- 点到点之间的距离

在 n 维空间中，$\boldsymbol{a}$ 与 $\boldsymbol{b}$ 两点之间的欧氏距离为：

$$\mathrm{D}(\boldsymbol{a},\boldsymbol{b}) = \|\boldsymbol{a}-\boldsymbol{b}\|$$

写成距离平方：

$$D^2(\boldsymbol{a},\boldsymbol{b}) = (\boldsymbol{a}-\boldsymbol{b})^T(\boldsymbol{a}-\boldsymbol{b}) = \sum_{k=1}^{n}(a_k-b_k)^2$$

其中，$\boldsymbol{a}$ 和 $\boldsymbol{b}$ 为 n 维向量，其第 k 个分量分别是 a_k 和 b_k。

- 点到点集之间的距离

在 n 维空间中，点 $\boldsymbol{a}$ 到点 $\boldsymbol{x}^{(i)}$之间的距离平方为：

$$D^2(\boldsymbol{a},\boldsymbol{x}^{(i)}) = \sum_{k=1}^{n}(a_k-\boldsymbol{x}_k^{(i)})^2$$

因此，点 $\boldsymbol{a}$ 到点集$\{\boldsymbol{x}^{(i)}\}_{i=1,2,\ldots,K}$之间的均方距离为：

$$\overline{D^2(\boldsymbol{a},\{\boldsymbol{x}^{(i)}\})} = \frac{1}{K}\sum_{i=1}^{K}D^2(\boldsymbol{a},\boldsymbol{x}^{(i)}) = \frac{1}{K}\sum_{i=1}^{K}\left\{\sum_{k=1}^{n}(a_k-x_k^{(i)})^2\right\}$$

- 类内距离

n 维空间中同一类内各模式样本点集$\{\boldsymbol{x}^{(i)}\}_{i=1,2,\ldots,K}$，其内部各点的均方距离为$\overline{D^2(\{\boldsymbol{x}^{(j)}\},\{\boldsymbol{x}^{(i)}\})}$，其中$i,j=1,2,\ldots,K,i\neq j$，即：

$$\overline{D^2(\{\boldsymbol{x}^{(j)}\},\{\boldsymbol{x}^{(i)}\})}=\frac{1}{K}\sum_{j=1}^{K}\left[\frac{1}{K-1}\sum_{\substack{i=1\\ i\neq j}}^{K}\sum_{k=1}^{n}(x_k^{(j)}-x_k^{(i)})^2\right]$$

可证明：

$$\overline{D^2}=2\sum_{k=1}^{n}\sigma_k^2$$

其中σ_k^2为$\{\boldsymbol{x}^{(i)}\}$在第 k 个分量上的无偏方差，即：

$$\sigma_k^2=\frac{1}{K-1}\sum_{i=1}^{K}(x_k^{(i)}-\overline{x_k})^2$$

其中$\overline{x_k}=\dfrac{1}{K}\sum_{i=1}^{K}x_k^{(i)}$为$\{\boldsymbol{x}^{(i)}\}$在第 k 个分量方向上的均值。

[*证明作为练习*]

- 类内散布矩阵

考虑一类内模式点集$\{\boldsymbol{x}^{(i)}\}_{i=1,2,\ldots,K}$，其类内散布矩阵为：

$$\boldsymbol{S}=\sum_{i=1}^{K}\{(\boldsymbol{x}^{(i)}-\boldsymbol{m})(\boldsymbol{x}^{(i)}-\boldsymbol{m})^{T}\}$$

其中

$$\boldsymbol{m}=\frac{1}{K}\sum_{i=1}^{K}\boldsymbol{x}^{(i)}$$

对属于同一类的模式样本，类内散布矩阵表示各样本点围绕其均值周围的散布情况。

- 类间距离和类间散布矩阵

在考虑有两个以上的类别，如集合$\{\boldsymbol{a}^{(i)}\}$和$\{\boldsymbol{b}^{(j)}\}$时，类间距离对类别的可分性起着重要作用，此时应计算：

$\overline{D^2(\{\boldsymbol{a}^{(i)}\},\{\boldsymbol{b}^{(j)}\})}_{i=1,2,\ldots,K_a;\, j=1,2,\ldots,K_b}$。

为简化起见，常用两类样本各自质心间的距离作为类间距离，并假设两类样本出现的概率相等，则：

$$D^2=\sum_{k=1}^{n}(m_{1_k}-m_{2_k})^2$$

其中 $\boldsymbol{m}_1$ 和 $\boldsymbol{m}_2$ 为两类模式样本集各自的均值向量，m_{1_k} 和 m_{2_k} 为 $\boldsymbol{m}_1$ 和 $\boldsymbol{m}_2$ 的第 k 个分量，n 为维数。

写成矩阵形式：$\boldsymbol{S}_{b2}=(\boldsymbol{m}_1-\boldsymbol{m}_2)(\boldsymbol{m}_1-\boldsymbol{m}_2)^T$ 为两类模式的类间散布矩阵。

对三个以上的类别，类间散布矩阵常写成：

$$\begin{aligned}\boldsymbol{S}_b&=\frac{1}{2}\sum_{i=1}^{M}P(\omega_i)\sum_{j=1}^{M}P(\omega_j)(\boldsymbol{m}_i-\boldsymbol{m}_j)(\boldsymbol{m}_i-\boldsymbol{m}_j)^T\\&=\sum_{i=1}^{M}P(\omega_i)(\boldsymbol{m}_i-\boldsymbol{m}_0)(\boldsymbol{m}_i-\boldsymbol{m}_0)^T\\&\propto\sum_{i=1}^{M}N_i(\boldsymbol{m}_i-\boldsymbol{m}_0)(\boldsymbol{m}_i-\boldsymbol{m}_0)^T\end{aligned}$$

其中，$\boldsymbol{m}_0$ 为多类模式（如共有 M 类）分布的总体均值向量，即：

$$\boldsymbol{m}_0=E\{\boldsymbol{x}\}=\sum_{i=1}^{c}P(\omega_i)\boldsymbol{m}_i,\quad \forall\omega_i,\ i=1,2,\ldots,M$$

- 多类模式集散布矩阵

多类情况的类内散布矩阵，可写成各类的类内散布矩阵的先验概率的加权和，即：

$$\boldsymbol{S}_w = \sum_{i=1}^{M} P(\omega_i) E\{(\boldsymbol{x}-\boldsymbol{m}_i)(\boldsymbol{x}-\boldsymbol{m}_i)^T \mid \omega_i\} = \sum_{i=1}^{M} P(\omega_i) \boldsymbol{C}_i$$

其中 $\boldsymbol{C}_{\mathrm{i}}$ 是第 i 类的协方差矩阵。

有时，用多类模式总体分布的散布矩阵来反映其可分性，即：

$$\boldsymbol{S}_t = E\{(\boldsymbol{x}-\boldsymbol{m}_0)(\boldsymbol{x}-\boldsymbol{m}_0)^T\}, \quad \boldsymbol{x} \in \forall \omega_i, i = 1, 2, \ldots, M$$

其中，$\boldsymbol{m}_0$ 为多类模式分布的总体均值向量。

可以证明：$\boldsymbol{S}_t = \boldsymbol{S}_w + \boldsymbol{S}_b$，即总体散布矩阵是各类类内散布矩阵与类间散布矩阵之和。

- 对于 ω_i 和 ω_j 两类训练样本的特征选择

例：对于 ω_i 和 ω_j 两类训练样本，假设其均值向量为 $\boldsymbol{m}_i$ 和 $\boldsymbol{m}_j$，其 k 维方向的分量为 m_{ik} 和 m_{jk}，方差为 σ_{ik}^2 和 σ_{jk}^2，定义可分性准则函数：

$$G_k = \frac{(m_{ik} - m_{jk})^2}{\sigma_{ik}^2 + \sigma_{jk}^2},\ k = 1, 2, \ldots, n$$

则 G_K 为正值。G_K 值越大，表示测度值的第 k 个分量对分离 ω_i 和 ω_j 两类越有效。将 $\{G_K, k=1,2,\ldots,n\}$ 按大小排队，选出最大的 m 个对应的测度值作为分类特征，即达到特征选择的目的。

- 一般特征的散布矩阵准则

类内：$\boldsymbol{S}_w = \sum_{i=1}^{c} P(\omega_i) E\{(\boldsymbol{x}-\boldsymbol{m}_i)(\boldsymbol{x}-\boldsymbol{m}_i)^T \mid \omega_i\}$

类间：$\boldsymbol{S}_b = \sum_{i=1}^{c} P(\omega_i)(\boldsymbol{m}_i-\boldsymbol{m}_0)(\boldsymbol{m}_i-\boldsymbol{m}_0)^T$

直观上，类间离散度越大且类内离散度越小，则可分性越好。因此，可推导出散布矩阵准则采用如下形式：

行列式形式：$J_1 = \det(\boldsymbol{S}_w^{-1}\boldsymbol{S}_b) = \prod_i \lambda_i$

迹形式：$J_2 = tr(\boldsymbol{S}_w^{-1}\boldsymbol{S}_b) = \sum_i \lambda_i$

其中，λ_i 是矩阵 $\boldsymbol{S}_w^{-1}\boldsymbol{S}_b$ 的特征值。使 J_1 或 J_2 最大的子集可作为选择的分类特征。

- 离散的有限 K-L 展开式的形式设一连续的随机实函数 $\boldsymbol{x}(t)$，

$T_1 \le t \le T_2$，则 $\boldsymbol{x}(t)$可用已知的正交函数集$\{\varphi_j(t), j=1,2,\ldots\}$的线性组合来展开，即：

$$\begin{aligned} x(t) &= a_1\varphi_1(t) + a_2\varphi_2(t) + \cdots + a_j\varphi_j(t) + \cdots \\ &= \sum_{j=1}^{\infty} a_j\varphi_j(t), \quad T_1 \le t \le T_2 \end{aligned} \tag{1}$$

式中，a_j 为展开式的随机系数，$\varphi_j(t)$为一连续的正交函数，它应满足：

$$\int_{T_1}^{T_2} \varphi_n^{(t)} \tilde{\varphi}_m(t)dt = \begin{cases} 1, & if\ m = n \\ 0 & if\ m \ne n \end{cases}$$

其中 $\tilde{\varphi}_m(t)$ 为 $\varphi_m(t)$的共轭复数式。

将上式写成离散的正交函数形式，使连续随机函数 $\boldsymbol{x}(t)$和连续正交函数 $\varphi_j(t)$在区间 $T_1 \le t \le T_2$ 内被等间隔采样为 n 个离散点，即：

$$\boldsymbol{x}(t) \to \{x(1), x(2), \cdots, x(n)\}$$

$$\varphi_j(t) \to \{\varphi_j(1), \varphi_j(2), \cdots, \varphi_j(n)\}$$

写成向量形式：

$$\boldsymbol{x} = (x(1), x(2), \cdots, x(n))^T$$

$$\boldsymbol{\varphi}_j = (\varphi_j(1), \varphi_j(2), \cdots, \varphi_j(n))^T,\ j = 1, 2, \cdots, n$$

将式(1)取 n 项近似，并写成离散展开式：

$$\boldsymbol{x} = \sum_{j=1}^{n} a_j\boldsymbol{\varphi}_j = \boldsymbol{\Phi a}, \quad T_1 \le t \le T_2 \tag{2}$$

其中，$\boldsymbol{a}$ 为展开式中随机系数的向量形式，即：

$$\boldsymbol{a} = (a_1, a_2, \ldots, a_j, \ldots, a_n)^{\mathrm{T}}$$

$\boldsymbol{\Phi}$ 为 n x n 维矩阵，即：

$$\boldsymbol{\Phi}=(\boldsymbol{\varphi}_1,\boldsymbol{\varphi}_2,\cdots,\boldsymbol{\varphi}_n)=\begin{bmatrix}\varphi_1(1) & \varphi_2(1) & \cdots & \varphi_n(1)\\ \varphi_1(2) & \varphi_2(2) & \cdots & \varphi_n(2)\\ \cdots & \cdots & \cdots & \cdots\\ \varphi_1(n) & \varphi_2(n) & \cdots & \varphi_n(n)\end{bmatrix}$$

其中，每一列为正交函数集中的一个函数，小括号内的序号为正交函数的采样点次序。因此，$\boldsymbol{\Phi}$ 实质上是由 $\boldsymbol{\varphi}_{\mathrm{j}}$ 向量组成的正交变换矩阵，它将 $\boldsymbol{x}$ 变换成 $\boldsymbol{a}$。

- 正交向量集$\{\boldsymbol{\varphi}_j\}$的确定

设随机向量 $\boldsymbol{x}$ 的总体自相关矩阵为 $\boldsymbol{R} = \mathrm{E}\{\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\}$。由

$$\boldsymbol{x} = \sum_{j=1}^{n} a_j \boldsymbol{\varphi}_j = \boldsymbol{\Phi a}, \quad T_1 \le t \le T_2 \tag{1}$$

将 $\boldsymbol{x}=\boldsymbol{\Phi a}$ 代入 $\boldsymbol{R} = \mathrm{E}\{\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\}$，得：

$$\mathrm{R} = \mathrm{E}\{\boldsymbol{\Phi a}\, \boldsymbol{a}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\}=\boldsymbol{\Phi}(\mathrm{E}\{\boldsymbol{a}\, \boldsymbol{a}^{\mathrm{T}}\})\boldsymbol{\Phi}^{\mathrm{T}}$$

要求系数向量 $\boldsymbol{a}$ 的各个不同分量应统计独立，即应使$(a_1, a_2, \ldots, a_j, \ldots, a_n)$满足如下关系：

$$E(a_j a_k) = \begin{cases} \lambda_j & if\ j = k \\ 0 & if\ j \ne k \end{cases}$$

写成矩阵形式，应使：$\mathrm{E}\{\boldsymbol{a}\, \boldsymbol{a}^{\mathrm{T}}\} = \boldsymbol{D}_\lambda$，其中 $\boldsymbol{D}_\lambda$ 为对角形矩阵，其互相关成分均为 0，即：

$$\mathrm{D}_\lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & 0 & \cdots & 0 \\ 0 & & \lambda_{\mathrm{j}} & & 0 \\ 0 & \cdots & 0 & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & \lambda_{\mathrm{n}} \end{bmatrix}$$

则：

$$\boldsymbol{R} = \boldsymbol{\Phi D}_\lambda \boldsymbol{\Phi}^{\mathrm{T}}$$

由于 $\boldsymbol{\Phi}$ 中的各个向量 $\boldsymbol{\varphi}_j$ 都相互归一正交，故有：

$$\boldsymbol{R\Phi}=\boldsymbol{\Phi D}_\lambda \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}=\boldsymbol{\Phi D}_\lambda$$

其中，$\boldsymbol{\varphi}_j$ 向量对应为：

$$\boldsymbol{R}\boldsymbol{\varphi}_{\mathrm{j}}=\lambda_{\mathrm{j}}\boldsymbol{\varphi}_{\mathrm{j}}$$

可以看出，λ_j是 $\boldsymbol{x}$ 的自相关矩阵 $\boldsymbol{R}$ 的特征值，$\boldsymbol{\varphi}_j$是对应的特征向量。因为 $\boldsymbol{R}$ 是实对称矩阵，其不同特征值对应的特征向量应正交，即：

$$\boldsymbol{\varphi}_j^T\boldsymbol{\varphi}_k=\begin{cases}1 & if \;\; j=k\\ 0 & if \;\; j\neq k\end{cases}$$

由式(1)，K-L 展开式系数应为：

$$\boldsymbol{a}=\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{x}$$

- K-L 展开式系数的计算步骤

K-L 展开式系数可如下求得：

1. 求随机向量 $\boldsymbol{x}$ 的自相关矩阵：$\boldsymbol{R} = \mathrm{E}\{\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\}$
2. 求出矩阵 $\boldsymbol{R}$ 的特征值 λ_j 和对应的特征向量 $\boldsymbol{\varphi}_j$，$j = 1,2,\ldots,\mathrm{n}$，得矩阵：

$$\boldsymbol{\Phi} = (\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2, \cdots, \boldsymbol{\varphi}_n)$$

3. 计算展开式系数：

$$\boldsymbol{a} = \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{x}$$

- 问题：选取变换矩阵 $\boldsymbol{\Phi}$，使得降维后的新向量在最小均方差条件下接近原来的向量 $\boldsymbol{x}$

对于 $\boldsymbol{x}=\sum_{j=1}^{n}a_j\boldsymbol{\varphi}_j$，现仅取 m 项，对略去的系数项用预先选定的常数 b 代替，此时对 $\boldsymbol{x}$ 的估计值为：

$$\hat{\boldsymbol{x}}=\sum_{j=1}^{m}a_j\boldsymbol{\varphi}_j+\sum_{j=m+1}^{n}b\boldsymbol{\varphi}_j$$

则产生的误差为：

$$\Delta\boldsymbol{x}=\boldsymbol{x}-\hat{\boldsymbol{x}}=\sum_{j=m+1}^{n}(a_j-b)\boldsymbol{\varphi}_j$$

则 $\Delta\boldsymbol{x}$ 的均方误差为：

$$\overline{\varepsilon^2}=E\{\|\Delta\boldsymbol{x}\|\}^2=\sum_{j=m+1}^{n}\{E(a_j-b)^2\}$$

要使 $\overline{\varepsilon^2}$ 最小，对 b 的选择应满足：

$$\frac{\partial}{\partial b}[E(a_j-b)^2]=\frac{\partial}{\partial b}[E(a_j^2-2a_jb+b^2)]=-2[E(a_j)-b]=0$$

因此，$b=\mathrm{E}[a_j]$，即对省略掉的 a 中的分量，应使用它们的数学期望来代替，此时的误差为：

$$\begin{aligned}\overline{\varepsilon^2}&=\sum_{j=m+1}^{n}E[(a_j-E\{a_j\})^2]=\sum_{j=m+1}^{n}\boldsymbol{\varphi}_j^TE[(\boldsymbol{x}-E\{\boldsymbol{x}\})(\boldsymbol{x}-E\{\boldsymbol{x}\})^T]\boldsymbol{\varphi}_j\\&=\sum_{j=m+1}^{n}\boldsymbol{\varphi}_j^T\boldsymbol{C}_x\boldsymbol{\varphi}_j\end{aligned}$$

其中，$\boldsymbol{C}_\mathrm{x}$ 为 $\boldsymbol{x}$ 的协方差矩阵, $\{\boldsymbol{\varphi}_j\}$是正交向量，由拉格朗日法可导出 $\boldsymbol{\varphi}_j$ 为 $\boldsymbol{C}_\mathrm{x}$ 的特征值。

设 λ_j 为 $\boldsymbol{C}_\mathrm{x}$ 的第 j 个特征值，$\boldsymbol{\varphi}_j$ 是与 λ_j 对应的特征向量，则

$$\boldsymbol{C}_x\boldsymbol{\varphi}_j=\lambda_j\boldsymbol{\varphi}_j$$

由于

$$\boldsymbol{\varphi}_j^T \boldsymbol{\varphi}_j = 1$$

从而

$$\boldsymbol{\varphi}_j^T \boldsymbol{C}_x \boldsymbol{\varphi}_j = \lambda_j$$

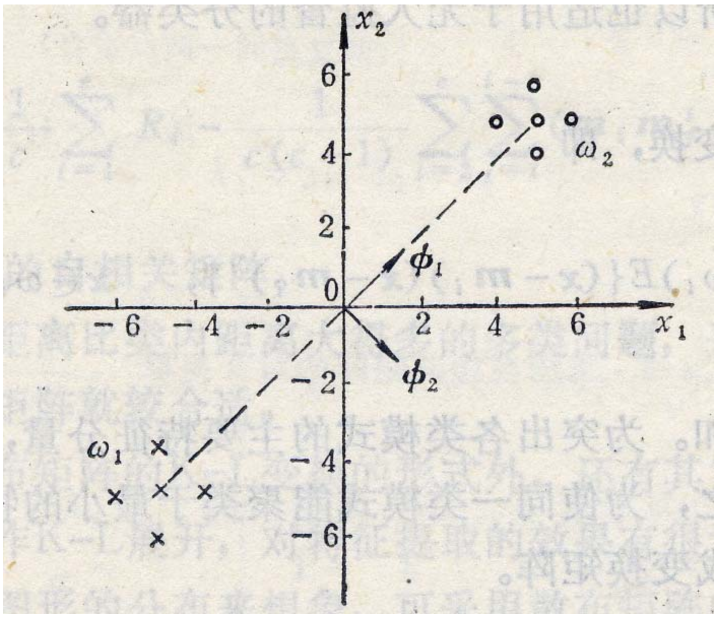
因此

$$\overline{\varepsilon^2} = \sum_{j=m+1}^{n} \boldsymbol{\varphi}_j^T \boldsymbol{C}_x \boldsymbol{\varphi}_j = \sum_{j=m+1}^{n} \lambda_j$$

由此可以看出，λ_j 值越小，误差也越小。

● K-L 变换实例

给定两类模式，其分布如图所示，试用 K-L 变换实现一维的特征提取（假定两类模式出现的概率相等）。



P(ω1)= P(ω2)=0.5, $\boldsymbol{m}=0.5*\frac{1}{5}\sum_{\boldsymbol{x}^j\in\omega_1}\boldsymbol{x}^j+0.5*\frac{1}{5}\sum_{\boldsymbol{x}^j\in\omega_2}^{N_i}\boldsymbol{x}^j=(0,0)^T$

符合 K-L 变换进行特征压缩的最佳条件。

因 P(ω1)= P(ω2)=0.5，故

$$\boldsymbol{R}=\sum_{i=1}^{2}P(\omega_i)E\{\boldsymbol{x}\boldsymbol{x}^T\}=\frac{1}{2}\left[\frac{1}{5}\sum_{\boldsymbol{x}^j\in\omega_1}^{5}\boldsymbol{x}^j(\boldsymbol{x}^j)^T\right]+\frac{1}{2}\left[\frac{1}{5}\sum_{\boldsymbol{x}^j\in\omega_2}^{5}\boldsymbol{x}^j(\boldsymbol{x}^j)^T\right]=\begin{pmatrix}25.4&25.0\\25.0&25.4\end{pmatrix}$$

解特征值方程$|\boldsymbol{R}-\lambda\boldsymbol{I}|=0$，求 $\boldsymbol{R}$ 的特征值。

由$(25.4-\lambda)^2-25.0^2=0$，得特征值 $\lambda_1=50.4$，$\lambda_2=0.4$

其对应的特征向量可由 $\boldsymbol{R}\boldsymbol{\varphi}_i=\lambda_i\boldsymbol{\varphi}_i$ 求得：

$$\boldsymbol{\varphi}_1=\frac{1}{\sqrt{2}}\begin{pmatrix}1\\1\end{pmatrix}，\quad \boldsymbol{\varphi}_2=\frac{1}{\sqrt{2}}\begin{pmatrix}1\\-1\end{pmatrix}$$

选 λ_1 对应的变换向量作为变换矩阵，由 $\boldsymbol{y}=\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{x}$ 得变换后的一维模式特征为：

$$\omega_1:\left\{-\frac{10}{\sqrt{2}},-\frac{9}{\sqrt{2}},-\frac{9}{\sqrt{2}},-\frac{11}{\sqrt{2}},-\frac{11}{\sqrt{2}}\right\}\ \omega_2:\left\{\frac{10}{\sqrt{2}},\frac{11}{\sqrt{2}},\frac{11}{\sqrt{2}},\frac{9}{\sqrt{2}},\frac{9}{\sqrt{2}}\right\}$$