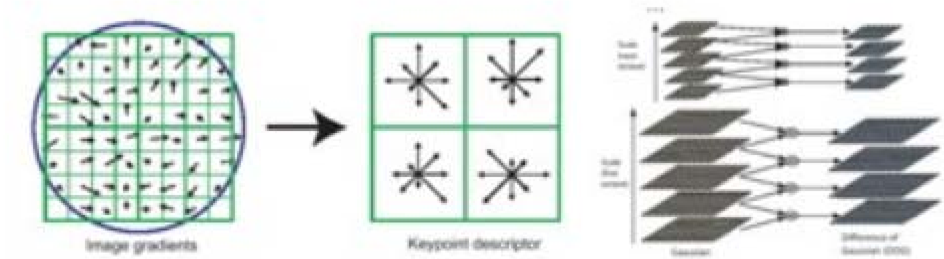
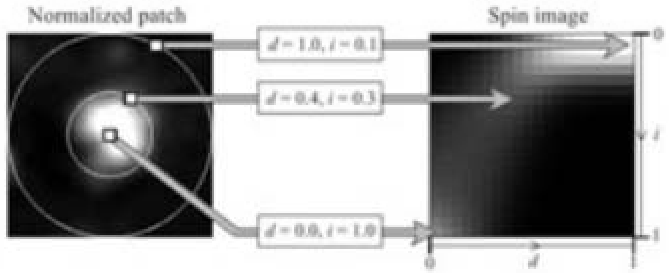
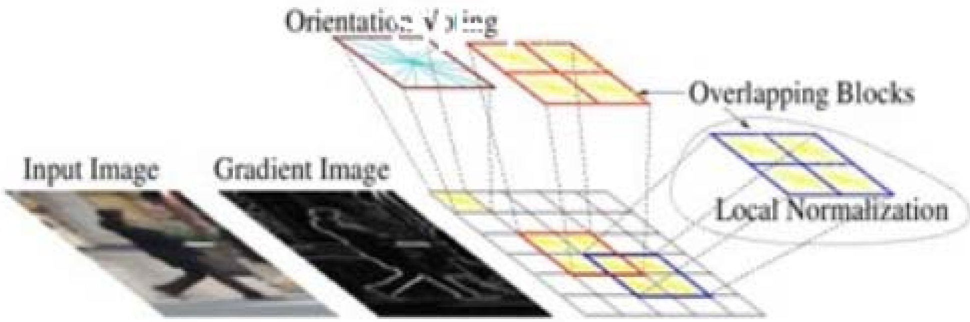
机器学习

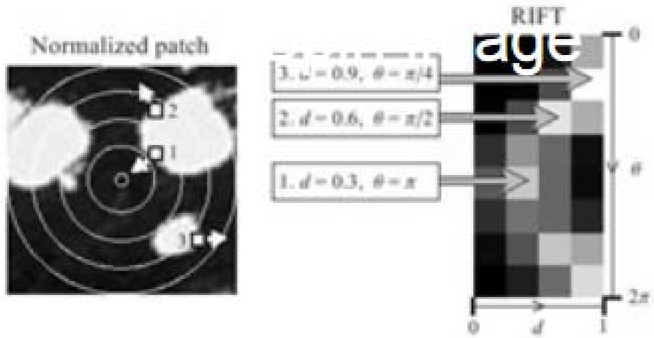
6. PCA & SVD

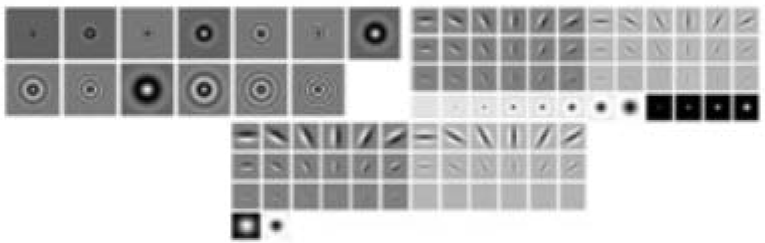
Computer vision features

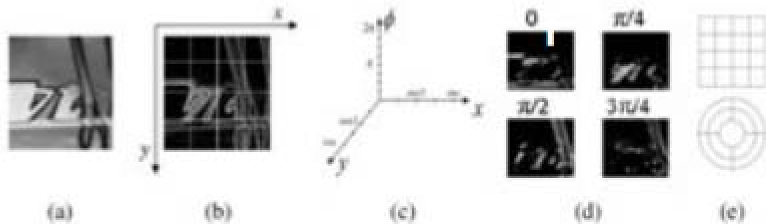






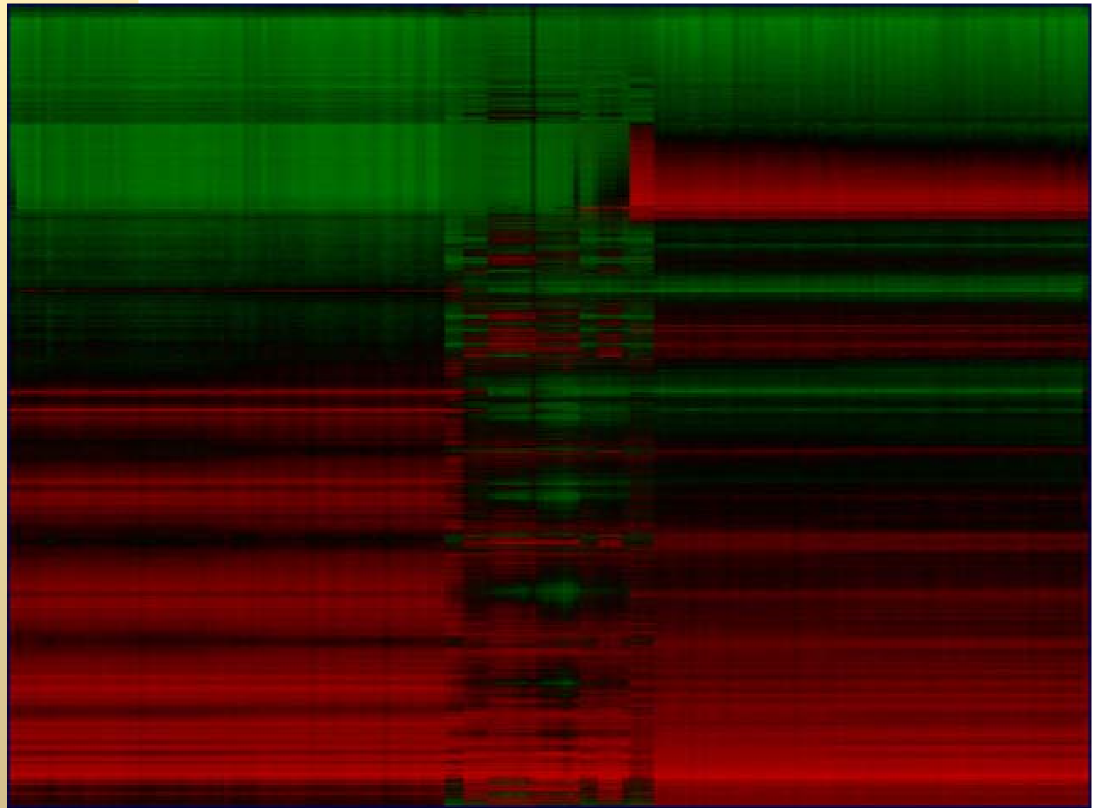


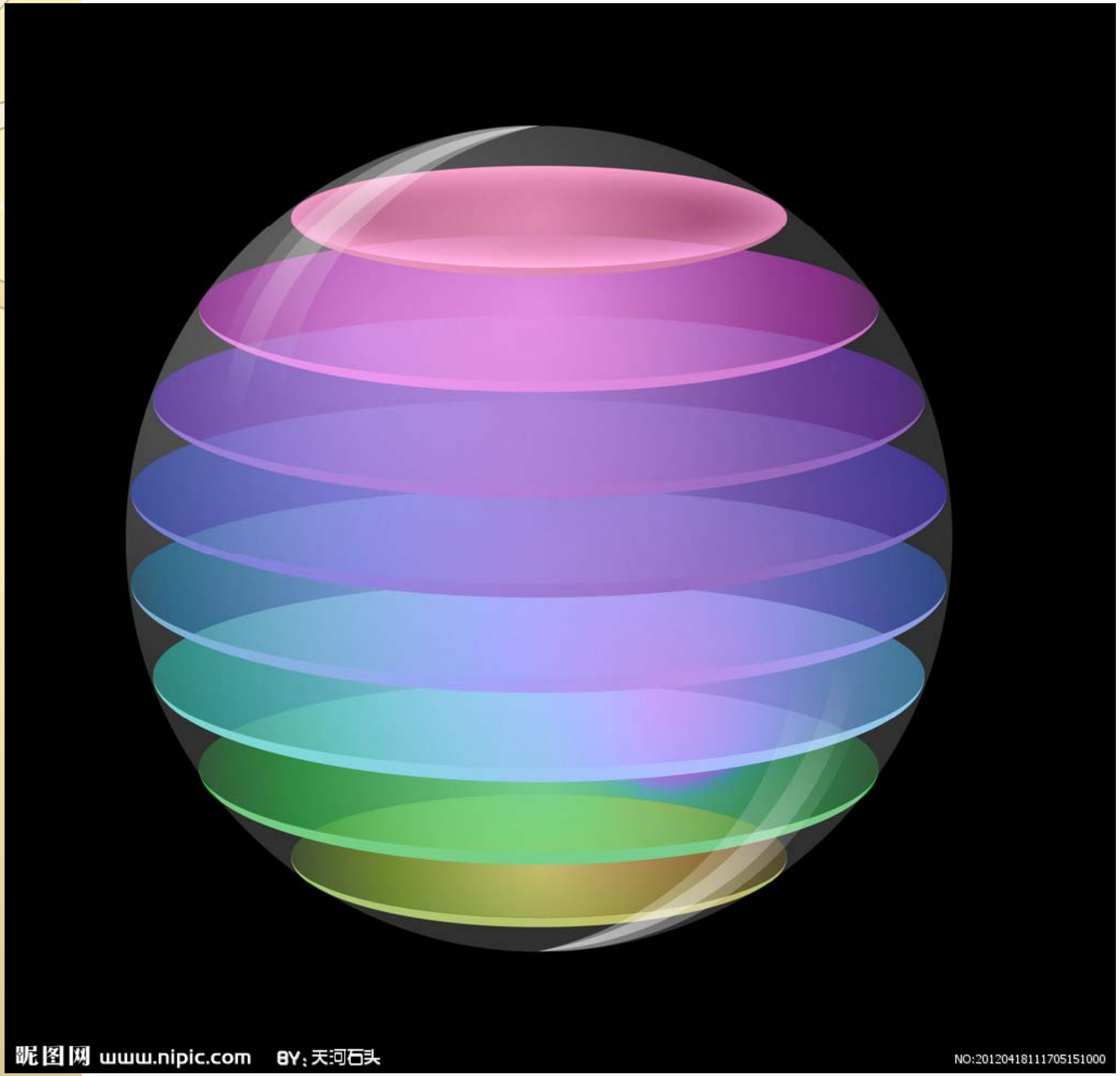




Slide Credit: Andrew Ng

"Feeding an Animal"




昵图网 www.nipic.com BY: 天河石头
NO:20120418111705151000


Big Data
STATISTICA

为什么要降维?

- 增加计算效率
- 提高计算精度

主要内容

- 子空间学习问题
- PCA基本思想
- SVD
- 典型应用
- 核PCA

主要内容

- 子空间学习问题
- PCA基本思想
- SVD
- 典型应用
- 核PCA

子空间学习

- 高维数据往往位于一个低维超平面上
- 发现这一低维超平面的基底将帮助我们有效挖掘数据内涵的特征与模式
- 应用：
  - 数据压缩
  - 数据可视化
  - 噪音消除

Subspace Learning

子空间学习

- 人脸识别
  - 匹配好的人脸具有典型的低维子空间特征

子空间学习

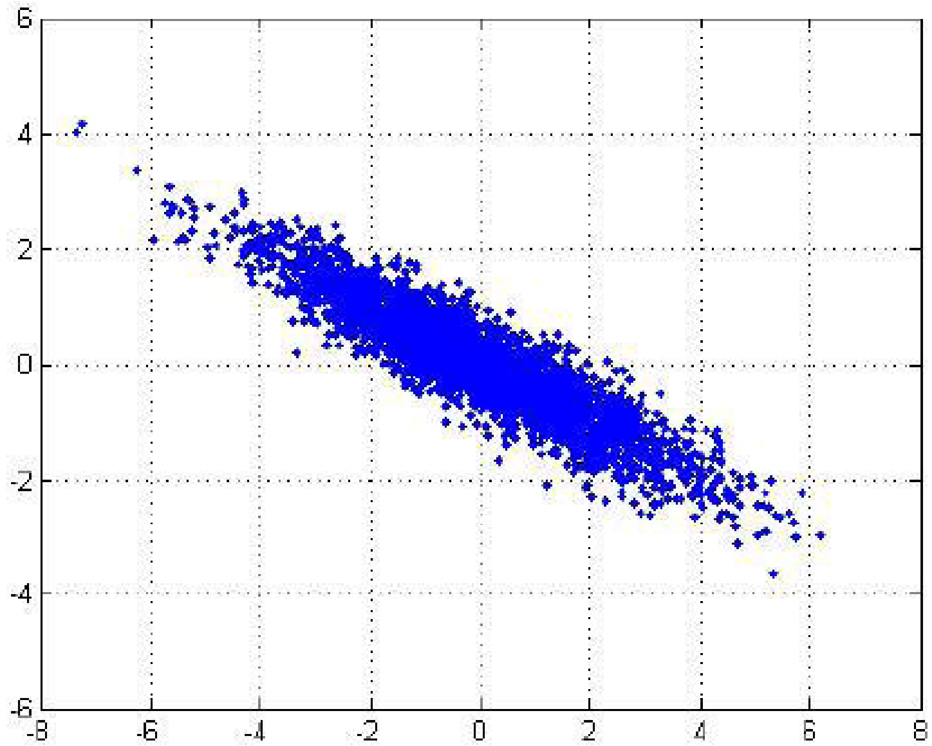
- 人脸识别
  - 匹配好的人脸具有典型的低维子空间特征
  - 可视为近似朗伯体，单个人近似本质维度为4
  - 多个人近似维度为9!
  - 奇妙之处在于?
- 方法1：找到每个人的人脸对应的子空间，然后计算测试人脸在哪个子空间投影距离最近
- 方法2：对所有人脸计算其子空间，然后再其子空间坐标投影，利用其低维坐标进行分类计算

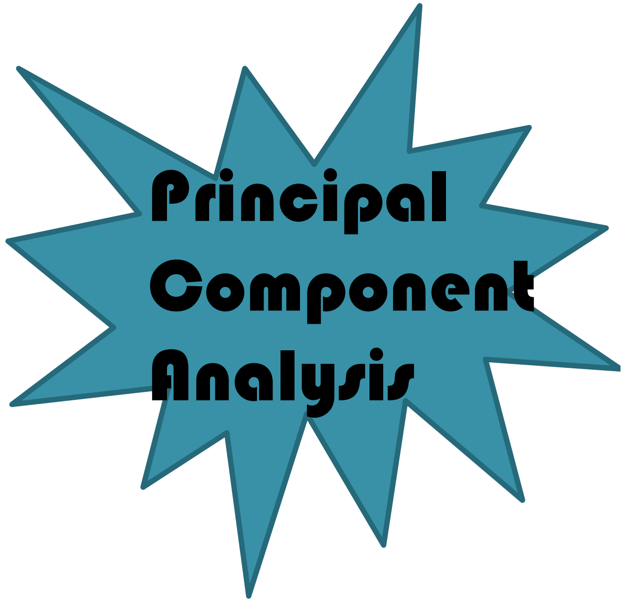
主要内容

- 子空间学习问题
- PCA基本思想
- SVD
- 典型应用
- 核PCA

PCA

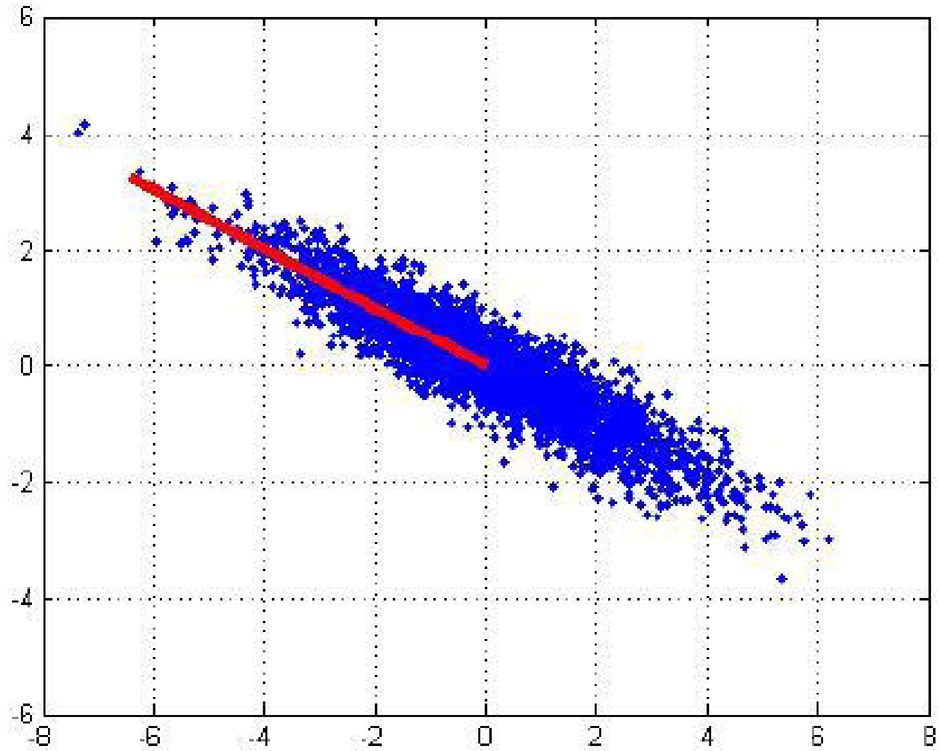
- 主成分分析
  - 目标：找到数据本质低维子空间的正交基





PCA

- 主成分分析(Principal component analysis)
  - 目标：找到数据本质低维子空间的正交基



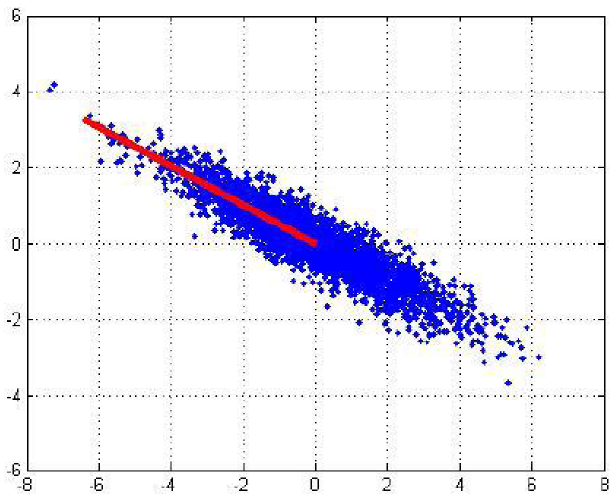
PCA

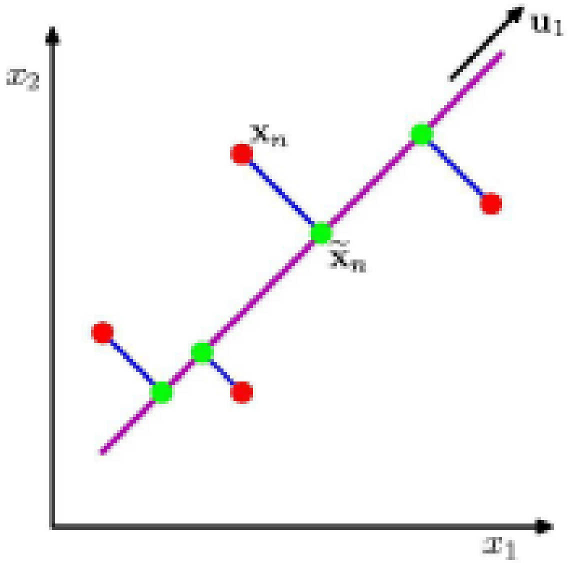
- 主成分分析(Principal component analysis)
  - 目标：找到数据本质低维子空间的正交基
  - 最关键的是?
  - 机器学习的基本元素?
  - Performance Measure!

- 三个切入点
  - 最大化投影数据方差
  - 最小化原数据与投影数据间距离
  - ?

PCA

- 最大化投影数据方差
- 最小化原数据与投影数据间距离





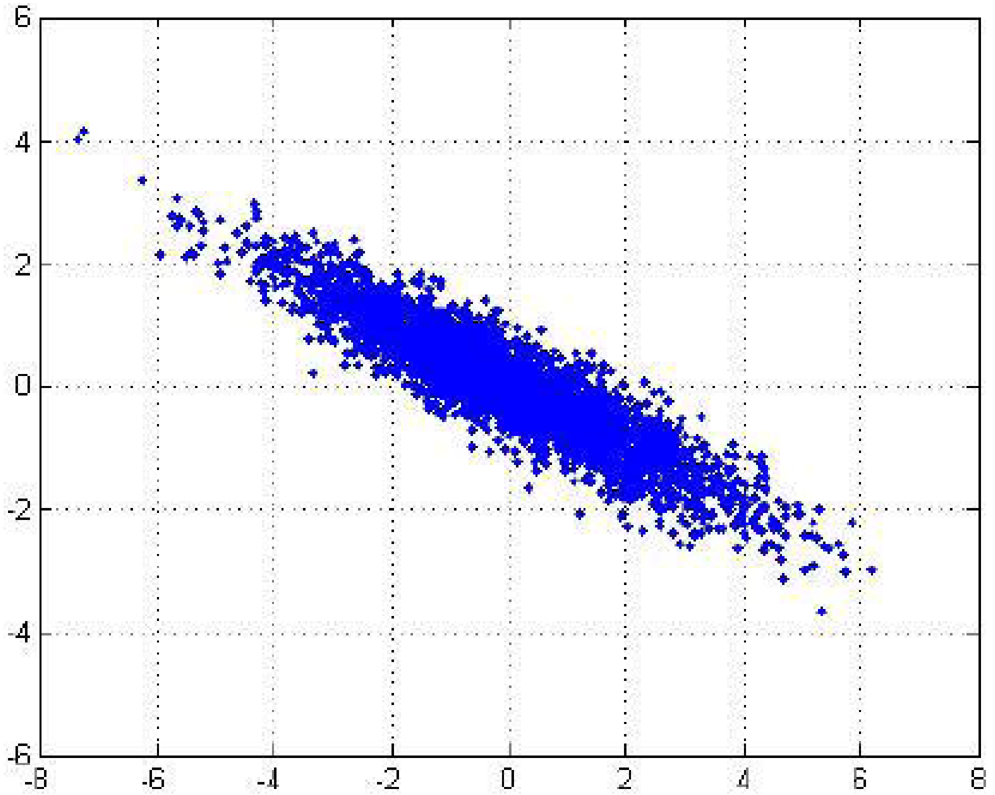
PCA

- 最大投影方差:
  - 思想：给定高维空间数据，找到低维子空间使得数据在此空间投影后，信息（方差）能尽可能大的保持
  - 表现度量：最大投影方差
- 最小投影距离:
  - 思想：给定高维空间数据，找到低维子空间使得数据在此空间投影后，投影数据与原数据之间ls距离尽可能小（尽可能重建原数据）
  - 表现度量：最小投影距离

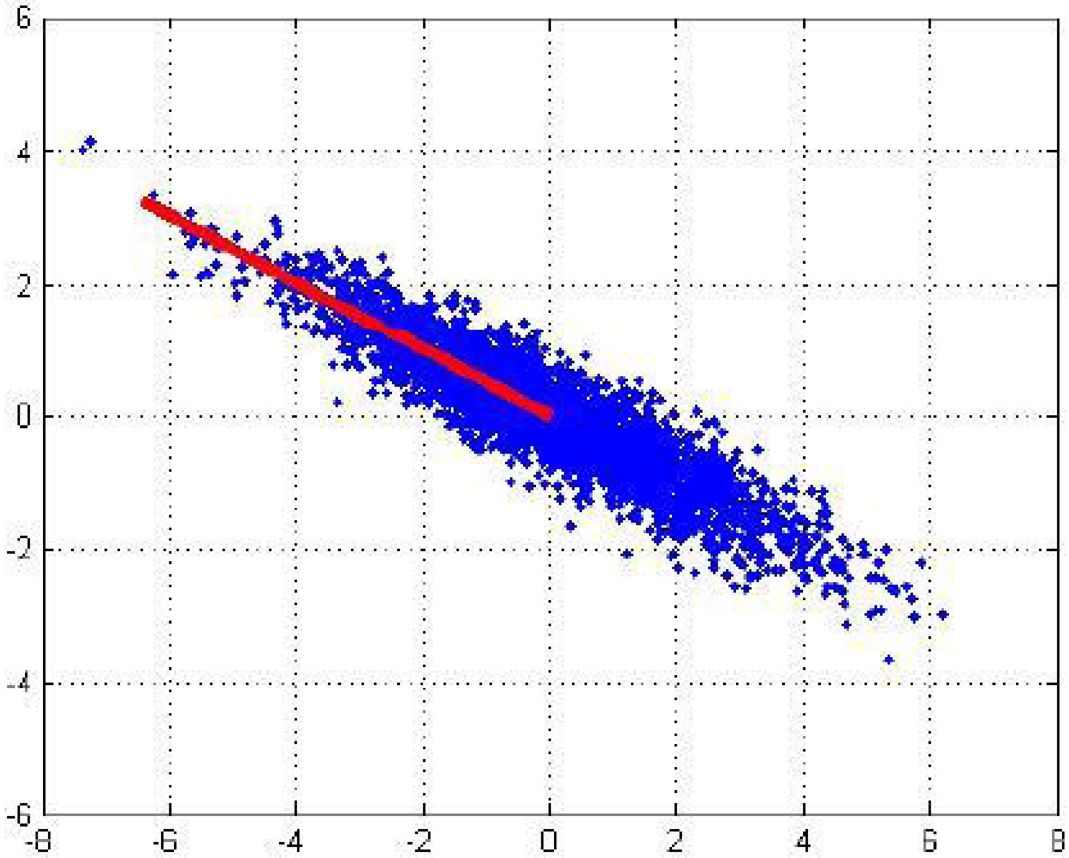
PCA

- 最大投影方差
- 数据居中处理!
- 第一主方向：最大方差投影方向
- 第一主成分：数据在第一主方向上的投影
- 之后的主方向
  - 依次与之前主方向正交
  - 数据投影方差尽可能最大

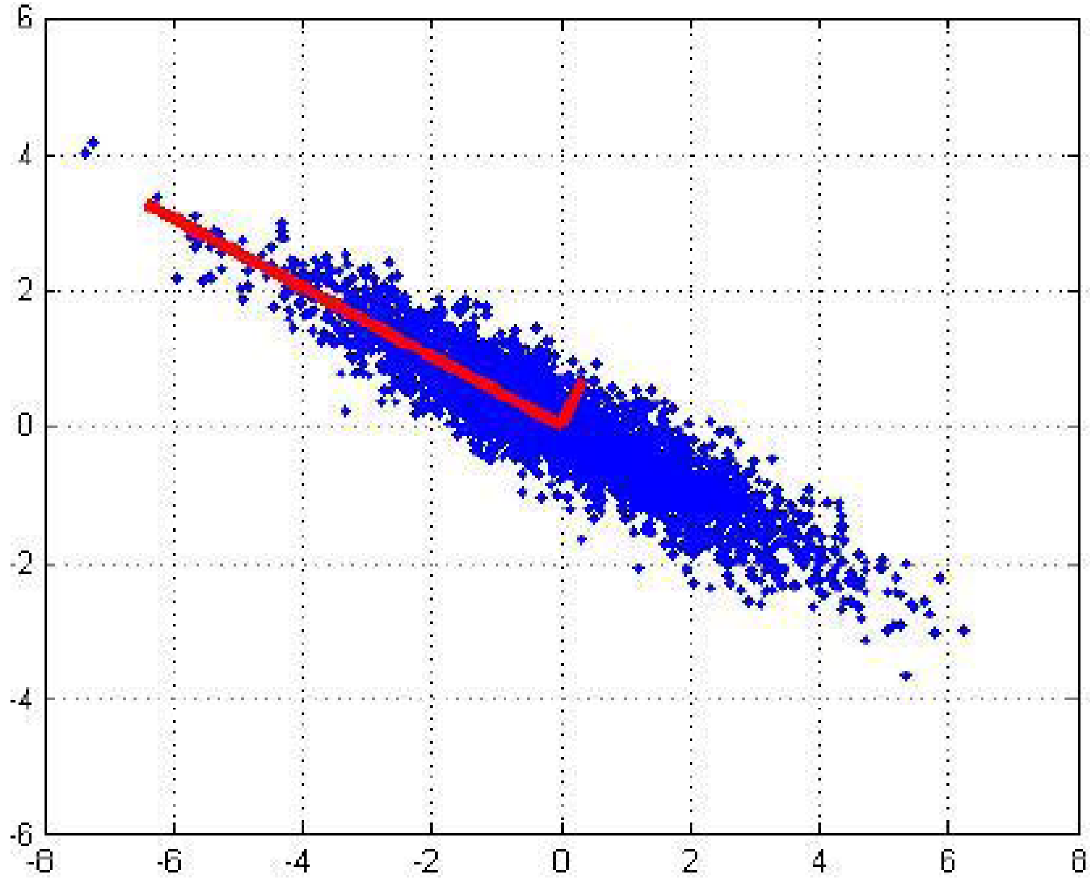
PCA



PCA



PCA



PCA

- 给定居中化数据{x1,...,xm}，计算第一主成分的目标函数为：

$$\mathbf{w}_1 = \arg\max_{\|\mathbf{w}\|=1} \frac{1}{m}\sum_{i=1}^{m}\{(\mathbf{w}^T\mathbf{x}_i)^2\}$$

- 第二主成分目标函数为：

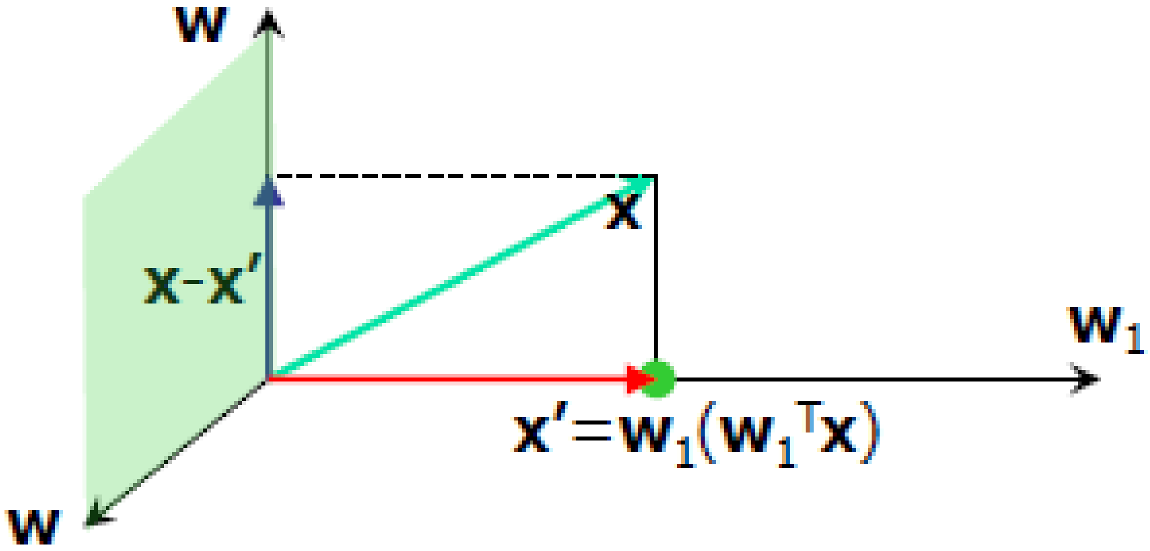
$$\mathbf{w}_2 = \arg\max_{\|\mathbf{w}\|=1} \frac{1}{m}\sum_{i=1}^{m}\{[\mathbf{w}^T(\mathbf{x}_i - \mathbf{w}_1\mathbf{w}_1^T\mathbf{x}_i)]^2\}$$

- 之后为?

PCA

$$\mathbf{w}_2 = \arg\max_{\|\mathbf{w}\|=1} \frac{1}{m}\sum_{i=1}^{m}\{[\mathbf{w}^T(\mathbf{x}_i - \mathbf{w}_1\mathbf{w}_1^T\mathbf{x}_i)]^2\}$$

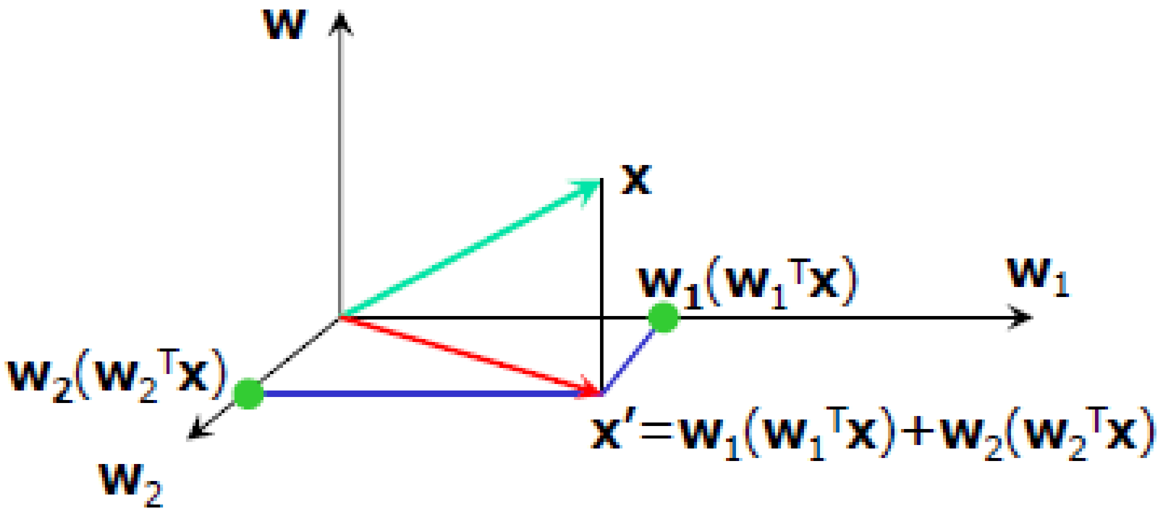
- 原理：



PCA

➢ 计算第k个主成分:

$$\mathbf{w}_k = \arg\max_{\|\mathbf{w}\|=1} \frac{1}{m} \sum_{i=1}^{m} \{[\mathbf{w}^T(\mathbf{x}_i - \sum_{j=1}^{k-1} \mathbf{w}_j \mathbf{w}_j^T \mathbf{x}_i)]^2\}$$



PCA

- 有无更简便的计算方法?
- 所有主成分能否一次性算出?

$$\mathbf{w}_1 = \arg\max_{\|\mathbf{w}\|=1} \frac{1}{m}\sum_{i=1}^{m}\{(\mathbf{w}^T\mathbf{x}_i)^2\}$$

- 能!
- PCA主方向 = 数据协方差矩阵特征向量
- 更大特征值 对应 更重要特征向量

PCA

- 主成分分析算法
  - 目标：计算数据k个主方向
  - 1. 数据居中
  - 2. 计算居中数据的协方差矩阵
  - 3. 计算协方差矩阵对应最大k个特征值的特征向量
  - 输出

PCA

- 切入点2：最小投影距离
- 尝试推导目标函数

Low-rank
Matrix
Factorization

主要内容

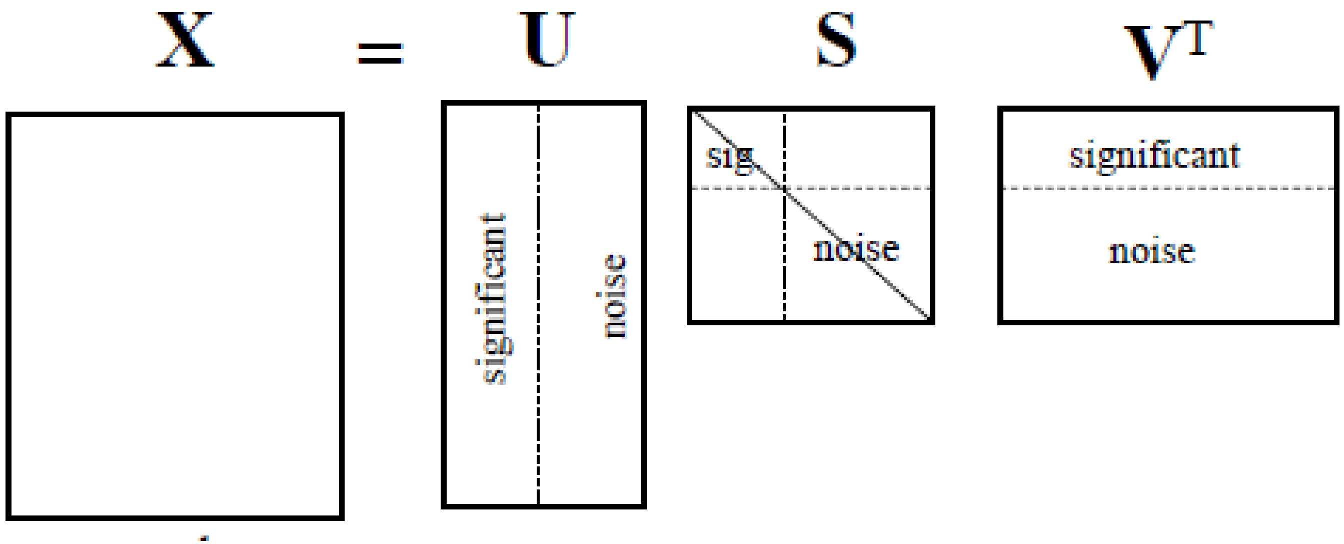
- 子空间学习问题
- PCA基本思想
- SVD
- 典型应用
- 核PCA

SVD

- 切入点三：奇异值分解

$$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_m]$$

Singular Value Decomposition



SVD

- 切入点一：最大投影方差
- 切入点三：最小投影距离(低秩矩阵分解)
- 切入点三：奇异值分解

同一问题，不同角度，完全等价！

SVD

- 切入点一：最大投影方差
- 切入点三：最小投影距离(低秩矩阵分解)
- 切入点三：奇异值分解

同一问题，不同角度，完全等价！

主要内容

- 子空间学习问题
- PCA基本思想
- SVD
- 典型应用
- 核PCA

典型应用

- 人脸识别
- 图像压缩
- 噪音去除

人脸识别

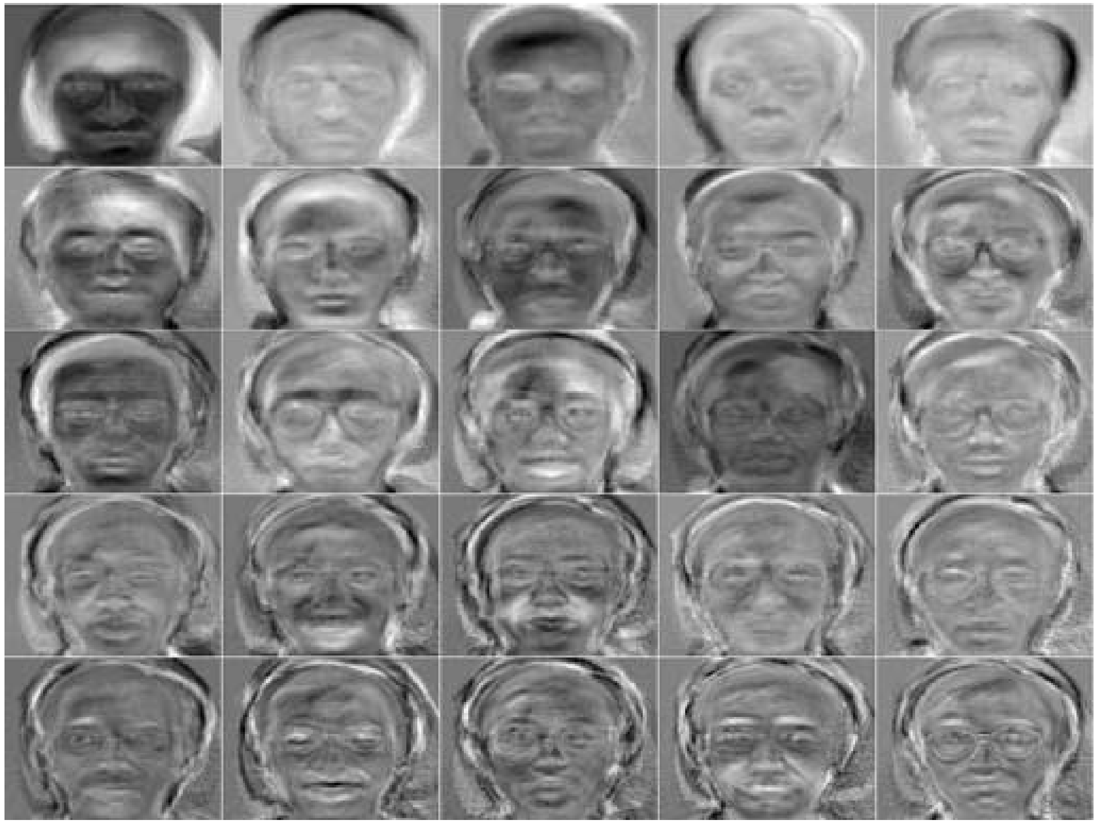
- 方法1：找到每个人的人脸对应的子空间，然后计算测试人脸在哪个子空间投影距离最近
- *方法2：对所有人脸计算其子空间，然后再其子空间坐标投影，利用其低维坐标进行分类计算*

- 特征脸：子空间基底向量

Eigenface

人脸识别

- 特征脸：子空间基底向量



人脸识别

- 数据： X
- 大小：图片尺寸*图片个数
- 协方差矩阵大小?
- 怎样算降维?
- 能否简化计算?

人脸识别

- 协方差矩阵XX^T
- 内积矩阵X^TX
- 用后者怎样获得前者信息?

$$L = X^T X \qquad \Sigma = X X^T$$

- 降维表示如何用主方向计算?

人脸识别

- 在子空间上的重建效果：

人脸识别

- 需要注意的:
  - 人脸须对准
  - 同样的人脸角度
  - 同样的尺寸

人脸识别

- 微笑特征脸

人脸识别

➢ 沮丧特征脸

图像压缩

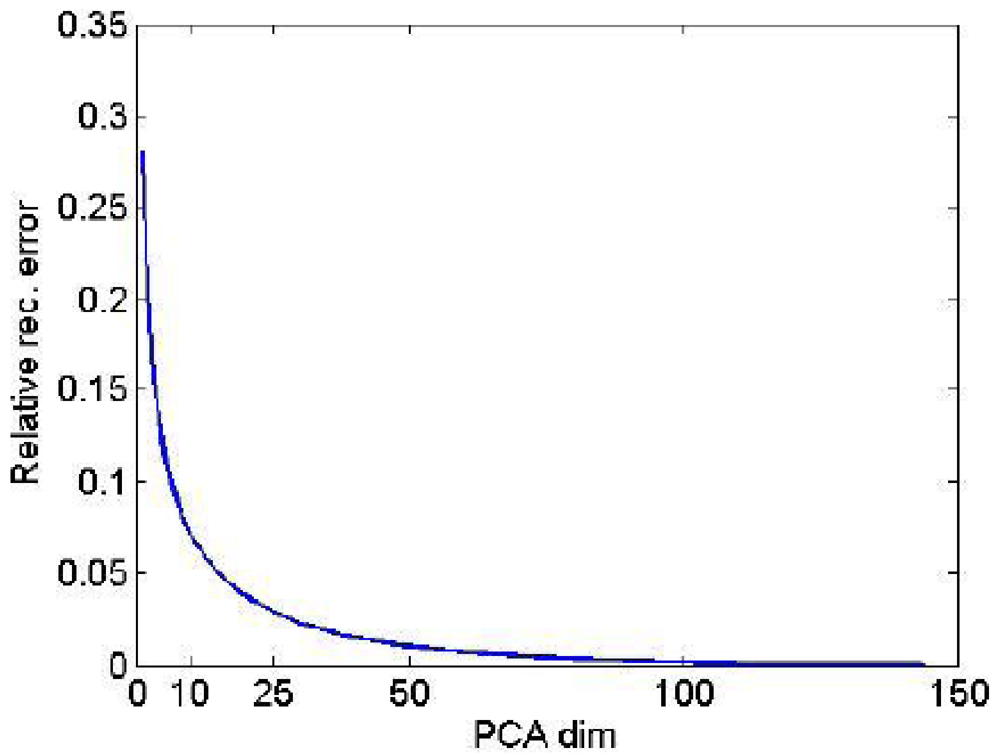
- 怎样压缩?

图像压缩

- 方法1：
  - 将图像分成小块
  - 如400*400图像，分成20*20小块， 则构成400*400矩阵
- 方法2：
  - 直接按行或按列！
- 构造原理?
- 为什么能压缩?

图像压缩

➢ 奇异值分布



0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Relative rec. error
0
10
25
50
100
150
PCA dim

图像压缩

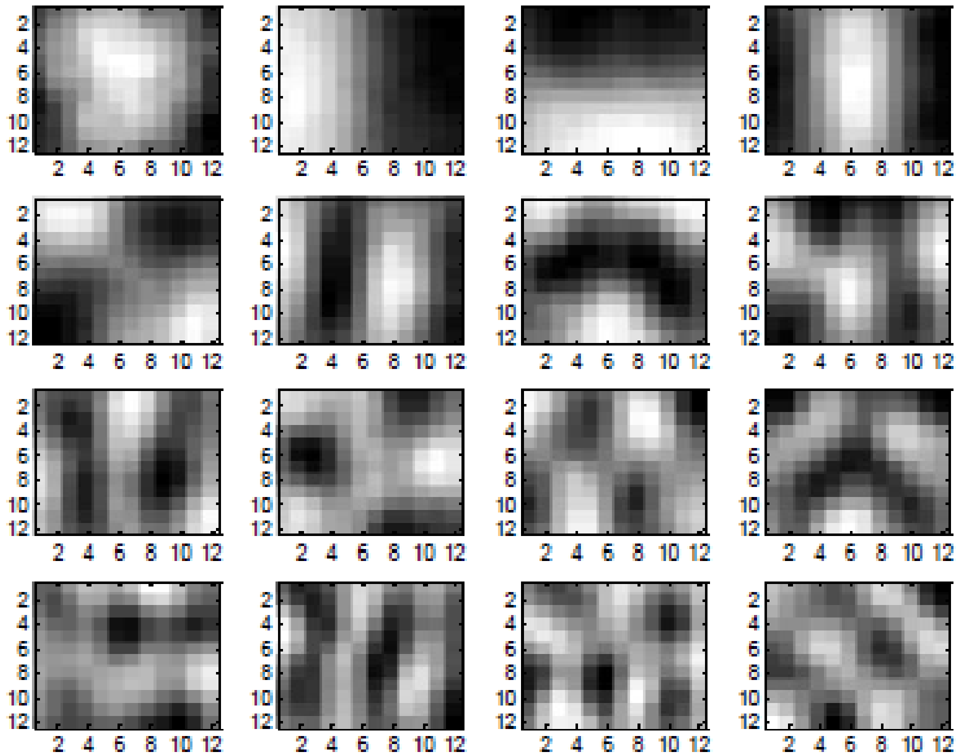
- 压缩效果：60维

图像压缩

- 压缩效果：16维

图像压缩

- 16个主方向

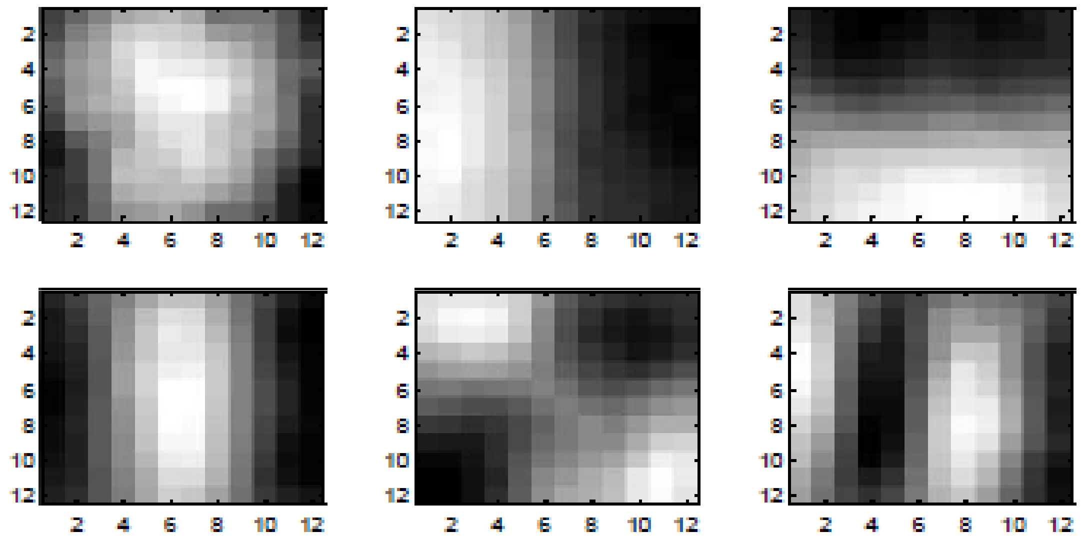


图像压缩

- 压缩效果：6维

图像压缩

- 6个主方向

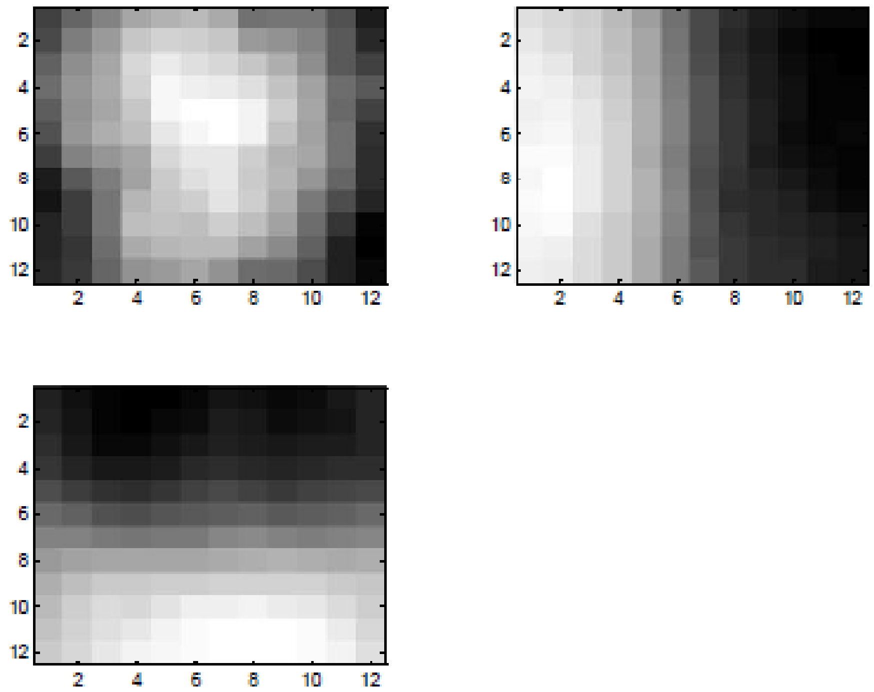


图像压缩

- 压缩效果：3维

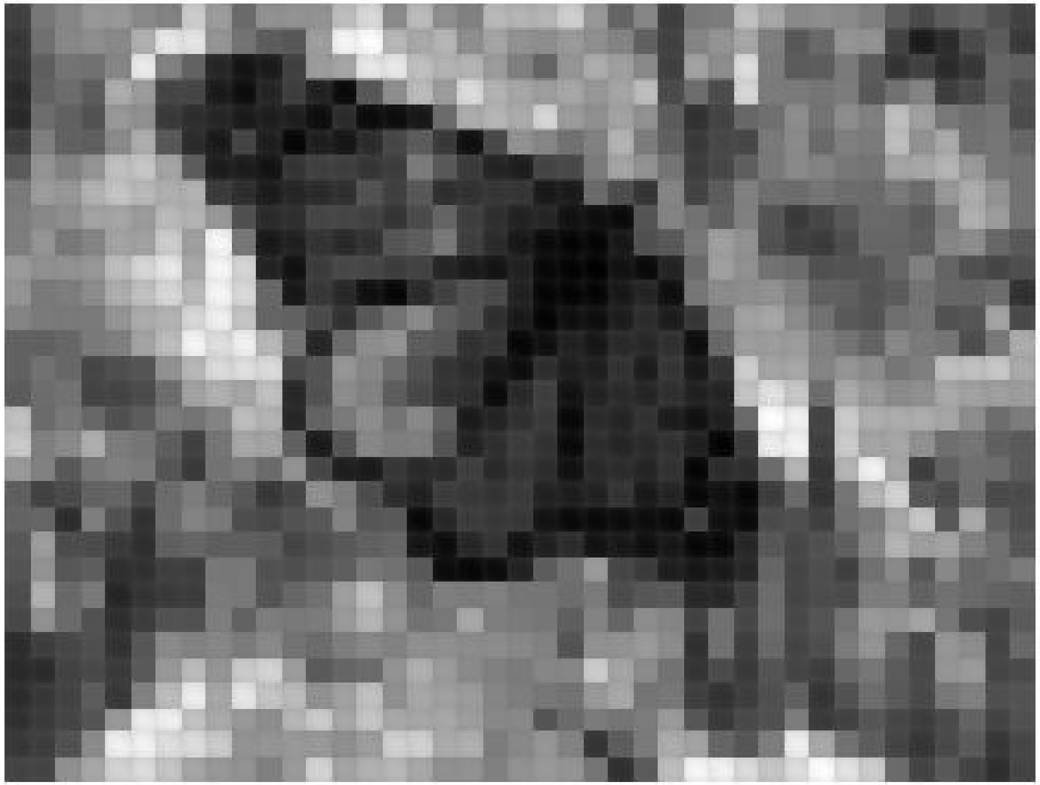
图像压缩

- 3个主方向



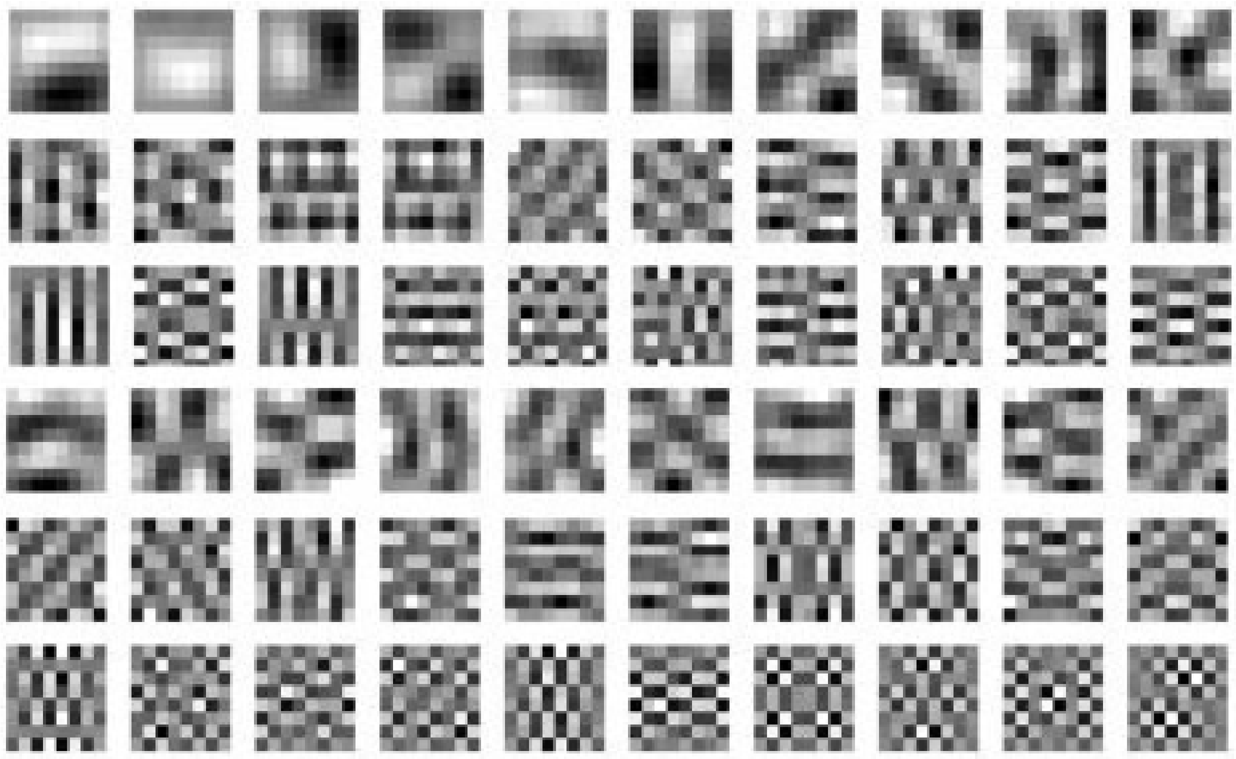
图像压缩

- 压缩效果：1维

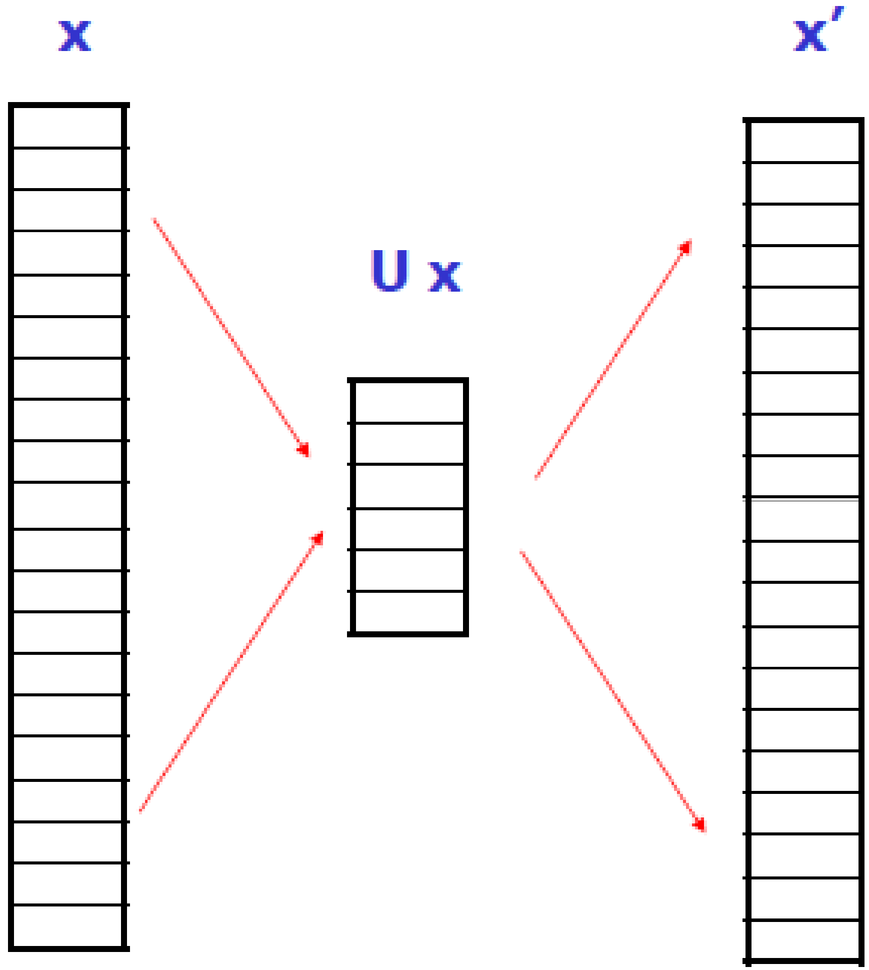


图像压缩

- 60个主方向



噪音去除



噪音去除

- 含噪图像

噪音去除

- 15个主成分后的去噪图像

要求

1. 子空间学习问题的重要性
2. PCA与SVD的基本原理
3. 维数爆炸

阅读：

[1] The Elements of Statistical Learning: Data Mining, Inference and Prediction. Hastie, Tibshirani, Friedman. Springer, 2008. 14.5 Principal Components

[2] Pattern Recognition and Machine Learning, Christopher , M. Bishop, Springer, 2006. 12.1 Principal Component Analysis； 12.3 Kernel PCA

编程作业2：

1. 实现PCA方法
2. 在一组自己设计的人工数据上尝试PCA计算效果，并从不同角度分析算法特点
3. 在人脸数据集上尝试PCA降维、重建效果（http://vision.ucsd.edu/~leekc/ExtYaleDatabase/ExtYaleB.html）