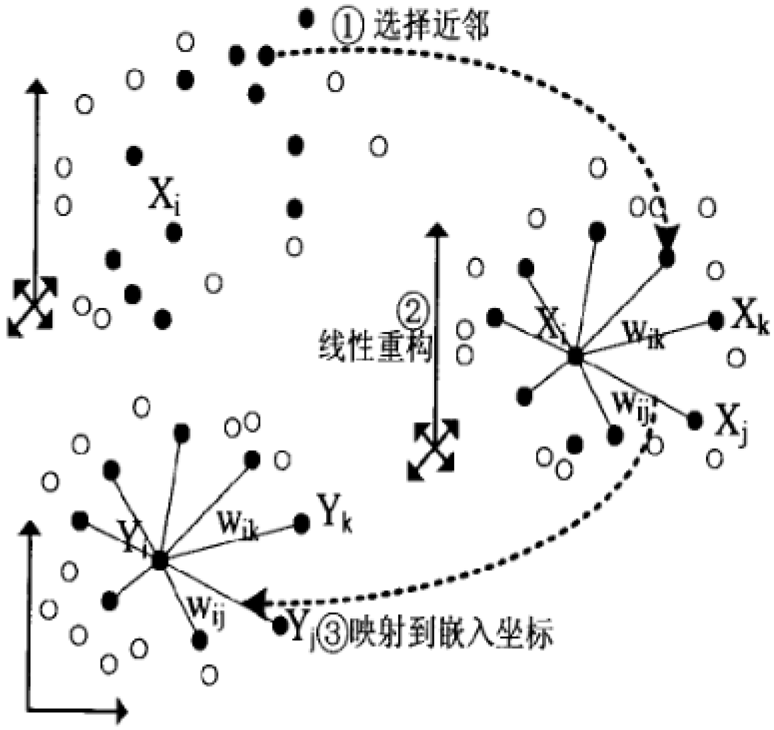
✓ 计算方法要点$\mathbf{M} = (\mathbf{I}-\mathbf{W})^T(\mathbf{I}-\mathbf{W})$

求二次型矩阵$\mathbf{M}_{ij} = \delta_{ij} - \mathbf{W}_{ij} - \mathbf{W}_{ji} + \sum_k \mathbf{W}_{ji}\mathbf{W}_{kj}$的特征值

δ_{ij}克罗内克函数，即$i = j$时，$\delta_{ij} = 1$，否则为0

具体推导过程略。

局部线性嵌入



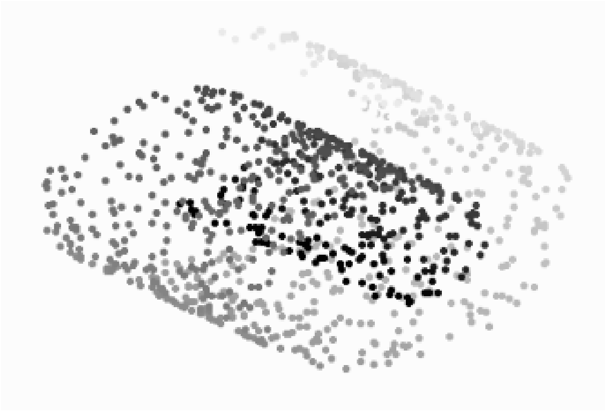
W阵在投影前后保持不变

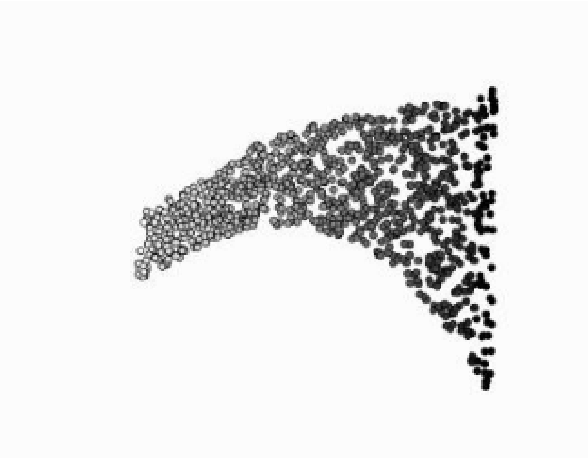
局部线性嵌入

- 找出每个点的临近点（即前k个近的点）
  - ✓ 计算每对点间的距离，找到前k个小的距离
  - ✓ 对于其它点，让$\mathbf{W}_{ij}=0$
  - ✓ 对每个点$\mathbf{x}_i$：创建一个临近点的位置表z_i，计算$\mathbf{z}_i=\mathbf{z}_i-\mathbf{x}_i$
- 根据约束条件计算让原始目标最小的权矩阵**W**
  - ✓ 计算局部协方差$\mathbf{C}=\mathbf{Z}\mathbf{Z}^T$，这儿$Z$是$z_i$组成的矩阵
  - ✓ 利用$\mathbf{CW}=\mathbf{I}$计算$\mathbf{W}$，这儿$\mathbf{I}$是$N\times N$单位矩阵
  - ✓ 对于非临近点，让$\mathbf{W}_{ij}=0$。对$\frac{\mathbf{W}}{\sum\mathbf{W}}$设置其它元素

局部线性嵌入

- 根据约束条件计算让低维优化目标的低维向量$\mathbf{y}_i$
  - ✓ 创建$\mathbf{M} = (\mathbf{I} - \mathbf{W})^T(\mathbf{I} - \mathbf{W})$
  - ✓ 计算$\mathbf{M}$的特征值和特征向量
  - ✓ 根据特征值大小排序
  - ✓ 将第q个最小的特征值所对应的特征向量，放置$\mathbf{y}$第$q+1$行（因为第1行为0）





大纲

特征选择和降维

线性判别分析 LDA

主成分分析 PCA

独立成分分析 ICA

局部线性嵌入 LLE

ISOMAP

等距特征映射ISOMAP

- 等距特征映射(Isometric Feature Mapping, ISOMAP)[Tenenbaum et al., Science2000]
  - ✓ 目标：映射后努力去保留相邻数据间的关系
  - ✓ 方法：通过检查所有点对间的距离和计算全局测地线的方法来最小化全局误差

ISOMAP方法是从MDS方法发展而来

多维缩放MDS

- 多维缩放(Multi-Dimensional Scaling, MDS)
  - ✓ 和PCA算法一样，MDS算法尝试寻找整个数据空间的一个线性近似，从而把数据嵌入到低维空间中
  - ✓ MDS算法在嵌入时尝试保留所有数据点对之间的距离(假定：这些距离已经测出)
  - ✓ 如果空间是欧式空间的话，这两种方法等价

已知高维上样本点两两之间的距离，尝试在低维上(通常是2维，但是可以是任意维)找到一组新的样本点，使降维后两点间的距离与它们在高维上的距离相等

多维缩放MDS

假设

- ✓ 原始数据集 $\mathbf{x_1}, \mathbf{x_2}, \dots, \mathbf{x_N} \in \mathbb{R}^M$
- ✓ 目标数据集 $\mathbf{z_1}, \mathbf{z_2}, \dots, \mathbf{z_N} \in \mathbb{R}^L$， $\mathrm{L} < \mathrm{M}$

目标函数

原空间距离　　　投影后空间距离

- ✓ Kruskal-shephard scaling(最小二乘法)

$$S_{KS}(\mathbf{z_1}, \mathbf{z_2}, \dots, \mathbf{z_N}) = \sum_{i \neq i'} \left(\boxed{d_{ii'}} - \boxed{\| \mathbf{z}_i - \mathbf{z}_{i'} \|}\right)^2$$

- ✓ Sammon mapping

$$S_{SM}(\mathbf{z_1}, \mathbf{z_2}, \dots, \mathbf{z_N}) = \sum_{i \neq i'} \frac{\left(d_{ii'} - \| \mathbf{z}_i - \mathbf{z}_{i'} \|\right)^2}{\boxed{d_{ii'}}}$$

对误差加权

产生非线性

基于相似度的MDS

假设

- ✓ 用数据点间的相似度替代距离
- ✓ 预处理：数据去中心化
- ✓ 相似度：$s_{ii'} = (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_{i'} - \overline{\mathbf{x}})^T$

目标函数

- ✓ 最小化：$\sum_{i \neq i'} \left(s_{ii'} - (\mathbf{z}_i - \overline{\mathbf{z}})(\mathbf{z}_{i'} - \overline{\mathbf{z}})^T \right)^2$

解法
- ✓ 用梯度下降法逼近(线性/非线性)投影(参考第7讲内容)
- ✓ 直接解析计算法线性投影

经典线性MDS的解析推导

解析思路

- ✓ 去中心化原始空间数据$(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M)$，计算点对距离**D**
- ✓ 求投影后空间数据点内积矩阵$\mathbf{B} = \boldsymbol{Z}^T\boldsymbol{Z}$
- ✓ 根据**B**，求解投影后数据$(\mathbf{z}_1, \mathbf{z}_2, \dots \mathbf{z}_M)$
- ✓ 得出投影矩阵

求和后等于0

$$dist_{ij} = \| \mathbf{z}_i - \mathbf{z}_j \| = |\mathbf{z}_i|^2 + \left|\mathbf{z}_j\right|^2 - 2|\mathbf{z}_i|\left|\mathbf{z}_j\right| = b_{ii} + b_{jj} \boxed{-2b_{ij}}$$

因为去中心化

$$\sum_{i=1}^{M} \mathbf{z}_i = 0 \qquad \sum_{i=1}^{M} b_{ij} = \sum_{j=1}^{M} b_{ij} = 0$$

$$\sum_{i=1}^{m} dist_{ij}^2 = b_{11} + b_{22} + ... + b_{mm} + mb_{jj} - 2(b_{1j} + b_{2j} + ... + b_{mj}) = tr(B) + mb_{jj}$$

$$\sum_{j=1}^{m} dist_{ij}^2 = b_{11} + b_{22} + ... + b_{mm} + mb_{ii} - 2(b_{i1} + b_{i2} + ... + b_{im}) = tr(B) + mb_{ii}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{m} dist_{ij}^2 = \sum_{i=1}^{m}(tr(B) + mb_{ii}) = 2m\, tr(B)$$

通过矩阵D计算矩阵B

$$\left.\begin{aligned} dist_{i.}^2 &= \frac{1}{m}\sum_{j=1}^{m} dist_{ij}^2 \\ dist_{.j}^2 &= \frac{1}{m}\sum_{i=1}^{m} dist_{ij}^2 \\ dist_{..}^2 &= \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} dist_{ij}^2 \end{aligned}\right\} \qquad b_{ij} = -\frac{1}{2}(dist_{ij}^2 - dist_{i.}^2 - dist_{.j}^2 + dist_{..}^2)$$

代入验证

$$\begin{aligned} & -\frac{1}{2}(dist_{ij}^2 - dist_{i.}^2 - dist_{.j}^2 + dist_{..}^2) \\ = & -\frac{1}{2}(b_{ii} + b_{jj} - b_{ij} - \frac{1}{m}(2tr(B) + mb_{ii} + mb_{jj} - 2tr(B))) \\ = & b_{ij} \end{aligned}$$

线性MDS的解析算法

算法流程

- ✓ 计算由高维上每对点相似度组成的矩阵$\mathbf{D}$，$\mathbf{D}_{ij} = \| \mathbf{x}_i - \mathbf{x}_j \|^2$
- ✓ 计算$\mathbf{J} = \mathbf{I}_N - \frac{1}{N}$（这儿$I_N$是$N \times N$单位矩阵，$N$是数据点个数）
- ✓ 计算$\mathbf{B} = -\frac{1}{2}\mathbf{J}\mathbf{D}\mathbf{J}^T$ $\Rightarrow$ 对B进行SVD分解
- ✓ 找到$\mathbf{B}$的L个最大特征值λ_i，和相对应特征向量$\boldsymbol{e}_i$
- ✓ 用特征值组成对角矩阵$\mathbf{V}$并且用特征向量作成矩阵P的列向量
- ✓ 计算嵌入$= \mathbf{P}\mathbf{V}^{\frac{1}{2}}$

流形空间

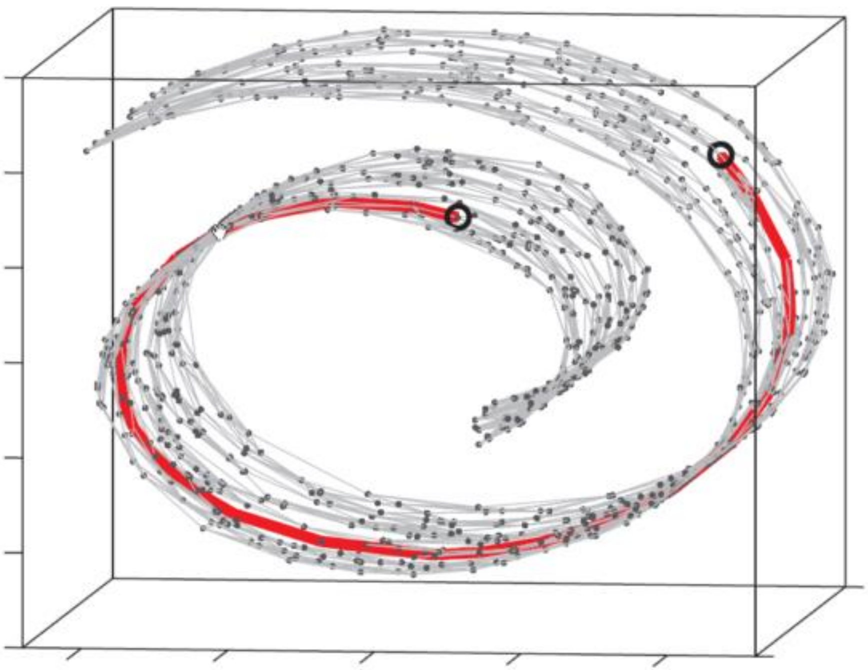
☐ 流形

- ✓ 圆：直角坐标系是二维的；极坐标系是一维的；
- ✓ 任何现实世界中的对象均可以看做是低维流形在高维空间的嵌入(嵌入可以理解为表达)

☐ 距离

- ✓ 球上两点的距离→测地距离，而不是欧式距离
- ✓ 生成数据的拓扑结构

测地线距离



测地线距离

- 高维空间欧式距离不适用
- 计算点对之间测地线距离
- 图最小路径算法
  - ✓ Dijkstra算法
  - ✓ 宽度优先搜索
- 思考
  - ✓ 计算中国各城市的经纬度距离
  - ✓ 在二维地图上表征出来

ISOMAP算法

- 创建所有点对之间的距离
- 确定每个点的临近点，并做成一个权表G
- 通过找最短路径法估计测地距离d_G
- 把经典MDS算法用于一系列d_G

思考和讨论

1. 思考有监督LDA和无监督PCA的区别。
2. 不相关性与独立性的数学差异。
3. 流形的概念。
4. 局部线性嵌入。
5. 思考三种无监督降维PCA, ICA, MDS的异同。

谢 谢 ！