14 Introduction to discriminant analysis

14.1 Introduction

In Chapter 12, principal component analysis (PCA) was introduced, which can be applied when you have M observations on N variables, denoted by Y_1 to Y_N. Recall that the aim of PCA is to create linear combinations of the N variables (principal components or axes), such that the first principal component (PC) has maximum variance, the second PC, the second largest variance, etc. The first PC, denoted by Z_1, is given by

$$Z_{i1} = c_{11}Y_{i1} + c_{12}\,Y_{i2} + \ldots + c_{1N}\,Y_{iN} \tag{14.1}$$

The index i refers to the observations, and all we need to find are the multiplication factors c_{ij}. However, if we already know that there is an *a priori* grouping structure in the M observations, then we can use discriminant analysis (DA) and take advantage of this additional information. As an example we will use a sparrow dataset (unpublished data, Chis Elphick, University of Connecticut). This dataset has seven body measurements from approximately 1100 saltmarsh sharp-tailed sparrows (*Ammodramus caudacutus*), e.g., size of the head, size of the wings, tarsus length, weight, etc. Over the course of the study, 10 different observers were involved in taking measurements. Some measured over 300 birds, whereas others measured much less. This gave a dataset with $M = 1100$ rows of measurements for $N = 7$ variables, plus an extra variable identifying the observer. We could apply a PCA on these data and use the identity of the observers as labels in the biplot, but this does not take advantage of the extra grouping information given by knowing who was the observer. Nor does it help answer the possible underlying question on whether the observers have influenced the results. PCA cannot answer this question, but discriminant analysis (DA) can.

Before applying DA on these data, we will look at some examples from the case study chapters. In Chapter 28, a zoobenthic dataset from an Argentinean salt marsh area is used. There are four species (variables), and measurements were taken across the seasons from three transects with 10 samples per transect (per season). The data matrix (per season) for the species data gives a 30-by-4 matrix plus an extra column with value 1, 2 or 3 to identify the transect where the sample was collected. And we can make the situation more complicated by combining the

data from two seasons resulting in a 60-by-4 data matrix. Assuming we are interested in how the relationships between the four species differ among the transect, we can use DA to investigate these relationships by discriminating between season, between transect, or between season and transect to see which are the most important in understanding the species relationships.

In Chapter 29, fatty acid concentrations in blubber of stranded dolphins are analysed. There are 31 fatty acids (variables), and 89 dolphins (observations) were measured. Although the chapter uses PCA to analyse these data (focussing on the relationship between different fatty acids), you can also ask the question whether the fatty acid values can be used to discriminate between male and female species, type of death and area of stranding. The data matrix is of dimension 89-by-31 and has three extra columns identifying sex, type of death and area. We could apply three different discriminant analyses: one for each question.

In Chapter 24, classification trees are applied on bird observations obtained by radar. The radar measures a large number of variables per bird, for example velocity, size of the target, etc. About 650 observations are used in the chapter. As well as the observations from the radar, field observations from the ground were available and these allowed the observations to be grouped by species, clutter, etc. The question is then whether we can only discriminate between birds and clutter, or whether we can also discriminate between species.

The common feature shared by all these datasets is that the dataset is of dimension M-by-N, and there is an extra column identifying groups of observations. The structure of the data is visualised in Table 14.1. In all the datasets the question is: 'Do the variables differ per group of observations, and if they do, which variables?' Stated differently, can we discriminate between *a priori* defined groups of observations using the variables? And which variables are the best at discriminating?

Table 14.1. Structure of the data for discriminant analysis. The variables Y_1 to Y_N are for each observation and 'Group' identifies either the observations made by different observers, male and female (g = 2), different transects, areas, etc.

Observation	Y_1	Y_2	...	Y_N	Group
1	10	16	...	21	1
2	21	18	...	52	1
3	31	41	...	2	1
4	12	15	...	34	1
5	1	10	...	1	2
6	12	20	...	2	2
...	...	...	...	...	...
...	...	...	...	...	...
N	15	21	...	6	G

To answer this question, we could apply one-way ANOVA on each of the N variables Y_j. The explanatory variable would be 'Group' and the response variable Y_j. The problem with this approach is that it is rather time consuming to apply N

one-way ANOVAs, and we are not taking advantage of the multivariate nature of the data. Discriminant analysis (DA), also called canonical variate analysis, although similar to PCA, uses all variables Y_1 to Y_N and extracts a linear combination of them. This linear combination is of the form:

$$Z_{i1} = \text{constant}_1 + w_{11}Y_{i1} + w_{12}Y_{i2} + \ldots + w_{1N}Y_{iN} \quad (14.2)$$

The unknown parameters are the constant and the multiplication factors (or weighting factors) w_{ij}. In PCA, the multiplication factors are chosen such that Z_1 has maximum variance. In DA the aim is different as we are interested in discrimination between the *a priori* defined groups of observations. To illustrate the underlying principle of DA, we use the sparrow data described earlier. Nine observers were involved in the sampling process and the observations consist of seven body measurements on each sparrow. A scatterplot for two variables is given in Figure 14.1. To keep the graph simple, we only used observations made by two observers, and they are identified by the symbols '+' and 'o'. If we apply PCA on these data, the first axis would probably go from somewhere in the lower left corner to the upper right corner as this would give a line that when all points are projected on it, has maximum variance. In DA, the objective is different; we look for a line that, when all points are projected on it, the observations of the same group are close to each other and observations from different groups are far away from each other. The dotted line in Figure 14.1 is a potential candidate. If we project all points on this line, the points with a '+' will be mostly on the left side and the observations with 'o' are mainly right of it.

So, how do we get this line? Just as in PCA it is a matter of finding the optimal rotation. We could define a criteria that measures the discrimination and try every possible rotation, but just as in PCA, it turns out that the solution can be obtained with an eigenvalue equation. Further axes can be calculated, and these are uncorrelated with each other. In this chapter, we will not present too much mathematical detail and refer you to Legendre and Legendre (1998), Huberty (1994) or Klecka (1980). Material presented here is mainly based on these three references.

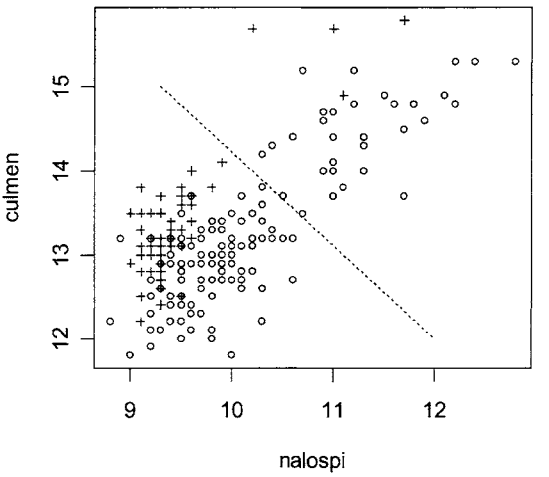

Figure 14.1. Example of the underlying principle of discriminant analysis. Measurements made by two observers for a sparrow dataset are used. The two observers are plotted using 'o' and '+'. The variables nalospi and culmen are variables measured by the observers. The dotted line shows a possible axis that gives maximum separation of the two groups, if all points are projected on it.

14.2 Assumptions

In PCA there were no underlying assumptions except that the relationships between the variables should be linear. In DA, there is a whole series of assumptions, and their validity dictates how useful the method is for your own data. These assumptions are as follows:

- The observations can be divided *a priori* into at least two groups. Each observation can only be in one group. This means that one should not apply clustering on the data to obtain groups, and then apply DA on the same data using the grouping structure obtained by clustering. A valid approach is to split up the data into two groups, apply clustering on one dataset, and use the results in the DA for the second dataset (Legendre and Legendre 1998).
- There are at least two observations per group. But a common recommendation is to have at least four or five times as many observations per group as the number of variables.
- The number of observations in the smallest group is larger than the number of variables. This assumption is stated in most books on DA, but interestingly half of the authors then give examples of where this assumption is violated.

- The variables are on a continuous scale. Because covariance matrices are calculated, you cannot use nominal variables in the dataset. You can convert them to dummy variables with zeros and ones, but this then violates other assumptions.
- Relationships between the variables are linear. This is because covariance matrices are used.
- There is no 100% collinearity between the variables. Always inspect the correlation matrix before applying DA. If variables have a correlation of 0.9 or higher, remove one of them.
- Within group variances are approximately similar between the groups. This is also called the homogeneity assumption. It means that the spread of a particular variable must be (approximately) the same for each group of observations. But different variables are allowed to have a different spreads. Cleveland dotplots or boxplots conditional on the grouping of observations can be used to assess this assumption (Chapter 4).
- During the calculations a covariance matrix for each group will be calculated and these will be pooled. Therefore, we also assume that the covariance matrices for the g groups are similar. This means that relationships between variables must be similar for different groups. It also means that a variable within a group is not allowed to have the same value (e.g., zero) for each observation as this prevents calculating the covariance matrix. Some programmes will still produce sensible results even if this assumption is not met.
- The hypothesis tests assume multivariate normality of the N variables within each group. This means that each variable must be (approximately) normally distributed within each group. Conditional histograms (Chapter 4) can be used to assess this assumption.
- The observations are independent. This means that time series, spatial data and before-after comparisons cannot be used.

If the normality and homogeneity assumptions do not hold, a logarithmic or square root transformation might help. Normality is required for the hypothesis tests, but not for the method itself (Hair et al. 1998). Violation of homogeneity in combination with small group sizes is seen as a serious problem. In such cases (multinomial) logistic regression might be a better alternative; the group variable is used as a response variable and the rest as explanatory variables and no conditions are imposed on the explanatory variables. Quadratic discriminant analysis is an alternative option if there is violation of homogeneity.

Equal group size is not required, but as with all statistical methods, common sense dictates that they should be similar. A ratio of largest group size versus the smallest group size of 9:1 has been suggested by some authors as the maximum value before one should decide not to apply DA.

Due to this long list of assumptions, some authors (e.g., McCune and Grace 2002) have suggested that DA has 'limited application in community ecology'. We do not agree with this, as DA can be used in about one quarter of the case study chapters in this book. Obviously, it is less useful if the dataset consists of a large number of plant species sampled at 200 sites, and 95% of the observations is

equal to zero as we are violating various assumptions. The answer to the question of whether DA is useful for your own data is simple: 'It all depends'. The example presented in the next section should help you decide.

14.3 Example

Recall that the sparrow data consist of approximately 1000 birds measured by 10 observers. The measured variables are the lengths of the wing (measured in two different ways, as the wing chord and as the flattened wing), leg (a standard measure of the tarsus), head (from the bill tip to the back of the skull), culmen (the top of the bill from the tip to where the feathering starts), nalospi (the distance from the bill tip to the nostril) and weight. The question we want to look at is whether the observers are producing similar measurements or, stated slightly differently, is there an observer effect? Discriminant analysis is one of the most appropriate methods to answer this (alternative methods are redundancy analysis and multivariate regression trees), but before we can apply DA, we need to verify the assumptions.

There are 10 observers, and the number of observations per observer was between 9 and 332. This means that we have a serious problem with unequal group sizes. If we use the (arbitrary) 9:1 rule, we need to drop the two observers that only made between 9 and 30 observations, from the analyses. Even without this rule, it is common sense not to compare results of an observer with only 9 observations with one with 332 observations.

The homogeneity assumption was checked using Cleveland dotplots or boxplots conditional on observer, and one such graph for flatwing is shown in Figure 14.2. It shows that the spread is approximately the same in each group, indicating homogeneity. The assumption also holds for the other variables (results are not shown here). To have faith in the hypotheses tests (which will be discussed later), the normality of each variable in each group is required. A conditional histogram for each variable (not shown here) indicates that this is a valid assumption.

Another assumption is that the variables are not collinear. The correlation coefficient between each pair of variables was calculated, and all were smaller than 0.75, except wingcrd and flatwing; their correlation coefficient was 0.99. This high correlation was entirely expected as the two variables represent the same thing. Hence, one of these variables should be dropped and we decided to use flatwing in the analysis.

A pairplot (not shown here) showed that relationships between all variables are approximately linear. The last assumption we need to check is the independence of the observations. Sampling took place in different months, and the variable weight shows a strong seasonal pattern, which means violation of the independence assumption. We de-seasonalised the weight variable by subtracting the monthly average (Chapter 16). The resulting dataset now complies with all assumptions, and we can apply linear discriminant analysis.

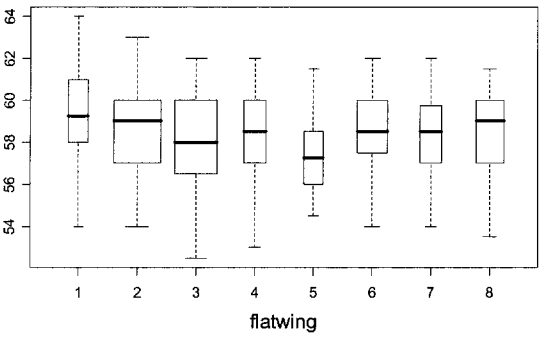

Figure 14.2. Boxplot of flatwing conditional on observer. Each number along the horizontal axis is an observer. The prime question is whether the spread is the same in each group. The width of a box is proportional to the number of observations made by an observer, but all observers (presented here) measured more than 55 birds.

The first discriminant function is given by

$$Z_{i1} = \text{constant}_i - 0.04 \times \text{flatwing}_i + 0.25 \times \text{tarsus}_i - 0.68 \times \text{head}_i - 1.62 \times \text{culmen}_i + 2.91 \times \text{nalospi}_i + 0.05 \times \text{wt}_i \quad (14.3)$$

The multiplication factors 0.04, .., 0.05 are called the unstandardised discrimination coefficients, and their interpretation is difficult as the original variables were not normalised (centred and divided by the standard deviation). If we normalise the variables prior to the analysis, it is easier to compare them with each other and the intercept also vanishes:

$$Z_{i1} = -0.08 \times \text{flatwing}_i + 0.17 \times \text{tarsus}_i - 0.47 \times \text{head}_i - 1.00 \times \text{culmen}_i + 1.55 \times \text{nalospi}_i + 0.07 \times \text{wt}_i \quad (14.4)$$

The multiplication factors obtained for normalised variables are called the standardised discrimination coefficients, and these can be used to assess which variables are important for the discrimination along the first axis. For example, the variable nalospi has a large positive multiplication factor and head and culmen have large negative factors. These three variables play an important role for discrimination along the first axis. All three variables indicate something about the size of the bird's head.

The standardised discrimination coefficients can either be obtained by standardising the variables prior to the analysis or by using a short-cut formula (Klecka 1980). Some authors mention that the (standardised) discrimination coefficients can be instable and instead advise using the correlation coefficients between the discriminant functions and each original variable. These are called canonical correlations, but note that terminology differs between software and authors.

Further axes are obtained, and the traditional graphical presentation of DA is to plot two axes against each other, in most cases Z_1 versus Z_2 as these explain most of the separation; see Figure 14.3. Each observation is plotted as a group number. If there is an observer effect, you would expect to see observations from the same group close to each other with a clear separation between groups. One way to enhance visual detection of group effects is to calculate the average group scores per axis and to plot these for example as large triangles (Figure 14.3). If the triangles are clearly separated, then there is a visual indication of a group effect. In this case, some triangles are separated, but we have seen examples with more separation. Instead of the scatterplot of Z_1 and Z_2, Krzanowski (1988) used tolerance intervals. These are presented in Figure 14.4 and show the group means again, but with these now identified by a number representing the group (observer). The circles around the group means represent the 90% tolerance regions where 90% of the whole population in a group is expected to lie (Krzanowksi 1988, pp. 374-375). These graphs are easier to interpret, as there is less clutter compared with plotting all the scores, but they are based on the normality assumption. The canonical correlation coefficients (these are the correlations between the original variables and the DA axes) are plotted in Figure 14.5 and indicate that the first axis is positively correlated with nalospi.

The DA graphs indicate that there is some marginal evidence of discrimination, possibly related to nalopsi. As this is probably the hardest variable to measure, it does not come as a surprise that it contributes the most to differences among individuals. The question is now whether the discrimination is significant, and this is discussed in Section 14.5. However, the fact that the groups are not clearly separated in the ordination diagrams indicates that even if the statistical tests indicate that there is an observer effect, it is not particularly strong.

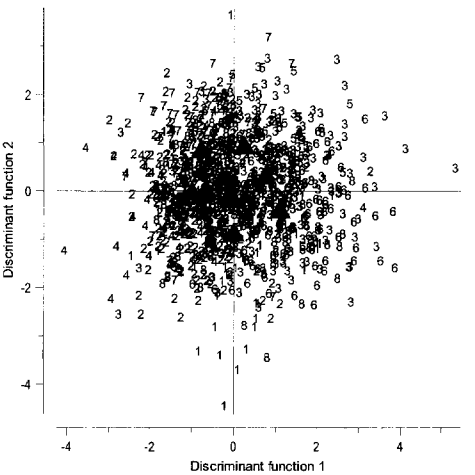

Figure 14.3. Scatterplot of the first discriminant function versus the second. Observations are represented by their group (=observer) number. The triangles represent the group averages.

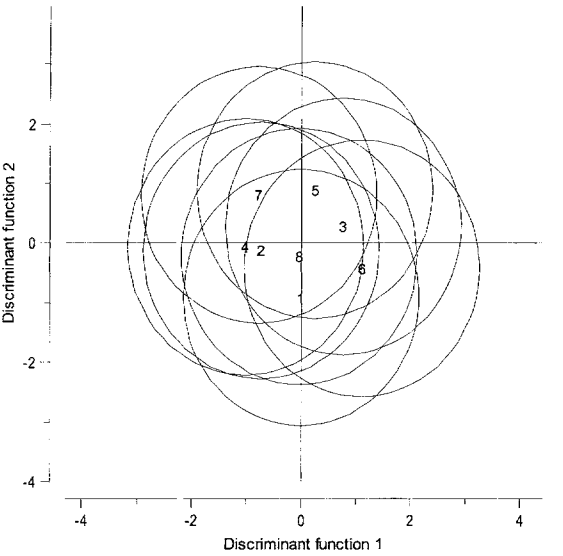

Figure 14.4. 90% tolerance intervals. The numbers refer to the observers, and the circles are the 90% tolerance intervals; they define the range in which 90% of the population values are found.

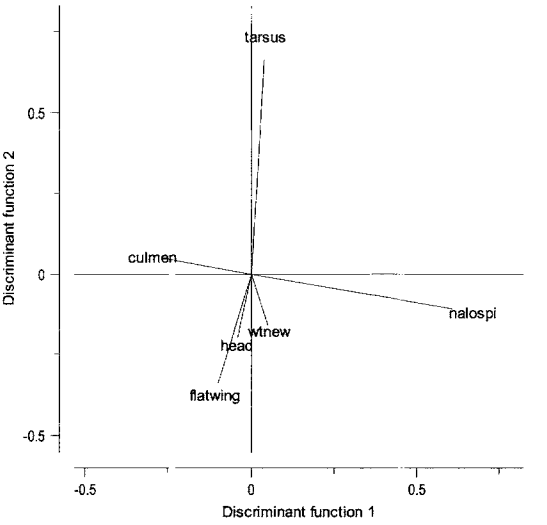

Figure 14.5. Correlation coefficients between the first two axes and each of the original variables.

14.4 The mathematics

The reader not interested in the principles of the underlying mathematics for linear discriminant analysis may skip this section. We follow the same notation and outline as Legendre and Legendre (1998). DA is based on three important matrices, and these are extensions of linear regression where we decomposed the total variation (SS_{total}) in $SS_{regression}$ and $SS_{residual}$, and used these in an F-test (Chapter 5). Here, we use multivariate extensions of these terms.

Define **X** as the matrix containing all the variables. The rows contain the M observations and the columns the N variables giving a matrix of dimension N-by-M. The N-by-N matrix measuring the overall variation in the data is given by

$$\mathbf{T} = (\mathbf{X} - \bar{\mathbf{X}})'(\mathbf{X} - \bar{\mathbf{X}})$$

It can easily be converted to the covariance (or correlation) matrix; simply divide **T** by $N - 1$:

$$\mathbf{S} = (\mathbf{X} - \bar{\mathbf{X}})'(\mathbf{X} - \bar{\mathbf{X}})/(N - 1)$$

The matrix **S** is the covariance (or correlation) matrix, and it represents the total variation in the data. Define $\mathbf{W}_j$ as the sum of squares for group j. This is calculated in the same way as **T**, but we only use data of group j. This matrix can be calculated for each group, giving $\mathbf{W}_1,..,\mathbf{W}_g$. Because we assumed homogeneity, we can pool the g within-group matrices:

$$\mathbf{W} = \mathbf{W}_1 + ... + \mathbf{W}_g$$

The matrix **W** can be converted into a covariance matrix by dividing it by N–g:

$$\mathbf{V} = \mathbf{W}/(N - g)$$

V represents the within group covariance. Having a total variation (**S**), within group variation (**V**), there is one last term we need, namely the between group variation. It is obtained by subtracting **W** from **T**:

$$\mathbf{B} = \mathbf{T} - \mathbf{W}.$$

It can be converted into a covariance matrix by dividing **B** by $g - 1$:

$$\mathbf{A} = \mathbf{B}/(g - 1)$$

The matrices **A** and **V** form the basis for the eigenvalue equation for linear discriminant analysis, or to be more precise: $\mathbf{V}^{-1}\mathbf{A}$. Because this matrix is not symmetrical the following eigenvalue equation is solved (Legendre and Legendre 1998, p. 621):

$$(\mathbf{A} - \lambda_k \mathbf{V})\mathbf{u}_k = 0$$

To make the axes orthogonal, the eigenvectors are rescaled: $\mathbf{C} = \mathbf{U}(\mathbf{U}'\mathbf{V}\mathbf{U})^{-0.5}$, where **U** contains the eigenvectors and **C** the scaled eigenvectors.

14.5 The numerical output for the sparrow data

Discriminant analysis comes with a whole suite of hypotheses, and full details on these tests can be found in Huberty (1994). These tests are based on the normality assumption, but this is a valid assumption for the sparrow data. Instead of giving a detailed explanation for each of these hypotheses and tests, we present the results of some of them within the context of the sparrow data. It is not only the statistical tests that make the numerical output of DA intimidating, there is also a lot of other information presented. We have grouped this information into three types: the importance of the axes, the importance of the variables and finally, classification.

Importance of the axes

It does not really matter which software is used as in essence they all produce similar output (although not necessarily identical). The first part of the output is as follows:

```
The number of variables is                6
The number of observations is             1102
The number of groups is                   8
The number of discriminant functions is          6
The number of rows (observations) containing missing values is    0
```

This information shows that there are 1102 observations and 8 groups (observers). The number of discriminant functions that one can calculate is the minimum of $g - 1$ and N (g is the number of groups and N the number of variables). In this case it is 6. As with most other multivariate statistical methods, DA cannot cope well with missing values. What it does with missing values depends on the software: It might delete the entire row or fill in an average value. For these data, there are no missing values. Further output is as follows:

```
Linear discriminant analysis is used. Observations per group:
1   56
2  332
3  271
4   73
5   54
6  135
7   67
8  114
```

So far we have only discussed linear discriminant analysis (see first line of the output). There are also extensions that can cope better with violation of homogeneity (quadratic discriminant analysis), but these are outside the scope of this book. The additional output (above) shows the number of observations per group. The problem of observers with small numbers of observations was solved by omit-

ting the two with the lowest values from the analysis. We now discuss how many axes to present. As in PCA, the eigenvalues indicate the importance of each axis (discriminant function). The information on eigenvalues is as follows:

Eigenvalues (=lambda)

axis	lambda	lambda as %	lambda cumulative %
1	0.572	67.792	67.792
2	0.171	20.320	88.112
3	0.078	9.249	97.361
4	0.013	1.575	98.936
5	0.007	0.867	99.804
6	0.002	0.196	100.000

It shows that the first two axes represent 88% of the variation, which is more than sufficient. Hence, there is no point in inspecting higher axes. Tests relevant for the number of axes give:

Dimensionality tests for group separation
H_0: No separation on any dimension
B0= 774.142
B0 is a chi-squared statistic with degrees of freedom: 42.000
Probability that a Chi-squared with larger value is found: 0.000

The statistical background for this test is described in Huberty (1994), and it shows that we can reject the null hypothesis that there is no separation on any dimension. This means that we need at least one discriminant function to describe group differences. Further tests (results are not shown here) all indicate that the separation along the first and second axes is significant.

To test the hypothesis that there is no overall group effects, three test statistics are available: the Wilks lambda statistic, the Barlett–Pillai statistic and the Hotelling–Lawley statistic. The first is the most popular. For the sparrow data, all three statistics indicate that there is a significant group effect. Again, in our experience these tests tend to reject the null hypothesis of no group separation even if the ordination diagrams do not show a clear separation.

Statistic	Value	F	Num df	Den df	*p*-value
Wilks lambda	0.493	19.817	42	5111	<0.001
Barlett–Pillai	0.604	17.509	42	6564	<0.001
Hotelling–Lawley	0.843	21.834	42	6524	<0.001

Importance of the variables

If the data contain a large number of variables, it may be interesting to identify which of the variables are responsible for the discrimination of the groups and which ones can be omitted. The discrimination between all the groups is measured by the total sum of Mahalanobis distances. It uses the distances between all the group means. In the backwards selection approach, one can leave out one variable in turn and the total sum of Mahalanobis distances between group means is

calculated again. An important variable, with respect to discrimination between groups, will give a large change in the total sum of Mahalanobis distances, compared with a less important variable. The information below shows the results of a backwards selection procedure.

Variables	Total Mah. distance	Dropped variable
6	61.9532	none
5	60.9435	wt
4	53.5066	flatwing
3	42.0006	head

Variables that were not dropped are tarsus, culmen and nalospi (as the DA was run with at least three variables). To decide how many variables to drop, a stop criterion needs to be used. One option is to make a so-called scree-plot, just as we did for PCA (not shown here); draw the total sum of Mahalanobis distances versus the number of variables and try to detect a cut-off point. This is similar to PCA where eigenvalues can be plotted versus the number of axes.

Classification

We can also apply classification in DA. The classification process is actually very simple. Once the discriminant functions are calculated, you can easily determine to which group average a particular observation is the closest. If an observation from group 1 is the closest to the group average of group 1, then it is classified correctly as group 1 (in this case 22 observations). But if it is closer to the group average of group 2, we classify it as group 2 (in this case three observations).

The classification table below indicates that from the 56 observations made by observer 1 (see group totals above), DA classified 22 of those to group 1 and 3 to group 2, 1 to group 3, etc. So, 39.25% of the observations of group 1 were classified correctly (=22/56).

	1	2	3	4	5	6	7	8
1	22	3	1	8	2	9	4	7
2	56	61	9	75	16	7	61	47
3	24	12	83	2	44	57	15	34
4	4	6	1	34	4	3	9	12
5	2	2	8	6	19	2	8	7
6	17	5	19	4	10	66	3	11
7	3	6	3	13	9	1	30	2
8	11	10	15	22	5	12	11	28

The percentages of correctly classified samples per group are as follows:

1	39.29
2	18.37
3	30.63
4	46.58
5	35.19

6 48.89
7 44.78
8 24.56

The problem with these numbers is that they were obtained using the same data that were used to create the classification rules. Tools exist to obtain more objective classification scores, and one option is a cross-validation process in which the data are split into two parts. The first dataset is used to derive classification rules and the second dataset for testing and obtaining classification scores (Tabachnick and Fidell 2001). Another option is the leave-one-out classification. An observation is left out, the classification rules are determined, and the left out observation is classified. This process is then applied on each observation in turn; see Huberty (1994) and references therein.

Several statistical programmes have routines for DA. Results obtained from these programmes can vary considerably due to different choices for scaling, standardisation, the centring of discriminant functions and the estimation method.