



大数据挖掘与统计学习

软件工程系
文化遗产数字化国家地方工程联合中心
可视化技术研究所
张海波
讲师/博士(后)



统计（机器）学习方法 分类

Supervised learning

Unsupervised learning

Semi-supervised learning

Reinforcement learning

三要素

方法 = 模型 + 策略 + 算法

模型

决策函数的集合: $\mathcal{F} = \{f \mid Y = f(X)\}$

参数空间 $\mathcal{F} = \{f \mid Y = f_{\theta}(X), \theta \in \mathbf{R}^n\}$

条件概率的集合: $\mathcal{F} = \{P \mid P(Y \mid X)\}$

参数空间 $\mathcal{F} = \{P \mid P_{\theta}(Y \mid X), \theta \in \mathbf{R}^n\}$

策略

损失函数：一次预测的好坏

风险函数：平均意义下模型预测的好坏

0-1损失函数 0-1 loss function

$$L(Y, f(X)) = \begin{cases} 1, & Y \neq f(X) \\ 0, & Y = f(X) \end{cases}$$

平方损失函数 quadratic loss function

$$L(Y, f(X)) = (Y - f(X))^2$$

绝对损失函数 absolute loss function

$$L(Y, f(X)) = |Y - f(X)|$$



对数损失函数 logarithmic loss function 或对数似然损失函数 loglikelihood loss function

$$L(Y, P(Y | X)) = -\log P(Y | X)$$

损失函数的期望

$$R_{\text{exp}}(f) = E_P[L(Y, f(X))] = \int_{\mathcal{X} \times \mathcal{Y}} L(y, f(x)) P(x, y) dx dy$$

风险函数 risk function 期望损失 expected loss

由 $P(x, y)$ 可以直接求出 $P(x|y)$,但不知道,

$$T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$$

经验风险 empirical risk , 经验损失 empirical loss

$$R_{\text{emp}}(f) = \frac{1}{N} \sum_{i=1}^N L(y_i, f(x_i))$$

经验风险最小化最优模型

$$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^N L(y_i, f(x_i))$$

当样本容量很小时，经验风险最小化学习的效果未必很好，会产生“过拟合over-fitting”

结构风险最小化 structure risk minimization，为防止过拟合提出的策略，等价于正则化（regularization），加入正则化项regularizer，或罚项 penalty term:

$$R_{\text{em}}(f) = \frac{1}{N} \sum_{i=1}^N L(y_i, f(x_i)) + \lambda J(f)$$

求最优模型就是求解最优化问题:

$$\longrightarrow \min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^N L(y_i, f(x_i)) + \lambda J(f)$$

算法

如果最优化问题有显式的解析式，算法比较简单
但通常解析式不存在，就需要数值计算的方法

解析解，是指通过严格的公式所求得解。即包含分式、三角函数、指数、对数甚至无限级数等基本函数的解的形式。

数值解，是采用某种计算方法，如有限元的方法，数值逼近，插值的方法，得到的解。别人只能利用数值计算的结果，而不能随意给出自变量并求出计算值。



经典降维方法（特征提取）：

SVD, TSVD, PCA, K-L, LDA等；

流形学习：

KPCA、LLE、LPP、ISOMAP等；

聚类：

k-means聚类、DBSCAN聚类、OPTICS聚类、AP聚类等



矩阵的奇异值分解-SVD

应用领域

1. 最优化问题;
 特征值问题;
 最小二乘问题;
 广义逆矩阵问题等.
2. 统计分析;
 信号与图像处理;
 系统理论和控制等.

矩阵的正交对角分解

若 A 是 n 阶实对称矩阵，则存在正交矩阵 Q ，使得

$$Q^T A Q = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n) \quad (1)$$

其中 $\lambda_i (i = 1, 2, \dots, n)$ 为矩阵 A 的特征值，而 Q 的 n 个列向量组成 A 的一个完备的标准正交特征向量系。

对于实的非对称矩阵 A ，不再有像式（1）的分解，但却存在两个正交矩阵 P 和 Q ，使 $P^T A Q$ 为对角矩阵，即有下面的**正交对角分解定理**。



定理 设 $A \in R^{n \times n}$ 非奇异, 则存在正交矩阵 P 和 Q ,
使得 $P^T A Q = \text{diag}(\alpha_1, \alpha_2, \dots, \alpha_n)$ (2)

其中 $\alpha_i > 0 (i = 1, 2, \dots, n)$

证 因为 A 非奇异, 所以 $A^T A$ 为实对称正定矩阵, 于是存在正交矩阵 Q 使得, $Q^T (A^T A) Q = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$

其中 $\lambda_i > 0 (i = 1, 2, \dots, n)$ 为 $A^T A$ 特征值

令 $\alpha_i = \sqrt{\lambda_i} (i = 1, 2, \dots, n)$, $\Lambda = \text{diag}(\alpha_1, \alpha_2, \dots, \alpha_n)$



则有 $Q^T (A^T A) Q = \Lambda^2$

或者 $(AQ \Lambda^{-1})^T AQ = \Lambda$

再令 $P = AQ \Lambda^{-1}$, 于是有

$$P^T P = (AQ \Lambda^{-1})^T (AQ \Lambda^{-1}) = I$$

即 P 为正交矩阵, 且使

$$P^T A Q = \Lambda = \text{diag}(\alpha_1, \alpha_2, \dots, \alpha_n)$$

改写式(2)为

$$A = P \cdot \text{diag}(\alpha_1, \alpha_2, \dots, \alpha_n) \cdot Q^T \quad (3)$$

称式(3)为 **正交矩阵 A 的正交对角分解**



引理:

1. 设 $A \in C_r^{m \times n} (r > 0)$, 则 $A^T A$ 是对称矩阵,
且其特征值是非负实数.
2. $\text{rank}(A^T A) = \text{rank} A$
3. 设 $A \in C_r^{m \times n} (r > 0)$, 则 $A = 0$ 的充要条件是

$$A^T A = 0$$



定义 设 A 是秩为 r 的 $m \times n$ 实矩阵, $A^T A$
的特征值为 $\lambda_i (i = 1, 2, \dots, r)$

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r > \lambda_{r+1} = \dots = \lambda_n = 0$$

则称 $\sigma_i = \sqrt{\lambda_i} (i = 1, 2, \dots, r)$ 为 A 的奇异值.

奇异值分解定理

设 A 是秩为 $r(r > 0)$ 的 $m \times n$ 实矩阵,

则存在 m 阶正交矩阵 U 与 n 阶正交矩阵 V ,

使得

$$U^T A V = \begin{bmatrix} \Sigma & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix} = S \quad \textcircled{1}$$

其中 $\Sigma = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_r)(i = 1, 2, \dots, r)$

$\sigma_1 \geq \dots \geq \sigma_r > 0$ 为矩阵 A 的全部奇异值.

证明 设实对称矩阵 A 的特征值为

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > \lambda_{r+1} = \cdots = \lambda_n = 0$$

则存在 n 阶正交矩阵 V ，使得

$$V^T(A^T A)V = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} = \begin{bmatrix} \Sigma^2 & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix} \quad (2)$$

将 V 分块为

$$V = (V_1 \quad V_2)$$

其中 V_1, V_2 分别是 V 的前 r 列与后 $n-r$ 列.

并改写②式为

$$\mathbf{A}^T \mathbf{A} \mathbf{V} = \mathbf{V} \begin{bmatrix} \boldsymbol{\Sigma}^2 & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix}$$

则有

$$\mathbf{A}^T \mathbf{A} \mathbf{V}_1 = \mathbf{V}_1 \boldsymbol{\Sigma}^2, \quad \mathbf{A}^T \mathbf{A} \mathbf{V}_2 = \mathbf{O} \quad \text{③}$$

由③的第一式可得

$$\mathbf{V}_1^T \mathbf{A}^T \mathbf{A} \mathbf{V}_1 = \boldsymbol{\Sigma}^2, \quad \text{或者} (\mathbf{A} \mathbf{V}_1 \boldsymbol{\Sigma})^T (\mathbf{A} \mathbf{V}_1 \boldsymbol{\Sigma}) = \mathbf{E}_r$$

由③的第二式可得

$$(\mathbf{A} \mathbf{V}_2)^T (\mathbf{A} \mathbf{V}_2) = \mathbf{O} \quad \text{或者} \mathbf{A} \mathbf{V}_2 = \mathbf{O}$$

令 $\mathbf{U}_1 = \mathbf{A} \mathbf{V}_1 \boldsymbol{\Sigma}^{-1}$, 则 $\mathbf{U}_1^T \mathbf{U}_1 = \mathbf{E}_r$, 即 \mathbf{U}_1 的 r 个列是两两正交的单位向量. 记



$$\mathbf{U}_1 = (\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_r)$$

因此可将 $\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_r$ 扩充成 \mathbb{C}^m 的标准正交基, 记增添的向量为 $\mathbf{u}_{r+1}, \cdots, \mathbf{u}_m$, 并构造矩阵

$$\mathbf{U}_2 = (\mathbf{u}_{r+1}, \cdots, \mathbf{u}_m)$$

则

$$\mathbf{U} = (\mathbf{U}_1, \mathbf{U}_2) = (\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_r, \mathbf{u}_{r+1}, \cdots, \mathbf{u}_m)$$

是 m 阶正交矩阵, 且有

$$\mathbf{U}_1^T \mathbf{U}_1 = \mathbf{E}_r, \quad \mathbf{U}_2^T \mathbf{U}_2 = \mathbf{O}$$

于是可得

$$\mathbf{U}^T \mathbf{A} \mathbf{V} = \mathbf{U}^T (\mathbf{A} \mathbf{V}_1, \mathbf{A} \mathbf{V}_2) = \begin{bmatrix} \mathbf{U}_1^T \\ \mathbf{U}_2^T \end{bmatrix} (\mathbf{U}_1 \boldsymbol{\Sigma}, \mathbf{O}) = \begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix}$$



$$A = U \begin{bmatrix} \Sigma & O \\ O & O \end{bmatrix} V^T = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^T + \cdots + \sigma_r \mathbf{u}_r \mathbf{v}_r^T$$

称上式为矩阵 A 的奇异值分解.

注意

在矩阵理论中，奇异值分解实际上是

“对称矩阵正交相似于对角矩阵”的推广。

奇异值分解中

$$\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_r, \mathbf{u}_{r+1}, \cdots, \mathbf{u}_m$$

是 $\mathbf{A}\mathbf{A}^T$ 的特征向量，而 \mathbf{V} 的列向量是 $\mathbf{A}^T\mathbf{A}$ 的特征向量，并且 $\mathbf{A}\mathbf{A}^T$ 与 $\mathbf{A}^T\mathbf{A}$ 的非零特征值完全相同。**但矩阵 \mathbf{A} 的奇异值分解不惟一。**

数值秩

在没有误差时，奇异值分解可以确定矩阵的秩。但是误差的存在使得确定变得非常困难。例如，考虑矩阵

$$\mathbf{A} = \begin{bmatrix} 1/3 & 1/3 & 2/3 \\ 2/3 & 2/3 & 4/3 \\ 1/3 & 2/3 & 1 \\ 2/5 & 1/5 & 3/5 \\ 3/7 & 1/7 & 4/7 \end{bmatrix}$$

因为第三列是前两列的和，所以 A 的秩是2.

如果不考虑到这个关系，
运用IEEE标准的双精度浮点计算模式，
用MATLAB命令SVD计算 A 的奇异值：

$$A = \begin{bmatrix} 1/3 & 1/3 & 2/3 \\ 2/3 & 2/3 & 4/3 \\ 1/3 & 2/3 & 1 \\ 2/5 & 1/5 & 3/5 \\ 3/7 & 1/7 & 4/7 \end{bmatrix}$$

```
format long e
```

```
A=[1/3,1/3,2/3;2/3,2/3,4/3;1/3,2/3,1;2/5,1/5,3/5;3/7,  
1/7,4/7];
```

```
D= svd(A)
```



- 计算结果为:

D =

2.421457493421318e+000

3.406534035359026e-001

1.875146052457622e-016

因为有“三”个非零奇异值，所以A的秩为“3”。然而，注意到在IEEE双精度的标准下，其中一个奇异值是微小的。也许应该将它看作零。因为这个原因，引入数值秩的概念。



如果矩阵 A 有 k 个“大”的奇异值，而其它都很“微小”，则称 A 的数值秩为 k 。

为了确定哪个奇异值是“微小”的，需要引入阈值或容忍度 ε 。就 **MATLAB** 而言，可以把

$$\text{tol} = n\sigma_1\text{eps} \quad (\text{eps} = 2.24 \times 10^{-16})$$

设为阈值，大于这个阈值的奇异值的数目就是 A 的数值秩，把小于这个阈值的奇异值看作零。

利用 **MATLAB** 的命令 **rank** 计算 A 的秩，它的结果是 2，就是这个道理。

求矩阵 $A = \begin{bmatrix} 0 & -1.6 & 0.6 \\ 0 & 1.2 & 0.8 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ 的奇异值分解

解: MATLAB程序为:

```
A=[0,-1.6,0.6;0 ,1.2,0.8;0,0,0;0,0,0]
```

```
[U,S,V]=svd(A)
```

计算结果

A =

0	-1.6000	0.6000
0	1.2000	0.8000
0	0	0
0	0	0

U =

0.8000	0.6000	0	0
-0.6000	0.8000	0	0
0	0	1.0000	0
0	0	0	1.0000

S =

2.0000	0	0
0	1.0000	0
0	0	0
0	0	0

V =

0	0	1.0000
-1.0000	0.0000	0
0.0000	1.0000	0



奇异值分解的几何意义

研究将一个空间映射到不同空间，特别是不同维数的空间时，例如超定或欠定方程组所表示的情况，就需要用**矩阵的奇异值**来描述算子对空间的作用了。

考察二维平面上的单位圆 $\{x \in R^2 : \|x\| = 1\}$

在映射A下的变换过程, 其中

$$A = \frac{1}{\sqrt{2}} \begin{bmatrix} \sqrt{3} & \sqrt{3} \\ -3 & 3 \\ 1 & 1 \end{bmatrix}$$

MATLAB程序为:

```
A=[sqrt(3)\sqrt(2),sqrt(3)\sqrt(2);-  
3\sqrt(2),3\sqrt(2); 1\sqrt(2),1\sqrt(2)]  
[U,S,V]=svd(A)
```



```
>> A=[sqrt(3)\sqrt(2), sqrt(3)\sqrt(2); -3\sqrt(2), 3\sqrt(2); 1\sqrt(2), 1\sqrt(2)]  
[U, S, V]=svd(A)
```

To get started, select "MATLAB Help" from the Help menu.

A =

0.8165	0.8165
-0.4714	0.4714
1.4142	1.4142

U =

-0.5000	-0.0000	-0.8660
0	-1.0000	0.0000
-0.8660	0.0000	0.5000

S =

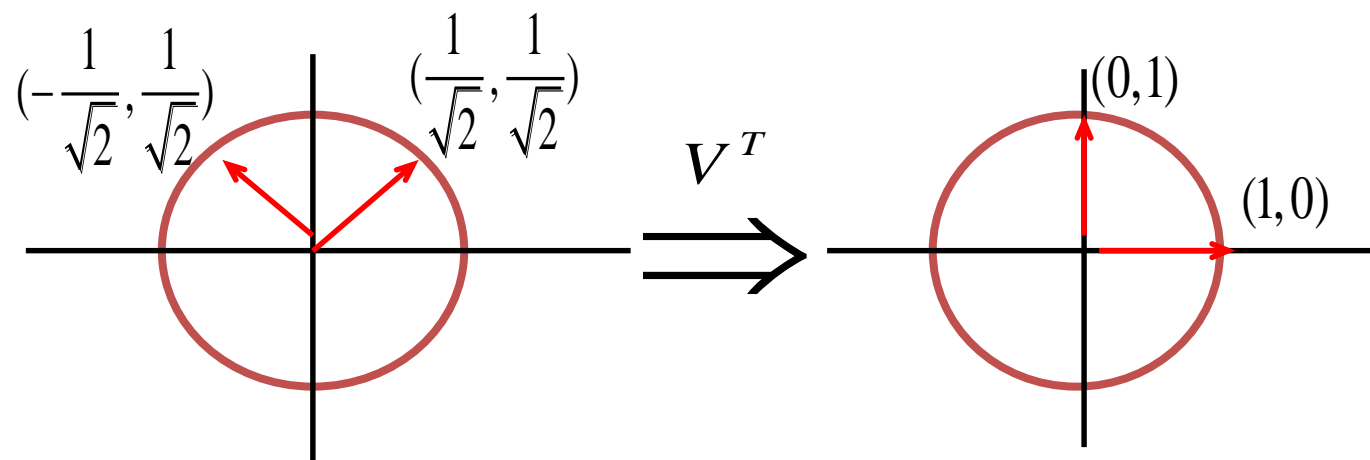
2.3094	0
0	0.6667
0	0

V =

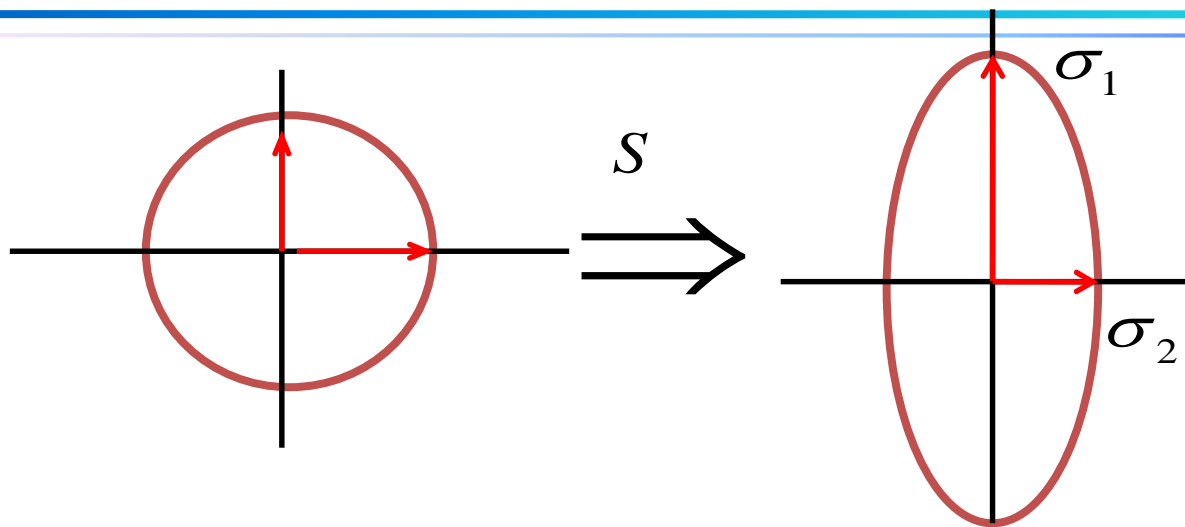
-0.7071	0.7071
-0.7071	-0.7071

```
>>
```

$$V^T = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



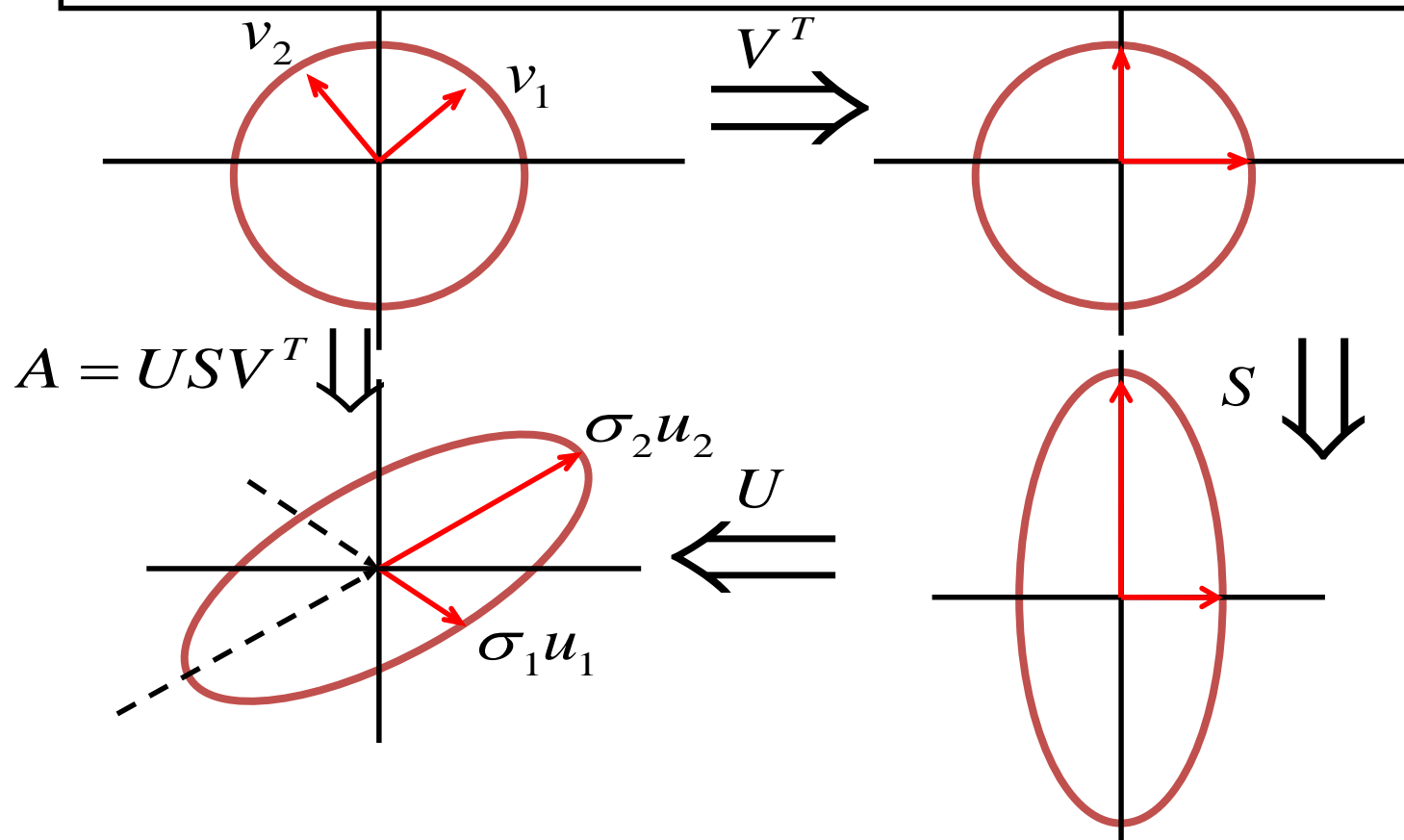
V是正交矩阵，表示二维空间的一个旋转



$$S \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \\ 0 & 0 \end{bmatrix}$$

S 将平面上的圆变换到三维空间坐标平面上的椭圆

V 是正交矩阵，表示二维空间的一个旋转



维空间坐标平面上的椭圆
 S 将平面上的圆变换到三

U 是正交矩阵，表示三维空间的一个旋转

当 A 是方阵时，其奇异值的几何意义是：

若 x 是 n 维单位球面上的一点，则 Ax 是一个 n 维椭球面上的点，其中椭球的 n 个半轴长正好是 A 的 n 个奇异值.

简单地说，在2维情况下， A 将单位圆变成了椭圆， A 的两个奇异值是椭圆的长半轴和短半轴.

奇异值分解的性质

设 \mathbf{A} 是秩为 $r (r > 0)$ 的 $m \times n$ 实矩阵, \mathbf{A} 的奇异值分解为:

$$\mathbf{A} = \mathbf{U} \mathbf{S} \mathbf{V}^T$$

即 $\mathbf{A} \mathbf{V} = \mathbf{U} \mathbf{S}$, 且

$$\mathbf{U} = (\mathbf{u}_1, \cdots, \mathbf{u}_r, \mathbf{u}_{r+1}, \cdots, \mathbf{u}_m)$$



$$S = \left[\begin{array}{ccc|ccc} \sigma_1 & & & & & \\ & \ddots & & & & \\ & & \sigma_r & & & \\ \hline & & & & & \\ & & & & & \\ & & & & & \end{array} \right]_{m \times n}$$

$$V = (\mathbf{v}_1, \cdots, \mathbf{v}_r, \mathbf{v}_{r+1}, \cdots, \mathbf{v}_n)$$

则



- (1) A 的非零奇异值的个数等于它的秩 r , 即
$$\text{rank}(A) = r$$
- (2) $\mathbf{u}_1, \cdots, \mathbf{u}_r$ 是 $C(A)$ 的标准正交基.
- (3) $\mathbf{u}_{r+1}, \cdots, \mathbf{u}_m$ 是 $N(A^T)$ 的标准正交基.
- (4) $\mathbf{v}_1, \cdots, \mathbf{v}_r$ 是 $C(A^T)$ 的标准正交基.
- (5) $\mathbf{v}_{r+1}, \cdots, \mathbf{v}_n$ 是 $N(A)$ 的标准正交基.



从上面的结论可以得到

$$C(A^T) = N(A)^\perp$$

$$C(A) = N(A^T)^\perp$$

$$\dim C(A) + \dim N(A) = n$$

$C(A)$ 与 $C(A^T)$ 同构



奇异值分解的特征

1. 奇异值分解可以降维

A 表示 n 个 m 维向量, 可以通过奇异值分解表示成 $m+n$ 个 r 维向量. 若 A 的秩 r 远远小于 m 和 n , 则通过奇异值分解可以降低 A 的维数. 可以计算出, 当 $r < \frac{mn}{m+n+1}$ 时, 可以达到降维的目的, 同时可以降低计算机对存储器的要求.

2. 奇异值对矩阵的扰动不敏感

特征值对矩阵的扰动敏感.

在数学上可以证明，奇异值的变化不会超过相应矩阵的变化，即对任何的相同阶数的实矩阵A、B的按从大到小排列的奇异值 α_i 和 ϖ_i 有

$$\sum |\alpha_i - \varpi_i| \leq \|A - B\|_2$$



3. 奇异值的比例不变性, 即 αA 的奇异值是 A 的奇异值的 $|\alpha|$ 倍.
4. 奇异值的旋转不变性. 即若 P 是正交阵, PA 的奇异值与 A 的奇异值相同.

奇异值的比例和旋转不变性特征在数字图象的旋转、镜像、平移、放大、缩小等几何变化方面有很好的应用.



5. 容易得到矩阵 A 的秩为 k ($k \leq r$)的一个最佳逼近矩阵.

奇异值的这个特征可以应用于信号的分解和重构, 提取有用信息, 消除信号噪声.

矩阵的秩 k 逼近

由矩阵 A 的奇异值分解

$$\mathbf{u}_1 \mathbf{v}_1^T, \mathbf{u}_2 \mathbf{v}_2^T, \dots, \mathbf{u}_r \mathbf{v}_r^T$$

可见, A 是矩阵

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0$$

的加权和, 其中权系数按递减排列:

$$\mathbf{A} = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^T + \dots + \sigma_r \mathbf{u}_r \mathbf{v}_r^T$$

显然，权系数大的那些项对矩阵A的贡献大

因此当舍去权系数小的一些项后，仍然能较好的矩阵 A，这一点在数字图像处理方面非常有用.

矩阵的秩 k 逼近定义为

$$A = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^T + \cdots + \sigma_k \mathbf{u}_k \mathbf{v}_k^T \quad (1 \leq k \leq r)$$

秩 r 逼近就精确等于A，而秩1逼近的误差最大.



在MATLAB中，秩 k 逼近的程序如下：

```
clear
A=[2,7,9,-5,4;-9,-9,5,3,-2;-2,5,-1,-3,5;-4,9,0,9,-4],
sumA=zeros(4,5);
k=3
[U,D,V]=svd(A);
for i=1:k
    sumA=sumA+ D(i,i)*U(:,i)*V(:,i)';
end
sumA
```



- 或者

clear

A=input('请输入矩阵A的值:A='),

sumA=zeros;

[U,D,V]=svd(A);

k=input('请输入k的值:')

for i=1:k

sumA=sumA+ D(i,i)*U(:,i)*V(:,i)';

end

sumA



File Edit View Web Window Help

To get started, select "MATLAB Help" from the Help menu.

```
>> clear
A=[2,7,9,-5,4;-9,-9,5,3,-2;-2,-2,5,-1,-3;5,-4,9,0,9,-4],
sumA=zeros(4,5);
k=3
[U,D,V]=svd(A);
for i=1:k
    sumA=sumA+ D(i,i)*U(:,i)*V(:,i)';
end
sumA
```

A =

2	7	9	-5	4
-9	-9	5	3	-2
-2	5	-1	-3	5
-4	9	0	9	-4

k =

3

sumA =

0.7887	7.1851	8.1303	-5.3248	4.9945
-8.3051	-9.1062	5.4990	3.1863	-2.5705
1.6726	4.4387	1.6371	-2.0152	1.9846
-4.1409	9.0215	-0.1012	8.9622	-3.8843

Ready



6. 奇异值的第六个特征是若**A**、**B**都有相同的奇异向量，
则

$$\|A - B\|_2 = \sum |\alpha_i - \varpi_i|$$

也就是说，我们可以通过控制奇异值的大小来控制两个矩阵空间的距离。

注：对奇异值按由大到小排列，取一定的阈值舍去部分较小的奇异值，会得到截断奇异值分解方法

主成分分析 (Principal Component Analysis, PCA)

基本思想

在研究中，**多变量问题**是经常会遇到的。变量太多，无疑会增加分析问题的难度与复杂性，而且在许多实际问题中，多个变量之间是具有一定的**相关关系**的。

因此，人们会很自然地想到，能否在相关分析的基础上，用**较少的新变量代替原来较多的旧变量**，而且使这些较少的新变量尽可能多地保留原来变量所反映的信息？



事实上，这种想法是可以实现的，**主成分分析方法**就是综合处理这种问题的一种强有力的工具。

主成分分析是**把原来多个变量划为少数几个综合指标的一种统计分析方法**。从数学角度来看，这是一种**降维**处理技术。



数学模型与几何解释

假设实际问题中有 p 个指标，我们把这 p 个指标看作 p 个随机变量，记为 x_1, x_2, \dots, x_p ，主成分分析就是要把这 p 个指标，转变为讨论 p 个指标的**线性组合**的问题，这些新的指标 $y_1, y_2, \dots, y_k (k \leq p)$ ，



原则：

保留主要信息量的充分反映原指标的信息，并且相互无关。这种由讨论多个指标降为少数几个综合指标的过程在数学上就叫做降维。



主成分分析通常的做法，是寻求原指标的
线性组合 y_i ：

$$y_1 = u_{11}x_1 + u_{21}x_2 + \cdots + u_{p1}x_p$$

$$y_2 = u_{12}x_1 + u_{22}x_2 + \cdots + u_{p2}x_p$$

.....

$$y_p = u_{1p}x_1 + u_{2p}x_2 + \cdots + u_{pp}x_p$$

满足如下的条件：

(1) 每个主成分的系数平方和为1（否则其方差可能为无穷大），即

$$u_1' u_1 = u_{1i}^2 + u_{2i}^2 + \cdots + u_{pi}^2 = 1$$

(2) 主成分之间相互无关，即无重叠的信息。即

$$\text{Cov}(y_i, y_j) = 0, \quad i \neq j, \quad i, j = 1, \cdots, p$$

(3) 主成分的方差依次递减，重要性依次递减，即

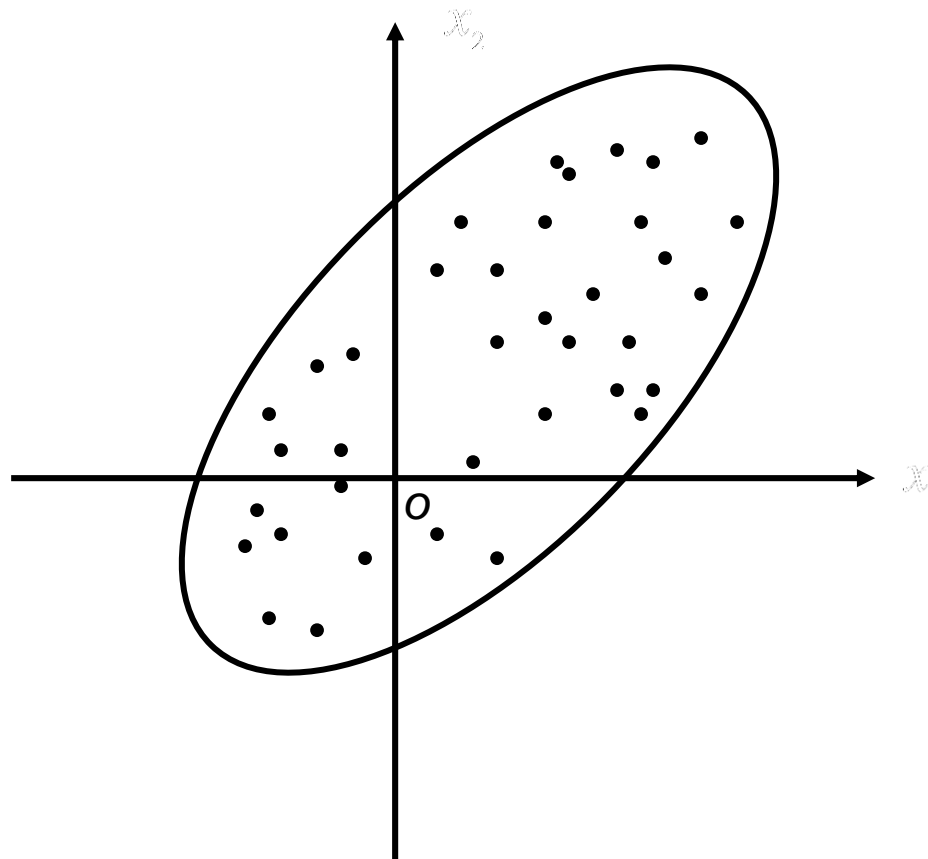
$$\text{Var}(y_1) \geq \text{Var}(y_2) \geq \cdots \geq \text{Var}(y_p)$$



二维空间中主成分的几何意义：设有 n 个样品，每个样品有两个观测变量 x_1 和 x_2 。在由变量 x_1 和 x_2 所确定的二维平面中， n 个样本点所散布的情况如椭圆状。由图可以看出这 n 个样本点无论是沿着 x_1 轴方向或 x_2 轴方向都具有较大的离散性，其离散的程度可以分别用观测变量 x_1 的方差和 x_2 的方差定量地表示。显然，如果只考虑 x_1 和 x_2 中的任何一个，那么包含在原始数据中的信息将会有较大的损失。

平移、旋转坐标轴

主成分分析的几何解释





将 x_1 轴和 x_2 轴先平移，再同时按逆时针方向旋

转 θ 角度，得到新坐标轴 F_1 和 F_2 ，则

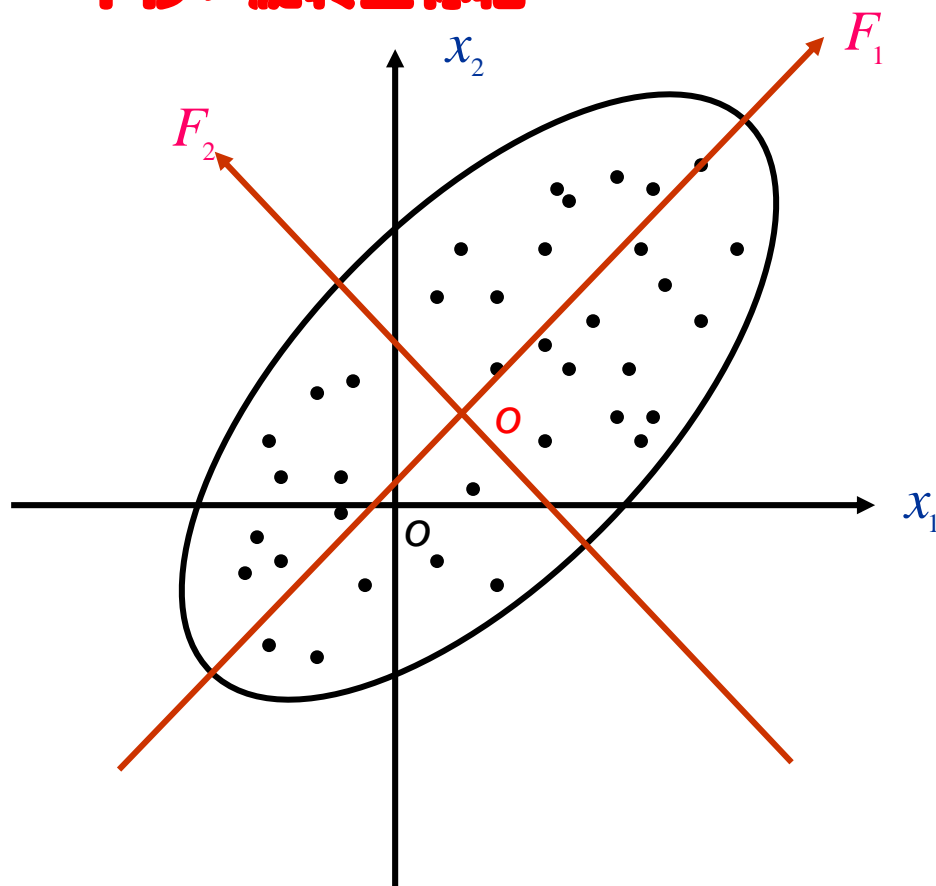
$$\begin{cases} y_1 = x_1 \cos \theta + x_2 \sin \theta \\ y_2 = -x_1 \sin \theta + x_2 \cos \theta \end{cases}$$

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \mathbf{U}' \mathbf{x}$$

\mathbf{U}' 为正交旋转变换矩阵

平移、旋转坐标轴

主成分分析的几何解释





旋转变换的目的是为了使得 n 个样品点在 F_1 轴方向上的离散程度最大，即 y_1 的方差最大。

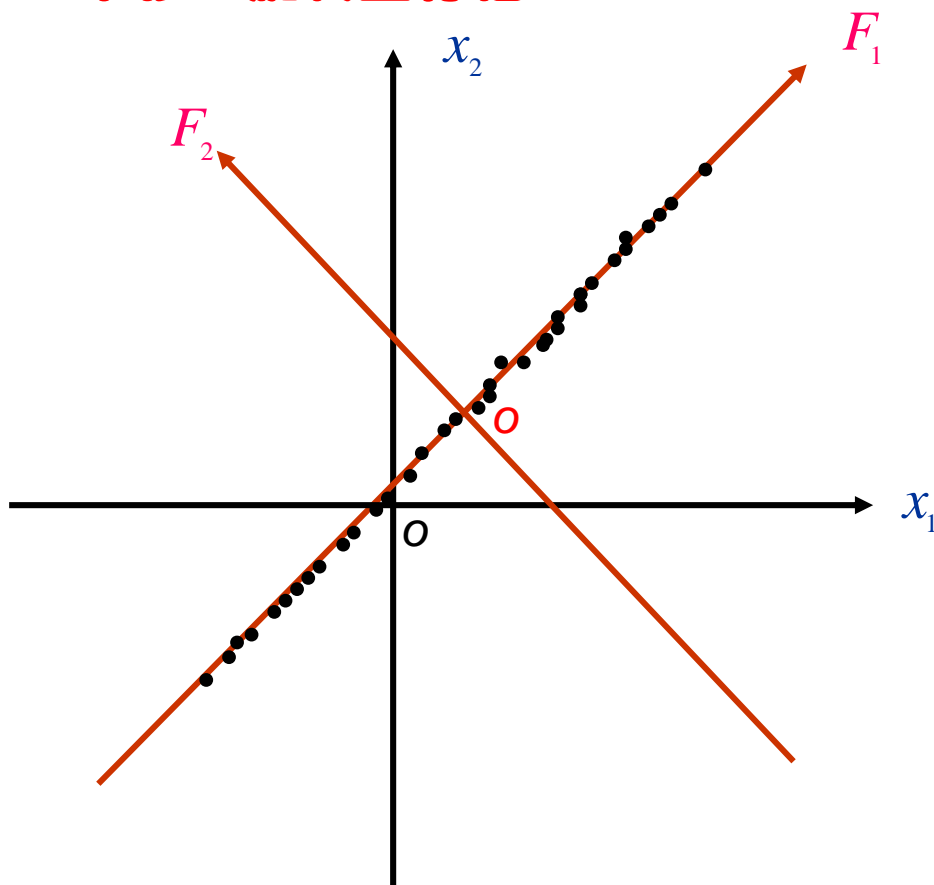
变量 y_1 代表了原始数据的大部分信息，在研究某些实际问题时，即使不考虑变量 y_2 也无损大局。

经过上述旋转变换原始数据的大部分信息集中到 F_1 轴上，对数据中包含的信息起到了浓缩作用。



平移、旋转坐标轴

主成分分析的几何解释





y_1 , y_2 除了可以对包含在 x_1 , x_2 中的信息起着浓缩作用之外, 还具有**不相关**的性质, 这就使得在研究复杂的问题时避免了信息重叠所带来的虚假性。

二维平面上的各点的**方差大部分都归结在 F_1 轴上**, 而 F_2 轴上的方差很小。

**y_1 和 y_2 称为原始变量 x_1 和 x_2 的综合变量。
 F 简化了系统结构, 抓住了主要矛盾。**



主成分的推导及性质

一、两个线性代数的结论

1、若A是p阶实对称阵，则一定可以找到正交阵U，
使

$$U^{-1}AU = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{bmatrix}$$

其中 $\lambda_i, i=1.2.\cdots p$ 是A的特征根。



2、若上述矩阵的特征根所对应的单位特征向量为 $\mathbf{u}_1, \dots, \mathbf{u}_p$

$$\text{令 } \mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_p) = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1p} \\ u_{21} & u_{22} & \cdots & u_{2p} \\ \vdots & \vdots & & \vdots \\ u_{p1} & u_{p2} & \cdots & u_{pp} \end{bmatrix}$$

则U是正交矩阵，即有

$$\mathbf{U}'\mathbf{U} = \mathbf{U}\mathbf{U}' = \mathbf{I}$$



主成分的推导

(一) 第一主成分

设 x 的协方差阵为

$$\Sigma_x = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{bmatrix}$$

由于 Σ_x 为非负定的对称阵，所以存在正交阵 U ，使得

$$U' \Sigma_x U = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{bmatrix}$$

其中 $\lambda_1, \dots, \lambda_p$ 为 Σ_x 的特征根，不妨假设 $\lambda_1 \geq \dots \geq \lambda_p$ 。

U是由特征根相对应的特征向量所组成的正交阵：

$$U = (u_1, \dots, u_p) = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1p} \\ u_{21} & u_{22} & \cdots & u_{2p} \\ \vdots & \vdots & & \vdots \\ u_{p1} & u_{p2} & \cdots & u_{pp} \end{bmatrix}$$

$$u_i = \left(u_{1i}, u_{2i}, \dots, u_{pi} \right)' \quad i = 1, 2, \dots, P$$



下面证明，由U的第一列元素所构成的原始变量的线性组合有最大的方差。

设有P维单位向量 $a_1 = (a_{11}, a_{21}, \dots, a_{p1})'$

$$y_1 = a_{11}x_1 + a_{21}x_2 + \dots + a_{p1}x_p = a_1'x$$

$$D(y_1) = a_1' \Sigma a_1 = a_1' U \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p \end{bmatrix} U' a_1$$



$$\begin{aligned} &= a_1' \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_p \end{bmatrix} \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p \end{bmatrix} \begin{bmatrix} \mathbf{u}'_1 \\ \mathbf{u}'_2 \\ \vdots \\ \mathbf{u}'_p \end{bmatrix} a_1 \\ &= \sum_{i=1}^p \lambda_i a_1' \mathbf{u}_i \mathbf{u}'_i a_1 \leq \lambda_1 \sum_{i=1}^p a_1' \mathbf{u}_i \mathbf{u}'_i a_1 \\ &= \lambda_1 a_1' \mathbf{U} \mathbf{U}' a_1 = \lambda_1 a_1' a_1 = \lambda_1 \end{aligned}$$

当 $a_1 = u_1$ 时, $y_1 = u_{11}x_1 + \cdots + u_{p1}x_p$, 且

$$\text{Var}(y_1) = u_1' \Sigma_x u_1 = \lambda_1.$$

所以当且仅当 $a_1 = u_1$ 时, y_1 有最大的方差 λ_1 .

y_1 称为第一主成分。

如果第一主成分的信息不够, 则需要寻找
第二主成分。



(二) 第二主成分

在约束条件 $\text{cov}(y_1, y_2) = 0$ 下，寻找第二主成分

$$y_2 = a_{12}x_1 + \cdots + a_{p2}x_p = a_2'x$$

因为 $\text{cov}(y_1, y_2) = \text{cov}(u_1'x, a_2'x)$
 $= u_1'\Sigma a_2 = \lambda_1 u_1' a_2 = 0$

所以 $a_2' u_1 = 0$



于是，对任意的 p 维向量 a_2 ，有

$$\begin{aligned} V(y_2) &= a_2' \Sigma a_2 = \sum_{i=1}^p \lambda_i a_2' u_i u_i' a_2 \\ &= \sum_{i=1}^p \lambda_i (a_2' u_i)^2 \leq \lambda_2 \sum_{i=2}^p (a_2' u_i)^2 \\ &= \lambda_2 \sum_{i=1}^p a_2' u_i u_i' a_2 = \lambda_2 a_2' U U' a_2 \\ &= \lambda_2 a_2' a_2 = \lambda_2 \end{aligned}$$



所以如果取线性变换：

$$y_2 = u_{12}x_1 + u_{22}x_2 + \cdots + u_{p2}x_p$$

则 y_2 的方差为 λ_2 次大，并且 y_1 和 y_2 线性无关。

类似地，可以得到方差逐步减少的 p 个线性无关的主成分：



小结：方差逐步减少的 p 个线性无关的主成分为

$$y_1 = u_{11}x_1 + u_{21}x_2 + \cdots + u_{p1}x_p$$

$$y_2 = u_{12}x_1 + u_{22}x_2 + \cdots + u_{p2}x_p$$

.....

$$y_p = u_{1p}x_1 + u_{2p}x_2 + \cdots + u_{pp}x_p$$

写为矩阵形式: $y = U'x$

$$U = (u_1, \dots, u_p) = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1p} \\ u_{21} & u_{22} & \cdots & u_{2p} \\ \vdots & \vdots & & \vdots \\ u_{p1} & u_{p2} & \cdots & u_{pp} \end{bmatrix}$$

$$x = (x_1, x_2, \dots, x_p)'$$



主成分的性质

1、均值 $E\mathbf{y} = E(\mathbf{U}'\mathbf{x}) = \mathbf{U}'\boldsymbol{\mu}$

2、原总体的总方差（或称为总惯量）等于不相关的主成分的方差之和

$$\sum_{i=1}^p \text{Var}(x_i) = \sum_{i=1}^p \sigma_{ii} = \sum_{i=1}^p \lambda_i$$

若存在 $m < p$ ，使得 $\sum_{i=1}^p \sigma_{ii} \approx \sum_{i=1}^m \lambda_i$ ，则 p 个原始变量所提供的总信息（总方差）的绝大部分只需用前 m 个主成分来代替。



3. 主成分 y_k 与原始变量 x_i 之间的相关系数 $\rho(y_k, x_i)$ 称为因子负荷量（或因子载荷量），并且

$$\rho(y_k, x_i) = \frac{\sqrt{\lambda_k} u_{ik}}{\sqrt{\sigma_{ii}}} \quad (k, i = 1, 2, \dots, p)$$

证明：

$$\rho(y_k, x_i) = \frac{\text{Cov}(y_k, x_i)}{\sqrt{\lambda_k} \sqrt{\sigma_{ii}}} = \frac{\text{Cov}(u'_k \mathbf{x}, \mathbf{e}'_i \mathbf{x})}{\sqrt{\lambda_k} \sqrt{\sigma_{ii}}},$$

其中 $\mathbf{e}'_i = (0, \dots, 0, 1, 0, \dots, 0)$ 。于是

$$\begin{aligned} \text{Cov}(u'_k \mathbf{x}, \mathbf{e}'_i \mathbf{x}) &= u'_k \text{Cov}(\mathbf{x}, \mathbf{x}) \mathbf{e}_i \\ &= u'_k \Sigma \mathbf{e}_i = \mathbf{e}'_i \Sigma u_k = \lambda_k \mathbf{e}'_i u_k = \lambda_k u_{ik} \end{aligned}$$

4、贡献率与累积贡献率

1) **贡献率**：第 i 个主成分的方差在全部方差中所占比重，称 $\lambda_i / \sum_{i=1}^p \lambda_i$ 为第 i 个主成分的贡献率，反映了第 i 个指标提供多大的信息，有多大的综合能力。

2) **累积贡献率**：前 k 个主成分共有多大的综合能力，用这 m 个主成分的方差和在全部方差中所占比重

$$\sum_{i=1}^m \lambda_i / \sum_{i=1}^p \lambda_i$$

来描述，称为累积贡献率。



累积贡献率大小反映 m 个主成分提取了 x_1, x_2, \dots, x_p

的多少信息，但没有表达某个变量被提取了多少信息，为此引入下述概念。

3) 将前 m 个主成分 y_1, y_2, \dots, y_m 对原始变量 x_i 的贡献率定义为 x_i 与 y_1, y_2, \dots, y_m 之间的相关系数的平方，即

$$v_i^{(m)} = \sum_{k=1}^m \frac{\lambda_k u_{ik}^2}{\sigma_{ii}}$$



4) 主成分个数的选择

进行主成分分析的目的之一是简化数据结构，用尽可能少的主成分 y_1, y_2, \dots, y_m ($m < p$) 代替原来的 p 个指标。在实际工作中，主成分个数的选取通常有两个标准

一个是按累积贡献率达到一定的程度(如 70%或 80% 以上) 来确定 m ；另一个先计算协方差矩阵或相关矩阵的特征值的均值 $\bar{\lambda}$ ，取大于 $\bar{\lambda}$ 的特征值的个数作为 m 。

大量实践表明，当 $p < 20$ 时，第一个标准容易取太多的主成分，第二个标准容易取太少的主成分，故最好将两者结合起来使用，并考虑 m 个主成分对 x_i 的贡献率。



例：设 x_1, x_2, x_3 的协方差矩阵为

$$\Sigma = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

解得特征根为

$$\lambda_1 = 5.83 \quad \lambda_2 = 2.00 \quad \lambda_3 = 0.17$$

相应的正交特征向量为

$$u_1 = \begin{bmatrix} 0.383 \\ -0.924 \\ 0.000 \end{bmatrix} \quad u_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \quad u_3 = \begin{bmatrix} 0.924 \\ 0.383 \\ 0.000 \end{bmatrix}$$

第一个主成分的贡献率为 $5.83 / (5.83 + 2.00 + 0.17) = 72.875\%$ ，尽管第一个主成分的贡献率并不小，但在本题中第一主成分不含第三个原始变量的信息，所以应该取两个主成分。



5 标准化变量的主成分及其性质

在实际问题中，不同的变量往往具有不同的量纲，为了消除由于量纲的不同可能带来的不合理的影响，常采用将变量标准化的方法。

记 $Ex_i = \mu_i$, $Var(x_i) = \sigma_i^2$ ，令

$$x_i^* = \frac{x_i - \mu_i}{\sigma_i}, i = 1, 2, \dots, p,$$

则标准化后的随机变量 $\mathbf{x}^* = (x_1^*, x_2^*, \dots, x_p^*)'$ 的协方差阵 Σ^* 就是原随机向量 \mathbf{x} 的相关阵 R 。从 R 出发求得的主成分 $\mathbf{y}^* = (y_1^*, y_2^*, \dots, y_p^*)'$ ，有与总体主成分相同的性质。

样本的主成分

在实际问题中，总体的协方差阵通常是未知的，需要由样本方差阵估计。记样本观测阵为

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \mathbf{x}'_{(1)} \\ \mathbf{x}'_{(2)} \\ \cdots \\ \mathbf{x}'_{(n)} \end{pmatrix}$$

$$\bar{x}_i = \frac{1}{n} \sum_{l=1}^n x_{li}, i = 1, 2, 3 \cdots p$$

则样本协方差阵和样本相关阵分别为

$$S = (s_{ij})_{p \times p} = \left(\frac{1}{n-1} \sum_{l=1}^n (x_{li} - \bar{x}_i)(x_{lj} - \bar{x}_j) \right)_{p \times p}$$

$$R = (r_{ij})_{p \times p} = \left(\frac{s_{ij}}{\sqrt{s_{ii}} * \sqrt{s_{jj}}} \right)_{p \times p}$$



一、样本主成分

假定每个变量的观测数据都已经标准化（标准化后的观测阵仍记为 X ），此时的样本协方差阵就是样本相关阵 $R = \frac{1}{(n-1)}X'X$ 。

仍记相关阵 R 的 p 个主成分为： y_1, y_2, \dots, y_p ；

特征值为： $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ ；

相应的单位正交特征向量为： u_1, u_2, \dots, u_p ；

$U = (u_1, u_2, \dots, u_p)$ 为正交矩阵。



1. 主成分得分阵

将第 t 个样品 $\mathbf{x}_{(t)} = (x_{t1}, x_{t2}, \dots, x_{tp})'$ 代入第 i 个样本主成分 $y_i = \mathbf{u}_i' \mathbf{x}$ 中, 经计算得到的值称为第 t 个样品在第 i 个样本主成分的得分, 记为 y_{it} , 即 $y_{it} = \mathbf{u}_i' \mathbf{x}_{(t)}$ 。

由此, 可构造如下的样本主成分得分阵:

$$Y = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{pmatrix} = \begin{pmatrix} y'_{(1)} \\ y'_{(2)} \\ \cdots \\ y'_{(n)} \end{pmatrix}$$



注意到

$$y_{(t)} = (y_{t1}, y_{t2}, \dots, y_{tp})' = U'x_{(t)} \quad (t = 1, 2, \dots, n)$$

所以，主成分得分阵和标准化后的原始观测阵之间满足：

$$Y = XU, \text{ or } X = YU'$$



由此可知新的综合变量（主成分） Y_1, Y_2, \dots, Y_p 彼此不相关，并且 Y_i 的方差为 λ_i ，则
 $Y_1 = \gamma_1' X, Y_2 = \gamma_2' X, Y_p = \gamma_p' X$ 分别称为第一、第二、
.....、第 p 个主成分。由上述求主成分的过程可知，主成分在几何图形中的方向实际上就是 R 的特征向量的方向，关于主成分分析的几何意义在前面已经讨论过；主成分的方差贡献就等于 R 的相应特征值。这样，我们在利用样本数据求解主成分的过程实际就转化为求相关阵或协方差阵的特征值和特征向量的过程。

主成分分析计算步骤

❖一、主成分的计算

表1 n个指标取值的一组样本数据

指标 样本	X_1	X_2	\dots	X_n
1	Y_{11}	Y_{12}	\dots	Y_{1n}
2	Y_{21}	Y_{22}	\dots	Y_{2n}
\dots	\dots	\dots	\dots	\dots
m	Y_{m1}	Y_{m2}	\dots	Y_{mn}

1. 对样本进行标准化处理

- **数据标准化**首先是**无量纲化**，因为不同指标的量纲通常是不完全相同的，为了使各指标之间具有可比性，必须消除指标的量纲。其次，数据的原始样本不一定满足 $E(\mathbf{X})=0$ ，因此必须对原始样本数据进行标准化处理，**以便使样本数据量纲为一，并且满足 $E(\mathbf{X})=0$ 。**

- 标准化处理的计算式为：

$$X_{ij} = \frac{Y_{ij} - \bar{Y}_j}{S_j}$$

$$\bar{Y}_j = \frac{1}{m} \sum_{i=1}^m Y_{ij} (j = 1, 2, \dots, n)$$

$$S_j^2 = \frac{1}{m-1} \sum_{i=1}^m (Y_{ij} - \bar{Y}_j)^2 (j = 1, 2, \dots, n)$$



- 经标准化处理后可得到标准化矩阵:

$$X = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ X_{m1} & X_{m2} & \cdots & X_{mn} \end{bmatrix}$$

2. 计算相关系数，得到相关矩阵

- 计算标准化后的每两个指标间的相关关系，得到**相关系数矩阵*****R***，即***n***个指标的协方差矩阵。即

$$R = \frac{1}{m-1} X'X = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{bmatrix}$$

$$r_{ij} = \frac{1}{m-1} \sum_{k=1}^m X_{ik} X_{jk} \quad (i, j = 1, 2, \dots, n)$$

3. 计算矩阵 R 的特征根及相应的特征向量

$$\begin{vmatrix} r_{11} - \lambda_1 & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} - \lambda_2 & \cdots & r_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} - \lambda_n \end{vmatrix} = 0$$



- 于是得到n个非负特征根

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0$$

- 从而得到对应于特征根的n个单位化特征向量，构成一个正交矩阵，记为 a ，则

$$a = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}$$

- a_{ij} 中的 i 为第 i 个主分量， j 为第 j 个分量。

4.计算主成分

- 对于m个样本中的第k个样本，根据
则可得n个主成分如下

$$\begin{bmatrix} Z_{k1} \\ Z_{k2} \\ \dots \\ Z_{kn} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} X_{k1} \\ X_{k2} \\ \dots \\ X_{kn} \end{bmatrix}$$

对于全部的 m 个样本，则有

$$\begin{bmatrix} z_{11} & z_{21} & \cdots & z_{m1} \\ z_{12} & z_{22} & \cdots & z_{m2} \\ \cdots & \cdots & \cdots & \cdots \\ z_{1n} & z_{2n} & \cdots & z_{mn} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} X_{11} & X_{21} & \cdots & X_{m1} \\ X_{12} & X_{22} & \cdots & X_{m2} \\ \cdots & \cdots & \cdots & \cdots \\ X_{1n} & X_{2n} & \cdots & X_{mn} \end{bmatrix}$$

- 即： $Z_0^T = aX_0^T$
- 整理得： $Z_0 = X_0 a^T$
- 式中 Z_0 —样本主成分， X_0 —标准化的样本。



二、样本主成分选择及原指标对主成分回归

- 1. 主成分选择

$\frac{\lambda_k}{\sum_{i=1}^n \lambda_i}$ — 第 k 个主成分的贡献率；

$\frac{\sum_{i=1}^r \lambda_i}{\sum_{i=1}^n \lambda_i}$ — 前 r 个主成分的累计贡献率。

2.原指标对主成分的回归

- 原指标对主成分的回归问题即为在 $X = BZ$ 中如何确定回归系数矩阵 B 的问题。
- 将 $Z = aX$ 两端分别左乘 a^T 变为 $X = a^T Z$ ，即得回归系数 $B = a^T$ 。
- 当取前 r 个主成分时， $X = a^T Z$ 为

$$\begin{bmatrix} X_1 \\ X_2 \\ \dots \\ X_n \end{bmatrix} = \begin{bmatrix} a_{11} & a_{21} & \dots & a_{r1} \\ a_{12} & a_{22} & \dots & a_{r2} \\ \dots & \dots & \dots & \dots \\ a_{1n} & a_{2n} & \dots & a_{rn} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ \dots \\ z_r \end{bmatrix}$$



三、主成分分析应注意的问题

- 由数理统计中的大数定理可知，随着样本容量的增大，它们的平均水平和离散程度将会趋于稳定，从而协方差矩阵也会趋于稳定，因此，主成分分析适宜于大样本容量的因素分析。一般来说，要求样本容量应大于指标个数的两倍（即 $m > 2n$ ）。

基于K-L变换的多类模式特征提取

特征提取的目的：

对一类模式：维数压缩。

对多类模式：维数压缩，突出类别的可分性。

卡洛南-洛伊（Karhunen-Loeve）变换（K-L变换）：

- * 一种常用的特征提取方法；
- * 最小均方误差意义下的最优正交变换；
- * 适用于任意的概率密度函数；
- * 在消除模式特征之间的相关性、突出差异性方面有最优的效果。

分为： 连续K-L变换 离散K-L变换

1. K-L展开式

设 $\{\mathbf{X}\}$ 是 n 维随机模式向量 \mathbf{X} 的集合, 对每一个 \mathbf{X} 可以用确定的完备归一化正交向量系 $\{\mathbf{u}_j\}$ 中的正交向量展开:

$$\mathbf{X} = \sum_{j=1}^{\infty} a_j \mathbf{u}_j \quad a_j: \text{随机系数};$$


用有限项估计 \mathbf{X} 时:

$$\hat{\mathbf{X}} = \sum_{j=1}^d a_j \mathbf{u}_j$$

引起的均方误差:


$$\xi = E[(\mathbf{X} - \hat{\mathbf{X}})^T (\mathbf{X} - \hat{\mathbf{X}})]$$

代入 $\mathbf{X}, \hat{\mathbf{X}}$, 利用 $\mathbf{u}_i^T \mathbf{u}_j = \begin{cases} 1, & j = i \\ 0, & j \neq i \end{cases}$



$$\xi = E\left[\sum_{j=d+1}^{\infty} a_j^2\right]$$

$$\xi = E\left[\sum_{j=d+1}^{\infty} a_j^2\right]$$

由 $\mathbf{X} = \sum_{j=1}^{\infty} a_j \mathbf{u}_j$ 两边  左乘 \mathbf{u}_j^T 得 $a_j = \mathbf{u}_j^T \mathbf{X}$

$$\xi = E\left[\sum_{j=d+1}^{\infty} \mathbf{u}_j^T \mathbf{X} \mathbf{X}^T \mathbf{u}_j\right]$$

$$= \sum_{j=d+1}^{\infty} \mathbf{u}_j^T E[\mathbf{X} \mathbf{X}^T] \mathbf{u}_j$$

$$= \sum_{j=d+1}^{\infty} \mathbf{u}_j^T \mathbf{R} \mathbf{u}_j$$

R: 自相关矩阵。

\mathbf{u}_j 为确定性向量

不同的 $\{\mathbf{u}_j\}$ 对应不同的均方误差， \mathbf{u}_j 的选择应使 ξ 最小。

利用拉格朗日乘数法求使 ξ 最小的正交系 $\{\mathbf{u}_j\}$ ，令

$$g(\mathbf{u}_j) = \sum_{j=d+1}^{\infty} \mathbf{u}_j^T \mathbf{R} \mathbf{u}_j - \sum_{j=d+1}^{\infty} \lambda_j (\mathbf{u}_j^T \mathbf{u}_j - 1) \quad \lambda_j: \text{拉格朗日乘数}$$

$$g(\mathbf{u}_j) = \sum_{j=d+1}^{\infty} \mathbf{u}_j^T \mathbf{R} \mathbf{u}_j - \sum_{j=d+1}^{\infty} \lambda_j (\mathbf{u}_j^T \mathbf{u}_j - 1)$$

用函数 $g(\mathbf{u}_j)$ 对 \mathbf{u}_j 求导，并令导数为零，得

$$(\mathbf{R} - \lambda_j \mathbf{I}) \mathbf{u}_j = 0 \quad j = d + 1, \dots, \infty$$

——正是矩阵 \mathbf{R} 与其特征值和对应特征向量的关系式。

说明：当用 \mathbf{X} 的自相关矩阵 \mathbf{R} 的特征值对应的特征向量展开 \mathbf{X} 时，截断误差最小。

选前 d 项估计 \mathbf{X} 时引起的均方误差为

$$\xi = \sum_{j=d+1}^{\infty} \mathbf{u}_j^T \mathbf{R} \mathbf{u}_j = \sum_{j=d+1}^{\infty} \text{tr}[\mathbf{u}_j \mathbf{R} \mathbf{u}_j^T] = \sum_{j=d+1}^{\infty} \lambda_j$$

λ_j 决定截断的均方误差， λ_j 的值小，那么 ξ 也小。

因此，当用 \mathbf{X} 的正交展开式中前 d 项估计 \mathbf{X} 时，展开式中的 \mathbf{u}_j 应当是前 d 个较大的特征值对应的特征向量。

K-L变换方法:

对 R 的特征值由大到小进行排队:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq \lambda_{d+1} \geq \cdots$$

均方误差最小的 \mathbf{X} 的近似式:

$$\mathbf{X} = \sum_{j=1}^d a_j \mathbf{u}_j \quad \text{—— K-L展开式}$$

矩阵形式:

$$\mathbf{X} = \mathbf{U} \mathbf{a} \quad (1)$$

式中, $\mathbf{a} = [a_1, a_2, \cdots, a_d]^T$ $\mathbf{U}_{n \times d} = [\mathbf{u}_1, \cdots, \mathbf{u}_j, \cdots, \mathbf{u}_d]$

其中: $\mathbf{u}_j = [u_{j1}, u_{j2}, \cdots, u_{jn}]^T$

$$\mathbf{U}^T \mathbf{U} = \begin{bmatrix} \mathbf{u}_1^T \\ \mathbf{u}_2^T \\ \cdots \\ \mathbf{u}_d^T \end{bmatrix} [\mathbf{u}_1 \quad \mathbf{u}_2 \quad \cdots \quad \mathbf{u}_d] = \mathbf{I}$$

对式(1)两边左乘 \mathbf{U}^T :

$$\mathbf{a} = \mathbf{U}^T \mathbf{X} \quad \text{—— K-L变换}$$

系数向量 \mathbf{a} 就是变换后的模式向量。

2. 利用自相关矩阵的K-L变换进行特征提取

设 \mathbf{X} 是 n 维模式向量, $\{\mathbf{X}\}$ 是来自 M 个模式类的样本集, 总样本数目为 N 。将 \mathbf{X} 变换为 d 维 ($d < n$) 向量的方法:

第一步: 求样本集 $\{\mathbf{X}\}$ 的总体自相关矩阵 \mathbf{R} 。

$$\mathbf{R} = E[\mathbf{X}\mathbf{X}^T] \approx \frac{1}{N} \sum_{j=1}^N \mathbf{X}_j \mathbf{X}_j^T$$

第二步: 求 \mathbf{R} 的特征值 λ_j , $j = 1, 2, \dots, n$ 。对特征值由大到小进行排队, 选择前 d 个较大的特征值。

第三步: 计算 d 个特征值对应的特征向量 \mathbf{u}_j , $j = 1, 2, \dots, d$, 归一化后构成变换矩阵 \mathbf{U} 。

$$\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_d]$$

第四步: 对 $\{\mathbf{X}\}$ 中的每个 \mathbf{X} 进行 K-L 变换, 得变换后向量 \mathbf{X}^* :

$$\mathbf{X}^* = \mathbf{U}^T \mathbf{X}$$

d 维向量 \mathbf{X}^* 就是代替 n 维向量 \mathbf{X} 进行分类的模式向量。

利用K-L变换进行特征提取的优点：

- 1) 变换在均方误差最小的意义下使新样本集 $\{X^*\}$ 逼近原样本集 $\{X\}$ 的分布，既压缩了维数又保留了类别鉴别信息。
- 2) 变换后的新模式向量各分量相对总体均值的方差等于原样本集总体自相关矩阵的大特征值，表明变换突出了模式类之间的差异性。

$$C^* = E\{(X^* - M^*)(X^* - M^*)^T\} = \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_d \end{bmatrix}$$

- 3) C^* 为对角矩阵说明了变换后样本各分量互不相关，亦即消除了原来特征之间的相关性，便于进一步进行特征的选择。



K-L变换的不足之处:

- 1) 对两类问题容易得到较满意的结果。类别愈多，效果愈差。
- 2) 需要通过足够多的样本估计样本集的协方差矩阵或其它类型的散布矩阵。当样本数不足时，矩阵的估计会变得十分粗略，变换的优越性也就不能充分地显示出来。



总结

- 主成分分析PCA
 - Principle Component Analysis
- 通过K-L变换实现主成分分析

PCA的变换矩阵是协方差矩阵，**K-L**变换的变换矩阵可以有很多种（二阶矩阵、协方差矩阵、总类内离散度矩阵等等）。当**K-L**变换矩阵为协方差矩阵时，等同于**PCA**。

课下思考：**SVD**和**PCA**的区别与联系？

- **K-L变换特征提取思想**
 - 用映射（或变换）的方法把原始特征变换为较少的新特征
 - 降维
- **主成分分析(PCA)基本思想**
 - 进行特征降维变换，不能完全地表示原有的对象，能量总会有损失。
 - 希望找到一种能量最为集中的的变换方法使损失最小