MAST30027: Modern Applied Statistics

Assignment 3, 2023.

Due: 5pm Monday September 25th

- This assignment is worth 14% of your total mark.
- To get full marks, show your working including 1) R commands and outputs you use, 2) mathematics derivation, and 3) rigorous explanation why you reach conclusions or answers. If you just provide final answers, you will get zero mark.
- The assignment you hand in must be typed (except for math formulas), and be submitted using LMS as a single PDF document only (no other formats allowed). For math formulas, you can take a picture of them. Your answers must be clearly numbered and in the same order as the assignment questions.
- The LMS will not accept late submissions. It is your responsibility to ensure that your assignments are submitted correctly and on time, and problems with online submissions are not a valid excuse for submitting a late or incorrect version of an assignment.
- We will mark a selected set of problems. We will select problems worth $\geq 50\%$ of the full marks listed.
- If you need an extension, please contact the lecturer before the due date with appropriate justification and supporting documents. Late assignments will only be accepted if you have obtained an extension from the lecturer before the due date. To ensure that the lecturer responds to your extension request email before the due date, please contact 24h before the due date. Under no circumstances an assignment will be marked if solutions for it have been released.
- Also, please read the "Assessments" section in "Subject Overview" page of the LMS.

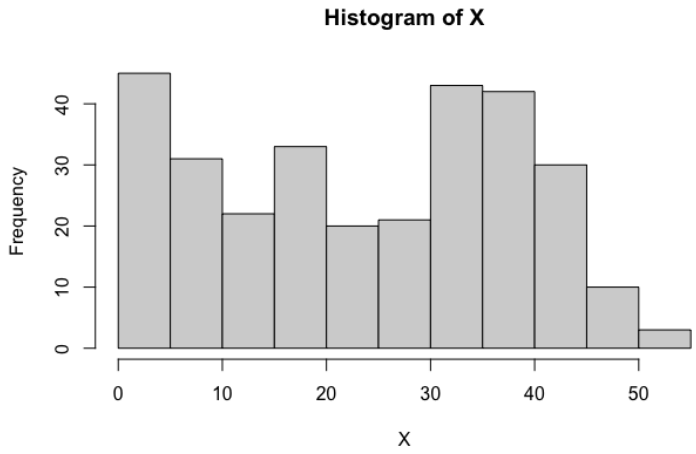
---

1. The file `assignment3_prob1_2023.txt` contains 300 observations. We can read the observations and make a histogram as follows.

```
> X = scan(file="assignment3_prob1_2023.txt", what=double())
Read 300 items
> length(X)
[1] 300
> hist(X)
```

We will model the observed data using a mixture of three Poisson distributions. Specifically, we assume the observations $X_1, \ldots, X_{300}$ are independent to each other, and each X_i follows this mixture model:

$$Z_i \sim \text{categorical } (\pi_1, \pi_2, 1-\pi_1-\pi_2),$$
$$X_i | Z_i = 1 \sim \text{Poisson}(\lambda_1),$$
$$X_i | Z_i = 2 \sim \text{Poisson}(\lambda_2),$$
$$X_i | Z_i = 3 \sim \text{Poisson}(\lambda_3).$$



The Poisson distribution has probability mass function

$$f(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}.$$

We aim to obtain MLE of parameters $\theta = (\pi_1, \pi_2, \lambda_1, \lambda_2, \lambda_3)$ using the EM algorithm.

(a) **(5 marks)** Let $X = (X_1, \ldots, X_{300})$ and $Z = (Z_1, \ldots, Z_{300})$. Derive the expectation of the complete log-likelihood, $Q(\theta, \theta^0) = E_{Z|X,\theta^0}[\log(P(X, Z|\theta))]$.

(b) **(3 marks)** Derive E-step of the EM algorithm.

(c) **(5 marks)** Derive M-step of the EM algorithm.

(d) **(5 marks) Note: Your answer for this problem should be typed. Handwritten solution or screen-captured R codes/figures won't be marked.**

Implement the EM algorithm and obtain MLE of the parameters by applying the implemented algorithm to the observed data, $X_1, \ldots, X_{300}$. Set EM iterations to stop when either the number of EM-iterations reaches 100 (max.iter = 100) or the incomplete log-likelihood has changed by less than 0.00001 ($\epsilon = 0.00001$). Run the EM algorithm two times with the following two different initial values and report estimators with the highest incomplete log-likelihood.

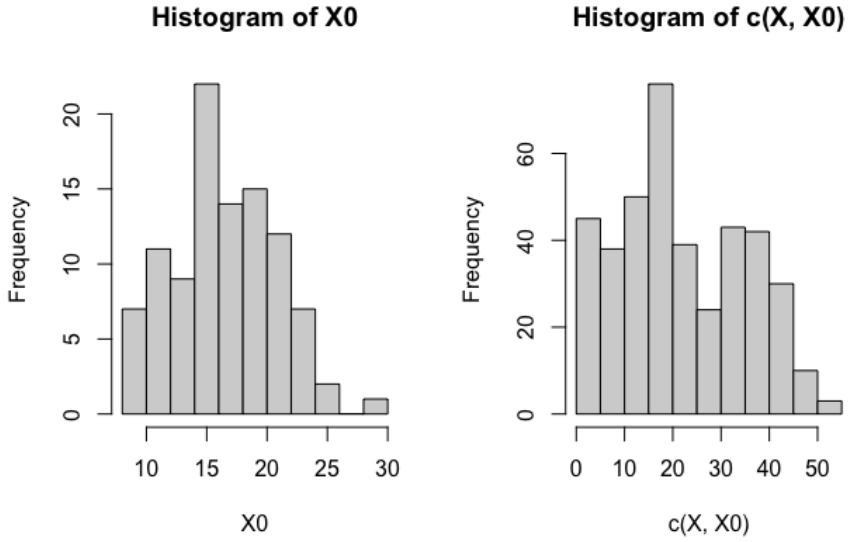
	π_1	π_2	λ_1	λ_2	λ_3
1st initial values	0.3	0.3	3	20	35
2nd initial values	0.1	0.2	5	25	40

For each EM run, check that the incomplete log-likelihoods increase at each EM-step by plotting them.

2. The file `assignment3_prob2_2023.txt` contains 100 observations. We can read the 300 observations from the problem 1 and the new 100 observations and make histograms as follows.

```
> X = scan(file="assignment3_prob1_2023.txt", what=double())
Read 300 items
```

```
> X0 = scan(file="assignment3_prob2_2023.txt", what=double())
Read 100 items
> length(X)
[1] 300
> length(X0)
[1] 100
> par(mfrow=c(1,2))
> hist(X0)
> hist(c(X,X0))
```



Let $X_1, \ldots, X_{300}$ and $X_{301}, \ldots, X_{400}$ denote the 300 observations from `assignment3_prob1_2023.txt` and the 100 observations from `assignment3_prob2_2023.txt`, respectively. We assume the observations $X_1, \ldots, X_{400}$ are independent to each other. We model $X_1, \ldots, X_{300}$ (from `assignment3_prob1_2023.txt`) using the mixture of three Poisson distributions (as we did in the problem 1), but we model $X_{301}, \ldots, X_{400}$ (from `assignment3_prob2_2023.txt`) using one of the three Poisson distributions. Specifically, for $i = 1, \ldots, 300$, X_i follows this mixture model:

$$Z_i \sim \text{categorical } (\pi_1, \pi_2, 1 - \pi_1 - \pi_2),$$
$$X_i | Z_i = 1 \sim \text{Poisson}(\lambda_1),$$
$$X_i | Z_i = 2 \sim \text{Poisson}(\lambda_2),$$
$$X_i | Z_i = 3 \sim \text{Poisson}(\lambda_3),$$

and for $i = 301, \ldots, 400$,

$$X_i \sim \text{Poisson}(\lambda_2).$$

We aim to obtain MLE of parameters $\theta = (\pi_1, \pi_2, \lambda_1, \lambda_2, \lambda_3)$ using the EM algorithm.

(a) **(5 marks)** Let $X = (X_1, \ldots, X_{400})$ and $Z = (Z_1, \ldots, Z_{300})$. Derive the expectation of the complete log-likelihood, $Q(\theta, \theta^0) = E_{Z|X,\theta^0}[\log(P(X, Z|\theta))]$.

(b) **(5 marks)** Derive E-step and M-step of the EM algorithm.

(c) **(5 marks) Note: Your answer for this problem should be typed. Hand-written solution or screen-captured R codes/figures won't be marked.**

Implement the EM algorithm and obtain MLE of the parameters by applying the implemented algorithm to the observed data, $X_1, \ldots, X_{400}$. Set EM iterations to stop when either the number of EM-iterations reaches 100 (max.iter = 100) or the incomplete log-likelihood has changed by less than 0.00001 ($\epsilon = 0.00001$). Run the EM algorithm two times with the following two different initial values and report estimators with the highest incomplete log-likelihood.

	π_1	π_2	λ_1	λ_2	λ_3
1st initial values	0.3	0.3	3	20	35
2nd initial values	0.1	0.2	5	25	40

For each EM run, check that the incomplete log-likelihoods increase at each EM-step by plotting them.