ASSIGNMENT 1 (STAT3500/7500)

(a) [15 marks]

Consider an observed random sample of size n, $w_1, \ldots, w_n$, from a normal distribution $N(\mu, \sigma^2)$.

To the 75 observations in the dataset Data-A1a.csv apply the EM algorithm to fit via maximum likelihood the two-component normal mixture density with common variances,

$$f(w; \boldsymbol{\Psi}) = \sum_{i=1}^{2} \pi_i \, \phi(w; \mu_i, \sigma^2),$$

where

$$\phi(w; \mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\{-\tfrac{1}{2}(w-\mu)^2/\sigma^2\}$$

and

$$\boldsymbol{\Psi} = (\pi_1, \mu_1, \mu_2, \sigma^2)^T.$$

To this end,

(i) [1/2 mark]

Specify the EM framework.

(ii) [1/2 mark]

Write down the expressions for the E- and M-steps. on the $(k+1)$th iteration of the EM algorithm.

(iii) [3 marks]

Use an available program to fit this mixture model via the EM algorithm such as MClust, FlexMix, and EMMIX, which may be found on CRAN. Explicitly give the starting or starting points tried in your fitting of the EM algorithm and the stopping criterion adopted.

(iv) [3 marks]

Let $\hat{\boldsymbol{\Psi}}$ be the ML estimate of $\boldsymbol{\Psi}$ obtained in (a) above. Plot the fitted two-component normal mixture density $f(w; \hat{\boldsymbol{\Psi}})$ on top of a histogram of the $n = 75$ data points.

Choose the number of bins N for the histogram by consideration of

$$n \approx 2^{N-1}$$

and/or using the formula,

$$\text{bin width} \approx \frac{2 \times \text{Sample IQR}}{n^{1/3}},$$

to guide in the choice of the number of bins N.

(v) [2 marks]

Carry out a chi-squared goodness-of-fit test to assess the adequacy of the fit of the two-component normal mixture model with common variances to the $n = 75$ data points.

(vi) [2 marks]

Fit to this dataset by maximum likelihood via the EM algorithm a two-component normal mixture model with now unequal component variances. Take the component variances to be arbitrary (that is, do not constrain them to be equal now) so that this mixture density is given by

$$f(w; \boldsymbol{\Psi}) = \sum_{i=1}^{2} \pi_i \, \phi(w; \mu_i, \sigma_i^2),$$

where

$$\boldsymbol{\Psi} = (\pi_1, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)^T.$$

(vii) [2 marks]

Use the nonparametric bootstrap to obtain standard errors of the estimates so obtained for the parameters $\pi_1, \mu_1, \mu_2, \sigma_1^2$, and σ_2^2.

(viii) [2 marks]

Use the parametric bootstrap to obtain standard errors of the estimates so obtained for the parameters $\pi_1, \mu_1, \mu_2, \sigma_1^2$, and σ_2^2.

(b) [10 marks]

Consider the dataset Data-A1b.csv with $n = 100$ four-dimensional observations.

(i) [4 marks]

Fit a g-component normal mixture model with a common covariance matrix for its four-dimensional components for $g = 1, g = 2$, and $g = 3$. Plot the clusters obtained for $g = 2$ and $g = 3$ in separate figures, displaying two of the variables at a time in each plot.

(ii) [2 marks]

Carry out a test of exact size 0.05 of the null hypothesis $H_0 : g = 1$ versus $H_1 : g = 2$ using a resampling approach.

(iii) [2 marks]

Use the bootstrap with $B = 99$ bootstrap replications to test the null hypothesis $H_0 : g = 2$ versus $H_1 : g = 3$.

(iv) [2 marks]

Use the Bayesian information criterion (BIC) to decide on the choice between $g = 2$ and $g = 3$ components.