第 7 章 判别分析

经典的“线性判别分析”(Linear Discriminant Analysis，简记 LDA)最早由 Fisher(1936)与 Mahalanobis(1936)提出，可用于二分类或多分类问题。

Smith(1946)将其推广到“二次判别分析”(Quadratic Discriminant Analysis，简记 QDA)。

7.1 贝叶斯决策理论

假设训练数据为$\left\{\mathbf{x}_i, y_i\right\}_{i=1}^{n}$，而$y_i$的可能取值为$y_i \in \left\{1, 2, \cdots, K\right\}$，共分为$K$类($K$ classes)。

如果知道条件概率$P(y_i = k \mid \mathbf{x} = \mathbf{x}_i)$，其中$k = 1, 2, \cdots, K$；则最优预测应最大化此条件概率，也称为**后验概率**(posterior probability)：

$$\max_k \; P(y_i = k \mid \mathbf{x} = \mathbf{x}_i) \qquad (7.1)$$

上式表明，对于y_i的预测，应选择$\hat{y}_i = k$，使得后验概率$P(y_i = k \mid \mathbf{x} = \mathbf{x}_i)$最大化。

这种决策方式称为**贝叶斯最优决策**(Bayes optimal decision)。

由此所得的决策边界，称为**贝叶斯决策边界**(Bayes decision boundary)。

使用贝叶斯最优决策，所能达到的错误率，称为**贝叶斯错误率**(Bayes error rate) 或“贝叶斯风险”(Bayes risk)。

贝叶斯错误率为可能的最低错误率，故也称为**最优贝叶斯错误率**(optimal Bayes rate)。

例 对于二分类问题，假设 $P(y_i=1|\mathbf{x}=\mathbf{x}_i)=0.7$，而 $P(y_i=0|\mathbf{x}=\mathbf{x}_i)=0.3$。此时，最优贝叶斯决策应预测 $y_i=1$。此预测错误的概率为 0.3，即“贝叶斯错误率”或“贝叶斯风险”。这是可能的最低预测错误率，常作为参照系，以考量模型的预测效果。反之，如果预测 $y_i=0$，预测错误的概率高达 0.7。

为进行贝叶斯最优决策，需要知道$P(y_i = k \mid \mathbf{x} = \mathbf{x}_i)$。

直接估计$P(y_i = k \mid \mathbf{x} = \mathbf{x}_i)$可能比较困难，因为满足条件“$\mathbf{x} = \mathbf{x}_i$”的观测值一般不多。

为此，使用贝叶斯公式，在给定$\mathbf{x} = \mathbf{x}_i$的情况下，事件“$y_i = k$”的后验概率可写为

$$
\begin{aligned}
P(y_i = k \mid \mathbf{x} = \mathbf{x}_i) &= \frac{P(\mathbf{x} = \mathbf{x}_i \mid y_i = k)P(y_i = k)}{P(\mathbf{x} = \mathbf{x}_i)} \\
&\equiv \frac{f_k(\mathbf{x}_i)\pi_k}{P(\mathbf{x} = \mathbf{x}_i)}
\end{aligned} \tag{7.2}
$$

其中，$\pi_k \equiv P(y_i = k)$为看到数据“$\mathbf{x} = \mathbf{x}_i$”之前的**先验概率**(prior probability)；

$f_k(\mathbf{x}_i) \equiv P(\mathbf{x} = \mathbf{x}_i \mid y_i = k)$为给定类别“$y_i = k$”的条件概率密度(class-conditional density of $\mathbf{x}_i$ in class k)。

给定$\mathbf{x}=\mathbf{x}_i$，为比较第k类与第l类$(k \neq l)$的后验概率，将二者相除，可得**后验几率**(posterior odds)，也称为“似然比”(likelihood ratio)：

$$\frac{P(y_i = k \mid \mathbf{x}=\mathbf{x}_i)}{P(y_i = l \mid \mathbf{x}=\mathbf{x}_i)} = \frac{\dfrac{f_k(\mathbf{x}_i)\pi_k}{P(\mathbf{x}=\mathbf{x}_i)}}{\dfrac{f_l(\mathbf{x}_i)\pi_l}{P(\mathbf{x}=\mathbf{x}_i)}} = \frac{f_k(\mathbf{x}_i)\pi_k}{f_l(\mathbf{x}_i)\pi_l} \tag{7.3}$$

只要后验几率$\dfrac{P(y_i = k \mid \mathbf{x}=\mathbf{x}_i)}{P(y_i = l \mid \mathbf{x}=\mathbf{x}_i)} > 1$，即可预测为第$k$类。

而“$\dfrac{P(y_i = k \mid \mathbf{x} = \mathbf{x}_i)}{P(y_i = l \mid \mathbf{x} = \mathbf{x}_i)} = 1$”，则为**贝叶斯决策边界**(Bayes decision boundary)。

例 对于二分类问题$y \in \{0, 1\}$，假设只有一个特征变量x。 对于$y = 0$的数据，x的条件分布为$x \mid y = 0 \sim N(-2,1)$；对于$y = 1$的数据，$x$的条件分布为$x \mid y = 1 \sim N(2,1)$，参见图 7.1。

若假设先验概率相等，即$\pi_0 = P(y=0) = P(y=1) = \pi_1$，则

$$\frac{P(y_i = 1 \mid \mathbf{x} = \mathbf{x}_i)}{P(y_i = 0 \mid \mathbf{x} = \mathbf{x}_i)} = \frac{f_1(\mathbf{x}_i)\pi_1}{f_0(\mathbf{x}_i)\pi_0} = \frac{f_1(\mathbf{x}_i)}{f_0(\mathbf{x}_i)} \tag{7.4}$$

由此可知，后验几率取决于两类数据的条件概率密度$f_1(\mathbf{x}_i)$与$f_0(\mathbf{x}_i)$之比。

故贝叶斯决策边界为$f_1(\mathbf{x}_i) = f_0(\mathbf{x}_i)$，参见图 7.1 中的垂直虚线。

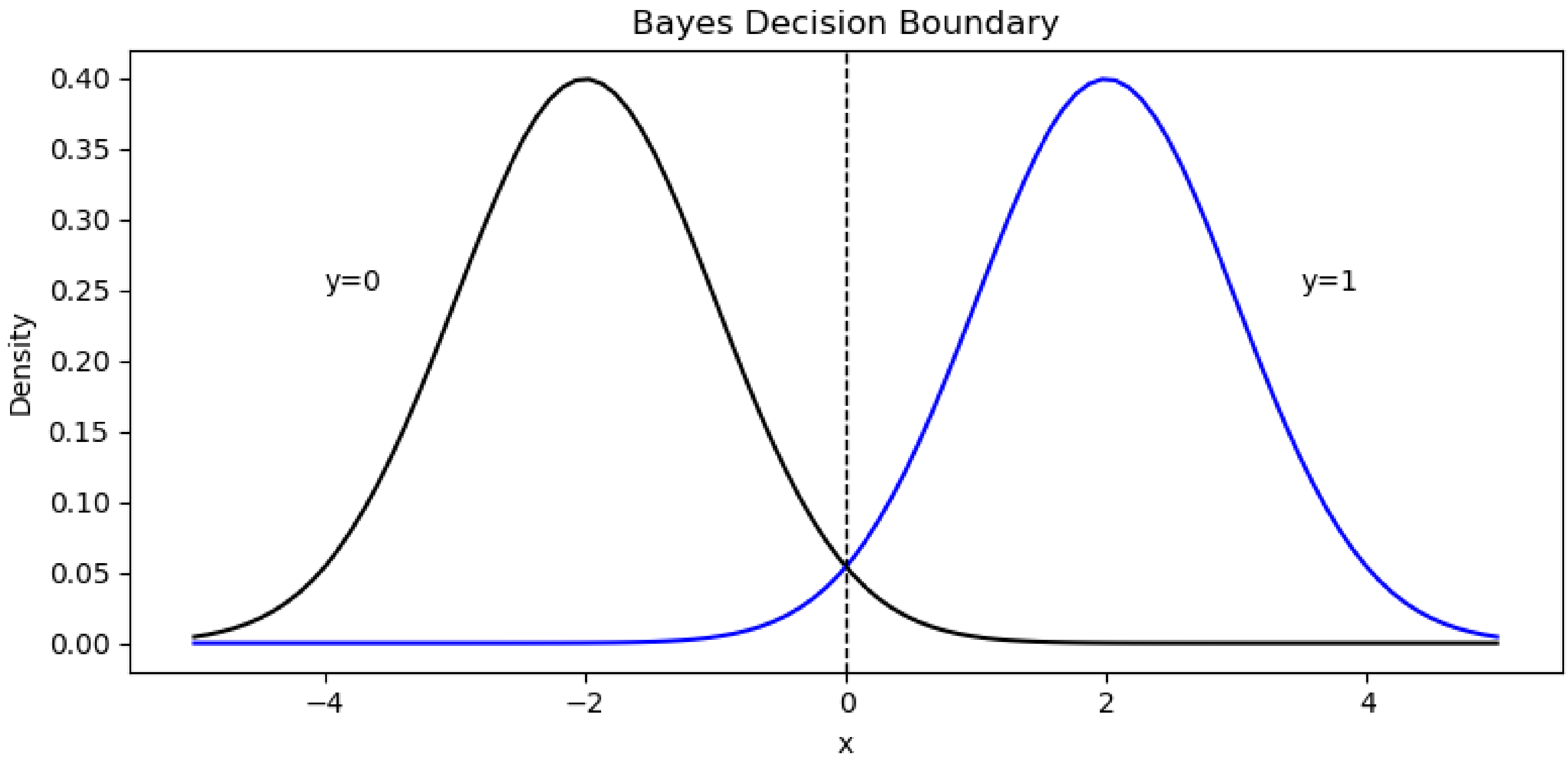

图 7.1　单变量的贝叶斯决策边界

7.2 线性判别分析

在一般的高维情况下，Mahalanobis(1936)假定，在给定类别“$y_i = k$”的情况下，$\mathbf{x}_i$服从p维多元正态分布$N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$，其概率密度可写为

$$f_k(\mathbf{x}_i) = \frac{1}{(2\pi)^{p/2} \left|\boldsymbol{\Sigma}_k\right|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k)\right\} \tag{7.5}$$

其中，$\boldsymbol{\mu}_k$为期望，$\boldsymbol{\Sigma}_k$为协方差矩阵，而$\left|\boldsymbol{\Sigma}_k\right|$为$\boldsymbol{\Sigma}_k$的行列式。

为简化计算，进一步假设所有类别的协方差矩阵均相等，即 $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma},\ \forall\, k$。该假设称为“同方差假定”(homoskedastic assumption)。

将表达式(7.5)代入后验几率方程(7.3)，并注意到$\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}\ \ (\forall\, k)$可得：

$$\frac{P(y_i = k \mid \mathbf{x} = \mathbf{x}_i)}{P(y_i = l \mid \mathbf{x} = \mathbf{x}_i)} = \frac{\pi_k \exp\left\{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)\right\}}{\pi_l \exp\left\{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)\right\}} \tag{7.6}$$

将上式取对数，即可得到**对数后验几率**(log posterior odds)：

$$\ln\frac{P(y_i=k\,|\,\mathbf{x}=\mathbf{x}_i)}{P(y_i=l\,|\,\mathbf{x}=\mathbf{x}_i)}$$

$$=\ln\frac{\pi_k}{\pi_l}-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu}_k)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_k)+\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu}_l)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_l)$$

$$=\underbrace{\left[\ln\frac{\pi_k}{\pi_l}-\frac{1}{2}\boldsymbol{\mu}_k'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k+\frac{1}{2}\boldsymbol{\mu}_l'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_l\right]}_{constant}+\underbrace{\mathbf{x}_i'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k-\boldsymbol{\mu}_l)}_{linear} \qquad (7.7)$$

若令对数几率等于 0，即得到在第k类与第l类之间的“决策边界”(decision boundary)：

$$\ln\frac{P(y_i=k\,|\,\mathbf{x}=\mathbf{x}_i)}{P(y_i=l\,|\,\mathbf{x}=\mathbf{x}_i)}=\underbrace{\left[\ln\frac{\pi_k}{\pi_l}-\frac{1}{2}\boldsymbol{\mu}_k'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k+\frac{1}{2}\boldsymbol{\mu}_l'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_l\right]}_{constant}+\underbrace{\mathbf{x}_i'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k-\boldsymbol{\mu}_l)}_{linear}=0 \quad (7.8)$$

由于此决策边界为线性函数，故名**线性判别分析**(Linear Discriminant Analysis，简记 LDA)。

进一步，将对数几率按照类别k与l，合并同类项可得：

$$
\begin{aligned}
&\ln \frac{P(y_i = k \mid \mathbf{x} = \mathbf{x}_i)}{P(y_i = l \mid \mathbf{x} = \mathbf{x}_i)} \\
&= \underbrace{\left[\ln \pi_k - \frac{1}{2}\boldsymbol{\mu}_k' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \mathbf{x}_i' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k \right]}_{\equiv \delta_k(\mathbf{x}_i)} - \underbrace{\left[\ln \pi_l - \frac{1}{2}\boldsymbol{\mu}_l' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_l + \mathbf{x}_i' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_l \right]}_{\equiv \delta_l(\mathbf{x}_i)} \\
&\equiv \delta_k(\mathbf{x}_i) - \delta_l(\mathbf{x}_i)
\end{aligned}
\tag{7.9}
$$

其中，$\delta_k(\mathbf{x}_i) \equiv \ln \pi_k - \frac{1}{2}\boldsymbol{\mu}_k' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \mathbf{x}_i' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k$ 称为**线性判别函数** (linear discriminant function)。

由方程(7.9)可知，最优决策规则为选择类别k，使得线性判别函数最大化：

$$\max_{k} \ \delta_k(\mathbf{x}_i) \qquad (7.10)$$

在实践中，需要估计线性判别函数$\delta_k(\mathbf{x}_i)$中的未知总体参数π_k，$\boldsymbol{\mu}_k$与$\boldsymbol{\Sigma}$。根据训练数据$\{\mathbf{x}_i, y_i\}_{i=1}^n$，只要计算这些总体参数的相应样本估计量即可。

先验概率π_k的估计量为

$$\hat{\pi}_k = \frac{n_k}{n} \quad (k = 1, \cdots, K) \tag{7.11}$$

其中，n_k为训练样本中第k类数据的样例数。

每类数据的期望值$\boldsymbol{\mu}_k$之估计量为样本均值$\hat{\boldsymbol{\mu}}_k$：

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{n_k} \sum_{y_i = k} \mathbf{x}_i \quad (k = 1, \cdots, K) \tag{7.12}$$

第k类数据的协方差矩阵$\boldsymbol{\Sigma}_k$之估计量为样本协方差矩阵$\hat{\boldsymbol{\Sigma}}_k$：

$$\hat{\boldsymbol{\Sigma}}_k = \frac{1}{n_k - 1}\sum_{y_i = k}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)' \equiv \frac{\mathbf{S}_k}{n_k - 1} \tag{7.13}$$

其中，$\mathbf{S}_k \equiv \sum_{y_i = k}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)'$称为第 k 类数据的“散度矩阵”

(scatter matrix)，而$(n_k - 1)$为自由度(在估计$\hat{\boldsymbol{\mu}}_k$时损失了一个自由度)。

在每类数据的协方差矩阵均相等的假设下(即$\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$)，可用每类数据的样本协方差矩阵$\hat{\boldsymbol{\Sigma}}_k$之加权平均(权重为每类数据在样本中的比重，经过自由度调整)，来估计整个样本的协方差矩阵：

$$\hat{\boldsymbol{\Sigma}} = \sum_{k=1}^{K} \hat{\boldsymbol{\Sigma}}_k \cdot \underbrace{\left(\frac{n_k - 1}{n - K}\right)}_{\text{权重}} = \sum_{k=1}^{K} \frac{\mathbf{S}_k}{n_k - 1} \cdot \left(\frac{n_k - 1}{n - K}\right) = \frac{1}{n - K} \sum_{k=1}^{K} \mathbf{S}_k \tag{7.14}$$

7.3 二次判别分析

线性判别分析假设所有类别的协方差矩阵均相同，即$\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$，该假设可能与现实数据相悖。

Smith(1946)放松了此假定，允许不同类别的协方差矩阵不同。

此时，仍可考虑第k类与第l类($k \neq l$)之间的对数几率：

$$\ln \frac{P(y_i = k \mid \mathbf{x} = \mathbf{x}_i)}{P(y_i = l \mid \mathbf{x} = \mathbf{x}_i)}$$

$$= \ln \frac{\pi_k}{\pi_l} - \frac{1}{2} \ln \left(\frac{|\mathbf{\Sigma}_k|}{|\mathbf{\Sigma}_l|} \right) - \frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)' \mathbf{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) + \frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_l)' \mathbf{\Sigma}_l^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_l) \tag{7.15}$$

由方程“$\ln \frac{P(y_i = k \mid \mathbf{x} = \mathbf{x}_i)}{P(y_i = l \mid \mathbf{x} = \mathbf{x}_i)} = 0$”所决定的决策边界为二次(型)函数，故称为**二次判别分析**(Quadratic Discriminant Analysis，简记 QDA)。

进一步，可将对数几率的表达式，按照类别k与l合并同类项：

$$
\begin{aligned}
&\ln \frac{P(y_i = k \mid \mathbf{x} = \mathbf{x}_i)}{P(y_i = l \mid \mathbf{x} = \mathbf{x}_i)} \\
&= \underbrace{\left[\ln \pi_k - \frac{1}{2}\ln\left|\boldsymbol{\Sigma}_k\right| - \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k) \right]}_{\equiv \delta_k(\mathbf{x}_i)} \\
&\quad - \underbrace{\left[\ln \pi_l - \frac{1}{2}\ln\left|\boldsymbol{\Sigma}_l\right| - \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_l)'\boldsymbol{\Sigma}_l^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_l) \right]}_{\equiv \delta_l(\mathbf{x}_i)} \qquad (7.16) \\
&\equiv \delta_k(\mathbf{x}_i) - \delta_l(\mathbf{x}_i)
\end{aligned}
$$

其中，$\delta_k(\mathbf{x}_i) \equiv \ln \pi_k - \frac{1}{2}\ln\left|\boldsymbol{\Sigma}_k\right| - \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)$称为**二次判别函数**(quadratic discriminant function)。

最优决策规则为选择类别k，使得二次判别函数$\delta_k(\mathbf{x}_i)$最大化。在实践中，可使用$\hat{\mathbf{\Sigma}}_k = \frac{1}{n_k - 1}\sum_{y_i=k}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)'$估计$\mathbf{\Sigma}_k$。

7.4 费雪线性判别分析

与 Mahalanobis(1936)基于正态分布的判别分析不同，Fisher(1936)从数据降维的角度来考虑线性判别问题，称为**费雪线性判别法**(Fisher Linear Discriminant Analysis)。

基本思想：能否将特征向量$\mathbf{x}_i$作适当的线性组合$\mathbf{w}'\mathbf{x}_i$，使得数据变得更容易分离？

首先考虑二分类问题。

假设训练样本为$\left\{\mathbf{x}_i, y_i\right\}_{i=1}^n$，其中$\mathbf{x}_i = (x_{i1} \cdots x_{ip})'$为$p$维特征向量，而响应变量$y_i \in \left\{1, 2\right\}$分为两类。

由于p可能很大，为高维数据，故考虑特征向量的线性组合$z_i = \mathbf{w}'\mathbf{x}_i$，其中$\mathbf{w} = (w_1 \cdots w_p)'$为此线性组合的“权重”(weights)。

然后，通过此一维标量z_i来进行样本分类。

为图示方便，不妨暂时假设只有两个特征变量，即 $p=2$。

样本点$\mathbf{x}_i=(x_{i1}\ x_{i2})'$，其坐标$(x_{i1},x_{i2})$可分别视为向横轴($x_1$轴)与纵轴($x_2$轴)的“投影”(projection)。

类似地，在几何上可将内积$\mathbf{w}'\mathbf{x}_i$解释为向量$\mathbf{x}_i$朝向量$\mathbf{w}$所作的投影(参见图 7.2)。

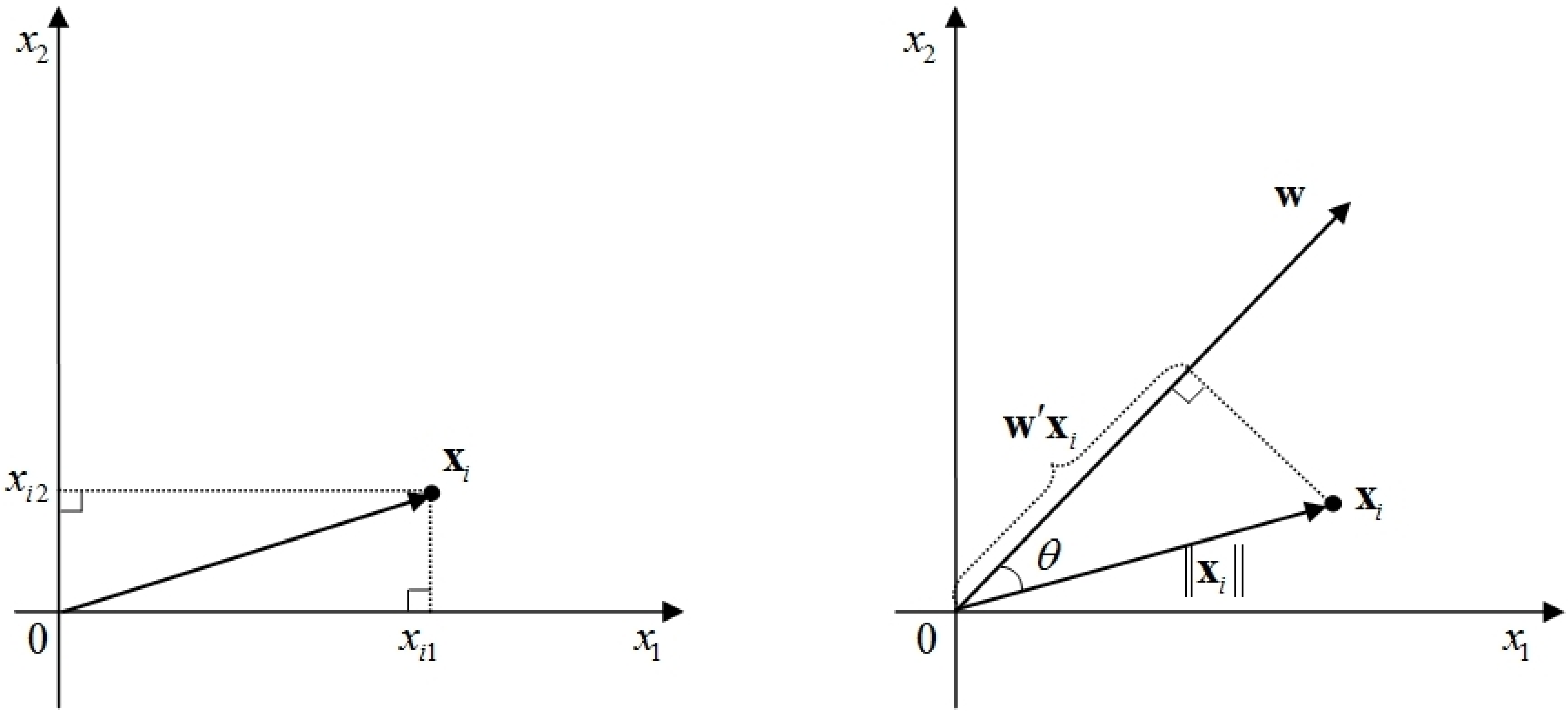

图 7.2　特征向量$\mathbf{x}_i$向坐标轴(左图)与向量$\mathbf{w}$(右图)的投影

记特征向量$\mathbf{x}_i$与权重向量$\mathbf{w}$的夹角为θ，则根据线性代数可知：

$$\cos\theta = \frac{\mathbf{w}'\mathbf{x}_i}{\|\mathbf{w}\| \cdot \|\mathbf{x}_i\|} \qquad (7.17)$$

在将$\mathbf{x}_i$朝向量$\mathbf{w}$作投影时，向量$\mathbf{w}$的长度$\|\mathbf{w}\|$无关紧要，故不妨令$\|\mathbf{w}\| = 1$。故上式可写为

$$\mathbf{w}'\mathbf{x}_i = \|\mathbf{x}_i\| \cos\theta \qquad (7.18)$$

$\mathbf{w}'\mathbf{x}_i$正是$\mathbf{x}_i$朝向量$\mathbf{w}$所作的投影。希望找到一个投影方向$\mathbf{w}$，使得在将特征向量$\left\{\mathbf{x}_i\right\}_{i=1}^n$投影之后，最容易地将两类样本区分开，参见图 7.3。

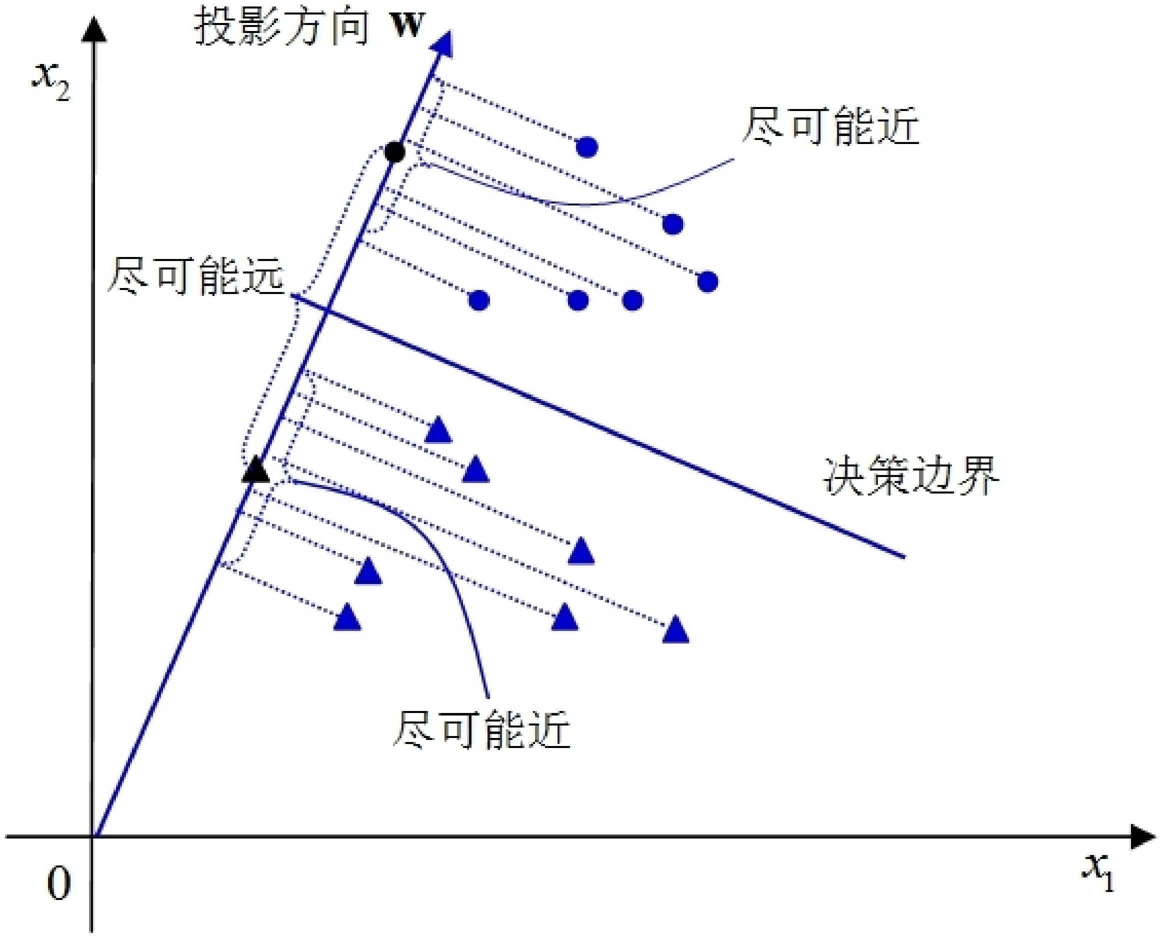

图 7.3　费雪判别法示意图

在图 7.3 中，分别以圆点与三角形表示两类数据。希望投影之后两类样本的中心位置(图中黑圆点与黑三角)离得尽可能远，而两类样本内部的点之间，则离得尽可能地近。

在给定“组内方差”(within-class variance)的情况下，希望“组间方差”(between-class variance)最大。

首先考虑组间方差。分别记第 1 类与第 2 类数据之特征向量的样本均值为$\hat{\boldsymbol{\mu}}_1$与$\hat{\boldsymbol{\mu}}_2$：

$$\hat{\boldsymbol{\mu}}_1 = \frac{1}{n_1}\sum_{y_i=1}\mathbf{x}_i,\quad \hat{\boldsymbol{\mu}}_2 = \frac{1}{n_2}\sum_{y_i=2}\mathbf{x}_i \tag{7.19}$$

其中，n_1为第 1 类数据的样例数，而n_2为第 2 类数据的样例数。

两类数据的中心位置经过投影变换之后，分别变为$\overline{z}_1 \equiv \mathbf{w}'\hat{\boldsymbol{\mu}}_1$与$\overline{z}_2 \equiv \mathbf{w}'\hat{\boldsymbol{\mu}}_2$。因此，投影之后两类样本中心位置的差距为

$$\overline{z}_1 - \overline{z}_2 = \mathbf{w}'(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2) \qquad (7.20)$$

进一步，可将投影后中心位置之间距离的平方视为“组间方差”(between-class variance)：

$$(\overline{z}_1 - \overline{z}_2)^2 = \mathbf{w}'(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)\left[\mathbf{w}'(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)\right]' = \mathbf{w}'\underbrace{(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)'}_{\equiv \mathbf{S}_B}\mathbf{w} \equiv \mathbf{w}'\mathbf{S}_B\mathbf{w}$$

(7.21)

在上式中，虽然组间方差$(\bar{z}_1 - \bar{z}_2)^2$为标量，但在形式上依然可写为二次型(为后面推导方便)，其中二次型矩阵$\mathbf{S}_B \equiv (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)'$称为“组间散度矩阵”(between-class scatter matrix)。

对于第 1 类数据，其投影之后的“组内方差”(within-class variance)可写为

$$
\begin{aligned}
\hat{s}_1 &= \sum_{y_i=1} (z_i - \overline{z}_1)^2 = \sum_{y_i=1} (\mathbf{w}'\mathbf{x}_i - \mathbf{w}'\hat{\boldsymbol{\mu}}_1)^2 \\
&= \sum_{y_i=1} \mathbf{w}'(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_1)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_1)'\mathbf{w} \\
&= \mathbf{w}'\left[\sum_{y_i=1} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_1)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_1)'\right]\mathbf{w} \\
&= \mathbf{w}'\mathbf{S}_1\mathbf{w}
\end{aligned}
\tag{7.22}
$$

其中，$\mathbf{S}_1 \equiv \sum_{y_i=1} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_1)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_1)'$为第 1 类数据在特征空间的散度矩阵(scatter matrix in feature space)。

类似地，第 2 类数据投影之后的组内方差可写为

$$\hat{s}_2 = \sum_{y_i=2} (z_i - \overline{z}_2)^2 = \mathbf{w}' \left[\sum_{y_i=2} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_2)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_2)' \right] \mathbf{w} = \mathbf{w}' \mathbf{S}_2 \mathbf{w} \tag{7.23}$$

其中，$\mathbf{S}_2 \equiv \sum_{y_i=2} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_2)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_2)'$为第 2 类数据在特征空间的散度矩阵。因此，两类数据投影之后的组内方差之和为

$$\hat{s}_1 + \hat{s}_2 = \mathbf{w}'\mathbf{S}_1\mathbf{w} + \mathbf{w}'\mathbf{S}_2\mathbf{w} = \mathbf{w}'(\mathbf{S}_1 + \mathbf{S}_2)\mathbf{w} \equiv \mathbf{w}'\mathbf{S}_W\mathbf{w} \tag{7.24}$$

其中，$\mathbf{S}_W \equiv \mathbf{S}_1 + \mathbf{S}_2$称为“组内散度矩阵”(within-class scatter matrix)。

费雪线性判别的最优化目标为，在给定组内方差的情况下，最大化组间方差。故可将最优化问题的目标函数写为

$$\max_{\mathbf{w}} \ J(\mathbf{w}) = \frac{(\bar{z}_1 - \bar{z}_2)^2}{\hat{s}_1 + \hat{s}_2} = \frac{\mathbf{w}'\mathbf{S}_B\mathbf{w}}{\mathbf{w}'\mathbf{S}_W\mathbf{w}} \qquad (7.25)$$

此目标函数也称为“费雪准则”(Fisher criterion)。

最优解$\hat{\mathbf{w}}$与其长度无关。如果$\hat{\mathbf{w}}$是最优解，则对于任意$\alpha \neq 0$，$\alpha\hat{\mathbf{w}}$也是最优解(可在上式的分子与分母同时消去α^2)。

不失一般性，令分母$\mathbf{w}'\mathbf{S}_W\mathbf{w} = 1$。

将上述无约束的最大化问题等价地写为有约束的最大化问题：

$$\begin{aligned}&\max_{\mathbf{w}}\ \ \mathbf{w}'\mathbf{S}_B\mathbf{w}\\&s.t.\ \ \mathbf{w}'\mathbf{S}_W\mathbf{w}=1\end{aligned}\tag{7.26}$$

为求解此约束极值问题，引入拉格朗日乘子函数：

$$L(\mathbf{w},\lambda)=\mathbf{w}'\mathbf{S}_B\mathbf{w}+\lambda(1-\mathbf{w}'\mathbf{S}_W\mathbf{w})\tag{7.27}$$

将上式对$\mathbf{w}$求偏导数，根据二次型的向量微分规则，并注意到$\mathbf{S}_B$与$\mathbf{S}_W$均为对称矩阵，可得一阶条件：

$$\frac{\partial L(\mathbf{w},\lambda)}{\partial \mathbf{w}} = 2\mathbf{S}_B\mathbf{w} - 2\lambda\mathbf{S}_W\mathbf{w} = \mathbf{0} \tag{7.28}$$

经移项整理可得：

$$\mathbf{S}_B\mathbf{w} = \lambda\mathbf{S}_W\mathbf{w} \tag{7.29}$$

在上式两边同时左乘$\mathbf{S}_W^{-1}$可得：

$$\underbrace{(\mathbf{S}_W^{-1}\mathbf{S}_B)}_{=\mathbf{A}}\mathbf{w} = \lambda\mathbf{w} \tag{7.30}$$

上式可视为 “$\mathbf{A}\mathbf{w} = \lambda\mathbf{w}$”。这正是线性代数的“特征值问题”(eigenvalue problem)，其中λ为特征值，$\mathbf{w}$为相应的特征向量(eigenvector)，而$\mathbf{A} = \mathbf{S}_W^{-1}\mathbf{S}_B$。故最优解为矩阵$\mathbf{S}_W^{-1}\mathbf{S}_B$的特征值与特征向量。

矩阵$\mathbf{S}_W^{-1}\mathbf{S}_B$的特征向量通常不止一个。最优解究竟是哪个特征值及其相应的特征向量呢？将方程(7.29)代入目标函数(7.26)可得：

$$\max_{\mathbf{w}} \ \mathbf{w}'\mathbf{S}_B\mathbf{w} = \mathbf{w}'\lambda\mathbf{S}_W\mathbf{w} = \lambda\underbrace{\mathbf{w}'\mathbf{S}_W\mathbf{w}}_{=1} = \lambda \tag{7.31}$$

其中，根据约束条件$\mathbf{w}'\mathbf{S}_W\mathbf{w} = 1$。

目标函数的最大值正是矩阵$\mathbf{S}_W^{-1}\mathbf{S}_B$的特征值$\lambda$。为了最大化此目标函数，应该选择最大的特征值，记为λ_1，并记其特征向量为$\mathbf{a}_1$(须将特征向量$\mathbf{a}_1$标准化，使得$\mathbf{a}_1'\mathbf{S}_W\mathbf{a}_1 = 1$)。

由此可得，最优投影方向为$\hat{\mathbf{w}} = \mathbf{a}_1'$，并称$z_i = \hat{\mathbf{w}}'\mathbf{x}_i$为**线性判别变量**，简称**线性判元**(linear discriminant)。

对于此具体问题，还有更简洁的解法。将组间散度矩阵$\mathbf{S}_B$的表达式$\mathbf{S}_B \equiv (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)'$代入方程(7.29)可得，

$$\lambda \mathbf{S}_W \mathbf{w} = \mathbf{S}_B \mathbf{w} = (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)\underbrace{(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)'\mathbf{w}}_{=c\in\mathbb{R}} \equiv c(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2) \tag{7.32}$$

其中，记$(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)'\mathbf{w}$为某常数$c \in \mathbb{R}$。

在上式两边同时左乘$\dfrac{1}{\lambda}\mathbf{S}_W^{-1}$可得，

$$\mathbf{w}=\frac{c}{\lambda}\mathbf{S}_W^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) \qquad (7.33)$$

显然，常数$\dfrac{c}{\lambda}$并不影响向量$\mathbf{w}$的方向，而我们只在乎$\mathbf{w}$的方向，故最优解可写为

$$\hat{\mathbf{w}}=\mathbf{S}_W^{-1}(\hat{\boldsymbol{\mu}}_1-\hat{\boldsymbol{\mu}}_2) \qquad (7.34)$$

严格来说，仍应将$\hat{\mathbf{w}}$标准化，使得$\hat{\mathbf{w}}'\mathbf{S}_W\hat{\mathbf{w}}=1$。

由此所得的最佳投影$z_i = \hat{\mathbf{w}}' \mathbf{x}_i$，即为**线性判元**(linear discriminant)或**线性判别得分**(linear discriminant score)。

最佳投影方向$\hat{\mathbf{w}}$称为**线性判别载荷**(linear discriminant loading)，**线性判别系数**(linear discriminant coefficients)或**判别坐标**(discriminant coordinate)。

对于一个新的样本点$\mathbf{x}_0$，可根据其线性判别得分$z_0 = \hat{\mathbf{w}}' \mathbf{x}_0$与两类数据投影的中心位置$\bar{z}_1 = \hat{\mathbf{w}}' \hat{\boldsymbol{\mu}}_1$与$\bar{z}_2 = \hat{\mathbf{w}}' \hat{\boldsymbol{\mu}}_2$的距离远近进行分类，即归入距离更近的那一类。

在具体操作上，如果z_0比$(\hat{\boldsymbol{\mu}}_1+\hat{\boldsymbol{\mu}}_2)/2$的投影更大，则可将$\mathbf{x}_0$归入第 1 类，即

$$z_0=\hat{\mathbf{w}}'\mathbf{x}_0\geq\hat{\mathbf{w}}'(\hat{\boldsymbol{\mu}}_1+\hat{\boldsymbol{\mu}}_2)/2 \qquad (7.35)$$

反之，则将$\mathbf{x}_0$归入第 2 类。进一步，还可定义如下**线性分类函数**(linear classification function)：

$$\mathbf{x}'\hat{\mathbf{w}}-\frac{1}{2}(\hat{\boldsymbol{\mu}}_1+\hat{\boldsymbol{\mu}}_2)'\hat{\mathbf{w}}=\underbrace{-\frac{1}{2}(\hat{\boldsymbol{\mu}}_1+\hat{\boldsymbol{\mu}}_2)'\mathbf{S}_W^{-1}(\hat{\boldsymbol{\mu}}_1-\hat{\boldsymbol{\mu}}_2)}_{constant}+\underbrace{\mathbf{x}'\mathbf{S}_W^{-1}(\hat{\boldsymbol{\mu}}_1-\hat{\boldsymbol{\mu}}_2)}_{linear}$$

$$(7.36)$$

然后，根据线性分类函数的取值正负对$\mathbf{x}$进行分类。

7.5 费雪线性判别与基于正态的线性判别之关系

在进行基于正态的线性判别分析时，如果假设先验概率相等，则等价于费雪的线性判别分析。

在进行正态判别分析时，根据对数后验几率的取值正负来分类。

下面证明，费雪判别分析的线性分类函数，等价于正态判别分析的对数后验几率函数。

在进行基于正态的线性判别分析时，第 1 类与第 2 类数据的“对数后验几率”(log posterior odds)的样本估计值为：

$$\widehat{\ln\frac{P(y=1\,|\,\mathbf{x})}{P(y=2\,|\,\mathbf{x})}}=\left(\ln\frac{\pi_1}{\pi_2}-\frac{1}{2}\hat{\boldsymbol{\mu}}_1'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}_1+\frac{1}{2}\hat{\boldsymbol{\mu}}_2'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}_2\right)+\mathbf{x}'\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_1-\hat{\boldsymbol{\mu}}_2)$$

$$=\left(-\frac{1}{2}\hat{\boldsymbol{\mu}}_1'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}_1+\frac{1}{2}\hat{\boldsymbol{\mu}}_2'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}_2\right)+\mathbf{x}'\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_1-\hat{\boldsymbol{\mu}}_2)$$

(7.37)

其中，假设先验概率$\pi_1=\pi_2$，故$\ln\dfrac{\pi_1}{\pi_2}=0$。

将上式与费雪判别法的线性分类函数(7.36)进行对比。

在两类数据的协方差矩阵相等的假设下，$\hat{\boldsymbol{\Sigma}}^{-1}$与$\mathbf{S}_W^{-1}$仅相差$(n-K)$倍，故可忽略其差别。

不妨令$\mathbf{S}_W^{-1}=\hat{\boldsymbol{\Sigma}}^{-1}$，则上式的一次项与线性分类函数(7.36)的一次项相等，即$\mathbf{x}'\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_1-\hat{\boldsymbol{\mu}}_2)=\mathbf{x}'\mathbf{S}_W^{-1}(\hat{\boldsymbol{\mu}}_1-\hat{\boldsymbol{\mu}}_2)$。

进一步，线性分类函数(7.36)的常数项也与方程(7.37)的常数项相等：

$$
\begin{aligned}
-\frac{1}{2}(\hat{\boldsymbol{\mu}}_1+\hat{\boldsymbol{\mu}}_2)'\mathbf{S}_W^{-1}(\hat{\boldsymbol{\mu}}_1-\hat{\boldsymbol{\mu}}_2) &= -\frac{1}{2}\left(\hat{\boldsymbol{\mu}}_1'+\hat{\boldsymbol{\mu}}_2'\right)\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_1-\hat{\boldsymbol{\mu}}_2) \\
&= -\frac{1}{2}\left[\hat{\boldsymbol{\mu}}_1'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}_1 \cancel{-\hat{\boldsymbol{\mu}}_1'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}_2} + \cancel{\hat{\boldsymbol{\mu}}_2'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}_1} - \hat{\boldsymbol{\mu}}_2'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}_2\right] \\
&= -\frac{1}{2}\hat{\boldsymbol{\mu}}_1'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}_1 + \frac{1}{2}\hat{\boldsymbol{\mu}}_2'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}_2
\end{aligned}
\tag{7.38}
$$

由此可知，如果假设先验概率相等，则费雪线性判别与基于正态的线性判别等价，使用二者所得的决策边界完全相同。

此结论在多分类问题中依然成立(Johnson and Wichern, 2007)。

费雪判别法可视为线性判别的特例。

费雪判别法作为一种“有监督”(supervised)的降维方法，依然具有独特的价值。

费雪判别法的降维效果往往优于“非监督”(unsupervised)的降维方法，比如主成分分析(参见第 16 章)。

如果先验概率不相等，则这两种判别法的决策边界并不相同。

在软件的实际操作中，一般同时进行费雪线性判别分析与基于正态的线性判别分析，前者作为降维可视化工具，提供线性判别变量 $(lda_1,\cdots,lda_{K-1})$；而后者则用于计算后验概率 $P(y_i=k\,|\,\mathbf{x}_i)$。

7.6 多分类问题的费雪判别分析

考虑将二分类问题的费雪判别分析推广到多分类问题。

对于二分类问题，只有一个最佳投影方向$\mathbf{w}$，以及相应的线性判元$\mathbf{w}'\mathbf{x}$。

对于K分类问题，记$y_i \in \{1,\cdots,K\}$，则一般可以有$(K-1)$个最佳投影方向$\{\mathbf{w}_1,\cdots,\mathbf{w}_{K-1}\}$，以及相应的$(K-1)$个线性判元(linear discriminants)$\{\mathbf{w}_1'\mathbf{x},\cdots,\mathbf{w}_{K-1}'\mathbf{x}\}$。

但线性判元的个数也受到属性个数p(即特征向量$\mathbf{x}$的维度)的限制。

线性判元的个数为$\min\{K-1, p\}$，即$(K-1)$与p二者之更小者。

这是因为如果属性个数$p<(K-1)$，则$(K-1)$个p维的投影向量$\{\mathbf{w}_1,\cdots,\mathbf{w}_{K-1}\}$之间必然存在严格多重共线性(根据线性代数知识，在$p$维空间中，一个线性无关的向量组最多只包含$p$个向量)，使得$\{\mathbf{w}_1,\cdots,\mathbf{w}_{K-1}\}$包含多余(可由其他向量线性表出)的向量。

在大多数情况下$p\geq(K-1)$，故为叙述方便，假设存在$(K-1)$个线性判元。

这$(K-1)$个线性判元意味着，将原来的p维特征向量$\mathbf{x}_i$(p可能较大)，通过适当的$(K-1)$个线性组合，寻找$(K-1)$个最佳投影方向(通常要求这些线性判元之间互不相关)，降到更低的$(K-1)$维。

比如，对于三分类问题，$K=3$，则有$(K-1)=2$个线性判元。

此时，每个样本点$\{\mathbf{x}_i\}_{i=1}^n$均有两个线性判别得分，记为(lda_1, lda_2)；然后可在线性判元(lda_1, lda_2)的二维空间中画图，进行可视化分析。

即使对于K较大的多分类问题，通常也可以只在第一与第二线性判元(lda_1, lda_2)的二维空间画图展示，因为第一线性判元对于组间方差的贡献最大，第二线性判元对于组间方差的贡献其次，以此类推。

如果需要，也可在(lda_1, lda_2, lda_3)的三维空间展示。

可以证明，对于多分类问题的费雪判别分析，其最优解为矩阵$\mathbf{S}_W^{-1}\mathbf{S}_B$的系列特征值$\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_s > 0$(其中$s \leq \min(K-1, p)$，因为有些特征值可能为 0)，以及相应的特征向量$\left\{\hat{\mathbf{w}}_1, \cdots, \hat{\mathbf{w}}_s\right\}$；但$\mathbf{S}_W$与$\mathbf{S}_B$的定义有所不同(参见本章附录)。

进一步，第一线性判元对于(投影前原始数据)组间方差的贡献率为 $\hat{\lambda}_1 \big/ (\hat{\lambda}_1 + \cdots + \hat{\lambda}_s)$，而第二线性判元对于组间方差的贡献率为 $\hat{\lambda}_2 \big/ (\hat{\lambda}_1 + \cdots + \hat{\lambda}_s)$，以此类推。

通常，第一与第二线性判元对于组间方差的累积贡献率 $(\hat{\lambda}_1 + \hat{\lambda}_2) \big/ (\hat{\lambda}_1 + \cdots + \hat{\lambda}_s)$可能就比较高，甚至接近于 1，从而达到降维的目的。

在统计软件中，一般称“$(\hat{\lambda}_1 + \cdots + \hat{\lambda}_s)$”为“迹”（trace），即对角矩阵$[\hat{\lambda}_1, \cdots, \hat{\lambda}_s]$的主对角线元素之和。

在对一个新样本点$\mathbf{x}_0$进行分类时，首先将它投影到$(K-1)$维的线性判元之空间，然后考察它与K类样本投影后之中心位置的距离远近(使用欧氏距离来度量)，将其归入最近的那一类数据即可。

在直观思想上，多分类问题的费雪判别分析与二分类问题类似，但前者的代数推导较为繁琐(参见本章附录)。

在不同的假定下，判别分析后来又发展出“正则判别分析”(regularized discriminant analysis)、“灵活判别分析”(flexible discriminant analysis)、“混合判别分析”(mixture discriminant analysis)等，参见 Hastie et al. (2009)，在此从略。

7.7 判别分析的 Python 案例

以 `iris` 数据为例，介绍判别分析的 Python 操作。

* 详见教材，以及配套 Python 程序（现场演示）。

附录

A 7.1 总体中的多分类费雪判别分析

首先考虑在总体中进行多分类问题的费雪线性判别分析(假设所有总体参数为已知)，然后再推广到样本数据中。

记第k类数据的总体均值为$\boldsymbol{\mu}_k = \mathrm{E}(\mathbf{x} \mid y = k)$，其中$k = 1, \cdots, K$；而所有数据的总体均值为$\boldsymbol{\mu} = \dfrac{1}{K}\sum_{k=1}^{K}\boldsymbol{\mu}_k$(简单起见，假设各类数据所占比重相同)。

对于投影方向$\mathbf{w}$，考虑线性组合$z=\mathbf{w}'\mathbf{x}$。对于第k类数据，此线性组合的均值为

$$\mu_z^{(k)} \equiv \mathrm{E}(z \mid y=k) = \mathbf{w}'\,\mathrm{E}(\mathbf{x} \mid y=k) = \mathbf{w}'\boldsymbol{\mu}_k \tag{7.39}$$

对于所有数据，线性组合z的均值为

$$\overline{\mu}_z = \frac{1}{n}\sum_{k=1}^{K}\mu_z^{(k)} = \frac{1}{n}\sum_{k=1}^{K}\mathbf{w}'\boldsymbol{\mu}_k = \mathbf{w}'\left(\frac{1}{n}\sum_{k=1}^{K}\boldsymbol{\mu}_k\right) = \mathbf{w}'\boldsymbol{\mu} \tag{7.40}$$

假设各类数据的协方差矩阵均相等，即$\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_K = \boldsymbol{\Sigma}$，则线性组合$\mathbf{z} = \mathbf{w}'\mathbf{x}$的协方差矩阵为(无论为哪类数据)：

$$\mathrm{Var}(z) = \mathrm{Var}(\mathbf{w}'\mathbf{x}) = \mathbf{w}'\,\mathrm{Var}(\mathbf{x})\mathbf{w} = \mathbf{w}'\boldsymbol{\Sigma}\mathbf{w} \tag{7.41}$$

作为对组间方差的度量，考虑投影后每类数据的中心($\mu_z^{(k)}$)与所有数据的中心($\overline{\mu}_z$)之距离的平方和：

$$\sum_{k=1}^{K}(\mu_z^{(k)}-\bar{\mu}_z)^2=\sum_{k=1}^{K}(\mathbf{w}'\boldsymbol{\mu}_k-\mathbf{w}'\boldsymbol{\mu})^2=\mathbf{w}'\left[\sum_{k=1}^{K}(\boldsymbol{\mu}_k-\boldsymbol{\mu})(\boldsymbol{\mu}_k-\boldsymbol{\mu})'\right]\mathbf{w}\equiv\mathbf{w}'\mathbf{B}\mathbf{w}$$

其中，二次型矩阵$\mathbf{B}\equiv\sum_{k=1}^{K}(\boldsymbol{\mu}_k-\boldsymbol{\mu})(\boldsymbol{\mu}_k-\boldsymbol{\mu})'$为组间散度矩阵。

在总体中，费雪准则(Fisher criterion)的目标函数可写为

$$\max_{\mathbf{w}}\ J(\mathbf{w})=\frac{\mathbf{w}'\mathbf{B}\mathbf{w}}{\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w}} \qquad (7.42)$$

由于$\boldsymbol{\Sigma}$为总体协方差矩阵，故为对称正定矩阵，可将其对角化，即$\boldsymbol{\Sigma} = \mathbf{P}'\boldsymbol{\Lambda}\mathbf{P}$，其中$\boldsymbol{\Lambda}$为对角矩阵，且其主对角线元素(即特征值)均大于 0；而$\mathbf{P}$为正交矩阵，包含相应的特征向量。

进一步，由于$\boldsymbol{\Sigma}$为对称正定矩阵，故可定义其“平方根矩阵”(square root matrix)$\boldsymbol{\Sigma}^{1/2}$，满足$\boldsymbol{\Sigma}^{1/2} \cdot \boldsymbol{\Sigma}^{1/2} = \boldsymbol{\Sigma}$；以及其逆矩阵$\boldsymbol{\Sigma}^{-1/2}$。

为简化目标函数的分母，作如下变换：

$$\mathbf{v} \equiv \boldsymbol{\Sigma}^{1/2}\mathbf{w}, \quad \mathbf{w} \equiv \boldsymbol{\Sigma}^{-1/2}\mathbf{v} \qquad (7.43)$$

可将目标函数(7.42)的分母变为更简单的平方和形式：

$$\mathbf{w}'\mathbf{\Sigma}\mathbf{w} = (\mathbf{\Sigma}^{-1/2}\mathbf{v})'\mathbf{\Sigma}(\mathbf{\Sigma}^{-1/2}\mathbf{v}) = \mathbf{v}'\mathbf{\Sigma}^{-1/2}\mathbf{\Sigma}\mathbf{\Sigma}^{-1/2}\mathbf{v} = \mathbf{v}'\mathbf{v} \qquad (7.44)$$

其中，$\mathbf{v}'\mathbf{v}$ 为向量内积，即平方和。另一方面，目标函数(7.42)的分子可写为

$$\mathbf{w}'\mathbf{B}\mathbf{w} = (\mathbf{\Sigma}^{-1/2}\mathbf{v})'\mathbf{B}(\mathbf{\Sigma}^{-1/2}\mathbf{v}) = \mathbf{v}'\mathbf{\Sigma}^{-1/2}\mathbf{B}\mathbf{\Sigma}^{-1/2}\mathbf{v} \qquad (7.45)$$

因此，最大化问题变为

$$\max_{\mathbf{v}} \; J(\mathbf{v}) = \frac{\mathbf{v}'\mathbf{\Sigma}^{-1/2}\mathbf{B}\mathbf{\Sigma}^{-1/2}\mathbf{v}}{\mathbf{v}'\mathbf{v}} \qquad (7.46)$$

由于最优的$\mathbf{v}$与其长度$\|\mathbf{v}\|$无关，故此无约束优化问题等价于以下约束极值问题：

$$\begin{aligned} \max_{\mathbf{v}} \ \ & \mathbf{v}'\mathbf{\Sigma}^{-1/2}\mathbf{B}\mathbf{\Sigma}^{-1/2}\mathbf{v} \\ s.t. \ \ & \mathbf{v}'\mathbf{v}=1 \end{aligned} \tag{7.47}$$

引入拉格朗日乘子函数：

$$L(\mathbf{v},\lambda)=\mathbf{v}'\mathbf{\Sigma}^{-1/2}\mathbf{B}\mathbf{\Sigma}^{-1/2}\mathbf{v}+\lambda(1-\mathbf{v}'\mathbf{v}) \tag{7.48}$$

其中，λ为拉格朗日乘子。对$\mathbf{v}$求偏导数，根据向量微分规则，可得一阶条件：

$$\frac{\partial L(\mathbf{v},\lambda)}{\partial \mathbf{v}} = 2\boldsymbol{\Sigma}^{-1/2}\mathbf{B}\boldsymbol{\Sigma}^{-1/2}\mathbf{v} - 2\lambda\mathbf{v} = \mathbf{0} \qquad (7.49)$$

经移项整理可得：

$$(\boldsymbol{\Sigma}^{-1/2}\mathbf{B}\boldsymbol{\Sigma}^{-1/2})\mathbf{v} = \lambda\mathbf{v} \qquad (7.50)$$

故最优解$\mathbf{v}$为矩阵$(\boldsymbol{\Sigma}^{-1/2}\mathbf{B}\boldsymbol{\Sigma}^{-1/2})$的特征向量，而$\lambda$为相应的特征值。

记$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_s > 0$为$(\boldsymbol{\Sigma}^{-1/2}\mathbf{B}\boldsymbol{\Sigma}^{-1/2})$的$s$个非零特征值(其中$s \leq \min(K-1, p)$)，其相应的特征向量为$\mathbf{e}_1,\cdots,\mathbf{e}_s$(将每个特征向量$\mathbf{e}$标准化，使得$\mathbf{e}'\mathbf{e} = 1$)。

将一阶条件(7.50)代入目标函数(7.47)可得：

$$\max_{\mathbf{v}} \ \mathbf{v}'\mathbf{\Sigma}^{-1/2}\mathbf{B}\mathbf{\Sigma}^{-1/2}\mathbf{v} = \lambda \underbrace{\mathbf{v}'\mathbf{v}}_{=1} = \lambda$$

其中，根据约束条件 $\mathbf{v}'\mathbf{v} = 1$。故目标函数的最大值正是矩阵 $(\mathbf{\Sigma}^{-1/2}\mathbf{B}\mathbf{\Sigma}^{-1/2})$ 的特征值 λ。

为了最大化此目标函数，应该选择最大的特征值 λ_1 及其相应的特征向量 $\mathbf{e}_1$。根据定义，$\mathbf{w} \equiv \mathbf{\Sigma}^{-1/2}\mathbf{v}$，故 $\mathbf{w}_1 \equiv \mathbf{\Sigma}^{-1/2}\mathbf{e}_1$。

“第一线性判元”(first linear discriminant)$\mathbf{w}_1'\mathbf{x}$的方差为

$$\mathrm{Var}(\mathbf{w}_1'\mathbf{x}) = \mathbf{w}_1' \mathrm{Var}(\mathbf{x})\mathbf{w}_1 = \mathbf{w}_1'\boldsymbol{\Sigma}\mathbf{w}_1 = \mathbf{e}_1'\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1/2}\mathbf{e}_1 = \mathbf{e}_1'\mathbf{e}_1 = 1 \tag{7.51}$$

类似地，选择第二大的特征值λ_2及其相应的特征向量$\mathbf{e}_2$，则可得到“第二线性判元”(second linear discriminant)$\mathbf{w}_2'\mathbf{x} = (\boldsymbol{\Sigma}^{-1/2}\mathbf{e}_2)'\mathbf{x}$。

不难证明，第二线性判元的方差也为 1，即$\mathrm{Var}(\mathbf{w}_2'\mathbf{x}) = 1$。

更重要的是，第一线性判元与第二线性判元并不相关：

$$
\begin{aligned}
\mathrm{Cov}(\mathbf{w}_1'\mathbf{x}, \mathbf{w}_2'\mathbf{x}) = \mathbf{w}_1' \mathrm{Var}(\mathbf{x})\mathbf{w}_2 &= (\mathbf{\Sigma}^{-1/2}\mathbf{e}_1)'\mathbf{\Sigma}(\mathbf{\Sigma}^{-1/2}\mathbf{e}_2) \\
&= \mathbf{e}_1'\mathbf{\Sigma}^{-1/2}\mathbf{\Sigma}\mathbf{\Sigma}^{-1/2}\mathbf{e}_2 = \mathbf{e}_1'\mathbf{e}_2 = 0
\end{aligned}
\tag{7.52}
$$

其中，$\mathbf{e}_1$与$\mathbf{e}_2$分别为属于不同特征值的特征向量，故根据线性代数知识，二者正交，即$\mathbf{e}_1'\mathbf{e}_2 = 0$。

类似地，可定义“第三线性判元”(third linear discriminant) $\mathbf{w}_3'\mathbf{x} = (\mathbf{\Sigma}^{-1/2}\mathbf{e}_3)'\mathbf{x}$，以此类推。而且同理可证，所有的线性判元之间均互不相关。

以上结果还可进一步简化。由于λ与$\mathbf{e}$为矩阵$(\boldsymbol{\Sigma}^{-1/2}\mathbf{B}\boldsymbol{\Sigma}^{-1/2})$的特征值与特征向量，故

$$(\boldsymbol{\Sigma}^{-1/2}\mathbf{B}\boldsymbol{\Sigma}^{-1/2})\mathbf{e} = \lambda\mathbf{e} \qquad (7.53)$$

在上式两边同时左乘$\boldsymbol{\Sigma}^{-1/2}$可得：

$$\boldsymbol{\Sigma}^{-1}\mathbf{B}(\boldsymbol{\Sigma}^{-1/2}\mathbf{e}) = \lambda(\boldsymbol{\Sigma}^{-1/2}\mathbf{e}) \qquad (7.54)$$

矩阵$\boldsymbol{\Sigma}^{-1}\mathbf{B}$拥有与$(\boldsymbol{\Sigma}^{-1/2}\mathbf{B}\boldsymbol{\Sigma}^{-1/2})$相同的特征值$\lambda$，而相应的特征向量为$\mathbf{w} = \boldsymbol{\Sigma}^{-1/2}\mathbf{e}$。故只需求解$\boldsymbol{\Sigma}^{-1}\mathbf{B}$的特征值与相应的特征向量，即可得到$(\lambda, \mathbf{w})$。

A 7.2 样本中的多分类费雪判别分析

在实践中我们并不知道$\boldsymbol{\Sigma}$与$\mathbf{B}$，但只要使用其样本估计值$\hat{\boldsymbol{\Sigma}}$与$\mathbf{S}_B$即可。

考虑训练样本$\left\{\mathbf{x}_i, y_i\right\}_{i=1}^n$，其中$\mathbf{x}_i = (x_{i1} \cdots x_{ip})'$为$p$维特征向量，而响应变量$y_i \in \left\{1, \cdots, K\right\}$分为$K$类。

希望求得系列最佳投影方向$\left\{\hat{\mathbf{w}}_1, \cdots, \hat{\mathbf{w}}_s\right\}$，以及相应的“线性判元”(linear discriminants)$\left\{\hat{\mathbf{w}}_1'\mathbf{x}, \cdots, \hat{\mathbf{w}}_s'\mathbf{x}\right\}$，其中$s \le \min\left\{K-1, p\right\}$。

记整个样本数据的均值为

$$\hat{\boldsymbol{\mu}} \equiv \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i \tag{7.55}$$

记第k类样本的均值为

$$\hat{\boldsymbol{\mu}}_k \equiv \frac{1}{n_k}\sum_{y_i=k}\mathbf{x}_i \quad (k=1,\cdots,K) \tag{7.56}$$

定义样本的组间散度矩阵为

$$\mathbf{S}_B \equiv \sum_{k=1}^{K} n_k(\hat{\boldsymbol{\mu}}_k-\hat{\boldsymbol{\mu}})(\hat{\boldsymbol{\mu}}_k-\hat{\boldsymbol{\mu}})' \tag{7.57}$$

定义第k类样本的组内散度矩阵为

$$\mathbf{S}_k \equiv \sum_{y_i=k} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)' \quad (k = 1, \cdots, K) \quad (7.58)$$

整个样本的组内散度矩阵为

$$\mathbf{S}_W = \sum_{k=1}^{K} \mathbf{S}_k \equiv \sum_{k=1}^{K} \sum_{y_i=k} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)' \quad (7.59)$$

在每类数据的协方差矩阵均相等的假设下，整个样本的协方差矩阵为 $\hat{\mathbf{\Sigma}} = \dfrac{1}{n-K}\mathbf{S}_W$。

将$(\hat{\mathbf{\Sigma}}, \mathbf{S}_B)$作为$(\mathbf{\Sigma}, \mathbf{B})$的样本估计量，代入附录 A7.1 在总体中的多分类费雪判别分析，即可得相应的最优解$\left\{\hat{\mathbf{w}}_1, \cdots, \hat{\mathbf{w}}_s\right\}$与线性判元$\left\{\hat{\mathbf{w}}_1'\mathbf{x}, \cdots, \hat{\mathbf{w}}_s'\mathbf{x}\right\}$。

A 7.3 线性判元对于组间方差的贡献率

考虑不同的线性判元对于投影前数据的组间方差之贡献率。

在总体中，将投影前数据的组间方差记为

$$\Delta^2 \equiv \sum_{k=1}^{K} (\boldsymbol{\mu}_k - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_k - \boldsymbol{\mu}) \qquad (7.60)$$

其中，$(\boldsymbol{\mu}_k - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_k - \boldsymbol{\mu})$为第$k$类数据的中心($\boldsymbol{\mu}_k$)到所有数据的中心($\boldsymbol{\mu}$)之统计距离的平方(squared statistical distance)。

记矩阵$\boldsymbol{\Sigma}^{-1}\mathbf{B}$的所有特征值为$\lambda_1,\cdots,\lambda_p$，其中$\lambda_1 \geq \cdots \geq \lambda_s > 0$为非零特征值，而$\lambda_{s+1} = \cdots = \lambda_p = 0$为零特征值。

记$\mathbf{P} \equiv (\mathbf{e}_1 \cdots \mathbf{e}_p)$为相应的特征向量所构成的正交矩阵。考虑以下$p$维随机向量：

$$\mathbf{z} \equiv \begin{pmatrix} z_1 \\ \vdots \\ z_s \\ \vdots \\ z_p \end{pmatrix} = \begin{pmatrix} \mathbf{e}_1'\boldsymbol{\Sigma}^{-1/2}\mathbf{x} \\ \vdots \\ \mathbf{e}_s'\boldsymbol{\Sigma}^{-1/2}\mathbf{x} \\ \vdots \\ \mathbf{e}_p'\boldsymbol{\Sigma}^{-1/2}\mathbf{x} \end{pmatrix} \equiv \mathbf{P}'\boldsymbol{\Sigma}^{-1/2}\mathbf{x} \qquad (7.61)$$

其中，$\mathbf{z}$的前s维分量为相应的线性判元。对于第k类数据，$\mathbf{z}$的均值为

$$\begin{aligned}\boldsymbol{\mu}_{\mathbf{z}}^{(k)} &\equiv (\mu_{z_1}^{(k)} \cdots \mu_{z_p}^{(k)})' \equiv \mathrm{E}(\mathbf{z} \mid y = k) \\ &= \mathrm{E}(\mathbf{P}'\boldsymbol{\Sigma}^{-1/2}\mathbf{x} \mid y = k) = \mathbf{P}'\boldsymbol{\Sigma}^{-1/2}\,\mathrm{E}(\mathbf{x} \mid y = k) = \mathbf{P}'\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}_k\end{aligned} \tag{7.62}$$

而对于所有数据，$\mathbf{z}$的均值为

$$\boldsymbol{\mu}_{\mathbf{z}} \equiv (\mu_{z_1} \cdots \mu_{z_p})' \equiv \mathrm{E}(\mathbf{z}) = \mathrm{E}(\mathbf{P}'\boldsymbol{\Sigma}^{-1/2}\mathbf{x}) = \mathbf{P}'\boldsymbol{\Sigma}^{-1/2}\,\mathrm{E}(\mathbf{x}) = \mathbf{P}'\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu} \tag{7.63}$$

故投影后每类数据的中心与所有数据的中心之离差可写为

$$\underbrace{\left(\boldsymbol{\mu}_{\mathbf{z}}^{(k)}-\boldsymbol{\mu}_{\mathbf{z}}\right)}_{\text{投影后}}=\mathbf{P}'\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}_k-\mathbf{P}'\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}=\mathbf{P}'\boldsymbol{\Sigma}^{-1/2}\underbrace{\left(\boldsymbol{\mu}_k-\boldsymbol{\mu}\right)}_{\text{投影前}} \quad (7.64)$$

在上式两边同时左乘$\boldsymbol{\Sigma}^{1/2}\mathbf{P}$，可将投影前的离差表达为投影后的离差之函数：

$$\underbrace{\left(\boldsymbol{\mu}_k-\boldsymbol{\mu}\right)}_{\text{投影前}}=\boldsymbol{\Sigma}^{1/2}\mathbf{P}\underbrace{\left(\boldsymbol{\mu}_{\mathbf{z}}^{(k)}-\boldsymbol{\mu}_{\mathbf{z}}\right)}_{\text{投影后}} \quad (7.65)$$

将上式代入投影前数据的组间方差表达式(7.60)可得：

$$
\begin{aligned}
\Delta^2 &= \sum_{k=1}^{K} (\boldsymbol{\mu}_k - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_k - \boldsymbol{\mu}) \\
&= \sum_{k=1}^{K} (\boldsymbol{\mu}_{\mathbf{z}}^{(k)} - \boldsymbol{\mu}_{\mathbf{z}})' \mathbf{P}' \boldsymbol{\Sigma}^{1/2} \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}^{1/2} \mathbf{P} (\boldsymbol{\mu}_{\mathbf{z}}^{(k)} - \boldsymbol{\mu}_{\mathbf{z}}) \qquad (7.66) \\
&= \sum_{k=1}^{K} (\boldsymbol{\mu}_{\mathbf{z}}^{(k)} - \boldsymbol{\mu}_{\mathbf{z}})' (\boldsymbol{\mu}_{\mathbf{z}}^{(k)} - \boldsymbol{\mu}_{\mathbf{z}})
\end{aligned}
$$

其中，由于$\mathbf{P}$为正交矩阵，故$\mathbf{P}'\mathbf{P} = \mathbf{I}$(单位矩阵)。

这表明，投影前数据的组间方差等于投影后数据的组间方差。

下面，考虑第一线性判元，即$\mathbf{z}_1 = \mathbf{e}_1'\boldsymbol{\Sigma}^{-1/2}\mathbf{x}$，对于组间方差$\Delta^2$的贡献。

由于$\boldsymbol{\mu}_{\mathbf{z}}^{(k)} = (\mu_{z_1}^{(k)} \cdots \mu_{z_p}^{(k)})'$，而$\boldsymbol{\mu}_{\mathbf{z}} = (\mu_{z_1} \cdots \mu_{z_p})'$，故

$$
\begin{aligned}
\Delta^2 &= \sum_{k=1}^{K} (\boldsymbol{\mu}_{\mathbf{z}}^{(k)} - \boldsymbol{\mu}_{\mathbf{z}})'(\boldsymbol{\mu}_{\mathbf{z}}^{(k)} - \boldsymbol{\mu}_{\mathbf{z}}) \\
&= \sum_{k=1}^{K} (\mu_{z_1}^{(k)} - \mu_{z_1}, \cdots, \mu_{z_p}^{(k)} - \mu_{z_p}) \begin{pmatrix} \mu_{z_1}^{(k)} - \mu_{z_1} \\ \vdots \\ \mu_{z_p}^{(k)} - \mu_{z_p} \end{pmatrix} \\
&= \sum_{k=1}^{K} \left[(\mu_{z_1}^{(k)} - \mu_{z_1})^2 + \cdots + (\mu_{z_p}^{(k)} - \mu_{z_p})^2 \right] \\
&= \sum_{k=1}^{K} (\mu_{z_1}^{(k)} - \mu_{z_1})^2 + \cdots + \sum_{k=1}^{K} (\mu_{z_p}^{(k)} - \mu_{z_p})^2
\end{aligned}
\tag{7.67}
$$

其中，$\sum_{k=1}^{K}(\mu_{z_1}^{(k)}-\mu_{z_1})^2$即为第一线性判元$z_1=\mathbf{e}_1'\mathbf{\Sigma}^{-1/2}\mathbf{x}$对于组间离差$\Delta^2$的贡献，以此类推。

进一步，对于第k类数据，第一线性判元z_1的均值为

$$\begin{aligned}\mu_{z_1}^{(k)} &= \mathrm{E}(z_1 \mid y=k) = \mathrm{E}(\mathbf{e}_1'\mathbf{\Sigma}^{-1/2}\mathbf{x} \mid y=k) \\ &= \mathbf{e}_1'\mathbf{\Sigma}^{-1/2}\,\mathrm{E}(\mathbf{x} \mid y=k) = \mathbf{e}_1'\mathbf{\Sigma}^{-1/2}\boldsymbol{\mu}_k\end{aligned} \tag{7.68}$$

而对于所有数据，第一线性判元z_1的均值为

$$\mu_{z_1} = \mathrm{E}(z_1) = \mathrm{E}(\mathbf{e}_1'\mathbf{\Sigma}^{-1/2}\mathbf{x}) = \mathbf{e}_1'\mathbf{\Sigma}^{-1/2}\,\mathrm{E}(\mathbf{x}) = \mathbf{e}_1'\mathbf{\Sigma}^{-1/2}\boldsymbol{\mu} \tag{7.69}$$

因此，$(\mu_{z_1}^{(k)} - \mu_{z_1}) = \mathbf{e}_1'\mathbf{\Sigma}^{-1/2}\boldsymbol{\mu}_k - \mathbf{e}_1'\mathbf{\Sigma}^{-1/2}\boldsymbol{\mu} = \mathbf{e}_1'\mathbf{\Sigma}^{-1/2}(\boldsymbol{\mu}_k - \boldsymbol{\mu})$，故

$$
\begin{aligned}
\sum_{k=1}^{K}(\mu_{z_1}^{(k)} - \mu_{z_1})^2 &= \sum_{k=1}^{K}\mathbf{e}_1'\mathbf{\Sigma}^{-1/2}(\boldsymbol{\mu}_k - \boldsymbol{\mu})(\boldsymbol{\mu}_k - \boldsymbol{\mu})'\mathbf{\Sigma}^{-1/2}\mathbf{e}_1 \\
&= \mathbf{e}_1'\mathbf{\Sigma}^{-1/2}\left(\sum_{k=1}^{K}(\boldsymbol{\mu}_k - \boldsymbol{\mu})(\boldsymbol{\mu}_k - \boldsymbol{\mu})'\right)\mathbf{\Sigma}^{-1/2}\mathbf{e}_1 \\
&= \mathbf{e}_1'\underbrace{\mathbf{\Sigma}^{-1/2}\mathbf{B}\mathbf{\Sigma}^{-1/2}\mathbf{e}_1}_{=\lambda_1\mathbf{e}_1} \\
&= \lambda_1\mathbf{e}_1'\mathbf{e}_1 = \lambda_1
\end{aligned}
\tag{7.70}
$$

其中，$\mathbf{e}_1$相应的特征值为λ_1，满足$(\mathbf{\Sigma}^{-1/2}\mathbf{B}\mathbf{\Sigma}^{-1/2})\mathbf{e}_1 = \lambda_1\mathbf{e}_1$，且$\mathbf{e}_1'\mathbf{e}_1 = 1$。

类似地，可以证明，$\sum_{k=1}^{K}(\mu_{z_2}^{(k)} - \mu_{z_2})^2 = \lambda_2$，以此类推。因此，组间方差可分解为

$$\begin{aligned}\Delta^2 &= \sum_{k=1}^{K}(\mu_{z_1}^{(k)} - \mu_{z_1})^2 + \cdots + \sum_{k=1}^{K}(\mu_{z_p}^{(k)} - \mu_{z_p})^2 \\ &= \lambda_1 + \cdots + \lambda_s + \underbrace{\lambda_{s+1} + \cdots + \lambda_p}_{=0} = \lambda_1 + \cdots + \lambda_s\end{aligned} \tag{7.71}$$

其中，λ_1为第一线性判元z_1对组间方差的贡献，λ_2为第二线性判元z_2对组间方差的贡献，以此类推。

从贡献率的角度，第一线性判元 z_1 对组间方差的贡献率为 $\hat{\lambda}_1 / (\hat{\lambda}_1 + \cdots + \hat{\lambda}_s)$，而第二线性判元 z_2 对组间方差的贡献率为 $\hat{\lambda}_2 / (\hat{\lambda}_1 + \cdots + \hat{\lambda}_s)$。

从累积贡献率的角度，第一线性判元 z_1 与二线性判元 z_2 对组间方差的累积贡献率为 $(\hat{\lambda}_1 + \hat{\lambda}_2) / (\hat{\lambda}_1 + \cdots + \hat{\lambda}_s)$，以此类推。