Chapter 6

Discriminant Function Analysis

Suppose that we are given a learning set $\mathbf{x}$ of multivariate observations, where $\mathbf{x} \in \mathbb{R}^{n \times p}$. That is, n multivariate observations, with p variables. In addition, there is a variable $p+1$ which is a *class* variable taking values in $\mathcal{C} = \{C_1, \ldots, C_m\}$. That is, each observation comes from one of m pre-defined classes. In this set up, there are p *classification* variables and m groups or classes. Suppose that there are n_j observations from group C_j, for $j = 1, \ldots, m$. For example, there is a data set containing information of several measurements from skulls found in Egypt. There are 150 skulls. They are from 5 different periods, 30 skulls from each period. There is one class variable (the period) and four classification variables (the measurements). The question is whether the age of the skull can be inferred from the measurement variables. The data may be represented by:

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_m \end{pmatrix}$$

where m is the number of groups and

$$\mathbf{x}_j = \begin{pmatrix} x_{1j1} & \ldots & x_{1jp} \\ \vdots & & \vdots \\ x_{n_j j1} & \ldots & x_{n_j jp} \end{pmatrix}$$

and $n = \sum_{j=1}^{m} n_j$. These observations are described as *labelled observations*. There are two main goals:

- **Discrimination** Use the information in a learning set of labelled observations to construct a *classifier* (or *classification rule*) that will separate the predefined classes as much as possible.
- **Classification** Given a set of measurements on a new *unlabelled* observation, use the classifier to decide which class the observation belongs to.

There are two basic methods for discriminant analysis; the *maximum likelihood* method and *Fisher's Linear Discriminant Function* method. The maximum likelihood method may be used when the

probability distribution of each population is known; the linear discriminant function method is used when the probability distribution is unknown.

6.1 The Maximum Likelihood Discriminant Rule

The *maximum likelihood rule* is used when the probability distribution, or at least the parametric family of probability distributions, is known for each population. Unknown parameters are estimated by the training data and the estimates plugged in. Then a new observation $\underline{x}$ is allocated to group j if $\widehat{L}_j(\underline{x}) = \max_m \widehat{L}_m(\underline{x})$, where $\widehat{L}_k : k = 1, \ldots, m$ is the estimated likelihood function (when the parameter estimates have been plugged in). It is assumed that the situation where there are two groups which maximise the likelihood will not arise. If it does, a classification cannot be determined.

Example 6.1 (Normal Populations, same covariance structure)**.**

Assume that $\underline{X}_j \sim N(\underline{\mu}_j, C)$ for $j = 1, \ldots, m$. That is, from group j, the observations are independent identical multivariate normal, with mean vector $\underline{\mu}_j$ and covariance matrix C. The covariance matrix is assumed to be the *same* for each group. Then

$$L_j(\underline{x}) = \frac{1}{(2\pi)^{p/2}|C|^{1/2}} \exp\left\{-\frac{1}{2}(\underline{x} - \underline{\mu}_j)^t C^{-1}(\underline{x} - \underline{\mu}_j)\right\},$$

For an observation $\underline{x}$, finding the j that maximises $L_\cdot(\underline{x})$ is equivalent to finding the j that maximises

$$\mathcal{L}_j(\underline{x}) := \ln L_j(\underline{x}) = -\frac{p}{2}\ln(2\pi) - \frac{1}{2}\ln|C| - \frac{1}{2}(\underline{x} - \underline{\mu}_j)^t C^{-1}(\underline{x} - \underline{\mu}_j).$$

If the parameters are unknown, they are estimated from the training examples. The expectation vectors $\underline{\mu}_j$ are estimated by the sample average $\overline{\underline{x}}_j$ for group j and the covariance matrix C is estimated by S, the *pooled* covariance matrix from *all* the observations. When classifying a new observation, the problem is then to find the j which *minimises* the *Mahalanobis distance* from the observation $\underline{x}$ to the centre of group j. Recall the definition of the Mahalanobis distance:

$$D_j^2 = (\underline{x} - \overline{\underline{x}}_j)^t S^{-1}(\underline{x} - \overline{\underline{x}}_j).$$

New observations are *classified* as belonging to group j for which D_j is smallest.

6.1.1 The Bayes Discriminant Rule

This is almost the same as likelihood, except that there is a prior probability over classes; if X_{p+1} is the class variable, then

$$\mathbb{P}(X_{p+1} = C_i) = \pi_i.$$

The *posterior probability* for class C_i given $X = (X_1 \ldots, X_p)$ is then

$$\mathbb{P}(X_{p+1} = C_i | X = x) \propto \pi_i L_i(x).$$

The observation is then classified as class $C_i; i = \text{argmax}_j \pi_j L_j(x)$ (these are estimated by plugging in the appropriate parameter estimates).

6.2 The Linear Discriminant Function

Suppose we have two classes, C_1 and C_2. Suppose we have two normal populations, $\underline{X}_1 \sim N(\underline{\mu}_1, C)$ and $\underline{X}_2 \sim N(\underline{\mu}_2, C)$. Let f_1 and f_2 denote the respective densities. Set

$$\mathbb{L}(\underline{x}) := \ln \frac{f_1(\underline{x})}{f_2(\underline{x})}$$

Then

$$\begin{aligned}
\mathbb{L}(\underline{x}) &= -\frac{1}{2}(\underline{x} - \underline{\mu}_1)^t C^{-1}(\underline{x} - \underline{\mu}_1) + \frac{1}{2}(\underline{x} - \underline{\mu}_2)^t C^{-1}(\underline{x} - \underline{\mu}_2) \\
&= (\underline{\mu}_1 - \underline{\mu}_2)C^{-1}\underline{x} - \frac{1}{2}\underline{\mu}_1^t C^{-1}\underline{\mu}_1 + \frac{1}{2}\underline{\mu}_2 C^{-1}\underline{\mu}_2 \\
&= (\underline{\mu}_1 - \underline{\mu}_2)C^{-1}\underline{x} - \frac{1}{2}(\underline{\mu}_1^t + \underline{\mu}_2^t)C^{-1}(\underline{\mu}_1 - \underline{\mu}_2) \\
&= (\underline{\mu}_1 - \underline{\mu}_2)^t C^{-1}(\underline{x} - \overline{\underline{\mu}}) = b_0 + \underline{b}^t \underline{x}
\end{aligned}$$

where

$$\overline{\underline{\mu}} = \frac{1}{2}(\underline{\mu}_1 + \underline{\mu}_2).$$

This is a linear function, where

$$\left\{ \begin{array}{l} \underline{b} = C^{-1}(\underline{\mu}_1 - \underline{\mu}_2) \\ b_0 = -\frac{1}{2}\left\{\underline{\mu}_1^t C^{-1}\underline{\mu}_1 - \underline{\mu}_2^t C^{-1}\underline{\mu}_2\right\} \end{array} \right. \tag{6.1}$$

The function $\mathbb{L}$ is known as the *Linear Discriminant Function* (LDF). It partitions the space $\mathbb{R}^p$ into disjoint classification regions $\mathcal{R}_1$ and $\mathcal{R}_2$. If $\underline{x}$ falls into $\mathcal{R}_1$, then the observation is classified as belonging to C_1. If it falls into $\mathcal{R}_2$, then it is classified as belonging to C_2.

6.3 Misclassification Probability

$\mathbb{L}(X)$ is linear in X and therefore, conditioned on the class, is a Gaussian random variable. Its means for each class and variance can be computed quite easily. Let $\mathcal{C}$ denote the class variable. Then, using $\mathbb{L}(X) = (\mu_1 - \mu_2)' C^{-1}(X - \frac{1}{2}(\mu_1 + \mu_2))$,

$$\mathbb{E}[\mathbb{L}(X)|\mathcal{C} = C_1] = \frac{1}{2}(\mu_1 - \mu_2)'C^{-1}(\mu_1 - \mu_2) \qquad \mathbb{E}[\mathbb{L}(X)|\mathcal{C} \in C_1] = -\frac{1}{2}(\mu_1 - \mu_2)'C^{-1}(\mu_1 - \mu_2).$$

Its variance, conditioned on the class, is:

$$\text{Var}(\mathbb{L}(X)|\mathcal{C} = C_1) = \text{Var}(\mathbb{L}(X)|\mathcal{C} = C_2) = (\mu_1 - \mu_2)'C^{-1}CC^{-1}(\mu_1 - \mu_2) = (\mu_1 - \mu_2)'C^{-1}(\mu_1 - \mu_2).$$

Define

$$\Delta^2 = (\underline{\mu}_1 - \underline{\mu}_2)^t C^{-1}(\underline{\mu}_1 - \underline{\mu}_2).$$

then

$$\mathbb{L}(X)|\{\mathcal{C} = C_1\} \sim N(\frac{1}{2}\Delta^2, \Delta^2) \qquad \mathbb{L}(X)|\{\mathcal{C} = C_2\} \sim N(-\frac{1}{2}\Delta^2, \Delta^2).$$

Let M denote the event of misclassification. The *misclassification probabilities* for individuals from the respective groups is therefore:

$$\mathbb{P}(M|\mathcal{C} = C_1) = \mathbb{P}(\mathbb{L}(X) < 0|\mathcal{C} \in C_1), \qquad \mathbb{P}(M|\mathcal{C} = C_2) = \mathbb{P}(\mathbb{L}(X) > 0|\mathcal{C} \in C_2)$$

where, for a random variable $Y \sim N(\frac{1}{2}\Delta^2, \Delta^2)$,

$$\mathbb{P}(\mathbb{L}(\underline{X}) < 0|\mathcal{C} = C_1) = \mathbb{P}(Y < 0) = \mathbb{P}(Z < -\frac{1}{2}\Delta) = \Phi\left(-\frac{\Delta}{2}\right)$$

and

$$\mathbb{P}(\mathbb{L}(X) > 0|\mathcal{C} = C_2) = \mathbb{P}(Y < 0) = \mathbb{P}(Z < -\frac{1}{2}\Delta) = \Phi\left(-\frac{\Delta}{2}\right).$$

□

A graph of $\mathbb{P}(M|\mathcal{C} = C_i)$ against Δ shows a downward sloping curve. It has value $\frac{1}{2}$ when $\Delta = 0$ (the two populations are identical) and tends to 0 as Δ increases. In other words, the greater the distance between the two population means, the less likely one is to misclassify $\underline{x}$.

6.4 Quadratic Discrimination

When populations are normal, but the covariance matrices are not equal, the maximum likelihood technique leads to *quadratic* discriminant functions.

Theorem 6.2. *Suppose that $\underline{X}_j \sim N(\underline{\mu}_j, C_j)$ (that is, a p-variate observation from population j has multivariate normal distribution with mean vector $\underline{\mu}_j$ and covariance matrix C_j). Suppose that $\overline{\underline{x}}_j$ and S_j are the maximum likelihood estimates of the mean and covariance matrix for population j. Then the maximum likelihood discrimination rule, where the estimates are used in place of the true parameter values, allocates a new observation $\underline{x}$ to population j if and only if*

$$(\underline{x} - \overline{\underline{x}}_j)^t (S_k^{-1} - S_j^{-1})(\underline{x} - \overline{\underline{x}}_j) + (\overline{\underline{x}}_j - \overline{\underline{x}}_k)^t S_k^{-1}(2\underline{x} - (\overline{\underline{x}}_j + \overline{\underline{x}}_k)) + \ln \frac{|S_k|}{|S_j|} > 0, \quad k \neq j. \tag{6.2}$$

Proof The log likelihood for population j is

$$l_j(\underline{x}) = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln|S_j| - \frac{1}{2}(\underline{x} - \overline{\underline{x}}_j)S_j^{-1}(\underline{x} - \overline{\underline{x}}_j).$$

The result follows from straightforward arithmetic manipulation. □

Corollary 6.3. *If $m = 2$ (two populations) and $C_1 = C_2 = C$ and this model is used, with $S = \frac{1}{n_1+n_2}W$, then the maximum likelihood method allocates a new observation $\mathbf{x}$ to population 1 if and only if*

$$(\overline{\underline{x}}_1 - \overline{\underline{x}}_2)^t W^{-1}(\underline{x} - \frac{1}{2}(\overline{\underline{x}}_1 + \overline{\underline{x}}_2)) > 0.$$

Proof Straightforward exercise. □

6.5 Fisher's Discriminant Function

We now consider *Fisher's Discriminant Function*, which is based on ANOVA (sums of squares). It will be clear that the likelihood discrimination rule for multivariate normal observations, where each population has the same covariance, the likelihood discriminant is the same as Fisher's discriminant.

Fisher's idea was to look for appropriate linear combinations of the variables:

$$Z = \sum_{k=1}^{p} a_k X_k$$

to maximise the distance between the various groups. Fisher (1936) suggested taking the linear combination that maximises the F ratio in the ANOVA table. Let $n = \sum_{j=1}^{m} n_j$, $\overline{z} = \frac{1}{n}\sum_{k=1}^{m}\sum_{l=1}^{n_k} z_{lk}$, $\overline{z}_k = \frac{1}{n_k}\sum_{l=1}^{n_k} z_{lk}$. The ANOVA is

source	$d.f.$	mean square	f
$SSB = \sum_{j=1}^{m} n_j(\overline{z}_j - \overline{z})^2$	$m-1$	$M_B = \frac{SSB}{m-1}$	$\frac{M_B}{M_E}$
$SSE = \sum_{j=1}^{m}\sum_{k=1}^{n_j}(z_{kj} - \overline{z}_j)^2$	$n-m$	$M_E = \frac{SSE}{n-m}$	
$SST = SSB + SSE = \sum_{j=1}^{m}\sum_{k=1}^{n_j}(z_{kj} - \overline{z})^2$	$n-1$		

Let T, W and B be the matrices for *Total*, *Within* (or error) and *Between* classes sums of squares defined by

$$T_{ab} = \sum_{j=1}^{m}\sum_{k=1}^{n_j}(x_{kja} - \overline{x}_a)(x_{kjb} - \overline{x}_b),$$

$$W_{ab} = \sum_{j=1}^{m}\sum_{k=1}^{n_j}(x_{kja} - \overline{x}_{.ja})(x_{kjb} - \overline{x}_{.jb}),$$

$$B_{ab} = \sum_{j=1}^{m} n_j(\overline{x}_{.ja} - \overline{x}_{..a})(\overline{x}_{.jb} - \overline{x}_{..b}),$$

where x_{kja} denotes observation k from sample j for variable a, $\overline{x}_a = \frac{1}{mn_j}\sum_{j=1}^{m}\sum_{k=1}^{n_j} x_{kja}$ and $\overline{x}_{ja} = \frac{1}{n_j}\sum_{k=1}^{m} x_{kja}$. As described before, T denotes 'total', B denotes 'between groups' and W as 'within groups', so W with suitable normalisation is an estimate of the error covariance. Note that

$$B = T - W.$$

Then it turns out that (this is one of the tutorial exercises) that Fisher's rule amounts to choosing a vector $\underline{a} \in \mathbb{R}^p$ that maximises the ratio

$$\frac{\underline{a}^t B \underline{a}}{\underline{a}^t W \underline{a}}.$$

Then the discriminant function is $Z = \sum_{j=1}^{p} a_j X_j$.

Definition 6.4. *The linear function Z satisfying $Z(\mathbf{x}) = \sum_{j=1}^{p} a_j x_j$ is called Fisher's* linear discriminant function. *The linear combination $\underline{a}^t\underline{x}$ is also called the* first canonical variate.

Theorem 6.5. *The vector $\underline{a}$ that maximises $\frac{\underline{a}^t B\underline{a}}{\underline{a}^t W\underline{a}}$ is the eigenvector corresponding to the largest eigenvalue of the $p \times p$ matrix $W^{-1}B$.*

Proof (an exercise - details in the exercise set) □

Let $\overline{\underline{x}}_j$ denote the mean vector for population (or group) j. Using Fisher's linear discriminant function, the rule is to assign a p- variate observation $\mathbf{x}$ to the class for which $|\mathbf{a}^t(\mathbf{x} - \overline{\mathbf{x}}_j)|$ is lowest.

Exercise Consider two populations j and k with mean vectors $\overline{\underline{x}}_j$ and $\overline{\underline{x}}_k$ respectively and assume that the populations have been labelled such that $\underline{a}^t\overline{\underline{x}}_j \geq \underline{a}^t\overline{\underline{x}}_k$. Then, for any $\underline{x} \in \mathbb{R}^p$,

$$|\underline{a}^t(\underline{x} - \underline{x}_j)| < |\underline{a}^t(\underline{x} - \underline{x}_k)| \implies \underline{a}^t\left(\underline{x} - \frac{1}{2}(\overline{\underline{x}}_j + \overline{\underline{x}}_k)\right) > 0.$$

□

This enables the following interpretation of Fisher's linear discriminant rule. The set

$$H_{jk} = \left\{\underline{x} \in \mathbb{R}^p | \underline{a}^t\left(\underline{x} - \frac{1}{2}(\overline{\underline{x}}_j + \overline{\underline{x}}_k)\right) = 0\right\}$$

defines a hyperplane perpendicular to the vector $\underline{a}$. This hyperplane divides $\mathbb{R}^p$ into two disjoint half spaces; the mean $\overline{\underline{x}}_j$ lies in one and the mean $\overline{\underline{x}}_k$ lies in the other.

By considering all pairs of populations (j, k) with $1 \leq j \leq m$ and $1 \leq k \leq m$, Fisher's linear discriminant function splits $\mathbb{R}^p$ into m disjoint regions

$$\mathbb{R}^p = \mathcal{R}_1 \cup \ldots \cup \mathcal{R}_m$$

by considering all $p(p-1)/2$ hyperplanes H_{jk}. The region $\mathcal{R}_j$ corresponds to the region where an observation $\underline{x}$ will be classified as belonging to population j. These hyperplanes are all perpendicular to the vector $\underline{a}$.

To find $\mathcal{R}_j$, drop a line from $\overline{\underline{x}}_j$, perpendicular to the vector $\underline{a}$, to the line through the origin containing the point $\underline{a}$. Denote the point of intersection by $\underline{y}_j$. From the $m-1$ hyperplanes $H_{j1}, \ldots, H_{jm}$ (there is no hyperplane H_{jj}), find the two with smallest distance from $\underline{y}_j$, on either side of that point. The region $\mathcal{R}_j$ is the region bounded by these two hyperplanes.

6.6 Canonical Discriminant Functions

Fisher's technique may be extended quite easily to obtain more discriminant functions, to sharpen up the clasification. Let $s = \min(p, m-1)$ and let $\lambda_1 > \ldots > \lambda_s$ be the first s eigenvalues of $W^{-1}B$ and let $(a_{i1}, \ldots, a_{ip})^t$ denote the eigenvector corresponding to eigenvalue λ_i and set

$$Z_i(\underline{x}) = \sum_{k=1}^{p} a_{ik} x_k.$$

Then Z_i is known as the ith canonical discriminant function. It turns out (proof omitted) that the ith eigenvalue is the ratio of the within group sum of squares to the between group sum of squares for

Z_1 is the combination that gives the largest M_B/M_W ratio, subject to the constraint that $\sum a_{1k}^2 = 1$.

Z_2 is the combination that gives the largest M_B/M_W ratio, subject to constraints that $\sum_k a_{2k}^2 = 1$ and $\sum_{jk} a_{1j} S_{jk} a_{2k} = 0$; i.e. Z_2 is statistically uncorrelated with Z_1.

For $i \geq 2$, Z_i is the linear combination that gives the largest M_B/M_W ratio, subject to the constraints that $\sum_{k=1}^{p} a_{jk}^2 = 1$, $j = 1, \ldots, p$ and $\sum_{\alpha,\beta} a_{j\alpha} a_{k\beta} S_{\alpha\beta} = 0$ for all $1 \leq j < k \leq i$.

Where discriminant analysis is useful, the first few functions ought to be sufficient to show the group differences. Hopefully, sufficiently few will be required so that they can be used to represent the group differences graphically.

Important Remark The value $s = \min(p, m-1)$ is the maximum number of canonical discriminant functions *available*; this is the *rank* of B as is easily checked and hence, if $s < p$ all remaining eigenvalues $\lambda_{s+1} = \ldots = \lambda_p = 0$.

Significance Tests The Hotelling T^2 test may be used to test for a significant difference between the mean values for any pair of groups. Other tests, which are variants of this test, may be used to detect overall significant differences between the means for the m groups.

χ^2 test In addition, let $(\lambda_j)_{j=1}^s$ denote the eigenvalues of the matrix $W^{-1}B$. Then, approximately,

$$\phi_j^2 := \left(n - 1 - \frac{p+m}{2}\right) \ln(1+\lambda_j) \sim \chi^2_{p+m-2j}.$$

A large value substantiates the claim that there are significant differences of the mean vectors between the groups. Alternatively, $\phi_j^2 + \ldots + \phi_s^2$ may be used, the χ^2 having the d.f. $\sum_{k=j}^s (p+m-2k)$.

Warnings

1. The χ^2 test does not seem to be robust if assumptions $\underline{X}_j \sim N(\mu_j, C)$ are relaxed. This contrasts with univariate analysis, where the results seem to be robust when assumptions of normality are relaxed.

2. Even if the data is normal, the statistical values for λ_j may appear in the wrong order, if the variance is large. The test does not take this possibility into account. A large value for an eigenvalue further down on the list that happens by chance will give a wrong impression of the significance of all the eigenvalues; the test has a greater chance of wrongly indicating significance than the nominal significance level.

Example 6.6 (Egyptian Skulls)**.**

The matrices W, T, $B = T - W$ can be obtained and the matrix $W^{-1}B$ calculated and its eigenvalues computed. These turn out to be $\lambda_1 = 0.437$, $\lambda_2 = 0.035$, $\lambda_3 = 0.015$, $\lambda_4 = 0.002$. The corresponding eigenvectors may be calculated, giving (up to scaling) canonical discriminant functions

$$Z_1 = 0.127X_1 - 0.037X_2 - 0.145X_3 - 0.0083X_4$$
$$Z_2 = 0.039X_1 + 0.210X_2 - 0.068X_3 - 0.077X_4$$
$$Z_3 = 0.093X_1 - 0.025X_2 + 0.015X_3 - 0.295X_4$$
$$Z_4 = 0.149X_1 - 0.000X_2 + 0.133X_3 + 0.067X_4$$

The eigenvalue λ_1 is much larger than the others; most of the sample differences are described by Z_1 alone. Large values correspond to skulls which are tall and narrow with long jaws and short nasal heights.

The *means* and *standard deviations* for the discriminant function Z_1 may be computed for the five samples. They are

	group	mean	standard deviation
I:	Earl predynastic	-0.029	0.097
II:	Late predynastic	-0.043	0.071
III:	12th and 13th dynasties	-0.099	0.075
IV:	Ptolemaic	-0.143	0.080
V:	Roman	-0.167	0.095

This discriminant function shows a clear trend in the mean. It is decreasing over time, indicating *on average* shorter broader skulls, with short jaws, but relatively larger nasal heights. But this is very much an *average* change; the standard deviation is rather large. When the 150 skulls are classified according to the group to which they are closest according to the Mahalanobis distance, rather many are wrongly classified. The following table, known as a *confusion table*, gives the number of objects which the classifier places in each class, from each class. The diagonal entries indicate the number that are correctly classified, the off-diagonals those that are incorrectly classified.

	number allocated to each group					
source group	I	II	III	IV	V	total
I	12	8	4	4	2	30
II	10	8	5	4	3	30
III	4	4	15	2	5	30
IV	3	3	7	5	12	30
V	2	4	4	9	11	30

□

Allowing for Additional Information

Suppose, for example, there are two groups and it is known that many more will fall into group 1 than into group 2. In that case, if an individual is allocated to each group, it makes sense to bias the allocation procedure in favour of group 1. The procedure of allocating an individual to the group with the smallest Mahalanobis distance is then modified, by taking into account prior probabilities of group membership.

Stepwise Discriminant Function Analysis

The standard approach to discriminant function analysis is to decide in advance the number of variables to be used. Alternatively, a *stepwise* approach may be adopted, when there are a very large number of variables, adding in the 'best' variable at each stage, until it is found that adding in extra variables

does not lead to better discrimination.

The main problem with stepwise discriminant function analysis is that it introduces bias. Given enough variables, it is likely that some combination of them will produce significant discriminant functions by chance alone.

To check that the results are valid, it might then be a good idea to (for example, with the Egyptian skull data) allocate the 150 skulls to the groups I,II,III,IV,V purely at random and see if the procedure is able to detect a pattern. If it can detect a pattern with the randomised data, then there is clearly a problem.

Jackknife Classification

A particular individual will necessarily affect the statistical average of the 'correct' group for that individual. To check that the classification procedure works, it is therefore probably better to remove that individual from the computations of sample means and sample covariance matrix, and then allocate the individual based on the analysis from which that individual has been removed. When the data set is reasonably large, this does not make much difference in practise.

6.7 LDA using Multiple Regression Techniques

The results on LDA can also be obtained using linear regression techniques. This may prove to be useful when we have a large number of variables and we would like to choose a subset of them for classification purposes. We may then employ LASSO to construct the classifier.

To use regression for LDA, create a variable Y which indicates which class the observations belong to, Then regress the feature variables X on Y.

Consider two classes, n_1 resp. n_2 in each class, items $1, \ldots, n_1$ belong to class 1 and items $n_1 + 1, \ldots, n_1 + n_2$ belong to class 2. let $Y_i = y_1$ for $i = 1, \ldots, n_1$ and y_2 for $i = n_1 + 1, \ldots, n_2$.

Let $X = \left(\frac{X_1}{X_2}\right)$ where X_1 and X_2 are respectively the $n_1 \times p$ and $n_2 \times p$ matrices containing the values of $(X_1, \ldots, X_p)$ for the observations for populations 1 and 2 respectively.

When classification is in view, we may use centred variables. Let

$$H = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^t$$

be the centring matrix and let

$$X^c = H_n X \qquad Y^c = H_n Y$$

so that the columns of X^c have mean zero and Y^c has mean 0. Therefore

$$T = X^{ct}X^c.$$

Regressing gives the OLS estimator

$$\widehat{\beta} = (X^{ct}X^c)^{-1}X^{ct}Y^c.$$

Set

$$d = \frac{1}{n_1}X_1^t\mathbf{1}_{n_1} - \frac{1}{n_2}X_2^t\mathbf{1}_{n_2},$$

The vector d is a p-vector where the entries are the differences of the sample means of the two populations for each variable.

A straightforward computation gives:

$$B = \frac{n_1n_2}{n}dd^t$$

Let

$$S_{XX} = X_1^tH_{n_1}X_1 + X_2^tH_{n_2}X_2$$

Here

$$S_{XX;ab} = \sum_{k=1}^{n_1}(x_{k1a} - \overline{x}_{.1a})(x_{k1b} - \overline{x}_{.1b}) + \sum_{k=1}^{n_2}(x_{k2a} - \overline{x}_{.2a})(x_{k2b} - \overline{x}_{.2b});$$

For two classes, the matrix S_{XX} is the matrix W from earlier. Set

$$k = \frac{n_1n_2}{n}$$

Then

$$X^{ct}X^c = S_{XX} + kdd^t$$

This is the identity $T = B + W$.

$$X^{ct}Y^c = k(y_1 - y_2)d$$
$$Y^{ct}Y^c = k(y_1 - y_2)^2.$$

It follows that

$$\widehat{\beta} = k(y_1 - y_2)(S_{XX} + kdd^t)^{-1}d = k(y_1 - y_2)S_{XX}^{-1}(I_p + kdd^tS_{XX}^{-1})^{-1}d.$$

Recall the matrix result:

$$(A + uv^t)^{-1} = A^{-1} - \frac{(A^{-1}u)(v^tA^{-1})}{1 + v^tA^{-1}u}.$$

Set $A = I_p$, $u = kd$, $v = S_{XX}^{-1}d$, then

$$(I_p + kdd^t S_{XX}^{-1})^{-1} = I_p - \frac{kdd^t S_{XX}^{-1}}{1 + kd^t S_{XX}^{-1} d} = \frac{I_p}{1 + kd^t S_{XX}^{-1} d}$$

from which

$$\widehat{\beta} = \frac{k(y_1 - y_2)}{n - 2 + T^2} \widehat{\Sigma}_{XX}^{-1} d$$

where $\widehat{\Sigma}_{XX} = \frac{1}{n-2} S_{XX}$ and

$$T^2 = kd^t \widehat{\Sigma}_{XX}^{-1} d = \frac{n_1 n_2}{n} (\overline{X}_1 - \overline{X}_2)^t \widehat{\Sigma}_{XX}^{-1} (\overline{X}_1 - \overline{X}_2)$$

is the *Hotelling T^2 statistic* for testing $\mu_1 = \mu_2$.

Recall the formulae for linear discriminant analysis (6.1) Note that $D^2 = d^t \widehat{\Sigma}_{XX}^{-1} d$ is proportional to the estimate of Δ and

$$\widehat{\beta} \propto \widehat{\Sigma}_{XX}^{-1} (\overline{X}_1 - \overline{X}_2) = \widehat{b}.$$

Variable Selection High dimensional data contains highly correlated variables. The equivalence between LDA and linear regression means that exactly the same techniques may be employed for making a selection; stepwise regression or other techniques that have not yet been encountered, such as LARS (least angle regression) and LASSO.

6.7.1 Logistic Discrimination

Consider two classes. Starting from

$$\log \frac{L_1(x)}{L_2(x)} = b_0 + b^t x$$

where

$$b = \Sigma_{XX}^{-1} (\mu_1 - \mu_2)$$

$$b_0 = -\frac{1}{2} (\mu_1^t \Sigma_{XX}^{-1} \mu_1 + \mu_2^t \Sigma_{XX}^{-1} \mu_2)$$

and using $\mathbb{P}(C_1|x) \propto L_1(x)$, $\mathbb{P}(C_2|x) \propto L_2(x)$ so that $\mathbb{P}(C_2|x) = 1 - \mathbb{P}(C_1|x)$, it follows that

$$\text{logit} p(C_1|x) = b_0 + b^t x$$

which is of the form of a logistic regression model. The logistic approach to discrimination assumes this linear model, estimates the parameters by logistic regression and assigns the observation to whichever category has the higher estimated likelihood.

6.8 Implementation in R

Implementation in R is straightforward, using (for example) the MASS library. This is illustrated using the `skulls.dat` data set.

```
www =
"https://www.mimuw.edu.pl/~noble/courses/MultivariateStatistics/data/
skulls.dat"
skulls <- read.table(www,header=T)
library("MASS")
fit <- lda(Year ~ MB + BH + BL + NH, data=skulls,
na.action="na.omit",
CV=TRUE)
```

'lda' stands for 'linear discriminant analysis'. The variable 'Year' is to be explained in terms of MB, BH, BL and NH. The 'na.action' refers to how R should treat a value that is not a number. The command 'CV = TRUE' generates the predictions. These are jacknifed (i.e. 'leave one out'). The `.$class` item gives the classes assigned to the skulls.

```
> head(fit$class)
[1] -1850 -4000 -3300 -4000 -1850 -200
Levels: -4000 -3300 -1850 -200 150
```

From the first 11 skulls, it is clear that they are not perfectly classified. To assess the accuracy of prediction, the following may help:

```
> ct <- table(skulls$Year, fit$class)
> ct

       -4000 -3300 -1850 -200 150
  -4000     9    10     5    4   2
  -3300    11     7     5    4   3
  -1850     6     4    12    2   6
  -200      3     3     7    5  12
  150       2     4     4   10  10
> diag(prop.table(ct, 1))
    -4000     -3300     -1850      -200       150
0.3000000 0.2333333 0.4000000 0.1666667 0.3333333
> sum(diag(prop.table(ct)))
[1] 0.2866667
```

Note that 'leave one out' is a more reliable method and this has substantially affected the accuracy of the prediction. The last item gives the total percentage correct.

Quadratic discriminant analysis may be carried out by substituting the `lda` command for `qda`. Quadratic discriminant analysis in this example does not give as good classification.

```
> fit <- qda(Year ~ MB + BH + BL + NH, data=na.omit(skulls), CV = TRUE,
prior=c(1,1,1,1,1)/5)
> ct <- table(skulls$Year, fit$class)
> ct

       -4000 -3300 -1850 -200 150
  -4000     8    12     4    4   2
  -3300    11     5     4    6   4
  -1850     4     5     6   11   4
  -200      2     3     2   14   9
  150       3     4     5   11   7
 > sum(diag(prop.table(ct)))
[1] 0.2666667
```

If one wants to obtain Fisher's canonical discriminant functions, this is not possible with the 'jacknifed' method; one needs to define a training data set. In this case, it is the whole data set. try

```
> fit2 <- lda(Year~MB + BH + BL + NH, data = skulls,CV=FALSE)
> fit2
Call:
lda(Year ~ MB + BH + BL + NH, data = skulls, CV = FALSE)

Prior probabilities of groups:
-4000 -3300 -1850  -200   150
  0.2   0.2   0.2   0.2   0.2

Group means:
            MB       BH       BL       NH
-4000 131.3667 133.6000 99.16667 50.53333
-3300 132.3667 132.7000 99.06667 50.23333
-1850 134.4667 133.8000 96.03333 50.56667
-200  135.5000 132.3000 94.53333 51.96667
150   136.1667 130.3333 93.50000 51.36667
```

```
Coefficients of linear discriminants:
           LD1         LD2         LD3           LD4
MB  0.12667629  0.03873784  0.09276835  0.1488398644
BH -0.03703209  0.21009773 -0.02456846 -0.0004200843
BL -0.14512512 -0.06811443  0.01474860  0.1325007670
NH  0.08285128 -0.07729281 -0.29458931  0.0668588797

Proportion of trace:
   LD1    LD2    LD3    LD4
0.8823 0.0809 0.0326 0.0042
```

These coefficients give the discriminant functions listed above. Discriminant analysis requires *training* data, which is used to construct the classifier, followed by data to be classified. Once the classifier has been constructed, classification is made using:

```
 > pred <- predict(fit2,skulls[,1:4])
```

The classes to which the objects are assigned are found in `pred$class`.

```
> ct2 <- table(skulls$Year,pred$class)
> ct2

       -4000 -3300 -1850 -200 150
  -4000    12     8     4    4   2
  -3300    10     8     5    4   3
  -1850     4     4    15    2   5
  -200      3     3     7    5  12
  150       2     4     4    9  11
```

Discriminant Function Analysis: Written Exercises

1. Let x_{ijk} denote observation i ($i = 1, \ldots, n_j$) from population j ($j = 1, \ldots, m$), for variable k ($k = 1, \ldots, p$). Let W denote the matrix with entries given by

$$W_{ab} = \sum_{j=1}^{m} \sum_{i=1}^{n_j} (x_{ija} - \bar{x}_{.ja})(x_{ijb} - \bar{x}_{.jb})$$

and let B denote the matrix with entries given by

$$B_{ab} = \sum_{j=1}^{m} n_j (\bar{x}_{.ja} - \bar{x}_{..a})(\bar{x}_{.jb} - \bar{x}_{..b}).$$

and T the matrix with entries

$$T_{ab} = \sum_{j=1}^{m} \sum_{i=1}^{n_j} (x_{ija} - \bar{x}_{..a})(x_{ijb} - \bar{x}_{..b}).$$

where . denotes the index that has been averaged over and the data matrix is:

$$\mathbf{x} = \left(\begin{array}{ccc} x_{111} & \ldots & x_{11p} \\ \vdots & & \vdots \\ x_{n_1 11} & \ldots & x_{n_1 1p} \\ \hline x_{121} & \ldots & x_{12p} \\ \vdots & & \vdots \\ x_{n_2 21} & \ldots & x_{n_2 2p} \\ \hline \vdots & \vdots & \\ \hline x_{1m1} & \ldots & x_{1mp} \\ \vdots & & \vdots \\ x_{n_m m1} & \ldots & x_{n_m mp} \end{array} \right)$$

Let $n = \sum_{j=1}^{m} n_j$ and let

$$H = I_n - \frac{1}{n} \mathbf{1}\mathbf{1}^t$$

where $\mathbf{1} = (1, \ldots, 1)^t$, the n vector where each entry is 1.

(a) Show that

$$T = W + B.$$

(b) Show that

$$T = \mathbf{X}^t H \mathbf{X}.$$

2. Let $\underline{a} \in \mathbb{R}^p$ be the vector defining Fisher's linear discriminant function; namely, the unit vector that maximises $\frac{\underline{a}^t B \underline{a}}{\underline{a}^t W \underline{a}}$. Show by computing partial derivatives

$$\frac{\partial}{\partial a_i}\left(\frac{\underline{a}^t B \underline{a}}{\underline{a}^t W \underline{a}}\right), \quad i = 1, \ldots, p$$

that it satisfies

$$W^{-1} B \underline{a} = \frac{\underline{a}^t B \underline{a}}{\underline{a}^t W \underline{a}} \underline{a}.$$

From this, note that it must be an *eigenvector* of $W^{-1}B$. Why is Fisherś linear discriminant function given by the eigenvector corresponding to the *largest* eigenvalue?

3. Consider a situation where an individual is chosen from $m = 2$ populations. Let $\bar{\underline{x}}_1$ and $\bar{\underline{x}}_2$ denote the sample mean vectors for populations 1 and 2 respectively. Let $\underline{d} = \bar{\underline{x}}_1 - \bar{\underline{x}}_2$. Let n_1, n_2 denote the number of observations from populations 1 and 2 respectively and let $n = n_1 + n_2$.

   (a) Show that

   $$B = \frac{n_1 n_2}{n} \underline{d}\underline{d}^t.$$

   (b) Hence show that in this case, for any eigenvalue λ of $W^{-1}B$ and corresponding eigenvector $\underline{a}$ such that $\underline{a}^t \underline{d} \neq 0$,

   $$W^{-1} \underline{d} = \frac{n}{n_1 n_2} \frac{\lambda}{\underline{d}^t \underline{a}} \underline{a}.$$

   (c) Use this to show that there is exactly one non zero eigenvalue of $W^{-1}B$, which is therefore the largest one, and that $W^{-1}\underline{d}$ is an eigenvector with this eigenvalue.

4. Consider two bivariate normal distributions, with true means $\underline{\mu}_1 = (0, 0)^t$ and $\underline{\mu}_2 = (1, 0)^t$ and true covariances

$$C_1 = \begin{pmatrix} \frac{1}{4} & 0 \\ 0 & 1 \end{pmatrix} \qquad C_2 = \begin{pmatrix} 4 & 0 \\ 0 & \frac{1}{4} \end{pmatrix}$$

respectively. The corresponding Maximum Likelihood Discriminant Rule is based on a division of $\mathbb{R}^2$ into two disjoint regions $\mathcal{R}_1$ and $\mathcal{R}_2$.

   (a) Give the equation of the boundary separating the two regions.

   (b) Sketch the two regions.

   (c) To which population would you assign a new individual with measurements $(\frac{1}{2}, \frac{1}{2})^t$?

5. Let $\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_m \end{pmatrix}$ be a multivariate random sample on p variables taken from m populations, where n_j is the number of units taken from population j and $n = \sum_{j=1}^{p} n_j$ ($\mathbf{X}_j$, $j = 1, \ldots, m$ is a random sample from population j with entries x_{ijk} denoting observation on individual i from population j on variable k).

(a) Define the *group means* $\underline{\bar{x}}_1, \ldots \underline{\bar{x}}_m$, the *between group* sum of squares matrix B and the within group sum of squares matrix W. In what follows, assume that W^{-1} exists.

(b) Suppose that λ_1 is the largest eigenvalue of $W^{-1}B$ with associated eigenvector $\underline{a}_1 \in \mathbb{R}^p$.

i. Prove that

$$\lambda_1 = \frac{(\underline{a}_1^t B \underline{a}_1)}{\underline{a}_1^t W \underline{a}_1} \geq \frac{(\underline{b}^t B \underline{b})}{\underline{b}^t W \underline{b}}$$

for all non zero p vectors $\underline{b}$.

ii. Why does a large ratio $\frac{(\underline{b}^t B \underline{b})}{\underline{b}^t W \underline{b}}$ help to discriminate between groups?

(c) Consider $m = 2$, where both populations have bivariate normal distribution and both the covariance matrices are the same. Estimating this covariance matrix by $\frac{1}{n_1+n_2-2}W$, prove that the maximum likelihood discriminant rule allocates a new observation $\underline{x}$ to population 1 if and only if

$$(\underline{\bar{x}}_1 - \underline{\bar{x}}_2)^t W^{-1} \left(\underline{x} - \frac{1}{2}(\underline{\bar{x}}_1 + \underline{\bar{x}}_2) \right) > 0.$$

6. For two classes, show that Fisher's linear discriminant function may be obtained by OLS regression.

Short Answers

1. (a)

$$\begin{aligned} T_{ab} &= \sum_{j=1}^{m}\sum_{i=1}^{n_j}(x_{ija} - \bar{x}_{.ja} + \bar{x}_{.ja} - \bar{x}_{..a})(x_{ijb} - \bar{x}_{.jb} + \bar{x}_{.jb} - \bar{x}_{..b}) \\ &= \sum_{j=1}^{m}\sum_{i=1}^{n_j}(x_{ija} - \bar{x}_{.ja})(x_{ijb} - \bar{x}_{.jb}) + \sum_{j=1}^{m} n_j(\bar{x}_{.ja} - \bar{x}_{..a})(\bar{x}_{.jb} - \bar{x}_{..b}) \\ &\quad + \sum_{j=1}^{m}\sum_{i=1}^{n_j}(x_{ija} - \bar{x}_{.ja})(\bar{x}_{.jb} - \bar{x}_{.jb}) + \sum_{j=1}^{m}\sum_{i=1}^{n_j}(x_{ijb} - \bar{x}_{.jb})(\bar{x}_{.ja} - \bar{x}_{.ja}) \\ &= W_{ab} + B_{ab} \end{aligned}$$

   (b) Note that $H = H^t = HH^t$ so that

$$X^tHX = X^tHH^tX$$

$$(H^tX)_{ijk} = x_{ijk} - \bar{x}_{..k}$$

   so that

$$(X^tHH^tX)_{kl} = \sum_{j=1}^{m}\sum_{i=1}^{n_j}(x_{ijk} - \bar{x}_{..k})(x_{ijl} - \bar{x}_{..l}) = T_{kl}$$

   as required

2. Let $A(\underline{a}) = \frac{\underline{a}^tB\underline{a}}{\underline{a}^tW\underline{a}}$. Let $\tilde{\underline{a}}$ denote the vector that maximises A. Then

$$\frac{\partial}{\partial a_j}A = \frac{2}{2\underline{a}^tW\underline{a}}\left(B_{j.}\underline{a} - AW_{j.}\underline{a}\right).$$

   At the maximum this is zero, so that,

   at the maximum,

$$B\tilde{\underline{a}} = A(\tilde{\underline{a}})W\tilde{\underline{a}}$$

   where $A(\tilde{\underline{a}})$ is a scalar, so that

$$W^{-1}B\tilde{\underline{a}} = A(\tilde{\underline{a}})\tilde{\underline{a}}.$$

   By the definition of eigenvalue, it follows that $A(\tilde{\underline{a}})$ is an eigenvalue. The vector $\tilde{\underline{a}}$ may be taken such that $\sum_{j=1}^{p}\tilde{a}_j^2 = 1$, since multiplying a vector $\underline{b}$ by a scalar does not alter the value of $A(\underline{b})$.

   To show that the maximum of the expression is also the maximum eigenvalue of $W^{-1}B$ and $\tilde{\underline{a}}$ the corresponding eigenvector, consider any other eigenvector $\underline{b}$ with eigenvalue λ. Then

$$W^{-1}B\underline{b} = \lambda\underline{b}$$

   so

$$\underline{b}^tB\underline{b} = \lambda\underline{b}^tW\underline{b}$$

so

$$\lambda = A(\underline{b}).$$

By definition of $\underline{\tilde{a}}$ (a vector that maximises $A(\underline{a})$), it follows that

$$\lambda = A(\underline{b}) \leq A(\underline{\tilde{a}})$$

and the result follows.

3. (a)

$$B_{ab} = \sum_{j=1}^{m} (\bar{x}_{.ja} - \bar{x}_{..a})(\bar{x}_{.jb} - \bar{x}_{..b})$$

When $m = 2$, note that

$$\bar{x}_{..a} = \frac{n_1 \bar{x}_{.1a} + n_2 \bar{x}_{.2a}}{n}$$

so that, using $n_1 + n_2 = n$

$$\begin{aligned} B_{ab} &= n_1(\bar{x}_{.1a} - \bar{x}_{..a})(\bar{x}_{.1b} - \bar{x}_{..b}) + n_2(\bar{x}_{.2a} - \bar{x}_{..a})(\bar{x}_{.2b} - \bar{x}_{..b}) \\ &= n_1((1 - \frac{n_1}{n})\bar{x}_{.1a} - \frac{n_2}{n}\bar{x}_{.2a})((1 - \frac{n_1}{n})\bar{x}_{.1b} - \frac{n_2}{n}\bar{x}_{.2b}) \\ &\quad + n_2((1 - \frac{n_2}{n})\bar{x}_{.2a} - \frac{n_1}{n}\bar{x}_{.1a})((1 - \frac{n_2}{n})\bar{x}_{.2b} - \frac{n_1}{n}\bar{x}_{.1b}) \\ &= \frac{n_1 n_2^2 + n_2 n_1^2}{n^2}(\bar{x}_{.1a} - \bar{x}_{.2a})(\bar{x}_{.1b} - \bar{x}_{.2b}) \\ &= \frac{n_1 n_2}{n}(\underline{d}\underline{d}^t)_{ab} \end{aligned}$$

as required.

(b) Let $\underline{a}$ be an eigenvector of $W^{-1}B$ with $\underline{d}^t\underline{a} \neq 0$. Then

$$\lambda \underline{a} = W^{-1}B\underline{a} = \frac{n_1 n_2}{n} W^{-1}\underline{d}\underline{d}^t\underline{a}$$

so that

$$W^{-1}\underline{d} = \lambda \frac{n}{n_1 n_2 (\underline{d}^t\underline{a})}\underline{a}$$

as required.

(c) For any eigenvector $\underline{v}$, suppose $\underline{v}^t\underline{d} = 0$, then

$$W^{-1}B\underline{v} = W^{-1}\underline{d}(\underline{d}^t\underline{v}) = 0$$

so that the corresponding eigenvalue is 0.

Suppose that the eigenvalue is non-zero, then $\underline{v}^t\underline{d} \neq 0$ and

$$W^{-1}\underline{d} = \frac{n}{n_1 n_2}\frac{\lambda}{\underline{d}^t\underline{v}}\underline{v}$$

from previous part. This is the only eigenvector with non-zero eigenvalue; the other eigenvectors are orthogonal to this and hence have zero eigenvalue.

CONCLUSION: there is exactly one strictly positive eigenvalue and Fisher's linear discriminant function is given by

$$f(\underline{x}) = \underline{d}^t W^{-1} \underline{x}.$$

4.

$$|C_1| = \frac{1}{4}, \quad |C_2| = 1, \quad C_1^{-1} = \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}, \quad C_2^{-1} = \begin{pmatrix} \frac{1}{4} & 0 \\ 0 & 4 \end{pmatrix}.$$

Log likelihoods are

$$\mathcal{L}_1(x_1, x_2) = -\log(2\pi) - \frac{1}{2}\log\frac{1}{4} - \frac{1}{2}(4x_1^2 + x_2^2)$$

$$\mathcal{L}_2(x_1, x_2) = -\log(2\pi) - \frac{1}{2}(\frac{1}{4}(x_1 - 1)^2 + 4x_2^2)$$

For an observation (x_1, x_2) classify it according to the population for which the log likelihood is the largest.

(a)

$$\mathcal{L}_1(x_1, x_2) - \mathcal{L}_2(x_1, x_2) = 0$$

gives

$$(\log 2 - \frac{1}{8}) - \frac{15}{8}x_1^2 - \frac{1}{4}x_1 + \frac{3}{2}x_2^2 = 0$$

$$(\log 2 + \frac{2}{15}) - \frac{15}{8}(x_1 + \frac{1}{15})^2 + \frac{3}{2}x_2^2 = 0$$

(b) (draw the hyperbola)

(c)

$$\mathcal{L}_1(\frac{1}{2}, \frac{1}{2}) - \mathcal{L}_2(\frac{1}{2}, \frac{1}{2}) = (\log 2 + \frac{2}{15}) - \frac{17^2}{480} + \frac{3}{8} \simeq 0.5994 > 0$$

so assign the observation to population 1.

5. (a) $\underline{\bar{x}}_j$ is the vector with kth component $\bar{x}_{.jk} = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ijk}$.

B is the between population sum of squares matrix. It is a $p \times p$ matrix with elements

$$B_{ab} = \sum_{j=1}^{m} n_j(\bar{x}_{.ja} - \bar{x}_{..a})(\bar{x}_{.jb} - \bar{x}_{..b})$$

W is the within population, or error sum of squares matrix. It is a $p \times p$ matrix with elements

$$W_{ab} = \sum_{j=1}^{m}\sum_{i=1}^{n_j}(x_{ija} - \bar{x}_{.ja})(x_{ijb} - \bar{x}_{.jb})$$

(b) i. See Q3

ii. For a unit vector $\underline{b}$ that maximises the ratio, let $Z = \underline{b}^t \underline{X}$. This linear combination ensures that the largest proportion of the sum of squares is due to the difference between group means, hence giving the strongest clustering around the group means.

(c)

$$L_1(\underline{x}) = -\log 2\pi - \frac{1}{2}\log|C| - \frac{1}{2}(\underline{x} - \underline{\mu}_1)^t C^{-1}(\underline{x} - \underline{\mu}_1)$$

approximated by

$$\hat{L}_1(\underline{x}) = -\log 2\pi - \frac{1}{2}\log|W| - \frac{1}{2}(\underline{x} - \underline{x}_1)^t W^{-1}(\underline{x} - \underline{x}_1)$$

$$L_2(\underline{x}) = -\log 2\pi - \frac{1}{2}\log|C| - \frac{1}{2}(\underline{x} - \underline{\mu}_2)^t C^{-1}(\underline{x} - \underline{\mu}_2)$$

approximated by

$$\hat{L}_2(\underline{x}) = -\log 2\pi - \frac{1}{2}\log|W| - \frac{1}{2}(\underline{x} - \underline{x}_2)^t W^{-1}(\underline{x} - \underline{x}_2)$$

Allocate to group 1 if $\hat{L}_1(\underline{x}) > \hat{L}_2(\underline{x})$. That is

$$-(\underline{x}_1^t W^{-1}\underline{x}_1 - \underline{x}_2^t W^{-1}\underline{x}_2) + 2\underline{x}^t W^{-1}(\underline{x}_1 - \underline{x}_2) > 0$$

i.e. if

$$(2\underline{x} - (\underline{x}_1 + \underline{x}_2))\, W^{-1}(\underline{x}_1 + \underline{x}_2) > 0$$

as required.

6. (Lecture notes)