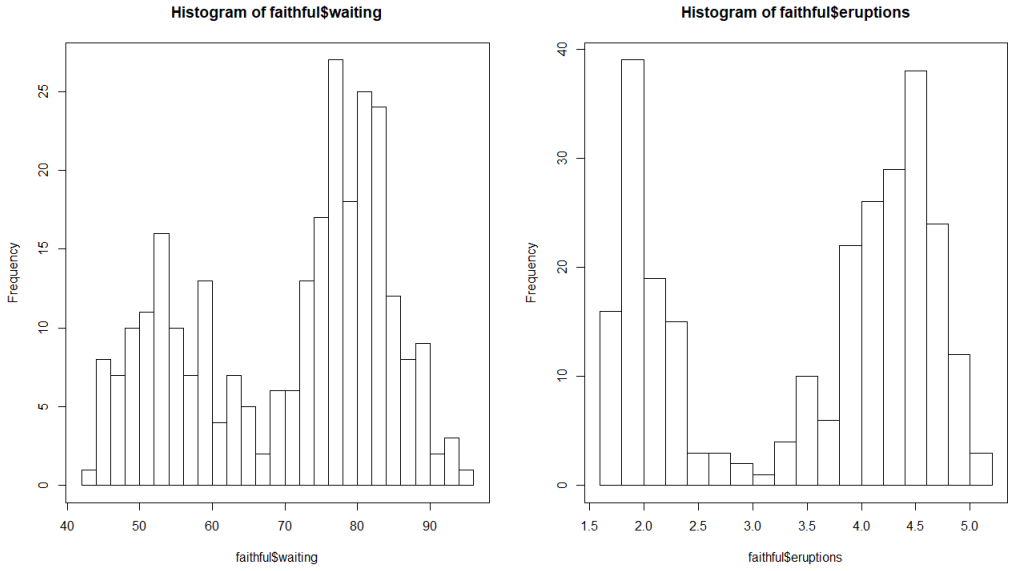
Statistical Computing: Homework 5

Due on May 17 (Wednesday) 23:30

1. The oldfaithfall data collect the waiting time between eruptions and the duration of the eruption for the Old Faithful geyser in Yellowstone National Park, USA. The data can be obtained from R (**faithful**) with 272 observations. In this homework, we ignore the temporal dependence in the data series and treat the 272 observations as independent data with 2 variates. The histogram plots of the two variables show that their joint distribution has more than one mode.

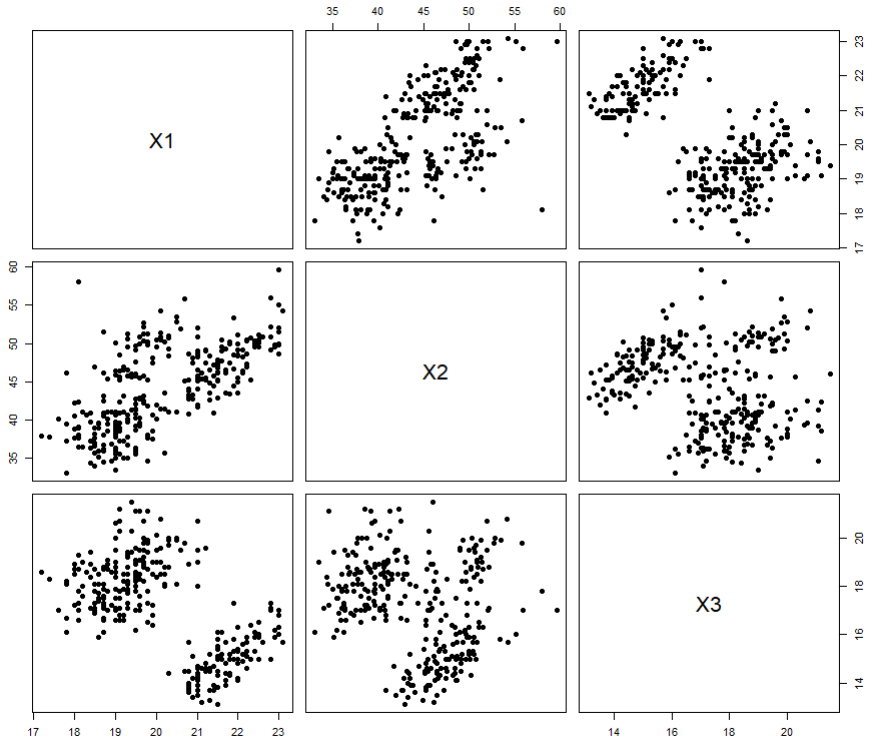


   Assume the 2-dimensional data follow a mixture model with 2 normal components. Apply the EM algorithm to fit the model and estimate the parameters.

   (a) Show your EM work and report your parameter estimates for each normal component.

   (b) Compare your estimated mixture pdf with the data histogram (in 1-dim or 2-dim) and make comments.

2. DataC consists of 3 variables (X_1-X_3). From the pairwise scatter plots, some subgroup (cluster) structures can be seen.



The goal of this problem is to identify possible clusters among data as well as to infer the distribution of each cluster. Such a goal can be achieved via a two-stage approach: first, identify the subgroup structures via usual clustering methods (such as k-means and hierarchical clustering) and then fit the model based on the data for each cluster. Here, you need to take the mixture modeling approach with parameters estimated via EM to achieve the goal.

(a) Fit a **mixture distribution** with k components to these data with 3 variates. Show your EM work and report your fitted model with parameter estimates for each component. Why this k? Justify your choice.

(b) (optional) You may try a two-stage approach. Their results might be helpful for determining k. You may also compare the results here with those obtained from (a). Make your comments.