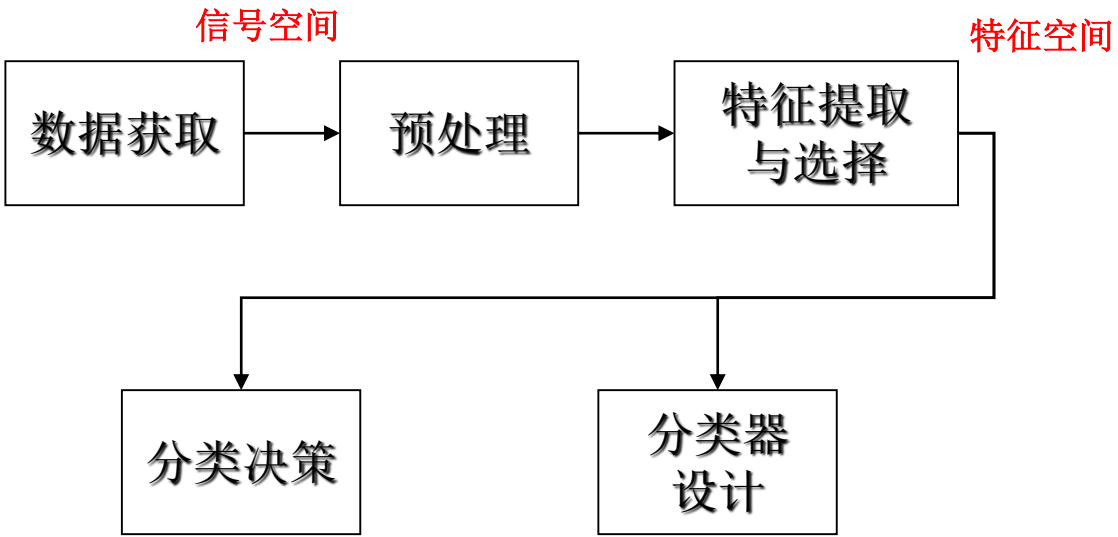
ZIXING AI
自兴人工智能

主成分分析
（Principal Component Analysis）

高琰

目录

引言



引言

- 两类提取有效信息、压缩特征空间的方法：特征提取和特征选择
- **特征提取** (extraction)：用映射（或变换）的方法把原始特征变换为较少的新特征

- **特征选择**(sele 从 m 个特征中选择 m_1 个，$m_1 < m$（人为选择、算法选择）有代表性，分类性能最好的特征

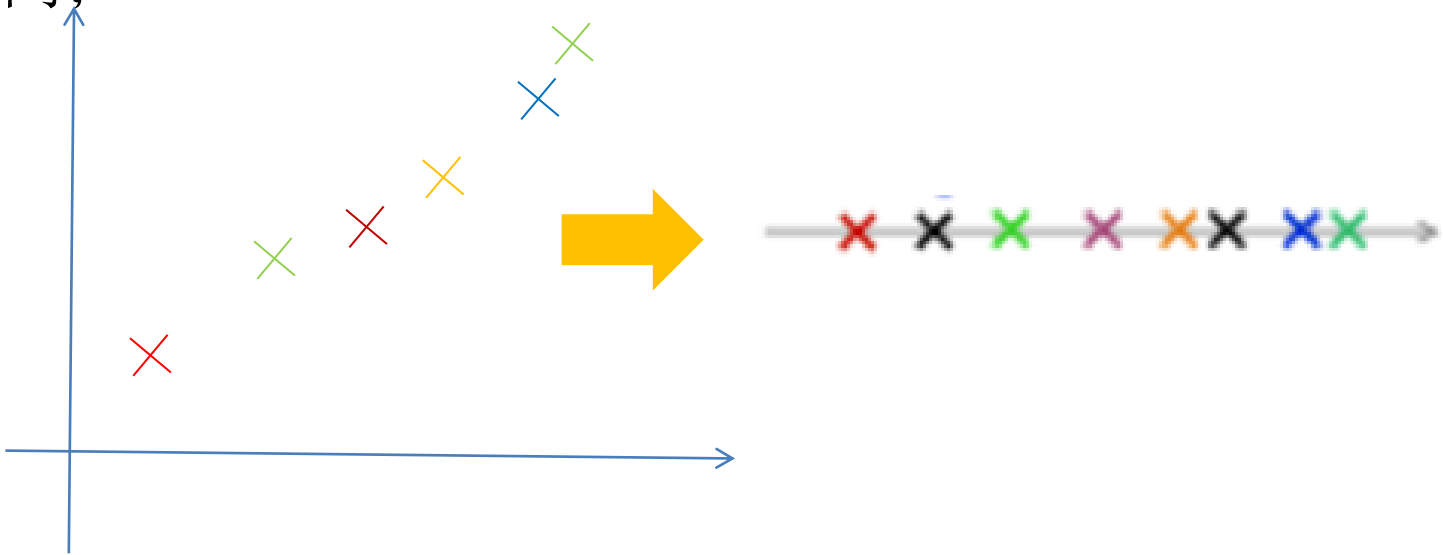
- 特征的选择与 将 m 个特征变为 m_2 个新特征 --- 二次特征 有理论能给出对任何问题都有效的特征选择与提取方法

主成分分析的作用

- 数据压缩
- 数据可视化
- 降维

主成分分析的作用-数据压缩

- PCA：n维数据集可以通过映射降成k维子空间；



主成分分析的作用-数据可视化

Data Visualization

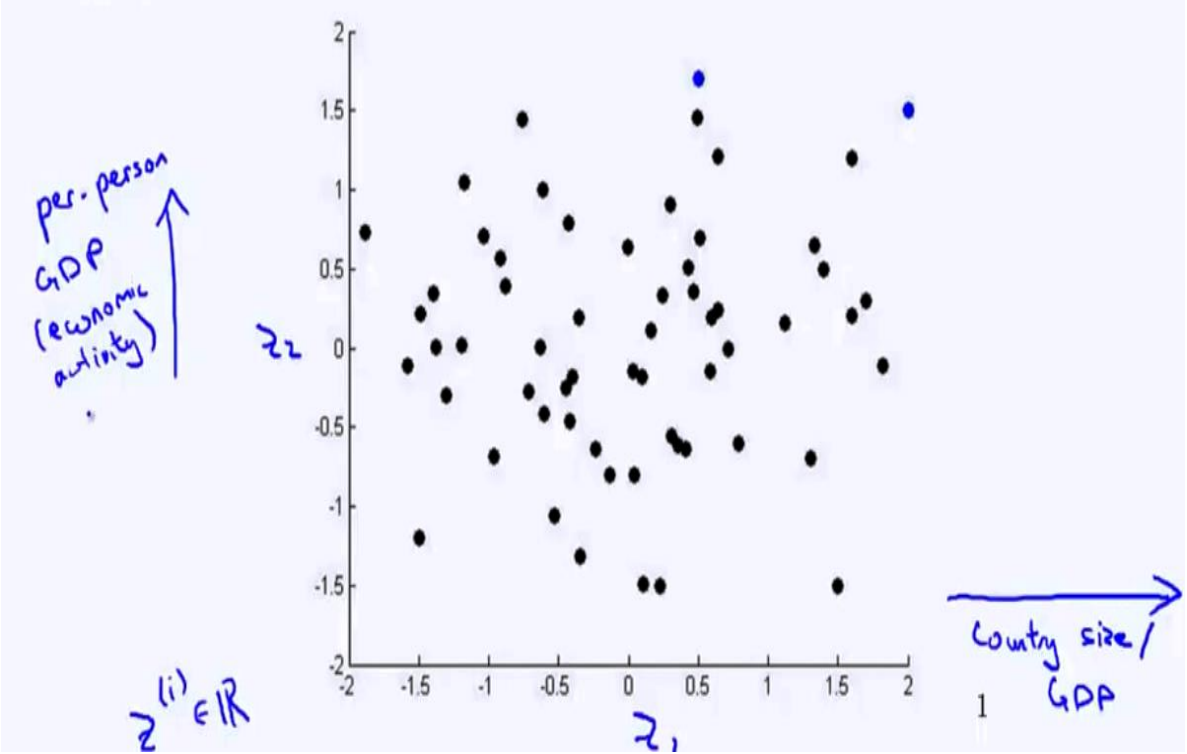
Country	GDP (trillions of US$)	Per capita GDP (thousands of intl. $)	Human Develop-ment Index	Life expectancy	Poverty Index (Gini as percentage)	Mean household income (thousands of US$)	...
Canada	1.577	39.17	0.908	80.7	32.6	67.293	...
China	5.878	7.54	0.687	73	46.9	10.22	...
India	1.632	3.41	0.547	64.7	36.8	0.735	...
Russia	1.48	19.84	0.755	65.5	39.9	0.72	...
Singapore	0.223	56.69	0.866	80	42.5	67.1	...
USA	14.527	46.86	0.91	78.3	40.8	84.3	...
...	...	...	...	...	...	...	

主成分分析的作用-数据可视化

Data Visualization

Country	z_1	z_2
Canada	1.6	1.2
China	1.7	0.3
India	1.6	0.2
Russia	1.4	0.5
Singapore	0.5	1.7
USA	2	1.5
...	...	...

主成分分析的作用-数据可视化



主成分分析的作用-降维

- 一般我们获取的原始数据维度都很高，比如1000个特征，在这1000个特征中可能包含了很多无用的信息或者噪声，真正有用的特征才100个，那么我们可以运用PCA算法将1000个特征降到100个特征。这样不仅可以去除无用的噪声，还能减少很大的计算量。

背景知识-方差、协方差

- **方差**（variance）-度量一组数据分散的程度；各个样本与样本均值的差的平方和的均值；

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})}{n-1}$$

背景知识-方差、协方差

- **协方差**（Covariance）-度量两个变量的线性相关程度；两个变量的协方差为0，则统计学上认为二者线性无关；大于0表示二者正相关，小于0表示二者负相关；

$$\operatorname{cov}(X,Y)=\frac{\sum_{i=1}^{n}(\mathrm{X_i}-\overline{X})(\mathrm{Y_i}-\overline{Y})}{n-1}$$

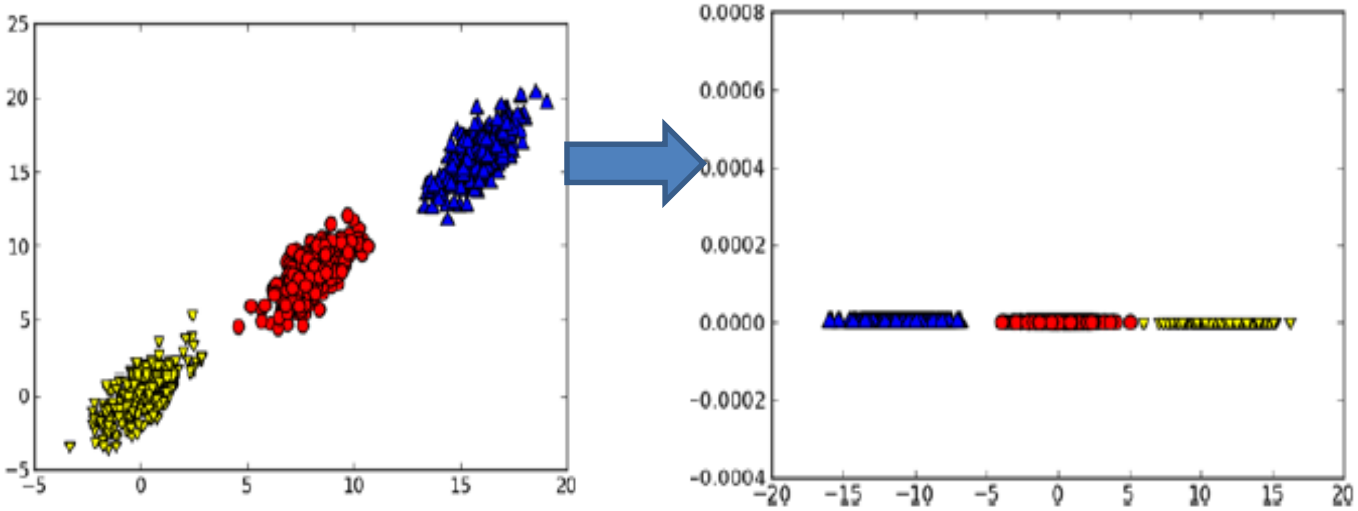
背景知识-协方差矩阵

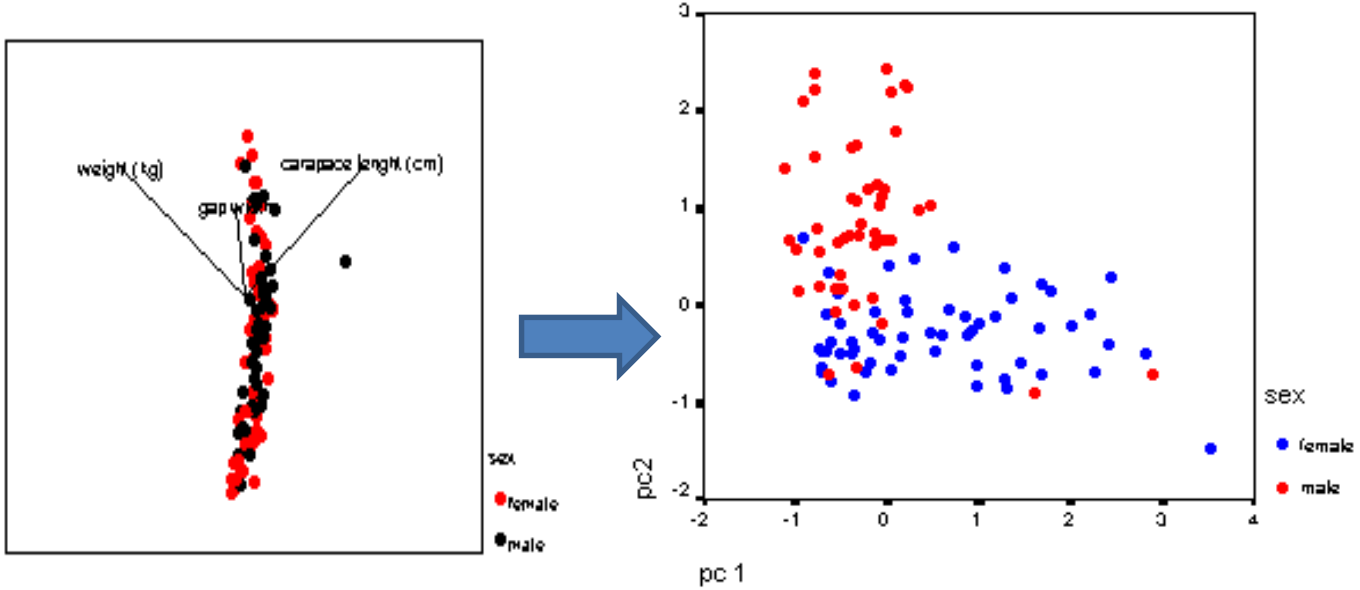
- **协方差矩阵**（Covariance matrix）：
  由数据集中两两变量的协方差组成；矩阵的第(i,j)个元素是数据集中第i个和第j个元素的协方差；如三维数据的协方差矩阵为：

$$c = \begin{bmatrix} \mathrm{cov}(x1,x1) & \mathrm{cov}(x1,x2) & \mathrm{cov}(x1,x3) \\ \mathrm{cov}(x2,x1) & \mathrm{cov}(x2,x2) & \mathrm{cov}(x2,x3) \\ \mathrm{cov}(x3,x1) & \mathrm{cov}(x3,x2) & \mathrm{cov}(x3,x3) \end{bmatrix}$$

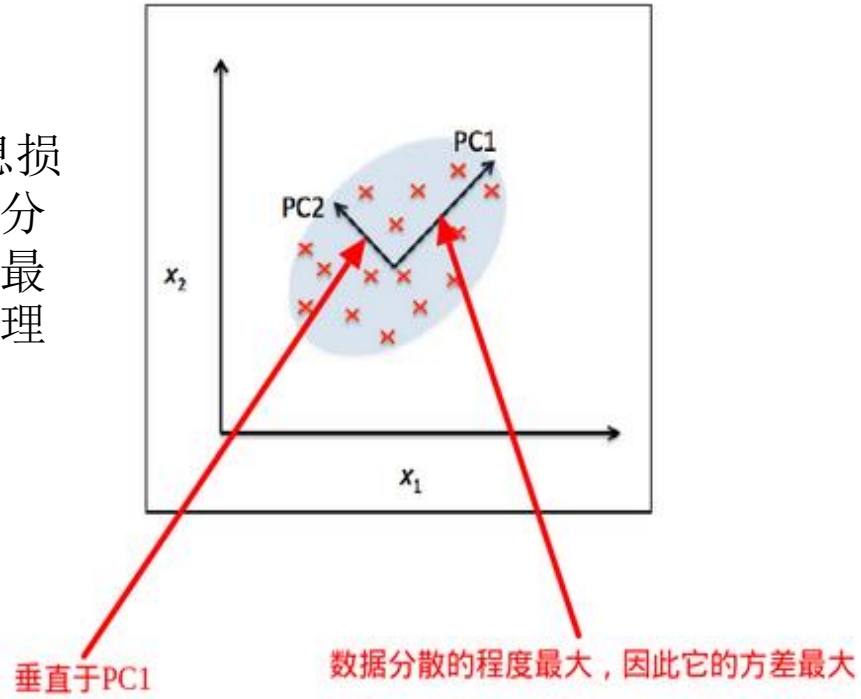
PCA的基本思想

- PCA可以把可能具有相关性的高维变量合成线性无关的低维变量，称为主成分；n维数据集可以通过映射降成k维子空间





•准则：压缩数据时让信息损失最小化，即第一个主成分是从数据差异最大（方差最大）的方向提取。依次处理



PCA的数学描述

- 原特征表示:$x=[x_1,\ldots\ldots,x_p]^T$
- 新特征表示:$x'=[x_1',\ldots\ldots,x_k']^T$

$$x_i' = \sum_{j=1}^{p} \alpha_{ij} x_j = \alpha_i^T x$$

$$x' = W^T x, \alpha_i^T \alpha_i = 1 \qquad \alpha_i = [\alpha_{i1},\ldots\ldots\alpha_{ip}]^T$$

$$W = [\alpha_1,\ldots\ldots\alpha_k] i = 1,\ldots k$$

PCA的公式推导

- 考虑第一个新特征x_1':

$$\mathrm{var}(x_1') = E(x_1'^2) - E(x_1')^2$$

$$= \mathrm{E}[\alpha_1^T x x^T \alpha_1] - \mathrm{E}[\alpha_1^T x] E(x^T \alpha_1)$$

$$= \alpha_1^T \Sigma \alpha_1$$

目标函数：$$\underset{\alpha_1}{\mathrm{argmax}}\ \mathrm{var}(x_1) = \underset{\alpha_1}{\mathrm{argmax}}\ \alpha_1^T \Sigma \alpha_1$$

PCA的公式推导

$$f(\alpha_1) = \alpha_1^T \Sigma \alpha_1 - \lambda\ (\alpha_1^T \alpha_1 - 1) = 0$$

对α_1求导：$\Sigma\alpha_1 - \lambda\alpha_1 = 0$

$$\mathrm{Var}(x_1) = \alpha_1^T \Sigma \alpha_1 = \lambda \alpha_1^T \alpha_1$$

α_1为Σ的最大特征向量对应的特征向量

PCA的公式推导

除了满足协方差最大外，第二个特征还要与第一个特征无关，即：

$$E(x_2' x_1') - E(x_2')E(x_1') = 0$$

$$E[\alpha_2^T x x^T \alpha_1] - E[\alpha_2^T x]E(x^T \alpha_1) = 0$$

$$\alpha_2^T \Sigma \alpha_1 = 0$$

$$f(\alpha_1) = \alpha_2^T \Sigma \alpha_2 - \lambda_2(\alpha_2^T \alpha_2 - 1) - \lambda_2' \alpha_2^T \alpha_1$$

$$\alpha_2^T \alpha_1 = 0$$

$$\Sigma \alpha_2 - \lambda_2 \alpha_2 - \lambda_2' \alpha_1 = 0$$

$$\Sigma\alpha_2 - \lambda_2\alpha_2 - \lambda_2{}'\alpha_1 = 0$$

$$\alpha_2{}^{\mathrm{T}}\Sigma\alpha_2 - \lambda_2\alpha_2{}^{\mathrm{T}}\alpha_2 - \lambda_2{}'\alpha_2{}^{\mathrm{T}}\alpha_1 = 0$$

$$\alpha_2^T\alpha_1 = 0$$

$$\alpha_2{}^{\mathrm{T}}\Sigma\alpha_2 - \lambda_2\alpha_2{}^{\mathrm{T}}\alpha_2 = 0 \quad \Rightarrow \quad \Sigma\alpha_2 - \lambda_2\alpha_2 = 0$$

α_2为Σ的第二大特征向量对应的特征向量

PCA计算过程

- 去除平均值
- 计算协方差矩阵：

$$\Sigma = \frac{1}{\mathrm{m}} \sum_{i=1}^{m} x^{(i)} \; x^{(i)\,\mathrm{T}}$$

- 计算特征向量和特征值：

$$\Sigma \alpha_1 = \lambda \alpha_1$$

- 将特征值从大到小排列
- 保留最上面的k个特征向量

$$\mathrm{W} = [\alpha_{1,\ldots\ldots}\alpha_k]$$

- 将数据转换到上述k个特征向量构建的新空间

$$\mathrm{X}' = W^{\mathrm{T}} X$$

PCA计算过程

Data =

x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

$\bar{x} = 1.81 \bar{y} = 1.91$

DataAdjust =

x	y
.69	.49
-1.31	-1.21
.39	.99
.09	.29
1.29	1.09
.49	.79
.19	-.31
-.81	-.81
-.31	-.31
-.71	-1.01

PCA计算过程

DataAdjust =

x	y
.69	.49
-1.31	-1.21
.39	.99
.09	.29
1.29	1.09
.49	.79
.19	-.31
-.81	-.81
-.31	-.31
-.71	-1.01

$$C = \begin{pmatrix} cov(x,x) & cov(x,y) \\ cov(y,x) & cov(y,y) \end{pmatrix}$$

$$cov = \begin{pmatrix} .616555556 & .615444444 \\ .615444444 & .716555556 \end{pmatrix}$$

PCA计算过程

$$eigenvalues = \begin{pmatrix} .0490833989 \\ 1.28402771 \end{pmatrix}$$

$$cov = \begin{pmatrix} .616555556 & .615444444 \\ .615444444 & .716555556 \end{pmatrix} \Rightarrow \begin{pmatrix} -.735178656 & -.677873399 \\ .677873399 & -.735178656 \end{pmatrix}$$

将特征值按照从大到小的顺序排序，选择其中最大的k个。
我们选择其中最大的那个，这里是1 .28402771，对应的特征向量是：
$(-0.677873399，735178656)^T$

PCA计算过程

$$eigenvalues = \begin{pmatrix} .0490833989 \\ 1.28402771 \end{pmatrix}$$

$$cov = \begin{pmatrix} .616555556 & .615444444 \\ .615444444 & .716555556 \end{pmatrix} \Rightarrow \begin{pmatrix} -.735178656 & -.677873399 \\ .677873399 & -.735178656 \end{pmatrix}$$

将特征值按照从大到小的顺序排序，选择其中最大的k个。
我们选择其中最大的那个，这里是1 .28402771，
对应的特征向量是：(-0.677873399，735178656)^T

PCA计算过程

$$FinalData(m*k) = DataAdjust(m*n) \times EigenVectors(n*k)$$

DataAdjust =

x	y
.69	.49
-1.31	-1.21
.39	.99
.09	.29
1.29	1.09
.49	.79
.19	-.31
-.81	-.81
-.31	-.31
-.71	-1.01

Transformed Data (Single eigenvector)

x
-.827970186
1.77758033
-.992197494
-.274210416
-1.67580142
-.912949103
.0991094375
1.14457216
.438046137
1.22382056

PCA的python实现

- loadDataset(filename,delim='\t')
- cov(?)建立协方差矩阵
  - cov(X,0) = cov(X) 除数是n-1(n为样本个数),除数n-1是为了得到协方差的无偏估计
  - cov(X,1) 除数是n
- linag.eig(?)求特征向量和特征值
- argsort :对特征值矩阵进行由大到小排序，返回对应排序后的索引排序

实验任务

- 对UCI的iris的数据集进行PCA处理
- 输出pca的特征值(由大到小排序)，填写下表:累计方差用np.cumsum(a)

序号	特征值	方差比	累计方差比
1			

- 画出方差比折线图,累计方差比折线图,
- K=2时，画出原始数据和重构数据散点图，画出数据在新特征空间散点图
- K=1时，画出重构数据折线图，画出数据在新特征空间折线图