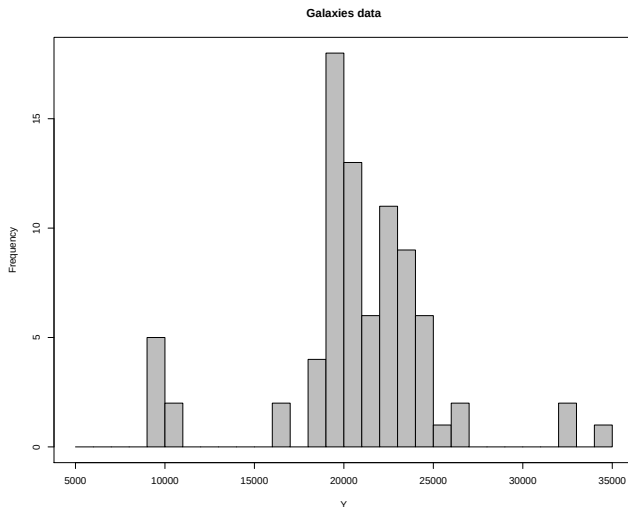


MATH 680: EXERCISE 4

THE EM ALGORITHM

The galaxies data

The data set `galaxies` in the `MASS` library in R contains the velocities in km/sec of 82 galaxies.



Task

1. Implement the mixtures of normals model from the lecture notes for the galaxies data with $K = 3, 4, 5$.
2. Compare the cases $K = 3, 4, 5$ using the BIC, where the favoured model is the one for which the quantity

$$-2\ell_Y(\hat{\theta}; \mathbf{y}) + \log n \times p$$

is smallest, where p is the total number of parameters .

The Student-t distribution

The Student-t location/scale family is based on the density

$$f_Z(z|\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\Gamma\left(\frac{1}{2}\right)} \frac{1}{\nu^{1/2}} \frac{1}{\left(1 + \frac{z^2}{\nu}\right)^{(\nu+1)/2}} \quad z \in \mathbb{R}$$

for degrees of freedom parameter $\nu > 0$, with the location-scale transform

$$y = \sigma z + \theta$$

used to generate general members of the family.

The Student-t distribution

The Student-t density has a *scale-mixture of normals* representation

$$f_Z(z|\nu) = \int_0^\infty f_{Z|X}(z|x) f_X(x|\nu) \, dx$$

where $Z|X = x \sim \text{Normal}(0, 1/x)$,

$$f_{Z|X}(z|x) = \left(\frac{x}{2\pi}\right)^{1/2} \exp\left\{-\frac{x}{2}z^2\right\} \quad z \in \mathbb{R}, x > 0$$

and $X \sim \text{Gamma}(\nu/2, \nu/2)$

$$f_X(x|\nu) = \frac{(\nu/2)^{(\nu/2)}}{\Gamma\left(\frac{\nu}{2}\right)} x^{\nu/2-1} \exp\left\{-\frac{\nu}{2}x\right\} \quad x > 0.$$

This representation facilitates estimation of parameters in the location/scale distribution using the EM algorithm.

Task

1. Write an EM algorithm to perform inference for location and scale parameters for the Student-t model for **fixed** parameter ν , and use it to fit the galaxies data with a single component; try fitting with different values of ν in the range (4,30).
2. Examine what extra steps are necessary if the parameter ν is also the focus of inference, and needs to be estimated using ML.