Assignment 7

Instructions: Make an R Notebook and in it answer the question or questions below. When you are done, hand in on Quercus the *output* from Previewing (or Knitting) your Notebook, probably an `html` or `pdf` file. An `html` file is easier for the grader to deal with. Do *not* hand in the Notebook itself. You want to show that you can (i) write code that will answer the questions, (ii) run that code and get some sensible output, (iii) write some words that show you know what is going on and that reflect your conclusions about the data. Your goal is to convince the grader that you *understand* what you are doing: not only doing the right thing, but making it clear that you know *why* it's the right thing.

Do *not* expect to get help on this assignment. The purpose of the assignments is for you to see how much *you* have understood. You will find that you also learn something from grappling with the assignments. The time to get help is after you watch the lectures and work through the problems from PASIAS, via tutorial and the discussion board, that is *before* you start work on the assignment. The only reasons to contact the instructor while working on an assignment are to report (i) something missing like a data file that cannot possibly be read, (ii) something *beyond your control* that makes it impossible to finish the assignment in time after you have started it.

There is a time limit on this assignment (you will see Quercus counting down the time remaining).

1. Earlier, we used MANOVA to investigate three treatments for OCD (obsessive-compulsive disorder). There were two response variables, counts of obsessive Thoughts and obsessive Actions over a certain time period. If you were happy with the assumptions behind the analysis, you found a slightly significant treatment effect, and it is our job here to investigate what kind of differences there are among the treatments. There were 30 observations.

   The data are in https://gaopinghuang0.github.io/assets/Rdata/OCD.dat.

   (a) Read in and display (some of) the data. Reminder: the data values were separated by tabs.

Solution:

`read_tsv`, therefore:

```
my_url <- "https://gaopinghuang0.github.io/assets/Rdata/OCD.dat"
ocd <- read_tsv(my_url)
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   Group = col_character(),
##   Actions = col_double(),
##   Thoughts = col_double()
## )
```

```
ocd
```

```
## # A tibble: 30 x 3
##    Group Actions Thoughts
##    <chr>   <dbl>    <dbl>
##  1 CBT         5       14
##  2 CBT         5       11
##  3 CBT         4       16
##  4 CBT         4       13
##  5 CBT         5       12
##  6 CBT         3       14
##  7 CBT         7       12
##  8 CBT         6       15
##  9 CBT         6       16
## 10 CBT         4       11
## # ... with 20 more rows
```

Thirty rows, ten from each treatment, as before.

Extra: as a reminder, the MANOVA looked like this:

```
ocd %>% select(Actions, Thoughts) %>% as.matrix() -> response
ocd.1 <- manova(response ~ Group, data = ocd)
summary(ocd.1)
```

```
##           Df  Pillai approx F num Df den Df  Pr(>F)
## Group      2 0.31845   2.5567      4     54 0.04904 *
## Residuals 27
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With a treatment effect that is only just significant, we might have a hard time seeing how the treatment groups are different.

(b) Run a discriminant analysis and display the results from it.

Solution:

```
ocd.1 <- lda(Group ~ Actions + Thoughts, data = ocd)
ocd.1
```

```
## Call:
## lda(Group ~ Actions + Thoughts, data = ocd)
##
## Prior probabilities of groups:
##                   BT                  CBT No Treatment Control
##            0.3333333            0.3333333            0.3333333
##
## Group means:
##                       Actions Thoughts
## BT                        3.7     15.2
## CBT                       4.9     13.4
## No Treatment Control      5.0     15.0
##
## Coefficients of linear discriminants:
##                  LD1        LD2
## Actions  -0.6030047 -0.4249451
## Thoughts  0.3352478 -0.3392631
##
```

```
## Proportion of trace:
##    LD1    LD2
## 0.8219 0.1781
```

(c) How many linear discriminants are there? Is that what you were expecting? Explain briefly.

Solution:

There are two response variables and three treatment groups, so the number of linear discriminants should be the smaller of 2 and $3 - 1$, namely 2, which is what happened, so it is what I was expecting.

(d) What would make a person have a *low* score on LD1, in terms of the original variables? Explain briefly.

Solution:

Look at the Coefficients of Linear Discriminants. The one for Actions is negative, and the one for Thoughts is positive. This means that a low score on LD1 would come from a person with a *high* number of obsessive Actions but a low number of obsessive Thoughts.

(e) What does your answer to the previous part suggest about the principal way in which the treatments differ? You might also wish to look at some other part of the output you have obtained so far.

Solution:

LD1 shows the principal way that the treatments differ: that the people in the study tend to have a high number of Actions and a low number of Thoughts (low score on LD1) or a low number of Actions and a high number of Thoughts (high score on LD1), depending on which treatment they got. That is why LD1 came out the way it did; anything else would have not separated out the treatments so well. (As you'll see later, "well" is a relative term here.)

Then you need to say something about *which* treatments will tend to go with high and low scores on LD1. For this, look at the Group Means in the output. BT is low Actions and high Thoughts (at least on average), so it will tend to be at the high end of LD1; CBT is high Actions and low Thoughts, so will tend to be at the low end of LD1.[1] The Control group is high on both; these will cancel each other out on LD1, so we'd expect the Control group people to be somewhere in the middle.

Extra: the coefficients for LD2 are both positive, so that will be large if both Thoughts and Actions are large, as the Control group is. This might, therefore, distinguish the Control group people from the rest. According to Proportion of Trace, though, LD2 doesn't have much to say, so we would probably be wise not to expect too much here.

(f) Obtain the LD1 and LD2 scores for each person, and plot them against each other with the treatments distinguished by colour.

Solution:

A couple of steps here. First, pass the `lda` output into `predict`, which will get "predictions" for everyone in the original dataset, including the LD scores:

```
p <- predict(ocd.1)
glimpse(p)
```

```
## List of 3
##  $ class    : Factor w/ 3 levels "BT","CBT","No Treatment Control": 2 2 1 2 2 1 2 3 3 2 ...
##  $ posterior: num [1:30, 1:3] 0.209 0.0842 0.4877 0.2867 0.1185 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:30] "1" "2" "3" "4" ...
##   .. ..$ : chr [1:3] "BT" "CBT" "No Treatment Control"
##  $ x        : num [1:30, 1:2] -0.46 -1.466 0.813 -0.192 -1.131 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:30] "1" "2" "3" "4" ...
##   .. ..$ : chr [1:2] "LD1" "LD2"
```

This is rather big if you display it all, but `glimpse` gives you a quick overview: the output is a `list`, with three things in it:

- `class`, the best guess at which treatment each person did, given their number of Thoughts and Actions
- `posterior`, the posterior probability of being in each treatment given that person's number of Thoughts and Actions, which gives a sense of uncertainty (we come back to this later)
- `x`, the discriminant scores.

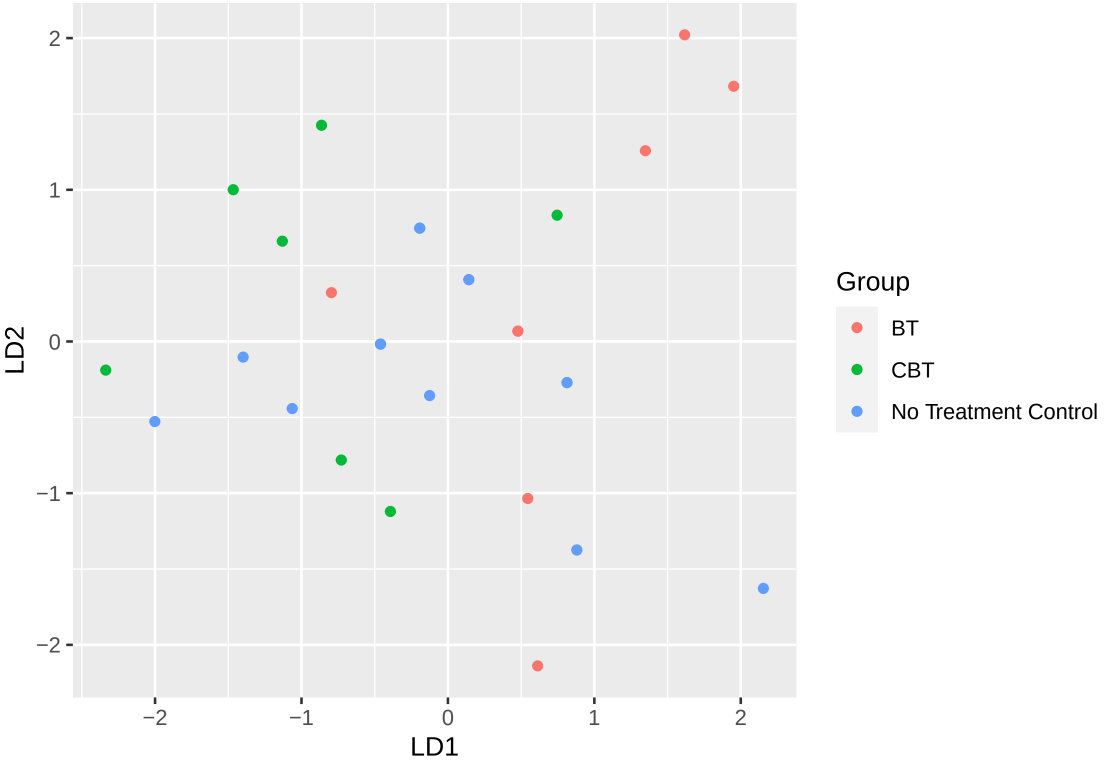
We want to plot `x`, therefore, but these are the discriminant scores only; we need to have the treatments as well, which were in `Group` in the original data frame `ocd`. This is like `augment` in regression, except that this doesn't work here (I tried it). What you do is to `cbind` what you need onto the original dataframe, in the same spirit as putting the predictions next to the values they are predictions for (which is really what we're doing here):

```
cbind(ocd, p$x)
```

```
##                     Group Actions Thoughts        LD1         LD2
## 1                     CBT       5       14 -0.4602010 -0.01736741
## 2                     CBT       5       11 -1.4659443  1.00042182
## 3                     CBT       4       16  0.8132992 -0.27094845
## 4                     CBT       4       13 -0.1924441  0.74684078
## 5                     CBT       5       12 -1.1306965  0.66115874
## 6                     CBT       3       14  0.7458083  0.83252282
## 7                     CBT       7       12 -2.3367058 -0.18873149
## 8                     CBT       6       15 -0.7279579 -0.78157561
## 9                     CBT       6       16 -0.3927101 -1.12083868
## 10                    CBT       4       11 -0.8629396  1.42536694
## 11                     BT       4       14  0.1428037  0.40757770
## 12                     BT       4       15  0.4780514  0.06831463
## 13                     BT       1       13  1.6165699  2.02167613
## 14                     BT       1       14  1.9518176  1.68241305
## 15                     BT       4       15  0.4780514  0.06831463
## 16                     BT       6       19  0.6130332 -2.13862792
## 17                     BT       5       13 -0.7954487  0.32189566
## 18                     BT       5       18  0.8807901 -1.37441972
## 19                     BT       2       14  1.3488130  1.25746794
## 20                     BT       5       17  0.5455423 -1.03515665
## 21 No Treatment Control       4       13 -0.1924441  0.74684078
## 22 No Treatment Control       5       15 -0.1249532 -0.35663049
## 23 No Treatment Control       5       14 -0.4602010 -0.01736741
## 24 No Treatment Control       4       14  0.1428037  0.40757770
## 25 No Treatment Control       6       13 -1.3984534 -0.10304945
## 26 No Treatment Control       4       20  2.1542903 -1.62800076
## 27 No Treatment Control       7       13 -2.0014580 -0.52799457
## 28 No Treatment Control       4       16  0.8132992 -0.27094845
## 29 No Treatment Control       6       14 -1.0632056 -0.44231253
## 30 No Treatment Control       5       18  0.8807901 -1.37441972
```

and then plot that:

```
cbind(ocd, p$x) %>%
  ggplot(aes(x=LD1, y=LD2, colour = Group)) +
  geom_point()
```



Comment on this coming up.

(g) Comment briefly on what you learn from the plot you just made, and how this makes sense (or does not) in the light of what you have seen before in this question.

Solution:

My immediate reaction is that there was no pattern at all (like the similar plot you made when you did a MANOVA on this data set), but there is a little something here. To get a sense of what to expect, go back to your answer about what makes LD1 large or small. This is what I said:

> LD1 shows the principal way that the treatments differ: that the people in the study tend to have a high number of Actions and a low number of Thoughts (low score on LD1) or a low number of Actions and a high number of Thoughts (high score on LD1), depending on which treatment they got. That is why LD1 came out the way it did; anything else would have not separated out the treatments so well. (As you'll see later, "well" is a relative term here.)
>
> Then you need to say something about *which* treatments will tend to go with high and low scores on LD1. For this, look at the Group Means in the output. BT is low Actions and high Thoughts (at least on average), so it will tend to be at the high end of LD1; CBT is high Actions and low Thoughts, so will tend to be at the low end of LD1. The Control group is high on both; these will cancel each other out on LD1, so we'd expect the Control group people to be somewhere in the middle.

So CBT should be at the low end of LD1 (on the left), and BT should be at the high end (right).

Is that what happened? I think the answer to this is somewhere between "kind of" and "not really", it being up to you where you think the graph falls between those. It is true that the green points (CBT) are mostly on the left, and the red points (BT) are mostly on the right, but there is a lot of intermingling of the treatment groups, and the Control group people are literally all over the place.

Extra: I suggested earlier that LD2 might separate the Control group from the rest. That has really not happened, though I suppose it *is* true that the majority of the blue points are in the bottom half of the plot.

(h) Obtain a table of the actual and predicted treatment groups for the 30 people. Are the results in your table consistent with your graph? Explain briefly.

Solution:

A couple of ways to go. One way is to use `table`. This takes two columns, which you can get from the right places using dollar signs:

```
table(ocd$Group, p$class)
```

```
## 
##                         BT CBT No Treatment Control
##   BT                     6   1                    3
##   CBT                    2   6                    2
##   No Treatment Control   3   5                    2
```

This is probably the easiest way. Another more `tidyverse` way involves gluing the two places where these columns are together using `cbind` (as you did for the graph), and then `count`ing:

```
cbind(ocd, class = p$class) %>%
  count(Group, class)
```

```
##                   Group                class n
## 1                    BT                   BT 6
## 2                    BT                  CBT 1
## 3                    BT No Treatment Control 3
## 4                   CBT                   BT 2
## 5                   CBT                  CBT 6
## 6                   CBT No Treatment Control 2
## 7 No Treatment Control                   BT 3
## 8 No Treatment Control                  CBT 5
## 9 No Treatment Control No Treatment Control 2
```

and if you want, you can then go one step further to make it look like the other one:

```
cbind(ocd, class = p$class) %>%
  count(Group, class) %>%
  pivot_wider(names_from = class, values_from = n)
```

```
## # A tibble: 3 x 4
##   Group                   BT   CBT `No Treatment Control`
##   <chr>                <int> <int>                  <int>
## 1 BT                       6     1                      3
## 2 CBT                      2     6                      2
## 3 No Treatment Control     3     5                      2
```

The two actual therapies BT and CBT are actually classified better than I was expecting; in each case 6 out of 10 of them were gotten right. But the Control group people were almost all gotten wrong. This is reasonably consistent with the graph; the BT and CBT people were mostly in one part of the graph (right or left respectively), so they were not confused with each other too much, while the Control people were all over the graph and were very easy to confuse with another group.

The point you need to make here is to show somehow that you know how this table is telling the same story as the plot, and to explain how you know that.

(i) Display the posterior probabilities next to the original data, and explain briefly how these are also consistent with your graph and table of the previous two parts (or are not consistent, if you think that's the case).

Solution:

Again use the `cbind` idea. In here, we don't need all the decimal places of the posterior probabilities; three or four is enough:

```
cbind(ocd, round(p$posterior, 3))
```

```
##                   Group Actions Thoughts    BT   CBT No Treatment Control
## 1                   CBT       5       14 0.209 0.410                0.381
## 2                   CBT       5       11 0.084 0.695                0.221
## 3                   CBT       4       16 0.488 0.172                0.340
## 4                   CBT       4       13 0.287 0.426                0.287
## 5                   CBT       5       12 0.119 0.606                0.275
## 6                   CBT       3       14 0.536 0.231                0.232
## 7                   CBT       7       12 0.029 0.683                0.287
## 8                   CBT       6       15 0.145 0.375                0.480
## 9                   CBT       6       16 0.180 0.290                0.530
## 10                  CBT       4       11 0.163 0.630                0.208
## 11                   BT       4       14 0.356 0.328                0.316
## 12                   BT       4       15 0.424 0.242                0.334
## 13                   BT       1       13 0.788 0.121                0.091
## 14                   BT       1       14 0.835 0.079                0.086
## 15                   BT       4       15 0.424 0.242                0.334
## 16                   BT       6       19 0.292 0.111                0.597
## 17                   BT       5       13 0.161 0.509                0.330
## 18                   BT       5       18 0.414 0.120                0.466
## 19                   BT       2       14 0.708 0.143                0.149
## 20                   BT       5       17 0.366 0.170                0.464
## 21 No Treatment Control       4       13 0.287 0.426                0.287
## 22 No Treatment Control       5       15 0.261 0.317                0.422
## 23 No Treatment Control       5       14 0.209 0.410                0.381
## 24 No Treatment Control       4       14 0.356 0.328                0.316
## 25 No Treatment Control       6       13 0.083 0.564                0.352
## 26 No Treatment Control       4       20 0.675 0.035                0.290
## 27 No Treatment Control       7       13 0.042 0.599                0.360
## 28 No Treatment Control       4       16 0.488 0.172                0.340
## 29 No Treatment Control       6       14 0.112 0.469                0.419
## 30 No Treatment Control       5       18 0.414 0.120                0.466
```

You will probably get the first ten rows and something to click to see the rest. Scroll down through them and see what you see. Note an example or two to support your point.

The picture we have been getting so far is that the groups are not very distinct. That shows up here by the posterior probabilities containing a lot of middling numbers, 0.2 up to 0.6 or so. It is almost never clear which group an observation really belongs to; even observation 2, a CBT that was correctly classified, still has a 30% chance of being something else, and that's one of the best ones. Say something that indicates that it is not clear really which treatment group any observation belongs to; I think it's clearest to use an example or two. (Down at the bottom, most of the control group observations look like one of the real treatments, BT for some and CBT for others; even the ones that were gotten correct, such as #30, are almost as likely to be something else.)

Extra: I wanted to try something else: to re-do the plot of LD scores with some kind of indication of which observations were correctly classified. Let's first gather what we need, rounding off the LD scores since we don't need all those decimals either:

```
cbind(ocd, round(p$x, 4), class = p$class)
```

```
##                   Group Actions Thoughts     LD1     LD2                class
## 1                   CBT       5       14 -0.4602 -0.0174                  CBT
## 2                   CBT       5       11 -1.4659  1.0004                  CBT
## 3                   CBT       4       16  0.8133 -0.2709                   BT
## 4                   CBT       4       13 -0.1924  0.7468                  CBT
## 5                   CBT       5       12 -1.1307  0.6612                  CBT
## 6                   CBT       3       14  0.7458  0.8325                   BT
## 7                   CBT       7       12 -2.3367 -0.1887                  CBT
## 8                   CBT       6       15 -0.7280 -0.7816 No Treatment Control
## 9                   CBT       6       16 -0.3927 -1.1208 No Treatment Control
## 10                  CBT       4       11 -0.8629  1.4254                  CBT
## 11                   BT       4       14  0.1428  0.4076                   BT
## 12                   BT       4       15  0.4781  0.0683                   BT
## 13                   BT       1       13  1.6166  2.0217                   BT
## 14                   BT       1       14  1.9518  1.6824                   BT
## 15                   BT       4       15  0.4781  0.0683                   BT
## 16                   BT       6       19  0.6130 -2.1386 No Treatment Control
## 17                   BT       5       13 -0.7954  0.3219                  CBT
## 18                   BT       5       18  0.8808 -1.3744 No Treatment Control
## 19                   BT       2       14  1.3488  1.2575                   BT
## 20                   BT       5       17  0.5455 -1.0352 No Treatment Control
## 21 No Treatment Control       4       13 -0.1924  0.7468                  CBT
## 22 No Treatment Control       5       15 -0.1250 -0.3566 No Treatment Control
## 23 No Treatment Control       5       14 -0.4602 -0.0174                  CBT
## 24 No Treatment Control       4       14  0.1428  0.4076                   BT
## 25 No Treatment Control       6       13 -1.3985 -0.1030                  CBT
## 26 No Treatment Control       4       20  2.1543 -1.6280                   BT
## 27 No Treatment Control       7       13 -2.0015 -0.5280                  CBT
## 28 No Treatment Control       4       16  0.8133 -0.2709                   BT
## 29 No Treatment Control       6       14 -1.0632 -0.4423                  CBT
## 30 No Treatment Control       5       18  0.8808 -1.3744 No Treatment Control
```

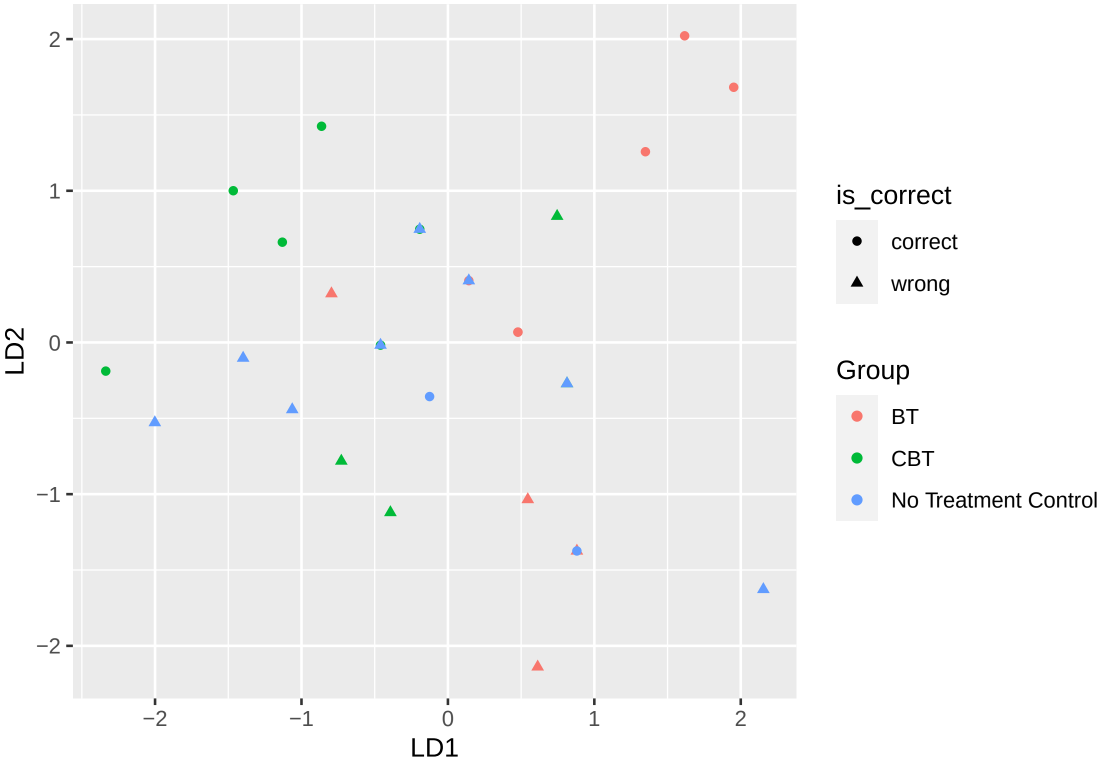
I want to use colour for the true `Group` again, so I need something else to indicate whether or not it was correct. This could be `shape`, shape of the plotting point. I also need to work out which ones actually *were* correct, which I'd better do first:

```r
cbind(ocd, round(p$x, 4), class = p$class) %>%
  mutate(is_correct = ifelse(Group == class, "correct", "wrong"))
```

```
##                   Group Actions Thoughts     LD1     LD2                class is_correct
## 1                   CBT       5       14 -0.4602 -0.0174                  CBT    correct
## 2                   CBT       5       11 -1.4659  1.0004                  CBT    correct
## 3                   CBT       4       16  0.8133 -0.2709                   BT      wrong
## 4                   CBT       4       13 -0.1924  0.7468                  CBT    correct
## 5                   CBT       5       12 -1.1307  0.6612                  CBT    correct
## 6                   CBT       3       14  0.7458  0.8325                   BT      wrong
## 7                   CBT       7       12 -2.3367 -0.1887                  CBT    correct
## 8                   CBT       6       15 -0.7280 -0.7816 No Treatment Control      wrong
## 9                   CBT       6       16 -0.3927 -1.1208 No Treatment Control      wrong
## 10                  CBT       4       11 -0.8629  1.4254                  CBT    correct
## 11                   BT       4       14  0.1428  0.4076                   BT    correct
## 12                   BT       4       15  0.4781  0.0683                   BT    correct
## 13                   BT       1       13  1.6166  2.0217                   BT    correct
## 14                   BT       1       14  1.9518  1.6824                   BT    correct
## 15                   BT       4       15  0.4781  0.0683                   BT    correct
## 16                   BT       6       19  0.6130 -2.1386 No Treatment Control      wrong
## 17                   BT       5       13 -0.7954  0.3219                  CBT      wrong
## 18                   BT       5       18  0.8808 -1.3744 No Treatment Control      wrong
## 19                   BT       2       14  1.3488  1.2575                   BT    correct
## 20                   BT       5       17  0.5455 -1.0352 No Treatment Control      wrong
## 21 No Treatment Control       4       13 -0.1924  0.7468                  CBT      wrong
## 22 No Treatment Control       5       15 -0.1250 -0.3566 No Treatment Control    correct
## 23 No Treatment Control       5       14 -0.4602 -0.0174                  CBT      wrong
## 24 No Treatment Control       4       14  0.1428  0.4076                   BT      wrong
## 25 No Treatment Control       6       13 -1.3985 -0.1030                  CBT      wrong
## 26 No Treatment Control       4       20  2.1543 -1.6280                   BT      wrong
## 27 No Treatment Control       7       13 -2.0015 -0.5280                  CBT      wrong
## 28 No Treatment Control       4       16  0.8133 -0.2709                   BT      wrong
## 29 No Treatment Control       6       14 -1.0632 -0.4423                  CBT      wrong
## 30 No Treatment Control       5       18  0.8808 -1.3744 No Treatment Control    correct
```

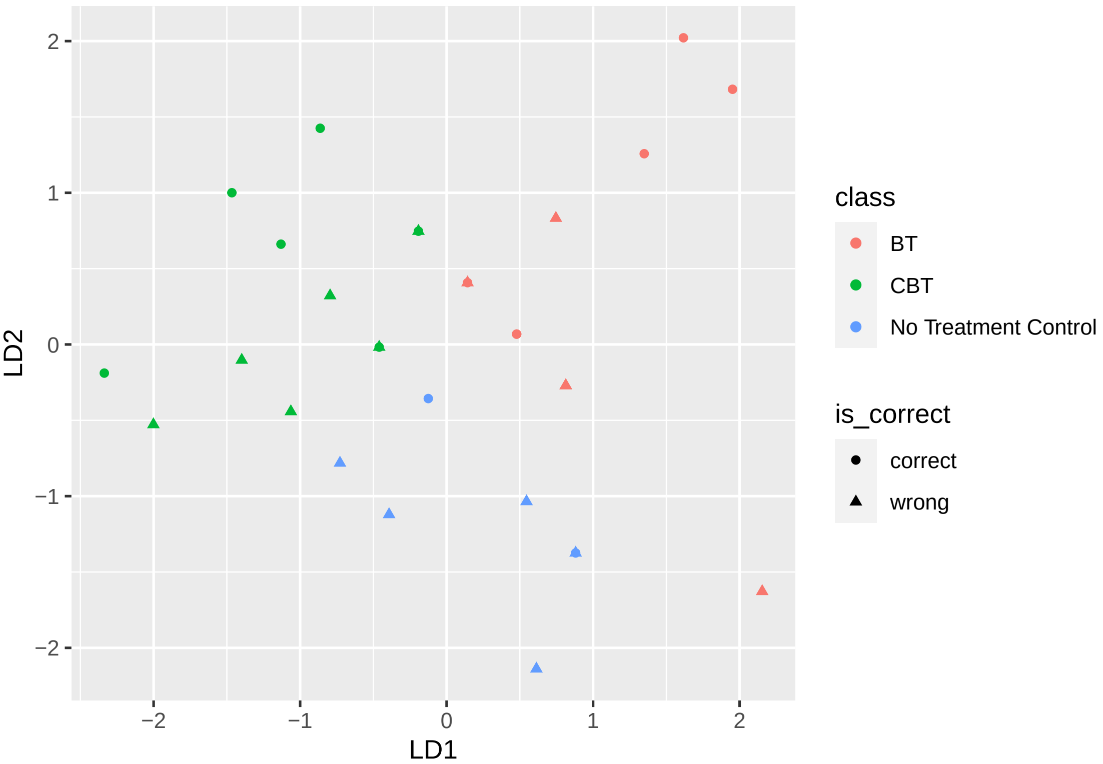
Check a few of those to see whether I got them right. Then:

```r
cbind(ocd, round(p$x, 4), class = p$class) %>%
  mutate(is_correct = ifelse(Group == class, "correct", "wrong")) %>%
  ggplot(aes(x = LD1, y = LD2, colour = Group, shape = is_correct)) + geom_point()
```



The correct ones are the circles, and the wrong ones are the triangles. The correct CBTs are on the top left, the correct BTs are mostly top right, and the correct Control is right in the middle. (There were two correct Controls, but I don't see the other one.) Another way to go at this is to use colour to represent the *predictions*, which is one small change:

```
cbind(ocd, round(p$x, 4), class = p$class) %>%
  mutate(is_correct = ifelse(Group == class, "correct", "wrong")) %>%
  ggplot(aes(x = LD1, y = LD2, colour = class, shape = is_correct)) + geom_point()
```



This shows where an observation would have to be to be predicted as each of the three treatments: on the right for BT, on the left for CBT, and bottom middle for Control. If a point is far enough to the right, it could be predicted as BT even with a negative LD2 score (that point bottom right).

The boundaries between the predicted regions are lines, which is the L of `lda`: “linear”.

I am now thinking about a contour plot or similar that shows the sizes of the posterior probabilities for various LD1 and LD2 scores, but that would require more inspiration than I have at the moment.

Notes

1. The first time I did this, I got it the wrong way around. Happens to us all.