



主成分分析

耿修瑞

中国科学院空间信息创新研究院

gengxr@sina.com.cn

2025.4

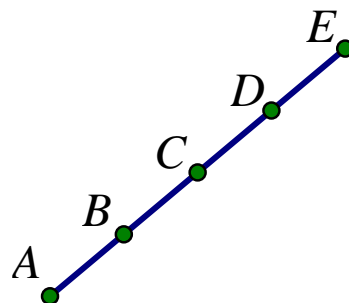
- 问题背景
- 基本概念
- 主成分变换

主成分分析首先是由卡尔·皮尔森在 1901 年引入的，但当时只针对非随机变量进行讨论。之后霍特林将此方法推广到随机向量的情形。↵

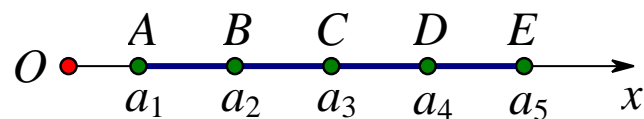
主成分分析的应用范围之广，可以从其在多个应用领域的各种不同命名窥出端倪。在信号处理领域，它通常称为离散 Karhunen-Loeve 变换(KLT)；在多元质量控制领域，它称之为 Hotelling 变换；在机械工程领域，它称之为本征正交分解 (POD)；在线性代数领域，它称之为奇异值分解 (SVD) 或者特征分解；在心理测量学领域，它称之为因子分析或 Eckart-Young theorem 或 Schmidt-Mirsky theorem；在气象科学领域，它称之为经验正交函数 (EOF)；在振动与噪声领域，它称之为经验特征函数分解或实证分析；在结构力学领域，它又称之为实验模态分析。↵



卡尔·皮尔逊(Karl Pearson, 1857~1936)，英国数学家，主成分分析创始人。↵

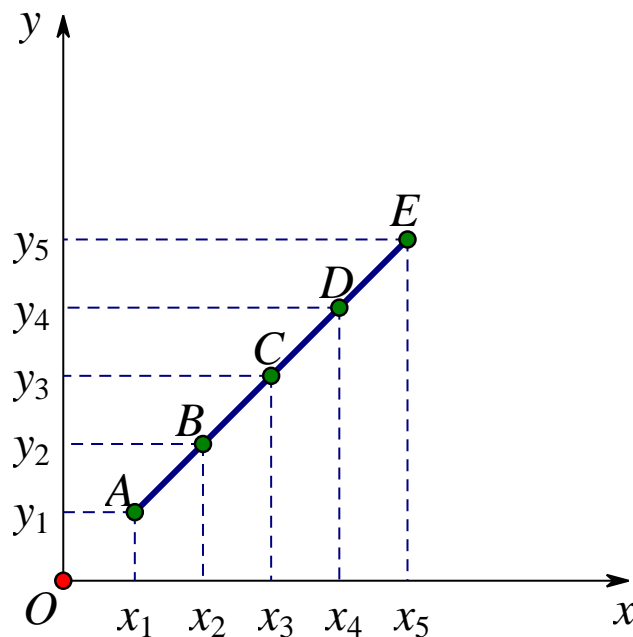


线段上的五个点，该如何对其进行定量描述



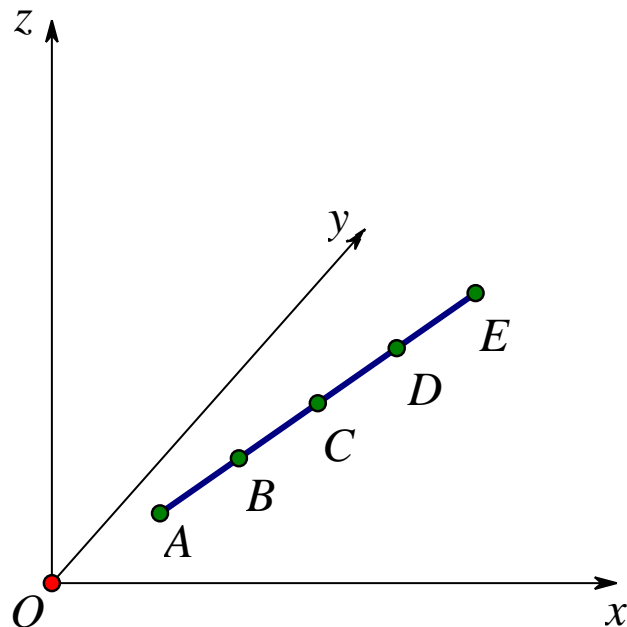
在实轴上，这五个点可以用它们的坐标定量描述，记为

$$\mathbf{x} = [a_1 \quad a_2 \quad a_3 \quad a_4 \quad a_5]$$



在平面直角坐标系中，这五个点可以描述如下

$$\mathbf{X} = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 \\ y_1 & y_2 & y_3 & y_4 & y_5 \end{bmatrix}$$



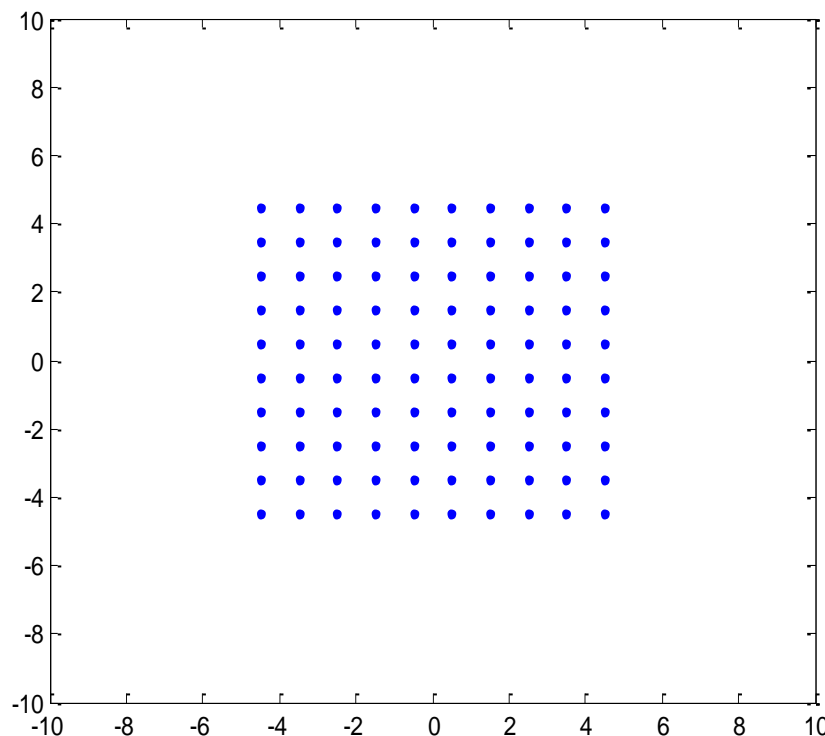
在三维空间直角坐标系中，这五个点可以描述如下

$$\mathbf{X} = \begin{bmatrix} \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \alpha_5 \\ \beta_1 & \beta_2 & \beta_3 & \beta_4 & \beta_5 \\ \gamma_1 & \gamma_2 & \gamma_3 & \gamma_4 & \gamma_5 \end{bmatrix}$$

上面几个例子表明，同一个描述对象，有不同的描述方式。

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & \cdots & x_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ x_{L1} & x_{L2} & \cdots & x_{LN} \end{bmatrix}$$

对于如上给定的数据，其描述的对象是什么呢？或者说，该数据的本征维度是多少呢？主成分分析就是回答这一问题的经典工具。



平面上均匀分布的散点，哪个方向信息量最大？

□ 样本均值: $Mean(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} \mathbf{x}^T \mathbf{1} = \bar{x}$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

□ 样本方差: $Var(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \|\tilde{\mathbf{x}}\|^2$

□ 样本标准差: $S(\mathbf{x}) = \sqrt{Var(\mathbf{x})}$

□ 样本协方差: $Cov(\mathbf{x}, \mathbf{y}) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n} \tilde{\mathbf{x}}^T \tilde{\mathbf{y}}$

□ 样本相关系数: $r(\mathbf{x}, \mathbf{y}) = \frac{Cov(\mathbf{x}, \mathbf{y})}{S(\mathbf{x})S(\mathbf{y})}$

$$\tilde{\mathbf{x}} = \begin{bmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{bmatrix}, \tilde{\mathbf{y}} = \begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{bmatrix}$$

■ 协方差矩阵: Σ

$$\Sigma = \begin{bmatrix} \text{COV}(\mathbf{x}_1, \mathbf{x}_1) & \text{COV}(\mathbf{x}_1, \mathbf{x}_2) & \cdots & \text{COV}(\mathbf{x}_1, \mathbf{x}_n) \\ \text{COV}(\mathbf{x}_2, \mathbf{x}_1) & \text{COV}(\mathbf{x}_2, \mathbf{x}_2) & \cdots & \text{COV}(\mathbf{x}_2, \mathbf{x}_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{COV}(\mathbf{x}_n, \mathbf{x}_1) & \text{COV}(\mathbf{x}_n, \mathbf{x}_2) & \cdots & \text{COV}(\mathbf{x}_n, \mathbf{x}_n) \end{bmatrix}$$

讨论两种协方差矩阵：波段协方差矩阵和像元协方差矩阵

■ 相关系数矩阵：C

$$C = \begin{bmatrix} 1 & \frac{\text{COV}(\mathbf{x}_1, \mathbf{x}_2)}{\sqrt{\text{COV}(\mathbf{x}_1, \mathbf{x}_1)}\sqrt{\text{COV}(\mathbf{x}_2, \mathbf{x}_2)}} & \dots & \frac{\text{COV}(\mathbf{x}_1, \mathbf{x}_n)}{\sqrt{\text{COV}(\mathbf{x}_1, \mathbf{x}_1)}\sqrt{\text{COV}(\mathbf{x}_n, \mathbf{x}_n)}} \\ \frac{\text{COV}(\mathbf{x}_2, \mathbf{x}_1)}{\sqrt{\text{COV}(\mathbf{x}_1, \mathbf{x}_1)}\sqrt{\text{COV}(\mathbf{x}_2, \mathbf{x}_2)}} & 1 & \dots & \frac{\text{COV}(\mathbf{x}_2, \mathbf{x}_n)}{\sqrt{\text{COV}(\mathbf{x}_2, \mathbf{x}_2)}\sqrt{\text{COV}(\mathbf{x}_n, \mathbf{x}_n)}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\text{COV}(\mathbf{x}_n, \mathbf{x}_1)}{\sqrt{\text{COV}(\mathbf{x}_1, \mathbf{x}_1)}\sqrt{\text{COV}(\mathbf{x}_n, \mathbf{x}_n)}} & \frac{\text{COV}(\mathbf{x}_n, \mathbf{x}_2)}{\sqrt{\text{COV}(\mathbf{x}_2, \mathbf{x}_2)}\sqrt{\text{COV}(\mathbf{x}_n, \mathbf{x}_n)}} & \dots & 1 \end{bmatrix}$$

使用相关系数矩阵可以克服量纲差异.

■数据中心化

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_L \end{bmatrix} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N], \quad \boldsymbol{\mu} = \frac{1}{N} \begin{bmatrix} X_1 \mathbf{1}_N \\ \vdots \\ X_L \mathbf{1}_N \end{bmatrix} = \frac{1}{N} \mathbf{X} \mathbf{1}_N = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i$$

中
心
化

$$\begin{aligned} \mathbf{X} &= [\mathbf{x}_1 - \boldsymbol{\mu}, \mathbf{x}_2 - \boldsymbol{\mu}, \dots, \mathbf{x}_N - \boldsymbol{\mu}] = \mathbf{X} - \boldsymbol{\mu} \mathbf{1}_N^T \\ &= \mathbf{X} - \frac{1}{N} \mathbf{X} \mathbf{1}_N \mathbf{1}_N^T = \mathbf{X} \left(\mathbf{I} - \frac{1}{N} \mathbf{1}_N \mathbf{1}_N^T \right) \end{aligned}$$

平移!



投影!

■ 协方差矩阵的计算

$$\begin{aligned}\mathbf{X} &= \begin{bmatrix} X_1 \\ \vdots \\ X_L \end{bmatrix} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N], \text{ 假设数据已经中心化, 即 } X_i \mathbf{1}_N = 0 \\ \Sigma &= \begin{bmatrix} \text{cov}(X_1, X_1) & \dots & \text{cov}(X_1, X_L) \\ \vdots & \ddots & \vdots \\ \text{cov}(X_L, X_1) & \dots & \text{cov}(X_L, X_L) \end{bmatrix} = \begin{bmatrix} \frac{1}{N} X_1 X_1^T & \dots & \frac{1}{N} X_1 X_L^T \\ \vdots & \ddots & \vdots \\ \frac{1}{N} X_L X_1^T & \dots & \frac{1}{N} X_L X_L^T \end{bmatrix} \\ &= \frac{1}{N} \begin{bmatrix} X_1 \\ \vdots \\ X_L \end{bmatrix} \begin{bmatrix} X_1^T, \dots, X_L^T \end{bmatrix} = \frac{1}{N} \mathbf{X} \mathbf{X}^T = \frac{1}{N} [\mathbf{x}_1, \dots, \mathbf{x}_N] \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_N^T \end{bmatrix} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i^T\end{aligned}$$

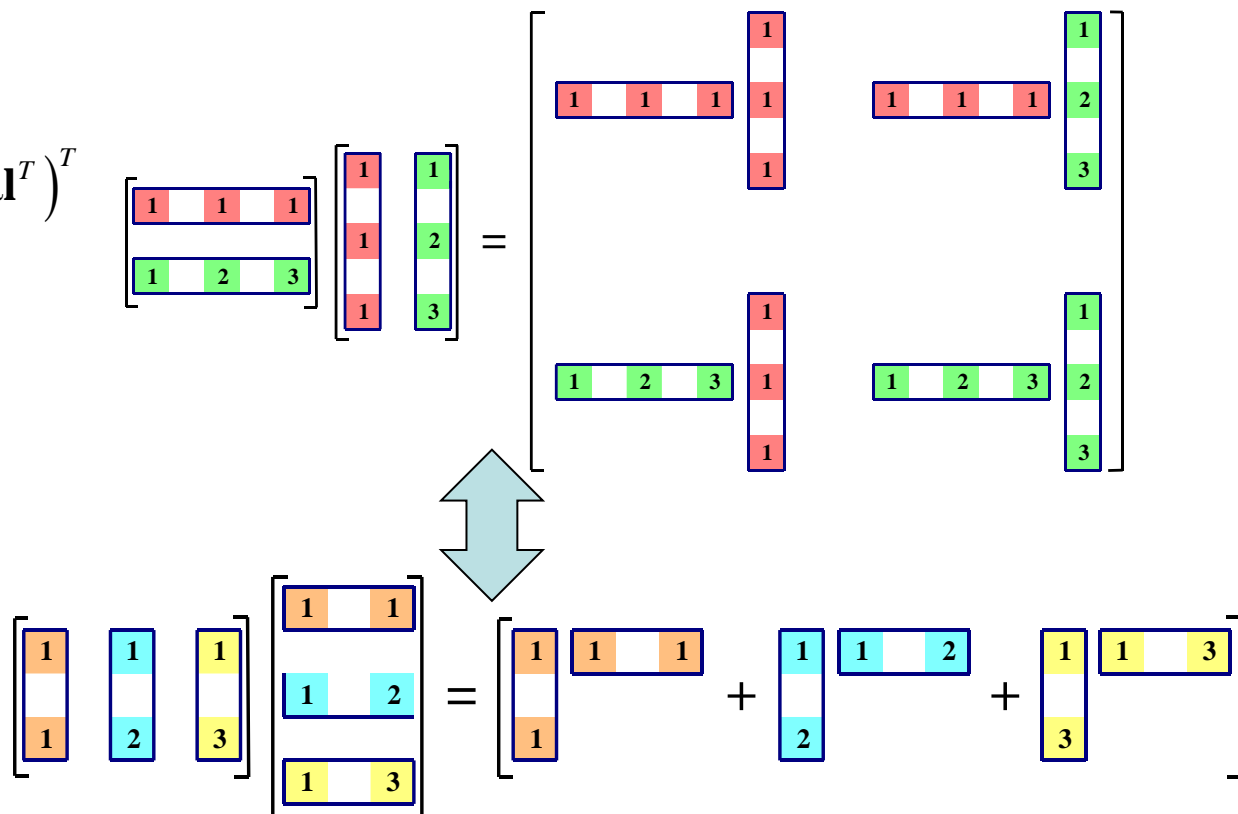
■ 协方差矩阵的计算及性质？

➤ 矩阵乘法

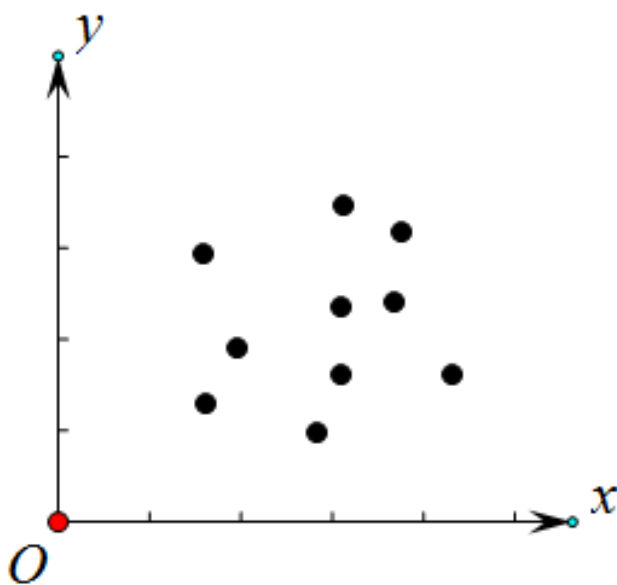
$$\begin{aligned}\Sigma &= \frac{1}{N}(\mathbf{X} - \boldsymbol{\mu}\mathbf{l}^T)(\mathbf{X} - \boldsymbol{\mu}\mathbf{l}^T)^T \\ &= \frac{1}{N}\mathbf{X}\mathbf{X}^T - \boldsymbol{\mu}\boldsymbol{\mu}^T\end{aligned}$$

➤ 像元法

$$\begin{aligned}\Sigma &= \frac{1}{N} \sum_{i=1}^N (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T \\ &= \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i^T - \boldsymbol{\mu}\boldsymbol{\mu}^T\end{aligned}$$



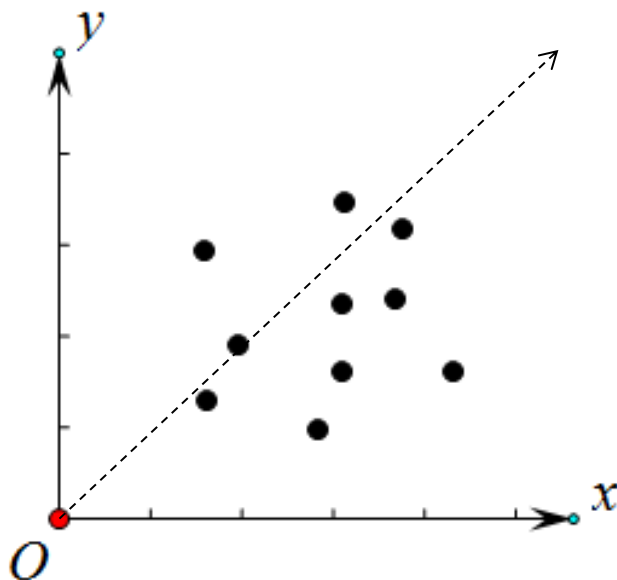
- 均值和方差分别描述的是单个随机变量的一阶统计特征和二阶统计特征（针对的是单特征数据）。
- 协方差和相关系数描述的是两个随机变量关系的二阶统计特征（针对的是多元数据）。



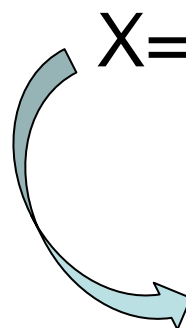
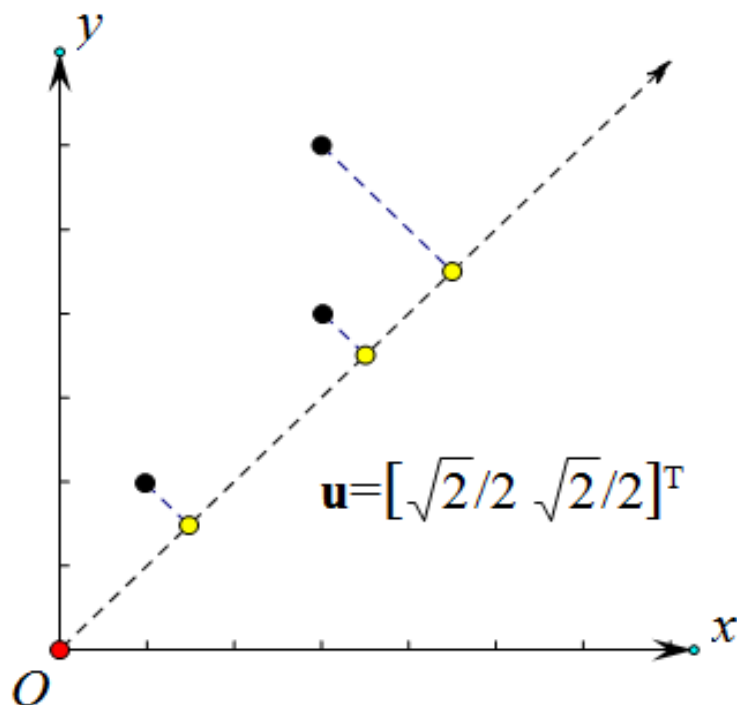
比如对于平面上分布的这些散点我们不能说这些数据的方差是多少！而只能说这些数据在某个方向投影之后的方差是多少！

一个灰度图像可以计算方差么？

■任意方向方差的计算？



- 以平面上三个点 $(1,2)$, $(3,4)$, $(3,6)$ 为例
计算此数据在任意方向的方差 (注意投影算子的
应用)


$$\mathbf{X} =$$

1	3	3
2	4	6

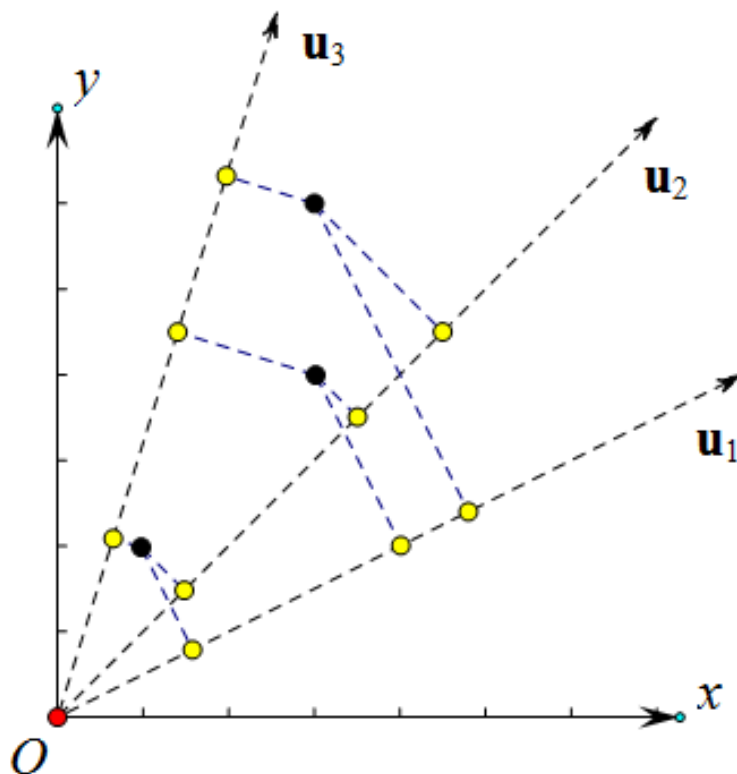
$$\mathbf{X} =$$

-1.33	0.67	0.67
-2	0	2

先投影，再计算
方差

$$\text{Var}(\mathbf{u}^T \mathbf{X})$$

- 以平面上三个点 $(1,2)$, $(3,4)$, $(3,6)$ 为例
计算此数据在任意方向的方差 (注意投影算子的
应用)



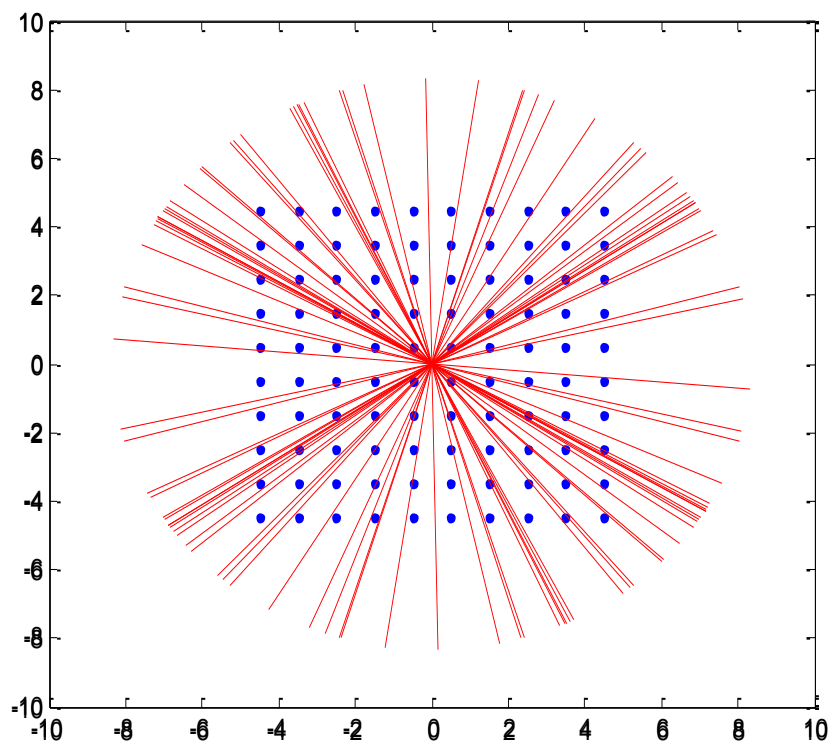
$$\text{Var}(\mathbf{u}_1^T \mathbf{X})$$

$$\text{Var}(\mathbf{u}_2^T \mathbf{X})$$

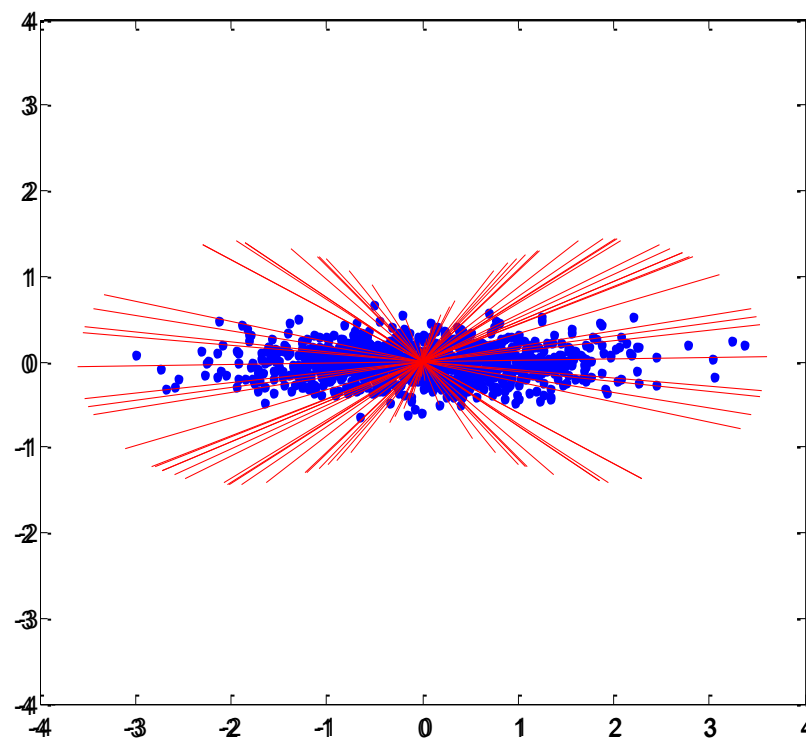
$$\text{Var}(\mathbf{u}_3^T \mathbf{X})$$

...

□ 估计一下该数据的方差分布情况？



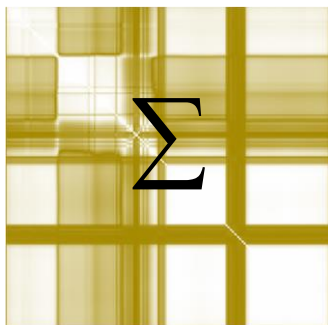
□ 估计一下该数据的方差分布情况？



如何得到方差最大的方向呢？

□ 数据任意方向方差的表达式

$$\text{Var}(\mathbf{u}^T \mathbf{X}) = \mathbf{u}^T \Sigma \mathbf{u}$$

$$\mathbf{u}^T \Sigma \mathbf{u} = \begin{bmatrix} u_1 & u_2 & \cdots & u_{L-1} & u_L \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{L-1} \\ u_L \end{bmatrix} \Sigma$$


根据上述公式，数据在任意方向的方差
均可以由数据的二阶统计量，协方差矩
阵显式表达出来！！！！

如何得到方差最大的方向呢？梯度下降（上升）法？

□考察目标：信息量（方差）

□模型：
$$\begin{cases} \max_{\mathbf{u}} \mathbf{u}^T \Sigma \mathbf{u} \\ \mathbf{u}^T \mathbf{u} = 1 \end{cases}$$

□模型的解（拉格朗日极值问题）：

$$\Sigma \mathbf{u} = \lambda \mathbf{u}$$

讨论：第2个以及后续的各个极值方差方向怎么求？

高斯分布的熵与方差之间关系

$$\begin{aligned} H[x] &= -\int \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \ln \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx \\ &= -\frac{1}{(2\pi\sigma^2)^{1/2}} \int \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \left(-\ln(\sqrt{2\pi}\sigma) - \frac{(x-\mu)^2}{2\sigma^2}\right) dx \\ &= -\frac{1}{(2\pi\sigma^2)^{1/2}} \cdot \ln(\sqrt{2\pi}\sigma) \int \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx + \frac{1}{(2\pi\sigma^2)^{1/2}} \int \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \frac{(x-\mu)^2}{2\sigma^2} dx \\ &= \frac{\ln(\sqrt{2\pi}\sigma)}{(2\pi\sigma^2)^{1/2}} \int \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx + \frac{1}{(2\pi\sigma^2)^{1/2}} \int \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \frac{(x-\mu)^2}{2\sigma^2} dx \\ &= \frac{\ln(\sqrt{2\pi}\sigma)}{(2\pi\sigma^2)^{1/2}} \sqrt{2\sigma} \int \exp\left\{-\left(\frac{x-\mu}{\sqrt{2\sigma}}\right)^2\right\} d\left(\frac{x-\mu}{\sqrt{2\sigma}}\right) + \frac{1}{(2\pi\sigma^2)^{1/2}} \sqrt{2\sigma} \int \exp\left\{-\left(\frac{x-\mu}{\sqrt{2\sigma}}\right)^2\right\} \frac{(x-\mu)^2}{2\sigma^2} d\left(\frac{x-\mu}{\sqrt{2\sigma}}\right) \\ &= \frac{\ln(\sqrt{2\pi}\sigma)}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-y^2} dy + \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-y^2} y^2 dy \\ &= \ln(\sqrt{2\pi}\sigma) + \frac{1}{\sqrt{\pi}} \cdot \frac{1}{2} \left(0 - \int_{-\infty}^{\infty} e^{-y^2} dy\right) \\ &= \ln(\sqrt{2\pi}\sigma) + \frac{1}{2} \\ &= \frac{1}{2} (\ln(2\pi\sigma^2) + 1) \end{aligned}$$

中国科学院空间信息处理与应用系统技术重点实验室

□ 主成分变换步骤：

输入： \mathbf{X}

1. 计算数据的均值向量且将数据中心化：

$$\boldsymbol{\mu} = \frac{1}{N} \mathbf{X} \mathbf{1} \quad \mathbf{X} = \mathbf{X} - \boldsymbol{\mu} \mathbf{1}^T$$

2. 计算协方差矩阵： $\boldsymbol{\Sigma} = \frac{1}{N} \mathbf{X} \mathbf{X}^T$

3. 计算协方差矩阵的特征值与特征向量： $\boldsymbol{\Sigma} \mathbf{U} = \mathbf{U} \boldsymbol{\Lambda}$

4. 主成分变换： $\mathbf{Y} = \mathbf{U}^T \mathbf{X}$

□ 例：求以下三个点的主成分方向。

(1,1), (2,1) 和 (3,3)

$$\mathbf{X} = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 1 & 3 \end{bmatrix}$$

1. 中心化:

$$\mathbf{X} = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 1 & 3 \end{bmatrix} \xrightarrow{\text{中心化}} \mathbf{X} = \begin{bmatrix} -1 & 0 & 1 \\ -2/3 & -2/3 & 4/3 \end{bmatrix}$$

2. 计算协方差矩阵: $\Sigma = \frac{1}{3} \mathbf{X} \mathbf{X}^T = \begin{bmatrix} 2/3 & 2/3 \\ 2/3 & 8/9 \end{bmatrix}$

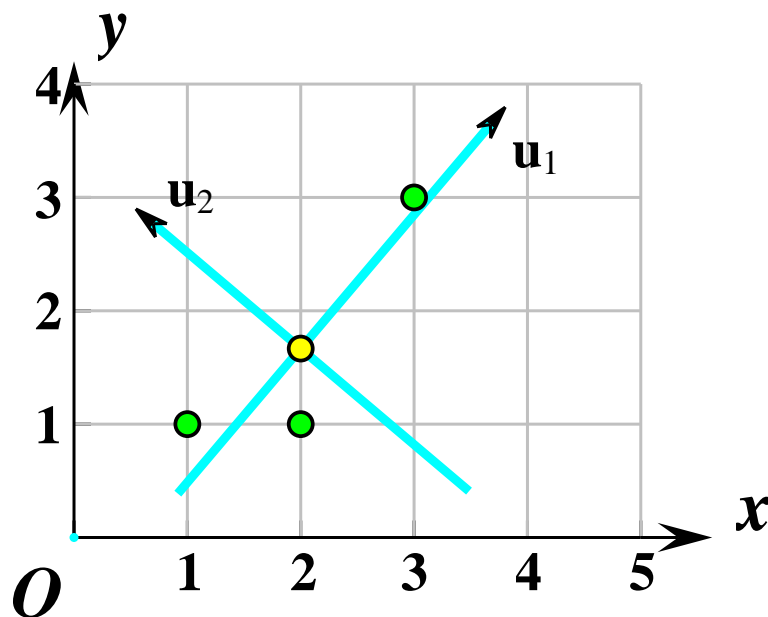
3. 计算协方差矩阵的特征值与特征向量:

$$\mathbf{D} = \begin{bmatrix} 0.1019 & 0 \\ 0 & 1.4536 \end{bmatrix} \quad \mathbf{U} = \begin{bmatrix} -0.7630 & 0.6464 \\ 0.6464 & 0.7630 \end{bmatrix}$$

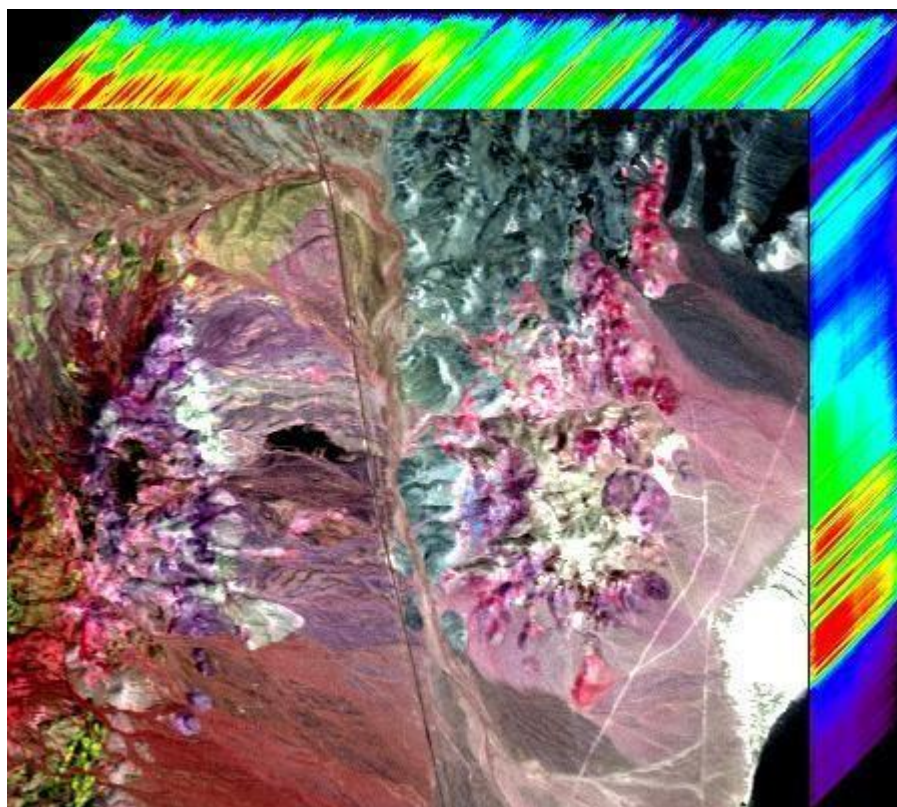
4. 主成分变换: $\mathbf{Y} = \mathbf{U}^T \mathbf{X} = \begin{bmatrix} 0.3321 & -0.4309 & 0.0988 \\ -1.1551 & -0.5087 & 1.6637 \end{bmatrix}$

□ 例：求以下三个点的主成分方向。

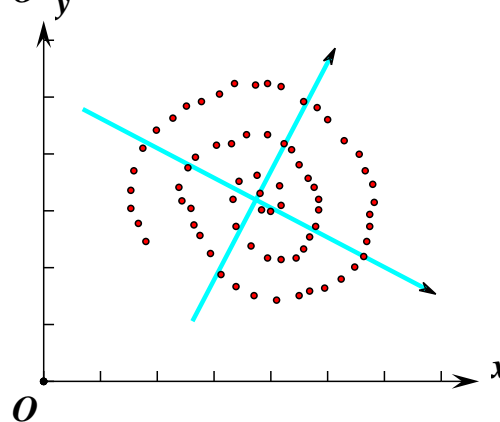
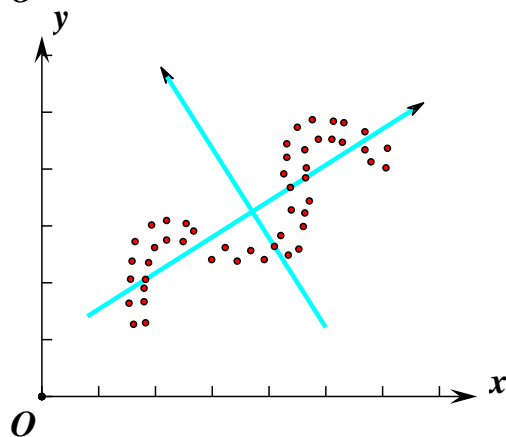
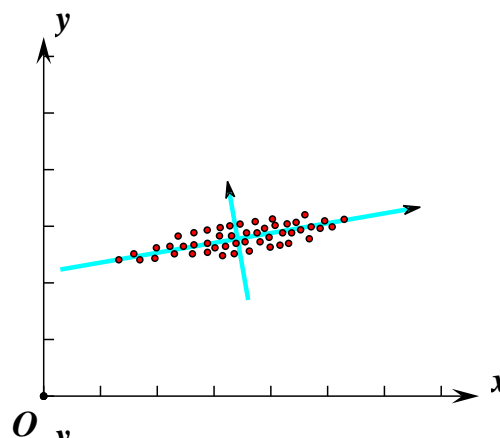
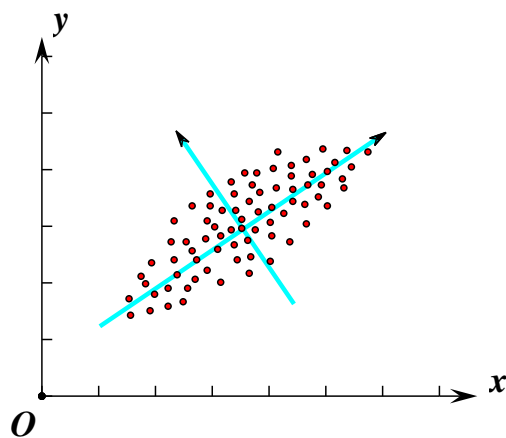
$(1,1)$, $(2,1)$ 和 $(3,3)$



□ 图像实例

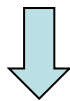


□ 一个现象：无论我们如何改变数据，任意两个主成分方向必然垂直？



□ 重新考察主成分变换的目标函数（方差）

$$\text{var}(\mathbf{u}^T \mathbf{X}) = \boxed{\mathbf{u}^T \Sigma \mathbf{u}} = \frac{1}{N} \mathbf{u}^T \mathbf{X} \mathbf{X}^T \mathbf{u} = \boxed{\frac{1}{N} \|\mathbf{u}^T \mathbf{X}\|^2}$$



这是 L 维特征空间的一个统计概念



这是 N 维样本空间的一个几何概念

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1,N-1} & x_{1N} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{L1} & x_{L2} & \cdots & x_{L,N-1} & x_{LN} \end{bmatrix}$$

关键在于如何看待 $\mathbf{X}^T \mathbf{u}$ 或 $\mathbf{u}^T \mathbf{X}$

□ 如何理解 $\mathbf{u}^T \mathbf{X}$

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1,N-1} & x_{1N} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{L1} & x_{L2} & \cdots & x_{L,N-1} & x_{LN} \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_L \end{bmatrix} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N]$$

$$\mathbf{u}^T \mathbf{X} = u_1 X_1 + u_2 X_2 + \cdots + u_L X_L = [\mathbf{u}^T \mathbf{x}_1, \mathbf{u}^T \mathbf{x}_2, \cdots, \mathbf{u}^T \mathbf{x}_N]$$

N 维空间的线性组合
: N 维空间的一个点

L 维空间在 \mathbf{u} 方向
的投影: 1维空间
上 N 个点

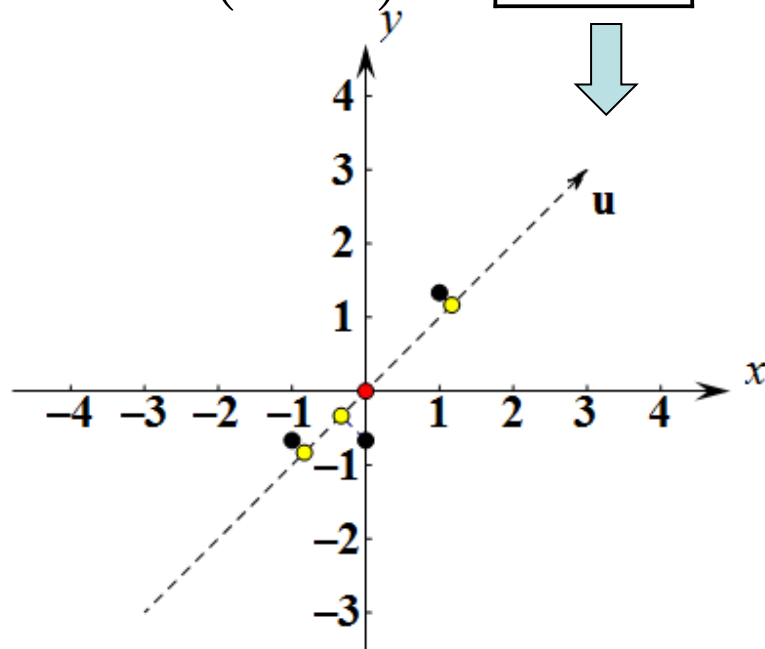
主成分变换



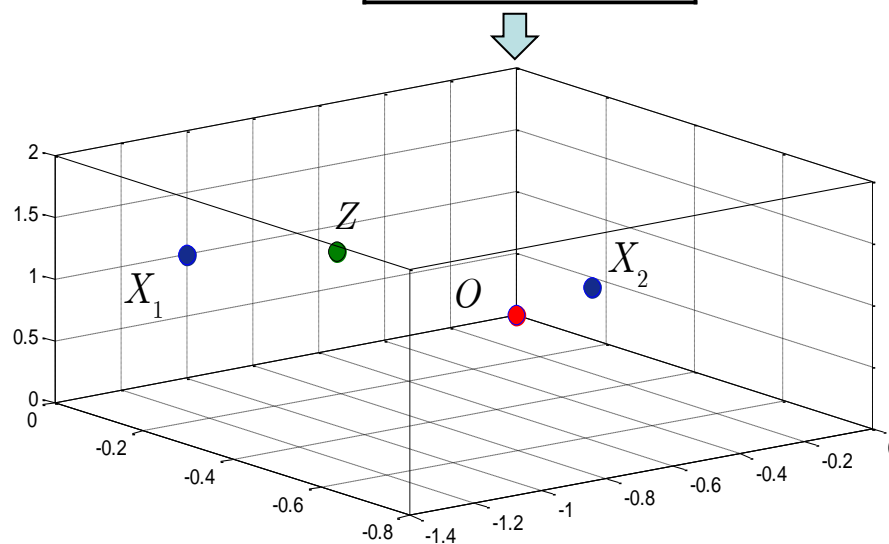
重新考察主成分变换的目标函数（方差）

$$\mathbf{X} = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 1 & 3 \end{bmatrix} \xrightarrow{\text{中心化}} \mathbf{X} = \begin{bmatrix} -1 & 0 & 1 \\ -2/3 & -2/3 & 4/3 \end{bmatrix}$$

$$\text{var}(\mathbf{u}^T \mathbf{X}) = \boxed{\mathbf{u}^T \Sigma \mathbf{u}} = \frac{1}{N} \mathbf{u}^T \mathbf{X} \mathbf{X}^T \mathbf{u} = \boxed{\frac{1}{N} \|\mathbf{u}^T \mathbf{X}\|^2}$$



$$\mathbf{X} = \begin{bmatrix} -1 & 0 & 1 \\ -2/3 & -2/3 & 4/3 \end{bmatrix} = [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3]$$



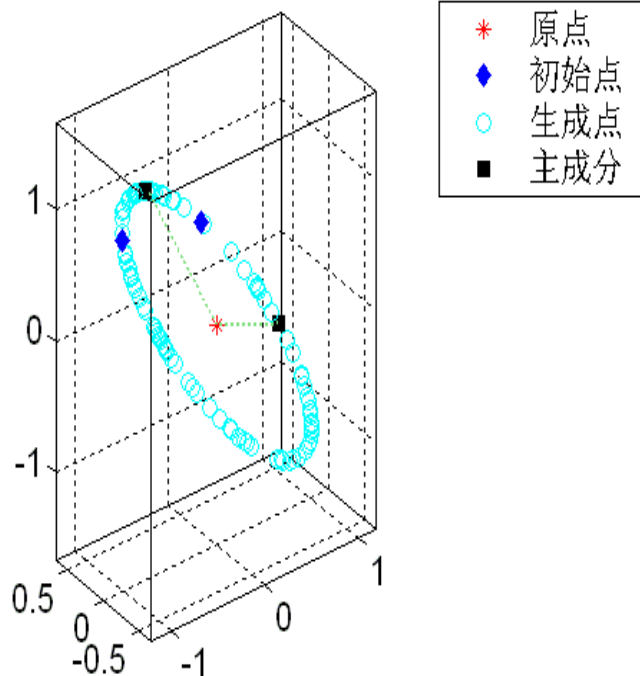
$$\mathbf{X} = \begin{bmatrix} -1 & 0 & 1 \\ -2/3 & -2/3 & 4/3 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \quad Z = \mathbf{u}^T \mathbf{X} = u_1 X_1 + u_2 X_2$$

□ 讨论 $\mathbf{X} = \begin{bmatrix} -1 & 0 & 1 \\ -2/3 & -2/3 & 4/3 \end{bmatrix}$

$$Z = \mathbf{u}^T \mathbf{X} = u_1 X_1 + u_2 X_2$$

随机选多个不同的方向 \mathbf{u} ,将得到多个不同的 \mathbf{Z} ,
所有的这些 \mathbf{Z} 将组成一个什么样的集合?

□ PCA几何解释



$$Y = \mathbf{u}^T \mathbf{X}$$

$$\mathbf{y} = Y^T$$

$$\mathbf{u} = (\mathbf{X}\mathbf{X}^T)^{-1} \mathbf{X}\mathbf{y}$$

$$\mathbf{u}^T \mathbf{u} = 1$$

$$\mathbf{y}^T \mathbf{X}^T (\mathbf{X}\mathbf{X}^T)^{-1} (\mathbf{X}\mathbf{X}^T)^{-1} \mathbf{X}\mathbf{y} = 1$$

$$f(\mathbf{y}) = \mathbf{y}^T \mathbf{X}^T (\mathbf{X}\mathbf{X}^T)^{-1} (\mathbf{X}\mathbf{X}^T)^{-1} \mathbf{X}\mathbf{y} - 1 = 0$$

$$\mathbf{H} = \mathbf{X}^T (\mathbf{X}\mathbf{X}^T)^{-1} (\mathbf{X}\mathbf{X}^T)^{-1} \mathbf{X}$$

- 子空间逼近解释：假设观测数据 \mathbf{X} 具有 p 个特征 n 个观测。且假设其均值向量为零向量。PCA的目的就是希望找到一个能够表征数据所有信息的低维子空间。假设 $\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_s$ 为该空间的一组标准正交基，记

$$\mathbf{Q} = [\mathbf{q}_1 \quad \mathbf{q}_2 \quad \dots \quad \mathbf{q}_s]$$

则可以构建该矩阵的投影矩阵

$$\mathbf{P}_Q = \mathbf{Q}\mathbf{Q}^\# = \mathbf{Q}(\mathbf{Q}^\mathrm{T}\mathbf{Q})^{-1}\mathbf{Q}^\mathrm{T} = \mathbf{Q}\mathbf{Q}^\mathrm{T}$$

将 \mathbf{X} 投影到 \mathbf{Q} 的列空间有

$$\mathbf{Y} = \mathbf{P}_Q \mathbf{X} = \mathbf{Q} \mathbf{Q}^T \mathbf{X}$$

如果 \mathbf{Q} 的列空间是 \mathbf{X} 的最佳逼近子空间，则必然满足 $\mathbf{Y} = \mathbf{Q} \mathbf{Q}^T \mathbf{X}$ 与 \mathbf{X} 尽量接近，因此有如下优化模型

$$\begin{cases} \min_{\mathbf{Q}} \left\| (\mathbf{I}_p - \mathbf{Q} \mathbf{Q}^T) \mathbf{X} \right\|_F^2 \\ s.t. \mathbf{Q}^T \mathbf{Q} = \mathbf{I}_s \end{cases}$$

又由于

$$\begin{aligned}\|(\mathbf{I}_p - \mathbf{Q}\mathbf{Q}^T)\mathbf{X}\|_F^2 &= \text{trace}(\mathbf{X}^T (\mathbf{I}_p - \mathbf{Q}\mathbf{Q}^T)(\mathbf{I}_p - \mathbf{Q}\mathbf{Q}^T)\mathbf{X}) \\ &= \text{trace}(\mathbf{X}^T (\mathbf{I}_p - \mathbf{Q}\mathbf{Q}^T)\mathbf{X}) = \text{trace}(\mathbf{X}^T \mathbf{X}) - \text{trace}(\mathbf{Q}^T \mathbf{X}\mathbf{X}^T \mathbf{Q})\end{aligned}$$

因此上述优化模型可以转化为

$$\begin{cases} \min_{\mathbf{Q}} \text{trace}(\mathbf{Q}^T \mathbf{X}\mathbf{X}^T \mathbf{Q}) \\ s.t. \mathbf{Q}^T \mathbf{Q} = \mathbf{I}_s \end{cases}$$

该模型的解归结为 $\mathbf{X}\mathbf{X}^T$ 的特征值与特征向量问题（思考？）

$$\begin{aligned} \text{trace}(\mathbf{Q}^T \mathbf{X} \mathbf{X}^T \mathbf{Q}) &= \text{trace}(\mathbf{Q}^T \mathbf{U} \mathbf{\Lambda} \mathbf{U}^T \mathbf{Q}) \\ &= \text{trace}(\mathbf{U}^T \mathbf{Q} \mathbf{Q}^T \mathbf{U} \mathbf{\Lambda}) = \text{trace}(\mathbf{Z} \mathbf{\Lambda}) = \sum_{i=1}^p \lambda_i z_{ii} \end{aligned}$$

其中：

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0 \quad \mathbf{Z} = \mathbf{U}^T \mathbf{Q} \mathbf{Q}^T \mathbf{U}$$

显然有：

$$z_{ii} \leq 1$$

又由于：

$$\sum_{i=1}^p z_{ii} = \text{trace}(\mathbf{Z}) = \text{trace}(\mathbf{U}^T \mathbf{Q} \mathbf{Q}^T \mathbf{U}) = \text{trace}(\mathbf{Q}^T \mathbf{U} \mathbf{U}^T \mathbf{Q}) = s$$

$$\begin{aligned} \mathbf{Z} &= \mathbf{U}^T \mathbf{Q} \mathbf{Q}^T \mathbf{U} \\ &= \mathbf{V} \mathbf{V}^T \quad (\text{其中 } \mathbf{V}^T \mathbf{V} = \mathbf{I}_{s \times s}) \end{aligned}$$

则 z_{ii} 即为 \mathbf{V} 的第 i 行行向量的2范数的平方，显然

$$z_{ii} \leq 1$$

因此为了使得 $trace(\mathbf{Q}^T \mathbf{X} \mathbf{X}^T \mathbf{Q})$ 达到最大值，必须有：

$$\begin{aligned} z_{11} = z_{22} = \cdots = z_{ss} &= 1 \\ z_{s+1,s+1} = z_{s+2,s+2} = \cdots = z_{pp} &= 0 \end{aligned} \quad (*)$$

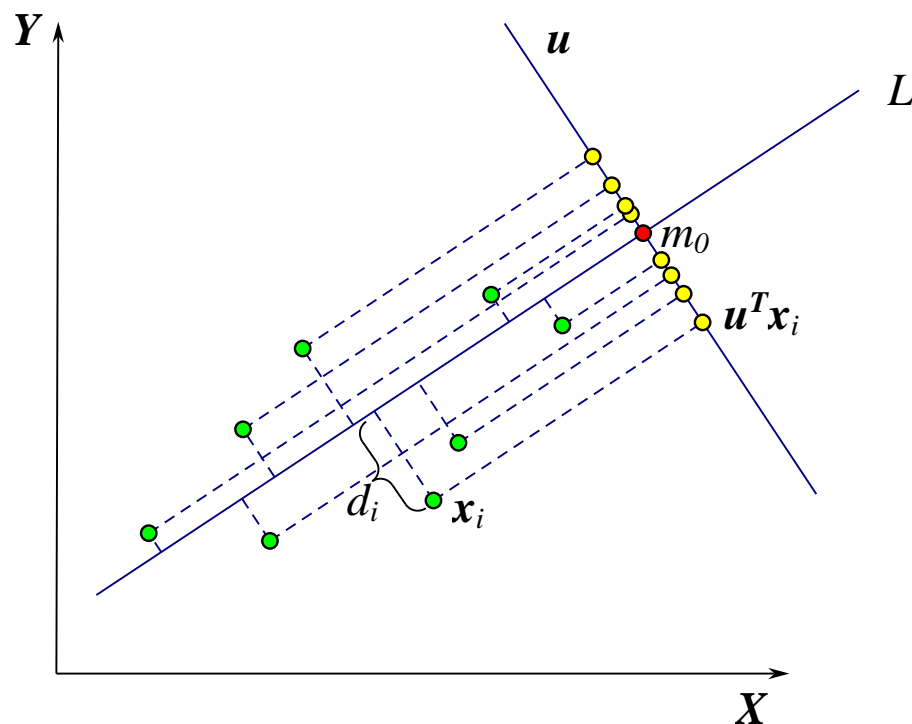
此时， $trace(\mathbf{Q}^T \mathbf{X} \mathbf{X}^T \mathbf{Q})$ 达到最大，且为：

$$trace(\mathbf{Q}^T \mathbf{X} \mathbf{X}^T \mathbf{Q}) = \sum_{i=1}^s \lambda_i$$

注意到，当 \mathbf{Q} 为特征向量矩阵 \mathbf{U} 的前 s 个特征向量构成的矩阵时，即 $\mathbf{Q} = \mathbf{U}(:, 1:s)$ 时，上面 $(*)$ 式成立。

子空间逼近与总体最小二乘

$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$



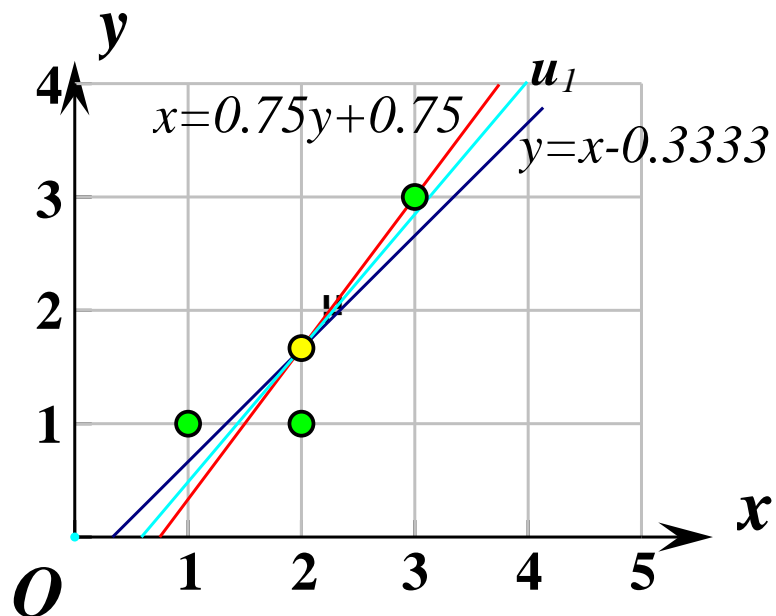
$$f(\mathbf{u}, m_0) = \frac{1}{n} \sum_{i=1}^n d_i^2 = \frac{1}{n} \sum_{i=1}^n (\mathbf{u}^T \mathbf{x}_i - m_0)^2$$

$$\frac{\partial f(\mathbf{u}, m_0)}{\partial m_0} = 2nm_0 - 2 \sum_{i=1}^n \mathbf{u}^T \mathbf{x}_i \stackrel{\text{令}}{=} 0$$

$$m_0 = \frac{1}{n} \sum_{i=1}^n \mathbf{u}^T \mathbf{x}_i$$

$$\begin{aligned} f(\mathbf{u}, m_0) &= \frac{1}{n} \sum_{i=1}^n \left(\mathbf{u}^T \mathbf{x}_i - \frac{1}{n} \sum_{i=1}^n \mathbf{u}^T \mathbf{x}_i \right)^2 \\ &= \frac{1}{n} \mathbf{u}^T \left(\sum_{i=1}^n \left(\mathbf{x}_i - \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \right) \left(\mathbf{x}_i - \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \right)^T \right) \mathbf{u} \\ &= \mathbf{u}^T \Sigma \mathbf{u} \end{aligned}$$

□ PCA与最小二乘法



□ PCA的概率解释

在前面的讨论中，PCA可以等价于下面目标函数的极值问题

$$\min_{\mathbf{Q}} \left\| (\mathbf{I}_p - \mathbf{Q}\mathbf{Q}^T) \mathbf{X} \right\|_F^2$$

值得注意的是，这里面用的是F范数，为什么呢？

$$\mathbf{Q} = \begin{bmatrix} q_{11} & \cdots & q_{1s} \\ \vdots & \ddots & \vdots \\ q_{p1} & \cdots & q_{ps} \end{bmatrix}$$

□ PCA的概率解释

令 $(\mathbf{I}_p - \mathbf{Q}\mathbf{Q}^T)\mathbf{X} = \mathbf{E}$, 假设 e_{ij} 独立同高斯分布, 即

$P(e_{ij}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{e_{ij}^2}{2\sigma^2}}$, 则子空间逼近的误差的联合概率密度函数为

$P(\mathbf{E}|\mathbf{Q}) = \prod_{i,j} P(e_{ij})$, 因此, 定义似然函数为

$L(\mathbf{Q}) = P(\mathbf{E}|\mathbf{Q}) = \prod_{i,j} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{e_{ij}^2}{2\sigma^2}\right)$, 于是

$l(\mathbf{Q}) = \ln L(\mathbf{Q}) = \sum_{i,j} \ln \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{i,j} e_{ij}^2$

$l(\mathbf{Q})$ 的最大化等价于 $\sum_{i,j} e_{ij}^2$ 的最小化

而 $\sum_{i,j} e_{ij}^2 = \|(\mathbf{I}_p - \mathbf{Q}\mathbf{Q}^T)\mathbf{X}\|_F^2$

□ 讨论

1. 两个协方差矩阵：样本，波段，实际中用哪个？
2. PCA得到的各个主成分相互正交么？
3. 协方差矩阵的特征值和特征向量的含义？
4. 协方差矩阵的特征向量矩阵的转置的特征值和特征向量的含义？
5. PCA的结果和原始数据之间的关系？
6. PCA的各个特征向量对应优化模型的什么类型的临界点（驻点）？
7. PCA的适用情况？

□ 小结

1. 当观测数据服从高斯分布时，主成分分析是最佳的降维或者特征提取手段。
2. 数据的协方差矩阵包含了数据的所有二阶统计信息，数据在任意方向的方差都可以由协方差矩阵和表征相应方向的单位向量解析表达。
3. 数据的主成分分析可以转化为数据协方差矩阵的特征值与特征向量分析。
4. 在几何上，任意的观测数据都对应一个样本空间的超椭球面，该超椭球面的各个长短轴点对应数据的各个主成分。
5. 从子空间逼近角度，主成分分析等价于总体最小二乘法，它可以认为是用一个低维超平面拟合给定散点。
6. 从概率角度，主成分分析要求数据服从高斯分布。且当数据服从高斯分布时，数据方差的大小等价于信息熵的大小。

□ 思考题1（1分）

超高维数据（数据维数远大于样本数）的降维该如何处理？

□ 思考题2（1分）

当数据在各个波段的噪声水平显著不同时，如何改进PCA，使其可以用于该数据的降维。



谢 谢

耿修瑞

中国科学院空间信息创新研究院

gengxr@sina.com.cn