机器学习

苏州大学计算机科学与技术学院
自然语言处理实验室
主讲：周夏冰
邮箱：zhouxiabing@suda.edu.cn

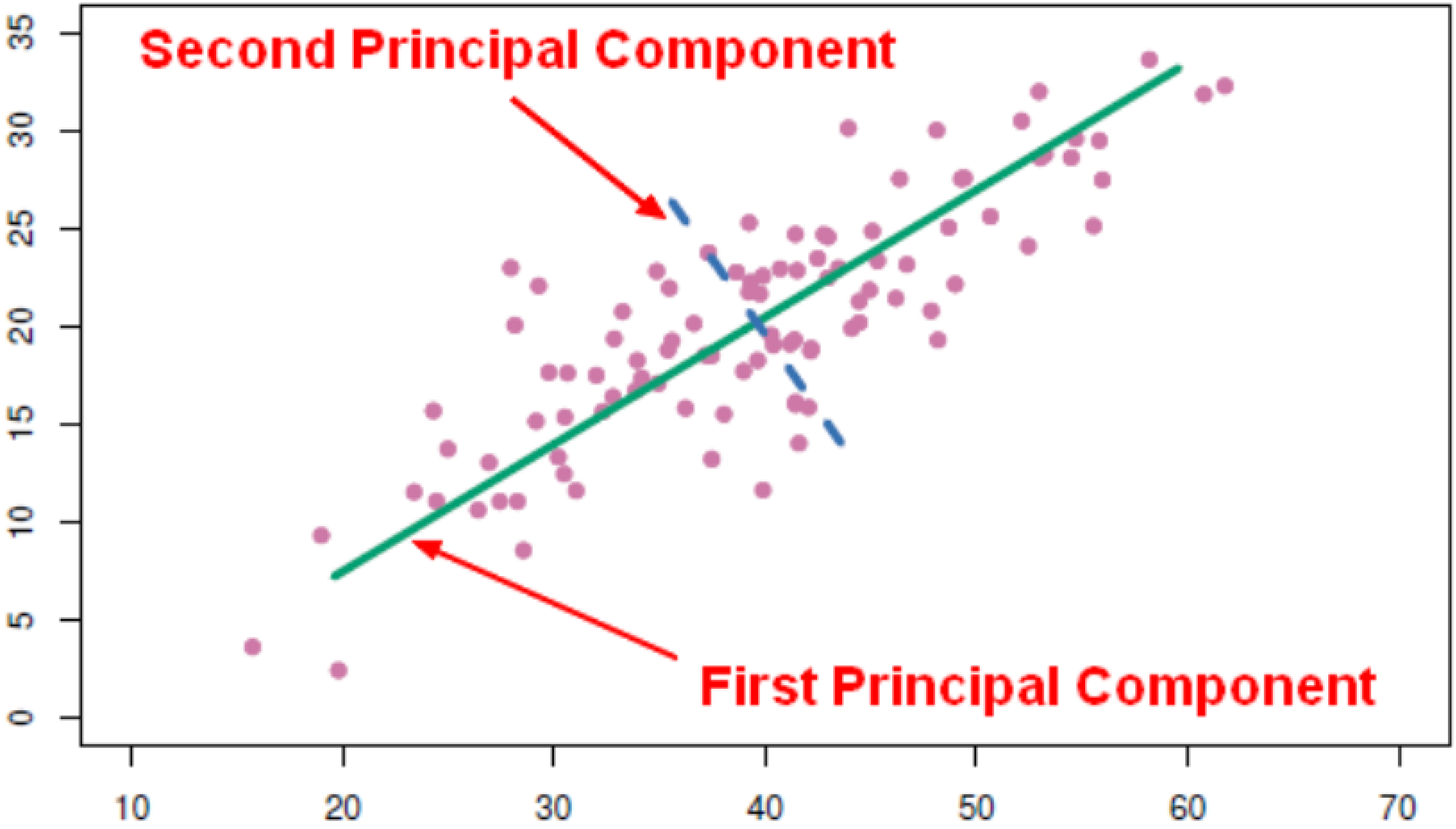
维数灾难

- 密采样
- 高维空间给距离计算带来了很大的麻烦；样本稀疏
- 降维
  - 与学习任务密切相关的也许仅仅是某个低维分布（高维空间的低维嵌入）
  - 多维缩放：保持低维空间样本距离与原来一致
  - 主成分分析：尽量减少信息损失
  - 线性判别分析：保留数据的类别差异

主成分分析

PCA

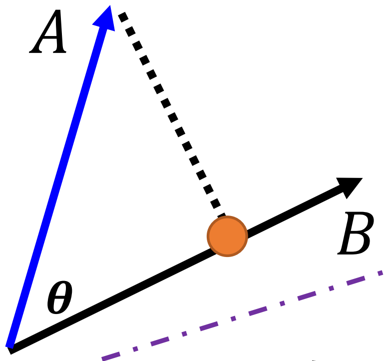
- 降维方法



Second Principal Component
First Principal Component
0
5
10
15
20
25
30
35
10
20
30
40
50
60
70

PCA

- $\boldsymbol{A \cdot B = |A||B|cos\theta}$
- **A在B上的投影**

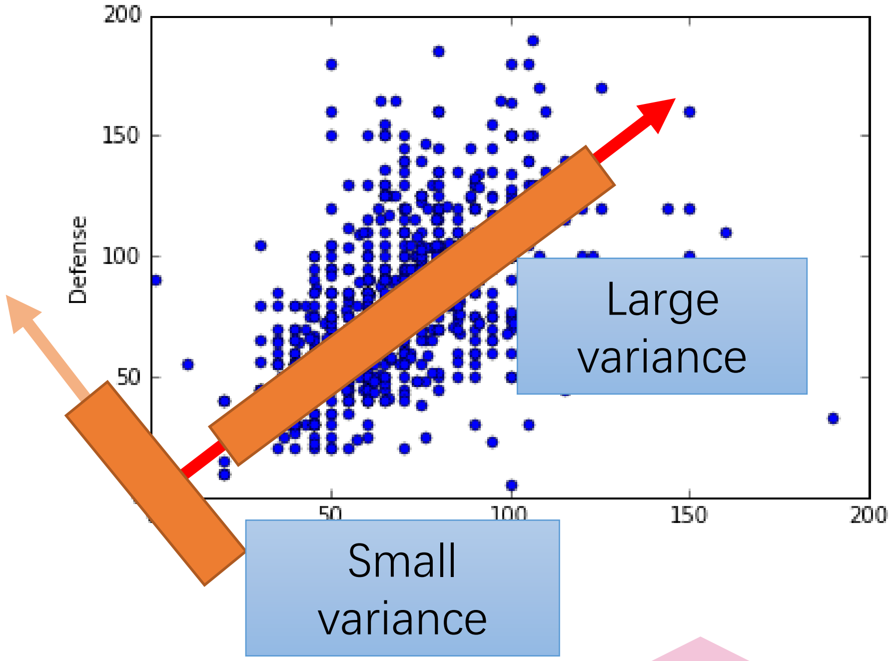


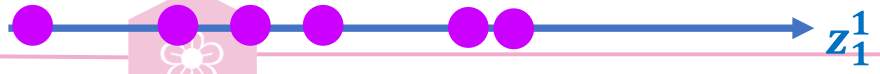
$|A|cos\theta$，当$|B|$=1

- $z = Wx$
- Reduce to 1-D: $z_1 = w^1 \cdot x$, $\|w^1\|_2 = 1$
- $\boldsymbol{Var(z_1) = \frac{1}{N}\sum_{z_1}(z_1 - \overline{z_1})^2}$

$$\boldsymbol{x_1 = \left[x_1^1, \cdots, x_1^d\right]^T}$$
$$\boldsymbol{w^1 = [w_1^1, \cdots, w_d^1]}$$
$$\boldsymbol{W = \begin{bmatrix} w^1 \\ w^2 \\ \cdots \\ w^{d'} \end{bmatrix}}$$

$$\boldsymbol{z_1 = \begin{bmatrix} z_1^1 \\ z_1^2 \\ \cdots \\ z_1^{d'} \end{bmatrix}}$$





PCA

$$z = Wx$$

Reduce to 1-D:

$$z_1 = w^1 \cdot x$$

$$z_2 = w^2 \cdot x$$

$$W = \begin{bmatrix} (w^1)^T \\ (w^2)^T \\ \vdots \end{bmatrix}$$

正交

x的点都映射在 w^1, 获得 z_1

z_1 的协方差越大越好

$$Var(z_1) = \frac{1}{N}\sum_{z_1}(z_1 - \bar{z_1})^2$$

$$\|w^1\|_2 = 1$$

同理，希望 z_2 的协方差越大越好

$$Var(z_2) = \frac{1}{N}\sum_{z_2}(z_2 - \bar{z_2})^2$$

$$\|w^2\|_2 = 1 \qquad w^1 \cdot w^2 = 0$$

PCA

$$z_1 = w^1 \cdot x$$

$$\bar{z_1} = \frac{1}{N}\sum z_1 = \frac{1}{N}\sum w^1 \cdot x = w^1 \cdot \frac{1}{N}\sum x = w^1 \cdot \bar{x}$$

$$Var(z_1) = \frac{1}{N}\sum_{z_1}(z_1 - \bar{z_1})^2$$

$$= \frac{1}{N}\sum_{x}(w^1 \cdot x - w^1 \cdot \bar{x})^2$$

$$= \frac{1}{N}\sum\left(w^1 \cdot (x - \bar{x})\right)^2$$

$$= \frac{1}{N}\sum(w^1)^T(x - \bar{x})(x - \bar{x})^T w^1$$

$$= (w^1)^T \frac{1}{N}\sum(x - \bar{x})(x - \bar{x})^T w^1$$

$$= (w^1)^T Cov(x) w^1$$

$$S = Cov(x)$$

Find w^1 maximizing

$$(w^1)^T S w^1$$

$$\|w^1\|_2 = (w^1)^T w^1 = 1$$

PCA

Find w^1 maximizing $(w^1)^T S w^1$ $\quad (w^1)^T w^1 = 1$

S是协方差矩阵，协方差矩阵是对称的，且半正定（特征值非负）

$$g(w^1) = (w^1)^T S w^1 - \alpha\left((w^1)^T w^1 - 1\right)$$

$$\boldsymbol{w^1} = [\boldsymbol{w^1_1}, \cdots, \boldsymbol{w^1_d}]$$

$$\left.\begin{array}{l} \partial g(w^1)/\partial w^1_1 = 0 \\ \partial g(w^1)/\partial w^1_2 = 0 \\ \vdots \end{array}\right\}$$

$$Sw^1 - \alpha w^1 = 0$$

$$Sw^1 = \alpha w^1$$

$$\boldsymbol{Au = \lambda u}$$

w^1 : 特征向量

$$(w^1)^T S w^1 = \alpha (w^1)^T w^1$$

$$= \alpha$$

找到S特征值最大的对应的特征向量即可

PCA

Find w^2 maximizing $(w^2)^T S w^2$ $\quad (w^2)^T w^2 = 1 \quad (w^2)^T w^1 = 0$

$$g(w^2) = (w^2)^T S w^2 - \alpha\left((w^2)^T w^2 - 1\right) - \beta\left((w^2)^T w^1 - 0\right)$$

$\partial g(w^2)/\partial w_1^2 = 0$

$\partial g(w^2)/\partial w_2^2 = 0$

$\vdots$

$$Sw^2 - \alpha w^2 - \beta w^1 = 0$$

$$(w^2)^T S w^2 - \alpha(1) - \beta(0) = 0$$

$$(w^2)^T S w^2 = \alpha$$

找到S特征值第二大的对应的特征向量即可

PCA

- 算法

$$S = \frac{1}{N}\sum(x-\overline{x})(x-\overline{x})^T$$

输入： 样本集 $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$;
低维空间维数 d'.

过程：

1: 对所有样本进行中心化: $\boldsymbol{x}_i \leftarrow \boldsymbol{x}_i - \frac{1}{m}\sum_{i=1}^{m}\boldsymbol{x}_i$;

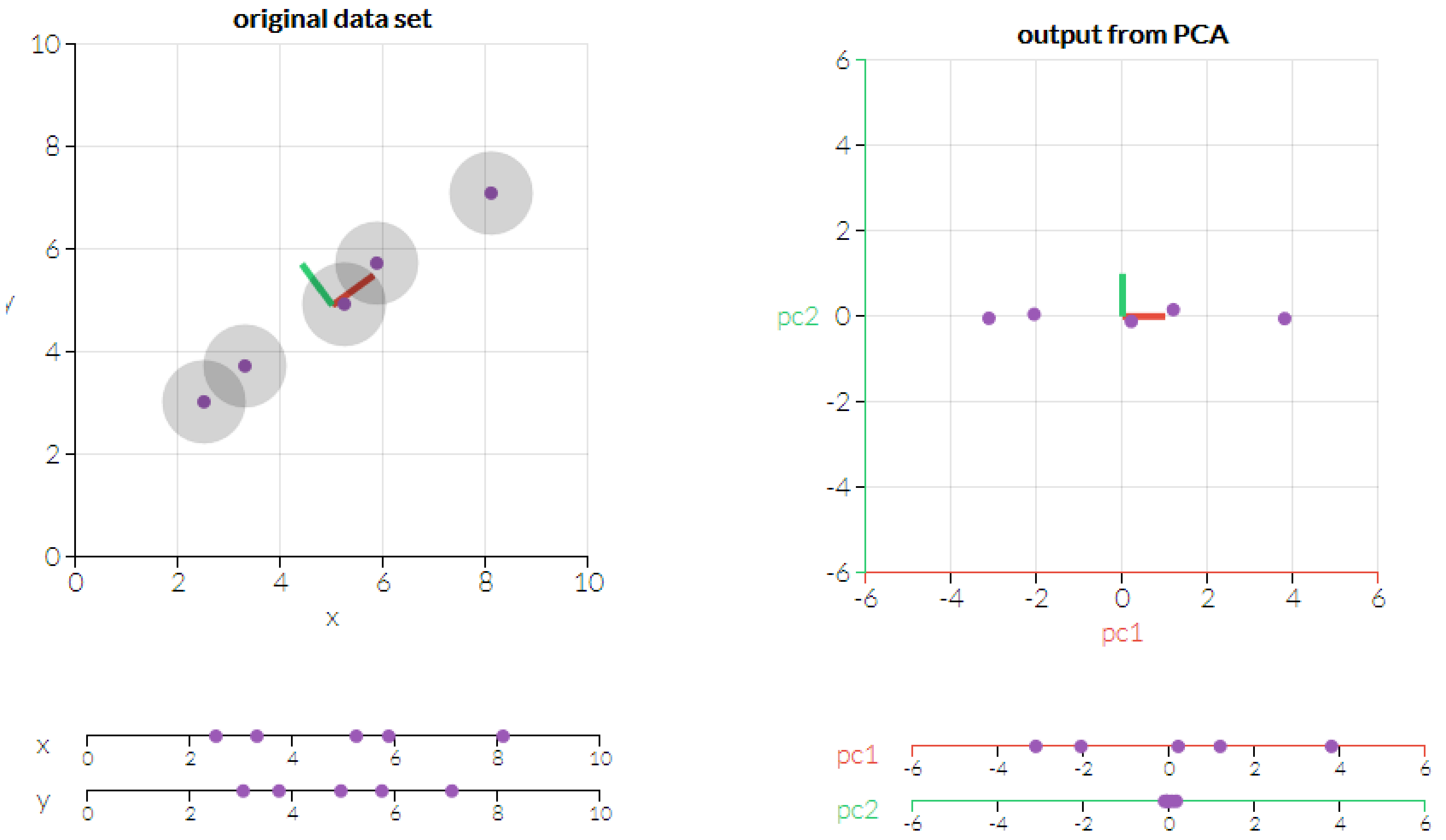
2: 计算样本的协方差矩阵 $\mathbf{X}\mathbf{X}^{\mathrm{T}}$;

3: 对协方差矩阵 $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ 做特征值分解;

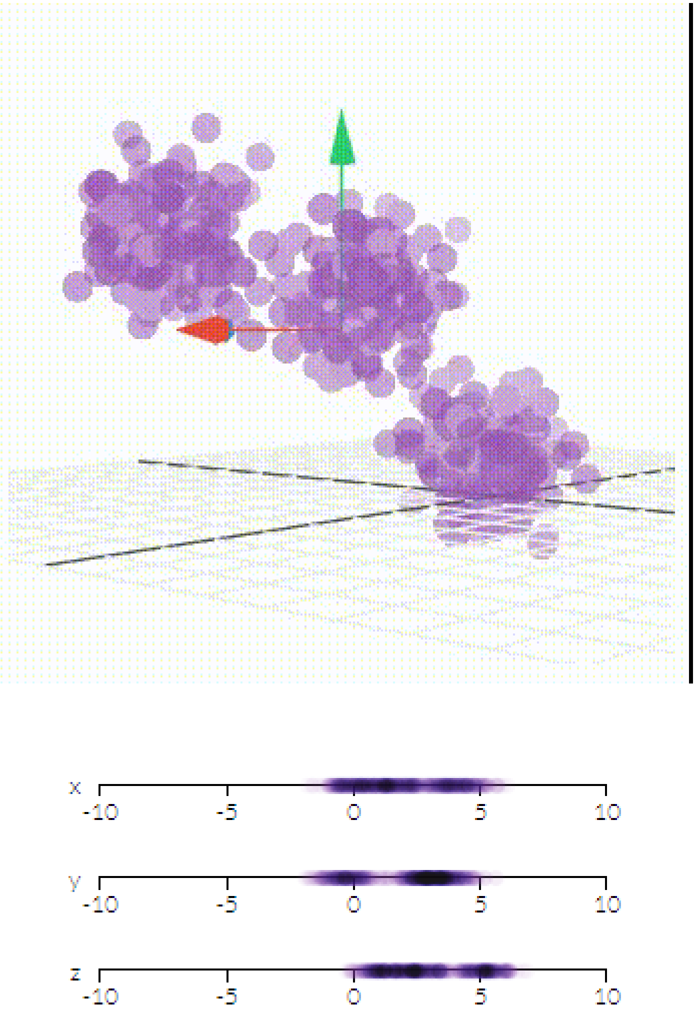
4: 取最大的 d' 个特征值所对应的特征向量 $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_{d'}$.

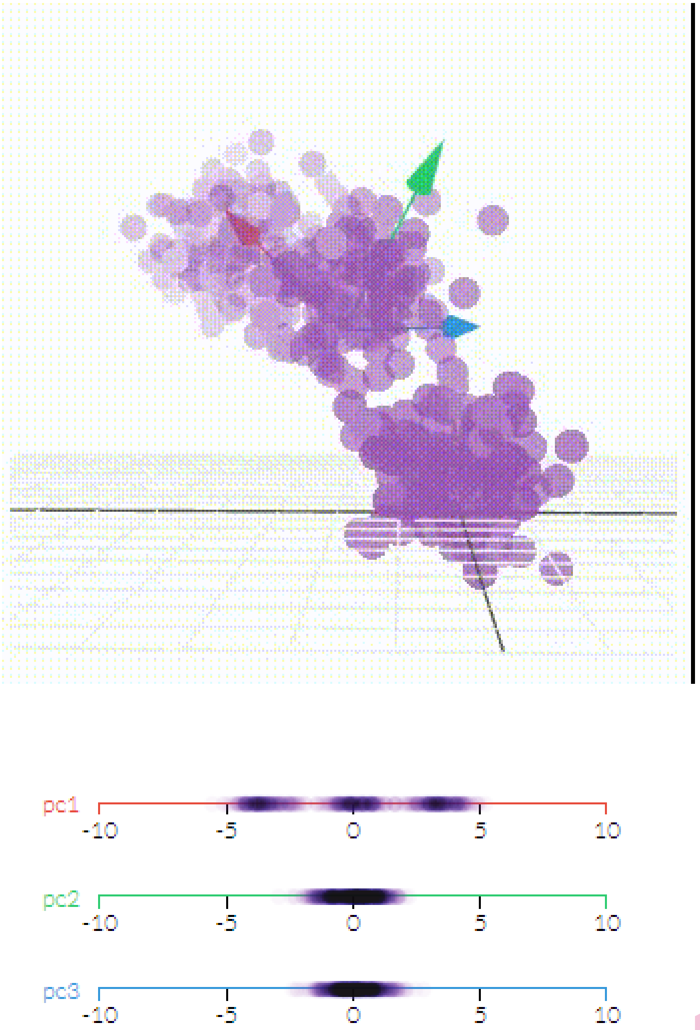
输出： 投影矩阵 $\mathbf{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_{d'})$.

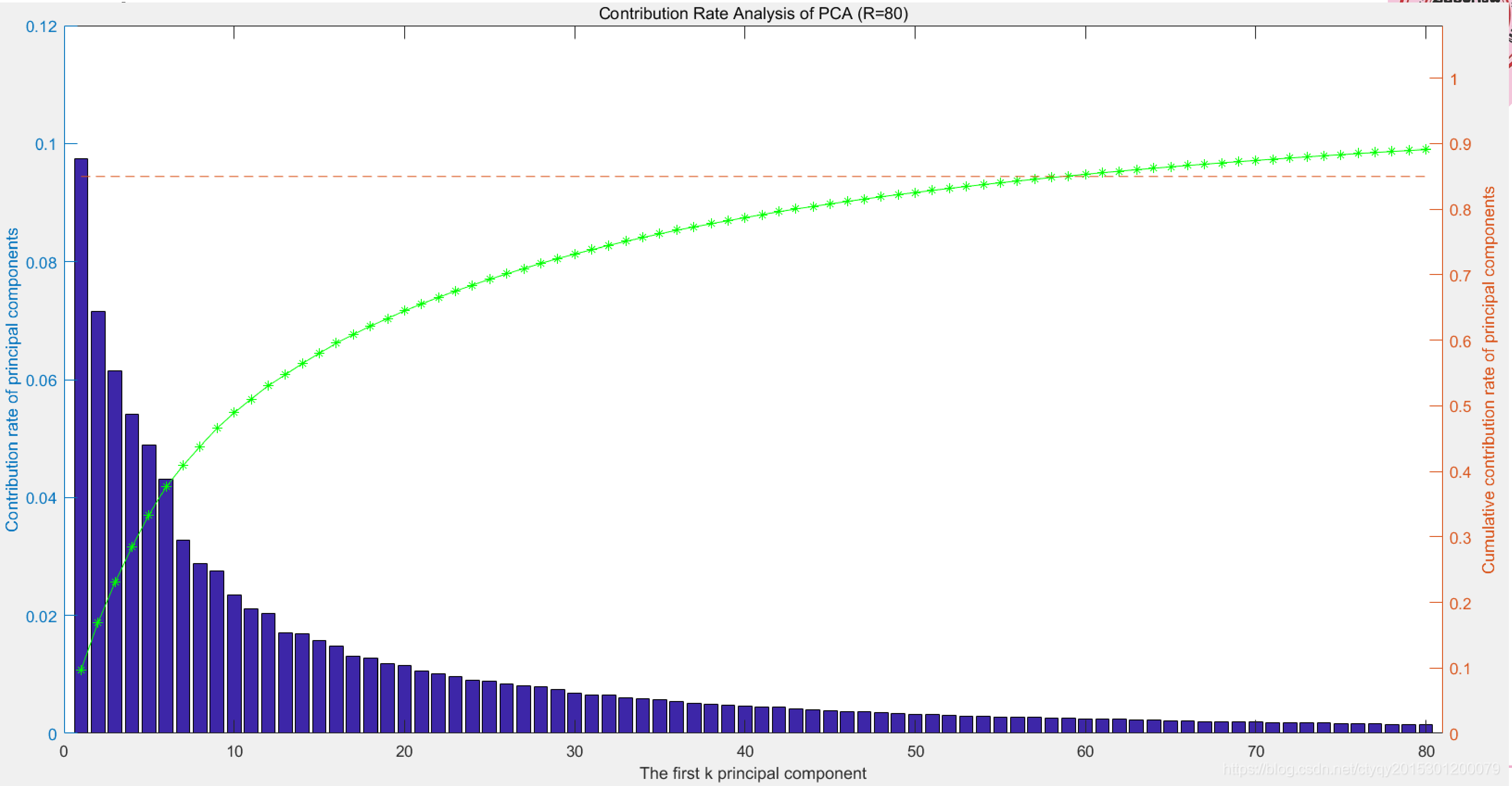
PCA图示



PCA图示






Contribution Rate Analysis of PCA (R=80)
0.12
0.1
0.08
0.06
0.04
0.02
0
Contribution rate of principal components
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Cumulative contribution rate of principal components
0
10
20
30
40
50
60
70
80
The first k principal component