模式识别

主成分分析

Principal Component Analysis

吴建鑫

南京大学人工智能学院，2023

与领域无关的特征提取

- domain independent feature extraction

- 第4讲. PCA
- 第5讲. 特征的归一化(normalization)
- 第5讲. FLD

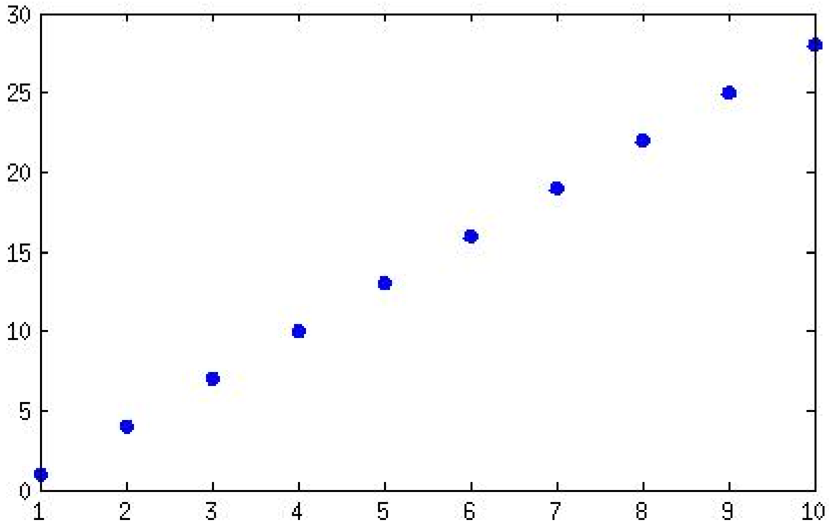
- 首先介绍PCA

目标

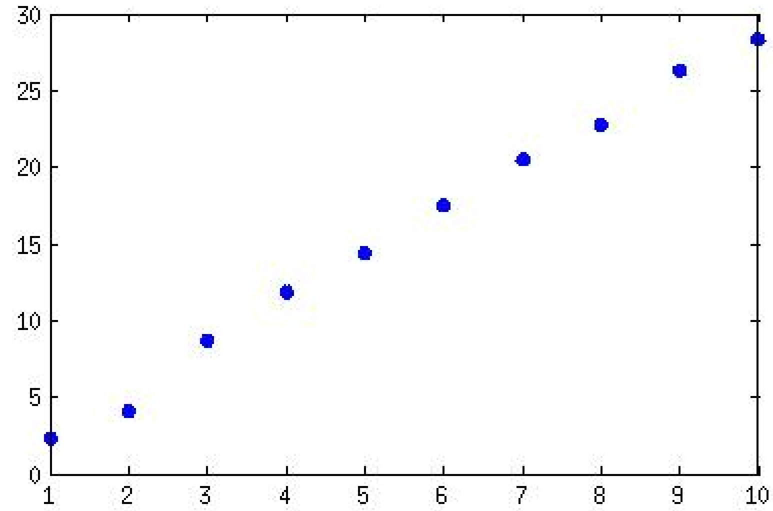
- ✓ 理解PCA的含义、目的、适用范围
- ✓ 熟记PCA的各个步骤，能实际应用PCA
- ✓ 了解PCA的各种相关解释，理解其含义
- ✓ 提高目标
  - 理解相关推导，能自主独立完成推导
  - 进一步能通过独立阅读理解更多PCA的含义、限制、解释等，并能应用到学习、研究中遇到的问题中去

PCA基础

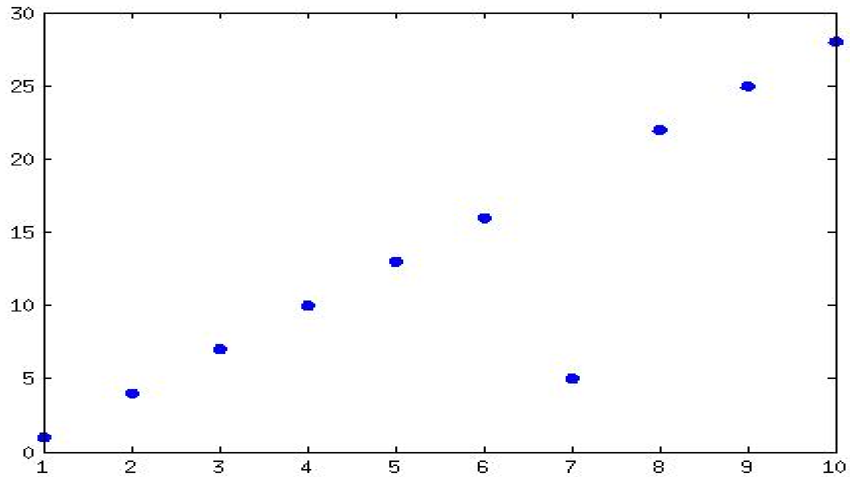
你的数据是多少维的？



0
5
10
15
20
25
30
1
2
3
4
5
6
7
8
9
10



0
5
10
15
20
25
30
1
2
3
4
5
6
7
8
9
10



0
5
10
15
20
25
30
1
2
3
4
5
6
7
8
9
10

常见的数据特点

✓ 数据各维度之间不是互相独立的

- 数据的内在维度(intrinsic dimensionality)通常远低于其表面维度
- 因此，需要降低数据维度(dimensionality reduction)
- PCA在降维方法中（可能）是最常用的一种

这是谁？

$$96 \times 108 = 10368?$$

Starting point: 零阶表示

- ✓ Zero-dimensional representation
- ✓ 不允许使用任何维度，如何最佳表示$\boldsymbol{x}$?
- ✓ 寻找某个固定(constant)的$\boldsymbol{m}$，使得

$$J_1(\boldsymbol{m}) = \min_{\boldsymbol{m}} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{m}\|^2$$

- ✓ 最优解：(证明？)

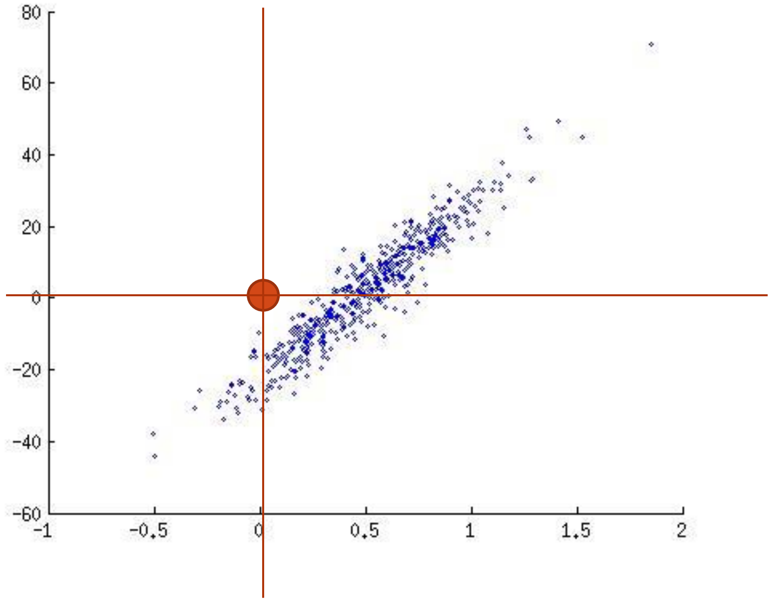
$$\boldsymbol{m}^* = \operatorname*{argmin}_{\boldsymbol{m}} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{m}\|^2 = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$$

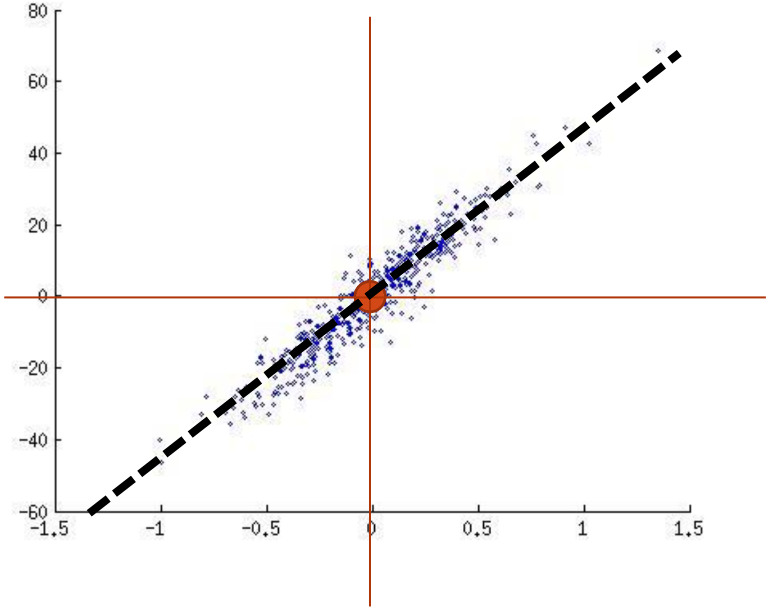
一维表示：数据维度间的线性关系

- ✓ 如同前面的例子
  - 数据是d维
  - 但是内在维度可能是m维的，$m < d$或者$m \ll d$
  - PCA用线性关系来降低维度

- ✓ $\boldsymbol{x} \in \mathbb{R}^d$: 原来的高维数据（随机变量）
  - 训练样本：$\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n$
- ✓ 假设$m = 1$, 用原数据的单个线性组合来表示
  - $y_i = \boldsymbol{w}^T \boldsymbol{x}_i + b$
  - $y_1, y_2, \dots, y_n$ -- 新的数据/特征(features)
  - 如何寻找最佳的$\boldsymbol{w}$？如何找到最佳的b？

Idea：选择什么方向？为什么？





✓什么方向最优？

形式化formalization：最大化方差

- ✓ 方差是衡量新特征包含信息多少的度量
  - 有时也称为能量energy
- ✓ 优化目标函数 $J_2(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{w}^T(\boldsymbol{x}_i - \overline{\boldsymbol{x}})\right\|^2$
- ✓ 发现问题了吗?
  - $J_2(\boldsymbol{w})$可以是无穷大或者无穷小！
  - 最常用的解决办法：加上限制条件 $\|\boldsymbol{w}\|^2 = \boldsymbol{w}^T\boldsymbol{w} = 1$

$$\underset{\boldsymbol{w}}{\operatorname{arg\,max}} \frac{1}{n}\sum_{i=1}^{n}\left\|(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T\boldsymbol{w}\right\|^2$$

$$\text{s.t.} \qquad \boldsymbol{w}^T\boldsymbol{w} = 1$$

  - s.t. –subject to,表示约束条件constraint(s)

简化simplification 变换transformation

$$\left\|(\boldsymbol{x}_i-\overline{\boldsymbol{x}})^T\boldsymbol{w}\right\|^2=\left((\boldsymbol{x}_i-\overline{\boldsymbol{x}})^T\boldsymbol{w}\right)^T\left((\boldsymbol{x}_i-\overline{\boldsymbol{x}})^T\boldsymbol{w}\right)$$
$$=\boldsymbol{w}^T(\boldsymbol{x}_i-\overline{\boldsymbol{x}})(\boldsymbol{x}_i-\overline{\boldsymbol{x}})^T\boldsymbol{w}$$

$$\frac{1}{n}\sum_{i=1}^{n}\left\|(\boldsymbol{x}_i-\overline{\boldsymbol{x}})^T\boldsymbol{w}\right\|^2=\boldsymbol{w}^T\sum_{i=1}^{n}\frac{1}{n}(\boldsymbol{x}_i-\overline{\boldsymbol{x}})(\boldsymbol{x}_i-\overline{\boldsymbol{x}})^T\boldsymbol{w}$$
$$=\boldsymbol{w}^T Cov(\boldsymbol{x})\boldsymbol{w}$$

优化optimization

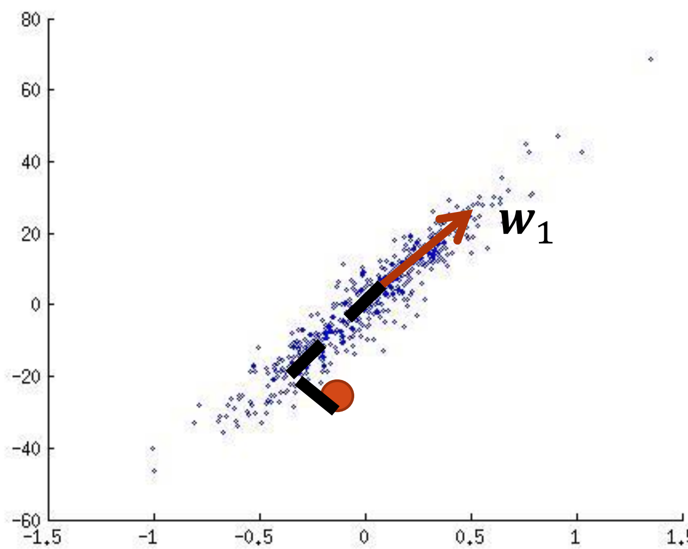
- ✓ 拉格朗日乘子法 Lagrange multipliers
  - 将有约束的优化问题转化为无约束的优化问题
- ✓ Lagrangian 拉格朗日函数

$$f(\boldsymbol{w},\lambda)=\boldsymbol{w}^T Cov(\boldsymbol{x})\boldsymbol{w}-\lambda(\boldsymbol{w}^T\boldsymbol{w}-1)$$

- ✓ λ：拉格朗日乘子Lagrange multiplier
- ✓ 最优的必要条件：$\frac{\partial f}{\partial \boldsymbol{w}}=\mathbf{0}$，$\frac{\partial f}{\partial \lambda}=0$
- ✓ $\frac{\partial f}{\partial \boldsymbol{w}}=2Cov(\boldsymbol{x})\boldsymbol{w}-2\lambda\boldsymbol{w}=\mathbf{0}$
  - 我们这里的前提条件是什么？
  - 应该想到用哪一个公式？
- ✓ $Cov(\boldsymbol{x})\boldsymbol{w}=\lambda\boldsymbol{w}$, $\quad\quad \boldsymbol{w}^T\boldsymbol{w}=1$!

选哪个特征向量？

- ✓ $J_2(\boldsymbol{w}) = \boldsymbol{w}^T Cov(\boldsymbol{x})\boldsymbol{w} = ?$
- ✓ $Cov(\boldsymbol{x})$是半正定的（如何证明？）
- ✓ 选取λ_1(即最大特征值)对应的特征向量$\boldsymbol{\xi}_1$为$\boldsymbol{w}_1$
  - 为什么？约束条件满足了吗？
- ✓ 怎样用$\boldsymbol{w}_1$来近似$\boldsymbol{x}$？
  - 投影！
  - $\boldsymbol{x} \approx \bar{\boldsymbol{x}} + (\boldsymbol{w}_1^T(\boldsymbol{x} - \bar{\boldsymbol{x}}))\boldsymbol{w}_1$
  - 所以，$y_i = \boldsymbol{w}_1^T(\boldsymbol{x} - \bar{\boldsymbol{x}})$
  - 那么，$b = ?$



J_1和J_2的等价关系

- 若干向量
  - $\boldsymbol{x}_i$：降维之前的向量
  - $\boldsymbol{w}_1^T(\boldsymbol{x}_i - \overline{\boldsymbol{x}})\boldsymbol{w}_1 = y_i\boldsymbol{w}_1$：降维之后的向量
  - $\widehat{\boldsymbol{x}}$：在原空间中重建的向量
  - 目前的重建关系：$\widehat{\boldsymbol{x}}_{\boldsymbol{i}} \approx \overline{\boldsymbol{x}} + y_i\boldsymbol{w}_1$
- J_1的目的是使得$\widehat{\boldsymbol{x}}_i$和$\boldsymbol{x}_i$尽可能相差小($\overline{\boldsymbol{x}}$固定为均值)
  - $J_1(\boldsymbol{w}, \boldsymbol{a}) = \sum_{i=1}^{n} \frac{1}{n} \|\boldsymbol{x}_i - (\overline{\boldsymbol{x}} + a_i\boldsymbol{w})\|^2$
  - $\boldsymbol{w}$：投影方向，a_i：投影系数
- 最小化J_1得到的a_i和$\boldsymbol{w}$与J_2得到的结果完全一致！
  - 试着去证明!

如果需要更多投影方向？

- ✓ What if we need $\boldsymbol{w}_2, \boldsymbol{w}_3, \ldots$
  - 新的投影方向需要继续保持“能量”
  - 但是需要限制
  - $\boldsymbol{w}_2 \perp \boldsymbol{w}_1$， $\boldsymbol{w}_3 \perp \boldsymbol{w}_2$， $\boldsymbol{w}_3 \perp \boldsymbol{w}_1$， …
- ✓ 在上述限制条件下
  - $\boldsymbol{w}_2 = \boldsymbol{\xi}_2$， $\boldsymbol{w}_3 = \boldsymbol{\xi}_3$， …
  - 重建系数: $\boldsymbol{w}_j^T(\boldsymbol{x} - \overline{\boldsymbol{x}})$
- ✓ 总之，

$$\boldsymbol{x} \approx \overline{\boldsymbol{x}} + (\boldsymbol{w}_1^T(\boldsymbol{x} - \overline{\boldsymbol{x}}))\boldsymbol{w}_1 + (\boldsymbol{w}_2^T(\boldsymbol{x} - \overline{\boldsymbol{x}}))\boldsymbol{w}_2 + \cdots$$

重建和原数据的关系

- 假设$n > d$，即数据比维数多
  - 进一步假设$Cov(\boldsymbol{x})$可逆
  - 如果$n < d$，那么情况如何、还能做PCA变换吗？
- $Cov(\boldsymbol{x})$是$d \times d$矩阵，有d个互相垂直的特征向量$\boldsymbol{\xi}_i$
  - 重建会有d个互相垂直的投影方向$\boldsymbol{w}_i$

$$\forall \boldsymbol{x}, \qquad \boldsymbol{x} = \bar{\boldsymbol{x}} + \sum_{i=1}^{d} \left(\boldsymbol{w}_i^T(\boldsymbol{x} - \bar{\boldsymbol{x}})\right)\boldsymbol{w}_i$$

  - 将$\boldsymbol{w}_i$拼成矩阵形式$W = [\boldsymbol{w}_1 \ \ \boldsymbol{w}_2 \ \ \cdots \boldsymbol{w}_d]$ $(d \times d)$
  - 投影系数是$W^T(\boldsymbol{x} - \bar{\boldsymbol{x}})$，投影方向是$W$
  - $\boldsymbol{x} = \bar{\boldsymbol{x}} + WW^T(\boldsymbol{x} - \bar{\boldsymbol{x}})$ （为什么？）
  - 重建是完全精确的（没有误差），为什么？

降维

- 很多时候，有些投影方向是噪声
  - 需要扔掉一些方向
  - 扔掉哪些？扔掉多少?
- 去掉特征值最小的那些
- 通常保持90%的能量

$$\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_T}{\lambda_1 + \lambda_2 + \cdots + \lambda_d} > 0.9$$

  - 寻找第一个T，使得上面的不等式成立

降维的损失

- 现在$\widehat{W} = [\boldsymbol{w}_1 \ \ \boldsymbol{w}_2 \ \ \cdots \boldsymbol{w}_T]\ (d \times T)$
- $\boldsymbol{x} - \widehat{\boldsymbol{x}} = \sum_{j=T+1}^{d} (\boldsymbol{w}_j^T (\boldsymbol{x} - \overline{\boldsymbol{x}}))\boldsymbol{w}_j = \sum_{j=T+1}^{d} \boldsymbol{e}_j$
  - 这个误差多大?
  - $\boldsymbol{e}_j^T \boldsymbol{e}_k = 0$ 如果 $j \neq k$ (为什么？)
  - $E\left(\|\boldsymbol{x} - \widehat{\boldsymbol{x}}\|^2\right) = \sum_{j=T+1}^{d} E\left(\left\|\boldsymbol{e}_j\right\|^2\right)$ (为什么？)
  - $E\left(\left\|\boldsymbol{e}_j\right\|^2\right) = \lambda_j$ (为什么？)
- 这样降维保证平均（期望）重建误差最小
  - 直接优化重建误差J_1得到同样的结果

小结：PCA变换的步骤

- ✓ 训练样本：$\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n$
- ✓ 计算得到$\overline{\boldsymbol{x}}$和$Cov(\boldsymbol{x})$
- ✓ 求得$Cov(\boldsymbol{x})$的特征值和特征向量
  - Matlab，R，octave，numpy, …
- ✓ 根据特征值选定T
- ✓ 根据T的值确定矩阵$\widehat{W}$

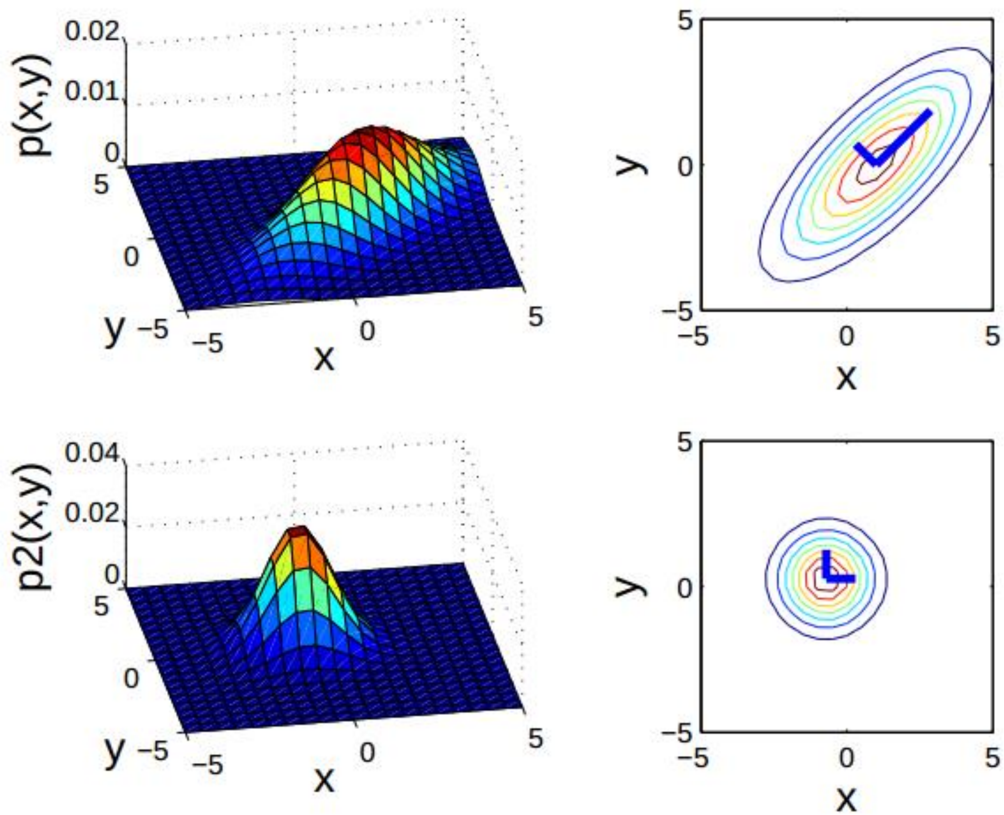
- ✓ 对任何数据$\boldsymbol{x}$，其新的经过PCA变换得到的特征是

$$\boldsymbol{y} = \widehat{W}^T(\boldsymbol{x} - \overline{\boldsymbol{x}})$$

重建则为 $\boldsymbol{x} \approx \widehat{\boldsymbol{x}} = \overline{\boldsymbol{x}} + \widehat{W}\boldsymbol{y}$

正态分布与PCA

PDF和等概率曲线



Generated from Kevin Murphy's toolbox

PCA vs. Gaussian

- ✓ $\boldsymbol{x}$服从D维高斯分布$N(\boldsymbol{\mu}, \Sigma)$
  - $p(\boldsymbol{x}) = (2\pi)^{-\frac{D}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}$
- ✓ 对$\boldsymbol{x}$进行PCA操作，结果是什么？
  - W =?
  - $\widehat{W}$ =?
  - PCA完成后的新特征是什么样子的？

PCA of Gaussian

- ✓ $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \Sigma)$
- ✓ 假设使用全部特征向量，则$\boldsymbol{y} = W^T(\boldsymbol{x} - \boldsymbol{\mu})$
  - $\Sigma = \sum_{i=1}^{d} \lambda_i \boldsymbol{\xi}_i \boldsymbol{\xi}_i^T = \sum_{i=1}^{d} \lambda_i \boldsymbol{w}_i \boldsymbol{w}_i^T = W \Lambda W^T$
    - Λ是一个对角矩阵，$\Lambda_{ii} = \lambda_i$
  - $WW^T = ?$　　$W^TW = ?$
  - $E\boldsymbol{y} = ?$
  - $Cov(\boldsymbol{y}) = ?$
- ✓ $\boldsymbol{y} \sim N(\boldsymbol{0}, \Lambda)$!
- ✓ PCA旋转了数据，使得新特征各个维度互不相关
  - 对高斯分布，不相关意味着互相独立！

PCA的优点

- ✓ 减少了数据量
  - 可以减少计算量，缩短训练、测试、识别时间
  - 可以减少所需的存储空间
    - 对大规模数据特别重要
  - 可能去除数据中的噪声
    - 所以可能提高系统的识别精确度
- ✓ 如果数据服从高斯分布
  - 完成PCA后，新特征各维度互不相关
  - 有利于模式识别

白化变换

✓ 协方差矩阵Σ，可以从训练样本估计

- PCA： $\boldsymbol{y} = W^T(\boldsymbol{x} - \boldsymbol{\mu})$
- $\Sigma = \sum_{i=1}^{d} \lambda_i \boldsymbol{\xi}_i \boldsymbol{\xi}_i^T = \sum_{i=1}^{d} \lambda_i \boldsymbol{w}_i \boldsymbol{w}_i^T = W\Lambda W^T$

✓ 白化变化(Whitening transform):

- $\boldsymbol{y} = (W\Lambda^{-\frac{1}{2}})^T(\boldsymbol{x} - \boldsymbol{\mu})$
- 如何计算$\Lambda^{-\frac{1}{2}}$?

✓ $\boldsymbol{y} \sim N(\mathbf{0}, I)$! (如果原数据是高斯)

- 各向同性

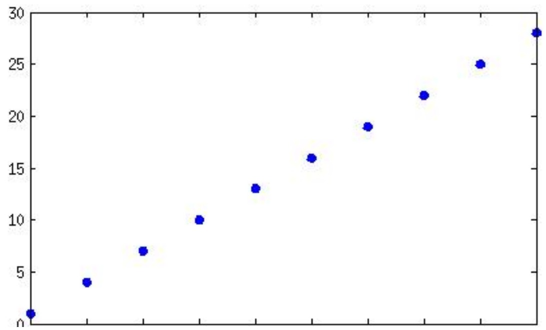
高斯假设

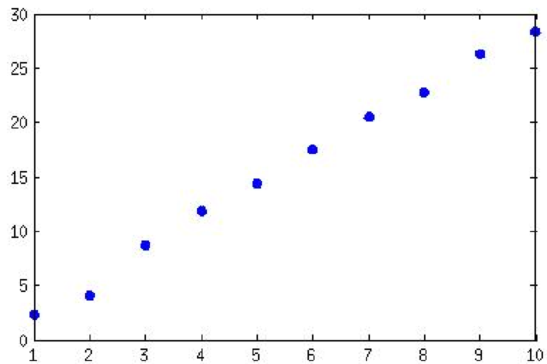
- PCA变换
  - PCA变换不一定要求$\boldsymbol{x}$服从高斯分布
  - $\boldsymbol{x}$不服从高斯分布时，$E(\boldsymbol{y}) = \mathbf{0}$，$Cov(\boldsymbol{y}) = \Lambda$，但$\boldsymbol{y}$不服从高斯分布
  - $\boldsymbol{x}$不服从高斯分布时，$\boldsymbol{y}$的各维度不相关，但不独立

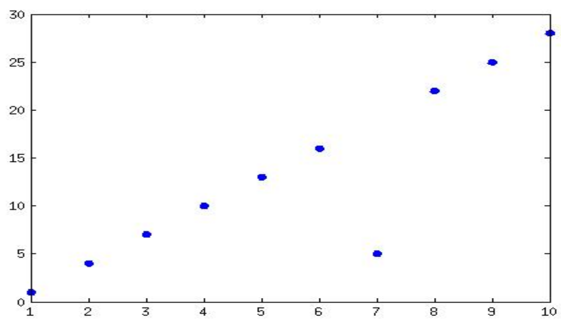
- 白化变换
  - 白化变换不一定要求$\boldsymbol{x}$服从高斯分布
  - $\boldsymbol{x}$不服从高斯分布时，$E(\boldsymbol{y}) = \mathbf{0}$，$Cov(\boldsymbol{y}) = I$，但$\boldsymbol{y}$不服从高斯分布
  - $\boldsymbol{x}$不服从高斯分布时，$\boldsymbol{y}$的各维度不相关，但不独立

Can I use PCA?

- ✓ 如果数据服从高斯分布
  - 单峰分布（unimodal distribution)
  - 白噪声(white noise)
    - $\boldsymbol{x} = \boldsymbol{x}' + \boldsymbol{\epsilon}$
    - $\boldsymbol{\epsilon} \sim \boldsymbol{N}(\boldsymbol{0}, \Gamma)$
    - 噪声独立于数据，噪声均值为**0**(zero mean)，噪声各维互相独立(Γ是对角阵)，噪声幅度有限(finite variance)
    - 此时PCA效果最佳
- ✓ 实际上，如果特征值服从指数递减即可
- ✓ 能处理离群值吗？（outlier）







特征值(82.5, 0)

方向 (3,1)

特征值 (83.713 0.42)

方向 (3.03 1)

特征值 (85.897 16.43)

方向 (3.43 1)

进一步的阅读

- ✓ 如果对本章的内容感兴趣，可以参考如下文献
  - DHS相关部分
    - 提示：使用DHS中的index（索引）和目录
  - PRML相关部分
  - 拉格朗日乘子法：
    - 经典教材：Convex Optimization， by Boyd and Vandenberghe
    - http://www.stanford.edu/~boyd/cvxbook/
    - 有电子书可以下载
    - 第五章
- ✓ https://cs.nju.edu.cn/wujx/paper/AAAI2023_AFM.pdf