第9节 数据降维

提纲

- 特征选择
- PCA方法
- 流形

为什么进行数据降维

- 有些特征具有相关性，因而在进行学习建模时具有冗余
- 可视化高维数据

监督特征选择

- 问题描述
  - 假设学习 f: X->Y，其中X=<x_1,x_2,…,x_d>, 其中并非每一维都有相关性
- 降维
  - 是从X中选择一个子集，或者给每一个被选特征一个权重
  - 找到一个能够表达问题的最好的子集

监督特征选择-选择特征集合

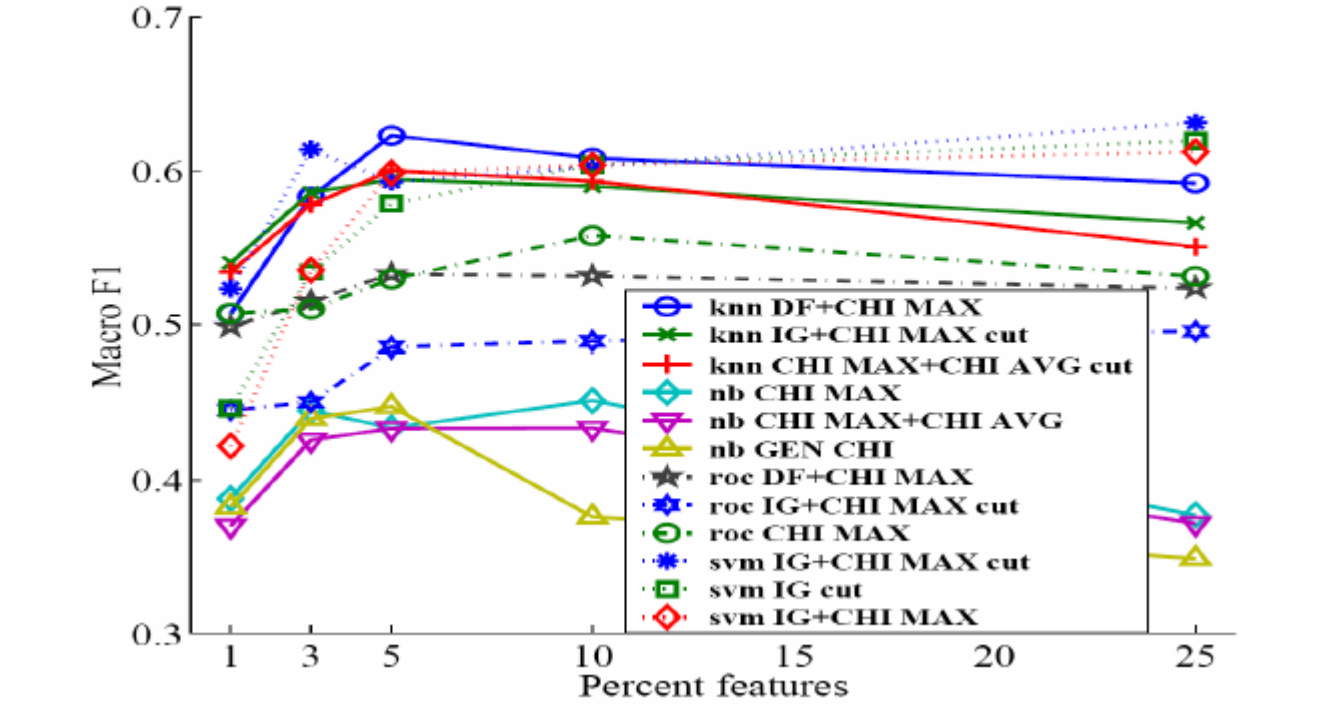
- 一般做法
  - Forward selection: 选择分类测试中得分最高的d个特征
    - 选择单个分值最高的特征X_k,
    - 在已选的特征的基础上，给剩余特征打分
      - E.g., 评估（$X_i \mid X_k$）= $E(X_i, Y \mid X_k)$
      - E.g., 评估（ $X_i \mid X_k$ ）= Accuracy(Predicting Y from X_i and X_k)
    - 重复此过程得到特征集合

监督特征选择

- 例子

Approximately 10^5 words in English

[Rogati&Yang, 2002]



提纲

- 特征选择
- PCA方法
- 流形

PCA

主成份分析（Principal Component Analysis, PCA ）是一种利用线性映射来进行数据降维的方法，并去除数据的相关性；且最大限度保持原始数据的方差信息。

PCA

一项著名的工作是美国的统计学家斯通(stone)在1947年关于国民经济的研究。他曾利用美国1929一1938年各年的数据，得到了17个反映国民收入与支出的变量要素，例如雇主补贴、消费资料和生产资料、纯公共支出、净增库存、股息、利息外贸平衡等等。

在进行主成份分析后，以97.4%的精度，用三新变量就取代了原17个变量的**方差**信息。根据经济学知识，斯通给这三个新变量分别命名为总收入f_1、总收入变化率f_2和经济发展或衰退的趋势f_3。

PCA 的直观解释

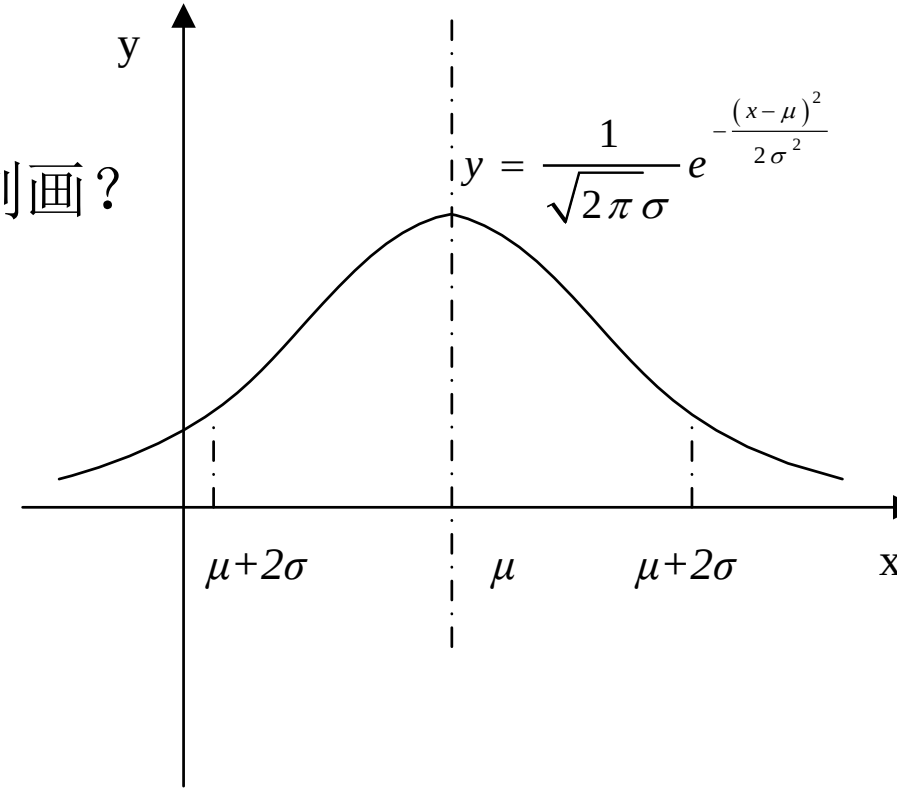
利用3维向量能够保持原始17维向量，97.4%的方差信息

（是在低维空间能够尽可能多保持原始空间数据的方差）

我们所讨论的问题中都有一个近似的假设，假定数据满足**高斯分布或者近似满足高斯分布**

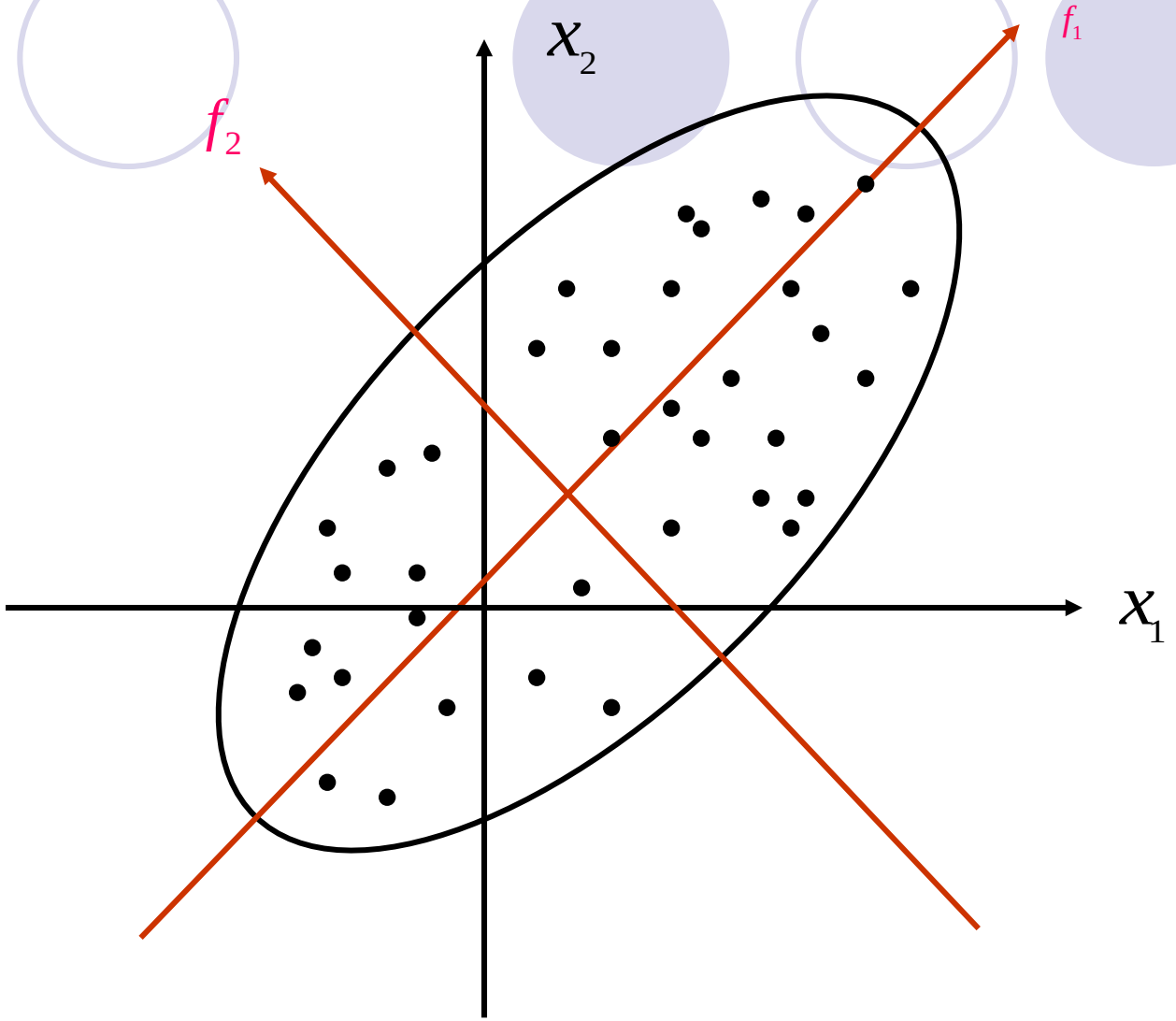
问题：高斯分布需要什么参数刻画？

均值，方差（离散程度）



平移、旋转坐标轴

主成份分析的几何解释

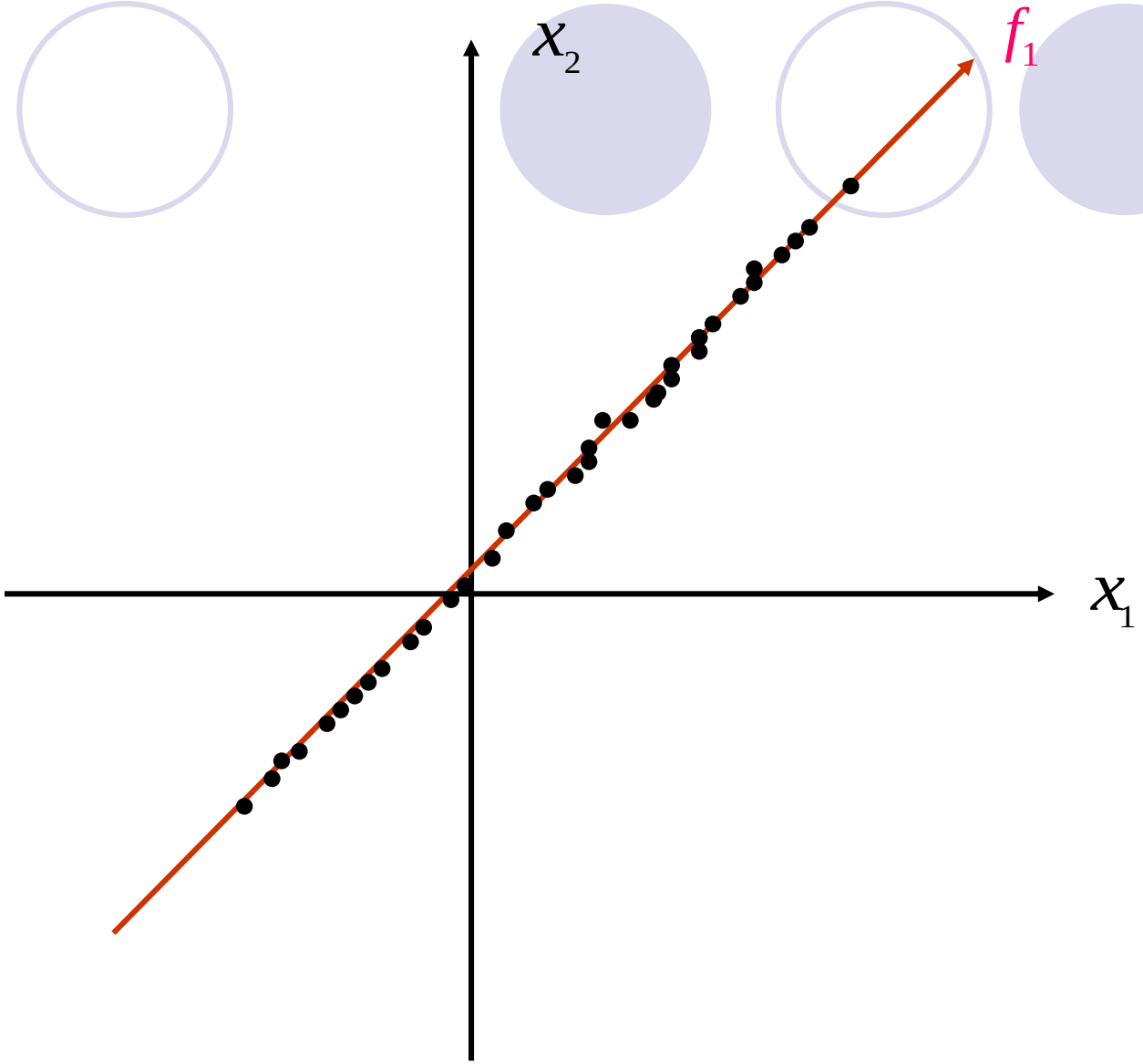


设有n个样本，每个样本有二维即x_1和x_2，在由x_1和x_2所确定的二维平面中，n个样本点所散布的情况如椭圆状。

PCA 的直观解释

- ✓ 由图可以看出这n个样本点沿着f_1 轴方向有最大的离散性，这是第一个主成份
- ✓ 为了去掉相关性，第二个主成份应该正交于第一个主成份
- ✓ 如果只考虑f_1和f_2 中的任何一个，那么包含在原始数据中的信息将会有损失。
- ✓ 根据系统精度的要求，可以只选择f_1

主成份分析的几何解释



✓主成份分析试图在力保数据信息丢失最少的原则下，去除数据的相关性，对高维空间的数据降维处理。

PCA 的计算

- 假设我们所讨论的实际问题中，X是p维变量，记为X_1，X_2，…，X_p，PCA就是要把这p个变量的问题，转变为讨论p个变量的线性组合的问题
- 这些新的分量f_1，f_2，…，f_k(k≤p)，按照保留主要信息量的原则充分反映原变量的信息，并且相互独立。

$$f_1 = u_{11}x_1 + u_{21}x_2 + \cdots + u_{p1}x_p$$

$$f_2 = u_{12}x_1 + u_{22}x_2 + \cdots + u_{p2}x_p$$

$$\cdots\cdots$$

$$f_k = u_{1k}x_1 + u_{2k}x_2 + \cdots + u_{pk}x_p$$

PCA 的计算

两个线性代数的结论

1、若A是p阶正定或者半正定实阵，则可以找到正交阵U，使

$$\mathbf{U}^T\mathbf{A}\mathbf{U}=\begin{bmatrix}\lambda_1 & 0 & \cdots & 0\\ 0 & \lambda_2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \lambda_p\end{bmatrix}_{p\times p}$$

其中 $\lambda_i, i=1.2.\cdots p$ 是A的特征根。

PCA 的计算

2、若上述矩阵的特征根所对应的单位特征向量为

$$\mathbf{u}_1, \cdots, \mathbf{u}_p$$

令 $$\mathbf{U} = (\mathbf{u}_1, \cdots, \mathbf{u}_p) = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1p} \\ u_{21} & u_{22} & \cdots & u_{2p} \\ \vdots & \vdots & & \vdots \\ u_{p1} & u_{p2} & \cdots & u_{pp} \end{bmatrix}$$

则实对称阵 $\mathbf{A}$ 属于不同特征根所对应的特征向量是正交的，即有 $\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathbf{I}$

PCA 的计算

协方差矩阵

$$\Sigma_x = \left(\frac{1}{n} \sum_{i=1}^{n} (x^{(i)} - \bar{x})(x^{(i)} - \bar{x})^T \right)_{p \times p}$$

$$x^{(i)} = \left(x_1^{(i)}, \ x_2^{(i)}, \cdots, \ x_p^{(i)} \right), i = 1,2, \cdots, \ n$$

PCA 的计算

第一步：由X的协方差阵Σ_x，求出其特征根，即解方程$|\boldsymbol{\Sigma}-\lambda\mathbf{I}|=0$，可得特征根 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$

第二步：求出协方差矩阵分别所对应的特征向量U_1，U_2，…，U_p，

$$\mathbf{U}=(\mathbf{u_1},\cdots,\mathbf{u_p})=\begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1p} \\ u_{21} & u_{22} & \cdots & u_{2p} \\ \vdots & \vdots & & \vdots \\ u_{p1} & u_{p2} & \cdots & u_{pp} \end{bmatrix}$$

PCA 的计算

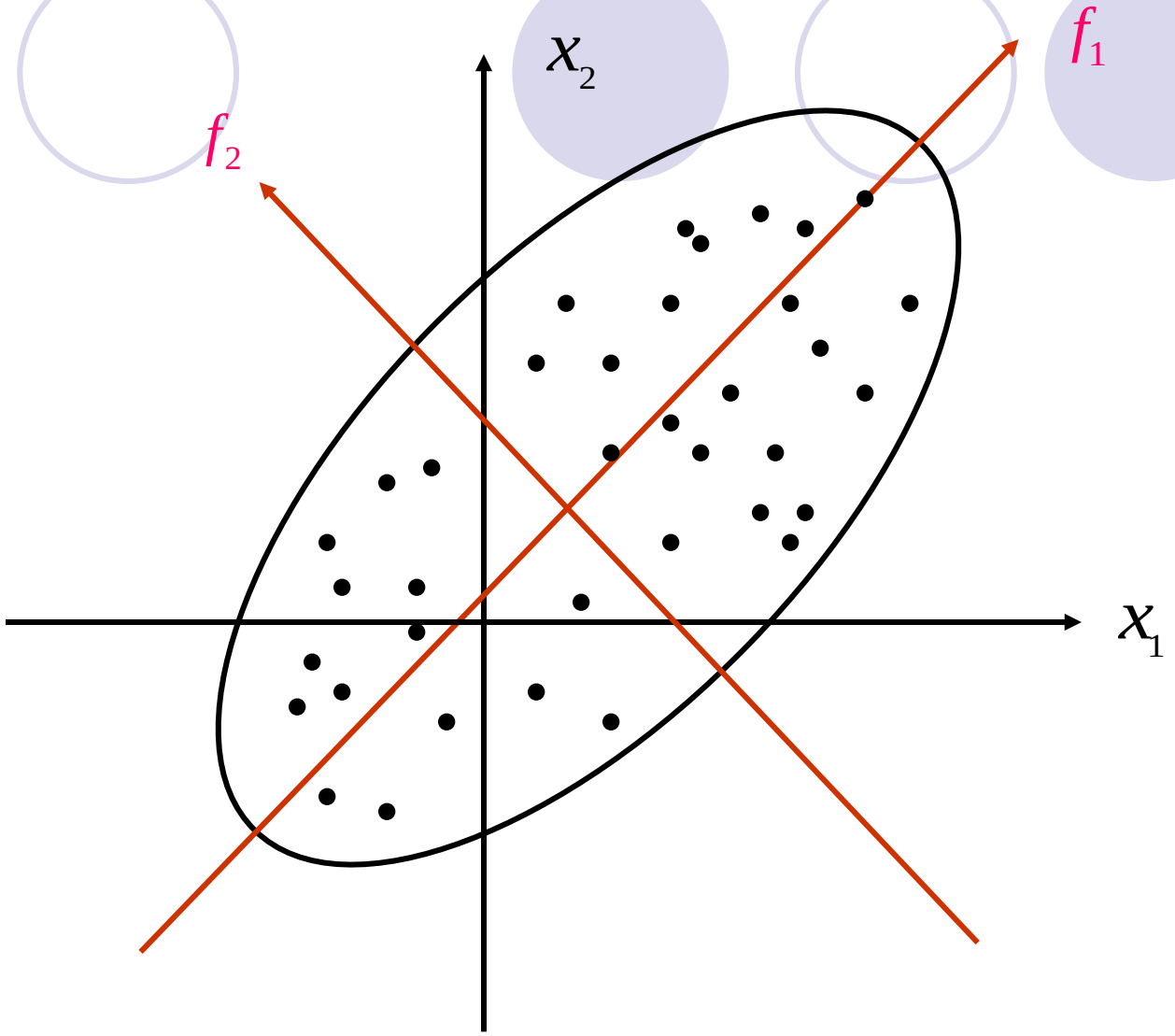
第三步：任一个样本的正交变换

$$f_{p\times 1} = f_{p\times p} \cdot x_{p\times 1}$$

$$f_1 = u_{11}x_1 + u_{21}x_2 + \cdots + u_{p1}x_p$$
$$f_2 = u_{12}x_1 + u_{22}X_2 + \cdots + u_{p2}x_p$$
$$\cdots\cdots$$
$$f_k = u_{1k}x_1 + u_{2k}x_2 + \cdots + u_{pk}x_p$$
$$\cdots\cdots$$
$$f_p = u_{1p}x_1 + u_{2p}x_2 + \cdots + u_{pp}x_p$$

平移、旋转坐标轴

主成份分析的几何解释



设有n个样本，每个样本有二维即x_1和x_2，在由x_1和x_2 所确定的二维平面中，n个样本点所散布的情况如椭圆状。

PCA 的计算

第四步：从所有变换成份中取K个主成分

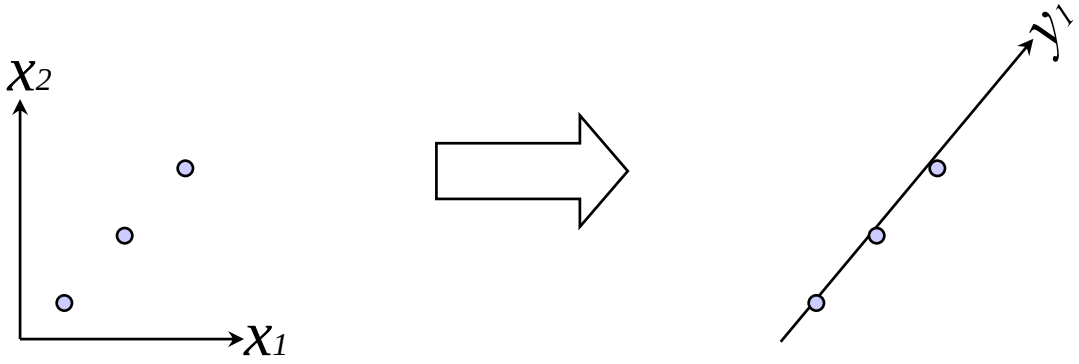
$$f_{k\times 1} = U_{k\times p} \cdot x_{p\times 1}$$

$$f_1 = u_{11}x_1 + u_{21}x_2 + \cdots + u_{p1}x_p$$
$$f_2 = u_{12}x_1 + u_{22}x_2 + \cdots + u_{p2}x_p$$
$$\cdots\cdots$$
$$f_k = u_{1k}x_1 + u_{2k}x_2 + \cdots + u_{pk}x_p$$

注：计算k个主成份之前。将原始数据的中心化值：

$$x^{(i)} = x^{(i)} - \overline{x}$$

PCA 的计算的例子



2D data

1D data

3个点(1,1)(2,2)(3,3)，特征向量？特征值？

PCA 的中主成分个数的选择

1）贡献率：第i个主成份的方差在全部方差中所占比重 $\lambda_i \Big/ \sum_{i=1}^{p} \lambda_i$ 称为贡献率，反映了原来i个特征向量的信息，有多大的提取信息能力。

2）累积贡献率：前k个主成份共有多大的综合能力，用这k个主成份的方差和在全部方差中所占比重

$$\sum_{i=1}^{k} \lambda_i \Big/ \sum_{i=1}^{p} \lambda_i$$

来描述，称为累积贡献率。

进行主成份分析的目的之一是希望用尽可能少的主成分f_1，f_2，…，f_k（k≤p）代替原来的P维向量。

PCA 给人脸数据降维

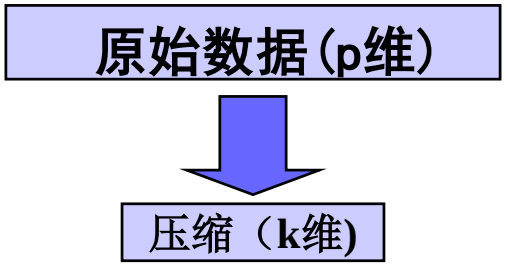
PCA 给人脸数据降维

- 计算过程为：

计算样本均值 m → 中心平移每个训练样本 x_i → 计算训练集合的样本协方差矩阵 → 对协方差矩阵进行特征值分解 → 取协方差矩阵的K个特征向量形成变换矩阵，进行降维

从p维空间到k维空间的投影

$(k < p)$！

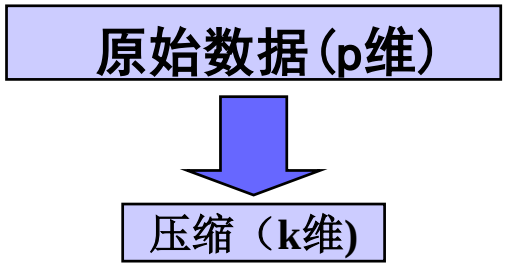


PCA：用于人脸降维

- 按照其所相应的特征值的大小对特征向量排序
- 选择头*k*个对应最大特征值的特征向量构成变换矩阵U_{pxk}

从p维空间到k维空间的投影

$(k < p)$！



$$y = U^T x$$

PCA 给人脸数据降维

特征人脸

原始数据的维数为64x64=4096

数据降维到8个主成份的可视化表示

提纲

- 特征选择
- PCA方法
- 流形

流形（Manifold）

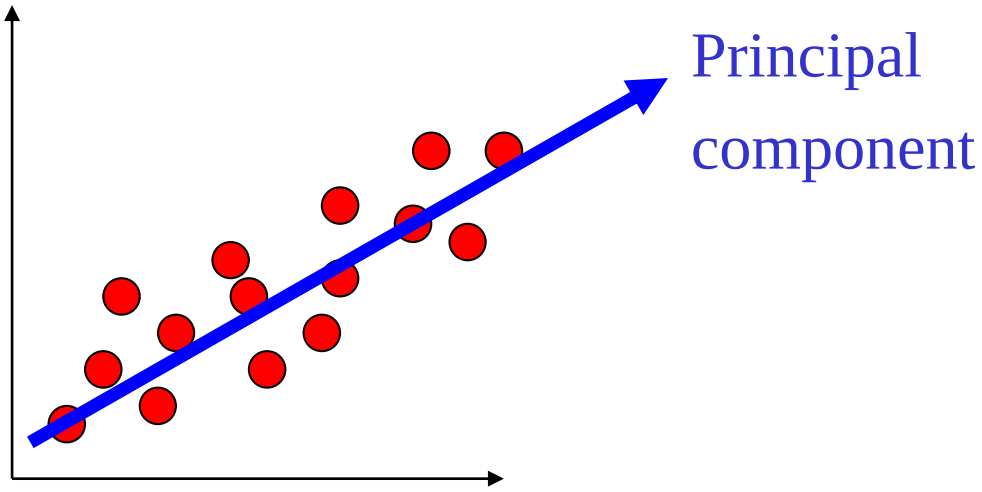
所谓**流形**（manifold）就是一般的几何对象的总称。**流形**包括各种维数的曲线曲面等。和一般的降维分析一样，**流形学习**把一组在高维空间中的数据在低维空间中重新表示。和以往方法不同的是，在**流形学习**中有一个假设，就是所处理的数据采样于一个潜在的**流形**上，或是说对于这组数据存在一个潜在的**流形**。

流形（Manifold）

- 降维
  - 特征选择:依据某一标准选择性质最突出的特征
  - 特征变换:经已有特征的某种变换获取约简特征
- 数据可视化和数据挖掘分析也需要降维
  - 通常降到2维或3维
  - 流形降维来观测数据的内在形状

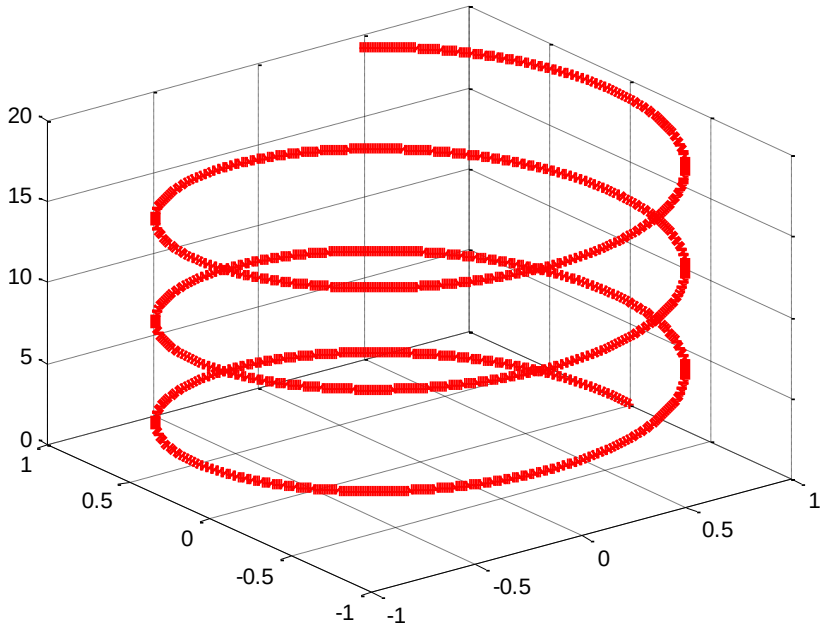
线性方法: (PCA)

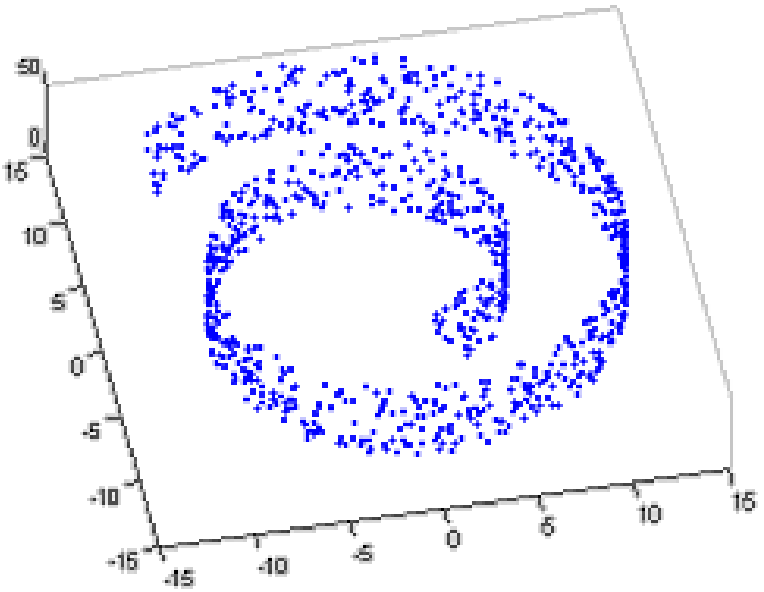
- PCA的目的：寻找能够表示采样数据的最好的投影子空间.
- PCA的求解：对样本的协方差矩阵进行特征值分解, 所求子空间为过样本均值, 以最大特征值所对应的特征向量为方向的子空间.



线性方法的不足

- 数据特征并不具备简单性
  - 例如: PCA 不能发现螺旋型数据，适合高斯分布
  - KPCA或许能解决主曲线问题，但曲面，立体？





流形（Manifold）

1 许多高维采样数据都是由少数几个隐含变量所决定的, 如人脸采样由光线亮度, 人离相机的距离, 人的头部姿势, 人的脸部肌肉等因素决定.

2 从认知心理学的角度, 心理学家认为人的认知过程是基于认知流形和拓扑连续性的.

R

流形（Manifold）-几种方法

- **局部线性嵌入(LLE).**

S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, vol. 290, pp. 2323--2326, 2000.

- **等距映射(Isomap).**

J.B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, vol. 290, pp. 2319--2323, 2000.

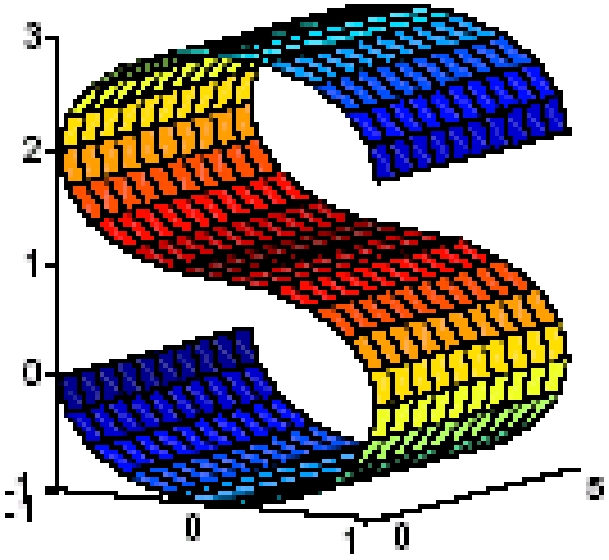
- 拉普拉斯特征映射(Laplacian Eigenmap).

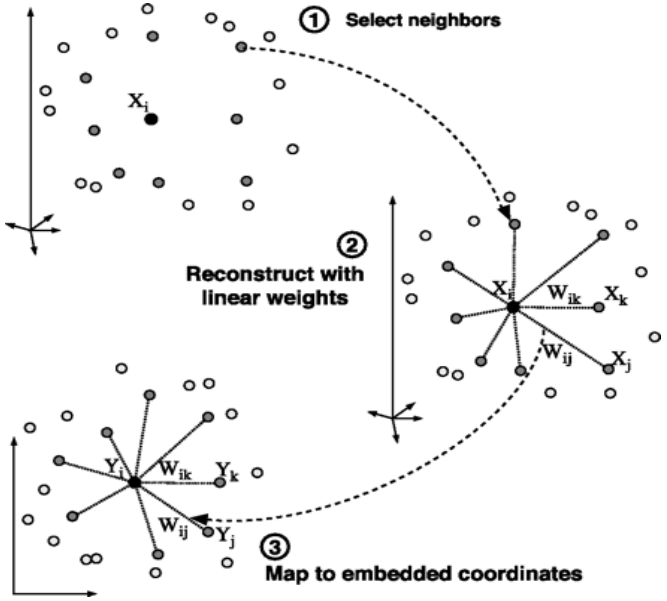
M. Belkin, P. Niyogi, Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation*, Vol. 15, Issue 6, pp. 1373 –1396, 2003 .

流形（Manifold）-LLE

- 前提假设：采样数据所在的低维流形在局部是线性的,即每个采样点可以用它的近邻点线性表示.

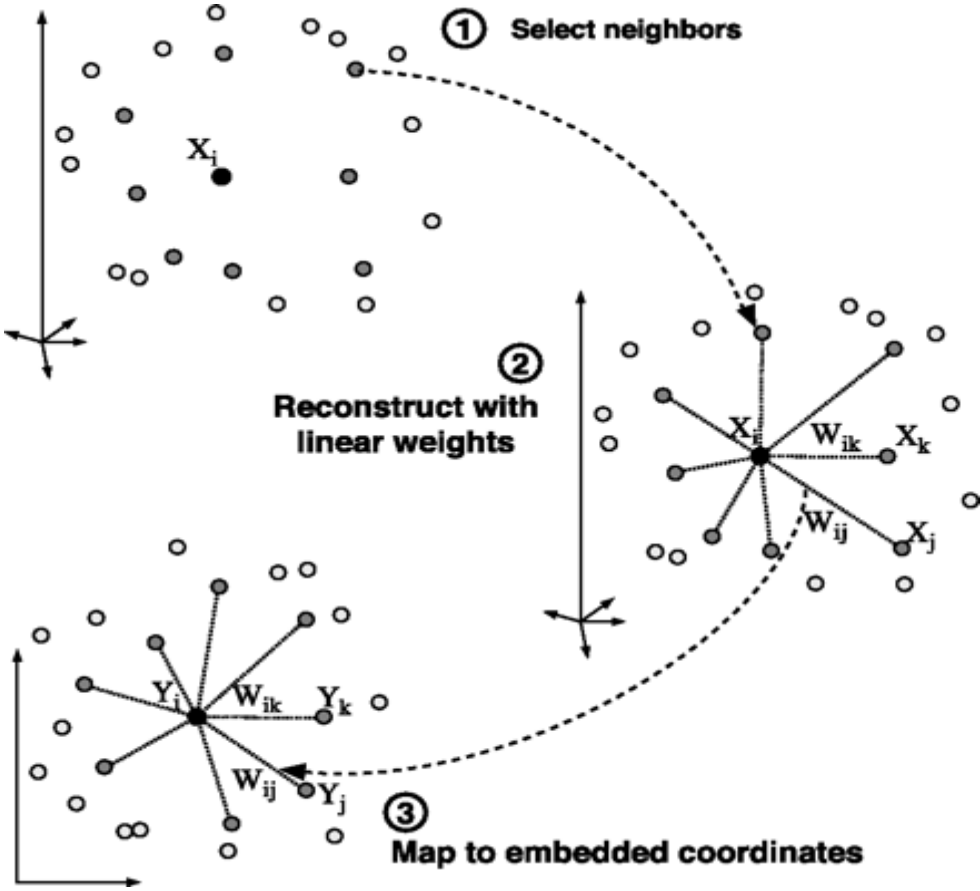
- 学习目标：在低维空间中保持每个邻域中的权值不变, 即假设嵌入映射在局部是线性的条件下, 最小化重构误差.





流形（Manifold）-LLE

- **假设即策略**
- 采样数据所在的低维流形在局部是线性的,即每个采样点可以用它的近邻点线性表示.
- 在低维空间中保持每个邻域中的权值不变。



流形（Manifold）-LLE 最小化重构误差

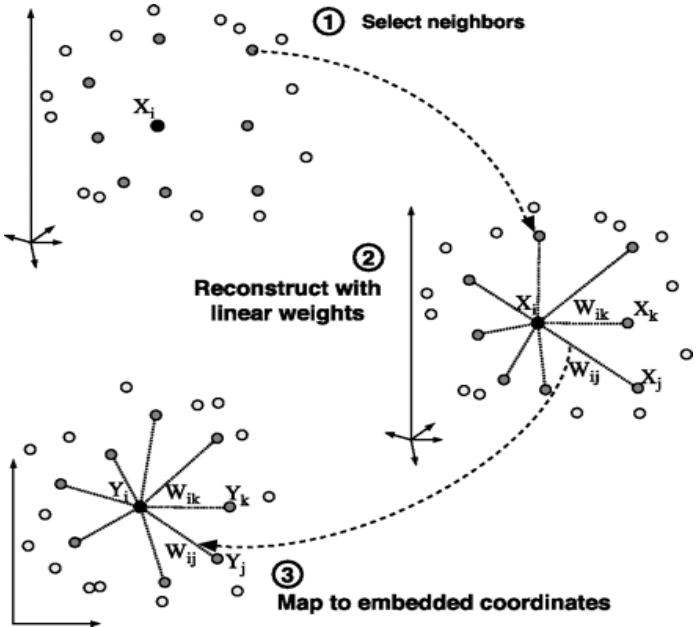
1. 计算每一个点 X_i的近邻点, 一般采用K 近邻或者 ε 邻域.

2 .计算权值 W_{ij}, 使得把 $x^{(i)}$用它的K个近邻点线性表示的误差最小, 即通过最小化 $\left\| x^{(i)} - W_{ij} x^{(j)} \right\|$ 来求出 W_{ij}.

3 .保持权值 W_{ij} 不变, 求 $x^{(i)}$在低维空间的映射 $y^{(i)}$, 使得低维重构误差最小.

$$\min \sum_i | x^{(i)} - W_{ij} x^{(j)} |^2$$

$$\min \sum_i | y^{(i)} - W_{ij} y^{(j)} |^2$$



流形（Manifold）-LLE 求解

1. 计算每一个点X_i 的近邻点（可采用K-NN的方法）.

2 .对于点X_i 和它的近邻点的权值W_{ij} ，

计算局部协方差矩阵 $C_{jk}=(X-\eta_j)^T \bullet (X-\eta_k)$, η_k为X的近邻点.

最小化 $\left\| x^{(i)} - W_{ij} x^{(j)} \right\|$ 得到：

$$w_j = \frac{\sum_k C_{jk}^{-1}}{\sum_{lm} C_{lm}^{-1}}$$

流形（Manifold）-LLE 求解

3. 求解低维流形嵌入

目标函数：$$\phi(Y)=\sum_i \left\| y^{(i)} - W_{ij} y^{(j)} \right\|^2 = \sum_{i,j} M_{ij} \left(\left(y^{(i)} \right)^T \cdot y^{(i)} \right)$$

$$M = (I - W)^{\mathrm{T}} (I - W)$$

中心化、归一化：$$\sum_i y^{(i)} = 0 \qquad \frac{1}{N} \sum_i \left(y^{(i)} \right)^T \cdot y^{(i)} = I$$

单位协方差矩阵

$$MY = \lambda Y$$

转化为求特征值与特征向量的问题，用PCA求解，低维嵌入Y是 *M* 的最小 k 个特征值对应的特征向量.

流形（Manifold）-LLE 最小化重构误差

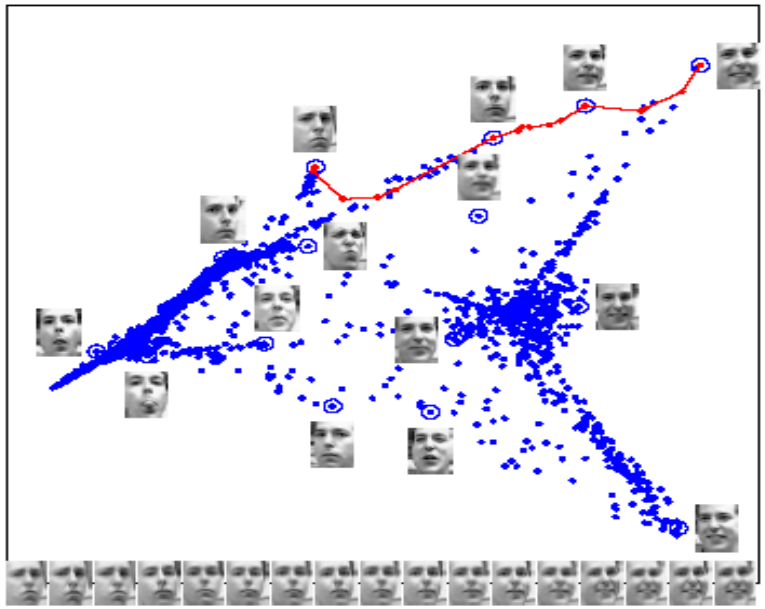
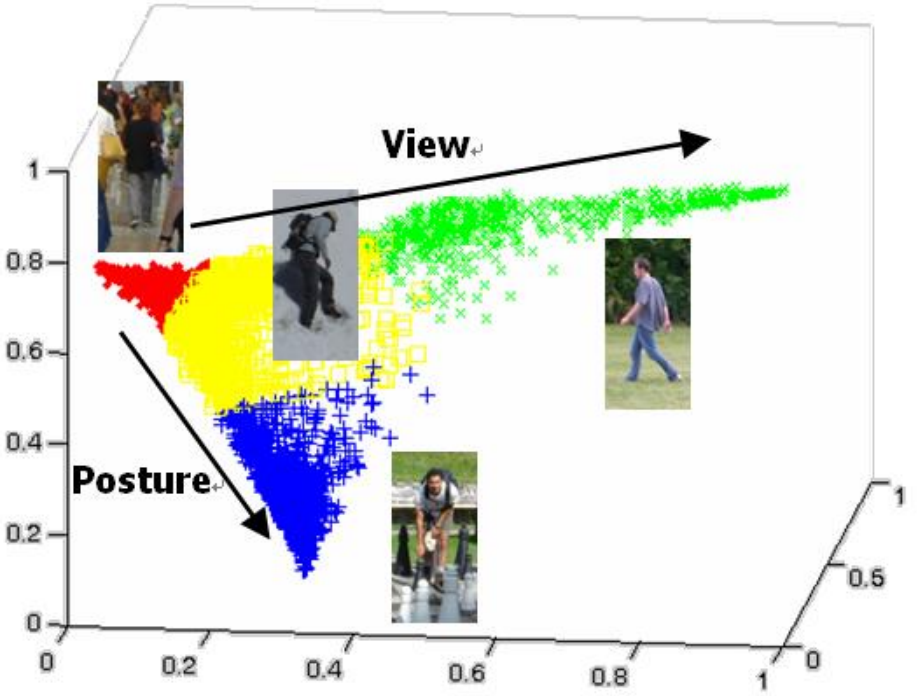


Figure 6: Images of faces mapped into the embedding space described by the first two coordinates of LLE, using $K = 12$ nearest neighbors. Representative faces are shown next to circled points at different points of the space. The bottom images correspond to points along the top-right path, illustrating one particular mode of variability in pose and expression. The data set had a total of $N = 1965$ grayscale images at 20×28 resolution ($D = 560$).



流形（Manifold）

- 流形学习作为一种非线性降维或数据可视化的方法已经在图像处理如人脸图像,手写数字图像, 语言处理方面得了利用.
- 将其作为一种监督的学习方法用于模式识别, 虽然有研究者涉足, 但是目前在这方面的工作还很有限.

流形（Manifold）相关的故事

Tenenbaum根本不是做与数据处理有关算法的人，他是做计算认知科学（computational cognition science）的。在做这个方法的时候，他还在stanford，2年就去了MIT开创一派，成了掌门人，他的组成长十分迅速。但是有趣的 ,在Isomap之后，他包括他在MIT带的学生就从来再也没有做过类似的工作。

他在参加 UCLA Alan Yuille 组织的一个summer school上说，我们经常忘了做研究的原始出发点是什么。他做Isomap就是为了找一个好的visual perception的方法，他还坚持了他的方向和信仰，computational cognition，他没有随波逐流。而由他引导起来的 manifold learning 却快速的发展成了一个新的方向。