EE 5111: Estimation

Jan - May 2021

Mini Project 2

March 22, 2021

1 Performance of MLE

In the following exercises, we evaluate the performance of MLE for the estimation of certain parameters of three different distributions which are of practical interest.

1. Consider the following model:

$$x_n \sim \text{Weibull}(k, \theta) \qquad n = 1, ..., N, \tag{1}$$

where k and θ are the shape and scale parameter of the Weibull RV with the probability density function (p.d.f.) given in (2).

$$f(x) = \begin{cases} \frac{k}{\theta}\left(\frac{x}{\theta}\right)^{k-1} e^{-(x/\theta)^k} & x \geq 0 \\ 0 & x < 0. \end{cases} \tag{2}$$

Compute the MLE of the parameter θ for $k = 1$ and $N = 1, 10, 10^2, 10^3, 10^4$. Generate data using $\theta = 2$.

2. The Gumbel distribution (Type I extreme value distribution) is widely used for characterising the distribution of the maximum of sequences of appropriately normalised independent and identically distributed random variables. Generate RVs $Y_n, n = 1, \cdots, N$ where $Y_n = \max\{X_{n1}, \cdots, X_{nK}\}$ and $X_{nk} \sim \exp(\lambda)$. Here, $\exp(\lambda)$ represents the exponential distribution with mean $\frac{1}{\lambda}$. Then for K *large*, the RV Y_n follows the Gumbel distribution whose p.d.f. is given by

$$f_Y(y) = \frac{1}{\sigma} \exp\left(\frac{-(y-\mu)}{\sigma}\right) \exp\left(-\exp\left(\frac{-(y-\mu)}{\sigma}\right)\right), \tag{3}$$

where $\mu = 14.9787$ and $\sigma = 5$ are the location parameter and the scale parameter respectively. For $\lambda = 0.2$, $K = 20$ generate samples of the form Y_n. Assuming that you know the values of μ, estimate the MLE of σ ($\hat{\sigma}_{gev}$) using only the samples $\{Y_n, n = 1, \cdots, N\}$.

3. The famous Pickands-Balkema-de Haan theorem proves that for any RV X_{nk} in the previous question, and a high threshold d ($d \to F^{-1}_{X_{nk}}(1)$), we can approximately characterize the excess value $\{X_{nk} - d | X_{nk} - d > 0\}$ by a generalized Pareto distribution (GPD) with

parameter σ which is same as the scale parameter in Question 2[1]. Let $\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{I}_{X_{nk}>d} = L$ be the number of RVs X_{nk} satisfying th required condition[2]. The p.d.f. of the GPD is given by

$$f_S(s) = \sigma^{-1}\exp\left(\frac{-s}{\sigma}\right). \tag{4}$$

Using the samples $\{X_{nk}; 1 \le n \le N, 1 \le k \le K\}$ where $X_{nk} \sim \exp(\lambda = 0.2)$, generate RVs[3] $\{S_\ell; 1 \le \ell \le L\}$ for $d = 23$ and compute the MLE of the GPD parameter σ ($\hat{\sigma}_{gpd}$). Compare this with the estimate of σ obtained from the previous question.

Present the following results for each of the above experiments[4].

(a) Tabulate the values of $\mathbb{E}[\hat{\theta}]$, $\mathbb{E}[\hat{\sigma}_{gev}]$ and $\mathbb{E}[\hat{\sigma}_{gpd}]$ against the number of samples N. What do you infer?

(b) Tabulate the values of $\mathrm{Var}(\hat{\theta})$, $\mathrm{Var}(\hat{\sigma}_{gev})$ and $\mathrm{Var}(\hat{\sigma}_{gpd})$ against the number of samples N. What do you infer?

(c) Plot the CDF of the estimates for $N = 1, 10, 100, 1000, 10000$ samples. Ensure that you take enough realizations to get a smooth CDF. What can you say about the CDF? Justify. Do you observe the following relations?

$$\sqrt{N}(\hat{\theta} - \theta) \sim \mathcal{N}(0, I(\theta)^{-1})$$

$$\sqrt{N}\,(\hat{\sigma}_{gev} - \sigma) \sim \mathcal{N}(0, I(\sigma)^{-1})$$

$$\sqrt{L}\,(\hat{\sigma}_{gpd} - \sigma) \sim \mathcal{N}(0, I(\sigma)^{-1})$$

Here, I denotes the Fisher information[5].

(d) Plot the PDF of the estimates for $N = 1, 10, 100, 1000, 10000$ samples. Ensure that you take enough realizations to get a smooth PDF. What can you say about the PDF convergence?

Submission

You are required to submit the solutions for this problem no later than March 23, 2021 (11:59 pm). Upload a compressed file containing the program/programs your four member team has written for the mini project along with a 1-2 page report in Moodle. The report should include the final expressions for the MLE and Fisher information you have used in all the three cases, the final plots for each of the questions (1)-(5) and a one-two line inference on the results observed. Note that you need not include any derivations in this report. The viva for each team will be conducted jointly on a date and time convenient for all the members.

[1] Here, $A_k|b_k > 0$ represents the RV A_k given the condition $b_k > 0$ is satisfied.

[2] Here $\mathbb{I}_E$ is the indicator function for the event E.

[3] Note that here you are asked to use all the samples satisfying $X_{nk} > d$ from the $K \times N$ samples you had used for generating maxima samples in the previous question.

[4] Whenever the closed form solution for MLE is not available, use Newton Raphson or any other appropriate numerical method and mention the same in your report.

[5] Feel free to use any numerical integration package to evaluate the Fisher information if you could not derive the closed form expression for the same.