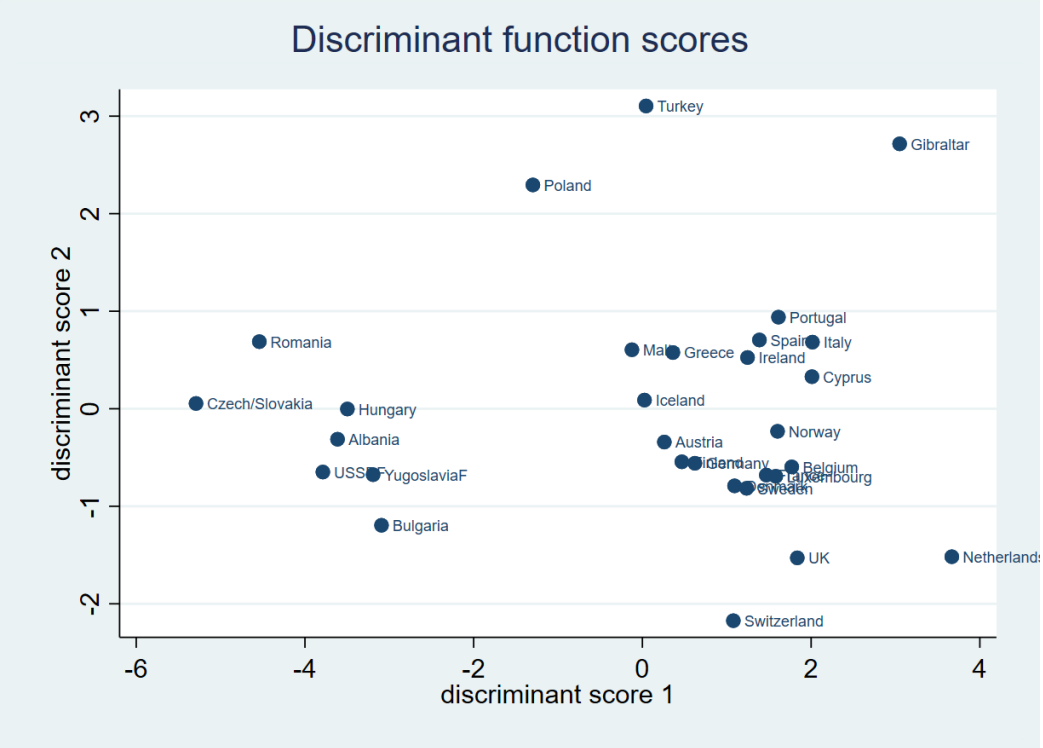
Additional Multivariate Exercise 15

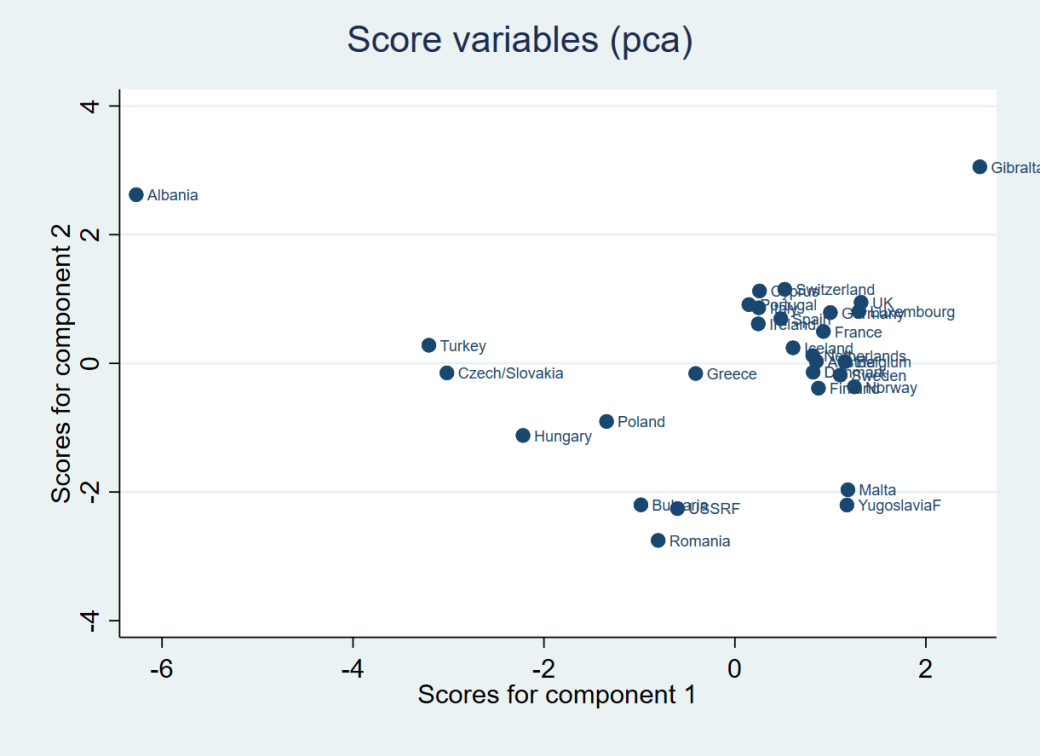
Data: The European employment data were described in the Manly textbook and also previously studied in Exercises 2 and 5.

Linear discriminant analysis: Although Minitab performs the same analysis (for LDA, with uniform priors) as Stata, the output is less useful to compare with the results of the Manly text. Minitab also prints multiple warnings, to an excessive degree when cross-validation is requested (and the results no longer correspond to those in Stata). Because Minitab also does not have the option of changing the uniform priors to for example data priors, its implementation is generally less useful, and will not be discussed further.

As mentioned in the text, due to the linear dependence between the variables (the values add up to 100), one of them needs to be excluded from analysis. Any of them will do, and as in the Manly we choose to exclude the last one (`TC`). With $g = 4$ and $p = 8$ variables, there are $\min(g-1, p) = 3$ canonical discriminant functions, each corresponding to eigenvectors of a specific matrix (detailed in both the Manly text and the lecture). Here the first eigenvalue is by far the largest (5.35). In order to get the values corresponding to the matrix eigenvalue problem displayed in Stata, one needs to use the `candisc` command. (The same results are obtained with the `discrim lda` command, but its output focuses on the classification performance.) Because the third eigenvalue is very small, we focus on the first two functions for our discussion. The graph below shows the values of these, in Stata labelled a score plot. Stata also offers a plot of the coefficients (a "loading plot"), but as discussed in Manly the coefficients seems less useful than the correlations with the original variables (Table 8.5 of Manly). With correlations beyond ± 0.5 of the first variable with `AGR,MIN` (both negative) and `SER,FIN,SPS` (all positive), the first variable can be interpreted as a contrast with employment in service type industries rather than traditional industries. The correlations with the second variable are smaller in magnitude (within ± 0.5), but the most extreme ones seem to represent a contrast between `AGR,CON` and `FIN,TC`. These interpretations are useful when viewing the score plot.



The figure shows the Eastern European countries well separated on the left; only Poland is somewhat close to the Western European countries (which are all clumped together with no real separation between EU and EFTA countries). This tells us that the Eastern European countries are characterized by relatively higher employment in traditional industries compared to service industries than in the Western European countries. The four "Other" countries are also in the Western group on the first variable, whereas two of them, Turkey and Gibraltar, have been separated out on the second variable (due to high values on `AGR` and `CON`, respectively). It seems as a fair conclusion that the separation between Eastern and Western European countries is better than in the corresponding plot from PCA (below), but the biggest difference in the separation of the Eastern European countries is probably that in the PCA score plot also Turkey and Malta (both "Other") are blended with these countries. Generally speaking, one would indeed expect a supervised classification (LDA), where the actual grouping is part of the algorithm, to perform better in terms of separating the groups than unsupervised clustering (from PCA), where the groups were not used at all.



We finally show some classification tables for LDA, with uniform priors (left) and data priors (right), and evaluated based on the data itself (this page) or after leave-one-out cross-validation (next page).

	Classified (uniform priors)				Classified (data priors)				
True Group	EFTA	EU	East	Other	EFTA	EU	East	Other	Total
EFTA	4	2	0	0	4	2	0	0	6
EU	3	9	0	0	1	11	0	0	12
Eastern	0	0	7	1	0	0	7	1	8
Other	1	1	0	2	0	2	0	2	4
Total	8	12	7	3	5	15	7	3	30

	LOO Classified (uniform priors)				LOO Classified (data priors)				
True Group	EFTA	EU	East	Other	EFTA	EU	East	Other	Total
EFTA	4	2	0	0	2	4	0	0	6
EU	3	8	0	1	2	10	0	0	12
Eastern	1	0	6	1	1	0	6	1	8
Other	2	1	1	0	1	2	1	0	4
Total	10	11	7	2	6	16	7	1	30

The misclassification naturally increases with cross-validation, from a low of 20% to a high of 40%. The difference between uniform and data priors is not that big, but the data priors tend to increase the differences between classfication group proportions (as one would expect). With cross-validation neither of the priors perform very well, at only 60% correctly classified countries; one should however remember that the classification problem gets more difficult with a larger number of groups. The comparison value for random choice is here 25% correctly classified units (not 50%, as for two categories). All classifications have trouble to distinguish between EU and EFTA countries, and indeed this is probably the distinction least likely to come out clearly in the employment statistics. Also the "Other" group appears difficult to classify, which is not surprising considering its less homogeneous character.