Classification Analysis

September 23, 2025

1 Introduction to Classification Analysis

Classification analysis is a major application of multivariate statistical methods. Its central goal is to **assign individuals (observations) to one of several predefined groups (populations)** based on observed measurements of multiple variables.

Objective:Given measurements on several variables (predictors, features), decide from which population a new observation most likely comes.

Examples:

- Medical diagnosis: classify patients into “diseased” or “healthy” groups using biomarkers.
- Marketing: assign customers to “high-value” or “low-value” segments based on purchasing patterns.
- Biology: classify species from morphological measurements.

Classification analysis assumes that:

1. Populations are **well defined** and observations are drawn from them.
2. The **group membership** of training samples is known.
3. The **group membership** of new observations is unknown and must be inferred.

2 Formal Statement of the Classification Problem

Suppose we have:

- g groups (populations).

- Group i has multivariate distribution $N_p(\boldsymbol{\mu}_i, \Sigma_i)$, where:
    - $\boldsymbol{\mu}_i$ is the mean vector.
    - Σ_i is the covariance matrix.
- A new observation $\mathbf{x}_0$ (a $p \times 1$ vector) is to be classified into one of the g groups.

Goal: Construct a classification rule (decision rule) that assigns $\mathbf{x}_0$ to one of the groups.

3 Classification Into Two Groups

When there are two populations, we can use a classification procedure due to Fisher (1936). The principal assumption for Fisher's procedure is that the two populations have the same covariance matrix ($\Sigma_1 = \Sigma_2$). Normality is not required.

We obtain a sample from each of the two populations and compute $\bar{\mathbf{y}}_1, \bar{\mathbf{y}}_2$, and $\mathbf{S}_{pl}$. A simple procedure for classification can be based on the discriminant function,

$$z = \mathbf{a}'\mathbf{y} = (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)\mathbf{S}_{pl}^{-1}\mathbf{y}$$

where $\mathbf{y}$ is the vector of measurements on a new sampling unit that we wish to classify into one of the two groups. For convenience we speak of classifying $\mathbf{y}$ rather than classifying the subject or object associated with $\mathbf{y}$.

To determine whether $\mathbf{y}$ is closer to $\bar{\mathbf{y}}_1$ or $\bar{\mathbf{y}}_2$, we check to see if z in is closer to the transformed mean $\bar{z}_1$ or to $\bar{z}_2$. Denote the two groups by G_1 and G_2. Fisher's (1936) linear classification procedure assigns $\mathbf{y}$ to G_1 if $z = \mathbf{a}'\mathbf{y}$ is closer to $\bar{z}_1$ than to $\bar{z}_2$ and assigns $\mathbf{y}$ to G_2 if $z = \mathbf{a}'\mathbf{y}$ is closer to $\bar{z}_2$.

Fisher's (1936) approach is essentially nonparametric because no distributional assumptions were made. However, if the two populations are normal with equal covariance matrices, then this method is (asymptotically) optimal; that is, the probability of misclassification is minimized.

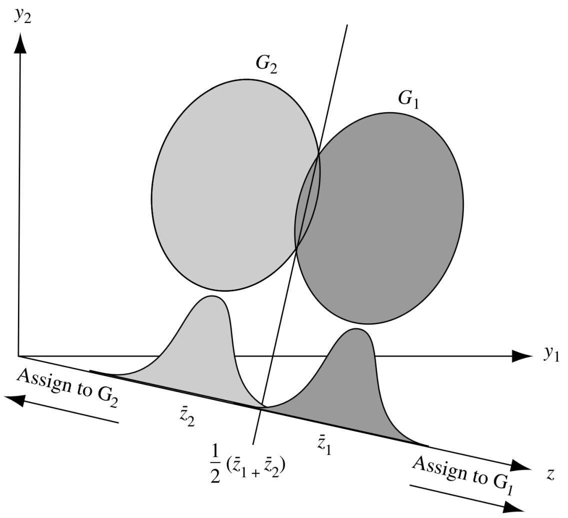


Figure 9.1. Fisher's procedure for classification into two groups.

For the configuration in Figure 9.1, we see that z is closer to $\overline{z}_1$ if

$$z > \tfrac{1}{2}(\overline{z}_1 + \overline{z}_2). \tag{9.2}$$

This is true in general because $\overline{z}_1$ is always greater than $\overline{z}_2$, which can easily be shown as follows:

$$\overline{z}_1 - \overline{z}_2 = \mathbf{a}'(\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2) = (\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2)'\mathbf{S}_{\text{pl}}^{-1}(\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2) > 0, \tag{9.3}$$

because $\mathbf{S}_{\text{pl}}^{-1}$ is positive definite. Thus $\overline{z}_1 > \overline{z}_2$. [If $\mathbf{a}$ were of the form $\mathbf{a}' = (\overline{\mathbf{y}}_2 - \overline{\mathbf{y}}_1)'\mathbf{S}_{\text{pl}}^{-1}$, then $\overline{z}_2 - \overline{z}_1$ would be positive.] Since $\frac{1}{2}(\overline{z}_1 + \overline{z}_2)$ is the midpoint, $z > \frac{1}{2}(\overline{z}_1 + \overline{z}_2)$ implies that z is closer to $\overline{z}_1$. By (9.3) the distance from $\overline{z}_1$ to $\overline{z}_2$ is the same as that from $\overline{\mathbf{y}}_1$ to $\overline{\mathbf{y}}_2$.

To express the classification rule in terms of $\mathbf{y}$, we first write $\frac{1}{2}(\overline{z}_1 + \overline{z}_2)$ in the form

$$\tfrac{1}{2}(\overline{z}_1 + \overline{z}_2) = \tfrac{1}{2}(\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2)'\mathbf{S}_{\text{pl}}^{-1}(\overline{\mathbf{y}}_1 + \overline{\mathbf{y}}_2). \tag{9.4}$$

Then the classification rule becomes: Assign $\mathbf{y}$ to G_1 if

$$\mathbf{a}'\mathbf{y} = (\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2)'\mathbf{S}_{\text{pl}}^{-1}\mathbf{y} > \tfrac{1}{2}(\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2)'\mathbf{S}_{\text{pl}}^{-1}(\overline{\mathbf{y}}_1 + \overline{\mathbf{y}}_2) \tag{9.5}$$

and assign $\mathbf{y}$ to G_2 if

$$\mathbf{a}'\mathbf{y} = (\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2)'\mathbf{S}_{\text{pl}}^{-1}\mathbf{y} < \tfrac{1}{2}(\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2)'\mathbf{S}_{\text{pl}}^{-1}(\overline{\mathbf{y}}_1 + \overline{\mathbf{y}}_2). \tag{9.6}$$

3.1 Problem setup, priors, and cost

Let there be two populations (groups) $G = 1, 2$. An observation $X \in \mathbb{R}^p$ has class-conditional densities

$$X \mid G = j \sim f_j(\mathbf{x}), \qquad j = 1, 2,$$

and prior probabilities

$$\pi_j = P(G = j), \qquad \pi_1 + \pi_2 = 1.$$

Define misclassification costs $C(i \mid j)$: cost of assigning to class i when true class is j. The expected cost of deciding class i given $\mathbf{x}$ is

$$R(i \mid \mathbf{x}) = \sum_{j=1}^{2} C(i \mid j)\, P(G = j \mid \mathbf{x}).$$

The Bayes rule (minimize expected cost) assigns $\mathbf{x}$ to class 1 iff

$$R(1 \mid \mathbf{x}) < R(2 \mid \mathbf{x}),$$

which after algebra is equivalent to comparing the likelihood ratio to a threshold. For equal costs $C(i \mid j) = 1$ when $i \neq j$ (and zero for correct classification), the Bayes rule reduces to assigning to the class with larger posterior probability:

$$\text{assign to 1} \quad \Longleftrightarrow \quad P(G = 1 \mid \mathbf{x}) > P(G = 2 \mid \mathbf{x}).$$

Using Bayes theorem,

$$P(G = j \mid \mathbf{x}) = \frac{f_j(\mathbf{x})\pi_j}{f_1(\mathbf{x})\pi_1 + f_2(\mathbf{x})\pi_2},$$

so the decision (for equal costs) is equivalent to

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > \frac{\pi_2}{\pi_1} \quad \text{(or equivalently compare } \ln(f_1\pi_1) - \ln(f_2\pi_2) > 0\text{)}.$$

If costs are unequal, we replace the RHS by $\dfrac{C(2 \mid 2) - C(1 \mid 2)}{C(1 \mid 1) - C(2 \mid 1)}$ (standard cost–ratio form). For simplicity most derivations below assume equal costs; priors π_j remain.

3.2 General multivariate normal case — likelihood ratio and log discriminant

Assume class-conditional densities are multivariate normal:

$$f_j(\mathbf{x}) = (2\pi)^{-p/2}|\Sigma_j|^{-1/2} \exp\Big(-\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_j)^\top \Sigma_j^{-1}(\mathbf{x} - \boldsymbol{\mu}_j)\Big).$$

Take the log of posterior numerator for class j: $\ell_j(\mathbf{x}) = \ln f_j(\mathbf{x}) + \ln \pi_j$. Then the log-likelihood ratio is

$$\Lambda(\mathbf{x}) \; = \; \ell_1(\mathbf{x}) - \ell_2(\mathbf{x})$$

$$= -\tfrac{1}{2}\ln|\Sigma_1| + \tfrac{1}{2}\ln|\Sigma_2| - \tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^\top \Sigma_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) + \tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^\top \Sigma_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) + \ln\frac{\pi_1}{\pi_2}.$$

Bayes decision: assign to class 1 iff $\Lambda(\mathbf{x}) > 0$ (or $>$ ln cost ratio if unequal costs).

This is the full, exact discriminant for the Gaussian model. It is quadratic in $\mathbf{x}$ in the general unequal covariance case (because of the $\mathbf{x}^\top \Sigma_j^{-1} \mathbf{x}$ terms).

3.3 Special case: equal covariance matrices

$\Sigma_1 = \Sigma_2 = \Sigma \rightarrow$ **Linear Discriminant Analysis (LDA)**

When $\Sigma_1 = \Sigma_2 = \Sigma$, the $\mathbf{x}^\top \Sigma^{-1} \mathbf{x}$ terms cancel. Simplify $\Lambda(\mathbf{x})$:

Starting from the quadratic terms,

$$\begin{aligned}
&-\tfrac{1}{2}(\mathbf{x}-\mu_1)^\top \Sigma^{-1}(\mathbf{x}-\mu_1) + \tfrac{1}{2}(\mathbf{x}-\mu_2)^\top \Sigma^{-1}(\mathbf{x}-\mu_2) \\
&\quad = -\tfrac{1}{2}\big(\mathbf{x}^\top \Sigma^{-1}\mathbf{x} - 2\mu_1^\top \Sigma^{-1}\mathbf{x} + \mu_1^\top \Sigma^{-1}\mu_1\big) + \tfrac{1}{2}\big(\mathbf{x}^\top \Sigma^{-1}\mathbf{x} - 2\mu_2^\top \Sigma^{-1}\mathbf{x} + \mu_2^\top \Sigma^{-1}\mu_2\big) \\
&\quad = (\mu_1-\mu_2)^\top \Sigma^{-1}\mathbf{x} - \tfrac{1}{2}\big(\mu_1^\top \Sigma^{-1}\mu_1 - \mu_2^\top \Sigma^{-1}\mu_2\big).
\end{aligned}$$

So the log discriminant reduces to a linear function in $\mathbf{x}$:

$$\Lambda(\mathbf{x}) = (\mu_1-\mu_2)^\top \Sigma^{-1}\mathbf{x} - \tfrac{1}{2}\big(\mu_1^\top \Sigma^{-1}\mu_1 - \mu_2^\top \Sigma^{-1}\mu_2\big) + \ln\frac{\pi_1}{\pi_2}.$$

Define the weight vector

$$\mathbf{w} = \Sigma^{-1}(\mu_1 - \mu_2).$$

Then the decision rule is a linear threshold test:

$$\mathbf{w}^\top \mathbf{x} + w_0 \gtrless 0,$$

with intercept

$$w_0 = -\tfrac{1}{2}(\mu_1^\top \Sigma^{-1}\mu_1 - \mu_2^\top \Sigma^{-1}\mu_2) + \ln\frac{\pi_1}{\pi_2}.$$

Geometry. The decision boundary $\{\mathbf{x} : \Lambda(\mathbf{x}) = 0\}$ is a hyperplane perpendicular to $\mathbf{w}$. If priors equal ($\pi_1 = \pi_2$), the intercept simplifies and the boundary is the hyperplane where projected distances to the class means are equal in the Mahalanobis sense.

3.4 Unequal covariance case → Quadratic Discriminant Analysis (QDA)

If $\Sigma_1 \neq \Sigma_2$ we cannot cancel quadratic terms. From the general $\Lambda(\mathbf{x})$ above:

$$\Lambda(\mathbf{x}) = -\tfrac{1}{2}\ln|\Sigma_1| + \tfrac{1}{2}\ln|\Sigma_2| - \tfrac{1}{2}(\mathbf{x}-\mu_1)^\top\Sigma_1^{-1}(\mathbf{x}-\mu_1) + \tfrac{1}{2}(\mathbf{x}-\mu_2)^\top\Sigma_2^{-1}(\mathbf{x}-\mu_2) \\ + \ln\frac{\pi_1}{\pi_2}.$$

Group the terms to expose the quadratic form in $\mathbf{x}$. Define matrices and vectors:

$$\begin{aligned} Q &= \tfrac{1}{2}(\Sigma_2^{-1} - \Sigma_1^{-1}), \\ \mathbf{q} &= \Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2, \\ c &= -\tfrac{1}{2}(\mu_1^\top\Sigma_1^{-1}\mu_1 - \mu_2^\top\Sigma_2^{-1}\mu_2) - \tfrac{1}{2}\ln\frac{|\Sigma_1|}{|\Sigma_2|} + \ln\frac{\pi_1}{\pi_2}. \end{aligned}$$

Then $\Lambda(\mathbf{x}) = \mathbf{x}^\top Q\mathbf{x} + \mathbf{q}^\top\mathbf{x} + c$. So the decision boundary $\Lambda(\mathbf{x}) = 0$ is in general a **quadratic surface** (ellipsoid, paraboloid, hyperboloid depending on Q). This is the QDA decision surface.

Tradeoff: QDA is more flexible (can capture different shapes) but requires more parameters (each Σ_j estimated) and thus more data; LDA is more stable when sample sizes are small.

3.5 Estimation from data (two groups)

In practice μ_j, Σ_j unknown → estimate from training data:

* sample means: $\hat{\mu}_j = \frac{1}{n_j}\sum_{i:\,G_i=j} x_i$. * sample covariances: $S_j = \frac{1}{n_j-1}\sum_{i:\,G_i=j}(x_i - \hat{\mu}_j)(x_i - \hat{\mu}_j)^\top$.

For LDA (equal covariance assumption), use pooled covariance

$$S_p = \frac{(n_1-1)S_1 + (n_2-1)S_2}{n_1+n_2-2}.$$

Plug $\hat{\mu}_j$ and S_p into formulas for $\mathbf{w}$ and w_0. Then classify new $\mathbf{x}$ by sign of $\hat{\mathbf{w}}^\top\mathbf{x} + \hat{w}_0$.

For QDA, plug $\hat{\mu}_j$ and S_j into $\Lambda(\mathbf{x})$. Regularization (shrinkage, ridge) may be used if covariance estimates are ill-conditioned.

3.6 Practical 6-step recipe for two-group classification (LDA preferred if Σ plausibly equal)

1. Compute sample means $\hat{\mu}_1, \hat{\mu}_2$ and sample covariances S_1, S_2.

2. Test or assess equality of covariances (Box's M, graphical checks). If equal, compute pooled S_p.

3. Compute $\hat{\mathbf{w}} = S_p^{-1}(\hat{\mu}_1 - \hat{\mu}_2)$.
4. Compute intercept $\hat{w}_0 = -\frac{1}{2}(\hat{\mu}_1^\top S_p^{-1}\hat{\mu}_1 - \hat{\mu}_2^\top S_p^{-1}\hat{\mu}_2) + \ln(\hat{\pi}_1/\hat{\pi}_2)$, where $\hat{\pi}_j = n_j/n$.
5. Classify new $\mathbf{x}$ by sign of $\hat{\mathbf{w}}^\top\mathbf{x} + \hat{w}_0$.
6. Evaluate via confusion matrix, cross-validation / leave-one-out, and compute estimated error.

4 Classification into several groups

Now extend to $g \geq 2$ groups. Many formulas generalize directly.

4.1 Bayes rule for multiple groups

Given class densities $f_j(\mathbf{x})$ and priors π_j, Bayes assigns $\mathbf{x}$ to the class with maximum posterior probability:

$$\text{assign to } j^* = \arg\max_{j=1,\dots,g} P(G = j \mid \mathbf{x}) = \arg\max_j f_j(\mathbf{x})\pi_j.$$

Equivalently choose j that maximizes the log-score $\ell_j(\mathbf{x}) = \ln f_j(\mathbf{x}) + \ln \pi_j$.

4.2 Multivariate normal model, common covariance → Linear Discriminant Analysis (g-class LDA)

Assume

$$X \mid G = j \sim N_p(\mu_j, \Sigma), \qquad j = 1, \dots, g,$$

with common Σ. Then

$$\ell_j(\mathbf{x}) = -\tfrac{1}{2}(\mathbf{x} - \mu_j)^\top\Sigma^{-1}(\mathbf{x} - \mu_j) - \tfrac{p}{2}\ln(2\pi) - \tfrac{1}{2}\ln|\Sigma| + \ln \pi_j.$$

Dropping the constant terms that are the same for all j, define the linear discriminant functions

$$\boxed{\delta_j(\mathbf{x}) \;=\; \mathbf{x}^\top\Sigma^{-1}\mu_j \;-\; \tfrac{1}{2}\mu_j^\top\Sigma^{-1}\mu_j \;+\; \ln \pi_j}$$

and assign $\mathbf{x}$ to $\arg\max_j \delta_j(\mathbf{x})$.

9.3.1 Equal Population Covariance Matrices: Linear Classification Functions

In this section we assume $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \cdots = \boldsymbol{\Sigma}_k$. We can estimate the common population covariance matrix by a pooled sample covariance matrix

$$\mathbf{S}_{\text{pl}} = \frac{1}{N-k}\sum_{i=1}^{k}(n_i - 1)\mathbf{S}_i = \frac{\mathbf{E}}{N-k},$$

where n_i and $\mathbf{S}_i$ are the sample size and covariance matrix of the ith group, $\mathbf{E}$ is the error matrix from one-way MANOVA, and $N = \sum_i n_i$. We compare $\mathbf{y}$ to each $\overline{\mathbf{y}}_i$, $i = 1, 2, \ldots, k$, by the distance function

$$D_i^2(\mathbf{y}) = (\mathbf{y} - \overline{\mathbf{y}}_i)'\mathbf{S}_{\text{pl}}^{-1}(\mathbf{y} - \overline{\mathbf{y}}_i) \tag{9.9}$$

and assign $\mathbf{y}$ to the group for which $D_i^2(\mathbf{y})$ is smallest.

We can obtain a linear classification rule by expanding (9.9):

$$\begin{aligned} D_i^2(\mathbf{y}) &= \mathbf{y}'\mathbf{S}_{\text{pl}}^{-1}\mathbf{y} - \mathbf{y}'\mathbf{S}_{\text{pl}}^{-1}\overline{\mathbf{y}}_i - \overline{\mathbf{y}}_i'\mathbf{S}_{\text{pl}}^{-1}\mathbf{y} + \overline{\mathbf{y}}_i'\mathbf{S}_{\text{pl}}^{-1}\overline{\mathbf{y}}_i \\ &= \mathbf{y}'\mathbf{S}_{\text{pl}}^{-1}\mathbf{y} - 2\overline{\mathbf{y}}_i'\mathbf{S}_{\text{pl}}^{-1}\mathbf{y} + \overline{\mathbf{y}}_i'\mathbf{S}_{\text{pl}}^{-1}\overline{\mathbf{y}}_i. \end{aligned}$$

The term $\mathbf{y}'\mathbf{S}_{\text{pl}}^{-1}\mathbf{y}$ on the right can be neglected since it is not a function of i and, consequently, does not change from group to group. The second term is a linear function of $\mathbf{y}$, and the third does not involve $\mathbf{y}$. We thus delete $\mathbf{y}'\mathbf{S}_{\text{pl}}^{-1}\mathbf{y}$ and obtain a *linear classification function*, which we denote by $L_i(\mathbf{y})$. If we multiply by $-\frac{1}{2}$ to agree with the rule based on the normal distribution and prior probabilities given in (9.12), our linear classification rule becomes: Assign $\mathbf{y}$ to the group for which

$$L_i(\mathbf{y}) = \overline{\mathbf{y}}_i'\mathbf{S}_{\text{pl}}^{-1}\mathbf{y} - \tfrac{1}{2}\overline{\mathbf{y}}_i'\mathbf{S}_{\text{pl}}^{-1}\overline{\mathbf{y}}_i, \qquad i = 1, 2, \ldots, k \tag{9.10}$$

is a *maximum* (we reversed the sign when multiplying by $-\frac{1}{2}$). To highlight the linearity of (9.10) as a function of $\mathbf{y}$, we can express it as

$$L_i(\mathbf{y}) = \mathbf{c}_i'\mathbf{y} + c_{i0} = c_{i1}y_1 + c_{i2}y_2 + \cdots + c_{ip}y_p + c_{i0},$$

where $\mathbf{c}_i' = \overline{\mathbf{y}}_i'\mathbf{S}_{\text{pl}}^{-1}$ and $c_{i0} = -\frac{1}{2}\overline{\mathbf{y}}_i'\mathbf{S}_{\text{pl}}^{-1}\overline{\mathbf{y}}_i$. To assign $\mathbf{y}$ to a group using this procedure, we calculate $\mathbf{c}_i$ and c_{i0} for each of the k groups, evaluate $L_i(\mathbf{y})$, $i = 1, 2, \ldots, k$, and allocate $\mathbf{y}$ to the group for which $L_i(\mathbf{y})$ is largest. This will be the same group for which $D_i^2(\mathbf{y})$ in (9.9) is smallest, that is, the group whose mean vector $\overline{\mathbf{y}}_i$ is closest to $\mathbf{y}$.

For the case of several groups, the optimal rule in (9.7) extends to:

$$\text{Assign } \mathbf{y} \text{ to the group for which } p_i f(\mathbf{y}|G_i) \text{ is maximum.} \tag{9.11}$$

With this rule, the probability of misclassification is minimized. If we assume normality with equal covariance matrices and with prior probabilities of group membership, $p_1, p_2, \ldots, p_k$, then $f(\mathbf{y}|G_i) = N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, and the rule in (9.11) becomes (with estimates in place of parameters): Calculate

$$L_i'(\mathbf{y}) = \ln p_i + \overline{\mathbf{y}}_i'\mathbf{S}_{\text{pl}}^{-1}\mathbf{y} - \tfrac{1}{2}\overline{\mathbf{y}}_i'\mathbf{S}_{\text{pl}}^{-1}\overline{\mathbf{y}}_i, \qquad i = 1, 2, \ldots, k \tag{9.12}$$

and assign $\mathbf{y}$ to the group with maximum value of $L_i'(\mathbf{y})$. Note that if $p_1 = p_2 = \cdots = p_k$, then (9.12), which optimizes the classification rate for the normal distribution, reduces to (9.10), which was based on the heuristic approach of minimizing the distance of $\mathbf{y}$ to $\overline{\mathbf{y}}_i$.

Example 9.3.1. For the football data of Table 8.3, the mean vectors for the three groups are as follows:

$$\begin{aligned}
\overline{\mathbf{y}}_1' &= (15.2, 58.9, 20.1, 13.1, 14.7, 12.3),\\
\overline{\mathbf{y}}_2' &= (15.4, 57.4, 19.8, 10.1, 13.5, 11.9),\\
\overline{\mathbf{y}}_3' &= (15.6, 57.8, 19.8, 10.9, 13.7, 11.8).
\end{aligned}$$

Using these values of $\overline{\mathbf{y}}_i$ and the pooled covariance matrix $\mathbf{S}_{\text{pl}}$, given in Example 8.5, the linear classification functions (9.10) become

$$\begin{aligned}
L_1(\mathbf{y}) &= 7.6y_1 + 13.3y_2 + 4.2y_3 - 1.2y_4 + 14.6y_5 + 8.2y_6 - 641.1,\\
L_2(\mathbf{y}) &= 10.2y_1 + 13.3y_2 + 4.2y_3 - 3.4y_4 + 13.2y_5 + 6.1y_6 - 608.0,\\
L_3(\mathbf{y}) &= 10.9y_1 + 13.3y_2 + 4.1y_3 - 2.7y_4 + 13.1y_5 + 5.2y_6 - 614.6.
\end{aligned}$$

We note that y_2 and y_3 have essentially the same coefficients in all three functions and hence do not contribute to classification of $\mathbf{y}$. These same two variables were eliminated in the stepwise discriminant analysis in Example 8.9.

We illustrate the use of these linear functions for the first and third observations in group 1. For the first observation, $\mathbf{y}_{11}$, we obtain

$$\begin{aligned} L_1(\mathbf{y}_{11}) &= 7.6(13.5) + 13.3(57.2) + 4.2(19.5) - 1.2(12.5) + 14.6(14.0) \\ &\quad + 8.2(11.0) - 641.1 = 582.124, \\ L_2(\mathbf{y}_{11}) &= 10.2(13.5) + 13.3(57.2) + 4.2(19.5) - 3.4(12.5) + 13.2(14.0) \\ &\quad + 6.1(11.0) - 608.0 = 578.099, \\ L_3(\mathbf{y}_{11}) &= 10.9(13.5) + 13.3(57.2) + 4.1(19.5) - 2.7(12.5) + 13.1(14.0) \\ &\quad + 5.2(11.0) - 614.6 = 578.760. \end{aligned}$$

We classify $\mathbf{y}_{11}$ into group 1 since $L_1(\mathbf{y}_{11}) = 582.1$ exceeds $L_2(\mathbf{y}_{11})$ and $L_3(\mathbf{y}_{11})$.

For the third observation in group 1, $\mathbf{y}_{13}$, we obtain

$$L_1(\mathbf{y}_{13}) = 567.054, \qquad L_2(\mathbf{y}_{13}) = 570.290, \qquad L_3(\mathbf{y}_{13}) = 569.137.$$

This observation is misclassified into group 2 since $L_2(\mathbf{y}_{13}) = 570.290$ exceeds $L_1(\mathbf{y}_{13})$ and $L_3(\mathbf{y}_{13})$. □

9.3.2 Unequal Population Covariance Matrices: Quadratic Classification Functions

The linear classification functions in Section 9.3.1 are based on the assumption $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \cdots = \boldsymbol{\Sigma}_k$. The resulting classification rules are sensitive to heterogeneity of covariance matrices. Observations tend to be classified too frequently into groups whose covariance matrices have larger variances on the diagonal. Thus the population covariance matrices should not be assumed to be equal if there is reason to suspect otherwise.

If $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \cdots = \boldsymbol{\Sigma}_k$ does not hold, the classification rules can easily be altered to preserve optimality of classification rates. In place of (9.9), we can use

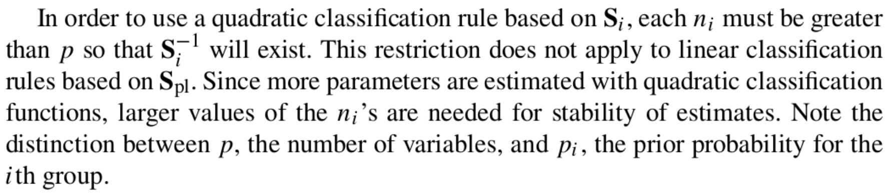
$$D_i^2(\mathbf{y}) = (\mathbf{y} - \overline{\mathbf{y}}_i)'\mathbf{S}_i^{-1}(\mathbf{y} - \overline{\mathbf{y}}_i), \qquad i = 1, 2, \ldots, k, \tag{9.13}$$

where $\mathbf{S}_i$ is the sample covariance matrix for the ith group. As before, we would assign $\mathbf{y}$ to the group for which $D_i^2(\mathbf{y})$ is smallest. With $\mathbf{S}_i$ in place of $\mathbf{S}_{\text{pl}}$, (9.13) cannot be reduced to a linear function of $\mathbf{y}$ as in (9.10) but remains a quadratic function. Hence rules based on $\mathbf{S}_i$ are called *quadratic classification rules*.

If we assume normality with unequal covariance matrices and with prior probabilities $p_1, p_2, \ldots, p_k$, then $f(\mathbf{y}|G_i) = N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, and the optimal rule in (9.11) based on $p_i f(\mathbf{y}|G_i)$ becomes: Assign $\mathbf{y}$ to the group for which

$$Q_i(\mathbf{y}) = \ln p_i - \tfrac{1}{2}\ln|\mathbf{S}_i| - \tfrac{1}{2}(\mathbf{y} - \overline{\mathbf{y}}_i)'\mathbf{S}_i^{-1}(\mathbf{y} - \overline{\mathbf{y}}_i) \tag{9.14}$$

is maximum. If $p_1 = p_2 = \cdots = p_k$ or if the p_i's are unknown, the term $\ln p_i$ is deleted.



In order to use a quadratic classification rule based on $\mathbf{S}_i$, each n_i must be greater than p so that $\mathbf{S}_i^{-1}$ will exist. This restriction does not apply to linear classification rules based on $\mathbf{S}_{\text{pl}}$. Since more parameters are estimated with quadratic classification functions, larger values of the n_i's are needed for stability of estimates. Note the distinction between p, the number of variables, and p_i, the prior probability for the ith group.

5 Bayes' Rule for Classification

The foundation of classification analysis is **Bayes' theorem**:

$$P(j|\mathbf{x}_0) = \frac{f_j(\mathbf{x}_0)\pi_j}{\sum_{k=1}^{g} f_k(\mathbf{x}_0)\pi_k}$$

where:
* $f_j(\mathbf{x}_0)$ = probability density function of group j. * π_j = prior probability of group j.

Bayes' Classification Rule: Assign $\mathbf{x}_0$ to the group i with the largest posterior probability $P(i|\mathbf{x}_0)$.

When misclassification costs are equal, this rule minimizes the overall error rate

6 Discriminant Functions

To implement Bayes' rule, we define **discriminant functions** $\delta_i(\mathbf{x})$.

Assign $\mathbf{x}_0$ to group i if:

$$\delta_i(\mathbf{x}_0) > \delta_j(\mathbf{x}_0), \quad \forall j \neq i$$

* The form of $\delta_i(\mathbf{x})$ depends on assumptions about Σ_i. * These lead to two common rules: Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA).

6.1 Linear Discriminant Analysis (LDA)

Assumption: All groups have the same covariance matrix,

$$\Sigma_1 = \Sigma_2 = \cdots = \Sigma_g = \Sigma$$

Then the discriminant function simplifies to:

$$\delta_i(\mathbf{x}) = \mathbf{x}^T \Sigma^{-1} \boldsymbol{\mu}_i - \tfrac{1}{2} \boldsymbol{\mu}_i^T \Sigma^{-1} \boldsymbol{\mu}_i + \ln(\pi_i)$$

* A linear function of $\mathbf{x}$, hence the name. * The classification boundaries between groups are linear hyperplanes.

6.2 Quadratic Discriminant Analysis (QDA)

Assumption: Each group has its own covariance matrix,

$$\Sigma_1 \neq \Sigma_2 \neq \ldots \neq \Sigma_g$$

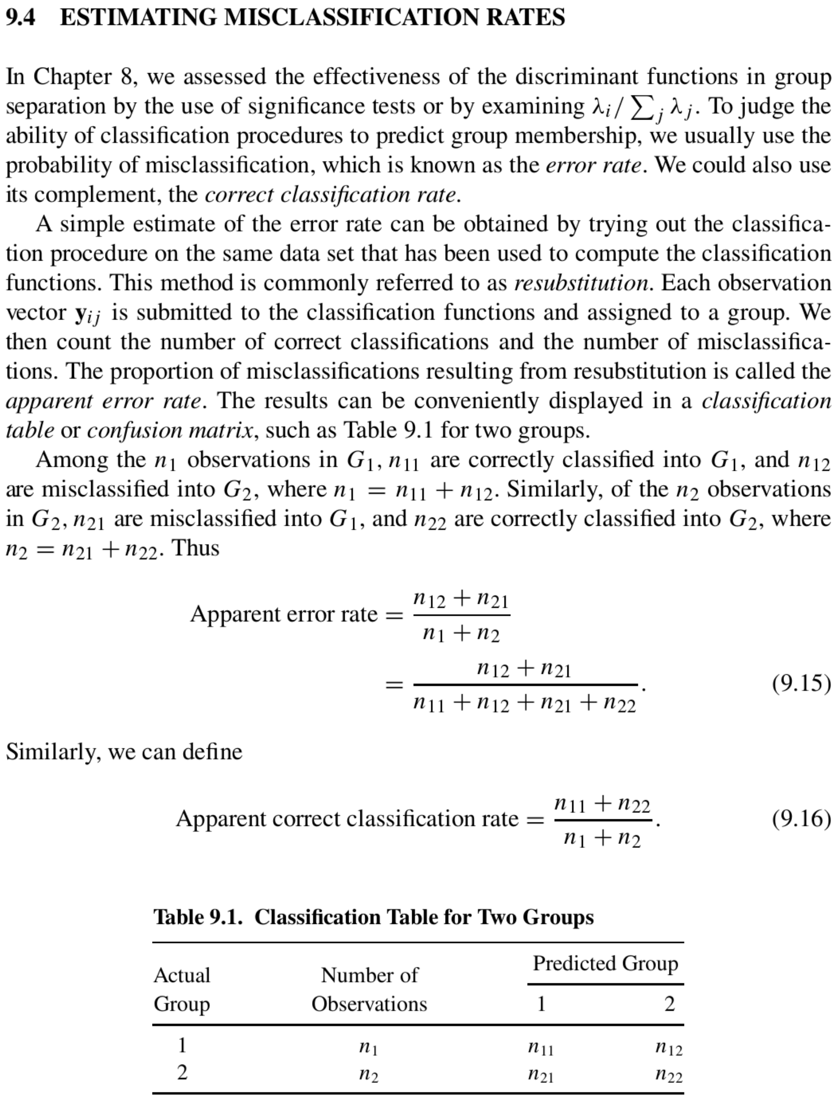
Then the discriminant function becomes:

$$\delta_i(\mathbf{x}) = -\frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^T \Sigma_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) + \ln(\pi_i)$$

* A quadratic function of **x**. * Boundaries between groups are quadratic surfaces (curved). * More flexible but requires larger sample sizes to estimate separate covariance matrices reliably.

7 Estimating Misclassification Rates

A central issue in classification analysis is the accuracy of the classification rule. No matter how carefully derived, classification functions will inevitably misclassify some observations when population distributions overlap. To assess the effectiveness of a classifier, we must estimate its misclassification rate.



9.4 ESTIMATING MISCLASSIFICATION RATES

In Chapter 8, we assessed the effectiveness of the discriminant functions in group separation by the use of significance tests or by examining $\lambda_i / \sum_j \lambda_j$. To judge the ability of classification procedures to predict group membership, we usually use the probability of misclassification, which is known as the *error rate*. We could also use its complement, the *correct classification rate*.

A simple estimate of the error rate can be obtained by trying out the classification procedure on the same data set that has been used to compute the classification functions. This method is commonly referred to as *resubstitution*. Each observation vector $\mathbf{y}_{ij}$ is submitted to the classification functions and assigned to a group. We then count the number of correct classifications and the number of misclassifications. The proportion of misclassifications resulting from resubstitution is called the *apparent error rate*. The results can be conveniently displayed in a *classification table* or *confusion matrix*, such as Table 9.1 for two groups.

Among the n_1 observations in G_1, n_{11} are correctly classified into G_1, and n_{12} are misclassified into G_2, where $n_1 = n_{11} + n_{12}$. Similarly, of the n_2 observations in G_2, n_{21} are misclassified into G_1, and n_{22} are correctly classified into G_2, where $n_2 = n_{21} + n_{22}$. Thus

$$\begin{aligned} \text{Apparent error rate} &= \frac{n_{12} + n_{21}}{n_1 + n_2} \\ &= \frac{n_{12} + n_{21}}{n_{11} + n_{12} + n_{21} + n_{22}}. \end{aligned} \tag{9.15}$$

Similarly, we can define

$$\text{Apparent correct classification rate} = \frac{n_{11} + n_{22}}{n_1 + n_2}. \tag{9.16}$$

Table 9.1. Classification Table for Two Groups

Actual	Number of	Predicted Group	
Group	Observations	1	2
1	n_1	n_{11}	n_{12}
2	n_2	n_{21}	n_{22}

8 Improved Estimates of Error Rates

When parameters are estimated from finite samples, the apparent error rate (i.e., the proportion of training observations misclassified by the fitted rule) is usually too optimistic.

Reason: Each observation influences the construction of the classifier (through estimated means and covariances). As a result, the classifier tends to fit the training data unusually well, underestimating the true error on new data.

Hence, we need methods to obtain less biased estimates of misclassification probability.

8.1 Partitioning the Sample

One approach is to split the sample into two subsets:

1. Training sample – used to construct the classification rule. 2. Test sample (or validation sample) – used to estimate the misclassification rate.

Procedure:

* Fit the classification rule using only the training data. * Apply it to the test data (which were not used in estimation). * Count the misclassifications and compute the proportion.

$$P\hat{M}C = \frac{\text{Number misclassified in test set}}{\text{Total size of test set}}$$

Advantages:

* Provides an unbiased estimate of classification performance.

Limitations:

* Reduces the effective sample size available for both training and testing. * If sample size is small, parameter estimates may be unstable, and test results unreliable.

8.2 Holdout Method

A more refined approach for small samples is the holdout (or “leave-one-out”) method.

Leave-One-Out Cross-Validation (LOOCV):

* Omit a single observation from the dataset. * Construct the classification rule using the remaining $n-1$ observations. * Classify the omitted observation. * Repeat for each observation in the dataset.

The estimated error rate is:

$$P\hat{M}C = \frac{\text{Number of misclassified observations across all trials}}{n}.$$

Advantages:

* Uses nearly all data for both training and testing. * Less biased than the apparent error rate.

Disadvantages:

* Computationally intensive (though feasible with modern computing). * Variability may still be high for small n.

Other variations: **k-fold cross-validation** (partition data into k folds, use $k-1$ for training and 1 for testing, rotate, then average error rates).

8.3 Subset Selection

Not all measured variables contribute meaningfully to classification. Some may be irrelevant, redundant, or noisy. Including such variables can:

* Inflate estimation error of covariance matrices. * Reduce classification accuracy. * Complicate interpretation.

Thus, **variable (subset) selection** is important.