Solution Sheet 5

The full R code will be made available in a separate file on the course homepage.

1. a) A call of the function `plot` shows the projection of the observations onto the discriminant (Figure 1). We can see that Group 3 attains lower values than does Group 4. Both groups can largely be separated, albeit not perfectly.

   b) Figure 2 shows the result. As we can see, groups 2, 3 and 4 can be separated relatively well by the first two discriminants. The first group is comparatively small, and it is difficult to distinguish from the other three. The third discriminant does not seem to aid the classification into these groups.

   c) Table 1 compares the predicted and true group membership for 2 groups. Groups 3 and 4 are easily separated in this manner (with about 95% of all observations correctly classified). Table 2 shows the 4-group classification. While groups 2, 3 and 4 can be easily separated, observations from group 1 are not so frequently recognized as such. This difficulty in classifying Group 1 is one we already saw in the images of the last part of this exercise (Fig. 2).

P \ T	3	4	total
3	28	2	30
4	2	22	24
total	30	24	54

Table 1: Columns correspond to true (T) and rows to predicted (P) group membership. We can see that about 95% of the observations are correctly classified.

P \ T	1	2	3	4	total
1	2	1	0	0	3
2	2	18	0	0	20
3	0	2	30	2	34
4	3	0	0	22	25
total	7	21	30	24	82

Table 2: Columns correspond to true (T) and rows to predicted (P) group membership. The first group is poorly classified, while the other threee are classified very well.

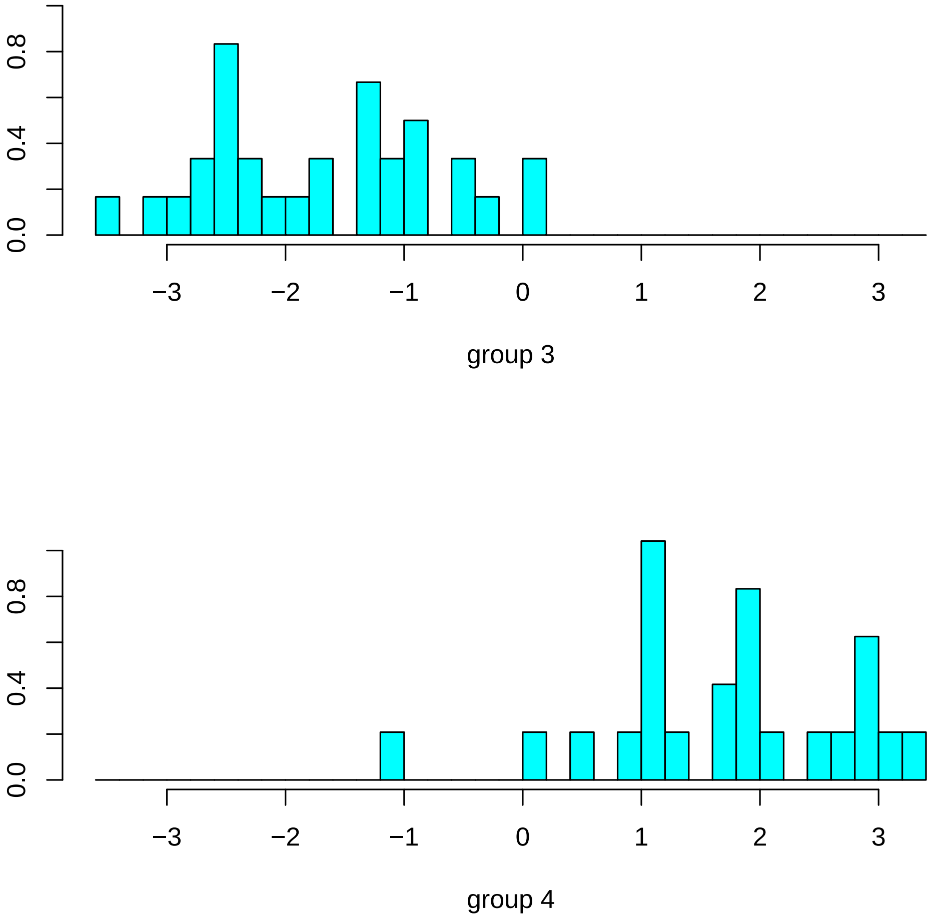


Figure 1: The observations are projected onto the discriminant. We can see that the two groups can largely be separated, albeit not perfectly.

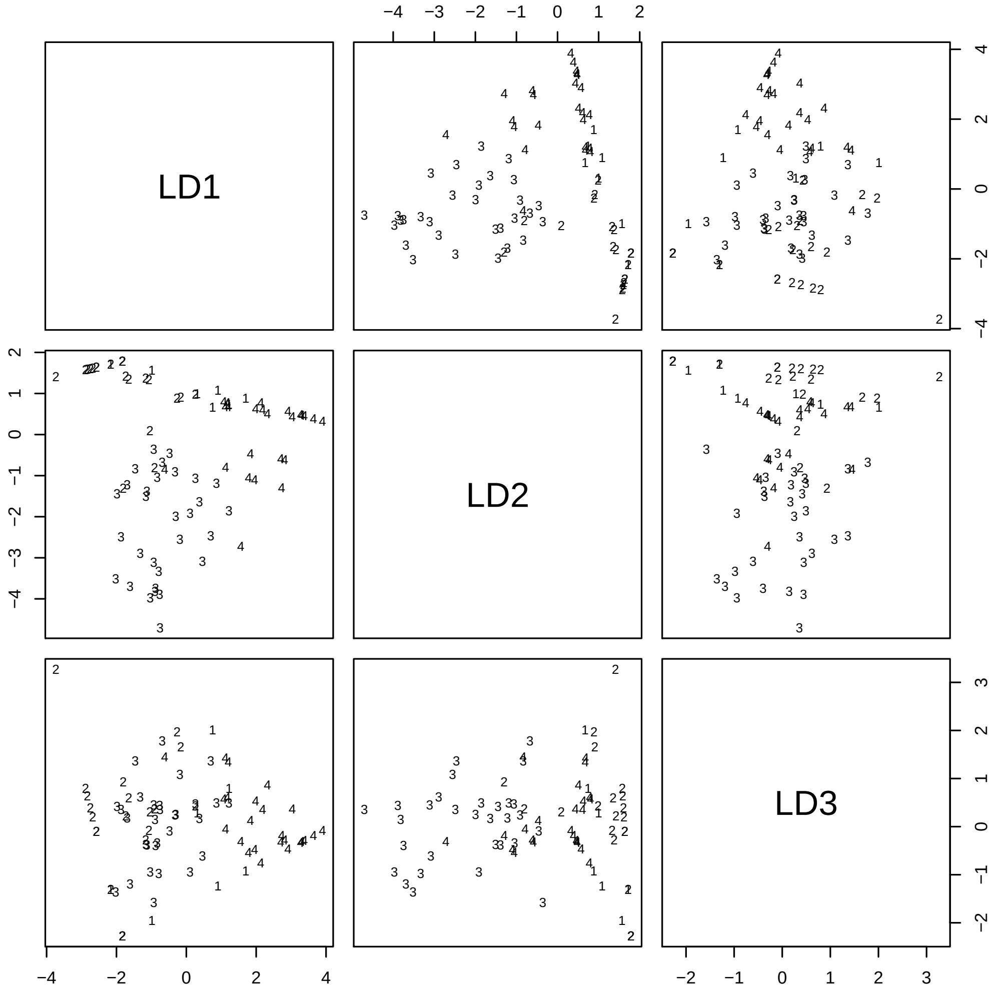


Figure 2: The projections of the observations onto two of the three discriminants.

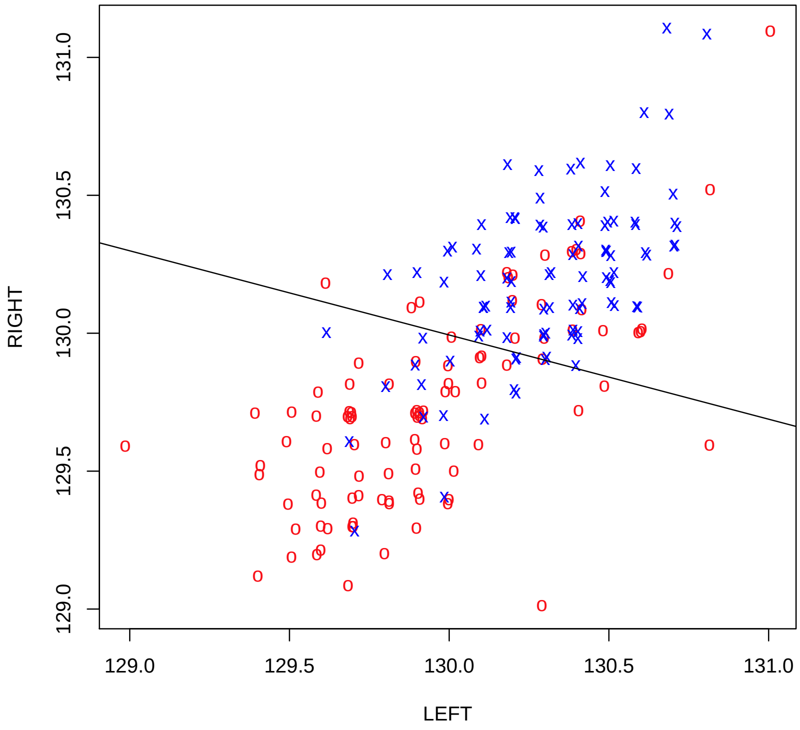
2. **a)** The two groups are not very well seperated by the variables `LEFT` and `RIGHT`, i.e., we have a relatively large overlap.

b)

```
          real
predicted  0  1
         0 74 18
         1 26 82
```

More than 25% of the original bills are falsely classified as forged. Roughly 20% of the forged bills are not recognized as being forged.

c) We solve for x_2, here `RIGHT`, and obtain $x_2 = -\frac{\alpha}{\beta_2} - \frac{\beta_1}{\beta_2}x_1$.



Everything below the line will be classified as original, everything above as forged.

d) The sensitivity is 0.82, the specificity 0.74.

e) With cross-validation:

```
          real
predicted  0  1
         0 73 18
         1 27 82
```

The sensitivity is 0.82, the specificity 0.73.

3. **a)** In lda, the covariance matrices are assumed to be the same for the different groups. Here, judging from the scatter plot in **2.c)**, the covariance matrices are similar in terms of principal directions and variability along these directions, so we do not expect qda to perform much better.

b) The cross-table for qda is (using cross-validation):

```
         real
predicted  0  1
        0 71 15
        1 29 85
```

The sensitivity is 0.85, the specificity 0.71. In comparison to lda, qda gives better sensitivity; the specificity on the other hand is worse for qda.

c) The cross-table for logistic regression is (using cross-validation):

```
         real
predicted  0  1
        0 74 19
        1 26 81
```

The sensitivity is 0.81, the specificity 0.74. The error rates are very similar as for lda.