Linear Discriminant Analysis

Dr. J. Kyle Roberts

Southern Methodist University
Simmons School of Education and Human Development
Department of Teaching and Learning

Background for LDA

- LDA is a method for identifying the “classification” of individuals based on a series of explanatory variables.
- For example, suppose we wanted to know how height and weight contribute to the classification of males and females.
- LDA does this by producing a series of $k - 1$ discriminants (we will discuss this more later) where k is the number of groups.
- Some call this “MANOVA turned around.”
- The number of linear discriminant functions is equal to the number of levels minus 1 $(k - 1)$.

Steps in Computing LDA Coefficients

- Calculate the variance/covariance matrix for each group
- Calculate the between and within group variance/covariance matrix for each group
- We then maximize V where:

$$V = \frac{\boldsymbol{a'S_b a}}{\boldsymbol{a'S_w a}}$$

- where $\boldsymbol{S_b}$ is the pooled between group covariance matrix and $\boldsymbol{S_w}$ is the pooled within group covariance matrix.
- In this case, the vector $\boldsymbol{a}$ that maximizes V is solved and we produce an “allocation rule” whereby we can determine the probability of belonging to a given category.

LDA in R

Consider the following dataset:

```
> set.seed(12346)
> life.data <- data.frame(live = factor(rep(0:1,
+     each = 10)), smoke = c(1, 1, 1, 1, 1, 0, 1,
+     0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0), weight = c(rnorm(1
+     230, 35), rnorm(10, 180, 20)), gender = rep(0:1,
+     10))
> head(life.data)
```

```
  live smoke   weight gender
1    0     1 278.7733      0
2    0     1 222.0984      1
3    0     1 261.2810      0
4    0     1 164.9118      1
5    0     1 192.0768      0
6    0     0 306.2575      1
```

Using `lda`

Make sure that you have loaded `library(MASS)`

```
> (m1 <- lda(live ~ smoke + weight + gender, life.data,
+      prior = c(0.5, 0.5)))
```

```
Call:
lda(live ~ smoke + weight + gender, data = life.data, prior = c(0.5,
    0.5))
Prior probabilities of groups:
  0   1
0.5 0.5

Group means:
  smoke   weight gender
0   0.7 236.8091    0.5
1   0.2 179.9384    0.5

Coefficients of linear discriminants:
               LD1
smoke  -1.85411052
weight -0.02590172
gender -0.23047917
```

Strength of the Linear Discriminants

- Recall from the previous analysis:

```
> m1$scaling

              LD1
smoke  -1.85411052
weight -0.02590172
gender -0.23047917
```

- In this instance, we see that "smoke" has the strongest associated weight with the first linear discriminant function.
- Remember that we only have one discriminant function since we are looking at $(k-1)$ functions.

Predictions from `lda`

```
> predict(m1)$posterior

            0           1
1  0.998568301 0.001431699
2  0.972809861 0.027190139
3  0.995764541 0.004235459
4  0.505556781 0.494443219
5  0.760935888 0.239064112
6  0.987372688 0.012627312
7  0.446472732 0.553527268
8  0.461321970 0.538678030
9  0.995414383 0.004585617
10 0.935122429 0.064877571
11 0.007548797 0.992451203
12 0.136903743 0.863096257
13 0.010730006 0.989269994
14 0.033147772 0.966852228
15 0.767760842 0.232239158
16 0.568412785 0.431587215
17 0.005079242 0.994920758
18 0.037425881 0.962574119
19 0.007797718 0.992202282
```

More Predictions from `lda`

```
> predict(m1)$class

 [1] 0 0 0 0 0 0 1 1 0 0 1 1 1 1 0 0 1 1 1 1
Levels: 0 1
```

These are the predicted scores (“0” or “1”) from the posterior weights.

Computing the Discriminant Score

```
> dis.score <- with(life.data, smoke * -1.85411 +
+     weight * -0.0259 + gender * -0.23048)
> cbind(predict(m1)$posterior, dis.score)

             0           1 dis.score
1  0.998568301 0.001431699 -9.074338
2  0.972809861 0.027190139 -7.836937
3  0.995764541 0.004235459 -8.621287
4  0.505556781 0.494443219 -6.355806
5  0.760935888 0.239064112 -6.828900
6  0.987372688 0.012627312 -8.162550
7  0.446472732 0.553527268 -6.256984
8  0.461321970 0.538678030 -6.281837
9  0.995414383 0.004585617 -8.588047
10 0.935122429 0.064877571 -7.458040
11 0.007548797 0.992451203 -4.313801
12 0.136903743 0.863096257 -5.579320
13 0.010730006 0.989269994 -4.461645
14 0.033147772 0.966852228 -4.941131
15 0.767760842 0.232239158 -6.844687
16 0.568412785 0.431587215 -6.461274
17 0.005079242 0.994920758 -4.147689
```

Computing the Huberty I index

- In lieu of a measure of effect size, we can compute the Huberty I index.
- The I is a ratio of the number of people correctly identified by the linear discriminant function relative to the total number of people in the study.
- This can be computed as follows

```
> preds <- predict(m1, method = "plug-in")$class
> table(preds, life.data$live)
```

```
preds 0 1
    0 8 2
    1 2 8
```

```
> 16/20
```

```
[1] 0.8
```

- Therefore, the I index for this linear function would be 0.80.

LDA Homework

In-class assignment for doing an LDA.

- Create a new dataset called `supplemental` in which you add scores for 5 new people on smoke, weight, and gender.
- Use the weights from the previous LDA to predict whether or not they will die before age 60.

	Smoke	Weight	Gender
Person 1	yes	258	F
Person 2	no	187	F
Person 3	yes	187	M
Person 4	no	360	M
Person 5	yes	155	M

- You can use `predict(m1, newdata=supplemental)` to run this new analysis.

Polytomous Outcomes

- Get the dataset at http://faculty.smu.edu/kyler/7314/hsandbeyond.txt

```
> hsb <- read.table("http://faculty.smu.edu/kyler/courses/7314/h
+     header = T)
> names(hsb)

[1] "gradlevl" "truancy"  "gpa"      "parent"
```

- For gradlevl, 1 means they did not graduate from HS, 2 means they graduated from HS, 3 means they graduated and went to college.
- truacny is the average number of times they were absent each 6 weeks
- gpa represents their gpa after grade 10
- parent represents whether or not their parents graduated from HS

Running the LDA for Polytomous Outcomes

```
> (hsbm1 <- lda(gradlevl ~ truancy + gpa + parent,
+     hsb, prior = c(1/3, 1/3, 1/3)))
```

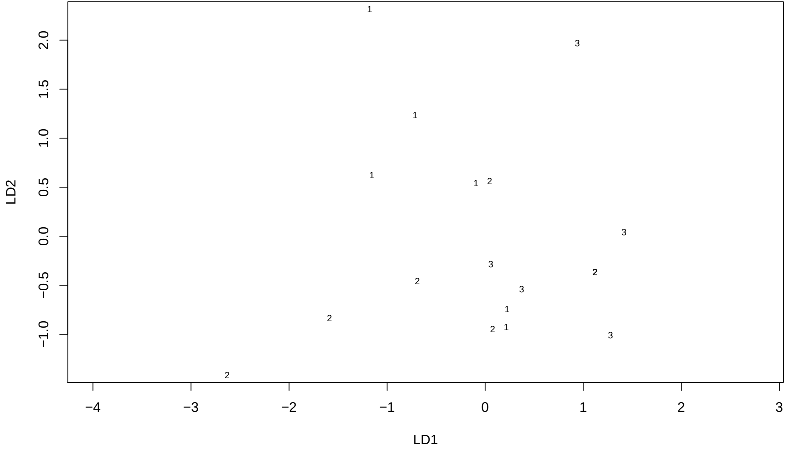
```
Call:
lda(gradlevl ~ truancy + gpa + parent, data = hsb, prior = c(1/3,
    1/3, 1/3))
Prior probabilities of groups:
        1         2         3
0.3333333 0.3333333 0.3333333

Group means:
   truancy  gpa    parent
1 3.333333 3.00 0.3333333
2 2.142857 3.10 0.4285714
3 1.600000 3.68 0.8000000

Coefficients of linear discriminants:
              LD1         LD2
truancy 0.1365433  1.05042667
gpa     1.4751735  2.03953466
parent  1.0634239 -0.07996984
```

Quiz - What does this figure represent?

```
> plot(hsbm1)
```



Computing the Huberty I index

```
> hsbpreds <- predict(hsbm1, method = "plug-in")$class
> table(hsbpreds, hsb$gradlevl)
```

```
hsbpreds 1 2 3
       1 4 1 0
       2 2 4 1
       3 0 2 4
```

- Therefore, the I index for this study would be $12/18{=}0.67$ or 67%.

In Class Assignment with `hsb` data

- In our original study, we looked at the ability of three variables `truacny`, `gpa`, and `parent` in classifying student `gradlevl`.
- What I would like for you to do now is to run three more `lda` analyses in which you have all possible 2-predictor combinations (e.g.,`truancy` and `gpa`; `truancy` and `parent`; `gpa` and `parent`) classifying `gradlevl`
- What (if anything) do you learn from this analysis?