Additional Multivariate Exercise 8

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

Principal components analysis: The eigenvalues decline slowly, and it takes four components to explain more than 80% of the variation. The Manly text only discusses the first two components, but for this solution we will include the first four components. The eigenvalues and eigenvectors (i.e., loadings for the components) are shown in the tables below. As also noted in the text, the 9th eigenvalue is exactly zero, due to the linear dependence between the variables (their values add up to 100%). In Stata, the 9th component is excluded from the listing.

Eigenanalysis of the Correlation Matrix

Eigenvalue	3.1123	1.8092	1.4962	1.0634	0.7103	0.3113	0.2934	0.2038	0.0000
Proportion	0.346	0.201	0.166	0.118	0.079	0.035	0.033	0.023	0.000
Cumulative	0.346	0.547	0.713	0.831	0.910	0.945	0.977	1.000	1.000

Eigenvectors

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
AGR	-0.511	0.023	-0.279	-0.016	0.024	-0.042	-0.164	-0.540	0.582
MIN	-0.375	-0.000	0.515	-0.114	-0.346	0.199	0.213	0.449	0.419
MAN	0.246	-0.432	-0.502	-0.058	0.234	-0.031	0.236	0.432	0.447
PS	0.316	-0.109	-0.294	-0.023	-0.854	0.206	-0.061	-0.155	0.030
CON	0.222	0.242	0.072	-0.783	-0.062	-0.503	-0.020	-0.031	0.129
SER	0.382	0.408	0.065	-0.169	0.267	0.673	0.175	-0.202	0.245
FIN	0.131	0.553	-0.096	0.489	-0.131	-0.406	0.458	0.027	0.191
SPS	0.428	-0.055	0.360	0.317	0.046	-0.158	-0.621	0.041	0.410
TC	0.205	-0.517	0.413	0.042	0.023	-0.142	0.492	-0.502	0.061

For interpretation of the components, one may display the loadings in a loadings plot. Default plots typically include the first two components, but a scatterplot matrix for any (larger) number of loadings may be used as well. Here we will base our interpretations on the values in the table. Note that components 1, 2 and 4 are shown with reversed sign compared to the textbook; sign switches are immaterial to the results.

1) The first component contrasts `AGR` and `MIN` with the other groups (because their coefficients have opposite signs and are mostly reasonably large). This is not surprising because we already had noted these sectors to include some countries well separated from the rest; we consider this below in the review of the score plot.

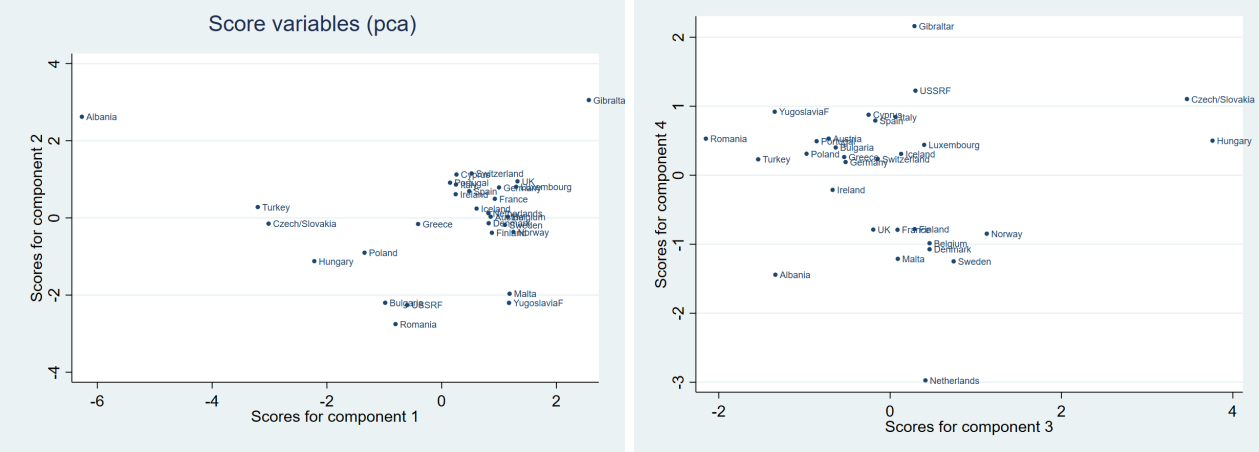
2) The second component contrasts `MAN` and `TC` with mostly `SER` and `FIN` (and to some extent `CON`).

3) The third component contrasts `MAN` (also to some extent `AGR` and `PS`) with `MIN`, `SPS` and `TC`.

4) The fourth component contrasts `CON` with `FIN` and `SPS`.

It is seen that all components have both positive and negative loadings of substantial size. This is not necessarily the case for other datasets.

Score plots in Minitab can be edited to use different symbols for different groups; in this way the four political groups can be added to the figure. There appears however to be no way to display the countries directly on the graph (it can be done as groups, but that is not so helpful with a large number of groups). Score plots in Stata can easily be annotated with subject id's — we show score plots of components 1 and 2 against each other, as well as for components 3 and 4.



The first score variable picks out Albania with a large negative score, and also shows a cluster of mostly Eastern European countries with slightly negative values, contrasting a dense cluster of Western European countries with slightly positive values. The second score variable additionally picks out Gibraltar and some (only in part the same as above) Eastern European countries. Taken together the two scores have put the Western European countries (and Cyprus) in one large cluster with other countries scattered around. The third score variable largely separates Czech/Slovakia and Hungary (due to their zero values for `MAN`) from the other countries, as we also saw with the distance methods. Finally, the fourth score variable distinguishes the Netherlands due to its very low value on `CON`, and also separates a group of (mostly) Scandinavian countries from the rest of the Western European countries. In summary, the four components reflect quite well the patterns we have seen with our previous explorations.

Factor analysis: For the factor analysis, we retain four factors. The unrotated factor loadings in the table below are just rescaled eigenvectors (multiplied by the square-root of eigenvalues), but they are normalized by the sum of the squared loadings being equal to the communality (which is less than 1, or more precisely 1 minus the uniqueness). Except for `PS`, the communalities are high, meaning that the variables are well explained by the four factors. It is worth noting that the 5th principal component has a high loading for this variable, so one would expect this lack of fit to be remedied by including an extra factor.

Unrotated Factor Loadings and Communalities

Variable	Factor1	Factor2	Factor3	Factor4	Communality
AGR	-0.902	0.032	-0.341	-0.017	0.932
MIN	-0.662	-0.001	0.630	-0.117	0.848
MAN	0.434	-0.581	-0.614	-0.060	0.907
PS	0.558	-0.147	-0.359	-0.024	0.462
CON	0.391	0.326	0.087	-0.807	0.918
SER	0.673	0.549	0.080	-0.174	0.791
FIN	0.231	0.744	-0.117	0.504	0.875
SPS	0.755	-0.074	0.441	0.327	0.877
TC	0.362	-0.695	0.505	0.043	0.871
Variance	3.1123	1.8092	1.4962	1.0634	7.4812
% Var	0.346	0.201	0.166	0.118	0.831

Because of the normalization within variables, we can now more easily look at how each variable is accounted for by the factors. Three of the well-explained variables (`AGR`, `CON` and `SPS`) are essentially explained by a single factor (adopting the convention from the text to focus on loadings beyond ± 0.5). The other five variables are essentially explained by two factors. Most of the factors have high loadings on more than two variables, we already discussed this above for the principal components.

We proceed with a varimax rotation with so-called Kaiser normalization, which affects how the rotation is performed (specifically, how the maximization of loadings within each factor is achieved). This is the only option available in Minitab, but the Stata default is without this normalization. The rotated factor loadings are shown below.

Rotated Factor Loadings and Communalities

Varimax Rotation

Variable	Factor1	Factor2	Factor3	Factor4	Communality
AGR	-0.851	0.266	-0.097	0.357	0.932
MIN	-0.108	0.860	-0.296	0.099	0.848
MAN	0.032	-0.892	-0.319	0.093	0.907
PS	0.192	-0.637	0.036	-0.137	0.462
CON	0.018	-0.042	-0.080	-0.954	0.918
SER	0.349	-0.153	0.479	-0.645	0.791
FIN	0.078	-0.004	0.932	0.012	0.875
SPS	0.911	-0.124	0.174	-0.040	0.877
TC	0.726	-0.034	-0.568	0.142	0.871
Variance	2.2593	2.0521	1.6571	1.5126	7.4812
% Var	0.251	0.228	0.184	0.168	0.831

There are indeed fewer large loadings, and all variables except `TC` are now essentially represented by a single factor. We can therefore more easily interpret the factors in terms of the original variables. The interpretations in the Manly text (here modified to account for sign changes, and for a switch between factors 2 and 3) are:

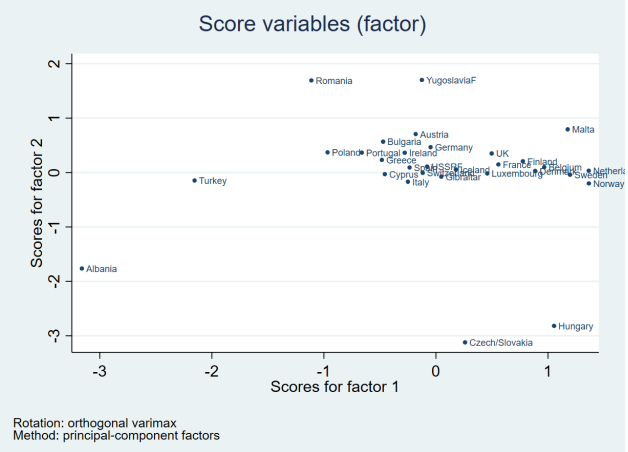
factor 1: social service and communication (`SPS,TC`) rather rural industries (`AGR`),

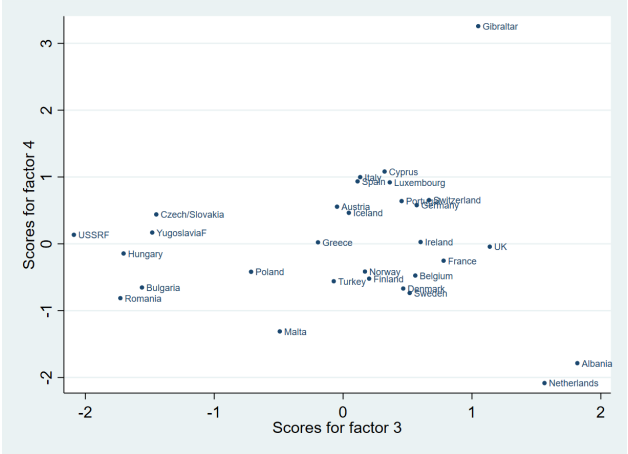
factor 2: mining (`MIN`) rather than manufacturing (`MAN`) and power supplies (`PS`),

factor 3: presence of finance industries (`FIN`), and lack of transport and communication (`TC`),

factor 4: lack of construction (`CON`) and service (`SER`) industries.

Minitab also displays the "Factor score coefficients"; these are the multipliers for the individual variables to obtain the scores for the observations (countries). Because of the rotation, these are no longer simple functions of the factor loadings. In fact, Stata offers two methods to estimate these coefficients; the "regression method" corresponds to the version used in the book. We show again score plots for factors 1 and 2, and for factors 3 and 4.





The plot of the first two scores does a remarkably good job of depicting the patterns and clusters previously seen: distances of Albania, Turkey, Czech/Slovakia and Hungary, from the other countries, even with some separation of the Scandinavian countries (however, no consistent separation of the Eastern European countries). The one missing feature is the separation of Gibraltar, and this is indeed the dominant feature among the scores for factor 4. The third factor separates the large group of Western and Eastern European countries into its expected (by now) clusters.

Further exploration is possible and may be quite interesting, but for the solution we stop here.