2021

# 补充内容：主成分分析

Principal components analysis：PCA

# 降维方法研究背景

- 为什么降维?

  维数灾难：样本的维度的增加而呈指数形式增长。

  降维的意义：克服维数灾难，获取本质特征，

  节省存储空间，去除无用噪声，

  实现数据可视化(大于三维的数据无法可视化)

- 降维方法分类

  **①线性降维方法**

  主成分分析(PCA), 判别分析(LDA), 多维尺度分析(MDS)

  **②非线性降维方法**

  流形学习，等距特征映射，局部线性嵌入

- 高维数据应用背景：
  - 特点：维度偏大，
  - 维度灾难：
    随着技术的进步，数据收集越来越容易，导致数据规模越来越大、复杂性越来越高，其维度（属性）通常可以达到成百上千维，甚至更高。
  - 实际情况：超过 3 维的数据，很难直观表示。
    面临存储、计算问题。
  - 高维数据中含有很多冗余信息，因此需要对其降维、简化处理。

# 数据介绍

表 1

数据维度

| 学生编号 | 语文 | 数学 | 物理 | 化学 |
|---|---|---|---|---|
| 1 | 90 | 140 | 99 | 100 |
| 2 | 90 | 97 | 88 | 92 |
| 3 | 90 | 110 | 79 | 83 |
| ． ． ． | ． ． ． | ． ． ． | ． ． ． | ． ． ． |

- 数据:
    - 样本数量(1, 2, . . . . )
    - 维度(特征个数)
    - 不同特征起的作用不一样
    - 维度过高时，对存储，计算分析带来困难
    - 需要去除冗余特征，保留重要的成分。

# Principal components analysis：PCA

- PCA:是一种对数据进行分析简化的技术

  1901年提出

  主成分分析:

  - 有效的找出数据中最“主要”的元素和结构，
  - 去除噪音和冗余，将原有的复杂数据降维简化
  - 揭示隐藏在复杂数据背后的简单结构。

  优点是简单，而且无参数限制

  - 应用范围极其广泛，从神经科学到计算机图形学都有它的用武之地。

  被誉为应用线性代数最有价值的结果之一

# 例子1

表 1

数据维度

| 学生编号 | 语文 | 数学 | 物理 | 化学 |
|---|---|---|---|---|
| 1 | 90 | 140 | 99 | 100 |
| 2 | 90 | 97 | 88 | 92 |
| 3 | 90 | 110 | 79 | 83 |
| · · · | · · · | · · · | · · · | · · · |

- 表1容易看出，数学、物理、化学，主要成分，其中数学是最重要的成分(区分度大)

表 2

| 学生编号 | 数学 | 物理 | 化学 | 语文 | 历史 | 英语 |
|---|---|---|---|---|---|---|
| 1 | 65 | 61 | 72 | 84 | 81 | 79 |
| 2 | 77 | 77 | 76 | 64 | 70 | 55 |
| 3 | 67 | 63 | 49 | 65 | 67 | 57 |
| 4 | 80 | 69 | 75 | 74 | 74 | 63 |
| 5 | 74 | 70 | 80 | 84 | 82 | 74 |
| 6 | 78 | 84 | 75 | 62 | 72 | 64 |
| 7 | 66 | 71 | 67 | 52 | 65 | 57 |
| 8 | 77 | 71 | 57 | 72 | 86 | 71 |
| 9 | 83 | 100 | 79 | 41 | 67 | 50 |
| · · · | · · · | · · · | · · · | · · · | · · · | · · · |

- 表2则维度大，数据多，分布散乱，无法直接看出其主成分。
- 如果把这些数据换一个空间表示，即换一个观察角度，找出主成分。

# 1 PCA概述

- PCA:是一种分析、简化数据集的技术
  - 提取、保留数据的主要成分，去掉数据的次要成分。
  - 高维数据降维方法。
- 原理：使用新的一组基(主成分)去重新描述数据空间

  新的基要能尽量揭示原有的数据间的关系
- PCA的目标:

  找到这样的“基\主元”，最大程度的去除冗余和噪音的干扰

# 1个样本数据--pca示例

- **n维样本数据** $x_1 = \begin{pmatrix} x_{11} \\ \vdots \\ x_{1n} \end{pmatrix}$
- **将其降到k维：** $y_1 = \begin{pmatrix} y_{11} \\ \vdots \\ y_{1k} \end{pmatrix}$
- **PCA的目标:寻找一组基函数** $P = \begin{pmatrix} \vec{p}_1 \\ \vdots \\ \vec{p}_k \end{pmatrix}$, $\vec{p}_k = (p_{k1}, \cdots, p_{kn})$为$n$维向量
- **满足以下公式：** $y_1 = Px_1$
- **如何求得 P ？**

# m个样本数据--pca示例

- **m个n维数据X** $X=(x_1,\cdots x_m)$,其中$x_i$为$n$维向量

- **将其降到k维：** $Y=(y_1,\cdots y_m)$, $y_m$为$k$维向量

- **PCA的目标:寻找一组基函数**

$$P=\begin{pmatrix}\vec{p}_1\\ \vdots \\ \vec{p}_k\end{pmatrix},\quad \vec{p}_k=(p_{k1},\cdots,p_{kn})\text{为}n\text{维向量}$$

- **满足以下公式：** $Y=PX$
- **如何求得 P ？**

  **同时满足：①尽量少损失原始数据信息，②又去除数据冗余**

- **用什么方法来衡量以上两点？？**

  **(方差：描述数据间分散程度,也即是满足第一条**

  **协方差：描述数据间的独立性)**

# 1.2 内积与投影：

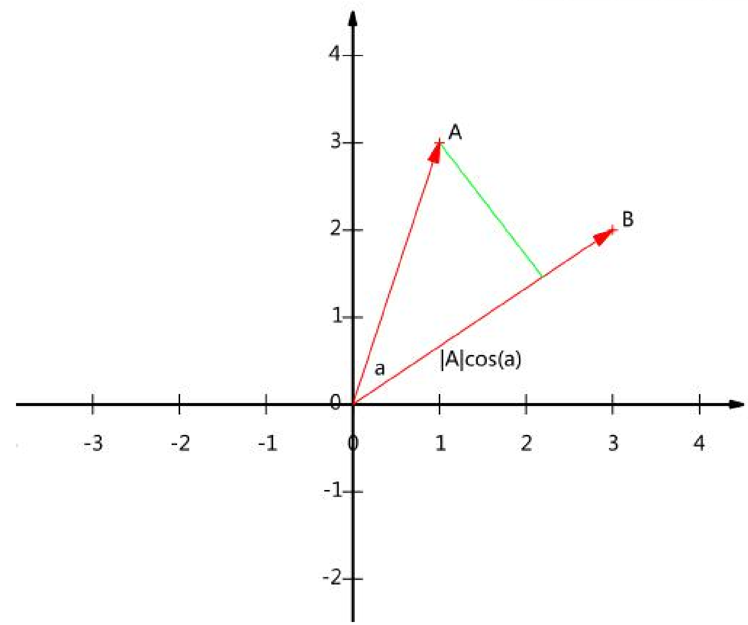
- 内积：二维向量 A , B 内积定义

$$A \cdot B = |A||B|\cos\theta$$

- 投影：假设 B 为单位向量，|B|=1,

$$A \cdot B = |A|\cos\theta$$

- A与B的**内积值**等于A向B所在直线投影的矢量长度

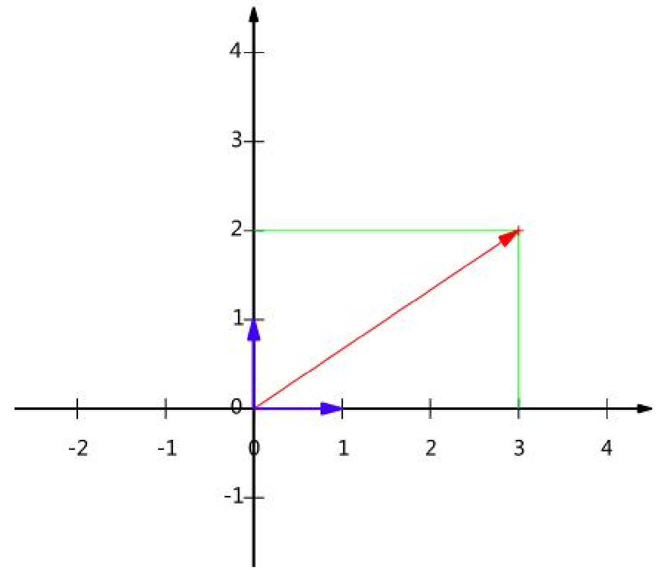


- 右图中二维坐标系的基$\vec{e}_1=(1,0),\vec{e}_2=(0,1)$
- 向量(3,2)完整表示为:

$$(3,2)=3\vec{e}_1+2\vec{e}_2$$

- 在x轴方向的投影为3,在y轴方向的投影为2
- 上式又可写为

$$(3,2)=<(3,2),\vec{e}_1>\vec{e}_1+<(3,2),\vec{e}_2>\vec{e}_2$$




因此，对于$n$维向量 $\vec{a}=(a_1,a_2,\cdots,a_n)$,

有 $$\vec{a}=\sum_{i=1}^{n}<\vec{a},e_i>e_i$$

注：任何两个线性无关的二维向量均可为二维空间的基

不一定正交(正交基的性质较好，因此通常选取正交)

# 其它正交基组成的坐标系例子

- 右图中以紫色向量为基构成的新坐标系:

$$\vec{e}_1' = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}), \vec{e}_2' = (-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$$

- 试计算向量(3,2)在新坐标系下的坐标:

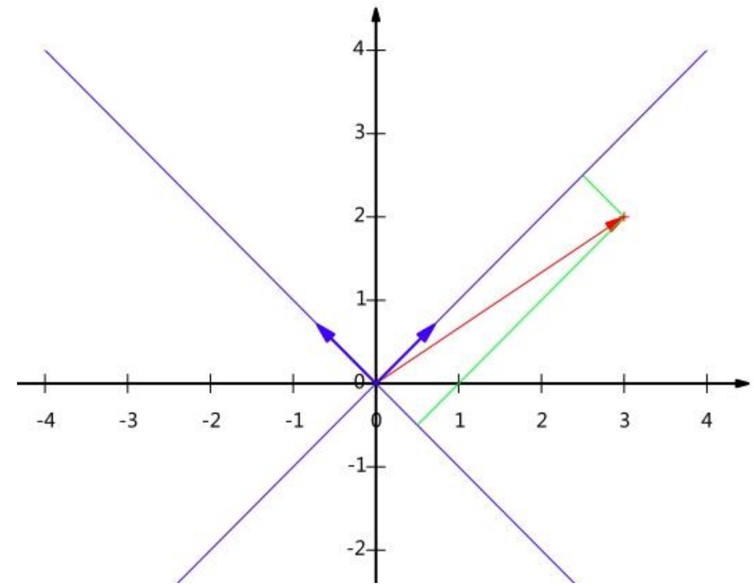
  即:分别计算在新坐标系下x,y轴方向的投影

$$x' = (3,2) \cdot (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}) = \frac{5}{\sqrt{2}}$$

$$y' = (3,2) \cdot (-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}) = -\frac{1}{\sqrt{2}}$$

- 因此新坐标为:原坐标在新基下的投影(内积)

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \vec{e}_1' \\ \vec{e}_2' \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$



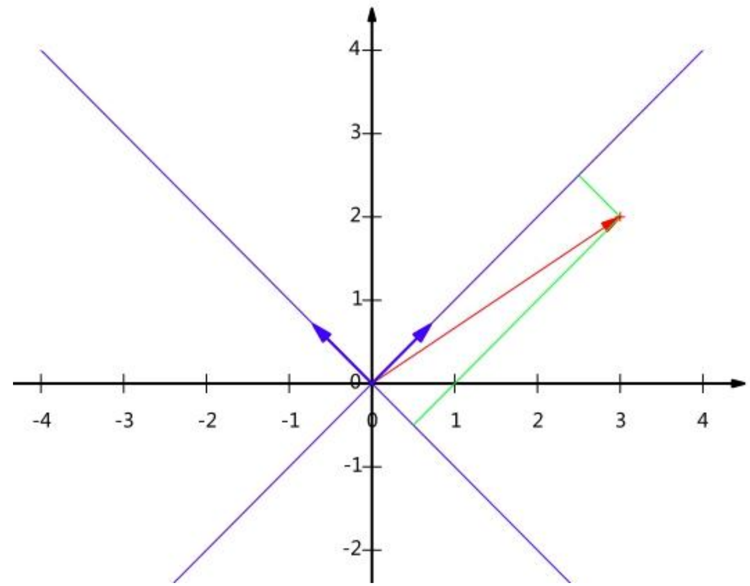
# 1.4 简单降维例子

- 给定本个样本数据(1,1),(2,2),(3,3)
- 计算其在新坐标系下的坐标:**二维数据降成一维**

$$\vec{e}_1' = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right), \vec{e}_2' = \left(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$$

$$\begin{pmatrix} \frac{2}{\sqrt{2}} & \frac{4}{\sqrt{2}} & \frac{6}{\sqrt{2}} \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} \vec{e}_1' \\ \vec{e}_2' \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}$$

$$= \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}$$




因此，选取合适的坐标基，高维数据在一定的情况下可以实现降维

(即选取 $\vec{e}_1' = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$

$$\begin{pmatrix} \frac{2}{\sqrt{2}} & \frac{4}{\sqrt{2}} & \frac{6}{\sqrt{2}} \end{pmatrix} = \begin{pmatrix} \vec{e}_1' \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}$$

$$= \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}$$

问题：对于给定的$n$维数据集：$(a_1,\cdots,a_M)$，

试图降到$R \prec n$维空间中？

$step\ 1$：选取$R$个$n$维线性无关向量$P=(p_1,\cdots p_R)$；

$step\ 2$：计算数据在该组向量基下的投影

$$\begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_R \end{pmatrix} \begin{pmatrix} a_1 & a_2 & \cdots & a_M \end{pmatrix} = \begin{pmatrix} p_1a_1 & p_1a_2 & \cdots & p_1a_M \\ p_2a_1 & p_2a_2 & \cdots & p_2a_M \\ \vdots & \vdots & \ddots & \vdots \\ p_Ra_1 & p_Ra_2 & \cdots & p_Ra_M \end{pmatrix}$$

**关键的问题：**

**如何选择基P才是最优的?**

**即，如何选取 R 个n维线性无关向量，最大程度保留原有的信息**

## 2 .1: 数据中心化

- 期望： $E=\frac{1}{N}\sum_i x_i$
- 方差： $\sigma=\frac{1}{N}\sum_i (x_i-E)^2$
- 数据中心化处理：

  数据集的每个属性减去其期望

注： 1. 特征的量纲和数值得量级都是不一样的，

2. 通过中心化处理，可以使得不同的特征具有相同的尺度

3.中心化对数据做平移，不改变数据的形状。

4 .简化计算

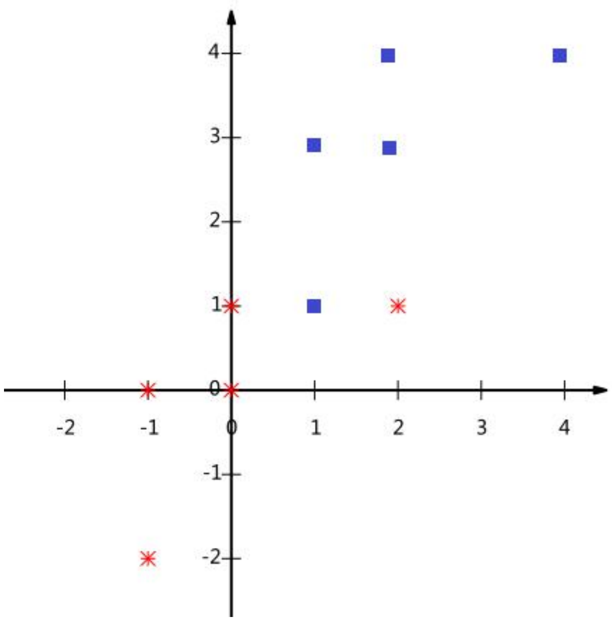
## 数据中心化处理：

**例：** 5 个二维数据，第一维均值为 2， 第二维均值为 3

第一行分别减去均值 2，第二行分别减去均值 3

$$\begin{pmatrix} 1 & 1 & 2 & 4 & 2 \\ 1 & 3 & 3 & 4 & 4 \end{pmatrix} \Rightarrow \begin{pmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{pmatrix}$$

原始数据　　　　　　　　　　中心化后数据




对应方差： $\sigma = \frac{1}{5}\sum_i (x_i - 2)^2$ 　　 $\sigma' = \frac{1}{5}\sum_i (x_i')^2$

例：$m$个中心化后样本(2个属性)数据协方差矩阵

$$X = \begin{pmatrix} a_1 a_2 & \cdots & a_m \\ b_1 b_2 & & b_m \end{pmatrix}$$

$$\frac{1}{m} XX^T = \begin{pmatrix} \frac{\sum_i a_i^2}{m} & \frac{\sum_i a_i b_i}{m} \\ \frac{\sum_i a_i b_i}{m} & \frac{\sum_i b_i^2}{m} \end{pmatrix}$$

- 注意区分样本的维度
- 与矩阵排列方式无关
- 下图为： 5 个 2 维属性(按行排)的样本集：

$$\begin{pmatrix} 1 & 1 & 2 & 4 & 2 \\ 1 & 3 & 3 & 4 & 4 \end{pmatrix}$$

- 下图为：m个n维属性(按列排)的样本集

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}_{m\times n}$$

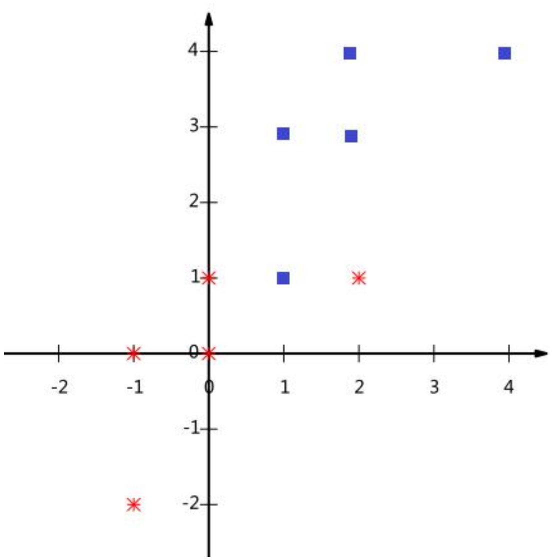


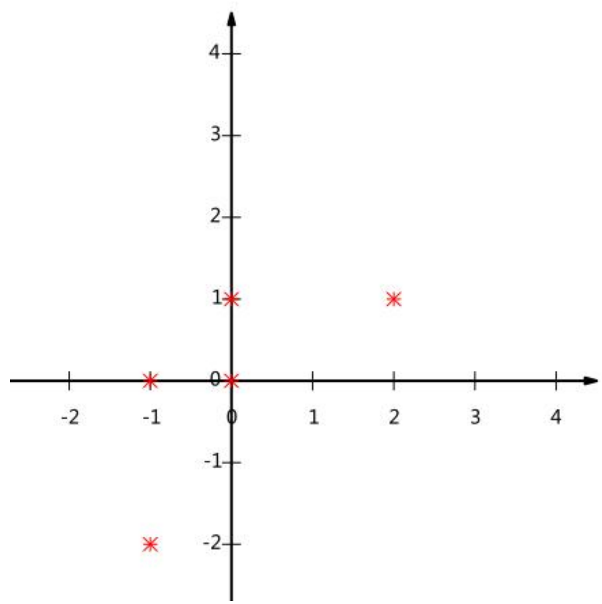
$$X = \begin{bmatrix} x_{11}-\mu_1 & x_{12}-\mu_2 & \cdots & x_{1n}-\mu_n \\ x_{21}-\mu_1 & x_{22}-\mu_2 & \cdots & x_{2n}-\mu_n \\ \cdots & \cdots & \cdots & \cdots \\ x_{m1}-\mu_1 & x_{m2}-\mu_2 & \cdots & x_{mn}-\mu_n \end{bmatrix}_{m\times n}$$

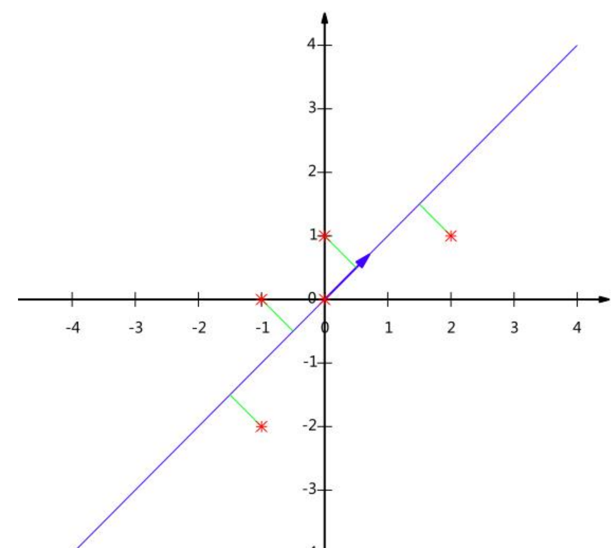
其中$\mu_i$是样本矩阵维度$i$的均值：$\mu_i = \frac{1}{m}\sum_{j=1}^{m} x_{ji}$

# 2 .2 PCA降维

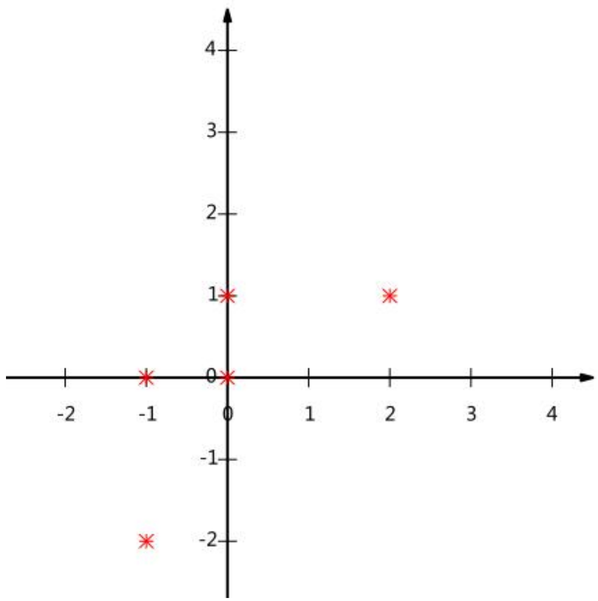
- 对于右图，中心化后的2维简单数据
  - 如何用一维表示这些数据，又尽量保留原始的信息?
- **问题转化**
  - 在二维平面中选择一个**方向**，将所有数据投影到直线上
- 如何选择该方向，且保留最多的原始信息?
  - 投影后的投影值尽可能分散

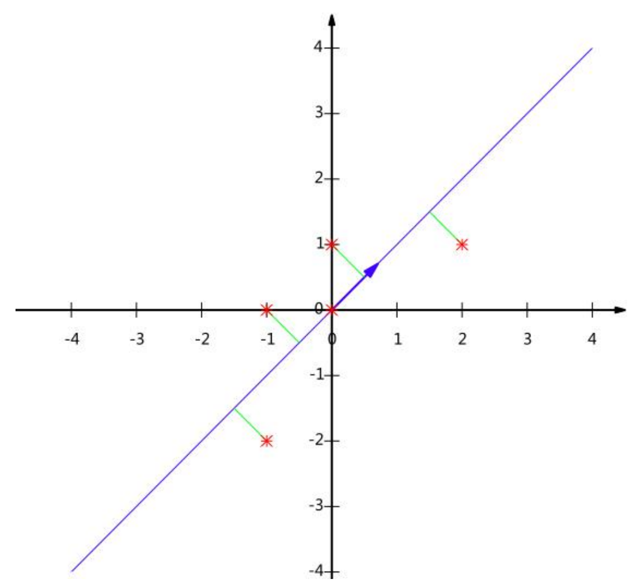


- **如何选择该方向，且保留最多的原始信息？**
  - 投影后的投影值尽可能分散
  - 向x轴投影：
    最左边的两个点会重叠，中间的两个点也会重叠，四个不相同的点投影后只剩两个，这是一种严重的信息丢失
  - 同理向y轴投影，也会出现信息丢失
  - **向通过第一象限和第三象限的斜线投影，数据较分散。**



## 2.3 方差：描述投影后投影值的分散程度

- 方差: $$\sigma^2 = \sum_i (x_i - \mu)^2 / n$$
- **PCA原理:**
  投影后投影值尽可能分散，
  而这种分散程度，可以用方差来描述
- **二维降一维，**问题转化:
  寻找一个一维基，
  所有数据变换为这个基上的坐标表示后，**方差值最大。**
- **高维降低维时**，需要添加其它的约束条件
  **协方差** 及协方差矩阵

# 2 .4 协方差矩阵：衡量特征间的独立性

- 样本特征间的协方差：　衡量任意 2 个特征间的互相独立性

$$\operatorname{cov}(x_i, x_j) = \frac{\sum_k (x_{ik} - \mu_\mathrm{i})(x_{jk} - \mu_j)}{n},$$

$\mu_\mathrm{i}$ : 第$i$个属性的均值

- 样本协方差矩阵：衡量n个特征间的互相独立性

$$C = [c_{ij}]_{n\times n} = [\operatorname{cov}(x_i, x_j)]_{n\times n}$$

**注：** 如果某2个维度间存在相关性，

说明从一个维度的值可以推测出另一个维度的值。

则该维度中有一个是多余的，可以把其中一个维度舍去。

- 三维降二维，问题转化：

  ①找第一个方向使投影后方差最大，尽量分散，不重叠

  ②找第二个方向(基)时，与第一个方向没有相关性

  “样本属性投影后尽可能互相独立”

  即，属性**投影后的协方差为 0**

  **③为了让协方差为0**

  **需要使第二个方向与第一个方向正交**

- 同理高维降低维……

- 将N维向量降为K维(K小于N )的**优化目标：**

  选择K个单位正交基，使原始数据投影到这组基后:

  **样本属性间协方差为0，方差尽可能大。**

**问题分析：**

例：$m$个中心化后样本(2个属性)数据的协方差矩阵

$$X=\begin{pmatrix} a_1 a_2 & \cdots & a_m \\ b_1 b_2 & & b_m \end{pmatrix}$$

$$\frac{1}{m}XX^T=\begin{pmatrix} \frac{\sum_i a_i^2}{m} & \frac{\sum_i a_i b_i}{m} \\ \frac{\sum_i a_i b_i}{m} & \frac{\sum_i b_i^2}{m} \end{pmatrix}$$

$C=\dfrac{XX^T}{m}$:样本属性的协方差矩阵

主对角线为方差，

非主对角线为协方差

$C = \frac{XX^T}{m}$:样本属性的协方差矩阵

主对角线为方差，

非主对角线为协方差

- 假设K个正交基构成的矩阵为 $P: K \times n$

  记投影后样本为 $Y = PX: K \times m$

  记Y的协方差矩阵为**D(依旧满足中心化)：对角矩阵**

(**满足协方差为0，方差尽可能大**)

$$D = \frac{YY^T}{m} = \frac{PXX^T P^T}{m} = PCP^T,$$ 根据特征值分解公式

**因此，P是能让原始协方差矩阵C对角化的一组基。**

# 进一步分析

- **P是能让原始协方差矩阵C对角化的一组基，且满足方差最大，协方差为0**

$$D=\frac{YY^T}{m}=\frac{PXX^TP^T}{m}=PCP^T\text{，根据特征值分解公式}$$

- 根据矩阵的迹的定义tr(A)：　矩阵对角线元素之和

$$tr(A)=\sum_i a_{ii}$$

- PCA: 求正交基，满足协方差矩阵的迹最大化

$$\boxed{\max_P \; tr(PCP'), \quad \text{s.t. } \mathrm{PP'}=\mathrm{I}}$$

$$\boxed{\max_{P}\ tr(PCP'),\quad \text{s.t. } \mathrm{PP}' = \mathrm{I}}$$

- 根据拉格朗日乘子法：求得拉格朗日函数

$$\boxed{L = tr(PCP') + \lambda(PP' - I)}$$

- 对拉格朗日函数关于 P 求偏导置 0

$$\boxed{\frac{\partial L}{\partial P} = CP' + \lambda P' = 0}$$

满足特征向量的关系式

- **因此**由C矩阵(协方差矩阵)的特征向量构成的矩阵，即是要求的变换矩阵 **P**

# PCA优化目标分析：

①寻找矩阵P，

满足 $PCP^T$ 是对角矩阵，且对角线元素从大到小依次排列；

②P的前K行组成的矩阵乘以X就使得X从N维降到了K维；

且满足(方差最大，协方差为０）优化条件。

协方差矩阵C是实对称矩阵，有一系列非常好的性质：

1）实对称矩阵不同特征值对应的特征向量必然正交。

2）存在正交矩阵 P ,使得 $P^{-1}CP = \Lambda$ ，且 $\Lambda$ 为对角矩阵，
主对角线元素为 C 的特征值， P 是 C 的特征向量构成的矩阵

# 2.6 PCA 求解目标

- **PCA问题：对样本进行降维，去掉冗余特征，保留主要成分**

  ① 降维后同一维度的方差最大

  ② 不同维度之间的相关性为0

- **转化后求解目标：**

  ①求投影矩阵 P：

  样本属性的协方差矩阵 C 的前K个最大的特征值对应的特征向量

  ②使用 P 对样本 X 进行降维

# PCA算法流程

- PCA算法流程

设有m个n维数据X

$$X=\{x_1,x_2,\cdots,x_m\}$$

**1.** 将原始数据按列组成n行m列矩阵

**2.**将X的每一行进行零均值化，即减去每一行的均值

$$\overline{X}=\left\{\begin{array}{ccc|cc} x_{11}-\mu_1 & x_{21}-\mu_1 & x_{31}-\mu_1 & \cdots & x_{m1}-\mu_1 \\ x_{12}-\mu_2 & x_{22}-\mu_2 & x_{32}-\mu_2 & \cdots & x_{m2}-\mu_2 \\ x_{13}-\mu_3 & x_{23}-\mu_3 & x_{33}-\mu_3 & \cdots & x_{m3}-\mu_3 \\ \hline \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{1n}-\mu_n & x_{2n}-\mu_n & x_{3n}-\mu_n & \cdots & x_{mn}-\mu_n \end{array}\right\}$$

**3.计算**协方差矩阵 $$\boxed{C=\frac{\overline{X}*\overline{X}^T}{m}}$$

**4.**求出协方差矩阵的前 K 个最大的特征值及对应的特征向量

**5.**将特征向量组成矩阵P

**6.**Y=PX即为降维到k维后的数据

# 2.7 PCA例子

- 对右边数据中心化处理 $\begin{pmatrix} 1 & 1 & 2 & 4 & 2 \\ 1 & 3 & 3 & 4 & 4 \end{pmatrix} \Longrightarrow \begin{pmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{pmatrix}$
- 计算其协方差矩阵C

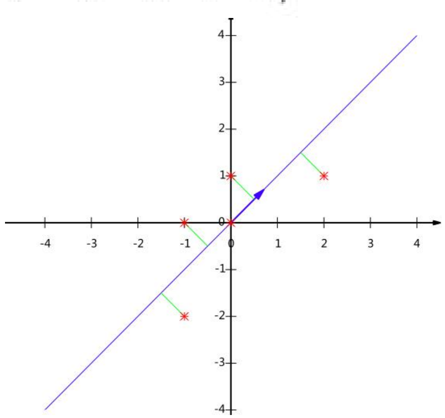
$$C = \frac{1}{5}\begin{pmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{pmatrix}\begin{pmatrix} -1 & -2 \\ -1 & 0 \\ 0 & 0 \\ 2 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \frac{6}{5} & \frac{4}{5} \\ \frac{4}{5} & \frac{6}{5} \end{pmatrix}$$

- 求C的特征值与特征向量(单位化) $\lambda_1 = 2, \lambda_2 = 2/5$ $\begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}, \begin{pmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$
- 投影矩阵P

$$P = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}$$

- 对数据进行降维

$$Y = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}\begin{pmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{pmatrix} = \begin{pmatrix} -3/\sqrt{2} & -1/\sqrt{2} & 0 & 3/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}$$



# PCA与SVD联系

- PCA求协方差矩阵C的最大的K个特征值对应的特征向量
- 协方差矩阵C，样本属性间的协方差

$$\boxed{C = \frac{XX^T}{m}}$$

- 当维度比较大时，直接计算协方差矩阵，计算量存储都很大
- SVD中的U,V分别为矩阵 $AA^T$与$A^T A$ 的特征向量
- **计算协方差矩阵C的特征向量，等价于求样本矩阵X的奇异矩阵**
- **直接省掉了计算协方差矩阵的计算量和存储**

# 2.8 最小二乘误差解释 PCA

- 普通二维向量的表示方式:

$$a=\begin{pmatrix}a_1\\a_2\end{pmatrix}=<a,e_1>e_1+<a,e_2>e_2$$
$$=(a^Te_1)e_1+(a^Te_2)e_2$$
$$=<a,e_1>\begin{pmatrix}1\\0\end{pmatrix}+<a,e_2>\begin{pmatrix}0\\1\end{pmatrix}$$

- n维向量 $x$ 可表示为

$$x=\sum_{i=1,\ n}<w_i,x>w_i$$

- PCA问题：给定n维样本 $x_i$ 求一组基函数

$$W=\{w_1,\cdots w_K\}(w_i\in R^n,n\text{维基向量，互相正交})$$

- $x_i$ 在该K个基函数的表示为

$$\hat{x}_i=\sum_{k=1,\ K}<w_k,x_i>w_k=\mathrm{W}\,\mathrm{W}^{\mathrm{T}}x_i$$

- 最小二乘误差 PCA，可视为求解以下优化问题：

$$\min_W \|x_i-\hat{x}_i\|^2,\quad \text{s.t. } \mathrm{W}^{\mathrm{T}}\mathrm{W}=\mathrm{I}$$

- 即等价于：

$$\min_{W} \| x_i - \sum_{k} < x_i, w_k > w_k \|_2^2, \qquad s.t. \ \ W^T W = I$$

- 对目标函数进行简化

$$\begin{aligned} \| x_i - WW^T x_i \|_2^2 &= -2x_i^T WW^T x_i \ + \ x_i^T WW^T WW^T x_i \\ &= -x_i^T WW^T x_i \\ &= -\|W^T x_i\|_2^2 \end{aligned}$$

- 原问题转化为

$$\max_{W} \| W^T x_i \|_2^2, \qquad \mathrm{s.t.}\, W^T W = I$$

- 将 $x_i$ 推广到m个样本集X,则最小二乘误差下的ＰＣＡ ：

$$\min_W \| X - \hat{X} \|^2 = \max_W tr(W^T X X^T W), \quad \text{s.t. } W^T W = I$$

- W即为所求的投影矩阵Ｐ：

  即由协方差矩阵的最大的Ｋ个特征向量组成

- **最新研究方向：L1范数下的ＰＣＡ问题：**

$$\max_W \| W^T X \|, \quad \text{s.t. } W^T W = I$$

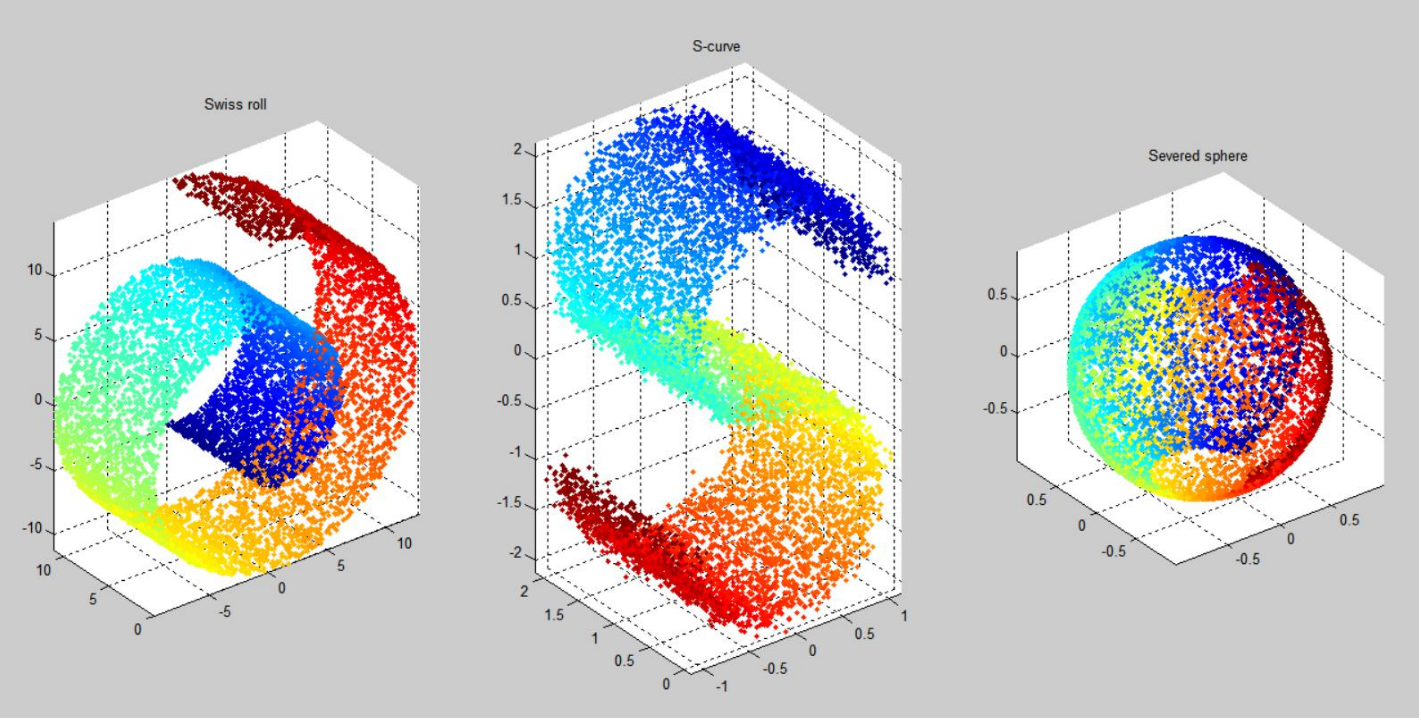
# 练习作业

- 请将以下 5 个4维样本降至 3 维

| | | | | |
|---|---|---|---|---|
| 0 | 4.9 | 3.0 | 1.4 | 0.2 |
| 1 | 4.7 | 3.2 | 1.3 | 0.2 |
| 2 | 4.6 | 3.1 | 1.5 | 0.2 |
| 3 | 5.0 | 3.6 | 1.4 | 0.2 |
| 4 | 5.4 | 3.9 | 1.7 | 0.4 |

# 程序作业

- 请使用PCA 算法将以下三个3维数据降至2维



# 程序作业

- 请使用PCA 算法将以下三个3维数据降至2维

```
% %% Swiss roll
% % Create the data
N = 10000;
% Noise term weighting (for Swiss roll and S-curve)
noise = 0.05;
% Size of points (for plotting)
sz = 10;
t = 3*pi/2 * (1 + 2*rand(N,1));
h = 11 * rand(N,1);
X = [t.*cos(t), h, t.*sin(t)] + noise*randn(N,3);

 figure('Position',[200,500,1000,1000],'WindowStyle','docked');
 subplot(1,3,1);
 scatter3(X(:,1),X(:,2),X(:,3),sz,t,'fill');
 axis('equal','tight');
 title('Swiss roll');
```

```
t = 3 * pi * (rand(N,1) - 0.5);
x = sin(t);
y = 2.0 * rand(N,1);
z = sign(t) .* (cos(t) - 1);
X = [x,y,z] + noise*randn(N,3);

subplot(1,3,2);
scatter3(X(:,1),X(:,2),X(:,3),sz,t,'fill');
axis('equal','tight');
title('S-curve');
```

```
p = rand(N,1) * (2 * pi - 0.55);
t = rand(N,1) * pi; % Sever the poles from the sphere.
indices = ((t < (pi - (pi / 8))) & (t > ((pi / 8))));
c = p(indices);
X = [sin(t(indices)) .* cos(p(indices)), sin(t(indices)) .* sin(p(indices)),cos(t(indices))];

subplot(1,3,3);
scatter3(X(:,1),X(:,2),X(:,3),sz,c,'fill');
axis('equal','tight');
title('Severed sphere');
```