第8章 基于K-L变换的特征提取

基于Karhunen-Loeve变换的特征提取

- K-L变换又称主分量分析，是一种正交变换，K-L变换常用来作为数据压缩和降维，学习这一节主要要掌握以下几个问题：
  - ✓1．什么是正交变换，这是在数字信号处理或其它课学习过的内容，很重要。
  - ✓2．K-L变换是一种最佳的正交变换，要弄清是什么意义的最佳，也就是说它最佳的定义。
  - ✓3．K-L变换的重要应用。

- **变换：**
  - ✓ 是把给定的测量集变换为新的特征集，如果变换方法选择的合适，那么变换域的特征与原始的输入样本相比将具有很高的信息压缩性能。这就意味着大多数与分类有关的信息被压缩到相对小的特征中，从而减少了必需的空间维数。
- **线性PCA：**
  - ✓ 一D维的列向量X，经过一个线性变换y=Ax,变成d维的向量y，d<<D,但y可以保留x分布上面的主要信息。
  - ✓ 原理：随机向量的正交归一化变换。把x变为y*,y*也是D维，但是协方差矩阵成了对角阵，表示各维独立，而且按照x协方差阵的本征值从大到小排序。本征值越小对应y*的维的分布就越趋近于一点，信息就越小。它们不是主分量，而是次要分量，可以抛弃，剩下的d维就是y了。

一、Karhunen-Loeve变换

➢ 正交变换概念

- ✓ 变换是一种工具，它的用途归根结底是用来描述事物，特别是描述信号用的;
- ✓ 描述事物的基本方法之一是将复杂的事物化成简单事物的组合, 或对其进行分解，分析其组成的成分;
- ✓ 变换的实质是一套度量用的工具，如用大尺子度量大的东西，用小尺子度量小的东西;
- ✓ 对某一套完整的工具就称为某种变换，如付里叶变换就是用一套随时间正弦、余弦信号作为度量工具，这些正弦，余弦信号的频率是各不相同的，才能度量出信号中相应的不同频率成分。

- 图1的信号只有一个单一频率的简谐信号

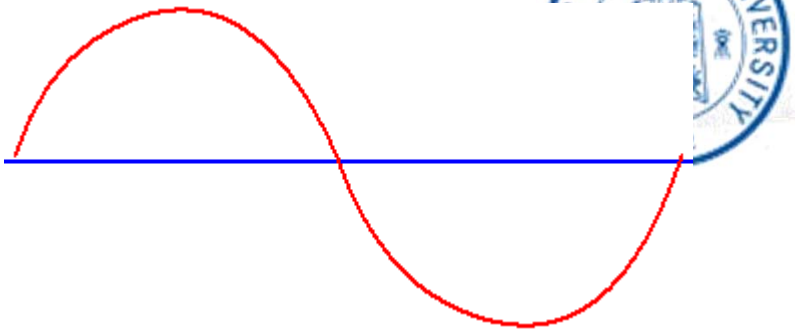


图1

- 图2(a)信号就不是一个简谐波号所描述的，它起码可以分解成图2(b)中的两个成分，一是基波，另一是三次谐波。

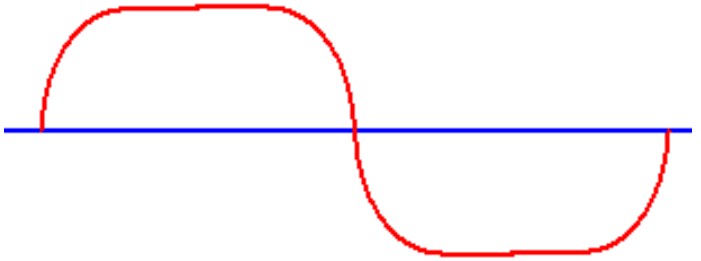


图2(a)

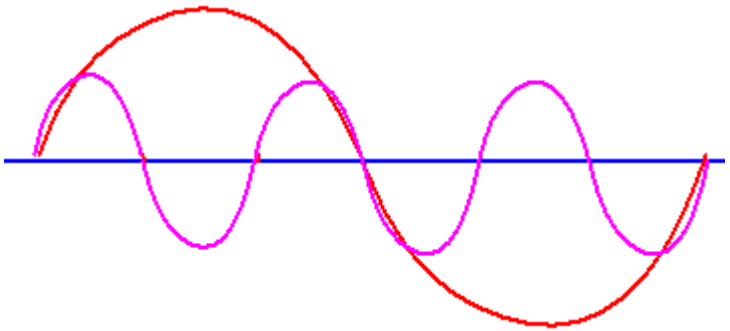


图2(b)

- 为了对复杂事物进行经济有效的描述，我们希望将其分解成相互独立的成分
- 用变换对信号进行分析，所使用的数学工具是点积。点积的实质就是两个信号中相同成分之间乘积之总和。

$$\int_{-\infty}^{\infty} F(t)G(t)dt$$

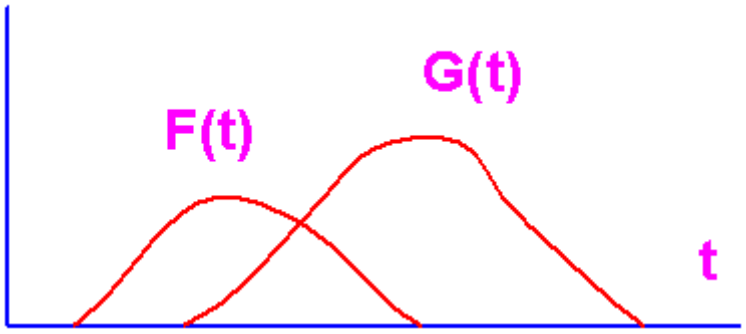


图3

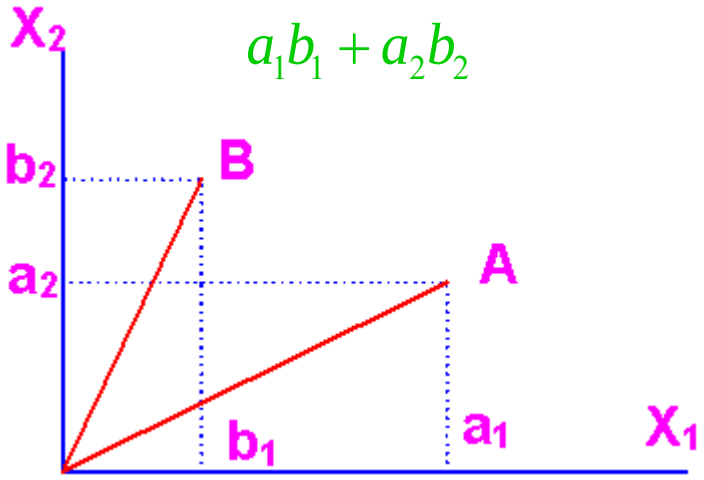
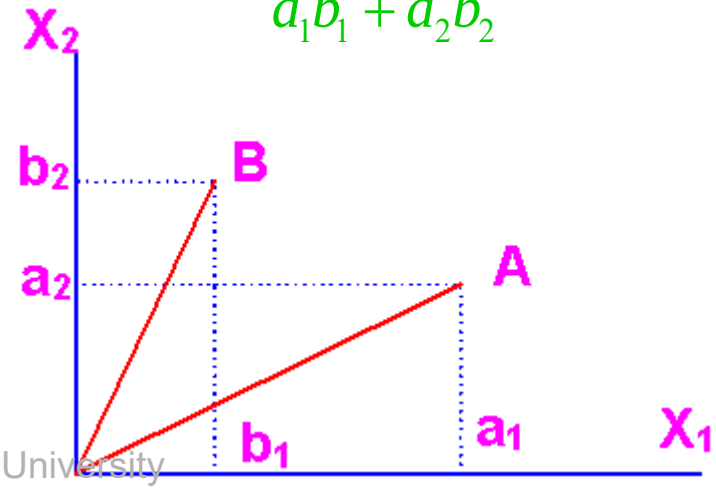


图4

- 点积运算的结果是一个数值，或大于零，小于零或等于零，等于零的情况在图中出现在A与B之间夹角为90°的情况，这表明B中没有A的成分，A中也没有B的成分，因此又称相互正交。由此知道作为一种变换，如果这种变换中的每一种成分与其它成分都正交时，它们之间的关系就相互独立了，每一种成分的作用是其它成分所不能代替的。如付里叶变换，频率为f的成分只能靠变换频率为f的成分去析取。另一方面也说明了这套变换必须是完备的，也就是它必须包含一切必要的成分。

$$a_1b_1 + a_2b_2$$



- 将这种变换中的每一成分，用一个向量 u_i 表示，i是下标，则正交变换可表示为：

$$u_i^T u_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

$$X = \sum_{i=1}^{\infty} c_i u_i$$

其中 c_i 是相应基 u_i 的相应成分

- 以样本特征向量在特征空间分布为原始数据，通过实行**Karhunen-Loeve**变换，找到维数较少的组合特征，达到降维的目的。由于样本的描述都是离散的向量，因此只讨论**Karhunen-Loeve**变换(以后称**K-L**变换)的离散情况；
- **K-L**变换的最佳：特征空间的降维，原特征空间是**D**维的，现希望降至**d**维**d<D**。要找的正交变换能使一组样本集的截断均方差的期望值为最小；
- 问题：给定一个训练样本集条件下要找一个好的正交变换，能使这种误差从总体上来说是最小。

- 衡量指标：均方误差最小

- **K-L变换是一种正交变换，即将一个向量X，在某一种坐标系统中的描述，转换成用另一种基向量组成的坐标系表示。这组基向量是正交的，其中每个坐标基向量用 u_j表示，j=1,…,∞，因此，一个向量X可表示成**

$$X = \sum_{j=1}^{\infty} c_j u_j$$

- 对一向量或一向量空间进行正交变换，可采用多种不同的正交坐标系，关键在于使用正交变换要达到的目的，不同的要求使用不同的正交变换。这里要讨论的是，如果我们将由上式X表示的无限多维基向量坐标系统改成有限维坐标系近似，即

$$\hat{X} = \sum_{j=1}^{d} c_j u_j$$

- $\hat{X}$表示 X的近似值或估计量，我们希望在同样维数条件下，使向量 X的估计量误差最小。确切地说是使所引起的均方误差为最小

$$\varepsilon = E\left[\left(X - \hat{X}\right)^T \left(X - \hat{X}\right)\right]$$

- 要找 ε 为最小是一个求极值的问题，就是求最佳的正交变换的基 $u_i, i=1,\cdots,\infty$ 。因此这是一个求条件极值的问题，一般方法是利用拉格朗日乘子法将条件数值转换成一个求无条件极值的问题。

- 至于对某一个数据**X**的相应 c_j 值，可以通过**X**与每一个基 u_j的点积来计算。由于不同的基之间是相互正交的，这个点积值就是 c_j 的值，即 $c_j = u_j^T x$。

$$C = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_d \end{pmatrix} = \begin{pmatrix} u_1^T \\ u_2^T \\ \vdots \\ u_d^T \end{pmatrix} X = UX$$

- 则U就是一个变换矩阵，其中每一行是某一个正交基向量的转置。由X计算C称为对X的分解。反过来，如果我们希望用C重构信号X，则根据 $X=\sum_{j=1}^{\infty} c_j u_j$

它是各个成分之和。如果将对应于每个基 u_i 的成分表示成 x_i ，则重构的信号又可表示成一个向量形式 $\hat{X}=(x_1,\cdots,x_d)^T$ 。

则 $$\hat{X}=\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix}=(u_1,u_2,\cdots,u_d)\begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_d \end{pmatrix}=U^T C$$

- 显然，$\hat{X}$ 与原向量X是有差别的，是原向量的一个近似，要使与X的差异越小，则要用更多维数的正交基。

- 如果将$X-\hat{X}=\sum_{j=d+1}^{\infty}c_j u_j$代入$\varepsilon$ 式可得到

$$\varepsilon = E\left[\sum_{j=d+1}^{\infty} c_j u_j^T . \sum_{i=d+1}^{\infty} c_i u_i\right]$$

$$\text{since} \qquad u_i^T u_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

$$\textit{therefore} \qquad \varepsilon = E\left[\sum_{j=d+1}^{\infty} c_j^2\right]$$

- 系数c_j 可以利用正交坐标系的特性得到。如令某一基向量u_j与向量**X**作点积，则有

$$c_j = X^T u_j = u_j^T X$$

- 由正交基特征得：$C_j = u_j^T X$

代入ε 式：

$$\varepsilon = E\left[\sum_{j=d+1}^{\infty} u_j^T X X^T u_j\right] = \sum_{j=d+1}^{\infty} u_j^T E\left[XX^T\right] u_j$$

令 $\Psi = E\left[XX^T\right]$

则 $\varepsilon = \sum_{j=d+1}^{\infty} u_j^T \Psi u_j$

欲使该均方误差ε为最小，就变成在确保正交变换的条件下，使ε达最小的问题，这可用拉格朗日乘子法求解。为此设一函数

$$g(u_j)=\sum_{j=d+1}^{\infty}u_j^T\Psi u_j-\sum_{j=d+1}^{\infty}\lambda_j\left[u_j^Tu_j-1\right]$$

令其对 $u_j, j=d+1,\cdots,\infty$ 求导数，可得：

$$(\Psi-\lambda_j I)u_j=0,\quad j=d+1,\cdots,\infty$$

可见向量 $u_j, j=d+1,\cdots,\infty$ 应是 Ψ 矩阵的特征值的特征向量，而此时截断误差为 $\varepsilon=\sum_{j=d+1}^{\infty}\lambda_j$。如将按其大小顺序排列，即 $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_d\cdots$

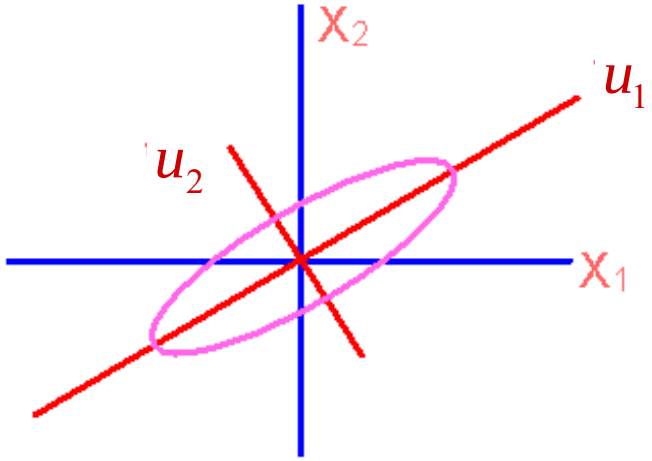
- 则取前**d**项特征值对应的特征向量组成的坐标系，可使向量的均方误差为最小。满足上述条件的变换就是**K-L**变换。

Note:

- K-L变换是一种独特的正交变换，它与一些常用的正交变换不同。最常见的正交变换如傅里叶变换、哈达玛变换、离散余弦变换等都是一种通用的正交变换，它们各自有固定的形式，如傅里叶变换的基是以频率为参数的函数族组成。而K-L变换的基并没有固定的形式，它是从对给定数据 {x}进行计算产生的。换句话说，给定的数据集不同，得到的K-L变换基函数也因此而不同。正是因为它对给定数据集{x}存在依赖关系，它能在降低维数时仍能较好地描述数据，因此是模式识别中降低特征空间维数的有效方法。但是由于它的正交基函数族是从训练样本集中计算出来的，因此并不存在一种对任何数据都适用的K-L变换基，一般的作法是先用一组训练数据计算出K-L变换基，然后用这组基来重构或分析其它数据。

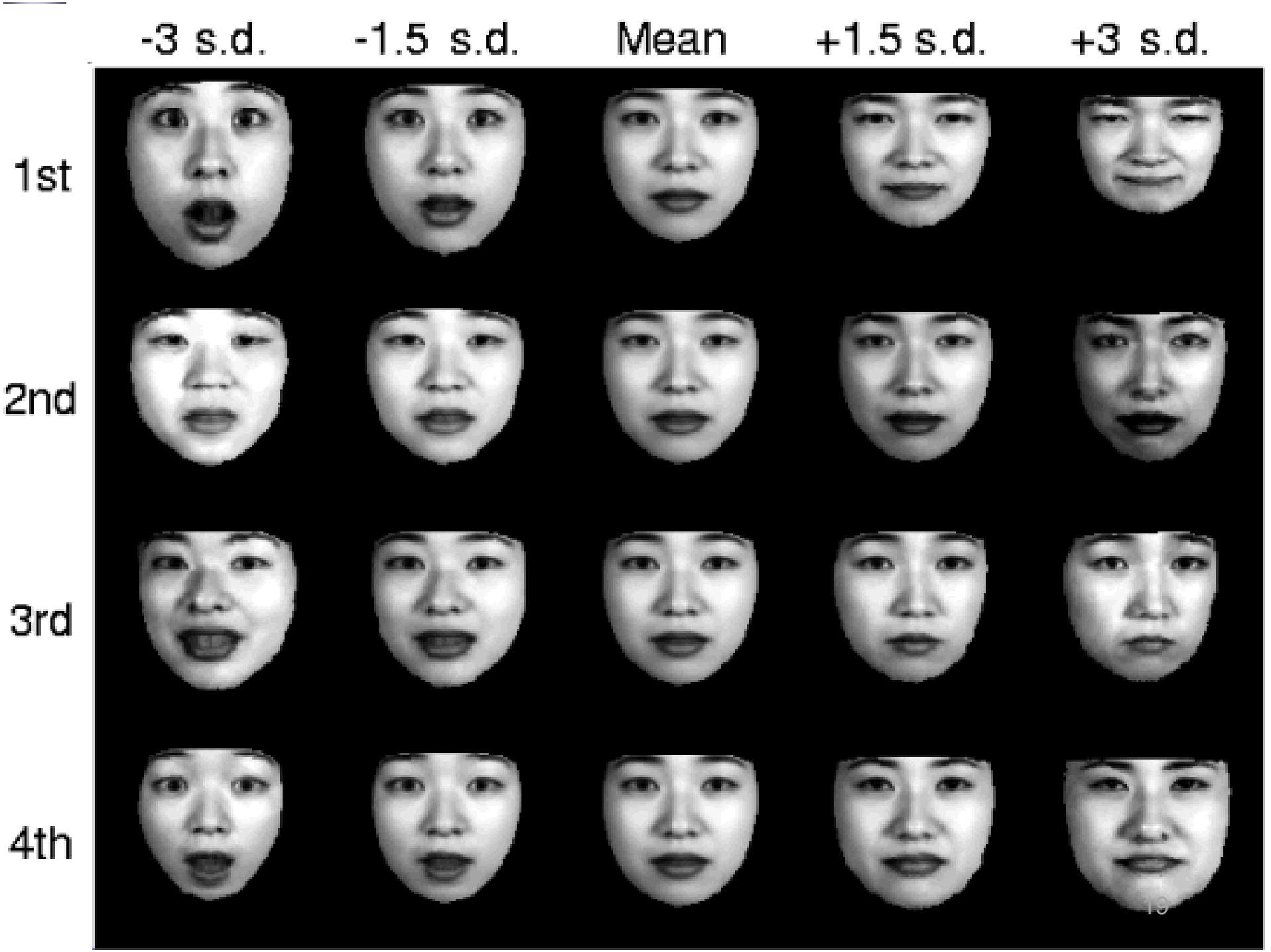
二、 K-L变换的性质

- 右图表示了一个二维空间中椭圆分布的样本集，在用**K-L**变换后新的坐标系中各分量的相关性消除。而在原坐标中 x_1, x_2 两个分量之间存在很明显的相关性。该图还反映了样本的 u_1 分量比较分散，因而对分类可能起较大作用，而 u_2 则对分类无太大作用，可以去掉。



K-L变换的一些典型应用

- 降维与压缩
- 构造参数模型
- 人脸识别
- 人脸图像合成


-3 s.d.
-1.5 s.d.
Mean
+1.5 s.d.
+3 s.d.
1st
2nd
3rd
4th

三、使用K-L变换进行特征提取

➢1、K-L坐标系的产生矩阵

$$E\left[XX^T\right]$$

$$S_\omega = \sum_{i=1}^{c} P_i \Sigma_i$$

其中 $\Sigma_i = E\left[(x-u_i)(x-u_i)^T\right]$

➢ 2、利用类均值向量提取特征

- ✓ 在很多**PR**问题中，类条件均值向量u_i包含有大量的分类信息。为了降低特征空间的维数，而又尽可能多地保持原有的分类信息，应选择这样一种变换：使变换后的**d**维特征空间中的类条件均值向量的各分量和其他的变换相比保持更多的分类信息。

- ✓ 各个类均值向量各分量的分类性能，不仅仅取决于它们和总体均值向量相应分量之间距离平方和的大小，而且还和该分量的方差以及分量间的相关程度有关。

- 为了估计各个分量（特征）对于分类的单独作用，可以先用计算出来的 $S_w = \sum_{i=1}^{c} P_i \Sigma_i$ 的**K-L**坐标系进行变换以消除原有各分量的相关性，同时考虑到 S_w 的本征值 λ_j 表示第 j 个分量的平均方差，为此用以下判据：

$$J(y_j) = (u_j^T S_b u_j) / \lambda_j$$

其中 $y_j = u_j^T X$ 表示在新坐标轴 u_j 上的分量

$$S_b = \sum_{i=1}^{c} P(\omega_1)(\mu_i - \mu)(\mu_i - \mu)^T$$

$$J(y_j) = (u_j^T S_b u_j) / (u_j^T S_w u_j)$$

可见 $J(y_j)$ 是类间离散度与类内离散度在 u_j 坐标轴的分量之比，$J(y_j)$ 越大，表明在新坐标系中该坐标轴包含的可分性信息越多。

- 为了降低特征空间的维数，可以简单地对各分量重新进行排序，使

$$J(y_1) \geq J(y_2) \geq \cdots J(y_d) \geq \cdots J(y_D)$$

并取前面**d**个最大的 $J(y_j)$ 值相对应的特征向量 $u_j, j = 1, \cdots, d$ 作为特征空间的基向量。

例：设有两类问题，其先验概率相等，即

$$P(\omega_1) = P(\omega_2) = \frac{1}{2}$$

样本均值向量分别为

$$\mu_1 = [4, 2]^T$$

$$\mu_2 = [-4, -2]^T$$

协方差矩阵分别是

$$\Sigma_1 = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 4 & 2 \\ 2 & 4 \end{bmatrix}$$

把维数从 2 压缩为 1

$$S_w = \sum_{i=1}^{c} P_i E_i \left[(x-\mu_i)(x-\mu_i)^T \right]$$

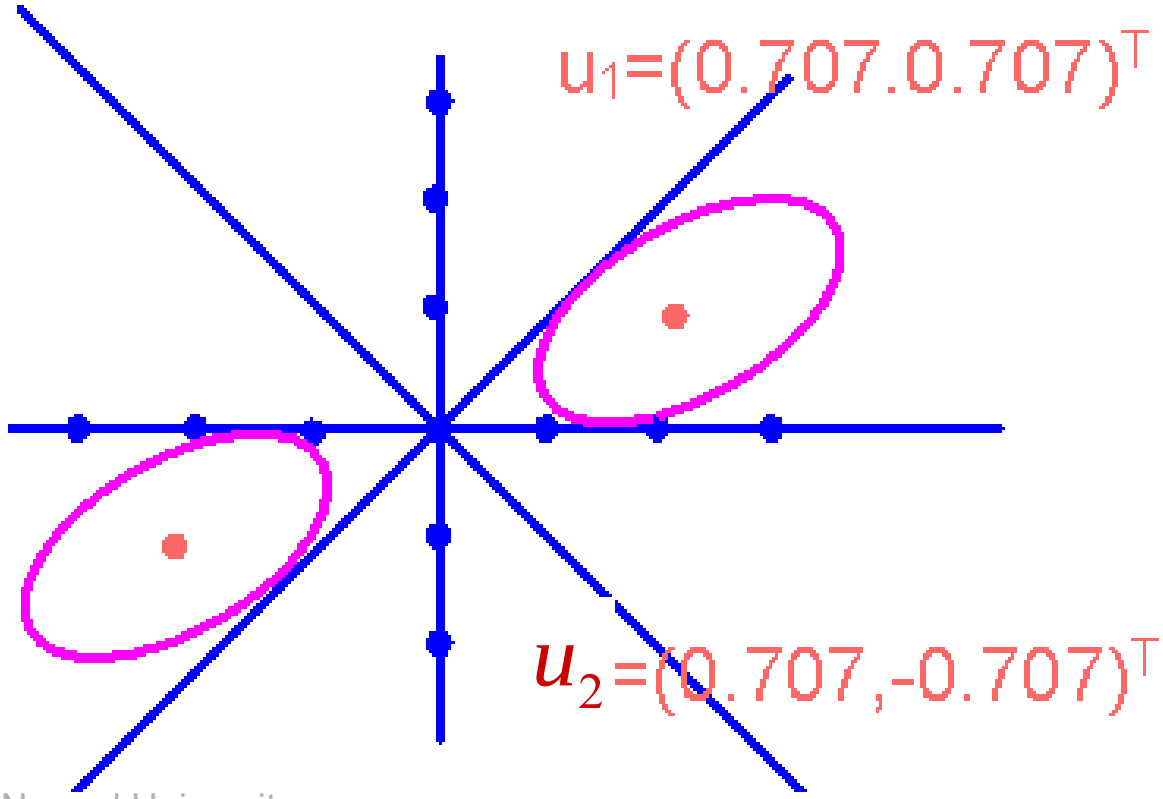
$$= \frac{1}{2}\Sigma_1 + \frac{1}{2}\Sigma_2 = \begin{bmatrix} 3.5 & 1.5 \\ 1.5 & 3.4 \end{bmatrix}$$

- 其特征值与特征向量：

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

$$\left.\begin{aligned} S_b &= \sum_{i=1}^{c} P_i(\mu_i - \mu)(\mu_i - \mu)^T \\ \mu &= \sum_{i=1}^{c} P_i \mu_i \end{aligned}\right\} \Rightarrow S_b = \begin{bmatrix} 16 & 8 \\ 8 & 4 \end{bmatrix}$$

$$J(y_j) = (u_j^T S_b u_j) / \lambda_j \qquad J(y_1) = 3.6 \qquad J(y_2) = 1$$

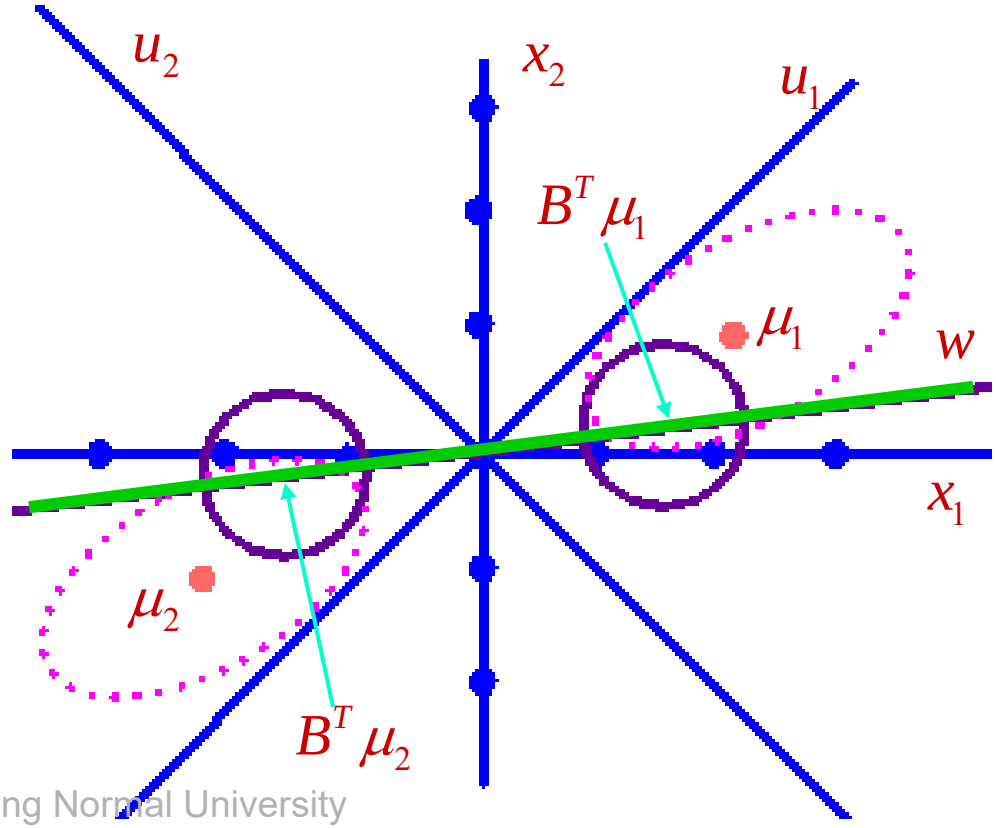


➢3、包含在类平均向量中判别信息的最优压缩

- ✓上述方法为了兼顾S_b和S_w，包含在类均值向量内的分类信息并没有全部利用。换句话说，类均值向量的判别信息在**K-L**坐标系的各个分量中都有反映，因此当$d \le D$时，上面方法总是不可避免地要损失掉一部分原来的类平均向量的判别信息。

- ✓如果只从类均值向量所包含的分类判别信息全部被利用这一点出发，应该选择包含均值向量连线方向在内的坐标系。这种方法一般不能满足各分量间互不相关的要求。

✓ 如果$S_w = I$，即它在特征空间中以超球体分布，就可做到既保持分量的不相关性，同时又能充分利用包含在类均值向量内的差别信息。



➢步骤：

✓ **Step1**：先用原坐标系中 S_w 作为产生矩阵，实行**K-L**变换**U**，将原有数据的相关性消除掉。然后用矩阵 $\Lambda^{-1/2}$进行归一化，使该 S'_w 矩阵变为单位矩阵。

白化变换： $B^T S_w B = I$

其中： $B = U\Lambda^{-1/2}$

经过**B**变换后的类间离散矩阵：

$$S'_b = B^T S_b B$$

- ✓ **Step2：** 以 S_b' 作为产生矩阵，做第二次**K-L**变换，由于S_b'的秩最多为**c-1**，所以S_b' 最多只有**c-1**个非零特征值。设共有**d**个非零特征值，$d \le c-1$，则该**d**个非零特征值就可表示类均值向量所包含的全部信息。设该**d**个特征值对应的特征向量为 V'：

$$V' = (v_1, \cdots, v_d)$$

- ✓ 从而把包含在类平均向量中的全部分类信息压缩为最小特征维数的变换是：

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

- ✓ 该方法在**c**不是很大时，有可能在更少维数的特征空间中保持原来的类平均向量的判别信息。

➢ 例：条件同上

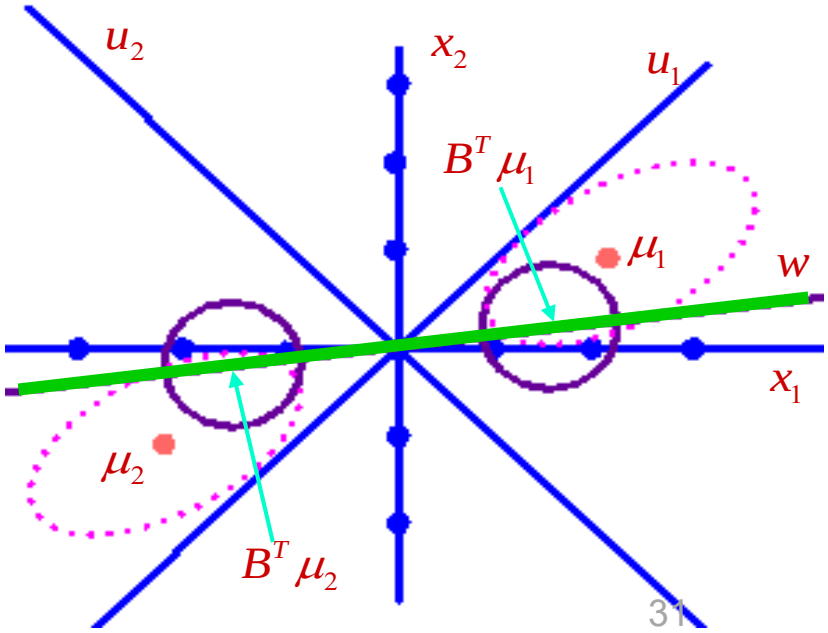
$$B = U\Lambda^{-1/2} = \begin{bmatrix} 0.707 & 0.707 \\ 0.707 & -0.707 \end{bmatrix}\begin{bmatrix} 0.447 & 0 \\ 0 & 0.707 \end{bmatrix}$$

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

$$S_b' = B^T S_b B = \begin{bmatrix} 3.6 & 1.897 \\ 1.897 & 1 \end{bmatrix}$$

$$\Lambda' = \begin{bmatrix} 4.6 & 0 \\ 0 & 0 \end{bmatrix} \qquad v = \begin{bmatrix} 0.884 \\ 0.466 \end{bmatrix}$$

$$w = Bv = \begin{bmatrix} 0.512 \\ 0.046 \end{bmatrix}$$



四、特征提取小结

- 基于对样本在特征空间分布的距离度量
  - ✓ 思想：通过原有特征线性组合而成新的特征向量，做到既降维，又能尽可能体现类间分离，类内聚集的原理；

- 从概率分布的差异出发，制定出反映概率分布差异的判据，以此确定特征如何提取。但需要有概率分布的知识，而且只有在概率分布具有简单形式是，计算才简便。
- 共同点： 判别函数的极值往往演变为找有关矩阵的特征值和特征向量，由相应的特征向量组成坐标系统基向量。

五、 特征选择

- 特征选择在概念上十分简单，即对原有特征进行删选优化。
- 特征选择:
  - ✓由原有**D**维特征所组成的特征空间中选出若干个特征，组成描述样本的新特征空间，即从原有的**D**维空间选取一个**d**维子空间(**d**＜**D**)，在该子空间中进行模式识别。

➢两个问题要解决

- ✓一是选择特性的标准，也就是选择前面讨论过的可分离性判据，以这些判据为准则，使所选择的**d**维子空间具有最大的可分离性。
- ✓二是要找出较好的特征选择方法，以在允许的时间内选择出一组最优的特征。所谓最优的特征组，就是要找到合适的特征的组合。
  - 最优搜索算法
  - 次优搜索算法