STAT 6240 - HW 2 - Due 2/11

Due Date: February 11th, Monday, 6 PM

Your homework submission should contain: (i) Your R code, (ii) outputs from R (including figures), (iii) answers to the questions.

You should use *R Markdown* or *R Notebooks*. **Do not print out your homework.** Save it as a pdf (or an html) file and upload it to Blackboard.

In this exercise you will code an E-M Algorithm for fitting data to a mixture of two Poissons.

1. Read the EM Tutorial on Blackboard. It can be found under Outline/Lecture 3 - Clustering.
2. You do not have to include the following derivations in the submitted homework file.

a. Calculate the log of the probability mass function for the Poisson distribution with parameter λ. Note that $\mathbb{P}(X = k|\lambda) = \frac{e^{-\lambda}\lambda^k}{k!}$.
b. For some fixed weights $\pi_{i,k}$ and observations x_i, compute the derivative of the weighted log-likelihood with respect to λ_k. The weighted log-likelihood is given by

$$\sum_{i=1}^{n} \pi_{i,k} \log \mathbb{P}(X = x_i|\lambda_k).$$

c. Set the derivative from the previous step equal to zero and solve for λ_k. This gives you the adjustment for the M-step.

3. Following the steps in the tutorial, write R codes for:

a. The initialization step. Note that the parameters for each cluster, λ_1 and λ_2 have to be positive.
b. The E-step. If you know what you are doing, this step should be straightforward. You just need to use the `dpois` function.
c. The M-step. Use your result from the question 2.

4. Combine all of the code in one big R code. Your code should take in a vector of Poisson observations and return the parameters of the model (λ_1, λ_2) and the expected cluster assignment for each sample $(\pi_{i,k})$.
5. Generate random Poisson samples with the following R code, and test how well your algorithm works in finding the clusters and the cluster parameters:

```
lambda1 <- 10
lambda2 <- 1.2

n1 <- 50
n2 <- 50

x1 <- rpois(n1,lambda1)
x2 <- rpois(n2,lambda2)
x <- c(x1,x2)
```

You should check how close the final estimates are to the actual parameters, and how well the clustering assignment is done. Note that the first 50 observations are from the first cluster and the

last 50 are from the second. If your code does not work properly, or if it returns `NA` values, try different iteration sizes (`max.iter` in the online tutorial) and different initialization methods.

6. Download the `DJI_vol.csv` dataset from Blackboard. In this dataset, you will find the total number of volatile days that the Dow Jones Index had for each month. Here, the number of volatile days is defined as the number of days in which the absolute value of the daily return is higher than 1%.
7. Apply your EM clustering algorithm on the dataset. What are the cluster estimates? Finally, plot the cluster assignments in time. Are there any interesting trends? Summarize your findings in a couple of sentences.