7
Grouped Multivariate Data: Multivariate Analysis of Variance and Discriminant Function Analysis

7.1 Introduction

Investigators in many disciplines frequently collect multivariate data on samples from different populations. In Chapter 5, for example, a set of data was introduced in which an archaeologist had made four measurements on Egyptian skulls from five different epochs. A variety of questions might be asked about such data and, correspondingly, there are a variety of (overlapping) approaches to their analysis. In many examples the prime interest will be in assessing whether the populations involved have different mean vectors on the measurements taken. For this, multivariate analogues of the familiar univariate t-test, *Hotelling's* T^2, or analysis of variance, *multivariate analysis of variance*, are available. A further question that is often of interest for grouped multivariate data is whether or not it is possible to use the measurements made to construct a classification rule derived from the original observations (the *training set*) that will allow new individuals having the same set of measurements, but no group label, to be allocated to a group in such a way that misclassifications are minimized. The relevant technique is now some form of *discriminant function analysis*.

In the next section we consider both the inference and classification questions for the two-group situation, and then in Section 7.3 move on to discuss data sets where there are more than two groups.

7.2 Two Groups: Hotellings T^2 Test and Fisher's Linear Discriminant Function Analysis

7.2.1 *Hotellings* T^2 *Test*

The data shown in Table 7.1 were originally collected by Colonel L.A. Waddell in southeastern and eastern Tibet. According to Morant (1923), the data consist of two groups of skulls: group one (type I), skulls 1–17, found in graves in Sikkim and neighboring areas of Tibet; group two (type II) consisting of the remaining 15 skulls picked up on battlefield in the Lhasa district and believed to be those of native

Table 7.1 Tibetan Skull Data (all measurements in mm). From Morant, G.M., *A First Study of the Tibetan Skull*, in Biometrika, Vol. 14, 1923, pp 193–260, by permission of the Biometrika Trustees

Obs	Length	Breadth	Height	Fheight	Fbreadth	Type
1	190.5	152.5	145.0	73.5	136.5	1
2	172.5	132.0	125.5	63.0	121.0	1
3	167.0	130.0	125.5	69.5	119.5	1
4	169.5	150.5	133.5	64.5	128.0	1
5	175.0	138.5	126.0	77.5	135.5	1
6	177.5	142.5	142.5	71.5	131.0	1
7	179.5	142.5	127.5	70.5	134.5	1
8	179.5	138.0	133.5	73.5	132.5	1
9	173.5	135.5	130.5	70.0	133.5	1
10	162.5	139.0	131.0	62.0	126.0	1
11	178.5	135.0	136.0	71.0	124.0	1
12	171.5	148.5	132.5	65.0	146.5	1
13	180.5	139.0	132.0	74.5	134.5	1
14	183.0	149.0	121.5	76.5	142.0	1
15	169.5	130.0	131.0	68.0	119.0	1
16	172.0	140.0	136.0	70.5	133.5	1
17	170.0	126.5	134.5	66.0	118.5	1
18	182.5	136.0	138.5	76.0	134.0	2
19	179.5	135.0	128.5	74.0	132.0	2
20	191.0	140.5	140.5	72.5	131.5	2
21	184.5	141.5	134.5	76.5	141.5	2
22	181.0	142.0	132.5	79.0	136.5	2
23	173.5	136.5	126.0	71.5	136.5	2
24	188.5	130.0	143.0	79.5	136.0	2
25	175.0	153.0	130.0	76.0	134.0	2
27	200.0	139.5	143.5	82.5	146.0	2
28	185.0	134.5	140.0	81.5	137.0	2
29	174.5	143.5	132.5	74.0	136.5	2
30	195.5	144.0	138.5	78.5	144.0	2
31	197.0	131.5	135.0	80.5	139.0	2
32	182.5	131.0	135.0	68.5	136.0	2

In dataframe `Tibet`.

soldiers from the eastern province of Khans. These skulls were of particular interest since it was thought at the time that Tibetans from Khans might be survivors of a particular fundamental human type, unrelated to the Mongolian and Indian types that surrounded them.

On each of the 32 skulls the following five measurements, all in millimeters, were recorded:

x_1: greatest length of skull (length),
x_2: greatest horizontal breadth of skull (breadth),
x_3: height of skull (height),
x_4: upper face height (fheight),
x_5: face breadth, between outermost points of cheek bones (fbreadth).

We assume the data are available as the dataframe `Tibet`.

The first task to carry out on these data is to test the hypothesis that the five-dimensional mean vectors of skull measurements are the same in the two populations from which the samples arise. For this we will use the multivariate analogue of Student's independent samples t-test, known as Hotelling's T^2 test, a test described in Display 7.1.

Display 7.1
Hotelling's T^2 Test

- If there are q variables, the null hypothesis is that the means of the variables in the first population equal the means of the variables in the second population.
- If $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are the mean vectors of the two populations the null hypothesis can be written as

$$H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2.$$

- The test statistic T^2 is defined as

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} D^2,$$

where n_1 and n_2 are the sample sizes in each group and D^2 is the generalized distance introduced in Chapter 1, namely

$$D^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2),$$

where $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$ are the two sample mean vectors and $\mathbf{S}$ is the estimate of the assumed common covariance matrix of the two populations, calculated from the two sample covariance matrix, $\mathbf{S}_1$ and $\mathbf{S}_2$ as

$$\mathbf{S} = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{n_1 + n_2 - 2}.$$

- Note that the form of the test statistic in the multivariate case is very similar to that for the univariate independent samples t-test, involving a difference between "means" (here mean vectors), and an assumed common "variance" (here a covariance matrix).

- Under H_0 (and when the assumptions given below hold), the statistic F given by

$$F = \frac{(n_1 + n_2 - q - 1)T^2}{(n_1 + n_2 - 2)q}$$

has a Fisher's F-distribution with q and $n_1 + n_2 - q - 1$ degrees of freedom.
- The T^2 test is based on the following assumptions:

1. In each population the variables have a multivariate normal distribution.
2. The two populations have the same covariance matrix.
3. The observations are independent.

As an exercise we will apply Hotelling's T^2 test to the skull data using the following R and S-PLUS® code, although we could also use the `manova` function as we shall see later:

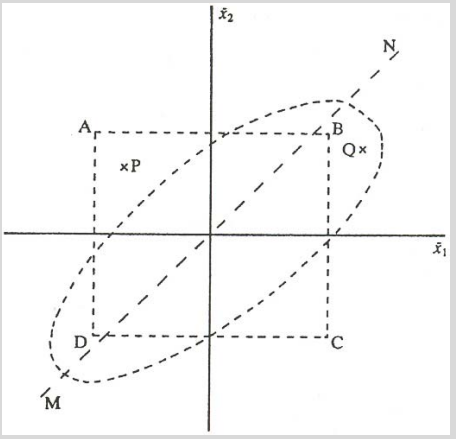
```
attach(Tibet)
m1<-apply(Tibet[Type==1,-6],2,mean)
m2<-apply(Tibet[Type==2,-6],2,mean)
l1<-length(Type[Type==1])
l2<-length(Type[Type==2])
x1<-Tibet[Type==1,-6]
x2<-Tibet[Type==2,-6]
S123<-((l1-1)*var(x1)+(l2-1)*var(x2))/(l1+l2-2)
T2<-t(m1-m2)%*%solve(S123)%*%(m1-m2)
Fstat<-(l1+l2-5-1)*T2/(l1+l2-2)*5
pvalue<-1-pf(Fstat,5,26)
```

Hotelling's T^2 takes the value 3.50 with the corresponding F statistic being 15.17 with 5 and 26 degrees of freedom. The associated p-value is very small, and we can conclude that there is strong evidence that the mean vectors of the two groups differ.

It might be thought that the results of Hotelling's T^2 test would simply reflect those that would be obtained using a series of univariate t-tests, in the sense that if no significant differences are found by the separate t-tests, then the T^2 test will inevitably lead to acceptance of the null hypothesis that the population mean vectors are equal. And, on the other hand, if any significant difference is found when using the t-tests on the individual variables, then the T^2 statistic must also lead to a significant result. But these speculations are not correct (if they were, the T^2 test would be a waste of time). It is entirely possible to find no significant difference for each separate t-test, but a significant result for the T^2 test, and vice versa. An illustration of how this can happen in the case of two variables is shown in Display 7.2.

Display 7.2
Univariate and Multivariate Tests for Equality of Means for Two Variables

- Suppose we have a sample of n observations on two variables, x_1 and x_2 and we wish to test whether the population means of the two variables μ_1 and μ_2 are both zero.
- Assume the mean and standard deviation of the x_1 observations are $\bar{x}_1$ and s_1, respectively, and of the x_2 observations, $\bar{x}_2$ and s_2.
- If we test separately whether each mean takes the value zero, then we would use two t tests. For example, to test $\mu_1 = 0$ against $\mu_1 \neq 0$ the appropriate test statistic is
$$t = \frac{\bar{x}_1 - 0}{s_1\sqrt{n}}.$$
- The hypothesis $\mu_1 = 0$ would be rejected at the α percent level of significance if $t < -t_{100\left(1-\frac{1}{2}\alpha\right)}$ or $t > t_{100\left(1-\frac{1}{2}\alpha\right)}$; that is, if $\bar{x}_1$ fell outside the interval $\left[-s_1 t_{100\left(1-\frac{1}{2}\alpha\right)}\Big/\sqrt{n}\right]$ where $t_{100\left(1-\frac{1}{2}\alpha\right)}$ is the $100\left(1-\frac{1}{2}\alpha\right)$ percent point of the t distribution with $n-1$ degrees of freedom. Thus the hypothesis would *not* be rejected if $\bar{x}_1$ fell *within* this interval.
- Similarly, the hypothesis $\mu_2 = 0$ for the variable x_2 would not be rejected if the mean, $\bar{x}_2$, of the x_2 observations fell within a corresponding interval with s_2 substituted for s_1.
- The multivariate hypothesis $[\mu_1, \mu_2] = [0, 0]$ would therefore not be rejected if *both* these conditions were satisfied.
- If we were to plot the point $(\bar{x}_1, \bar{x}_2)$ against rectangular axes, the area within which the point could like and the multivariate hypothesis not rejected is given by the rectangle $ABCD$ of the diagram below, where AB and DC are of length $2s_1 t_{100\left(1-\frac{1}{2}\alpha\right)}\sqrt{n}$ while AD and BC are of length $2s_2 t_{100\left(1-\frac{1}{2}\alpha\right)}\sqrt{n}$.



- Thus, a sample that gave the means $(\bar{x}_1, \bar{x}_2)$ represented by the point P would lead to acceptance of the multivariate hypothesis.
- Suppose, however, that the variables x_1 and x_2 are moderately highly correlated. Then all points (x_1, x_2) and hence $(\bar{x}_1, \bar{x}_2)$ should lie reasonably close to the straight line *MN* through the origin marked on the diagram.
- Hence samples consistent with the multivariate hypothesis should be represented by points $(\bar{x}_1, \bar{x}_2)$ that lie within a region encompassing the line *MN*. When we take account of the nature of the variation of bivariate normal samples that include correlation, this region can be shown to be an ellipse such as that marked on the diagram. The point P is *not* consistent with this region and, in fact, should be *rejected* for this sample.
- Thus, the inference drawn from the two separate univariate tests conflicts with the one drawn from a single multivariate test, and it is the wrong inference.
- A sample giving the $(\bar{x}_1, \bar{x}_2)$ values represented by point Q would give the other type of mistake, where the application of two separate univariate tests leads to the rejection of the null hypothesis, but the correct multivariate inference is that the hypothesis should *not* be rejected. (This explanation is taken with permission from Krzanowski, 1988.)

Having produced evidence that the mean vectors of skull types I and II are not the same, we can move on to the classification aspect of grouped multivariate data.

7.2.2 *Fisher's Linear Discriminant Function*

Suppose a further skull is uncovered whose origin is unknown, that is, we do not know if it is type I or type II. How might we use the original data to construct a classification rule that will allow the new skull to be classified as type I or II based on the same five measurements taken on the skulls in Table 7.1? The answer was provided by Fisher (1936) who approached the problem by seeking a linear function of the observed variables that provides maximal separation, in a particular sense, between the two groups. Details of Fisher's *linear discriminant function* are given in Display 7.3.

Display 7.3
Fisher's Linear Discriminant Function

- The aim is to find a way of classifying observations into one of two known groups using a set of variables, $x_1, x_2, \ldots, x_q$:
- Fisher's idea was to find a linear function z of the variables. $x_1, x_2, \ldots, x_q$;

$$z = a_1x_1 + a_2x_2 + \cdots + a_qx_q,$$

such that the ratio of the between-group variance of z to its within-group variance is maximized.

- The coefficients $\mathbf{a}' = [a_1, \ldots, a_q]$ have therefore to be chosen so that V, given by

$$V = \frac{\mathbf{a}'\mathbf{B}\mathbf{a}}{\mathbf{a}'\mathbf{S}\mathbf{a}},$$

is maximized, where $\mathbf{S}$ is the pooled within-group covariance matrix, and $\mathbf{B}$ is the covariance matrix of group means; explicitly,

$$\mathbf{S} = \frac{1}{n-2}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_j)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_j)',$$

$$\mathbf{B} = \sum_{i=1}^{2} n_i(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})',$$

where $\mathbf{x}'_{ij} = [x_{ij1}, x_{ij2}, \ldots, x_{ijq}]$ represents the set of q variable values for the jth individual in group i, $\bar{\mathbf{x}}_j$ is the mean vector of the jth group, and $\bar{\mathbf{x}}$ is the mean vector of all observations. The number of observations in each group is n_1 and n_2, with $n = n_1 + n_2$.

- The vector $\mathbf{a}$ that maximizes V is given by the solution of

$$(\mathbf{B} - \lambda\mathbf{S})\mathbf{a} = 0.$$

- In the two-group situation, the *single* solution can be shown to be

$$\mathbf{a} = \mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2).$$

- The allocation rule is now to allocate an individual with discriminate score z to group 1 if $z > (\bar{z}_1 + \bar{z}_2)/2$, where $\bar{z}_1$ and $\bar{z}_2$ are the mean discriminant scores in each group. (We are assuming that the groups are labelled such that $\bar{z}_1 > \bar{z}_2$.)
- Fisher's discriminant function also arises from assuming that the observations in group one have a multivariate normal distribution with mean vector $\boldsymbol{\mu}_1$ and covariance matrix $\boldsymbol{\Sigma}$ and those in group two have a multivariate distribution with mean vector $\boldsymbol{\mu}_2$ and, again, covariance matrix $\boldsymbol{\Sigma}$, and assuming that an individual with vector of scores $\mathbf{x}$ is allocated to group one if

$$MVN(\mathbf{x}, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}) > MVN(\mathbf{x}, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}),$$

where *MVN* is shorthand for the multivariate normal density function.
- Substituting sample values for population rules leads to the same allocation rule as that given above.
- The above is only valid if the prior probabilities of being in each group are assumed to be the same.
- When the prior probabilities are not equal the classification rule changes; for details, see Everitt and Dunn (2001).

The description given in Display 7.3 can be translated into R and S-PLUS code as follows:

```
m1<-apply(Tibet[Type==1,-6],2,mean)
m2<-apply(Tibet[Type==2,-6],2,mean)
l1<-length(Type[Type==1])
l2<-length(Type[Type==2])
x1<-Tibet[Type==1,-6]
x2<-Tibet[Type==2,-6]
S123<-((l1-1)*var(x1)+(l2-1)*var(x2))/(l1+l2-2)
a<-solve(S123)%*%(m1-m2)
z12<-(m1%*%a+m2%*%a)/2
```

This leads to the vector of discriminant function coefficients (**a** in Display 7.3) being

$$\mathbf{a}' = [-0.0893, 0.156, 0.005, -0.177, -0.177]$$

and the threshold value being -30.363. The resulting classification rule becomes: classify to type I if $-0.0893 \times$ length $+ 0.15 \times$ Breadth $+ 0.005 \times$ fheight $-0.177 \times$ fbreadth > -30.363 and Type II otherwise. The same results can be obtained using the `discrim` function in S-PLUS:

```
dis<-discrim(Type~Length+Breadth+Height+Fheight+Fbreadth,
    data=Tibet,
family=Classical("homoscedastic"),prior="uniform")
dis
```

This gives the results shown in Table 7.2. The results given previously are found from Table 7.2 by simply subtracting the two sets of linear coefficients to give the vector of discriminant function coefficients and the two constants to give the threshold value:

```
const<-coef(dis)$constants
const[2]-const[1]
coefs<-coef(dis)$linear.coefficients
coefs[,1]-coefs[,2]
```

By loading the MASS library in both R and S-PLUS, Fisher's linear discriminant analysis can be applied using the `lda` function as

```
library(MASS)
dis<-lda(Type~Length+Breadth+Height+Fheight+Fbreadth,
    data=Tibet,prior=c(0.5,0.5))
```

Suppose now we have the following observations on two new skulls:

	Length	Breadth	Height	Fheight	Fbreadth
Skull 1:	171.0	140.5	127.0	69.5	137.0
Skull 2:	179.0	132.0	140.0	72.0	138.5

Table 7.2 Results from `discrim` on Tibetan Skull Data

```
Group means:
      Length     Breadth     Height     Fheight     Fbreadth      N      Priors
1     174.82      139.35      132.00      69.824       130.35      17       0.5
2     185.73      138.73      134.77      76.467       137.50      15       0.5

Covariance Structure: homoscedastic
             Length      Breadth      Height      Fheight      Fbreadth
  Length     59.013        9.008      17.219       20.120        20.110
 Breadth                  48.261       1.077        4.339        30.046
  Height                               36.198        4.838         4.108
 Fheight                                           18.307        12.985
Fbreadth                                                         43.696

Constants:
     1           2
−514.9      −545.48

Linear Coefficients:
                 X1          X2
  Length     1.4683      1.5576
 Breadth     2.3611      2.2053
  Height     2.7522      2.7470
 Fheight     0.7753      0.9525
Fbreadth     0.1948      0.3722
```

and wish to classify them to be type I or type II. We can calculate each skull's discriminant score as follows;

Skull 1: $-0.0893 \times 171.0 + 0.156 \times 140.5 + 0.005 \times 127.0 - 0.177 \times 69.5 - 0.177 \times 137.0 = -29.27$

Skull 2: $-0.893 \times 179.0 + 0.156 \times 132.0 + 0.005 \times 140.0 - 0.177 \times 72.0 - 0.177 \times 138.5 = -31.95$

Comparing each score to the threshold value of -30.363 leads to classifying skull 1 as type I and skull 2 as type II. We can use the `predict` function applied to the object `dis` to do the same thing:

```
newdata<-rbind(c(171,140.5,127.0,69.5,137.0),c(179.0,132.0,
   140.0,72.0,138.5))
dimnames(newdata)<-list(NULL,c("Length","Breadth","Height",
   "Fheight","Fbreadth"))
newdata<-data.frame(newdata)
predict(dis,newdata=newdata)
```

to give the following classification probabilities:

	Skull 1	Skull 2
Prob(Type I):	0.77695	0.22305
Prob(Type II):	0.19284	0.80716

Fisher's linear discriminant function is optimal when the data arise from populations having multivariate normal distributions with the same covariance matrices.

When the distributions are clearly non-normal an alternative approach is *logistic discrimination* (see, e.g., Anderson, 1972), although the results of both this and Fisher's method are likely to be very similar in most cases. When the two covariance matrices are thought to be unequal, then the linear discriminant function is no longer optimal and a quadratic version may be needed. Details are given in Everitt and Dunn (2001).

The quadratic discriminant function has the advantage of increased flexibility compared to the linear version. There is, however, a penalty involved in the form of potential overfitting, making the derived function poor at classifying new observations. Friedman (1989) attempts to find a compromise between the data variability of quadratic discrimination and the possible bias of linear discrimination by adopting a weighted sum of the two called *regularized discriminant analysis*.

7.2.3 *Assessing the Performance of a Discriminant Function*

How might we evaluate the performance of a discriminant function? One obvious approach would be to apply the function to the data from which it was derived and calculate the misclassification rate (this approach is known as the "plug-in" estimate). We can do this for the Tibetan skull data in R and S-PLUS by again using the `predict` function as follows:

```
group<-predict(dis,method="plug-in")$class
#in S-PLUS use predict(dis,method="plug-in")$group
table(group,Type)
```

leading to the following counts of correct and incorrect classifications:

	Correct group	
Allocated	1	2
1	14	3
2	3	12

The misclassification rate is 19%. This technique has the advantage of being extremely simple. Unfortunately, however, it generally provides a very poor estimate of the actual misclassification rate. In most cases the estimate obtained in this way will be highly optimistic. An improved estimate of the misclassification rate of a discriminant function may be obtained in a variety of ways (see Hand, 1998, for details). The most commonly used of the alternatives available is the so-called "leaving-one-out method," in which the discriminant function is derived from just $n - 1$ members of the sample and then used to classify the member not included. The process is carried out n times, leaving out each sample member in turn. We will illustrate the use of this approach later in the chapter.

7.3 More Than Two Groups: Multivariate Analysis of Variance (MANOVA) and Classification Functions

7.3.1 *Multivariate Analysis of Variance*

MANOVA is an extension of univariate analysis of variance procedures to multidimensional observations. Details of the technique for a one-way design are given in Display 7.3.

Display 7.3
Multivariate Analysis of Variance

- We assume we have multivariate observations for a number of individuals from m different populations where $m \geqslant 2$ and there are n_i observations from population i.
- The linear model for observation x_{ijk}, the jth observation on variable k in group i, $k = 1, \ldots, q$, $j = 1, \ldots, n_i$, $i = 1, \ldots, m$ is

$$x_{ijk} = \mu_k + \alpha_{ik} + \varepsilon_{ijk},$$

  where μ_k is a general effect for the kth variable, α_{ik} is the effect of group i on the kth variable, and ε_{ijk} is a random disturbance term.
- The vector $\varepsilon_{ij} = [\varepsilon_{ij1}, \ldots, \varepsilon_{ijq}]$ is assumed to have a multivariate normal distribution with null mean vector and covariance matrix, $\boldsymbol{\Sigma}$, assumed to be the same in all m populations. The ε_{ij} of different individuals are assumed to be independent of one another.
- The hypothesis of equal mean vectors in the m populations can be written as

$$H_0: \alpha_{ik} = 0, \quad i = 1, \ldots, m, \quad k = 1, \ldots, q.$$

  The multivariate analysis of variance is based on two matrices, **H** and **E**, the elements of which are defined as follows:

$$h_{rs} = \sum_{i=1}^{k} n_i (\bar{x}_{ir} - \bar{x}_r)(\bar{x}_{is} - \bar{x}_s), \quad r, s = 1, \ldots, q,$$

$$e_{rs} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{x}_{ijr} - \bar{x}_{ir})(\bar{x}_{ijs} - \bar{x}_{is}), \quad r, s = 1, \ldots, q,$$

  where $\bar{x}_{ir}$ is the mean of variable r in group i, and $\bar{x}_r$ is the grand mean of variable r.
- The diagonal elements of **H** and **E** are, respectively, the between-groups sum of squares for each variable, and the within-group sum of squares for the variable.

- The off-diagonal elements of **H** and **E** are the corresponding sums of cross-products for pairs of variables.
- In the multivariate situation when $m > 2$ there is no single test statistic that is always the most powerful for detecting all types of departures from the null hypothesis of the mean vectors of the populations.
- A number of different test statistics have been proposed that may lead to different conclusions when used in the same data set, although on most occasions they will not.
- The following are the principal test statistics for the multivariate analysis of variance

  (a) *Wilks' determinantal ratio*

  $$\Lambda = \frac{|\mathbf{E}|}{|\mathbf{H} + \mathbf{E}|}$$

  (b) *Roy's greatest root*
  Here the criterion is the largest eigenvalue of $\mathbf{E}^{-1}\mathbf{H}$

  (c) *Lawley–Hotelling trace*

  $$t = \text{trace}(\mathbf{E}^{-1}\mathbf{H}).$$

  (d) *Pillai trace*

  $$v = \text{trace}[\mathbf{H}(\mathbf{H} + \mathbf{E})^{-1}].$$

- Each test statistic can be converted into an approximate F-statistic that allows associated p-values to be calculated. For details see Tabachnick and Fidell (2000).
- When there are only two groups all four test criteria above are equivalent and lead to the same F value as Hotelling's T^2 as given in Display 7.1.

We will illustrate the application of MANOVA using the data on skull measurements in different epochs met in Chapter 5 (see Table 5.8). We can apply a one-way MANOVA to these data and get values for each of the four test statistics described in Display 7.3 using the following R and S-PLUS code:

R

```
attach(skulls)
skulls.manova < −manova(cbind(MB, BH, BL, NH)∼EPOCH)
summary(skulls.manova, test = "Pillai")
summary(skulls.manova, test = "Wilks")
summary(skulls.manova, test = "Hotelling")
summary(skulls.manova, test = "Roy")
```

S-PLUS

```
attach(skulls)
skulls.manova < −manova(cbind(MB, BH, BL, NH)∼EPOCH)
summary(skulls.manova, test = "pillai")
summary(skulls.manova, test = "wilks")
summary(skulls.manova, test = "hotelling − lawley")
summary(skulls.manova, test = "roylargest")
```

The results are shown in Table 7.3. There is very strong evidence that the mean vectors of the five epochs differ.

The tests applied in MANOVA assume multivariate normality for the error terms in the corresponding model (see Display 7.3). An informal assessment of this assumption can be made using the chi-square plot described in Chapter 1 applied to the residuals from fitting the one-way MANOVA model; note that the residuals in this case are each four-dimensional vectors. The required R and S-PLUS instruction is

```
chisplot(residuals(skulls.manova))
```

This gives the diagram shown in Figure 7.1. There is no evidence of a departure from multivariate normality.

7.3.2 *Classification Functions and Canonical Variates*

When there is an interest in classification of multivariate observations where there are more than two groups, a series of classification functions can be derived based on the two-group approach described in Section 7.2. Details are given in Display 7.4 for the three group situation.

Table 7.3 Multivariate Analysis of Variance Results for Egyptian Skull Data

```
    Df Pillai Trace approx. F num df den df P-value
  EPOCH 4         0.35      3.51      16       580        0
Residuals 145
    Df Wilks Lambda approx. F num df den df P-value
  EPOCH 4         0.66       3.9      16     434.45       0
Residuals 145
    Df Hotelling-Lawley approx. F num df den df P-value
  EPOCH 4         0.48      4.23      16       562        0
Residuals 145
    Df Roy Largest approx. F num df den df P-value
  EPOCH 4         0.43     15.41       4       145        0
Residuals 145
```

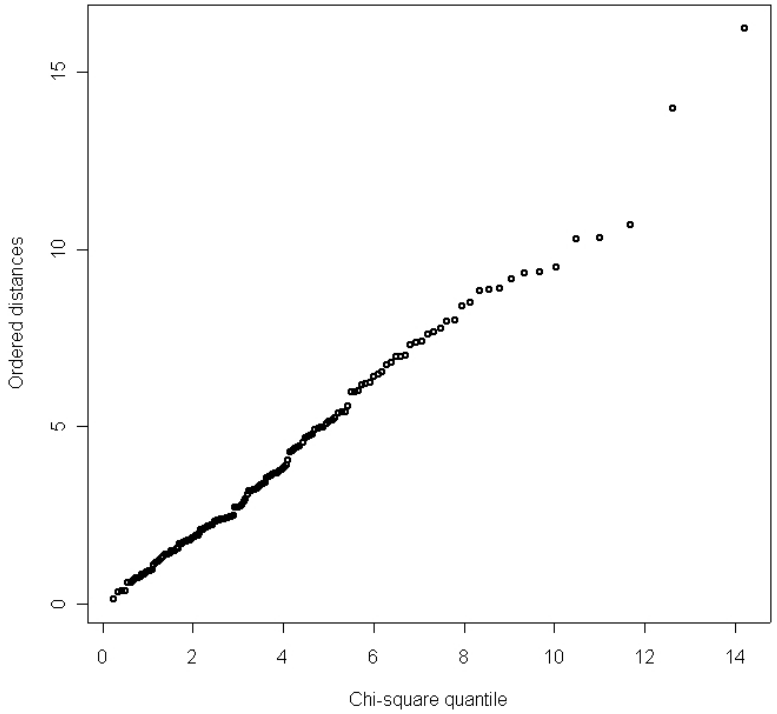


Figure 7.1 Chi-square plot of residuals from fitting one-way MANOVA model to Egyptian skull data.

Display 7.4
Classification Functions for Three Groups

- When more than two groups are involved, the rule for allocating to two multivariate normal distributions with the same covariance matrix can be applied to each pair of groups in turn to derive a series of classification functions.
- For three groups, for example, the sample versions of the functions would be:

$$h_{12}(\mathbf{x}) = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}\left[\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)\right],$$

$$h_{13}(\mathbf{x}) = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_3)'\mathbf{S}^{-1}\left[\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_3)\right],$$

$$h_{23}(\mathbf{x}) = (\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_3)'\mathbf{S}^{-1}\left[\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_2 + \bar{\mathbf{x}}_3)\right],$$

where $\mathbf{S}$ is the pooled within-groups covariance matrix calculated over *all* three groups.

- The classification rule now becomes:

  Allocate to G_1 if $h_{12}(x) > 0$ and $h_{13}(x) > 0$;
  Allocate to G_2 if $h_{12}(x) < 0$ and $h_{23}(x) > 0$;
  Allocate to G_3 if $h_{13}(x) < 0$ and $h_{23}(x) < 0$.

The classification functions allow observations to be classified optimally, but interest may also lie in identifying the dimensions of the multivariate space of the observed variables that are of most importance in distinguishing between the groups. For two groups the single dimension is given by Fisher's linear discriminant function, which, as described in Display 7.3, arises as the single solution of the equation

$$(\mathbf{B} - \lambda\mathbf{S})\mathbf{a} = 0.$$

When there are more than two groups however, this equation will have more then one solution, reflecting the fact that more than one direction is needed to describe the differences between the mean vectors of the groups. With g groups and q variables there will be $\min(q, g-1)$ solutions. These best separating dimensions are known as *canonical variates*. We can find them and the relevant classification for the Egyptian skull data by again using the `discrim` function (or alternatively the `lda` function although the options and output are not quite so comprehensive):

```
dis<-discrim(EPOCH~MB+BH+BL+NH,data=skulls,family=Canonical
    ("homoscedastic"))
dis
summary(dis)
```

An edited version of the results is shown in Table 7.4. To form the classification functions described in Display 7.4 we need to look at the "linear coefficients" and the "constants" in this table. For example, the classification function for epochs, c4000BC, and c3300BC, can be found as

```
const <- coef(dis)$constants
t12<-const[2] - const[1]
coefs <- coef(dis)$linear.coefficients
h12<-coefs[, 1] - coefs[, 2]
```

This gives the necessary vector of constants and threshold to form the first of the required classification functions; similarly, the remaining classification functions, `h13,h14,...,h45`, and thresholds, `t12,t13,...,t45`, can be found. They may then be applied to the four measurements on a new skull as indicated by the rule in Display 7.4 extended in an obvious way to the five-group situation, to classify the skull into one of the five epochs (see Exercise 7.4).

Table 7.4 Edited Results from Applying the `discrim` Function to the Egyptian Skull Data

```
Group means:
             MB       BH       BL       NH   N  Priors
c4000BC  131.37   133.60   99.167   50.533  30     0.2
c3300BC  132.37   132.70   99.067   50.233  30     0.2
c1850BC  134.47   133.80   96.033   50.567  30     0.2
 c200BC  135.50   132.30   94.533   51.967  30     0.2
 cAD150  136.17   130.33   93.500   51.367  30     0.2

Covariance Structure: homoscedastic
        MB       BH       BL       NH
MB  21.111    0.037    0.079    2.009
BH           23.485    5.200    2.845
BL                     24.179    1.133
NH                               10.153

Canonical Coefficients:
          dim1        dim2        dim3          dim4
MB    0.126676    0.038738  −0.092768  −0.14883986
BH  −0.037032    0.210098    0.024568   0.00042008
BL  −0.145125  −0.068114  −0.014749  −0.13250077
NH    0.082851  −0.077293    0.294589  −0.06685888

Singular Values:
  dim1    dim2     dim3     dim4
3.9255  1.189  0.75451  0.27062

Constants:
c4000BC   c3300BC   c1850BC    c200BC    cAD150
−914.53   −915.35   −925.34   −923.42   −914.12

Linear Coefficients:
    c4000BC  c3300BC  c1850BC  c200BC  cAD150
MB   6.0012   6.0515   6.1507  6.1850  6.2209
BH   4.7674   4.7316   4.8090  4.7380  4.6650
BL   2.9569   2.9617   2.8189  2.7647  2.7395
NH   2.1238   2.0938   2.1013  2.2583  2.2154

Plug-in classification table:
         c4000BC  c3300BC  c1850BC  c200BC  cAD150    Error  Posterior.Error
c4000BC       12        8        4       4       2  0.60000          0.60390
c3300BC       10        8        5       4       3  0.73333          0.68144
c1850BC        4        4       15       2       5  0.50000          0.62486
 c200BC        3        3        7       5      12  0.83333          0.73753
 cAD150        2        4        4       9      11  0.63333          0.51736
Overall                                             0.66000          0.63302
(from=rows,to=columns)

Rule Mean Square Error: 0.71928
(conditioned on the training data)

Cross-validation table:
       c4000BC  c3300BC  c1850BC  c200BC  cAD150    Error  Posterior.Error
c4000BC       9       10        5       4       2  0.70000          0.61844
c3300BC      11        7        5       4       3  0.76667          0.65822
c1850BC       6        4       12       2       6  0.60000          0.64549
 c200BC       3        3        7       5      12  0.83333          0.72218
 cAD150       2        4        4      10      10  0.66667          0.52258
Overall                                            0.71333          0.63338
(from=rows,to=columns)
```

The coefficients defining the canonical variates are to found under "canonical coefficients" in Table 7.4. Here, with five groups and four variables, there are four such variates. To see how the canonical variates discriminate between the groups, it is often useful to plot the group canonical variate means. For example, we can plot the means for the first two canonical variates as

```
dsfs1<-c(0.13,-0.04,-0.15,0.08)%*%t(skulls[,-1])
dsfs2<-c(0.04,0.21,-0.068,-0.08)%*%t(skulls[,-1])
m1<-
   c(mean(dsfs1[1:30]),mean(dsfs1[31:60]),mean(dsfs1[61:90]),
   mean(dsfs1[91:120]),mean(dsfs1[121:150]))
m2<-
   c(mean(dsfs2[1:30]),mean(dsfs2[31:60]),mean(dsfs2[61:90]),
   mean(dsfs2[91:120]),mean(dsfs2[121:150]))
plot(m1,m2,type="n",xlab="CV1",ylab="CV2",xlim=c(0.5,3))
text(m1,m2,labels=c("c4000BC","c3300BC","c1850BC","c200BC",
   "cAD150"))
```

The result is shown in Figure 7.2. The first canonical variate separates the two earliest epochs from the other three and the second separates c1850BC from the remaining four.

The "plug-in" estimate of the misclassification rate is shown in Table 7.4. Also shown is the more realistic "leave-out-one" or "cross-validation" estimate. There is a considerable amount of misclassification particularly for c200BC and cAD150.

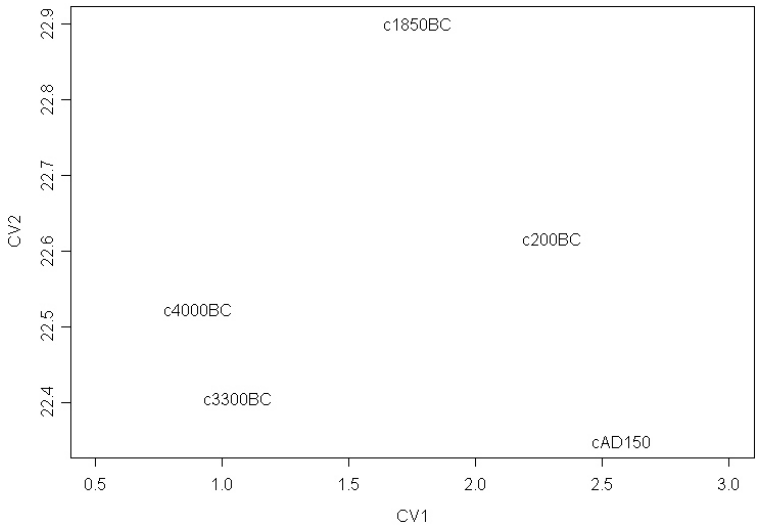


Figure 7.2 Epoch means for the first two canonical variates.

Table 7.5 SIDS data

Group	HR	BW	Factor68	Gesage
1	115.6	3060	0.291	39
1	108.2	3570	0.277	40
1	114.2	3950	0.390	41
1	118.8	3480	0.339	40
1	76.9	3370	0.248	39
1	132.6	3260	0.342	40
1	107.7	4420	0.310	42
1	118.2	3560	0.220	40
1	126.6	3290	0.233	38
1	138.0	3010	0.309	40
1	127.0	3180	0.355	40
1	127.7	3950	0.309	40
1	106.8	3400	0.250	40
1	142.1	2410	0.368	38
1	91.5	2890	0.223	42
1	151.1	4030	0.364	40
1	127.1	3770	0.335	42
1	134.3	2680	0.356	40
1	114.9	3370	0.374	41
1	118.1	3370	0.152	40
1	122.0	3270	0.356	40
1	167.0	3520	0.394	41
1	107.9	3340	0.250	41
1	134.6	3940	0.422	41
1	137.7	3350	0.409	40
1	112.8	3350	0.241	39
1	131.3	3000	0.312	40
1	132.7	3960	0.196	40
1	148.1	3490	0.266	40
1	118.9	2640	0.310	39
1	133.7	3630	0.351	40
1	141.0	2680	0.420	38
1	134.1	3580	0.366	40
1	135.5	3800	0.503	39
1	148.6	3350	0.272	40
1	147.9	3030	0.291	40

(*Continued*)

Table 7.6 (*Continued*)

Group	HR	BW	Factor68	Gesage
1	162.0	3940	0.308	42
1	146.8	4080	0.235	40
1	131.7	3520	0.287	40
1	149.0	3630	0.456	40
1	114.1	3290	0.284	40
1	129.2	3180	0.239	40
1	144.2	3580	0.191	40
1	148.1	3060	0.334	40
1	108.2	3000	0.321	37
1	131.1	4310	0.450	40
1	129.7	3975	0.244	40
1	142.0	3000	0.173	40
1	145.5	3940	0.304	41
2	139.7	3740	0.409	40
2	121.3	3005	0.626	38
2	131.4	4790	0.383	40
2	152.8	1890	0.432	38
2	125.6	2920	0.347	40
2	139.5	2810	0.493	39
2	117.2	3490	0.521	38
2	131.5	3030	0.343	37
2	137.3	2000	0.359	41
2	140.9	3770	0.349	40
2	139.5	2350	0.279	40
2	128.4	2780	0.409	39
2	154.2	2980	0.388	40
2	140.7	2120	0.372	38
2	105.5	2700	0.314	39
2	121.7	3060	0.405	41

7.4 Summary

Grouped multivariate data occur frequently in practice. The appropriate method of analysis depends on the question of most interest to the investigator. Hotelling's T^2 and MANOVA are used to assess formal hypothesis about population mean vectors. Where there is evidence of a difference then the construction of a classification rule

is often (but not always) of interest. A range of other discriminant procedures are available in the MASS library, and readers are encouraged to investigate.

Exercises

7.1 In a two-group discriminant situation, if members of one group have a y-value of -1 and those of the other group a value of 1, show that the coefficients in a regression of y on $x_1, x_2, \ldots, x_q$ are proportional to $\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$, the coefficients of Fisher's linear discriminant function.

7.2 In the two-group discrimination problem, suppose that

$$f_i(x) = \binom{n}{x} p_i^x (1 - p_i)^{n-x}, \qquad 0 < p_i < 1, \quad i = 1, 2,$$

where p_1 and p_2 are known. If π_1 and π_2 are the prior probabilities of the two groups, devise the classification rule using the approach described in Display 7.3.

7.3 The data shown in Table 7.5 were collected by Spicer et al. (1987) in an investigation of sudden infant death syndrome (SIDS). The two groups here consist of 16 SIDS victims and 49 controls. The Factor68 variable arises from spectral analysis of 24 hour recordings of electrocardiograms and respiratory movements made on each child. All the infants have a gestational age of 37 weeks or more and were regarded as full term.

(i) Construct Fisher's linear discriminant function using only the Factor68 and Birthweight variables. Show the derived discriminant function on a scatterplot of the data.

(ii) Construct the discriminant function based on all four variables and find an appropriate estimate of the misclassification rate.

(iii) How would you incorporate prior probabilities into your discriminant function?

7.4 Find all the classification functions for the Egyptian skull data and use them to allocate a new skull with the following measurements:

MB:	133.0
BH:	130.0
BL:	95.0
NH:	50.0