第九章 特征变换与降维表示

苏智勇
可视计算研究组
南京理工大学
suzhiyong@njust.edu.cn
https://zhiyongsu.github.io

主要内容

9.1 引言

- **特征选择**
  - 从D个特征中选出d个
- **特征变换**
  - 把D个特征变为d个新特征
  - 最常采用线性变换

$$\boldsymbol{y} = \boldsymbol{W}^{\mathsf{T}}\boldsymbol{x}$$

其中，$\boldsymbol{W}$是$D \times d$维矩阵，称作变换阵。通常，$d < D$。

9.2 基于类别可分性判据的特征变换

- **准则函数**（变换后的可分离判据）
  - $J_1 = \mathrm{tr}(\boldsymbol{W}^T(\boldsymbol{S}_w + \boldsymbol{S}_b)\boldsymbol{W})$
  - $J_2 = \mathrm{tr}\left[\left(\boldsymbol{W}^T\boldsymbol{S}_w\boldsymbol{W}\right)^{-1}(\boldsymbol{W}^T\boldsymbol{S}_b\boldsymbol{W})\right]$
  - $J_3 = \ln \dfrac{\left|\boldsymbol{W}^T\boldsymbol{S}_b\boldsymbol{W}\right|}{\left|\boldsymbol{W}^T\boldsymbol{S}_w\boldsymbol{W}\right|}$
  - $J_4 = \dfrac{\mathrm{tr}(\boldsymbol{W}^T\boldsymbol{S}_b\boldsymbol{W})}{\mathrm{tr}(\boldsymbol{W}^T\boldsymbol{S}_w\boldsymbol{W})}$
  - $J_5 = \dfrac{\left|\boldsymbol{W}^T(\boldsymbol{S}_w + \boldsymbol{S}_b)\boldsymbol{W}\right|)}{\left|\boldsymbol{W}^T\boldsymbol{S}_u\boldsymbol{W}\right|}$

9.2 基于类别可分性判据的特征变换

- **目标**

  求得最优W^*，使

$$\boldsymbol{W}^* = \arg\max_{\{\boldsymbol{W}\}} J(\boldsymbol{W}^T\boldsymbol{x})$$

- **结论**
  - 设矩阵$\boldsymbol{S}_w^{-1}\boldsymbol{S}_b$的本征值为$\lambda_1, \lambda_2, \cdots, \lambda_D$，且$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D$
  - 选前d个本征值对应的本征向量$\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_d$组成矩阵

$$\boldsymbol{W} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_d]$$

  即为最佳变换阵

9.2 基于类别可分性判据的特征变换

- **推导（以J_1为例）**
  - 优化问题

$$J_1 = \mathrm{tr}(\boldsymbol{W}^T(\boldsymbol{S}_w + \boldsymbol{S}_b)\boldsymbol{W})$$

$$\begin{aligned} &\max J_1(\boldsymbol{W}) \\ &\mathrm{s.t.}\ \mathrm{tr}(\boldsymbol{W}^T\boldsymbol{S}_w\boldsymbol{W}) = 1 \end{aligned}$$

  - 拉格朗日函数

$$g(\boldsymbol{W}) = J_1(\boldsymbol{W}) - \mathrm{tr}\left[\boldsymbol{\Lambda}\left(\boldsymbol{W}^T\boldsymbol{S}_w\boldsymbol{W} - \boldsymbol{I}\right)\right]$$

  - 令

$$\frac{\partial}{\partial \boldsymbol{W}} g(\boldsymbol{W}) = 0$$

9.2 基于类别可分性判据的特征变换

- **推导（以J_1为例）**
  - 整理可得

$$\boldsymbol{S}_w^{-1}\boldsymbol{S}_b\boldsymbol{W} = \boldsymbol{W}(\Lambda - \boldsymbol{I})$$

  - 考虑限制条件，可得

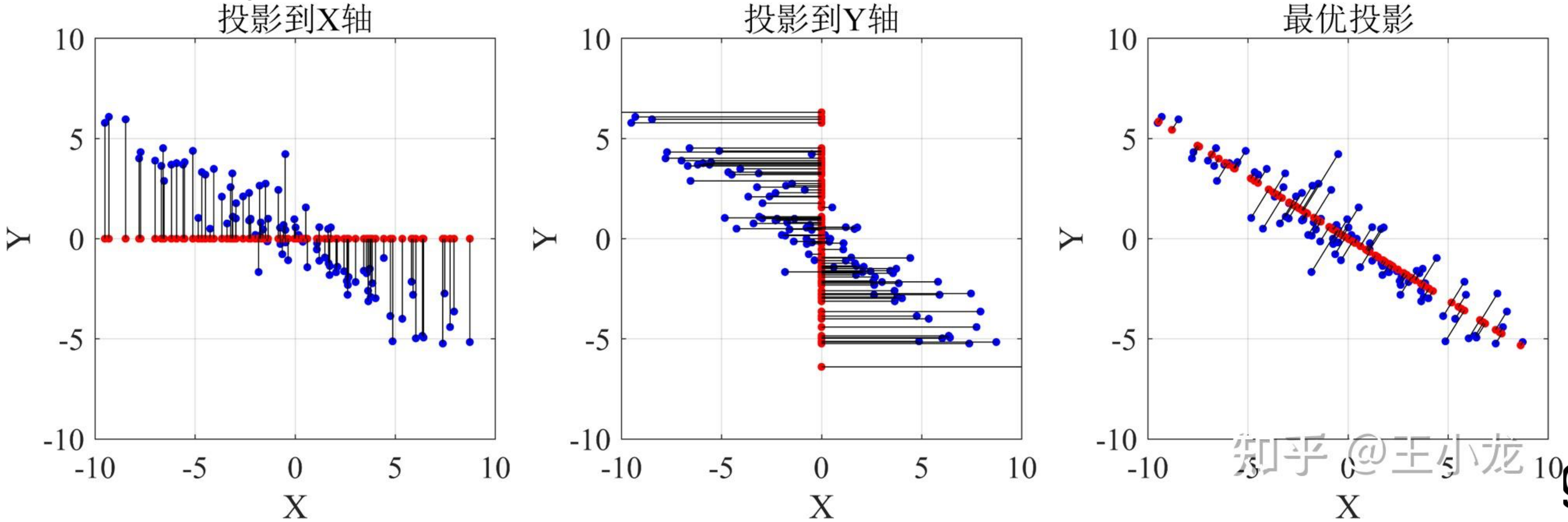
$$J_1(\boldsymbol{W}) = \mathrm{tr}\left(\boldsymbol{W}^T(\boldsymbol{S}_w + \boldsymbol{S}_b)\boldsymbol{W}\right) = \mathrm{tr}\left(\boldsymbol{W}^T\boldsymbol{S}_w\boldsymbol{W}\Lambda\right) = \mathrm{tr}\Lambda$$

  - 令对$D \times d$维变换矩阵

$$J_1(\boldsymbol{W}) = \sum_{i=1}^{d}(1 + \lambda_i)$$

9.3 主成分分析

- **目的**
  - 从一组特征中计算一组重要性从大到小排列的新特征，他们是原特征的线性组合，并且相互之间不相关（正交）
  - 把原有的多个指标（特征）转化成少数几个代表性较好的综合指标（新特征），这少数几个指标能反映原来指标大部分（如85%以上）的信息，且各个指标保持独立。



9.3 主成分分析

- **数学表示**
  - 原特征：$x_1, \cdots, x_p$
  - 新特征：

$$\xi_i = \sum_{j=1}^{p} \alpha_{ij} x_j = \boldsymbol{\alpha}_i^T \boldsymbol{x}, \quad i = 1, \cdots, p$$

矩阵形式：

$$\xi = \boldsymbol{A}^T \boldsymbol{x}$$

其中，为了统一ξ_i的尺度：$\boldsymbol{\alpha}_i^T \boldsymbol{\alpha}_i = 1$

9.3 主成分分析

- **第一主成分**ξ_1

$$\xi_1 = \sum_{j=1}^{p} \alpha_{1j} x_j = \boldsymbol{\alpha}_1^T \boldsymbol{x}$$

$$var(\xi_1) = E[\xi_1^2] - E[\xi_1]^2 = E[\boldsymbol{\alpha}_1^T x^T \boldsymbol{\alpha}_1] - E[\boldsymbol{\alpha}_1^T \boldsymbol{x}]E[\boldsymbol{x}^T \boldsymbol{\alpha}_1] = \boldsymbol{\alpha}_1^T \Sigma \boldsymbol{\alpha}_1$$

其中，Σ是$\boldsymbol{x}$的**协方差矩阵**

- 求下列拉格朗日函数的极值

$$f(\boldsymbol{\alpha}_1) = \boldsymbol{\alpha}_1^T \boldsymbol{\Sigma} \boldsymbol{\alpha}_1 - v\boldsymbol{\alpha}_1^T \boldsymbol{\alpha}_1,$$

得

$$\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 = v\boldsymbol{\alpha}_1$$

$$var(\xi_1) = \boldsymbol{\alpha}_1^T \boldsymbol{\Sigma} \boldsymbol{\alpha}_1 = v\boldsymbol{\alpha}_1^T \boldsymbol{\alpha}_1 = v$$

最优的$\boldsymbol{\alpha}_1$为$\boldsymbol{\Sigma}$的最大本质值对应的本征向量，ξ_1称作**第一主成分**

9.3 主成分分析

- **第二主成分**ξ_2

$$E[\xi_2\xi_1] - E[\xi_2]E[\xi_1] = 0$$

$$\boldsymbol{\alpha}_2^T\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 = 0$$

$$\boldsymbol{\alpha}_2^T\boldsymbol{\alpha}_1 = 0$$

$$\boldsymbol{\alpha}_2^T\boldsymbol{\alpha}_2 = 1$$

其中，$\boldsymbol{\alpha}_2$为$\boldsymbol{\Sigma}$的第二本征值对应的本征向量，ξ_2称作**第二主成分**

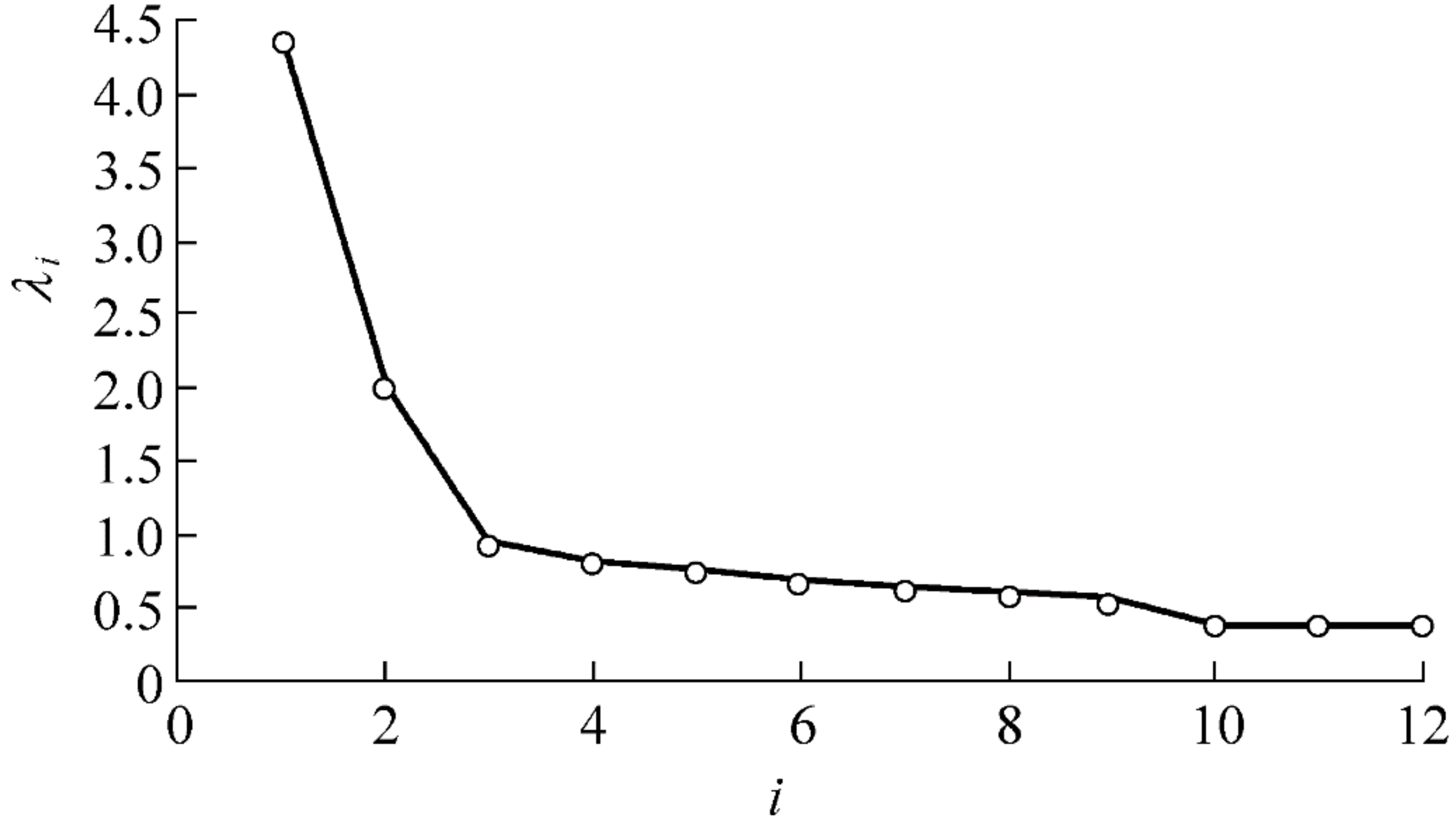
9.3 主成分分析

- 全部主成分方差之和等于各个原始特征方差之和：

$$\sum_{i=1}^{p} var(\xi_i) = \sum_{i=1}^{p} \lambda_i$$

- 全前k个主成分代表了数据全部方差的比例：

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i}$$



9.3 主成分分析

- PCA算法流程

输入： n 维样本集 $D=(x^{(1)},x^{(2)},\cdots,x^{(m)})$ ，要降维到的维数 n' 。

输出：降维后的样本集 D'

(1):对所有的样本进行中心化： $x^{(i)}=x^{(i)}-\frac{1}{m}\sum_{j=1}^{m}x^{(j)}$

(2):计算特征的协方差矩阵 XX^T

(3):对协方差矩阵 XX^T 进行特征值分解

(4):取出最大的 n' 个特征值对应的特征向量 $(w_1,w_2,\cdots,w_{(n')})$ ，将所有的特征向量标准化后，组成特征向量矩阵 W 。

(5):对样本集中的每一个样本 $x^{(i)}$ ，转化为新的样本 $z^{(i)}=W^Tx^{(i)}$

(6):得到输出样本集 $D'=(z^{(1)},z^{(2)},\cdots,z^{(m)})$

9.3 主成分分析

- PCA实例

假设我们得到的2维数据如下：

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

行代表了样例，列代表特征，这里有10个样例，每个样例两个特征。

9.3 主成分分析

- PCA实例

第一步分别求 x 和 y 的平均值，然后对于所有的样例，都减去对应的均值。这里 x 的均值是1.81， y 的均值是1.91，那么一个样例减去均值后即为 $(0.69, 0.49)$ ，得到

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

第二步，求特征协方差矩阵，如果数据是3维，那么协方差矩阵是

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

这里只有 x 和 y ，求解得

$$C = \begin{pmatrix} 0.616555556 & 0.615444444 \\ 0.615444444 & 0.716555556 \end{pmatrix}$$

对角线上分别是 x 和 y 的方差，非对角线上是协方差。协方差大于0表示 x 和 y 若有一个增，另一个也增；小于0表示一个增，一个减；协方差为0时，两者独立。协方差绝对值越大，两者对彼此的影响越大，反之越小。

9.3 主成分分析

- PCA实例

第三步，求协方差的特征值和特征向量，得到

$$eigenvalues = \begin{pmatrix} 0.490833989 \\ 1.28402771 \end{pmatrix}$$

$$eigenvectors = \begin{pmatrix} -0.735178656 & -0.677873399 \\ 0.677873399 & -0.735178656 \end{pmatrix}$$

上面是两个特征值，下面是对应的特征向量，这里的特征向量都归一化为单位向量。

第四步，将特征值按照从大到小的顺序排序，选择其中最大的 k 个，然后将其对应的 k 个特征向量分别作为列向量组成特征向量矩阵。

这里特征值只有两个，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,-0.735178656)。

第五步，将样本点投影到选取的特征向量上。假设样例数为 m，特征数为 n，减去均值后的样本矩阵为 $DataAdjust(m*n)$，协方差矩阵是 $n*n$，选取的 k 个特征向量组成的矩阵为 $EigenVectors(n*k)$。那么投影后的数据FinalData为

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

这里是

$$FinalData(10*1) = DataAdjust(10*2\text{矩阵}) \times \text{特征向量}(-0.677873399, -0.735178656)$$

得到结果是

Transformed Data (Single eigenvector)

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

这样，就将原始样例的 n 维特征变成了 k 维，这 k 维就是原始特征在 k 维上的投影。

9.4 Karhunen-Loève 变换

- **简介**
  - Karhunen-Loève 变换简称K-L变换，是一种常用的特征提取方法
  - 最基本的形式原理与主成分分析相同，两者都属于**正交变换**
  - PCA是无监督变换，是一种特殊的离散K-L变换
  - K-L变换可以实现**有监督**的特征提取，可以处理离散、连续的情况

9.4.1 K-L 变换

- **基本原理**
  - **函数的级数展开**：将函数用一组（正交）**基函数**展开，用**展开系数**表示原函数
  - **离散K-L展开**：把随机向量用一组**正交基向量**展开，用**展开系数**代表原向量
  - **基向量所张成的空间**：新的特征空间
  - **展开系数组成的向量**：新特征空间中的样本向量

9.4.1 K-L 变换

- **离散K-L展开**
  - 对随机向量$\boldsymbol{x}$，用确定的完备正交归一向量系$\boldsymbol{u}_j$，$j=1,2,\cdots,\infty$展开，得

$$\boldsymbol{x}=\sum_{j=1}^{\infty}c_j\boldsymbol{u}_j,\ \ c_j=\boldsymbol{u}_j^T\boldsymbol{x}$$

其中，$\boldsymbol{u}_i^T\boldsymbol{u}_j=\begin{cases}1,\ i=j\\0,\ i\neq j\end{cases}$

  - 只用有限项来逼近$\boldsymbol{x}$，即$\hat{\boldsymbol{x}}=\sum_{j=1}^{d}c_j\boldsymbol{u}_j$，其中，$\boldsymbol{x}$为$D$维度，且$d<D$，则与原向量的**均方误差**为：

$$e=E[(\boldsymbol{x}-\hat{\boldsymbol{x}})^T(\boldsymbol{x}-\hat{\boldsymbol{x}})]=E\left[\left(\sum_{j=1}^{\infty}c_j\boldsymbol{u}_j\right)^T\left(\sum_{j=d+1}^{\infty}c_j\boldsymbol{u}_j\right)\right]=E\left[\sum_{j=d+1}^{\infty}c_j^2\right]=E\left[\sum_{j=d+1}^{\infty}\boldsymbol{u}_j^T\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{u}_j\right]=\sum_{j=d+1}^{\infty}\boldsymbol{u}_j^TE[\boldsymbol{x}\boldsymbol{x}^T]\boldsymbol{u}_j$$

9.4.1 K-L 变换

- **离散K-L展开**
  - 记$\Psi = E[\boldsymbol{x}\boldsymbol{x}^T]$，则：$e = \sum_{j=d+1}^{\infty} \boldsymbol{u}_j^T \boldsymbol{\psi} \boldsymbol{u}_j$
  - 最小化均方误差

$$\min \ e = \sum_{j=d+1}^{\infty} \boldsymbol{u}_j^T \boldsymbol{\psi} \boldsymbol{u}_j$$

$$\text{s.t.} \ \boldsymbol{u}_j^T \boldsymbol{u}_j = 1, \ \forall j$$

  - 拉格朗日函数：$g(\boldsymbol{u}) = \sum_{j=d+1}^{\infty} \boldsymbol{u}_j^T \boldsymbol{\psi} \boldsymbol{u}_j - \sum_{j=d+1}^{\infty} \lambda \left[\boldsymbol{u}_j^T \boldsymbol{u}_j - 1 \right]$

9.4.1 K-L 变换

- **离散K-L展开**
  - 令$\dfrac{\partial}{\partial \boldsymbol{u}_j} g(u) = 0$，得$\left(\boldsymbol{\psi} - \lambda_j \boldsymbol{I}\right)\boldsymbol{u}_j = 0, \ j = d+1, \cdots, \infty$
  - 如令$d = 0$，则得$\boldsymbol{\psi}\boldsymbol{u}_j = \lambda_j \boldsymbol{u}_j, \ e = \sum_{j=d+1}^{\infty} \lambda_j, \ j = 1, \cdots, \infty$

  即：

  - 用**K-L变换的产生矩阵$\boldsymbol{\psi}$**的前d个本征值（从大到小排序）对应的本征向量作为基来展开$\boldsymbol{x}$时，截断误差在所有d维正交坐标展开中是最小的
  - $\boldsymbol{u}_j, \ j = 1, \cdots, d$张成了新特征空间，展开系数$c_j = \boldsymbol{u}_j^T \boldsymbol{x}$组成新特征向量

9.4.1 K-L 变换

- **K-L展开式的性质**
  - 信号的最佳（压缩）表达：均方误差最小
  - 新空间中的特征是互不相关
  - 用K-L变换坐标系表示原数据，表示熵最小
    - 即这种坐标系统下，样本的方差信息最大程度地集中在较少的维数上
  - 用本征值最小的K-L变换坐标来表示数据，总体熵最小
    - 本征值大的本征向量代表的是样本集中变化大的方向，即方差大的方向
    - 本征值小的本征向量对应样本分布集中的地方，这项方向方差小，均值可以更好的代表样本

9.4.2 用于监督模式识别的K-L变换

- 样本中没有类别信息时
  - 常用（二阶矩矩阵）$\boldsymbol{\psi} = E\left[\boldsymbol{x}\boldsymbol{x}^T\right]$
  - 如果去掉均值，可以用数据的协方差矩阵 $\boldsymbol{\Sigma} = E\left[\left(\boldsymbol{x}-\boldsymbol{\mu}\right)\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^T\right]$
- 样本类别已知时，三种典型策略**计算K-L坐标系**
  - 从类均值中提取判别信息
  - 包含在类平均向量中判别信息的最优压缩
  - 类中心化特征向量中分类信息的提取

9.4.2 用于监督模式识别的K-L变换

- **策略1：从类均值中提取判别信息**
  - 出发点
    - 消除特征各分量之间的相关性
    - 考查变换后各特征的类均值和方差，选择方差小、类均值与总体均值差别大的特征
  - 步骤
    - 用**总类内离散度矩阵**作为K-L展开的产生矩阵：$\boldsymbol{S}_w = \sum_{i=1}^{c} P_i \boldsymbol{\Sigma}_i$
    - 用S_w作K-L变换，得到本征值λ_i和本征向量$\boldsymbol{u}_i$；新特征y_i、各维新特征的方差λ_i，$i = 1, \cdots, D$

9.4.2 用于监督模式识别的K-L变换

- **策略1：从类均值中提取判别信息**
  - 步骤
    - 计算新特征各个分量的分类性能指标：

$$J(y_i) = \frac{\boldsymbol{u}_i^T \boldsymbol{S}_b \boldsymbol{u}_j}{\lambda_i},\ j = 1,\cdots,D$$

其中，$\boldsymbol{S}_b = \sum_{i=1}^{c} P(\omega_i)\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}\right)\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}\right)^T$为类间离散度矩阵，

用$J(y_i)$排序：$J(y_1) \geq J(y_2) \geq \cdots \geq J(y_d) \geq \cdots \geq J(y_D)$，选择前$d$个分量组成新的特征向量，相应的$\boldsymbol{u}_j$组成变换阵$\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_d]$

9.4.2 用于监督模式识别的K-L变换

- **策略1：从类均值中提取判别信息**

下面给出一个实例。设有一个两类问题，两类的先验概率相等，特征为二维向量，类均值向量分别为

$$\boldsymbol{\mu}_1 = [4,2]^{\mathrm{T}}$$
$$\boldsymbol{\mu}_2 = [-4,-2]^{\mathrm{T}}$$

协方差矩阵分别是

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 4 & 2 \\ 2 & 4 \end{bmatrix}$$

为了把维数从 2 压缩为 1，首先求 $\boldsymbol{S}_w$

$$\boldsymbol{S}_w = \frac{1}{2}\boldsymbol{\Sigma}_1 + \frac{1}{2}\boldsymbol{\Sigma}_2 = \begin{bmatrix} 3.5 & 1.5 \\ 1.5 & 3.5 \end{bmatrix}$$

它的本征值矩阵和本征向量分别是

$$\boldsymbol{\Lambda} = \begin{bmatrix} 5 & 0 \\ 0 & 2 \end{bmatrix}, \quad \boldsymbol{U} = \begin{bmatrix} 0.707 & 0.707 \\ 0.707 & -0.707 \end{bmatrix}$$

令

$$\boldsymbol{S}_b = \begin{bmatrix} 16 & 8 \\ 8 & 4 \end{bmatrix}$$

可计算得

$$J(x_1) = 3.6, \quad J(x_2) = 1$$

因此选 $\boldsymbol{u}_1 = [0.707, 0.707]^{\mathrm{T}}$ 作为一维的新特征，如图 8-3 所示。

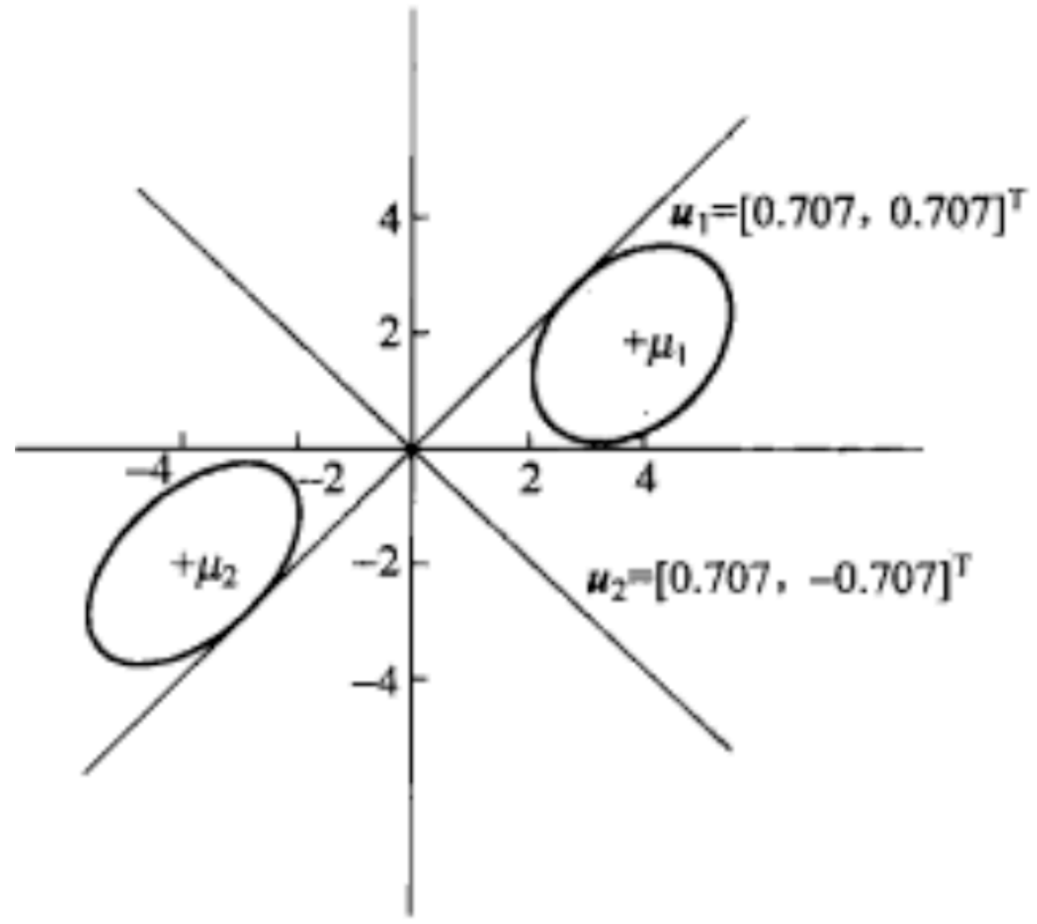


图 8-3　从类均值中提取判别信息的 K-L 变换举例

9.4.2 用于监督模式识别的K-L变换

- **策略2：包含在类平均向量中判别信息的最优压缩**
  - 出发点
    - 用最少的维数来保持原空间中类平均向量中的信息
    - 使特征间互不相关的前提下，最优压缩类均值向量中包含的分类信息
  - 步骤
    - 用**总类内离散度矩阵**$\boldsymbol{S}_w$做K-L变换，消除相关性 $\boldsymbol{U}^T\boldsymbol{S}_w\boldsymbol{U} = \Lambda$，令 $\boldsymbol{B} = \boldsymbol{U}\boldsymbol{\Lambda}^{-1/2}$，从而使 $\boldsymbol{B}^T\boldsymbol{S}_w\boldsymbol{B} = \boldsymbol{I}$（白化变化，之后再进行任何正交归一变换，类内离散度矩阵不变），变换后的类间离散度矩阵：$\boldsymbol{S}_b' = \boldsymbol{B}^T\boldsymbol{S}_b\boldsymbol{B}$

9.4.2 用于监督模式识别的K-L变换

- 策略2：包含在类平均向量中判别信息的最优压缩
  - 步骤
    - 用S_b'做K-L变换，以压缩包含在类平均向量中的信息。对于一个c类问题，$rank\left(S_b'\right) \geq c-1$，故最多有$d=c-1$个非零本征值 $\boldsymbol{V}'=\left[\boldsymbol{v}_1, \cdots, \boldsymbol{v}_d\right]$，总的变换阵是：

$$\boldsymbol{W}=\boldsymbol{U}\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{V}'$$

可以证明，两类情况下，这种特征提取得到的新特征方向就是Fisher线性判别器的最佳投影方向。

9.4.2 用于监督模式识别的K-L变换

- 策略2：包含在类平均向量中判别信息的最优压缩

再用上面的同一个例子来说明这种特征提取方法。因为只有两类，所以均值信息最优压缩的特征只有一维。接上面例子，可以得到

$$\boldsymbol{B}=\boldsymbol{U}\boldsymbol{\Lambda}^{-1/2}=\begin{bmatrix}0.707 & 0.707\\ 0.707 & -0.707\end{bmatrix}\begin{bmatrix}0.447 & 0\\ 0 & 0.707\end{bmatrix}$$

$$=\begin{bmatrix}0.316 & 0.5\\ 0.316 & -0.5\end{bmatrix}$$

$$\boldsymbol{S}_b'=\boldsymbol{B}^{\mathrm{T}}\boldsymbol{S}_b\boldsymbol{B}=\begin{bmatrix}3.6 & 1.897\\ 1.897 & 1\end{bmatrix}$$

$\boldsymbol{S}_b'$的本征值矩阵是

$$\boldsymbol{\Lambda}'=\begin{bmatrix}4.6 & 0\\ 0 & 0\end{bmatrix}$$

和非零本征值对应的本征向量是

$$\boldsymbol{v}=\begin{bmatrix}0.884\\ 0.466\end{bmatrix}$$

所以

$$\boldsymbol{w}=\boldsymbol{B}\boldsymbol{v}=\begin{bmatrix}0.512\\ 0.046\end{bmatrix}$$

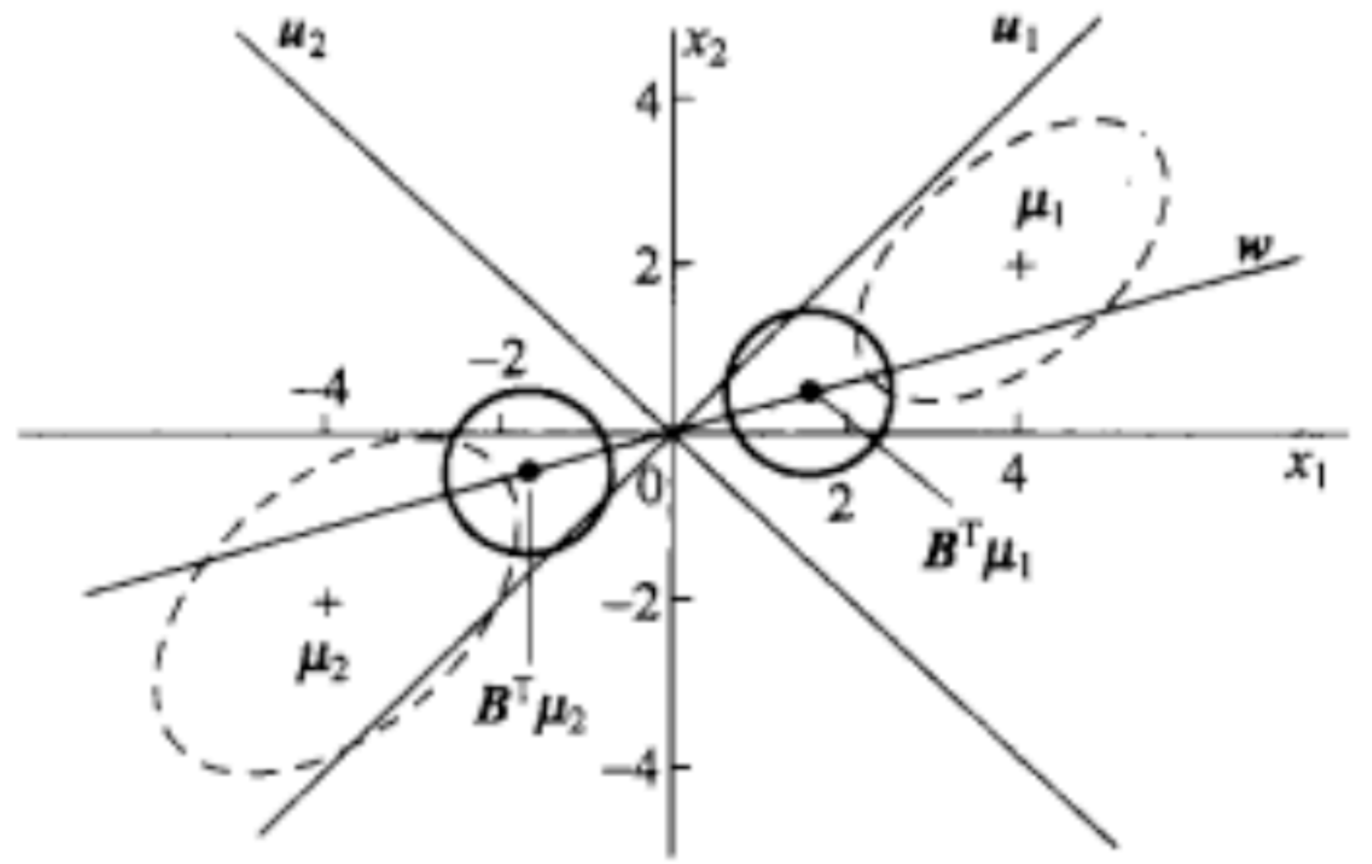


图 8-4 包含在类平均向量中判别信息的最优压缩示例

9.4.2 用于监督模式识别的K-L变换

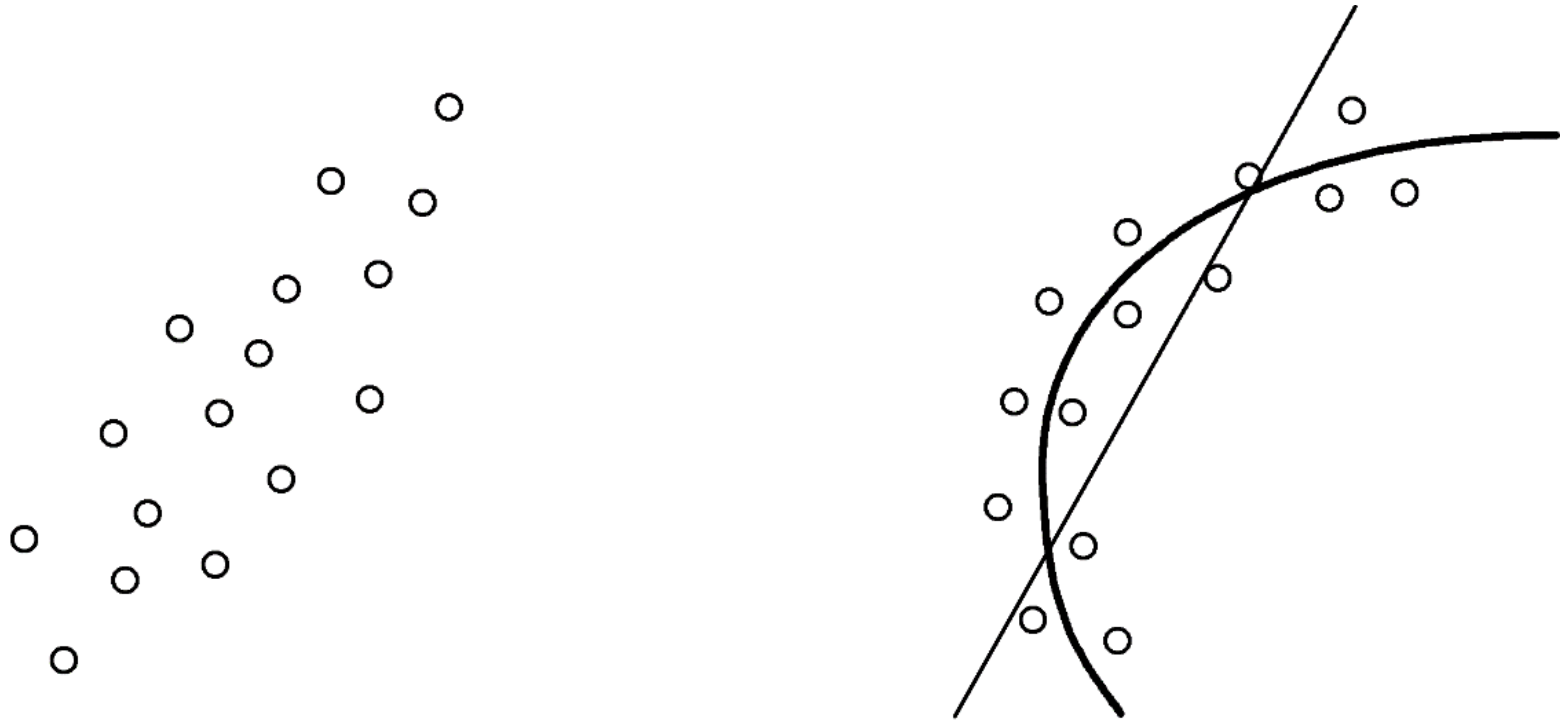
- **策略3：类中心化特征向量中分类信息的提取**
  - 出发点
    - 类中心化特征向量：把原来的样本向量中减去类均值，只考虑各类的协方差中可能包含的分类信息。
  - 步骤
    - 首先，用**总类内离散度矩阵**S_w做K-L变换，消除特征间相关性，考察各个新特征在各类中的方差r_{ij}（第i类第j个特征的方差）；归一化方差：$\tilde{r}_{ij} = p(\omega_i)\dfrac{r_{ij}}{\lambda_j}, i = 1,\cdots,c, j = 1,\cdots,D$；显然，归一化方差满足：$\sum_{i=1}^{c} \tilde{r}_{ij} = 1$

9.4.2 用于监督模式识别的K-L变换

- **策略3：类中心化特征向量中分类信息的提取**
  - 步骤
    - ✓定义**总体熵**来表示方差的分散程度：$J\left(x_j\right) = -\sum_{i=1}^{c} \widetilde{r}_{ij} \log \widetilde{r}_{ij}$，或：$J\left(x_j\right) = \prod_{i=1}^{c} \tilde{r}_{ij}$
    - ✓把K-L变换的新特征按照$J(x_j)$排序：$J(x_1) \leq J(x_2) \leq \cdots \leq J(x_d) \leq \cdots \leq J(x_D)$，选取其中前$d$个特征组成新的特征
  - 一般情况下，可以
    - ✓用均值分类信息的最优压获得$d' \leq c-1$个特征
    - ✓利用类中心化特征的方差信息获得另外$d-d'$个信息

9.5 非线性特征变换方法简介

- 进行特征提取和数据压缩，实际上是假定数据在高维空间中是沿着一定的方向分布的，这些方向能够用较小的维数来表示
- 采用线性变换进行特征提取是假设这种方向是线性的
- 但在某些情况下，数据可能会按照非线性规律分布，要提前这种规律，就要采用非线性变换



(a) 线性主轴　　(b) 非线性主轴

9.5.1 核主成分分析（KPCA）

- **基本思想**
  - 对样本进行非线性主成分分析
  - 根据**可再生希尔伯特空间**的性质，在变换空间中的**协方差矩阵**可以通过原空间中的核函数进行运算，从而绕过了复杂的非线性变换
- **步骤**
  - 通过核函数计算矩阵

$$K_{ij} = (\phi(\boldsymbol{x}_i \cdot \phi(\boldsymbol{x}_j))) = k(\boldsymbol{x}_i \cdot \boldsymbol{x}_j)$$

其中，n为样本数，x_i, x_j是原空间中的样本，$k(\cdot, \cdot)$与支持向量机中类似的核函数。$\phi(\cdot)$是非线性变换（无需知道具体形式或进行运算）

9.5.1 核主成分分析（KPCA）

- **步骤**
  - 解矩阵K的特征方程

$$\frac{1}{n}K\alpha = \lambda\alpha$$

▸并将得到的归一化本征向量$\boldsymbol{\alpha}^l, l = 1,2,\cdots$按照对应的本征值从大到小排列。本征向量的维数是$n$，向量的元素记为$\alpha^l = \left[\alpha_1^l, \alpha_2^l, \cdots, \alpha_n^l\right]$。根据需要选择前若干个本征值对应的本征向量作为非线性主成分。

▸第l个非线性主成分是：$v^l = \sum_{i=1}^{n} \alpha_i^l \phi(x_i)$，$\phi(x_i)$未知

9.5.1 核主成分分析（KPCA）

- **步骤**
  - 计算样本**在非线性主成分上的投影**。样本$\boldsymbol{x}$在第l个非线性主成分上的投影为

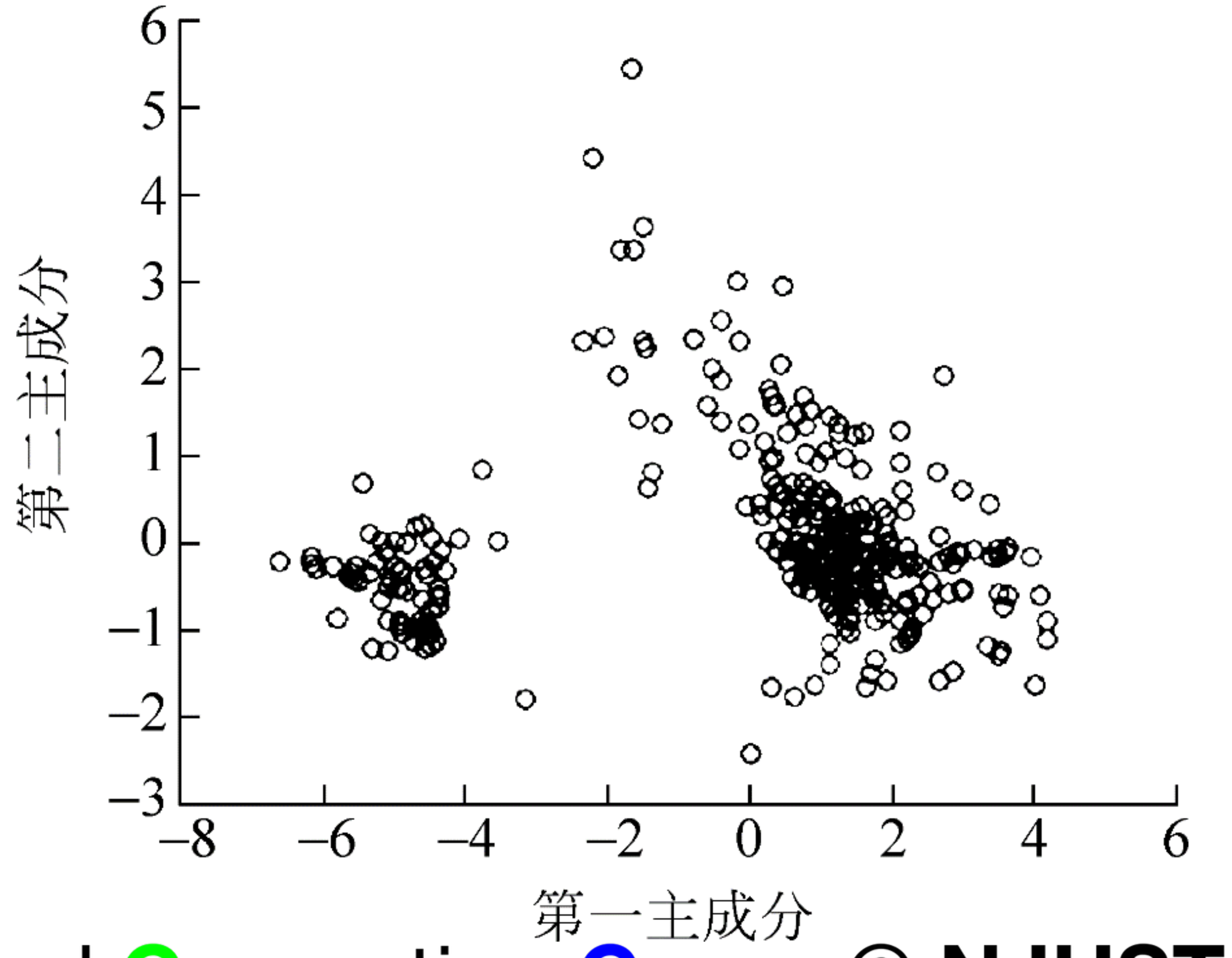
$$z^l(\boldsymbol{x}) = \left(\boldsymbol{v}^l \cdot \phi(\boldsymbol{x})\right) = \sum_{i=1}^{n} \alpha_i^l k(\boldsymbol{x}_i, \boldsymbol{x})$$

如果选择m个非线性主成分，则样本$\boldsymbol{x}$在前m个非线性主成分上的坐标构成样本在新空间的表示

$$\left[z^1(\boldsymbol{x}), \cdots, z^m(\boldsymbol{x})\right]^T$$

9.6 高维数据的低维可视化

- **定义**
  - 将高维空间的数据映射到二维平面来，而这种映射尽可能要反映原空间中样本的分布情况，或者使各样本间的距离关系保持不变。
- **举例**
  - 主成分分析
    - 利用第一、第二主成分构成二维平面



9.7 t-SNE降维可视化方法

- **t-SNE（t-distribution stochastic neighbor embedding，t分布随机近邻嵌入法）**
  - 本质是基于流形学习（manifold learning）的降维方法，即寻找高维数据中可能存在的低维流形
  - 利用**概率分布来度量样本间的距离**，将高维空间中的欧式距离转化为条件概率密度函数来表示样本间的相似度
  - 特点是能够保持样本间的局部结构，使得在高维数据中距离相近的点投影到低维中仍然相近
  - 常用作样本可视化分析
  - 作者网站：https://lvdmaaten.github.io/tsne/，各种实现、案例