Discriminant Analysis

James H. Steiger

Department of Psychology and Human Development
Vanderbilt University

Discriminant Analysis

1. Introduction
2. Classification in One Dimension
    - A Simple Special Case
3. Classification in Two Dimensions
    - The Two-Group Linear Discriminant Function
    - Plotting the Two-Group Discriminant Function
    - Unequal Probabilities of Group Membership
    - Unequal Costs
4. More than Two Groups
    - Generalizing the Classification Score Approach
    - An Alternate Approach: Canonical Discriminant Functions
    - Tests of Significance
5. Canonical Dimensions in Discriminant Analysis
6. Statistical Variable Selection in Discriminant Analysis

Introduction

There are two prototypical situations in multivariate analysis that are, in a sense, different sides of the same coin. Suppose we have identifiable groups, and they may (or may not) differ in their means (and possibly in their covariance structure) on one or more response measures.

- How can we test whether the groups are significantly different?
- If the groups are different, how can we construct a rule that allows us to accurately assign an individual to one of several groups, depending on their scores on the response measures?
- In this module, we will deal with the second problem, examining, in detail, a method known as *discriminant analysis*.
- However, the first problem, related to a technique known as MANOVA (Multivariate Analysis of Variance) is closely related to the first.

Classification in One Dimension

- There are many situations in which we measure a response variable on a group of people, objects, or situations, and then try to sort these into one or more groups depending on their score on that variable.
- Some examples? (C.P.)

Classification in One Dimension – Some Examples

- Your response variable is the color of a test strip. You try to sort individuals into:
    1. Pregnant
    2. Non-Pregnant
- Your response variable is a brief sensation of change of illumination in a very dark backround. You try to decide whether a very dim signal light is
    1. Present
    2. Not Present
- You have individuals who are either male or female, and you have their heights. You try to devise a rule that will, with the highest possible degree of accuracy, decide only on the basis of height whether a person is:
    1. Male
    2. Female

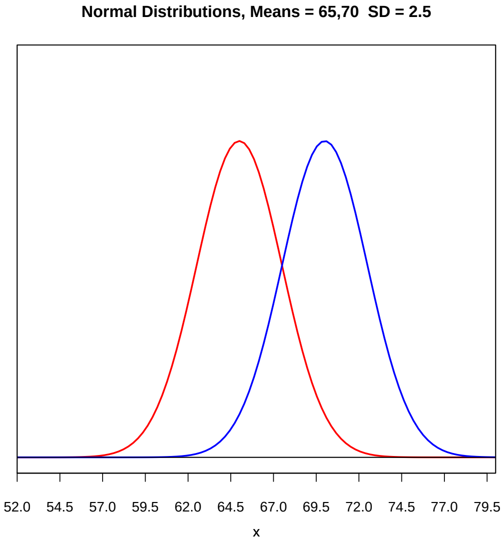
A Simple Special Case

- As a simple special case, suppose we consider the whole population of men and women, and imagine that we *knew* that both populations are normally distributed with standard deviations of 2.5, but men have a mean of 70, women of 65.
- Suppose that men and women occur with equal probability, and we randomly sample a person from the population. What is an *optimal decision rule* for deciding whether the person is male or female, given *only* the information about the person's height?

A Simple Special Case

- The rule we choose depends on what is, for us, optimal.
- For example, in this situation, there are two kinds of misclassification errors we can make:
  1. We can assign a person who is really Male to the Female group.
  2. We can assign a person who is really Female to the Male group.
- If these two types of errors have different costs, then this might effect our decision rule!

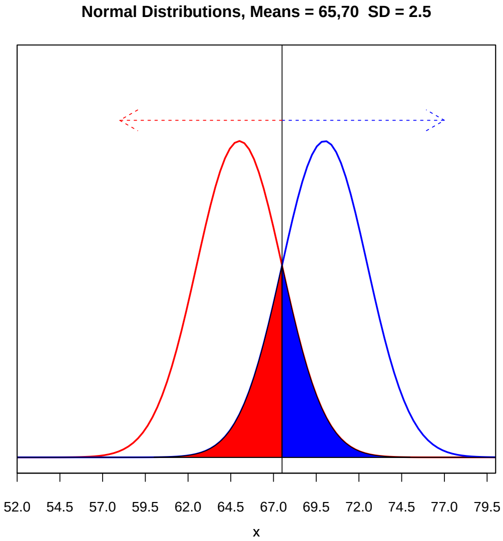
A Simple Special Case



Choosing a Decision Point

- Suppose we choose a decision point based on height. If a person's height is larger than a particular value, we decide they are male, otherwise we decide they are female.
- Where is the best place to put our decision point?
- Let's begin by putting our decision point exactly halfway between the means of the two distributions.
- I've colored in areas under the normal curves corresponding to the two types of misclassification errors.
- The blue area represents the probability of erroneously classifying a female as a male, the red area the probability of erroneously classifying a male as a female.

Choosing a Decision Point



Choosing a Decision Point

- In this case, it is fairly easy to see that moving the decision point slightly to the right or to the left will increase the overall probability of an error.
- So, if males and females are equally represented in the population, this is the optimal decision point.
- However if males and females are not equally represented, or if the costs of the two types of misclassification are different, then the point halfway between the two means would not necessarily be optimal.

Classification in Two Dimensions

- As an extension of our previous simple example, suppose we have two measurements on two or more distinct groups.
- For example, suppose we have heights and weights of a group of people, and we try to predict, on the basis of those data, whether the individuals are male or female.
- For simplicity, let's assume that heights and weights have a bivariate normal distribution for both men and women. For women, the mean vector is $\boldsymbol{\mu}_1 = (65, 135)'$, and for men it is $\boldsymbol{\mu}_2 = (70, 150)'$. Furthermore, assume that both groups have a common covariance matrix given by
$$\boldsymbol{\Sigma} = \begin{bmatrix} 6.25 & 43.75 \\ 43.75 & 625.00 \end{bmatrix}$$
- On the next slide, we plot a simulated data set representing 50 observations at random from both groups.

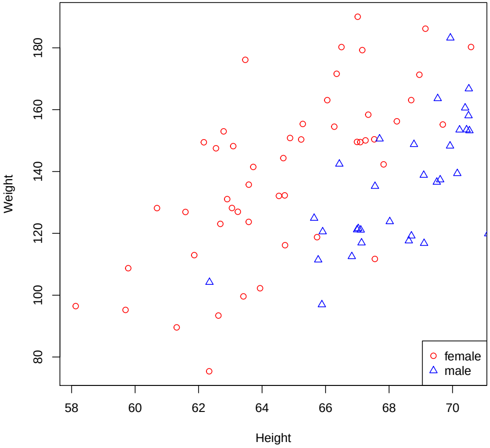
Classification in Two Dimensions

We'll create some data and plot it on the next slide. Here are the commands to create the data.

```
> set.seed(12345)
> mu1 <- c(65,135)
> mu2 <- c(70,150)
> Sigma <- matrix(c(6.25,.7*2.5*25,.7*2.5*25,625),2,2)
> g1 <- mvrnorm(50,mu1,Sigma)
> g2 <- mvrnorm(50,mu2,Sigma)
> group <- rbind(matrix(rep(1,50),50,1),matrix(rep(2,50),50,1))
> data <- rbind(g1,g2)
> data <- cbind(group,data)
> colnames(data) <- c("group","height","weight")
> height.data <- data.frame(data)
> attach(height.data)
```

Classification in Two Dimensions

```
> plot(height[1:50],weight[1:50],pch=1,col="red",xlab="Height",ylab="Weight")
> points(height[51:100],weight[51:100],pch=2,col="blue")
> legend("bottomright",c("female","male"),pch=c(1,2),col = c("red","blue"))
```



Classification in Two Dimensions

- We can see that the points tend to occupy different regions of the two-dimensional data space.
- *Linear* discriminant analysis would attempt to find a straight line that reliably separates the two groups.
- However, since the two groups overlap, it is not possible, in the long run, to obtain perfect accuracy, any more than it was in one dimension.
- In the long run, where should we draw our "line of demarcation"?

Classification in Two Dimensions

- Recall that, in the case of one variable, we put a line of demarcation perpendicular to a line connecting the two group means, at a point halfway between them.
- In two-group discriminant analysis, we do the same thing, except that it is now much more complicated.
  - First, we need to find a *direction* in two dimensional space along which the two groups differ maximally.
  - Next, we compute the mean value, along this direction, for each of the two groups.
  - We draw a connecting line, then draw a line perpendicular to its midpoint.
  - Any observation on the side of the line closer to the mean of group 1 is classified as belonging to group 1, otherwise it is classified as belonging to group 2.
- But this raises the key question — how do we find the direction in two dimensional space that maximally separates the two groups?

A Caveat

- There are a number of different ways of arriving at formulae that produce essentially the same result in discriminant analysis.
- Consequently, different computer programs or books may give different formulae that yield different numerical values for some quantities.
- This can be very confusing.

The Two-Group Linear Discriminant Function

- Suppose we have two groups to be classified, based on a linear function of the classifying variables in $\boldsymbol{x}$.
- Call the discriminant function $L = \boldsymbol{a}'\boldsymbol{x}$.
- We seek an $\boldsymbol{a}$ that produces maximally different mean scores for individuals in the two groups.
- It may be shown (see, e.g., Timm,*Applied Multivariate Analysis*, Equation 3.9.10) that the set of discriminant weights $\boldsymbol{a}_s$ that accomplishes maximal separation is given by

  $$\boldsymbol{a}_s = \boldsymbol{S}^{-1}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2) \tag{1}$$

  where $\boldsymbol{S}$ is the pooled unbiased estimator of the common covariance matrix Σ.

The Two-Group Linear Discriminant Function

- Using $\boldsymbol{a}_s$ as defined above, the mean difference in discriminant scores is

$$\begin{aligned}\overline{L}_1 - \overline{L}_2 &= \boldsymbol{a}_s'\overline{\boldsymbol{x}}_1 - \boldsymbol{a}_s'\overline{\boldsymbol{x}}_2 \\ &= \boldsymbol{a}_s'(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2) \\ &= (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)'\boldsymbol{S}^{-1}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2) \qquad (2)\end{aligned}$$

- The above expression is known as Mahalanobis' D^2, and is a measure of distance between two groups of scores.
- When we get to MANOVA, we shall see that this statistic is closely related to Hotelling's T^2 statistic used for testing the equality of two mean vectors.

Plotting the Two-Group Discriminant Function

- The linear weights for the discriminant function define the direction in two-dimensional space that most effectively discriminates between the two groups.

```
> centroid.1 <- c(mean(height[group==1]),mean(weight[group==1]))
> centroid.2 <- c(mean(height[group==2]),mean(weight[group==2]))
> xs <- c(centroid.1[1],centroid.2[1])
> ys <- c(centroid.1[2],centroid.2[2])
> mid.point <- (centroid.1 + centroid.2)/2
> mid.point <- matrix(mid.point,2,1)
> data.1 <- cbind(height[group==1],weight[group==1])
> data.2 <- cbind(height[group==2],weight[group==2])
> S <- (var(data.1)+var(data.2))/2
> xbar.1 <- matrix(centroid.1,2,1)
> xbar.2 <- matrix(centroid.2,2,1)
> a <- solve(S) %*% (xbar.1 - xbar.2)
> a

            [,1]
[1,] -1.29126337
[2,]  0.07880716
```

Plotting the Two-Group Discriminant Function

- We can use the linear weights in $\boldsymbol{a_s}$ to compute discriminant scores for each individual.
- If the ith individual has score vector $\boldsymbol{x}_i$, then that individual's discriminant score is $L_i = \boldsymbol{a}_s'\boldsymbol{x_i}$
- If we plot the discriminant weights as a line in 2-dimensional space, the discriminant scores are proportional to the projection of an individual's data point onto that line.
- This is not simple to visualize in this case — because height and weight are plotted with axes having different numerical scales, lines that are perpendicular do not appear to be at right angles on the plot.
- So I'll work at it in reverse.

Plotting the Two-Group Discriminant Function

- Where do we draw the line? The rule for assigning individuals to groups is to
  - Compute the discriminant score.
  - If an individual discriminant score is higher than the discriminant score computed at a *cutoff* point halfway between the two group centroids (i.e., at an overall weighted average score), then assign the individual to group 1, otherwise assign to group 2.
- The cutoff point is thus $c = \boldsymbol{a}_s'(\overline{\boldsymbol{x}}_1 + \overline{\boldsymbol{x}}_2)/2$, which can also be written as $c = \frac{1}{2}(\overline{\boldsymbol{x}}_1 + \overline{\boldsymbol{x}}_2)'\boldsymbol{S}^{-1}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)$.

```
> cutoff <- t(a) %*% mid.point
> cutoff

          [,1]
[1,] -75.47808
```

Plotting the Two-Group Discriminant Function

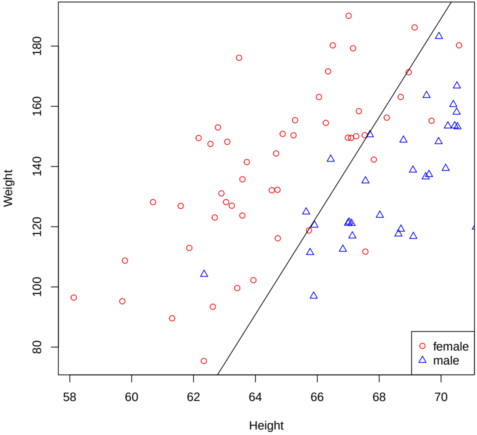
- The *cutoff line* for deciding whether to classify an observation as group 1 or group 2 is at the point $a_1x_1 + a_2x_2 = c$.
- This may be re-expressed in the classic form of a linear equation as

$$x_2 = -(a_1/a_2)x_1 + c/a_2 \tag{3}$$

that is, a straight line with a slope of $-(a_1/a_2)$ and an intercept of c/a_2.

Plotting the Two-Group Discriminant Function

```
> plot(height[1:50],weight[1:50],pch=1,col="red",xlab="Height",ylab="Weight")
> points(height[51:100],weight[51:100],pch=2,col="blue")
> legend("bottomright",c("female","male"),pch=c(1,2),col = c("red","blue"))
> abline(cutoff/a[2],-(a[1]/a[2]))
```



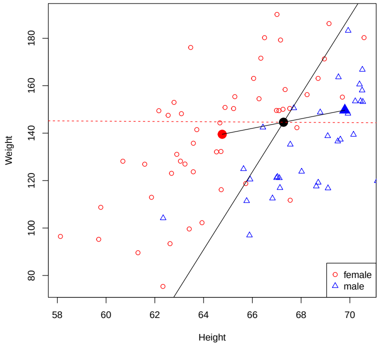
Plotting the Two-Group Discriminant Function

- The discriminant function is evaluated by projecting points onto the discriminant function line, which has a slope of a_2/a_1, and an intercept of 0. This line is not visible in the current plot, but we can make it visible by moving it upwards, so that it intersects with the midpoint between the two group centroids.
- It is convenient to use the point-slope function. I've written this function to plot a straight line that has a given slope and intersects with a given point.

```
> point.slope.line <- function(point,slope,col="black",lty=1)
+ {
+  x.0 <- point[1]
+  y.0 <- point[2]
+  intercept <- y.0 - slope*x.0
+  abline(intercept,slope,lty=lty,col=col)
+ }
```

Plotting the Two-Group Discriminant Function

```
> plot(height[1:50],weight[1:50],pch=1,col="red",xlab="Height",ylab="Weight")
> points(height[51:100],weight[51:100],pch=2,col="blue")
> legend("bottomright",c("female","male"),pch=c(1,2),col = c("red","blue"))
> abline(cutoff/a[2],-(a[1]/a[2]))
> points(centroid.1[1],centroid.1[2],pch=19,cex=2,col="red")
> points(centroid.2[1],centroid.2[2],pch=17,cex=2,col="blue")
> xs <- c(centroid.1[1],centroid.2[1])
> ys <- c(centroid.1[2],centroid.2[2])
> lines(xs,ys)
> points(mid.point[1],mid.point[2],pch=19,cex=2,col="black")
> point.slope.line(mid.point,a[2]/a[1],lty=2,col="red")
```



```
> discriminant.scores <- a[1]*height + a[2]*weight
> W.hat <- discriminant.scores - cutoff
```

Plotting the Two-Group Discriminant Function

- The red dotted line is the discriminant function line. The black solid line is the "line of demarcation" that also shows the correct direction to orthogonally project points onto the discriminant function line.
- The two lines are actually perpendicular, but do not appear so in the plot because the numerical scales in the plot are not the same.

Plotting the Two-Group Discriminant Function

- Using the `identify` function in R, we can identify points and also compute their discriminant functions.
- For example, point number 65 lies just above the midpoint, and just to the left of the demarcation line. Point number 32 lies just to the right of the demarcation line. To compute the amount by which the discriminant score is above or below the cutoff, I subtracted the cutoff value to generate a "decision score."

$$\begin{aligned}\hat{W}_i &= a_1 height_i + a_2 weight_i - c \\ &= a_1 height_i + a_2 weight_i - \frac{1}{2}(\overline{\boldsymbol{x}}_1 + \overline{\boldsymbol{x}}_2)' \boldsymbol{S}^{-1}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)\end{aligned}$$

- Using the decision scores, we classify an observation in group 1 if the decision score $\hat{W}_i$ is greater than 0.

```
> W.hat[65]

[1] -5.349976

> W.hat[32]

[1] 3.926848
```

Unequal Prior Probabilities

- Suppose that we somehow knew that groups 1 and 2 are unequally represented in the population with probabilities Pr(1) and Pr(2), respectively.
- Should this affect our decision rule?

Unequal Prior Probabilities

- Anderson's classification rule that minimizes the total probability of misclassification (TPM) uses the decision score to assign a person to group 1 if the decision score exceeds $\ln(\Pr(2)/\Pr(1))$, or, alternatively, if

$$W^* = \hat{W} - \ln(\Pr(2)/\Pr(1)) > 0$$

.

- Suppose we knew that, in our classification system, males were actually 9 times as likely to occur as females.
- Can you diagram the new decision line?

Unequal Prior Probabilities

- Anderson's classification rule that minimizes the total probability of misclassification (TPM) uses the decision score to assign a person to group 1 if the decision score exceeds $\ln(\Pr(2)/\Pr(1))$, or, alternatively, if

$$W^* = \hat{W} - \ln(\Pr(2)/\Pr(1)) > 0$$

.

- Suppose we knew that, in our classification system, males were actually 9 times as likely to occur as females.
- Can you diagram the new decision line?

Unequal Costs

- Suppose the costs of misclassification are unequal, and we wish to minimize the overall cost. A classification rule that minimizes the expected cost uses the decision score to assign a person to group 1 if the decision score exceeds $\log[(C(1|2)\,\Pr(2))/(C(2|1)\,\Pr(1))]$, or, alternatively, if

$$W^{**} = \hat{W} - \log[(C(1|2)\,\Pr(2))/(C(2|1)\,\Pr(1))] > 0$$

.

- Suppose we knew that, in our classification system, males and females were equally likely, but the cost of an error for misclassifying a male as a female is twice as great as the cost of an error for misclassifying a female as a male.
- Can you diagram the new decision line?

More than Two Groups

- Suppose you have $k > 2$ groups, and wish to discriminate between them.
- In this case, Anderson (1984, Chapter 6) has shown that the Bayes rule for classifying an observation is based on the same discriminant function defined previously, except now a pairwise function W_{ij} is computed for all pairs of groups.
- The classification rule becomes the following: Assign observation vector $\boldsymbol{x}$ to population i if $W_{i,j} > 0 \forall j \neq i$.
- It should be noted that $W_{ji} = -W_{ij}$, and that any $k-1$ linearly independent W_{ij} form a basis for the complete set of statistics if $p \geq (k-1)$. If $p < (k-1)$, then the space of the W_{ij} will have rank p, and the classification rule can be specified in terms of p scores.
- To compensate for unequal prior probabilities and/or unequal costs, we utilize the same correction factors for each W_{ij} that were described for the two-group case.

More than Two Groups

- Consider the case of 3 groups.
- The discriminant functions are:

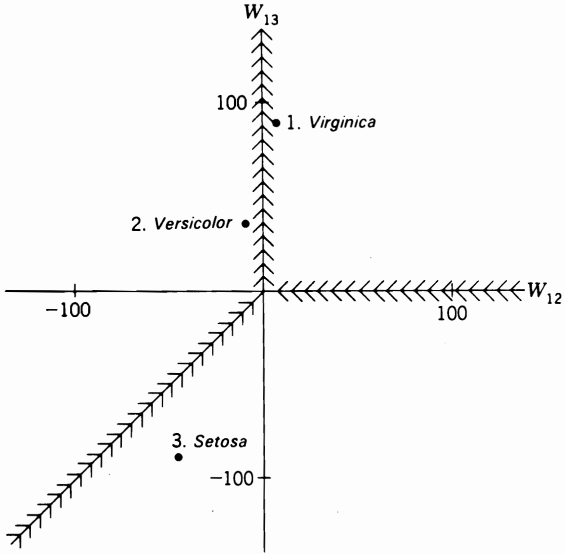
$$\begin{aligned}
W_{12} &= \boldsymbol{x}'\boldsymbol{S}^{-1}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2) - \frac{1}{2}(\overline{\boldsymbol{x}}_1 + \overline{\boldsymbol{x}}_2)'\boldsymbol{S}^{-1}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2) \\
W_{13} &= \boldsymbol{x}'\boldsymbol{S}^{-1}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_3) - \frac{1}{2}(\overline{\boldsymbol{x}}_1 + \overline{\boldsymbol{x}}_3)'\boldsymbol{S}^{-1}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_3) \qquad (4) \\
W_{23} &= \boldsymbol{x}'\boldsymbol{S}^{-1}(\overline{\boldsymbol{x}}_2 - \overline{\boldsymbol{x}}_3) - \frac{1}{2}(\overline{\boldsymbol{x}}_2 + \overline{\boldsymbol{x}}_3)'\boldsymbol{S}^{-1}(\overline{\boldsymbol{x}}_2 - \overline{\boldsymbol{x}}_3)
\end{aligned}$$

- Note that $W_{23} = W_{13} - W_{12}$. Because of this linear redundancy, we can devise a decision rule using only W_{12} and W_{13}.

More than Two Groups

- The classification rule is as follows: Classify $\boldsymbol{x}$ as from
  - Population 1 if $W_{12} > 0$ and $W_{13} > 0$.
  - Population 2 if $W_{12} < 0$ (i.e.,$W_{21} > 0$) and $W_{13} > W_{12}$ (i.e, $W_{23} = W_{13} - W_{12} > 0$).
  - Population 3 if $W_{13} < 0$ (i.e., $W_{31} > 0$) and $W_{13} < W_{12}$ (i.e, $W_{32} = W_{12} - W_{13} > 0$).

More than Two Groups



Canonical Discriminant Functions

- With two groups, an alternative way of computing the (single) classification function is the eigenvector of the matrix $\boldsymbol{B}^{-1}\boldsymbol{A}$, where $\boldsymbol{B}$ and $\boldsymbol{A}$ are multivariate (MANOVA) analogs of sum of squares within and sum of squares between computed in ANOVA.
- Before pursuing this approach, we digress to obtain background on the meaning of these two matrices, and how they relate to ANOVA and MANOVA.
- This background is in the lecture notes on ANOVA and MANOVA and the general linear model.
- After completing these notes, we will resume on the next slide.

Canonical Discriminant Functions

- Let $\boldsymbol{V}$ be the eigenvectors corresponding to the meaningful eigenvalues of $\boldsymbol{B}^{-1}\boldsymbol{A}$.
- Let $\boldsymbol{W} = \boldsymbol{B}/(N - q)$ be the pooled estimate of the within-groups covariance matrix.
- The "raw" discriminant weights $\boldsymbol{a}$ are normalized so that $\boldsymbol{v}'\boldsymbol{W}\boldsymbol{v} = 1$ for any column of V. That is, $\boldsymbol{a}_i = \boldsymbol{v}_i/\sqrt{(\boldsymbol{v}_i'\boldsymbol{W}\boldsymbol{v}_i}$.
- Commercial programs print "standardized" weights as an aid to interpretation. Over the years, there has been substantial controversy over the proper method to standardize the weights.
- In SPSS and Stata, the values in $\boldsymbol{a}$ currently are standardized by multiplying them by the variable standard deviations computed from $\boldsymbol{W}$. That is, $\boldsymbol{a}_s = (\mathbf{diag}\,(\boldsymbol{W}))^{1/2}\boldsymbol{a}$.

Eigenvalues and Canonical Correlations

- The eigenvalues λ_i of the matrix $\boldsymbol{BA}^{-1}$ are related to the canonical correlation between the set of group indicator variables and the variables used to discriminate between the groups by the relationship

$$r_i^2 = \frac{\lambda_i}{1 + \lambda_i} \tag{5}$$

- So, for example, if the first eigenvalue is 1, then the corresponding squared canonical correlation is $1/2$, and the canonical correlation is .7071.

Wilks' Λ

- There are a number of tests of significance in discriminant analysis and MANOVA.
- A primary test statistic that is a monotone function of the likelihood ratio statistic is Wilks' Λ, given by

$$\Lambda = \frac{|\boldsymbol{B}|}{|\boldsymbol{A} + \boldsymbol{B}|} = \frac{1}{|\boldsymbol{B}^{-1}\boldsymbol{A} + \boldsymbol{I}|} \tag{6}$$

- The determinant of a covariance matrix is sometimes referred to as the *generalized variance*, because it is equal to the square of the area (or volume) of an N-dimensional parallelogram with sides equal to the standard deviations of the variables.
- This explains why $|\boldsymbol{\Sigma}|^{-1/2}$ appears as a standardizing constant in the multivariate normal density.
- Under the assumption of multivariate normality and equality of covariance matrices, the distribution of Λ is known.

Wilks' Λ

- Since the determinant of a matrix is the product of its s nonzero eigenvalues, we have, from Equation 6:

$$\Lambda = \prod_{i=1}^{s} \frac{1}{1+\lambda_i} \tag{7}$$

- We are interested in which, if any, of the s dimensions are significant. In the context of discriminant functions, Wilks' Λ is more useful than the other three MANOVA test statistics, because it can be used on a subset of eigenvalues, as we see shortly.

Hotelling Trace Criterion

- This criterion is

$$\tau = \mathrm{Tr}(\boldsymbol{B}^{-1}\boldsymbol{A}) \tag{8}$$

Roy's Largest Root Criterion

- This criterion is a function of the largest eigenvalue λ_1 of $(\boldsymbol{B}+\boldsymbol{A})^{-1}\boldsymbol{A}$.
- The criterion is

$$\theta = \frac{1}{1+\lambda_1} \tag{9}$$

Pillai-Bartlett Trace Criterion

- This criterion uses the eigenvalues λ_i of $(\boldsymbol{B}+\boldsymbol{A})^{-1}\boldsymbol{A}$

$$\sum_{i=1}^{s} \frac{\lambda_i}{1+\lambda_i} \tag{10}$$

Hotelling Trace Criterion

- This criterion is

$$\tau = \mathrm{Tr}(\boldsymbol{B}^{-1}\boldsymbol{A}) \tag{11}$$

Comparing the Criteria

- If the null hypothesis of equal mean vectors is true, all 4 criteria have the same rejection rate.
- On the other hand, if the null hypothesis is false, there is no one uniformly most powerful test. Power for any procedure depends on how the mean vectors are aligned in multidimensional space.
- For example, if the mean vectors are in a straight line in multidimensional space, then they can be maximally separated along a single dimension, and Roy's greatest root criterion will be most powerful.
- On the other hand, according to Rencher (*Methods of Multivariate Analysis, 2nd Edition*), 2002, p. 177), when the pattern of means is relatively diffuse in multidimensional space, Roy's criterion is least powerful, the Pillai-Bartlett trace criterion and Wilks' Λ the most powerful.
- Wilks' Λ has the very substantial advantage of lending itself readily to sequential tests.
- In general, the criteria tend in practice to produce highly similar results for most data.

Canonical Dimensions in Discriminant Analysis

- If there are more than two groups, more than one classification function will be available.
- The eigenvectors of $\boldsymbol{B}^{-1}\boldsymbol{A}$ define the dimensions that maximally separate between the groups.
- In general, there will be $s = \min(p, k-1)$ canonical discriminant functions, where k is the number of groups and p the number of variables.
- A well-known example is Rencher's (2002, p. 279) football player data.

Canonical Dimensions in Discriminant Analysis

- The data in Rencher's Table 8.3 were collected by G. R. Bryce and R. M. Barker (Brigham Young University) as part of a preliminary study of a possible link between football helmet design and neck injuries.
- Six head measurements were made on each subject. There were 30 subjects in each of three groups: high school football players (group 1), college football players (group 2), and nonfootball players (group 3).
- The six variables are
  1. WDIM = head width at widest dimension,
  2. CIRCUM = head circumference,
  3. FBEYE = front-to-back measurement at eye level,
  4. EYEHD = eye-to-top-of-head measurement,
  5. EARHD = ear-to-top-of-head measurement,
  6. JAW = jaw width.

Canonical Dimensions in Discriminant Analysis

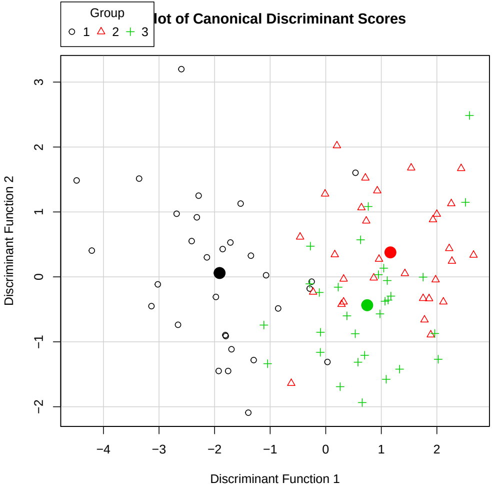
- The following code sets up the data for a standard analysis on the dimensions.

```
> library(car)
> library(MASS)
> source(
+   "http://www.statpower.net/R312/Steiger R Library Functions.txt")
> fb.data <- read.table(
+   "http://www.statpower.net/R312/football.txt",header=T,sep=",")
> ## Analyze FB data
> ## ###################################
> ## Create x,D,H,and Group matrices
> ######################################
> x <- as.matrix(fb.data[,2:7])
> Group <- as.matrix(fb.data[,1:1])
> D <- Make.D(Group)
> H <- Make.H(Group)
```

Canonical Dimensions in Discriminant Analysis

- Here is the plot of the scores.

```
> Plot.Discriminant.Scores(x,D,H,Group)
```



Canonical Dimensions in Discriminant Analysis

- Here is the table of statistical analyses.

```
> print(Canonical.Table(x,D,H))

     Fcn  Eigen  Prop CanCorr Lambda  F-Stat df1 df2
[1,]   1 1.9178 0.943  0.8107 0.3071 10.9941  12 164
[2,]   2 0.1159 0.057  0.3223 0.8961  1.9245   5  83
       prob
[1,] 0.0000
[2,] 0.0989
```

- The λ_i are the respective eigenvalues of $\boldsymbol{B}^{-1}\boldsymbol{A}$, and the squared canonical correlation between the scores on a dimension and the set of dummy variables representing the groups is given by

$$r_i^2 = \frac{\lambda_i}{1+\lambda_i}$$

- A test of significance is given for each dimension, as well as a proportion of the total of the eigenvalues.
- In this case, we find that the first dimension separates the groups very well, but the second canonical dimension is of limited use.

Canonical Dimensions in Discriminant Analysis

- What are the dimensions? The best way of evaluating and naming the dimensions is to examine the standardized discriminant weights.
- Below, we see that two of the variables are essentially unrepresented in the first canonical discriminant function
- This brings up the question of which variables actually contribute "significantly" to discrimination between the groups, which leads naturally to the topic of stepwise discriminant analysis.

```
> print(Standardized.Discriminant.Weights(x,D,H))

               [,1]          [,2]
WDIM    0.620641211 -0.9205833819
CIRCUM -0.006471485  0.0009114308
FBEYE  -0.004758090  0.0211450008
EYEHD  -0.718812268 -0.5997882273
EARHD  -0.396511561  0.3018196450
JAW    -0.507721826  0.9368744941
```

Statistical Variable Selection in Discriminant Analysis

- Wilks' Λ lends itself to stepwise evaluation of variables in discriminant analysis.
- The *partial* Λ for evaluating the contribution of a variable (or set of variables) $\boldsymbol{x}$ over and above a set $\boldsymbol{y}$ is given by

$$\Lambda_{x|y} = \frac{\Lambda_{x,y}}{\Lambda_y}$$

- An F-statistic is available for analyzing the statistical significance of a partial Λ, and can be used to evaluate whether a variable contributes significantly to group discrimination.
- Note, of course, that as in any stepwise procedure, this approach is subject to abuse and should ideally be used with caution.
- However, in some cases, one enters the analysis with a definite question. For example: Do measures of spatial ability, over and above math and verbal ability measures, add to our ability to discriminate between groups characterized by levels of high creative achievement?
- Forward and stepwise selection procedures work essentially the same here as in multiple regression. The full stepwise procedure, after adding a variable at each stage, deletes any previously added variables that have "become non-significant" as a result of the addition of the latest variable.

Statistical Variable Selection in Discriminant Analysis

Fortunately, there is a function that automates stepwise discriminant analysis:

```
> ## stepwise discriminant analysis
> library(klaR)
> options(digits=4,scipen=10,width=70)
> fit <- greedy.wilks(GROUP ~ .,
+                        data=fb.data,niveau = .10)
> fit

Formula containing included variables:

GROUP ~ EYEHD + WDIM + JAW + EARHD
<environment: 0x0000000017f3a060>


Values calculated in each step of the selection procedure:

   vars Wilks.lambda F.statistics.overall p.value.overall
1 EYEHD       0.4279                58.16       9.182e-17
2  WDIM       0.4003                24.96       2.604e-16
3   JAW       0.3383                20.38       6.677e-18
4 EARHD       0.3072                16.89       2.888e-18
  F.statistics.diff p.value.diff
1            58.162    9.182e-17
2             2.964    5.687e-02
3             7.791    7.766e-04
4             4.257    1.730e-02
```