Problem Sheet 04

Alexander Pieper (815402)
alexander.pieper@uni-potsdam.de

21. November 2021

Exercise 1

Let $(x_1, \ldots, x_n) \in (0, \infty)^n$ be a realisation of independent on $[0, \theta]$ uniformly distributed random variables $X_1, \ldots, X_n$. What is Maximum Spacing Estimator in this case? Using the data set provide on Moodle computer the unknown parameter θ via the Maximum spacing estimator for the three different sets of samples (note that they are of different sizes).

Answer:

$$X_1, \ldots, X_n \sim U(0, \theta) \Rightarrow \Theta = \mathbb{R}^+ \tag{1}$$

The Maximum Spacing Estimator is given by (Lecture 8, p.24):

$$\widehat{\theta_{MS}} = \underset{\theta \in \Theta}{\operatorname{argmax}} \prod_{i=1}^{n+1} D_i(\Theta), \tag{2}$$

where $D_i(\Theta) = F_\theta(X_{(i)}) - F_\theta(X_{(i-1)})$. Also we define $X_{(0)} = -\infty \to F_\theta(X_{(0)}) = 0$ and also $X_{(n+1)} = -\infty \to F_\theta(X_{(n+1)}) = 1$

The Distribution function of our uniform distribution is given as

$$F(t) = \begin{cases} 0 & \text{, for } t < 0 \\ \dfrac{t}{\theta} & \text{, for } t \in [0, \theta] \\ 1 & \text{, for } t > \theta. \end{cases} \tag{3}$$

Since all $X_{(i)} \in [0, \theta]$ we have $F_\theta(X_{(i)}) = \frac{X_{(i)}}{\theta}$ and thus

$$\widehat{\theta_{MS}} = \underset{\theta \in \Theta}{\operatorname{argmax}} \prod_{i=1}^{n+1} F_\theta\left(X_{(i)}\right) - F_\theta\left(X_{(i-1)}\right) \tag{4}$$

$$= \underset{\theta \in \Theta}{\operatorname{argmax}} log\left(\prod_{i=1}^{n+1} F_\theta\left(X_{(i)}\right) - F_\theta\left(X_{(i-1)}\right)\right) \qquad \text{same argmax} \tag{5}$$

$$= \underset{\theta \in \Theta}{\operatorname{argmax}} \sum_{i=1}^{n+1} log\left(F_\theta\left(X_{(i)}\right) - F_\theta\left(X_{(i-1)}\right)\right) \qquad \text{property of log()} \tag{6}$$

$$= \underset{\theta \in \Theta}{\operatorname{argmax}} \sum_{i=1}^{n} log\left(F_\theta\left(X_{(i)}\right) - F_\theta\left(X_{(i-1)}\right)\right) + log\left(F_\theta\left(X_{(n+1)}\right) - F_\theta\left(X_{(n)}\right)\right) \qquad \text{drag out last element} \tag{7}$$

$$= \underset{\theta \in \Theta}{\operatorname{argmax}} \sum_{i=1}^{n} log\left(\frac{X_{(i)}}{\theta} - \frac{X_{(i-1)}}{\theta}\right) + log\left(1 - \frac{X_{(n)}}{\theta}\right) \qquad F_\theta(X_{(n+1)}) = 1 \tag{8}$$

$$= \underset{\theta \in \Theta}{\operatorname{argmax}} \sum_{i=1}^{n} log\left(\frac{X_{(i)} - X_{(i-1)}}{\theta}\right) + log\left(1 - \frac{X_{(n)}}{\theta}\right) \tag{9}$$

To get the argmax, we calculate the first derivative w.r.t θ.

$$\frac{\partial}{\partial\theta}\sum_{i=1}^{n} log\left(\frac{X_{(i)}-X_{(i-1)}}{\theta}\right)+log\left(1-\frac{X_{(n)}}{\theta}\right)=\sum_{i=1}^{n}\left[-\frac{1}{\theta}\right]+\frac{X_{(n)}}{(\theta-X_{(n)})\theta} \tag{10}$$

$$=-\frac{n}{\theta}+\frac{X_{(n)}}{(\theta-X_{(n)})\theta} \tag{11}$$

Setting it equal to 0 and solving it for θ.

$$-\frac{n}{\theta}+\frac{X_{(n)}}{(\theta-X_{(n)})\theta}\overset{!}{=}0 \tag{12}$$

$$\Leftrightarrow \frac{n}{\theta}=\frac{X_{(n)}}{(\theta-X_{(n)})\theta} \tag{13}$$

$$\Leftrightarrow n=\frac{X_{(n)}}{(\theta-X_{(n)})} \qquad \theta>0 \tag{14}$$

$$\Leftrightarrow n(\theta-X_{(n)})=X_{(n)} \tag{15}$$

$$\Leftrightarrow \theta=\frac{X_{(n)}}{n}+X_{(n)} \tag{16}$$

$$\Rightarrow \widehat{\theta_{MS}}=\frac{n+1}{n}X_{(n)} \tag{17}$$

Now we have to verify that this is indeed a maximum. To do so, we compute the 2nd derivative and show that it it < 0, at least at $\widehat{\theta_{MS}}$.

$$\frac{\partial^2}{\partial\theta^2}\sum_{i=1}^{n} log\left(\frac{X_{(i)}-X_{(i-1)}}{\theta}\right)+log\left(1-\frac{X_{(n)}}{\theta}\right)=-\frac{n}{\theta^2}+\frac{X_{(n)}(X_{(n)}-2\theta)}{\theta^2\left(\theta-X_{(n)}\right)^2} \tag{18}$$

$$=-\frac{n}{\frac{(n+1)^2}{n^2}X_{(n)}^2}+\frac{X_{(n)}(X_{(n)}-2\frac{n+1}{n}X_{(n)})}{\frac{(n+1)^2}{n^2}X_{(n)}^2\left(\frac{n+1}{n}X_{(n)}-X_{(n)}\right)^2} \quad \text{using} \widehat{\theta_{MS}} \text{ as } \theta \tag{19}$$

$$=-\frac{n}{\frac{(n+1)^2}{n^2}X_{(n)}^2}+\frac{-\frac{n+1}{n}X_{(n)}^2}{\frac{(n+1)^2}{n^2}X_{(n)}^2\left(\frac{n+1}{n}X_{(n)}-X_{(n)}\right)^2} \tag{20}$$

$$=-\frac{n^3}{(n+1)^2X_{(n)}^2}+\frac{-X_{(n)}^2}{\frac{n+1}{n}X_{(n)}^2\left(\frac{1}{n}X_{(n)}\right)^2} \tag{21}$$

$$=-\frac{n^3}{(n+1)^2X_{(n)}^2}+\frac{-X_{(n)}^2}{\frac{n+1}{n^3}X_{(n)}^4} \tag{22}$$

$$=-\frac{n^3}{(n+1)^2X_{(n)}^2}-\frac{1}{\frac{n+1}{n^3}X_{(n)}^2} \tag{23}$$

$$=-\frac{n^3}{(n+1)X_{(n)}^2}-\frac{n^3}{(n+1)X_{(n)}^2} \tag{24}$$

$$=-\frac{2n^3}{(n+1)X_{(n)}^2}<0, \text{ since } n>0 \& X_{(n)}>0 \tag{25}$$

Hence, $\widehat{\theta_{MS}}=\frac{n+1}{n}X_{(n)}$ is a global maximum

Interpretation: The maximum spacing estimator for θ is a bit lager than the largest value in the sampleset. The Maximum Likelihood estimator of the same Problem is $\widehat{\theta_{ML}}=X_{(n)}$ [Ghosha & Jammalamadakab, 2000, p.74]. Therefore the difference between these two estimators is only the correction term $\frac{n+1}{n}$ in the MS Estimator.

Personally, I think in this case, this correction factor maxes a lot of sense, because when you only have very few samples (say 3), it is very likely that the true parameter θ is not the largest value, but a bit larger than that. Please see the Jupyter notebook in the next page for a visual interpretation on this estimator.

In [ ]:
```
import matplotlib.pyplot as plt
```

Firstly, we read the sampleset as lists of observations

In [ ]:
```
with open('sampleset_1_problemsheet4_ex1.txt', 'r') as handle:
    sampleset_1 = handle.readlines()
    sampleset_1 = [float(i.split('\n')[0]) for i in sampleset_1]
with open('sampleset_2_problemsheet4_ex1.txt', 'r') as handle:
    sampleset_2 = handle.readlines()
    sampleset_2 = [float(i.split('\n')[0]) for i in sampleset_2]
with open('sampleset_3_problemsheet4_ex1.txt', 'r') as handle:
    sampleset_3 = handle.readlines()
    sampleset_3 = [float(i.split('\n')[0]) for i in sampleset_3]
```

Now we need to calculate the $\widehat{\theta_{MS}} = \frac{n+1}{n} X_{(n)}$ for each of the datasets.

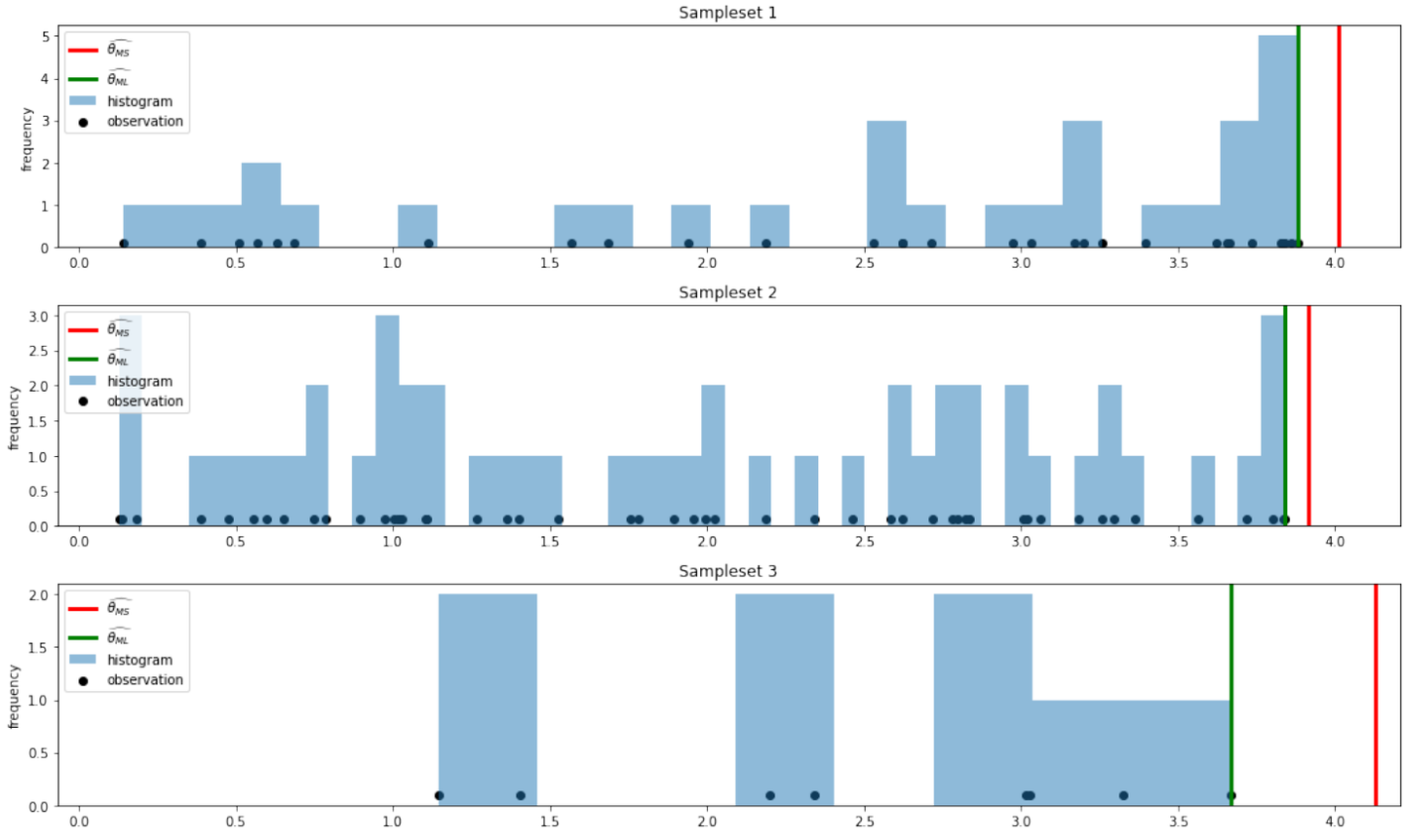
In [ ]:
```
theta_hats_ms = {}
theta_hats_ml = {}
i = 1
for set in [sampleset_1, sampleset_2, sampleset_3]:
    theta_hats_ms['sampleset_' + str(i)] = (len(set) + 1)/(len(set)) * sorted(set)[-1]
    theta_hats_ml['sampleset_' + str(i)] = sorted(set)[-1]
    i += 1
```

In [ ]:
```
print(f'The estimated Theta via Maximum Spacing method of the first sampleset is: {theta_hats_ms["sampleset_1"]}')
print(f'The estimated Theta via Maximum Spacing method of the second sampleset is: {theta_hats_ms["sampleset_2"]}')
print(f'The estimated Theta via Maximum Spacing method of the third sampleset is: {theta_hats_ms["sampleset_3"]}')
```

```
The estimated Theta via Maximum Spacing method of the first sampleset is: 4.0118133333333335
The estimated Theta via Maximum Spacing method of the second sampleset is: 3.91578
The estimated Theta via Maximum Spacing method of the third sampleset is: 4.1274
```

Visualisation

In [ ]:
```
fig, ax = plt.subplots(3,1,figsize = (15,9), sharex=True)
i = 0
for set in [sampleset_1, sampleset_2, sampleset_3]:
    ax[i].hist(set, bins = len(set), label = 'histogram', alpha = 0.5)
    ax[i].scatter(set, [0.1 for i in set], color = 'black', label = 'observation')
    ax[i].axvline(theta_hats_ms['sampleset_' + str(i+1)], color = 'red', label = r'$\widehat{\theta_{MS}}$', linewidth = 3)
    ax[i].axvline(theta_hats_ml['sampleset_' + str(i+1)], color = 'green', label = r'$\widehat{\theta_{ML}}$', linewidth =
    ax[i].set(title = 'Sampleset ' + str(i+1), ylabel = 'frequency')
    ax[i].tick_params(labelbottom=True)
    ax[i].legend(loc = 'upper left')
    i += 1
fig.tight_layout()
plt.show()
```



Analysis

The estimated parameters θ are relatively close to 4, for all three sample sets. Looking at the Histograms and the observations of each Sampleset, we can also see this property of the Maximum spacing Estimator, that the smaller the size n, the further away is θ from the largest value of the set $X_{(n)}$.

The MLE, which is $X_{(n)}$ for this parameter, would not take the sample size into consideration at all. (Ghosha & Jammalamadakab, p.74)

References

[Ghosha & Jammalamadakab, 2000] Kaushik Ghosha, S. Rao Jammalamadakab, A general estimation method using spacings, (2000), Elsevier - Journal of statistical planning and inference 93 (2001) 71–82