Multivariate Statistical Analysis

Pierpaolo De Blasi

University of Torino
email: pierpaolo.deblasi@unito.it
Moodle: math.i-learn.unito.it/msa

Problem set 3

deadline 31 May 2022 11.59pm

Exercise

The data set `swiss` (part of R library `datasets`) gives measures of socio-economic data on 47 Swiss provinces about 1888. Switzerland was entering a period kown as the "demographic transition"; i.e., its fertility was beginning to fall from high level generally found in underdeveloped countries to the lower level it has today. Together with a standardized fertility measure, variable `Fertility`, the data contains the following 5 indicators:

1. `Agriculture` % of males involved in agriculture as occupation
2. `Examination` % draftees receiving highest mark on army examination
3. `Education` % education beyond primary school for draftees
4. `Catholic` % "catholic" (as opposed to "protestant")
5. `Infant.Mortality` % live births who live less than 1 year.

The task is to cluster the observations by considering the 5 indicators only.

1. Obtain a 3 clusters solution via hierarchical clustering with Ward's method. Plot the observations in the space of the first two principal components with colors determined by cluster memberships. Add the projected cluster centroids. Give an interpretation to the clusters. What is the cluster assigned to province "V. De Geneve"?
2. Find a 3 clusters solution with k-means using the centroids found at previous point as initial choice of seeds. Plot the observations in the space of the first two principal components as done in the previous point and compare the k-means cluster solution with the hierarchical one.
3. Compute the average `Fertility` in the three clusters found via k-means. What is the cluster with the highest fertility? Can you make sense of this finding?
4. Consider now model-based clustering with Gaussian mixtures. Use R library `Mclust` to select the best model with 3 groups by BIC method. What model is chosen? Discuss.
5. Compare the k-means cluster solution with the model-based one. Total freedom in this. Can you make sense of the differences?