第二讲

主成份分析

Principal Component Analysis (PCA)

彭志科

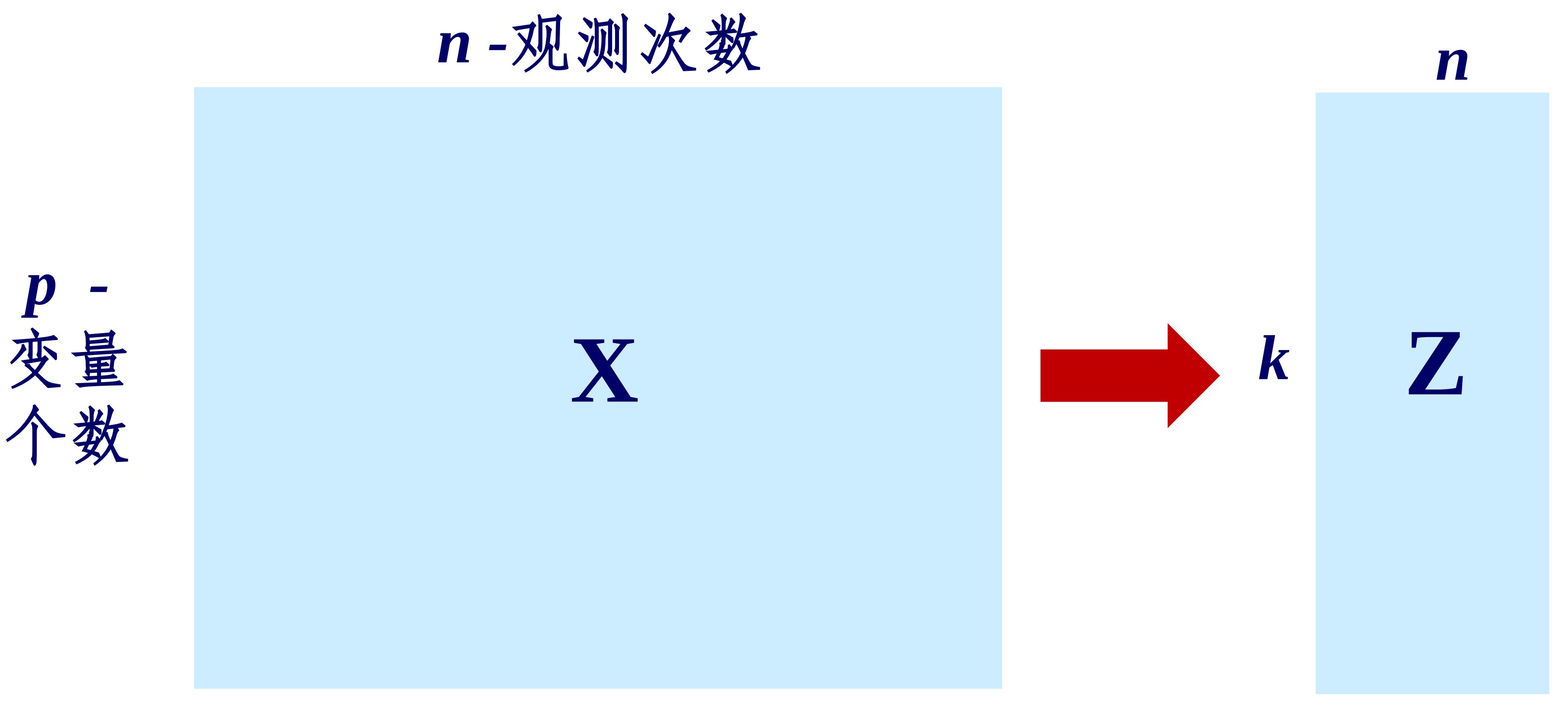
Email: z.peng@sjtu.edu.cn

上海交通大学

机械系统与振动国家重点实验室

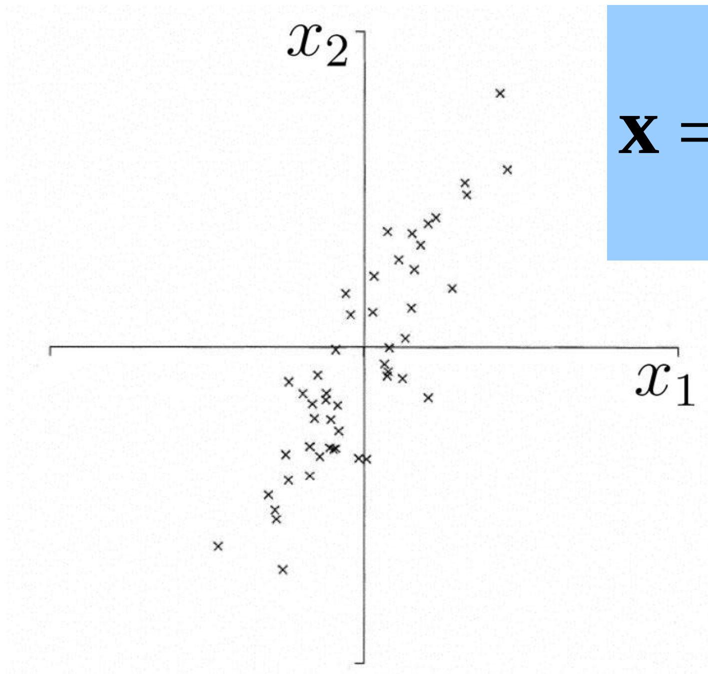
基本概念

- **PCA的基本目的**是对高维观测数据进行线性变换以构造出一组低维数据，在尽量不损失观测信息的同时，达到数据压缩的目的。



基本概念

示例



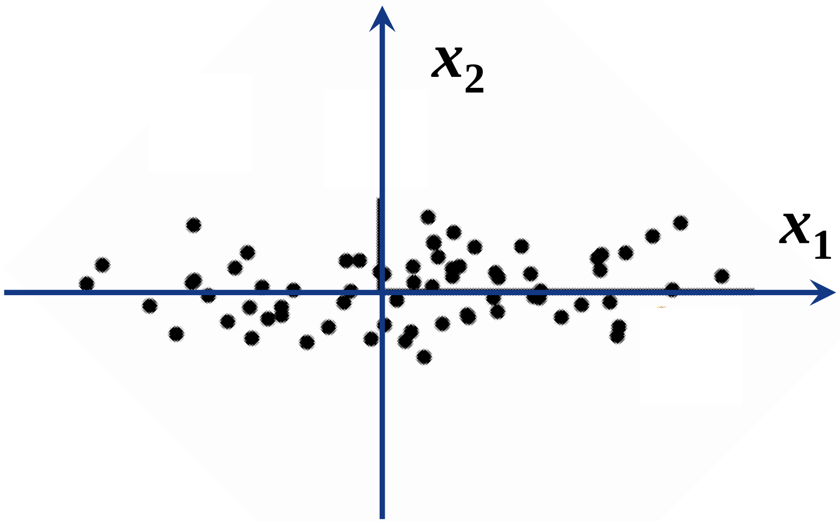
2D空间的***n***次观测

$$\mathbf{x} = \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ x_{21} & \cdots & x_{2n} \end{pmatrix}$$

问题： 能否从x_1和x_2的两个量的n次观测数据构造出一个变量，它可以最大的保留原始观测信息?

基本概念

信息量



预处理： $\bar{x} = 0$

$$\mathrm{var}(x) = \frac{1}{n}\sum_{k=1}^{n} x_k^2$$

$$\mathrm{var}(x) = \frac{1}{n}\langle x_k, x_k \rangle$$

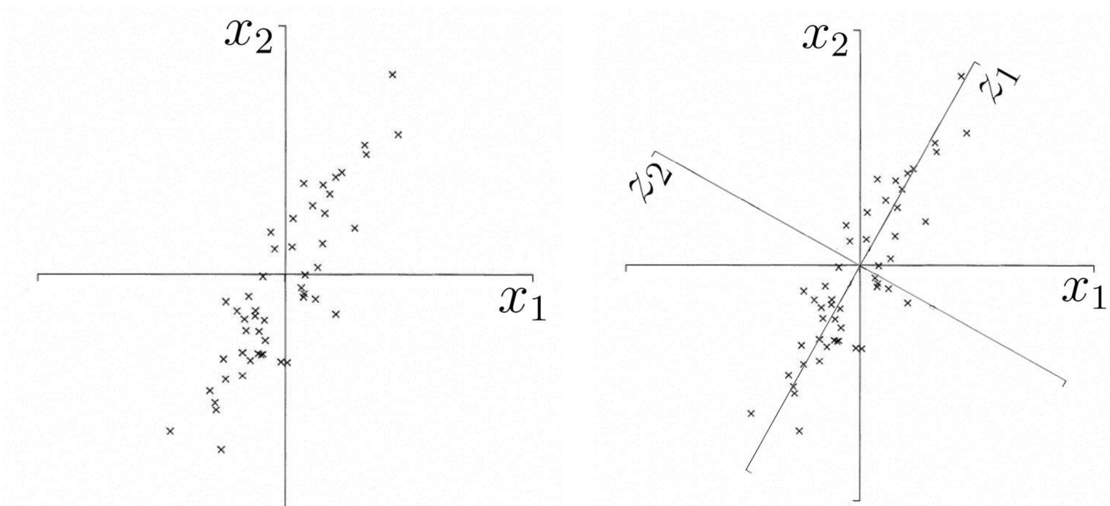
问题： 两个观测量x_1和x_2，谁携带的信息大?

答案： x_1， 因为 $\mathbf{var}(x_1) > \mathbf{var}(x_2)$

另外，$SNR = \dfrac{\mathrm{var}(signal)}{\mathrm{var}(noise)}$ ， $SNR(x_1) > SNR(x_2)$

基本概念

- **续问题：** 能否从x_1和x_2的两个量的n次观测数据构造出一个变量，它可以最大的保留原始观测信息？



答案：

$$z_1 = c_{11}x_1 + c_{12}x_2; (c_{11}, c_{12}) = \arg\max_{(\lambda_1,\lambda_2)} \left(\mathrm{var}(\lambda_1 x_1 + \lambda_2 x_2)\right)$$

z_1：第1个主成份；z_2：第2个主成份，垂直于z_1

主成份： 是一系列对观测数据的最小二乘线性拟合，且相互正交。

基本概念

高维情况

$$\begin{pmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_p \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \mathbf{X} & \vdots \\ x_{p1} & \cdots & x_{pn} \end{pmatrix} \qquad \begin{pmatrix} \boldsymbol{c}_1 \\ \vdots \\ \boldsymbol{c}_p \end{pmatrix} = \begin{pmatrix} c_{11} & \cdots & c_{1p} \\ \vdots & \mathbf{C} & \vdots \\ c_{p1} & \cdots & c_{pp} \end{pmatrix} \longrightarrow \begin{pmatrix} \boldsymbol{z}_1 \\ \vdots \\ \boldsymbol{z}_p \end{pmatrix} = \begin{pmatrix} z_{11} & \cdots & z_{1n} \\ \vdots & \mathbf{Z} & \vdots \\ z_{p1} & \cdots & z_{pn} \end{pmatrix}$$

$$\mathbf{Z} = \mathbf{C}\mathbf{X}$$

第1个主成份： $\boldsymbol{z}_1 = \boldsymbol{c}_1 \mathbf{X}$

其中 $(c_{11} \cdots c_{1p})$ 的选择，是要使 $\mathrm{var}(\boldsymbol{z}_1)$ 最大。

并且 $\langle \boldsymbol{c}_1, \boldsymbol{c}_1 \rangle = 1$

基本概念

高维情况

第k个主成份： $\boxed{z_k = c_k X}$

其中 $(c_{k1} \cdots c_{kp})$ 的选择，是要使在满足下列条件的情况下 $\mathrm{var}(z_k)$ 最大。

1) $\mathrm{cov}(z_k, z_l) = 0, \quad \text{for } k > l > 1$

$\rightarrow \frac{1}{n}\langle z_k, z_l \rangle = 0$

2) $\langle c_k, c_k \rangle = 1$

$\langle c_k, c_l \rangle = \delta(k-l)$

$\{c_i, i = 1, \ldots, p\}$

单位正交基

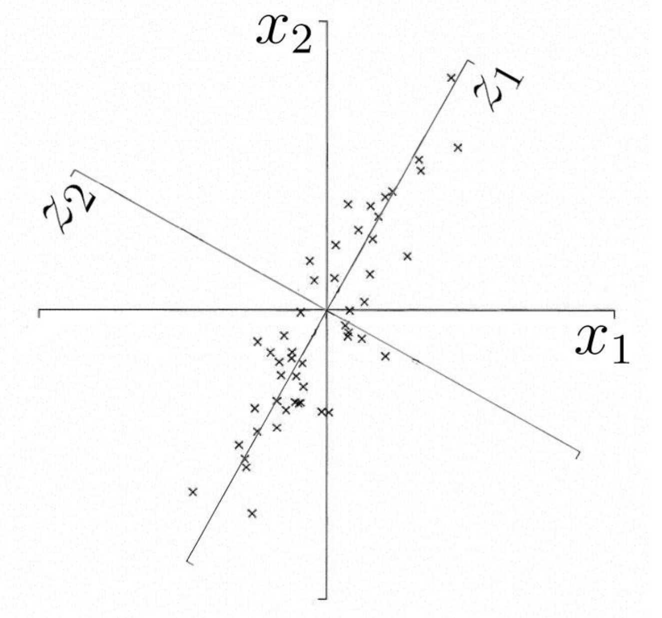
基本概念

$$\mathbf{Z} = \mathbf{C}\mathbf{X}$$

$$\{\boldsymbol{c}_i, i = 1, \ldots, p\}$$

单位正交基

PCA的实质是将高维观测数据投影到一组新的正交坐标基，使得观测数据的信息主要集中在少数几个坐标上。



算法 - I

如何确定变换矩阵 C $\mathbf{Z}=\mathbf{CX}$

$$\operatorname{cov}(\mathbf{Z})=\frac{1}{n}\mathbf{Z}\mathbf{Z}^T=\frac{1}{n}(\mathbf{CX})(\mathbf{CX})^T$$

$$=\frac{1}{n}\mathbf{C}\mathbf{X}\mathbf{X}^T\mathbf{C}^T$$

$$=\mathbf{C}\operatorname{cov}(\mathbf{X})\mathbf{C}^T$$

因为 $\operatorname{cov}(\mathbf{z}_k,\mathbf{z}_l)=0,\quad \text{for } k>l>1$

$\operatorname{cov}(\mathbf{Z})$ 必须为对角阵

算法 - I

选择变换矩阵 C 使得 $\operatorname{cov}(\mathbf{Z})$ 为对角阵

$$\operatorname{cov}(\mathbf{Z}) = \mathbf{C}\operatorname{cov}(\mathbf{X})\mathbf{C}^T$$

因为 $\operatorname{cov}(\mathbf{X}) = \mathbf{EDE}^T$，D为对角阵，E为特征向量

$$\mathbf{E}^{-1} = \mathbf{E}^T$$

$$\mathbf{EE}^T = \mathbf{I}$$

选择

$$\mathbf{C} = \mathbf{E}^T; \mathbf{E} = \mathbf{C}^T$$

$$\begin{aligned}\operatorname{cov}(\mathbf{Z}) &= \mathbf{C}\left(\mathbf{EDE}^T\right)\mathbf{C}^T \\ &= \left(\mathbf{CC}^T\right)\mathbf{D}\left(\mathbf{CC}^T\right) \\ &= \mathbf{D}\end{aligned}$$

Matlab Code

```
function [z,C,V] = pca(x)
[M,N] = size(x);
mx = mean(x,2);
x = x - repmat(mx,1,N);
co = 1 / (N-1) * x * x';
[C, V] = eig(co);
V = diag(V);
[temp, iInd] = sort(V);
V = V(iInd);
C = C(:,iInd);
z = C'*x;
```

算法 - II

方阵特征值分解

任意方阵X $\mathbf{X}v = \lambda v$ λ-特征值；v-特征向量

矩阵特征值分解 $\mathbf{X} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1}$ $\mathbf{\Lambda}$ -对角阵；**Q**-正交阵

奇异值分解 (SVD)

任意矩阵X $\mathbf{X}_{p\times n} = \mathbf{U}_{p\times p}\mathbf{\Lambda}_{p\times n}\mathbf{V}^{T}_{n\times n}$

U,V - 正交阵

$\mathbf{\Lambda}$ - 对角元非零 (奇异值)

算法 - II

SVD与PCA的关系

$$\mathbf{X}_{p\times n} = \mathbf{U}_{p\times p}\mathbf{\Lambda}_{p\times n}\mathbf{V}_{n\times n}^{T}$$

$$\Downarrow$$

$$\mathbf{Z} = \mathbf{U}_{p\times p}^{T}\mathbf{X}_{p\times n} = \mathbf{\Lambda}_{p\times n}\mathbf{V}_{n\times n}^{T}$$

$$\Downarrow$$

$$\operatorname{cov}(\mathbf{Z}) = \mathbf{\Lambda}_{p\times n}\mathbf{V}_{n\times n}^{T}\mathbf{V}_{n\times n}\mathbf{\Lambda}_{n\times p}^{T} = \mathbf{D}_{p\times p}$$

$$d_{ii} = \sigma_{ii}^{2}$$

$$\mathbf{Z} = \mathbf{U}\mathbf{X}$$

Matlab Code

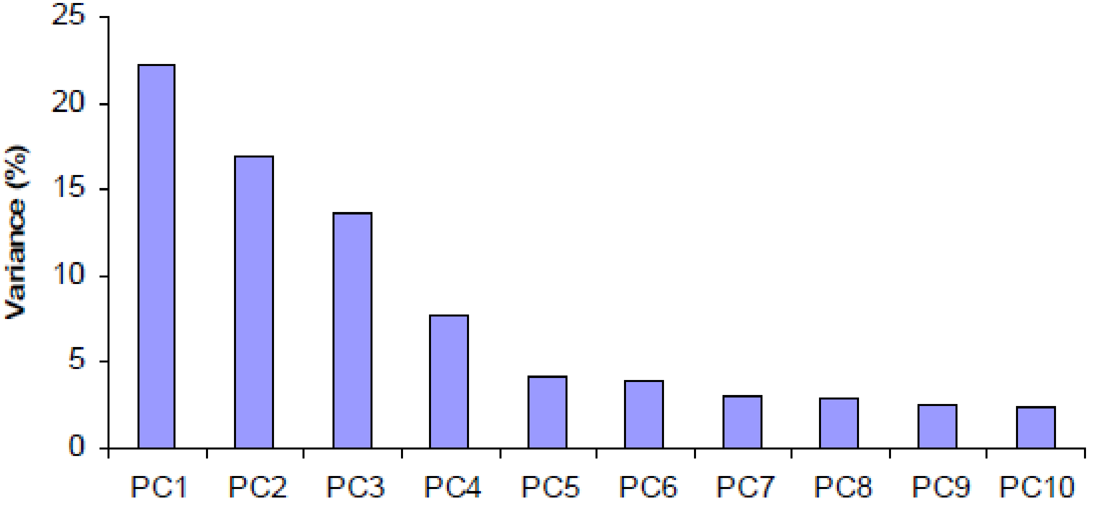
```
function [z,C,V] = pca_svd(x)

[M,N] = size(x);
mx = mean(x,2);
x = x - repmat(mx,1,N);
Y = x' / sqrt(N-1);
[u, S, C] = svd(Y);
S = diag(S);
V = S .* S;
z = C' * x;
```

算法 - II

PC个数的确定

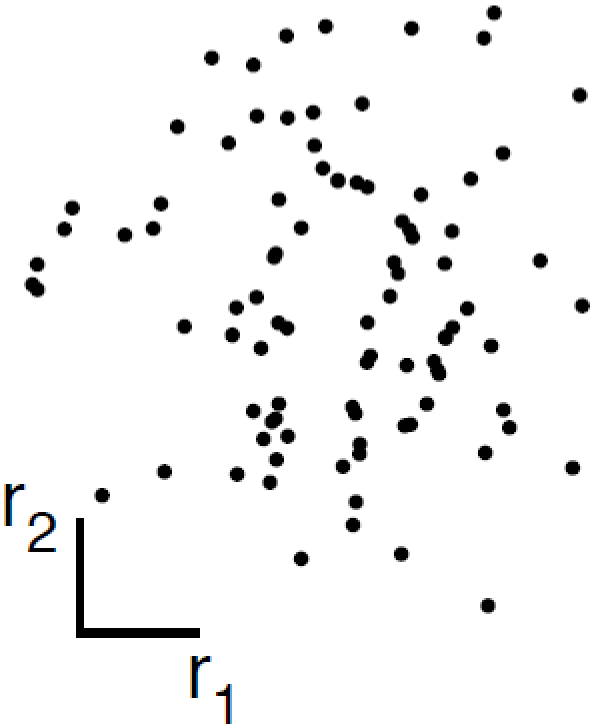
$$\mathbf{Z} = \mathbf{U}^T_{p\times p}\mathbf{X}_{p\times n} = \mathbf{\Lambda}_{p\times n}\mathbf{V}^T_{n\times n}$$

$$\mathbf{X}_{p\times n} \approx \mathbf{U}_{p\times r}\mathbf{\Lambda}_{r\times r}\mathbf{V}^T_{r\times n}$$
$$\approx \mathbf{U}_{p\times r}\mathbf{Z}_{r\times n}$$

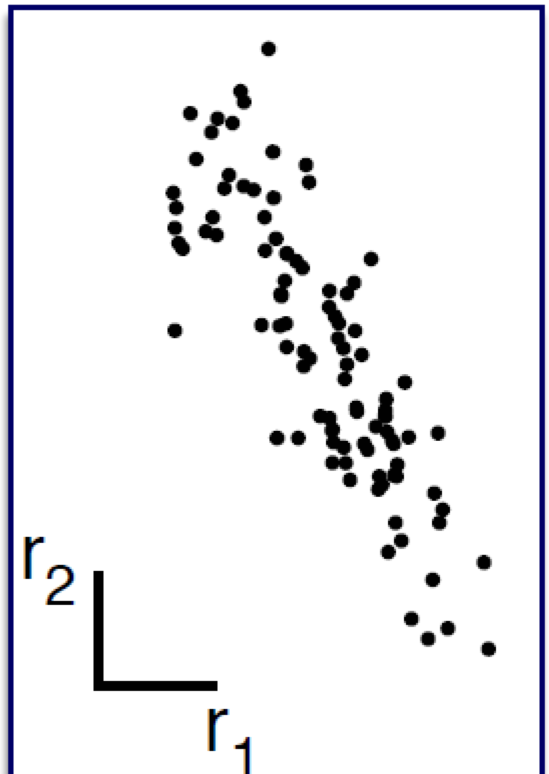


讨论

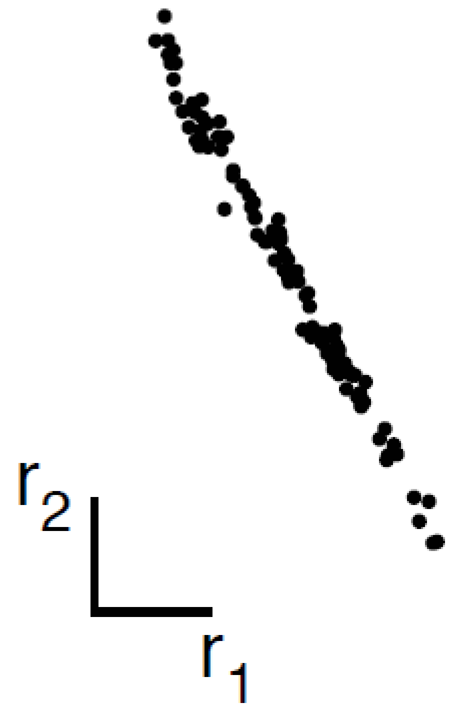
信息冗余度



低



中

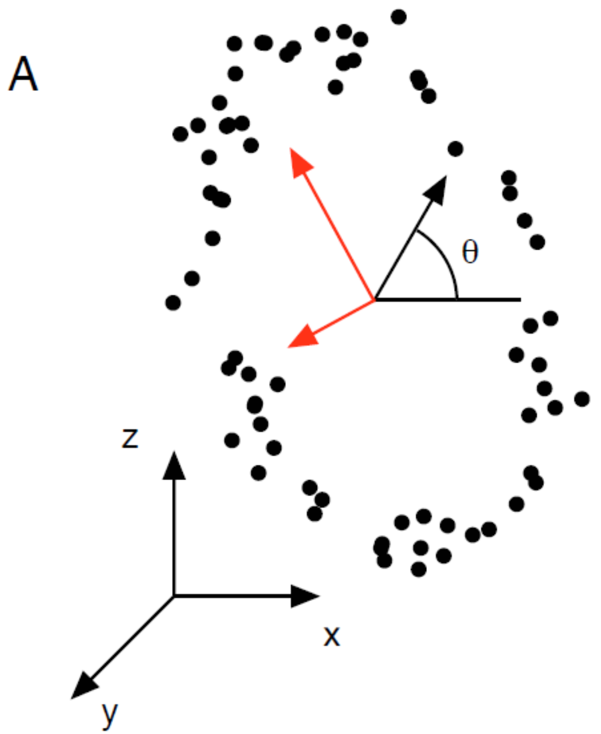


高

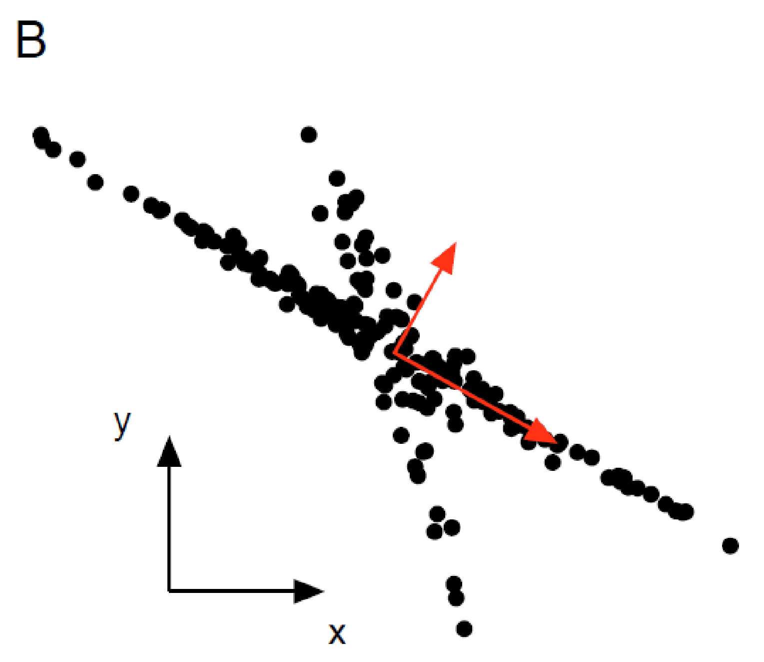
讨论

PCA优点： 非参数化 / 变换矩阵正交

PCA缺点



Kernel PCA



ICA

讨论

模态分析与PCA之间的关系

$$\mathbf{M}\ddot{x}+\mathbf{K}x=\mathbf{F}$$

M， K都为对称阵，存在同一正交阵P，使得

$$\mathbf{P}^{\mathrm{T}}\mathbf{M}\mathbf{P}=\begin{pmatrix} \widehat{m}_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \widehat{m}_n \end{pmatrix} \quad \mathbf{P}^{\mathrm{T}}\mathbf{K}\mathbf{P}=\begin{pmatrix} \widehat{k}_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \widehat{k}_n \end{pmatrix}$$

讨论

模态分析与PCA之间的关系

$$\mathbf{M}\ddot{\boldsymbol{x}} + \mathbf{K}\boldsymbol{x} = \mathbf{F}$$

令$\boldsymbol{y} = \mathbf{P}^{\mathrm{T}}\boldsymbol{x} \rightarrow \boldsymbol{x} = \mathbf{P}\boldsymbol{y}$

$$\Downarrow$$

$$\mathbf{M}\mathbf{P}\ddot{\boldsymbol{y}} + \mathbf{K}\mathbf{P}\boldsymbol{y} = \mathbf{F}$$

$$\Downarrow$$

$$\mathbf{P}^{T}\mathbf{M}\mathbf{P}\ddot{\boldsymbol{y}} + \mathbf{P}^{T}\mathbf{K}\mathbf{P}\boldsymbol{y} = \mathbf{P}^{T}\mathbf{F}$$

$$\Downarrow$$

$$\begin{pmatrix} \widehat{m}_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \widehat{m}_n \end{pmatrix} \ddot{\boldsymbol{y}} + \begin{pmatrix} \widehat{k}_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \widehat{k}_n \end{pmatrix} \boldsymbol{y} = \begin{pmatrix} \widehat{f}_1 \\ \vdots \\ \widehat{f}_n \end{pmatrix}$$

讨论

模态分析与PCA之间的关系

$$\begin{pmatrix} \widehat{m}_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \widehat{m}_n \end{pmatrix} \ddot{y} + \begin{pmatrix} \widehat{k}_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \widehat{k}_n \end{pmatrix} y = \begin{pmatrix} \widehat{f}_1 \\ \vdots \\ \widehat{f}_n \end{pmatrix}$$

假设

$$\|y_1\| > \|y_2\| \cdots > \|y_n\|$$

$$\boldsymbol{x}_{n\times 1} = \mathbf{P}_{n\times n}\boldsymbol{y}_{n\times 1}$$

$$\Downarrow$$

$k << n$ 有限元

$$\boldsymbol{x}_{n\times 1} \approx \widehat{\mathbf{P}}_{n\times k}\boldsymbol{y}_{k\times 1}$$

$$\mathbf{X}_{n\times R} \approx \widehat{\mathbf{P}}_{n\times k}\mathbf{Y}_{k\times R}$$

$\boldsymbol{R}$ – 采样点数

谢谢聆听
欢迎交流