STAT3019/M019/G019 Exercises 2

You can submit solutions on Moodle until Wednesday 24 January, 13:00, and will get personal feedback.

The workshops will be used for discussion of the Exercises.

1. Let $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_n$ be distributed according to the joint density

$$f(\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_n) = \prod_{i=1}^{n} \varphi(\mathbf{x}_1; \mathbf{a}_{\gamma(i)}, \mathbf{\Sigma}_{\gamma(i)}), \tag{1}$$

with $\gamma : I\!N_n \mapsto I\!N_K$ denoting the true distributions for the n observations, and all parameters unknown. For given $c(1), \ldots, c(n) \in I\!N_K,\ k \in I\!N_K$,
$n_k = \sum_{i=1}^{n} 1(c(i) = k)$ let

$$\mathbf{S}_k = \frac{1}{n_k} \sum_{c(i)=k} (\mathbf{x}_i - \hat{\mathbf{m}}_k^{Km})(\mathbf{x}_i - \hat{\mathbf{m}}_k^{Km})'.$$

Show that $(\hat{\mathbf{m}}_1^{Km}, \mathbf{S}_1), \ldots, (\hat{\mathbf{m}}_K^{Km}, \mathbf{S}_K)$ and $c(1), \ldots, c(n)$ minimising

$$\sum_{k=1}^{K} n_k \log \mathbf{S}_k$$

are ML-estimators for the parameters of model (1).

Note: This defines a clustering method for flexible covariance matrices, although in practice nowadays people use mixture models in this case.

Hint: You can use that for a single multivariate Gaussian model, i.e., (1) with $K = 1$, $(\hat{\mathbf{m}}_1^{1m}, \mathbf{S}_1)$ are ML-estimators for mean and covariance matrix, and also that in this case

$$\sum_{i=1}^{n} (\mathbf{x}_i - \hat{\mathbf{m}}_1^{1m}) \mathbf{S}_1^{-1} (\mathbf{x}_i - \hat{\mathbf{m}}_1^{1m})' = pn.$$

2. Use the gap statistic method to estimate the number of clusters for the olive oil data and for artificial dataset 2. Run this several times to see whether the results (which again depend on initialisation of K-means) are stable.

3. The R-function `clusGap` offers various options, for example `spaceH0="scaledPCA"` (specifying a rotation and transformation of the uniform reference distribution) and `method="firstSEmax"` (see help page).

   How would you design an experiment or a simulation study to find out which of these versions is the best in finding the correct number of clusters?

   Obviously you can run some comparative experiments if you want.

4. Show that Manhattan, Simple Matching and Jaccard distance fulfill the triangle inequality. Find a counter example the shows that the correlation dissimilarity does not fulfill the triangle equality.

   Can the triangle inequality be violated for d_G with continuous values only and d_l the absolute difference (i.e., as in the Manhattan distance) if there are missing values?