模式识别与计算机视觉

距离度量

稀疏

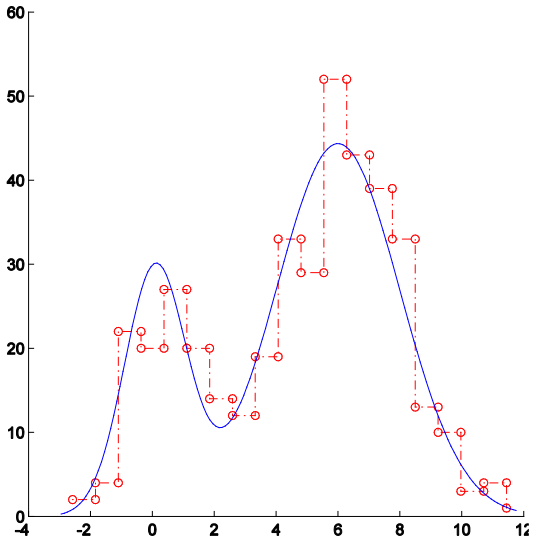
张振宇

南京大学智能科学学院

2025

非参数估计

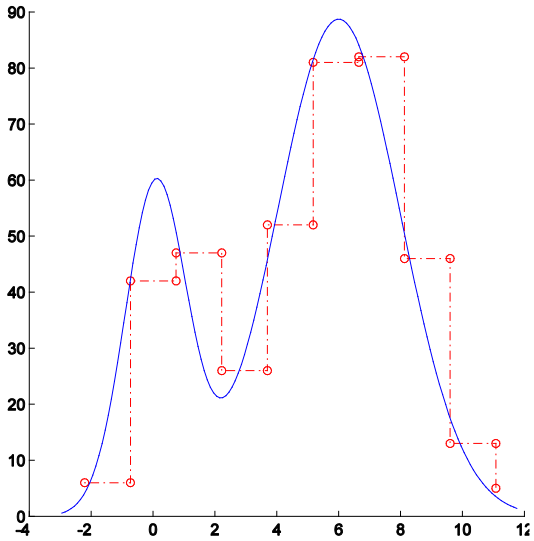
- ✓ 常用的参数形式基本都是单模single modal的，不足以描述复杂的数据分布：即应该直接以训练数据自身来估计分布
  - 例如直方图histogram，基于计数counting

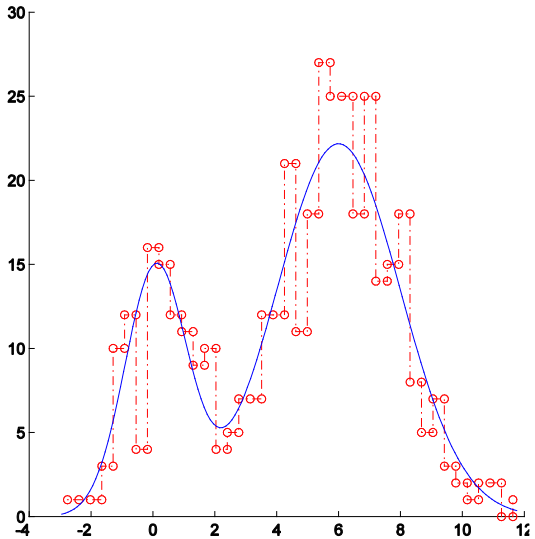


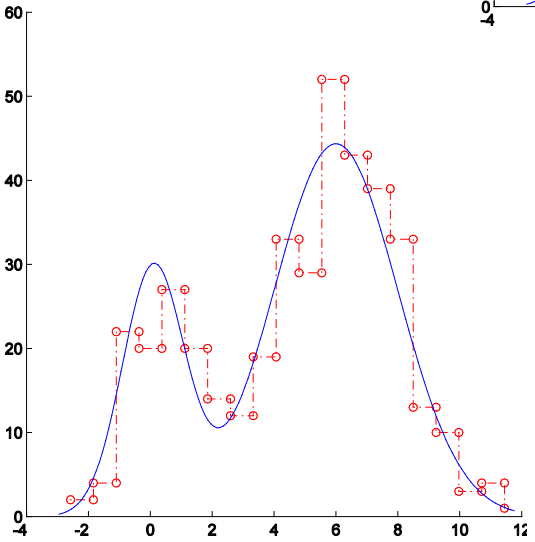
有很多问题：

- 多维怎么办？
- 怎么确定bin的个数？
- 连续？
- 需要保存数据吗？

Bin个数（宽度）的影响


90
80
70
60
50
40
30
20
10
0
-4
-2
0
2
4
6
8
10


30
25
20
15
10
5
0
-4
-2
0
2
4
6
8
10


60
50
40
30
20
10
0
-4
-2
0
2
4
6
8
10

维数灾难

- ✓ Curse of dimensionality
  - 以直方图为例，需要保存的参数是什么？
  - 如果每维n个参数，那么d维应该保存多少个参数？
  - 如果$n = 4, d = 100$，那么应该保存多少个参数？
  - $4^{100} = 2^{200} \approx 10^{60}$！那么，需要多少样例来学习？
    - $1G = 10^9$
- ✓ 不仅局限于直方图、非参数估计，在参数估计、以及很多其他统计学习方法中都是如此

Kernel density estimation

- ✓ KDE，注意这里的kernel与SVM中的核含义不完全一致，其要求的条件也不完全一致
- ✓ 核密度估计的基本形式

$$p(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}\mathrm{K}\left(\frac{x - x_i}{h}\right)$$

- ✓ 核函数K（非负、积分为1、对称）
- ✓ 带宽h，控制核的宽度，决定估计的平滑程度。

Kernel density estimation

- ✓ KDE，注意这里的kernel与SVM中的核含义不完全一致，其要求的条件也不完全一致
- ✓ 举例：Parzen window（一维，使用高斯核）

$$p(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{(2\pi h^2)^{\frac{1}{2}}}\exp\left(-\frac{|x-x_i|^2}{2h^2}\right)$$

问题：

- 连续吗？
- 多维：多个维度乘积
- 需要保存数据吗？
  - 存储和计算实际代价大
  - 无穷多的参数
- 怎么确定h？

$$\hat{f}(\mathbf{x}) = \frac{1}{n\det(\mathbf{H})}\sum_{i=1}^{n}K\left(\mathbf{H}^{-1/2}(\mathbf{x}-\mathbf{X}_i)\right)$$

$$\hat{f}(\mathbf{x}) = \frac{1}{nh_1h_2\cdots h_d}\sum_{i=1}^{n}\prod_{j=1}^{d}K\left(\frac{x_j - X_{ij}}{h_j}\right)$$

决策

Decision,或预测prediction

决策、预测

- ✓ 当inference完成之后，如果给定输入$\boldsymbol{x}$
  - 应当给出什么样的输出？
  - 怎么给出？
- ✓ 点估计：
  - 根据参数得到后验概率$p(y|\boldsymbol{x};\boldsymbol{\theta})$
  - 根据其给出结果（如分类，如何输出？）
- ✓ Bayesian decision
  - 输出，也是一个随机变量，称为预测分布predictive distribution
  - 结果通常根据其期望决定，同时还可以给出方差

点估计下的例子

- ✓ 在0-1风险时，选择后验概率最大的那个类别
  $\text{argmax}_i p(y = i|\boldsymbol{x}; \boldsymbol{\theta})$
- ✓ 在discriminant function观点下，可以定义函数

$$g_i(\boldsymbol{x}) = p(y = i|\boldsymbol{x}; \boldsymbol{\theta}) = \frac{p(\boldsymbol{x}|y = i; \boldsymbol{\theta})p(y = i)}{p(\boldsymbol{x}; \boldsymbol{\theta})}$$

- ✓ 或者定义为 $g_i(\boldsymbol{x}) = p(\boldsymbol{x}|y = i; \boldsymbol{\theta})p(y = i)$，为什么？
- ✓ 或者定义为

$$g_i(\boldsymbol{x}) = \ln\big(p(\boldsymbol{x}|y = i; \boldsymbol{\theta})\big) + \ln(p(y = i))$$

距离度量

Metric

距离度量和相似度度量

- ✓ 基于训练集$\{\boldsymbol{x}_i, y_i\}_{i=1}^{n}$，得到映射：$\mathcal{X} \mapsto \mathcal{Y}$
- ✓ 从数据中任取两个样例$\boldsymbol{x}, \boldsymbol{y}$，如何判断相互关系？
  - K近邻，PCA，FLD——使用距离作为不相似性度量
  - KDE——使用核函数作为相似性度量

- ✓ 相似度高，不相似度就低；反之亦然
- ✓ RBF（高斯）核：
- ✓ $$\mathrm{K}_{RBF}(\boldsymbol{x}, \boldsymbol{y}) = \exp(-\gamma \|\boldsymbol{x} - \boldsymbol{y}\|^2)$$
- ✓ 为什么RBF核可以度量相似度？

距离度量

- 距离是一种度量不相似度的指标
- 度量（metric）是一个函数
- 欧氏距离是最常用的距离度量：
- $d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\| = \sqrt{\sum_{i=1}^{d}(x_i - y_i)^2}$
- 满足以下性质：
- i. $d(\boldsymbol{x}, \boldsymbol{y}) \geq 0$ 非负性
- ii. $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$ 对称性
- iii. $d(\boldsymbol{x}, \boldsymbol{y}) = 0 \leftrightarrow \boldsymbol{x} = \boldsymbol{y}$ 同一性
- iv. $d(\boldsymbol{x}, \boldsymbol{z}) \leq d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{y}, \boldsymbol{z})$ 三角不等式

向量范数和度量

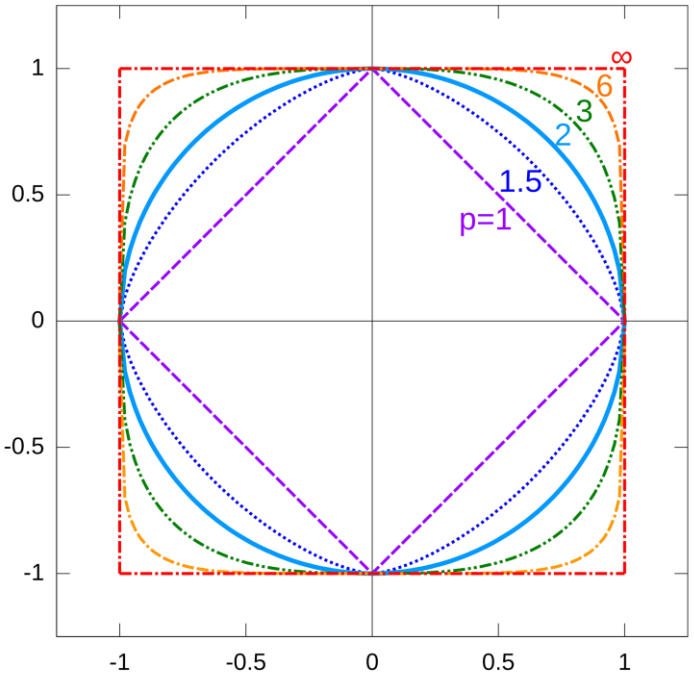
- 欧氏距离和向量范数的关系： $d(\boldsymbol{x},\boldsymbol{y}) = \|\boldsymbol{x}-\boldsymbol{y}\|$
- 只考虑实向量，向量范数f满足三个性质：
- i. $f(c\boldsymbol{x}) = |c|f(\boldsymbol{x})$ 齐次性
- ii. $f(\boldsymbol{x}) = 0 \leftrightarrow \boldsymbol{x} = 0$ 非负性
- iii. $f(\boldsymbol{x}+\boldsymbol{y}) \leq f(\boldsymbol{x}) + f(\boldsymbol{y})$ 三角不等式
- 容易证明 $f(\boldsymbol{0}) = 0, f(-\boldsymbol{x}) = f(\boldsymbol{x}), f(\boldsymbol{x}) \geq 0$

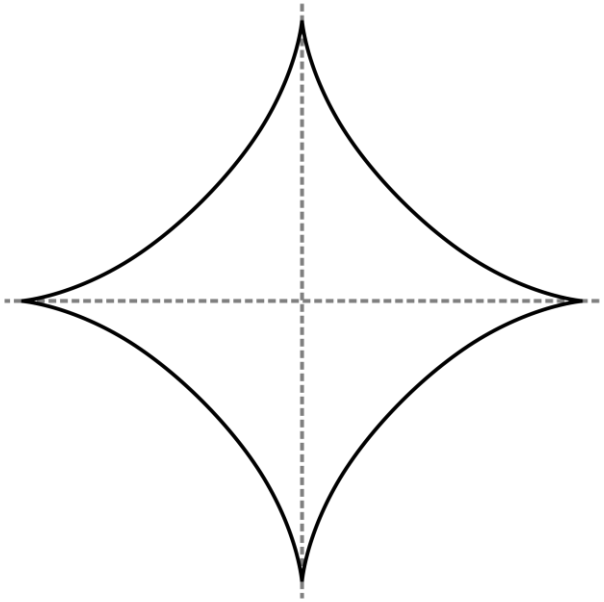
- 容易证明$f(\boldsymbol{x}-\boldsymbol{y})$满足度量的所有条件
- 向量范数可以诱导出度量

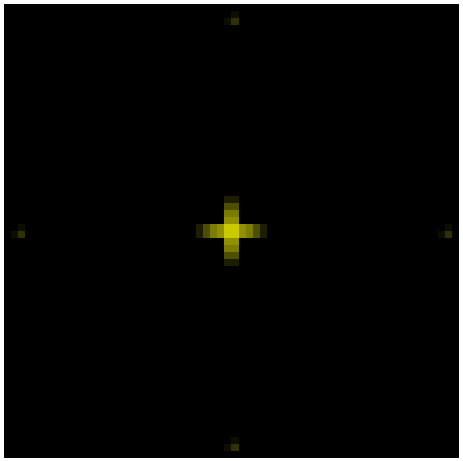
l_p范数和l_p度量

- ✓ l_p范数在$p \geq 1$时有定义：
- ✓ $$\|\boldsymbol{x}\|_p = (\sum_{i=1}^{d} |x_i|^p)^{\frac{1}{p}}$$
- ✓ 容易证明l_p范数满足齐次性和非负性
- ✓ 闵可夫斯基不等式
- ✓ $$(\sum_{i=1}^{d} (x_i + y_i)^p)^{\frac{1}{p}} \leq (\sum_{i=1}^{d} x_i^p)^{\frac{1}{p}} + (\sum_{i=1}^{d} y_i^p)^{\frac{1}{p}}$$
- ✓ 为什么要限定$p \geq 1$？
- ✓ $p < 1$时，会出现什么问题？

l_p范数和l_p度量







l_p范数和l_p度量

- ✓ $p \to \infty$

$$\|x\|_\infty = \max_{1 \le i \le d} |x_i| \qquad \|x\|_p = M \cdot \left(\sum_{i=1}^{n} \left(\frac{|x_i|}{M} \right)^p \right)^{1/p}$$

- ✓ $p = 0$

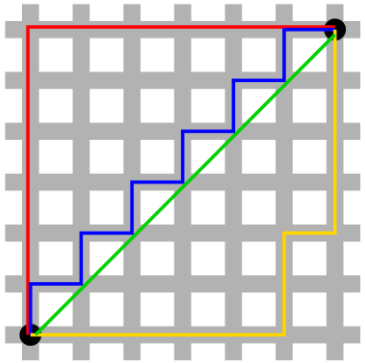
$$\|x\|_0 = \sum_{i=1}^{d} \mathbf{1}(x_i \neq 0) \qquad \lim_{p \to 0^+} \|x\|_p = \exp\left(\frac{1}{d} \sum_{i=1}^{d} \ln |x_i| \right)$$

- ✓ 满足范数的三点性质吗？

l_p范数和l_p度量

- ✓ 当$p \geq 1$时l_p范数导出l_p度量

$$d_p(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_p$$



- ✓ l_1距离又被称为曼哈顿距离

- ✓ l_∞距离衡量了任意一维的最大距离：$d_\infty(\boldsymbol{x}, \boldsymbol{y}) = \max_i\{|x_i - y_i|\}$

- ✓ 当$p > q > 0$时：

$$\|\boldsymbol{x}\|_p \leq \|\boldsymbol{x}\|_q$$

距离度量学习

- ✓ 马氏距离
  - 令G为任意$d \times d$的正定矩阵，那么
  - $f(\boldsymbol{x}) = \|G\boldsymbol{x}\|$定义了一个有效的向量范数
  - 对应的距离度量为$\|G(\boldsymbol{x} - \boldsymbol{y})\|$
  - 常用平方距离：
  - $d_A^2(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x} - \boldsymbol{y})^T A(\boldsymbol{x} - \boldsymbol{y})$，$A = G^T G$
- ✓ 为什么要引入马氏距离？
  - 假设数据服从高斯分布
  - 白化变换$\Sigma^{-\frac{1}{2}}(\boldsymbol{x} - \bar{\boldsymbol{x}})$

距离度量学习

- ✓ 度量学习：
  - 学习一个好的矩阵A来度量数据
  - $\|\boldsymbol{x}-\boldsymbol{y}\|_A^2=(\boldsymbol{x}-\boldsymbol{y})^T A(\boldsymbol{x}-\boldsymbol{y})$
- ✓ 如何确定优化目标？
  - 非常常用的思路是三元组损失（triplet loss）

$$\mathcal{L}_{\text{triplet}}=\sum_{(a,p,n)}\max\left(0,d(a,p)-d(a,n)+\alpha\right)$$

  - 深度度量学习常用Siamese Network

稀疏机器学习

Sparse machine learning

能识别吗？

我居北海君南海

黄沙白雾昼常昏

为什么？

- ✓ 我们靠什么能够从字的半边猜出正确的字来？

- ✓ 我们为什么不能靠什么从字的半边猜出正确的字来？

稀疏性

- ✓ 很多方法不需要对数据分布有显式假设
  - 最近邻，距离度量学习
  - PCA、FLD只影响最优性

- ✓ 有时候数据是具有稀疏性的
  - 人脸图像
  - 明暗光图像
  - 监控视频 …

什么是稀疏？

- ✓ 向量$\boldsymbol{x}$
  - 计算其非0元素的个数
- ✓ 那么，图像（例如人脸图像）是稀疏的吗？在什么意义上？
  - 不是在原来的空间（每一维代表一个像素）
  - 而是在某种更有效的（通常高维但稀疏的）表达方式上，如
    - 人脸图像在光照条件变化是存在于一个低维子空间中
    - 视频监控图像中背景是低频（变化慢）而前景是高频（变化快）的
    - 语音信号中的语音和噪声所处的频段，…
  - 总之，需要正确理解稀疏sparse的含义

稀疏PCA

- ✓ 如果一个向量中许多维是0，我们称它是稀疏的
- ✓ 稀疏机器学习的目标：
  - 给定一个向量$\boldsymbol{x}$，$\boldsymbol{x} \to \boldsymbol{y}$，$\boldsymbol{y}$是稀疏的
- ✓ 给定训练集$\{\boldsymbol{x}_i\}$，令新的表示为$\boldsymbol{y}_i$

- ✓ 在PCA中，$\boldsymbol{y}_i = E_d^T(\boldsymbol{x}_i - \overline{\boldsymbol{x}})$
- ✓ 如何使PCA表示稀疏？
  - 一个正则化项
  - $\min_{\boldsymbol{y}\_i} \left\| \boldsymbol{y}_i - E_d^T(\boldsymbol{x}_i - \overline{\boldsymbol{x}}) \right\|^2 + \lambda \|\boldsymbol{y}_i\|_0$

稀疏PCA

- ✓ 上述优化目标有什么问题？
  - l_0 “范数”不连续，不可微
  - E_d 已被证明导致最小的重构误差，鼓励稀疏性会带来更大的重构误差

- ✓ l_1 约束是一个好的替代
  - 凸函数
  - 连续性
  - 除0位置以外梯度存在

l_1范数诱导稀疏性

- ✓ 受限等距性（RIP）条件：
  - 测量矩阵$A, m \times n$且$m \ll n$
  - S稀疏信号：$\|\boldsymbol{x}\|_0 \leq s$
  - 若存在常数$\delta_s \in (0,1)$，使得

$$(1-\delta_s)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1+\delta_s)\|x\|_2^2$$

  - 则称A满足s阶RIP条件， δ_s为RIP常数
- ✓ 若A满足RIP条件，则下面两个目标等价

$$\arg\min_x \|x\|_0 \quad \text{s.t.} \quad Ax = y \qquad \arg\min_x \|x\|_1 \quad \text{s.t.} \quad Ax = y$$

过完备字典

- ✓ 给定$p \times k$矩阵D，$\boldsymbol{x}$与其新的表示$\boldsymbol{\alpha}$
- ✓ 考虑线性关系：

$$\boldsymbol{x} \approx D\boldsymbol{\alpha}$$

- ✓ 如果D已知，且希望重构是稀疏的：

$$\min_{\alpha_i} \|\boldsymbol{x}_i - D\boldsymbol{\alpha}_i\|^2 + \lambda\|\boldsymbol{\alpha}_i\|_1$$

- ✓ 我们称D为字典
  - $p > k$ 欠完备字典
  - $p < k$ 过完备字典

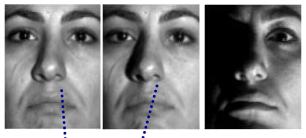
人脸识别上的应用

- ✓ 对全班150名同学采集人脸照片，每人100张，分辨率100*100
- ✓ 每张图片拉成向量，维度$p = 10000$
- ✓ 每人每张图片向量依次排列，得到字典项$k = 15000$
- ✓ 得到过完备字典D，大小$p \times k$
- ✓ 给定某个ID图像向量$\boldsymbol{x}_i$，求解

$$\min_{\alpha_i} \|\boldsymbol{x}_i - D\boldsymbol{\alpha}_i\|^2 + \lambda\|\boldsymbol{\alpha}_i\|_1$$

人脸识别上的应用

- ✓ 为什么求解$\min_{\alpha_i} \|\boldsymbol{x}_i - D\boldsymbol{\alpha}_i\|^2 + \lambda\|\boldsymbol{\alpha}_i\|_1$可以实现人脸识别？
  - 同ID一致性假设
  - 找到对应的稀疏系数所属ID
  - 非零系数在某个ID内明显最多，测试图像则属于该类



$$\boldsymbol{A}_i = [\ \vert\ \vert\ \ldots\ ] \in \mathbb{R}^{m \times n_i}$$

$$\boldsymbol{y} \approx x_{i,1} + x_{i,2} + \ldots + x_{i,n} = \boldsymbol{A}_i \boldsymbol{x}_i$$

字典学习

- ✓ 从数据集中学习一个字典D，使得所有样本$\{\boldsymbol{x}_i\}$能被字典中的原子稀疏表示，即

$$\forall i,\ \mathbf{x}_i \approx \mathbf{D}\alpha_i, \quad \|\alpha_i\|_0 \leq s.$$

- ✓ 需要联合优化字典和系数

$$\min_{\mathbf{D},\{\alpha_i\}} \sum_i \|\mathbf{x}_i - \mathbf{D}\alpha_i\|_2^2 + \lambda \sum_i \|\alpha_i\|_1$$

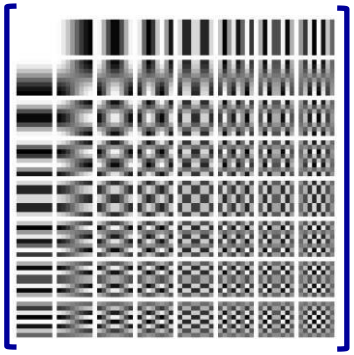
应用场景

- ✓ 图像压缩

Compression – JPEG

y

(Patches of) …
input image



A DCT basis

x coefficients

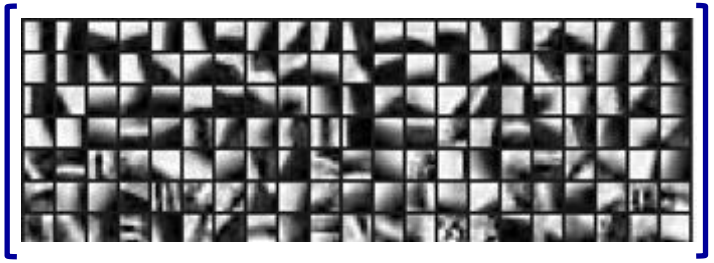
应用场景

- ✓ 图像压缩

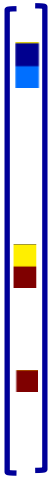
Compression – Learned dictionary

y

(Patches of) …
input image



A Learned dictionary



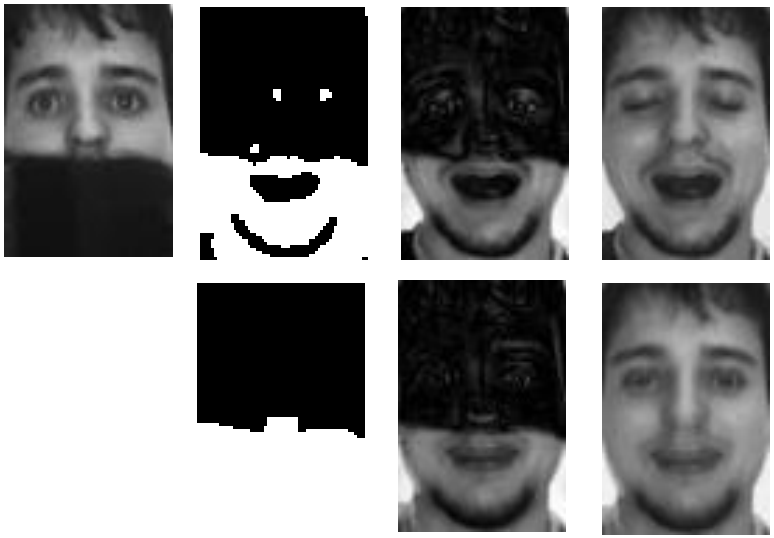
x coefficients

应用场景

- ✓ 图像去噪&修复
  - 学习干净图像的字典，用稀疏表示抑制噪声，利用字典补全局部信息
- ✓ 特征提取
  - 代替SIFT等手工特征
- ✓ MRI加速成像
  - 结合压缩感知，用少量测量和稀疏字典快速重建图像。

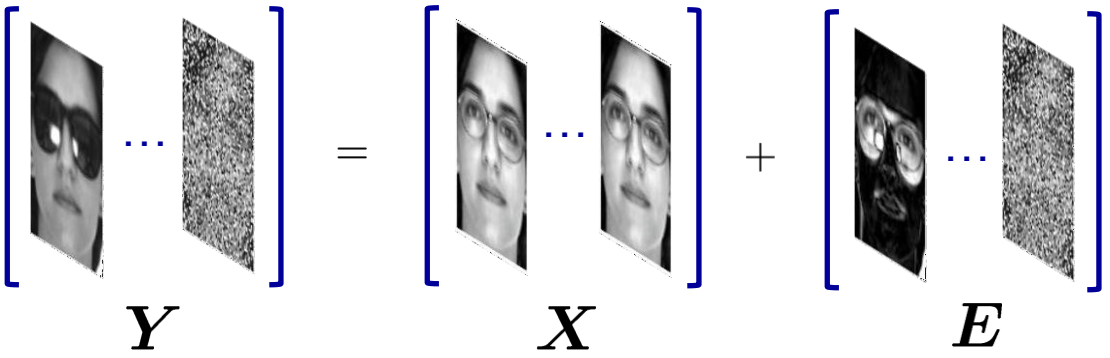
稀疏向量的推广

- ✓ 如果在二维结构上具有稀疏性，会有怎样的特点？
  - 低秩矩阵（Low-Rank Matrix）



稀疏向量的推广

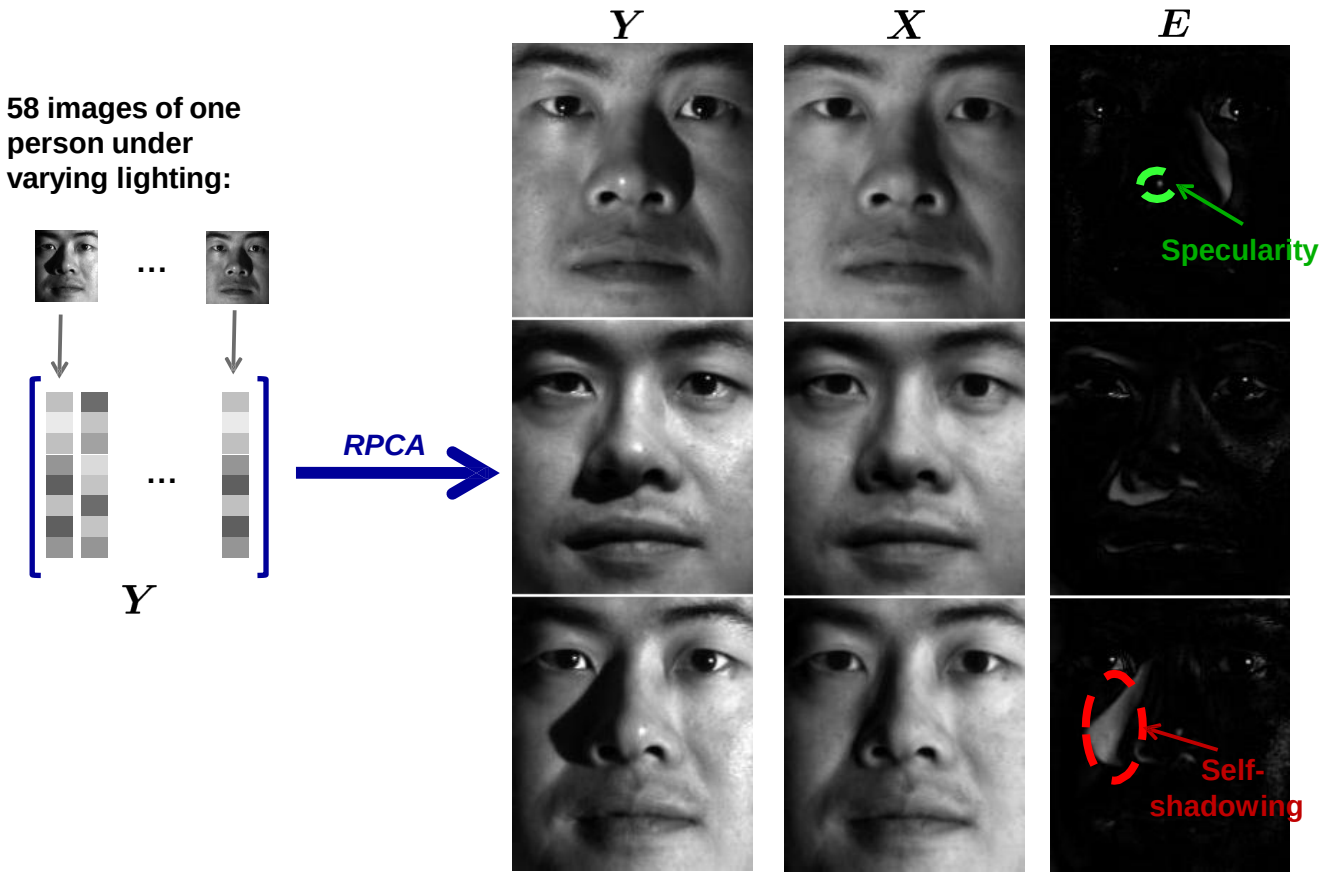
- ✓ 如果在二维结构上具有稀疏性，会有怎样的特点？
  - 低秩矩阵（Low-Rank Matrix）



Given $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{E}$, with $\boldsymbol{X}$ low-rank, $\boldsymbol{E}$ sparse, recover $\boldsymbol{X}$.

稀疏向量的推广

- ✓ 如果在二维结构上具有稀疏性，会有怎样的特点？
  - 低秩矩阵（Low-Rank Matrix）



Tutorial: Yi Ma et al.

- ✓ http://www.eecs.berkeley.edu/~yang/courses/ECCV2012/ECCV12-lecture1.pdf