Discriminant Function Analysis

Joseph J. Luczkovich

March 18, 2014

0.1 Introduction

Sometimes, the multivariate data we possess is useful for predicting the classification or group membership of a new case that comes to us. The multivariate data from the known groups can serve as a **training set** when predicting the group membership of new cases. For example, we may be coastal biologists trying understand where an oil spill may have floated underwater and caused damage to the ecological community. We would like to group sample locations in which collections of multiple species have been made, some of which are known to be polluted and others are known not to be polluted. Which stations are most similar, can they be divided into groups that are meaningful (polluted and non-polluted) based on the multivariate species data? If we can discriminate the polluted and non-polluted with a discriminant function analysis, can we then take similar measurements of species at a new site suspected of being polluted and predict its class membership, i.e., this is a collection of species from a polluted site? Other examples might be classifying students applying to college into accept" and "reject" groups based on objective criteria such as SAT scores, grade point average, ranking in high school class, etc.; classifying hatchery-reared and wild-spawned fishes based on multiple chemical constituents of their otolith microchemistry; seeing if various drilling techniques, geological features, and water quality measurements can predict if ground water violation has occurred at a new fracking site. Many remote sensing land use classifications are based on a discriminant function approach: if you know certain land use categories (wetland, forested land, agricultural land, developed land, etc.) have particular spectral properties in a satellite image, you can use this training set and LDFA to classify a new remotely sensed land image. Note that this is different than in the case of clustering, where the categorical groupings are not known in advance.

The math is the reverse of MANOVA, which used categorical predictors variables (group membership) to test for differences between multiple continuous response variables. In **Linear Discriminant Function Analysis (LDFA)**, the investigator wants to discriminate between *a priori* defined, mutually exclusive groups in such a way as to minimize misclassification. The investigator has the goal of predicting with some accuracy which group a new case will be classified in or assigned to using p data variables from n objects. The figure

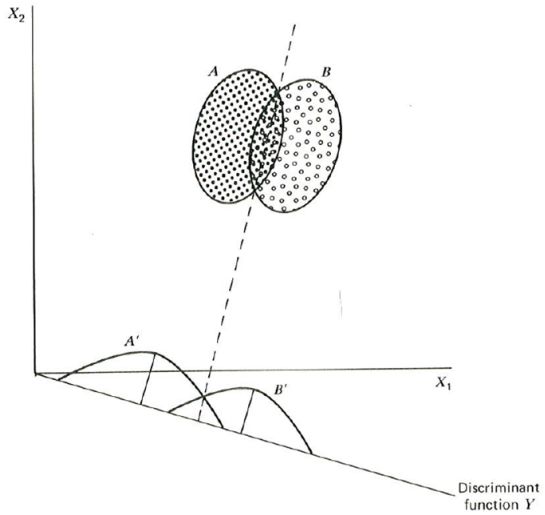


Figure 10.1-1. Graphical illustration of two-group discriminant analysis.

Figure 1: Graphical Illustration of the two-group discriminant function analysis. Variables X1 and X2 were measured on multiple indivduals in groups A and B. A new discriminant function Y is shown as an axis that maximizes the differences between groups A and B.

(Figure 1) shows a diagram that explains this LDFA concept visually for two groups. Two groups are easier to visualize than multiple groups, Two mutually exclusive groups A and B are determined in advance. Next, n individuals from each group are measured for p variables, here just two, called $X1$ and $X2$, which are then used to create a new linear combination of the variables, Y. This new linear combination Y of the original variables is the **discriminant function**, and it is chosen to maximize the difference between the weighted averages of the two groups (note that the distribution of A and B when projected on the Y axis are not overlapping much. Any other Y axis selected would have a greater overlap between A and B.

In terms of matrix math, LDFA is a way of computing the vector of discriminant scores Y for each case using:

$$Y = \beta' X$$

Where Y is a vector of discriminant scores, β' is a 1 x p vector of discriminant weights (like β coefficients in linear regression), and X is a n x p matrix containing the measured values for each of the n individuals and p independent variables.

0.1.1 Assumptions of LDFA

The multivariate data are assumed to be:

- Linear (transform if not)
- No outliers (remove or transform)
- Homoscedacity of variance (transform if not)
- Multivariate normal (transform if not)
- No multi-collinearity in the predictor variables

0.2 Example of LFDFA in R

Here we will use a set of continuous variables to predict a categorical data (classification). In this example, I will predict the class membership of species of fish from Finland. First we will open the "MASS" or "Modern Applied Statistics with S" package. We will also open the lattice package for 3-D graphics. Finally, we will load the data **FinnishFish.csv** from a local file (this was also a data set on your last exam).

```
> library ("MASS", lib.loc="C:/Program Files/R/R-3.0.2/library")
> library("lattice", lib.loc="C:/Program Files/R/R-3.0.2/library")
> Finnish_Fish <- read.csv("C:/Users/Joseph/Dropbox/CRM7008/DFA/Finnish_Fish.csv")
>
```

This will be our training set for the LDFA. Here are the summarized data:

```
> summary(Finnish_Fish)
```

```
  Observation         Species      Weight.g        SLength.cm
 Min.   :  1.0   Bream    :35   Min.   :   0.0   Min.   : 7.50
 1st Qu.: 40.5   Parkki   :11   1st Qu.: 120.0   1st Qu.:19.05
 Median : 80.0   Perch    :56   Median : 272.5   Median :25.20
 Mean   : 80.0   Pike     :17   Mean   : 398.7   Mean   :26.25
 3rd Qu.:119.5   Roach    :20   3rd Qu.: 650.0   3rd Qu.:32.70
 Max.   :159.0   Smelt    :14   Max.   :1650.0   Max.   :59.00
                 Whitefish: 6   NA's   :1
   FLength.cm      Tlength.cm    Height.Length    Width.Length
 Min.   : 8.40   Min.   : 8.80   Min.   :14.50   Min.   : 8.70
 1st Qu.:21.00   1st Qu.:23.15   1st Qu.:24.25   1st Qu.:13.40
 Median :27.30   Median :29.40   Median :27.10   Median :14.60
 Mean   :28.42   Mean   :31.23   Mean   :28.31   Mean   :14.12
 3rd Qu.:35.50   3rd Qu.:39.65   3rd Qu.:37.60   3rd Qu.:15.30
 Max.   :63.40   Max.   :68.00   Max.   :44.50   Max.   :20.90

      Sex
```

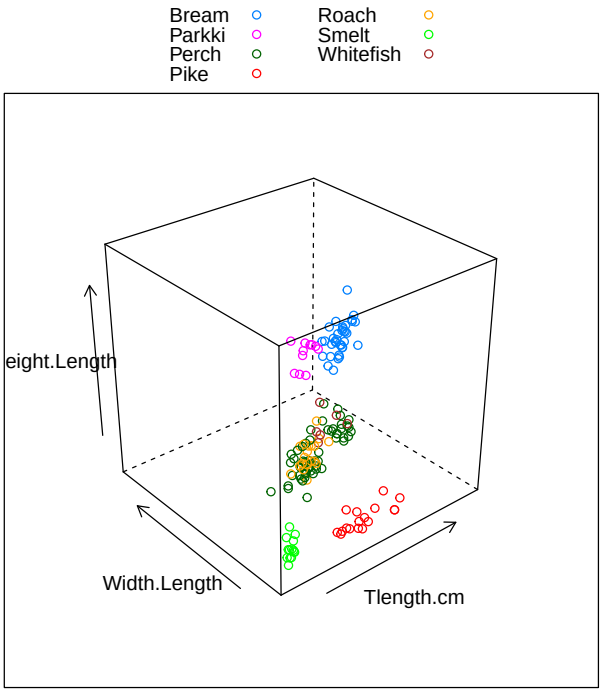
```
 Min.   :0.0000
 1st Qu.:0.0000
 Median :0.0000
 Mean   :0.2361
 3rd Qu.:0.0000
 Max.   :1.0000
 NA's   :87
```

The "Observation" columm is simply a case number, and the summary is not interesting. There are different numbers of fish in each species (Number of cases is given under "Species"). The remaining variables are:

- Weight.g = the weight of each specimen in grams
- SLength.cm = the standard length of each specimen in cm
- FLength.cm = the fork length of each specimen in cm
- TLength.cm = the total length of each specimen in cm
- Height.Length = The ratio of body height to total length
- Width.Length = The ratio of body width to total length
- sex = sex of specimen

Let's make a 3D plot of the variables. We will use the Lattice 3D graphic function **cloud()** where the plot is specified by a formula z x*y:

```
> attach(Finnish_Fish)
> cloud(Height.Length~Tlength.cm*Width.Length,groups=Species,
+       xlim=c(0,100),ylim=c(0,100),auto.key=list(columns = 2))
```



The colors in the graph represent the different fish species. Note how they do not overlap much in the 3D space, at least for some species. This is a good thing, because we can now develop our LDFA and it will be significant, at least for the non-overlapping species. We should be able to discriminate species based on these three measurements. These fishes tend to have very different lengths, widths:length ratio, and height:length ratio. In biology, this kind of measurement display is called **morphometry** and most species differ in their morphology. This is due to evolutionary selection pressures acting on the different species of fish, so they all have different ecological niches. If they don't vary much, they may be closely related, have very similar selection pressures, or be the same species.

Now let's do a linear discriminant function analysis. The command in R is **lda()** where a y x1+x2+x3 formula is used to specify the predictors.

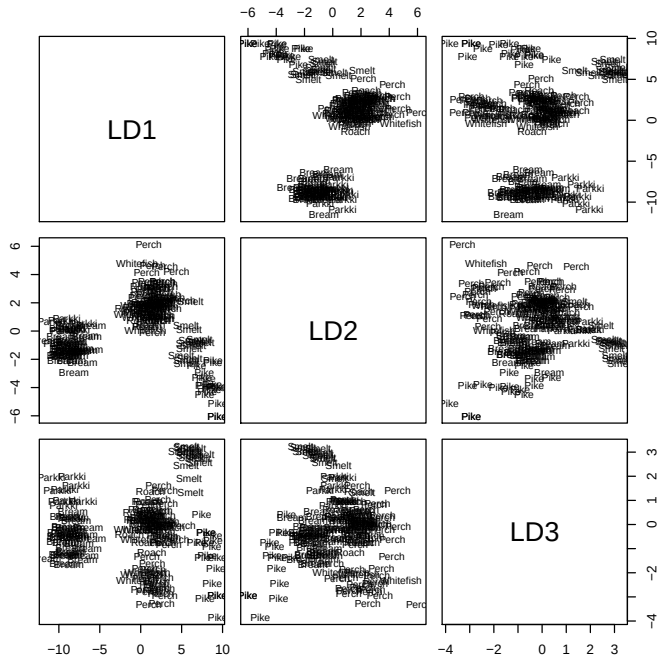
```
> F<-cbind(Height.Length,Tlength.cm,Width.Length)
> group<-factor(Finnish_Fish$Species)
> fit_lda<-lda(group~Height.Length+Tlength.cm+Width.Length)
> print(fit_lda)

Call:
lda(group ~ Height.Length + Tlength.cm + Width.Length)

Prior probabilities of groups:
```

```
     Bream     Parkki      Perch       Pike      Roach      Smelt  Whitefish
0.22012579 0.06918239 0.35220126 0.10691824 0.12578616 0.08805031 0.03773585

Group means:
          Height.Length Tlength.cm Width.Length
Bream           39.52571   38.35429     14.13143
Parkki          39.30909   22.79091     14.08182
Perch           26.25714   29.57143     15.83929
Pike            15.84118   48.71765     10.43529
Roach           26.73500   24.97000     14.60500
Smelt           16.88571   13.03571     10.22143
Whitefish       29.20000   34.31667     15.90000

Coefficients of linear discriminants:
                      LD1         LD2         LD3
Height.Length -0.76383137 -0.07723803  0.04172403
Tlength.cm     0.04004809 -0.09364970 -0.11403301
Width.Length   0.42303453  0.96147310 -0.35995770

Proportion of trace:
   LD1    LD2    LD3
0.8624 0.1053 0.0323
```

The output first reports the model we specified: Species group is linear function of Height.Length, TLength.cm, Width.Length. Next, the prior probabilities of each group are reported. This is simply the proportion of each species in the 159 observation (i.e, divide the number of bream by the total of all fish measured you get 0.22, or if you picked a fish up at random, this is the probability of it being correctly assigned to that species by chance alone). Next, we see a listing of the group means for each species and each variable. Finally, the β coefficients for the three LDFs are provided. Just like in principal component analysis (PCA), there are only three LDF discriminants, because we only included three predictors in the model. You cannot have more LDF's than predictor variables. Most of the variance will be explained on the first LDF, also like in PCA, with less and less variance explained on each one thereafter. The last line of the output explains the proportion of variance explained on each LDF: here LD1 explained 86 percent of the variance, and LD2 explained 10 percent. The output object in R has all of the above plus the **discriminant scores** for each case, the predicted vector Y, stored in it, so we can plot that.

Let's make a plot of discriminant scores for each case the Linear Discriminant Functions:

```
> plot(fit_lda)
```



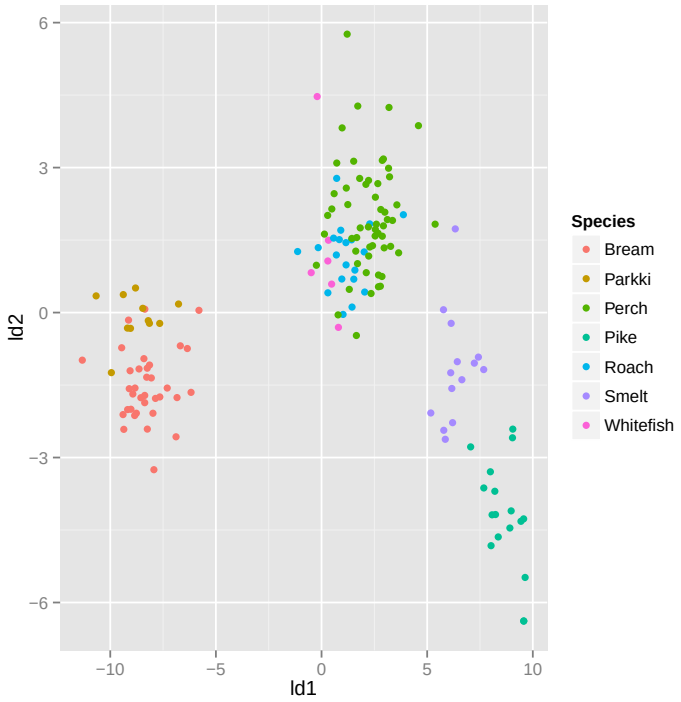
You can see in this scatterplot matrix plot that LD1 vs LD2 plot has the best separation among the fish species groups. There are three big clouds (the species identities are given as text annotations on the graphs). We can see on LDI vs. LD2 that pike and smelt (right side, high LD1 scores, low LDF2 scores) are very different from bream and pakrki (left side, low LD1 scores), and perch and whitefish form an intermediate group (but with high LD2 scores). There is some group separation in LD1 versus LD3 as well.

0.2.1 Discriminant Scores

The output above, using the LD coefficients, can be used to compute the discrimiant scores for each case. The LD1, LD2, and LD3 coefficents are multipled by the variables measured foir each of the fishes (159 cases) to get discriminant scores for each case. This is done in a single step in R by use of the **predict()** command. We can use predict(fitlda) to store these scores in a output object called fish.scores. We can bind these discriminant scores to the orginal data and plot them. A nicer plot of LD1 vs LD2 discriminant scores can be made using ggplot2:

```
> library("ggplot2", lib.loc="C:/Users/Joseph/Documents/R/win-library/3.0")
> fish.scores<-predict(fit_lda)
> ld1<-fish.scores$x[,1]
> ld2<-fish.scores$x[,2]
```

```
> ld3<-fish.scores$x[,3]
> fish<-cbind(Finnish_Fish,ld1,ld2,ld3)#adds the ld1,2,3scores to data
> qplot(ld1,ld2,geom=c("point"),color=Species,data=fish)
```



0.2.2 Accuracy of the Discriminant Function

The next thing to do is to check the accuracy of the LDFA model. Here we will use the original training set to check the classification accuracy:

```
> table(Species,fish.scores$class)
```

Species	Bream	Parkki	Perch	Pike	Roach	Smelt	Whitefish
Bream	35	0	0	0	0	0	0
Parkki	1	10	0	0	0	0	0
Perch	0	0	48	0	7	0	1
Pike	0	0	0	17	0	0	0
Roach	0	0	8	0	11	0	1
Smelt	0	0	0	0	0	14	0
Whitefish	0	0	1	0	3	0	2

This table, often called a ”confusion matrix”, lists the species classifications (columns) from the LDFA by their actual species memberships (rows). The diagonal are the number correctly classified for each species by the LDFA. The off-diagonals are the misclassifications. We incorrectly classified:

- parkki as bream (1 case),
- perch as roach (7 cases), whitefish (1 cases)
- roach as perch (8 cases), whitefish (1 cases)
- whitefish as perch (1 cases), roach (3 cases)

This is pretty accurate, but there were misclassifications. Most of the misclassifications occurred among members of the same Genus. In fact, the accuracy or correct classification rate can be computed by summing the misclassified fishes, and diving by the total:

```
> misclass<-1+10+8+1+3+7+1+1#sum of the misclassified fishes
> class<-35+10+48+17+11+14+2#sum along diagonal, correct classifications
> accuracy<-1-(misclass/(misclass+class))
> accuracy

[1] 0.8106509
```

This shows the overall accuracy of 81 percent. This is good, but we used the training set to test the classification accuracy. It would have been better to hold out some fish measurements, develop the LDFA on the training set, and then test the discriminant function accuracy on the held out group of fishes.

0.2.3 Classifying New Cases

We can use the Coefficients of the LDs to predict group membership of a new cases. This woruld be where we could repeat the accuracy analysis above with cases we held in reserve. Or we can predict group membership of unknown cases using the LDFA. In the exmaple here, we measured four new fishes, but we don't know which species they are:

```
> New_fish <- read.csv("C:/Users/Joseph/Dropbox/CRM7008/DFA/New_fish.csv")
> New_fish

  Observation Species Weight.g SLength.cm FLength.cm Tlength.cm Height.Length
1          NA      NA      200         20         22         23            10
2          NA      NA      400         40         44         46            20
3          NA      NA      300         25         26         32            45
4          NA      NA      290         24         26         31            42
  Width.Length Sex
1            8  NA
2           16  NA
3           15  NA
4           14  NA

> fish.id<-predict(fit_lda, newdata=New_fish)
> print(fish.id)
```

```
$class
[1] Pike   Pike   Parkki Bream
Levels: Bream Parkki Perch Pike Roach Smelt Whitefish

$posterior
         Bream       Parkki        Perch         Pike        Roach        Smelt
1 3.064081e-80 3.480926e-85 6.850427e-22 9.953519e-01 7.431312e-24 4.648054e-03
2 4.162263e-52 2.405691e-57 2.250159e-01 7.654556e-01 1.299429e-05 9.515368e-03
3 3.436561e-01 6.563439e-01 1.099678e-44 6.498862e-97 9.522802e-38 2.982452e-75
4 6.256754e-01 3.743246e-01 4.513201e-37 1.392634e-82 1.288274e-30 5.702635e-63
     Whitefish
1 1.003024e-30
2 1.452284e-07
3 2.629412e-33

$x
         LD1        LD2        LD3
1  11.071297 -3.6957718  2.3756820
2   7.738365  1.0696895 -2.7094987
3 -12.341127 -0.5116385  0.2900218
4 -10.512715 -1.1477478  0.6388405
```

In this output, we can see that the new group of fish were classified into the species membership based on the size measurements alone we gave the LDFA. First in "class", there is a listing of the classification of the four new fish The new fishes were identified as: Pike (posterior probability = 0.99), Pike (posterior probability = 0.76), Parkki (posterior probability = 0.66) and Bream (posterior probability = 0.63). The posterior probabilities for correctly classifying each of the four new fish are given in the table "posterior". Finally, there is a listing of the discriminant scores for the four new fish: LD1, LD2, LD3. You can see by examining these scores that the first two would be classified as pike, while the next two work be parkki and bream.