

# 模式识别

主成分分析

Principal Component Analysis

吴建鑫

南京大学计算机系，2017

# 与领域无关的特征提取

- ✓ domain independent feature extraction
- ✓ 第4章. PCA
- ✓ 第5章. 特征的归一化(normalization)
- ✓ 第5章. FLD
- ✓ 首先介绍PCA

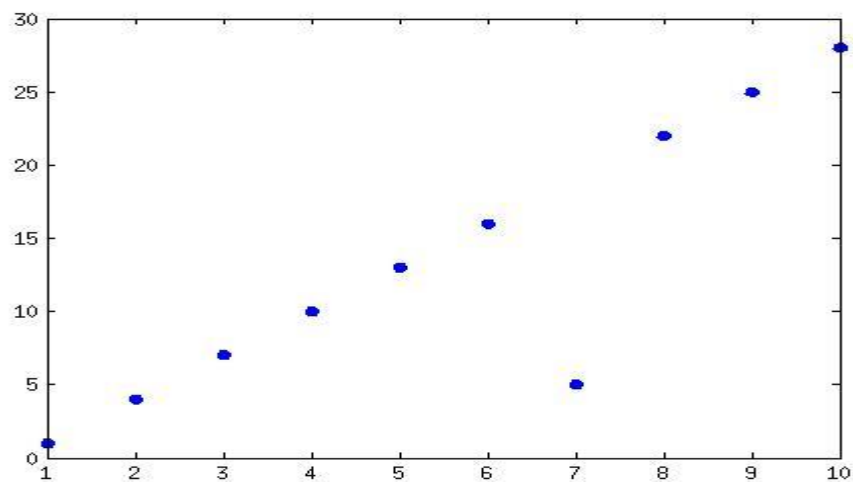
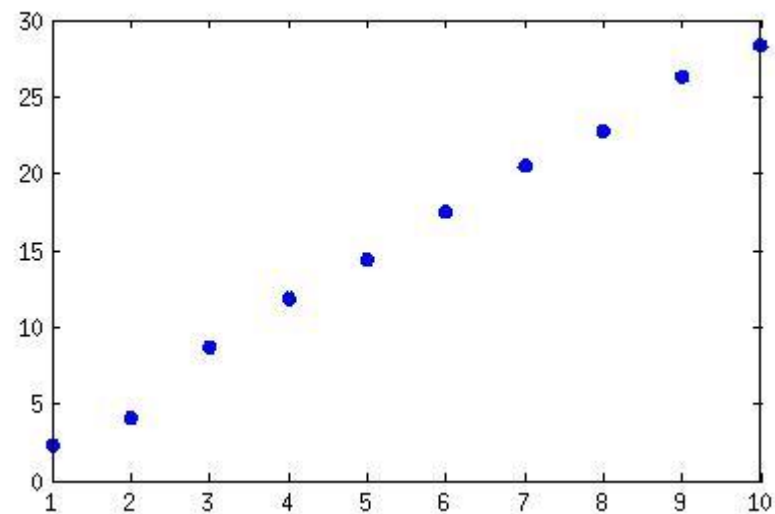
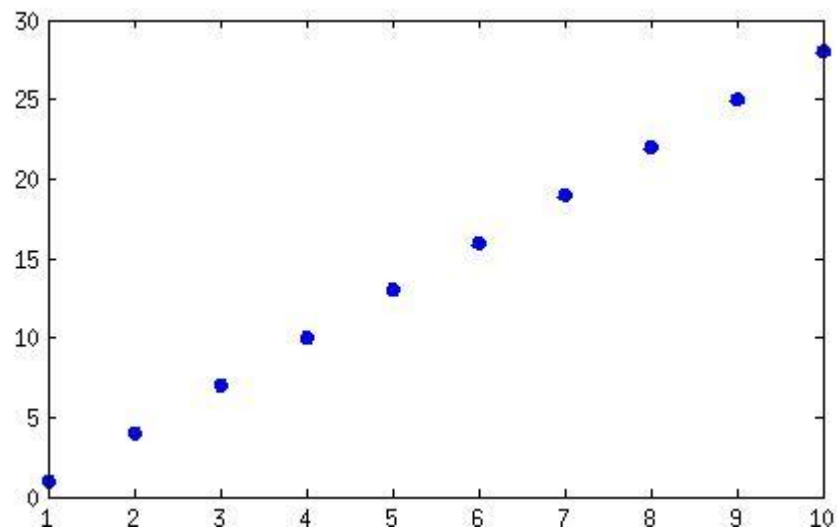
# 目标

- ✓ 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？



# 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  $\mathbf{x}$ ?

- ✓ 寻找某个固定 (constant) 的  $\mathbf{m}$ ，使得

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

- ✓ 最优解：(证明?)

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

# 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

✓  $\mathbf{x} \in \mathbb{R}^d$ ：原来的高维数据（随机变量）

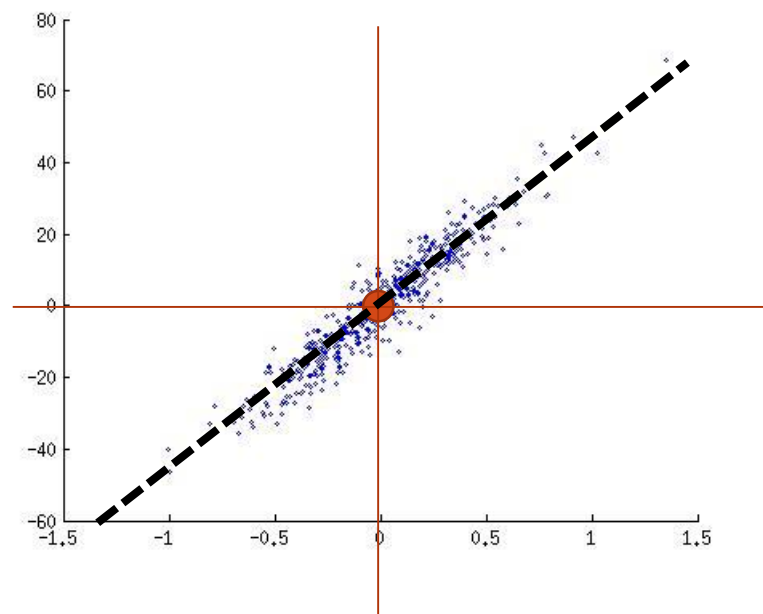
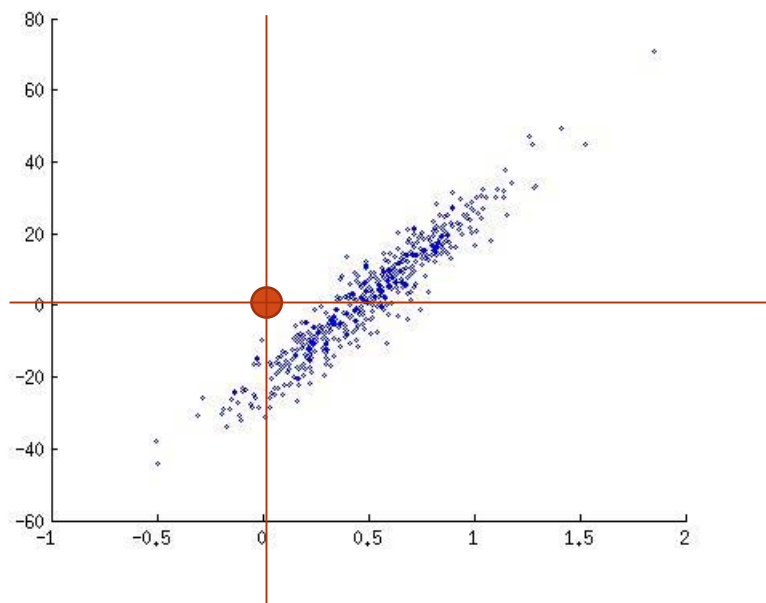
- 训练样本： $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$

✓ 假设 $m = 1$ ，用原数据的单个线性组合来表示

- $y_i = \mathbf{w}^T \mathbf{x}_i + b$
- $y_1, y_2, \dots, y_n$  -- 新的数据/特征(features)
- 如何寻找最佳的 $\mathbf{w}$ ？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  $J_2(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n \|\mathbf{w}^T (\mathbf{x}_i - \bar{\mathbf{x}})\|^2$

✓ 发现问题了吗?

- $J_2(\mathbf{w})$  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  $\|\mathbf{w}\|^2 = \mathbf{w}^T \mathbf{w} = 1$

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

$$\text{s. t.} \quad \mathbf{w}^T \mathbf{w} = 1$$

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

# 简化simplification 变换 transformation

$$\begin{aligned}\|(\mathbf{x}_i - \bar{\mathbf{x}})^T \mathbf{w}\|^2 &= ((\mathbf{x}_i - \bar{\mathbf{x}})^T \mathbf{w})^T ((\mathbf{x}_i - \bar{\mathbf{x}})^T \mathbf{w}) \\ &= \mathbf{w}^T (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T \mathbf{w}\end{aligned}$$

$$\begin{aligned}\frac{1}{n} \sum_{i=1}^n \|(\mathbf{x}_i - \bar{\mathbf{x}})^T \mathbf{w}\|^2 &= \mathbf{w}^T \sum_{i=1}^n \frac{1}{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T \mathbf{w} \\ &= \mathbf{w}^T \text{Cov}(\mathbf{x}) \mathbf{w}\end{aligned}$$

# 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

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

- ✓  $\lambda$ : 拉格朗日乘子 Lagrange multiplier

- ✓ 最优的必要条件:  $\frac{\partial f}{\partial \mathbf{w}} = \mathbf{0}, \quad \frac{\partial f}{\partial \lambda} = 0$

- ✓  $\frac{\partial f}{\partial \mathbf{w}} = 2\text{Cov}(\mathbf{x})\mathbf{w} - 2\lambda\mathbf{w} = \mathbf{0}$

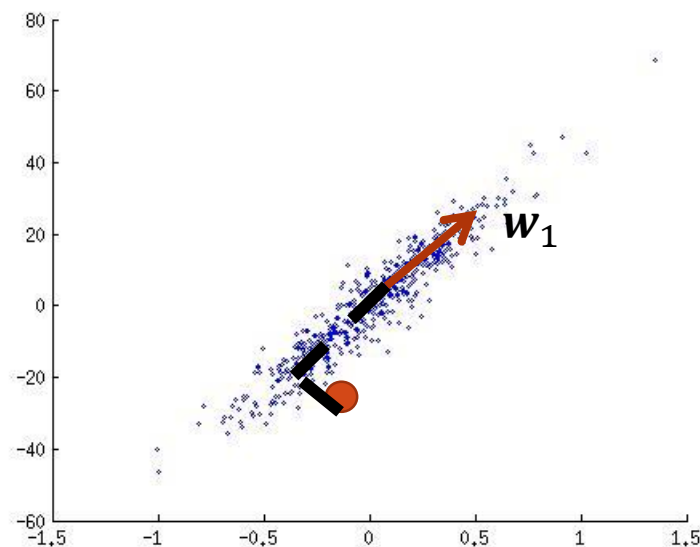
- 我们这里的前提条件是什么?

- 应该想到用哪一个公式?

- ✓  $\text{Cov}(\mathbf{x})\mathbf{w} = \lambda\mathbf{w}, \quad \mathbf{w}^T \mathbf{w} = 1!$

# 选哪个特征向量？

- ✓  $J_2(\mathbf{w}) = \mathbf{w}^T \text{Cov}(\mathbf{x}) \mathbf{w} = ?$
- ✓  $\text{Cov}(\mathbf{x})$  是半正定的（如何证明？）
- ✓ 选取  $\lambda_1$ （即最大特征值）对应的特征向量  $\xi_1$  为  $\mathbf{w}_1$ 
  - 为什么？约束条件满足了吗？
- ✓ 怎样用  $\mathbf{w}_1$  来近似  $\mathbf{x}$  ?
  - 投影！
  - $\mathbf{x} \approx \bar{\mathbf{x}} + (\mathbf{w}_1^T (\mathbf{x} - \bar{\mathbf{x}})) \mathbf{w}_1$
  - 所以，  $y_i = \mathbf{w}_1^T (\mathbf{x} - \bar{\mathbf{x}})$
  - 那么，  $b = ?$



# $J_1$ 和 $J_2$ 的等价关系

## ✓ 若干向量

- $\mathbf{x}_i$ : 降维之前的向量
- $\mathbf{w}_1^T(\mathbf{x}_i - \bar{\mathbf{x}})\mathbf{w}_1 = y_i\mathbf{w}_1$ : 降维之后的向量
- $\hat{\mathbf{x}}$ : 在原空间中重建的向量
- 目前的重建关系:  $\hat{\mathbf{x}}_i \approx \bar{\mathbf{x}} + y_i\mathbf{w}_1$

## ✓ $J_1$ 的目的是使得 $\hat{\mathbf{x}}_i$ 和 $\mathbf{x}_i$ 尽可能相差小( $\bar{\mathbf{x}}$ 固定为均值)

- $J_1(\mathbf{w}, \mathbf{a}) = \sum_{i=1}^n \frac{1}{n} \|\mathbf{x}_i - (\bar{\mathbf{x}} + a_i\mathbf{w})\|^2$
- $\mathbf{w}$ : 投影方向,  $a_i$ : 投影系数

## ✓ 最小化 $J_1$ 得到的 $a_i$ 和 $\mathbf{w}$ 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  $\mathbf{w}_2, \mathbf{w}_3, \dots$

- 新的投影方向需要继续保持“能量”
- 但是需要限制
- $\mathbf{w}_2 \perp \mathbf{w}_1, \quad \mathbf{w}_3 \perp \mathbf{w}_2, \quad \mathbf{w}_3 \perp \mathbf{w}_1, \quad \dots$

✓ 在上述限制条件下

- $\mathbf{w}_2 = \boldsymbol{\xi}_2, \quad \mathbf{w}_3 = \boldsymbol{\xi}_3, \quad \dots$
- 重建系数:  $\mathbf{w}_j^T (\mathbf{x} - \bar{\mathbf{x}})$

✓ 总之,

$$\mathbf{x} \approx \bar{\mathbf{x}} + (\mathbf{w}_1^T (\mathbf{x} - \bar{\mathbf{x}})) \mathbf{w}_1 + (\mathbf{w}_2^T (\mathbf{x} - \bar{\mathbf{x}})) \mathbf{w}_2 + \dots$$

# 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  $Cov(\mathbf{x})$  可逆
  - 如果  $n < d$ ，那么情况如何、还能做PCA变换吗？
- ✓  $Cov(\mathbf{x})$  是  $d \times d$  矩阵，有  $d$  个互相垂直的特征向量  $\xi_i$ 
  - 重建会有  $d$  个互相垂直的投影方向  $\mathbf{w}_i$

$$\forall \mathbf{x}, \quad \mathbf{x} = \bar{\mathbf{x}} + \sum_{i=1}^d (\mathbf{w}_i^T (\mathbf{x} - \bar{\mathbf{x}})) \mathbf{w}_i$$

- 将  $\mathbf{w}_i$  拼成矩阵形式  $W = [\mathbf{w}_1 \ \mathbf{w}_2 \ \cdots \ \mathbf{w}_d]$  ( $d \times d$ )
- 投影系数是  $W^T (\mathbf{x} - \bar{\mathbf{x}})$ ，投影方向是  $W$
- $\mathbf{x} = \bar{\mathbf{x}} + WW^T (\mathbf{x} - \bar{\mathbf{x}})$  (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  $\hat{W} = [\mathbf{w}_1 \ \mathbf{w}_2 \ \cdots \ \mathbf{w}_T] \ (d \times T)$
- ✓  $\mathbf{x} - \hat{\mathbf{x}} = \sum_{j=T+1}^d (\mathbf{w}_j^T (\mathbf{x} - \bar{\mathbf{x}})) \mathbf{w}_j = \sum_{j=T+1}^d \mathbf{e}_j$ 
  - 这个误差多大?
  - $\mathbf{e}_j^T \mathbf{e}_k = 0$  如果  $j \neq k$  (为什么?)
  - $E(\|\mathbf{x} - \hat{\mathbf{x}}\|^2) = \sum_{j=T+1}^d E(\|\mathbf{e}_j\|^2)$  (为什么?)
  - $E(\|\mathbf{e}_j\|^2) = \lambda_j$  (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

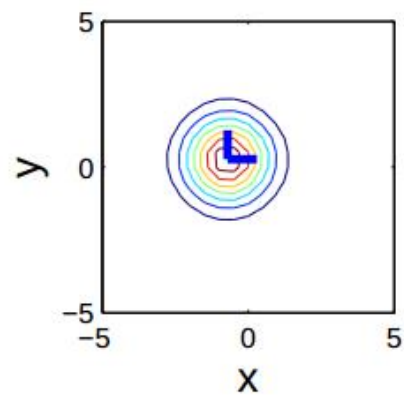
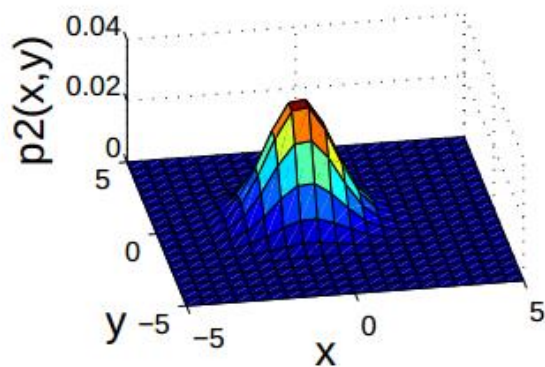
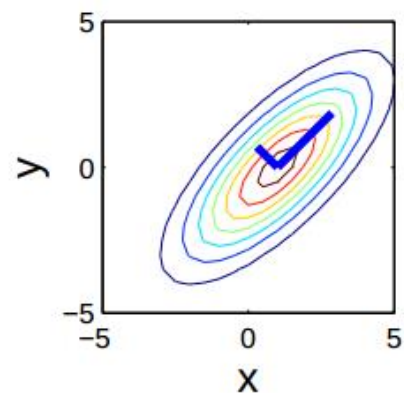
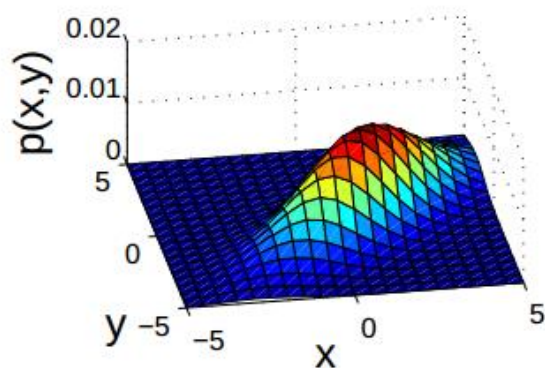
# 小结：PCA变换的步骤

- ✓ 训练样本：  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$
- ✓ 计算得到  $\bar{\mathbf{x}}$  和  $Cov(\mathbf{x})$
- ✓ 求得  $Cov(\mathbf{x})$  的特征值和特征向量
  - Matlab, R, octave, ...
- ✓ 根据特征值选定  $T$
- ✓ 根据  $T$  的值确定矩阵  $\hat{W}$
- ✓ 对任何数据  $\mathbf{x}$ ，其新的经过PCA变换得到的特征是
$$\mathbf{y} = \hat{W}^T (\mathbf{x} - \bar{\mathbf{x}})$$
重建则为  $\mathbf{x} \approx \hat{\mathbf{x}} = \bar{\mathbf{x}} + \hat{W} \mathbf{y}$

# 正态分布与PCA

---

# PDF和等概率曲线



# PCA vs. Gaussian

✓  $\mathbf{x}$  服从  $D$  维高斯分布  $N(\boldsymbol{\mu}, \Sigma)$

- $p(\mathbf{x}) = (2\pi)^{-\frac{D}{2}} |\Sigma|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}$

✓ 对  $\mathbf{x}$  进行 PCA 操作，结果是什么？

- $W = ?$
- $\hat{W} = ?$
- PCA 完成后的新特征是什么样子的？

# PCA of Gaussian

- ✓  $\mathbf{x} \sim N(\boldsymbol{\mu}, \Sigma)$
- ✓ 假设使用全部特征向量, 则  $\mathbf{y} = W^T(\mathbf{x} - \boldsymbol{\mu})$ 
  - $\Sigma = \sum_{i=1}^d \lambda_i \boldsymbol{\xi}_i \boldsymbol{\xi}_i^T = \sum_{i=1}^d \lambda_i \mathbf{w}_i \mathbf{w}_i^T = W \Lambda W^T$ 
    - $\Lambda$  是一个对角矩阵,  $\Lambda_{ii} = \lambda_i$
  - $WW^T = ? \quad W^TW = ?$
  - $E\mathbf{y} = ?$
  - $Cov(\mathbf{y}) = ?$
- ✓  $\mathbf{y} \sim N(\mathbf{0}, \Lambda)$ !
- ✓ 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 $\Sigma$ ，可以从训练样本估计

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

✓ 白化变化 (Whitening transform):

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

✓  $\mathbf{y} \sim N(\mathbf{0}, I)$  !

- 各向同性

# 高斯假设

## ✓ PCA变换

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

## ✓ 白化变换

- 白化变换不一定要要求 $\mathbf{x}$ 服从高斯分布
- $\mathbf{x}$ 不服从高斯分布时,  $E(\mathbf{y}) = \mathbf{0}$ ,  $Cov(\mathbf{y}) = I$ , 但 $\mathbf{y}$ 不服从高斯分布
- $\mathbf{x}$ 不服从高斯分布时,  $\mathbf{y}$ 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)

- $\mathbf{x} = \mathbf{x}' + \boldsymbol{\epsilon}$

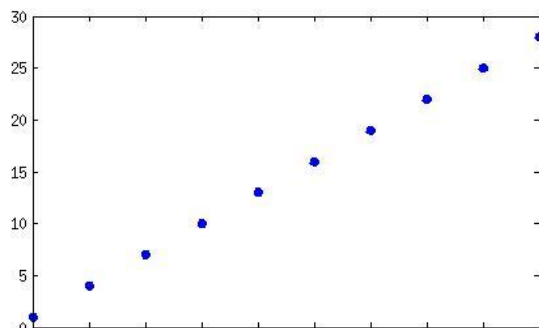
- $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \Gamma)$

- 噪声独立于数据，噪声均值为 $\mathbf{0}$  (zero mean)，噪声各维互相独立 ( $\Gamma$  是对角阵)，噪声幅度有限 (finite variance)

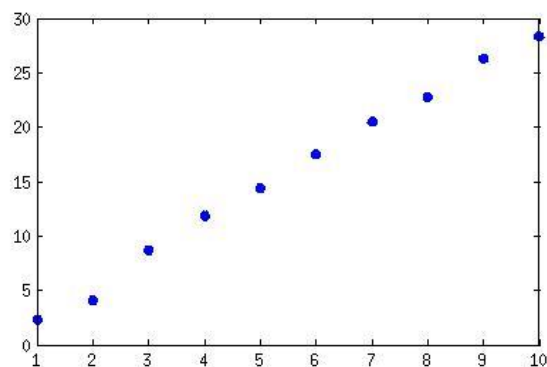
- 此时PCA效果最佳

## ✓ 实际上，如果特征值服从指数递减即可

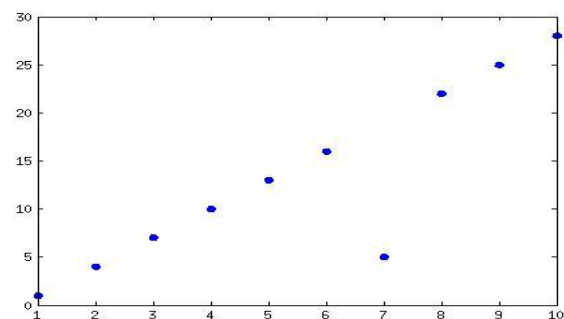
## ✓ 能处理离群值吗？ (outlier)



特征值 (825, 0)  
方向 (3, 1)



特征值 (837.13 0.42)  
方向 (3.03 1)



特征值 (858.97 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