Modern Statistics and Big Data Analysis, Exercises 8

Deadline: End of Friday 8 December.

1. (4 points) On Virtuale you find the data set `phonemes1000.dat`. This data set is about speech recognition. It comprises 1000 observations that are log periodograms of spoken phonemes extracted from digital recordings of a number of speakers. These data originally were used for supervised classification, so in fact there is knowledge which phoneme is spoken in which observation. The phonemes are transcribed as follows: "sh" as in "she", "dcl" as in "dark", "iy" as the vowel in "she", "aa" as the vowel in "dark", and "ao" as the first vowel in "water". A periodogram gives the strength of a signal at different frequency levels. For more detail, in case you're interested, see

   ```
   https://en.wikipedia.org/wiki/Periodogram
   https://en.wikipedia.org/wiki/Spectral_density_estimation
   ```

   The data set can be read by

   ```
   phonemes1000 <- read.table("phonemes1000.dat",header=TRUE)
   ```

   Variables 1-256 correspond to the different frequencies, variable 257 (`phonemes1000[,257]`, a string variable) gives the phonemes that can be interpreted as "true clusters" here. This should not be used for any analysis until the end, where you can use it to compare it to your clusterings. Until that point, work with

   ```
   phonemes256 <- as.matrix(phonemes1000[,1:256])
   ```

   The data can be interpreted as functional with the frequencies (variables) taking the role of the time axis. The data set is an example from The Elements of Statistical Learning, obtained from

   ```
   https://web.stanford.edu/~hastie/ElemStatLearn/
   ```

   but cut down to 1000 observations to save you some computing time (the original data set has 4509 observations).

   Represent the data in terms of a suitable B-spline basis. Show how well two exemplary observations are approximated in this way. Run a functional principal component analysis, run a `funFEM`-clustering, and for comparison run a cluster analysis of your choice on the functional principal component scores. Visualise your results suitably, and compare the clustering results with the true phonem classes. You are asked to make your own decisions about tuning choices such as the size of the B-spline basis, number of principal components etc. You are not necessarily asked to compare different choices, but of course you can if you want. It would be nice to give a short reason for your decisions.

2. (2 points) Consider knots at $s_1 = 1,\ s_2 = 2,\ s_3 = 3,\ s_4 = 4$. Define a B-spline B of order $d = 3$ so that $B(1) = 0,\ B(4) = 0,\ B(x) = 0$ for $x \leq 1$ and $x \geq 4$, but $B(x) > 0$ for $x \in (1, 4)$, also $\max_x B(x) = 1$. This means that you are asked to specify the second order polynomials between any two of the knots that define B.

3. (4 points) The R-function `plotcluster` in package `fpc` implements a visualisation of a given classification by a number of projection methods that attempt to show the classes as separated as possible. This can be used in cluster analysis for visualisation of a found clustering; it can improve over principal components in the sense that the data is projected on basis vectors that maximise the clustering information rather than the projected variance.

   The methods are described in the paper
   Hennig, C. (2004) Asymmetric linear dimension reduction for classification. Journal of Computational and Graphical Statistics 13, 930-945,
   which I have put on Virtuale.

   You are asked to focus on two methods (see help page of `plotcluster`): `method="dc"` implements “discriminant coordinates”. They are shortly explained in Section 3 of the paper but go in fact back to Fisher and his LDA as explained on p.322 of the course slides; you may also consider Sec. 4.3.3 of the Hastie, Tibshirani, and Friedman book, available for free download on

   `https://web.stanford.edu/~hastie/ElemStatLearn//`

   `method="awc"` is a method for showing a so called “homogeneous class” separated from a non-homogeneous class. This can be used in cluster analysis to optimally show a single cluster (the number of which can be specified in `plotcluster`) separated from all other clusters, and this can be done for all clusters. This is explained in the paper in Section 4.2 (although it may require to read the material before in order to understand this).

   (a) Explain in your own words what discriminant coordinates (“dc”) and asymmetric weighted discriminant coordinates (“awc”) are, and how they work.

   (b) For the optimal 10-clusters Gaussian mixture clustering of the olive oil data (Example 6.5 on the course slides) show 2-dimensional discriminant coordinates, and asymmetric weighted discriminant coordinates for all clusters[1]. Comment on how these plots compare to the principal components plot in terms of showing the separation of the clusters.

   (c) Considering the phoneme data from question 1 and the `funFEM`-clustering, compare the plot of the first two dimensions of the Fisher discriminating subspace from `funFEM` with what you get when applying discriminant cordinates using `plotcluster` to the data set of the coefficients of the full dimensional B-spline basis and the `funFEM`-clustering.

[1] The plotcluster-function produces 2-dimensional plots, although in principle further dimensions are available, which can be obtained by the discrproj function in fpc.