Mini Project 2 report

C Gayathri, Gokul Mohanraj, Shikhar Prakash, Saurabh Vaishampayan

March 23, 2021

1 Question 1

When k = 1, Weibull distribution reduces to an exponential distribution:

$$f(x) = \begin{cases} \frac{1}{\theta} exp^{-\frac{x}{\theta}} & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{1}$$

To calculate MLE:

$$\implies \hat{\theta} = \frac{\sum x_i}{N} \tag{2}$$

Results obtained:

N=1:MLE = 0.3396115561765516

N=10:MLE = 2.6519381627234724

N=100:MLE = 1.8841639212200298

N=1000:MLE = 2.037913965391342

N=10000:MLE = 2.013999435429514

Thus one can observe that as N increases, the estimate becomes closer to the real value.

2 Question 2

Here we are given the Gumbel Distribution (Type I extreme value distribution). The PDF of the distribution is given by

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

After generating the required exponential random variables and then taking their maximum to generate our data points, we are now tasked with finding the Maximum Likelihood Estimate(MLE) of Sigma. The theoretical value of sigma is given as 5 and the value of is given as 14.9787. To find the MLE, we first need to find the log-likelihood function for this PDF.

$$L(\theta) = -Nlog(\sigma) - \frac{1}{\sigma}((\sum_{i=1}^{N} y_i) - n\mu) - \sum_{i=1}^{N} exp\Big(-\frac{(y_i-\mu)}{\sigma}\Big) \tag{4}$$

There is no direct closed form expression when we try to maximise this expression with respect to $\theta by taking the first derivative. We can however simplify the problem into solving these two equations simultaneously.$

$$\sigma_{gev} = \bar{y} - \frac{\sum_{i=1}^{N} y_i exp\Big(-\frac{y_i}{\sigma_{gev}}\Big)}{\sum_{i=1}^{N} exp\Big(-\frac{y_i}{\sigma_{gev}}\Big)}; \mu = -\sigma_{gev} ln\Big[\frac{1}{N}\sum_{i=1}^{N} exp\Big(-\frac{y_i}{\sigma_{gev}}\Big)\Big] \tag{5}$$

Here, since we are concerned with only the value of gev, we can use the first equation to iteratively solve for the value of gev. For a good starting point, we use a property of the Gumbel Distribution.

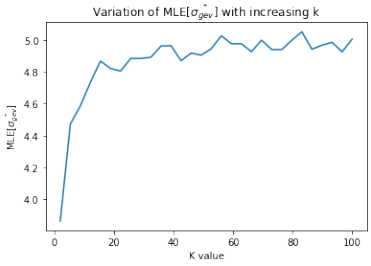
Property:

The standard deviation of Yn () is related to gev. as gev = 6/.

Using this as an initial estimate, we compute the RHS of the expression. Then we take the average of this new and the old beta for faster convergence. Even though the code has an upper limit on number of iterations at 100, there's also a threshold (set at 0.00000001) and hence the code stops within 10-12 iterations.

As mentioned, we noticed another observation that was noticed when trying to solve this problem. Here we keep the number of samples fixed at N=10,000 and plot the variation of the MLE of gev for increasing k (varied from 2 to 100)

[4]:
```
#Plot
```



Inferences from Question 2.

1. As the value of the number of samples increases, there seems to be an overall trend of convergence of the MLE towards the theoretical value of 5. The exact values are however heavily data dependent.

2. Another interesting observation was notices in this problem. For k=20, when the estimator expectation was plotted in problem 3, the gev showed a convergence to 4.85 for large values of N. The reason for this anomaly is that as mentioned in the problem statement, the value of K has to be large and a value of 20 proves to be insufficient for this purpose. To illustrate this point, we plotted the graph of varying k with the MLE. As predicted, for higher values of K, the algorithm converges to 5 albeit with a lot of data dependent fluctuations.

3 Question 3

Given one iteration of X, which is of shape NxK, find the relevant MLE for the different parameters.

$\hat{\theta} = \frac{\sum_{n,k=1} X_{nk}}{N}$

$\hat{\sigma_{gev}}$ is obtained from the previous question

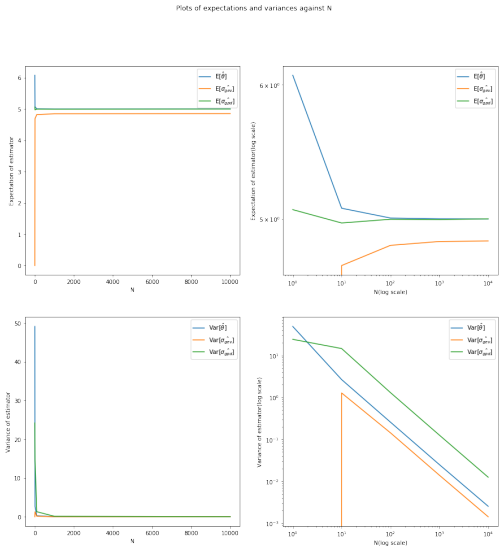
$\hat{\sigma_{gpd}} = \frac{\sum_l S_l}{L}$.

For constructing pdfs, cdfs and extracting mean and variance data, we take an ensemble of size 10000 for each value of N.

Parts a and b.

Here we print and plot the expectations and variances of the estimators for different values of N. We report our observations in tabular form in the next textwrite cell. Also, we also get the ensemble of estimators for different values of N, which will be used to get the empirical cdfs and pdfs.

[8]:
```
#Plots of expectations and variances against N
```



3.0.1 Tables for the expectations and variances against N

Ensemble size = 10000

Table for expectations

N/Estimator	$\hat{\theta}$	$\hat{\sigma_{gev}}$	$\hat{\sigma_{gev}}$
1	5.950329901248642	0.0	5.033953227534965
10	5.043588356025091	4.694070682196424	4.996358277655083

N/Estimator	$\hat{\theta}$	$\sigma_{\hat{gev}}$	$\sigma_{\hat{gev}}$
100	5.003858218579122	4.822220341981129	5.015853250839413
1000	4.999944635807824	4.849626615792224	4.996163777915849
10000	4.9998879762576465	4.852413282803332	5.001149746187272

Table for variance

N/Estimator	$\hat{\theta}$	$\sigma_{\hat{gev}}$	$\sigma_{\hat{gev}}$
1	45.02023265731358	0.0	24.76322371874191
10	2.580213342554128	1.2919900989448299	14.11663342556236
100	0.2477669982837358	0.14727024328073085	1.293763196613956
1000	0.02445510433380918	0.014610281029785855	0.12486491073293028
10000	0.0025150038728337584	0.0014424768522638954	0.012753353921647488

3.1 Observations for parts a and b

1. Regarding values for expectations, one can observe that the values of expectations for all the estimators are very close to 2. The mean gets closer to 5 as N increases, and we can say with fair amount of confidence that all the estimators are asymptotically unbiased with N.
2. Regarding values for variances: For all of the estimators, one can see that the variance decreases with N. We have plotted the estimators vs N in log log scale, and one can clearly see that the variance decreases as roughly 10 times as N is increased by a factor of 10. The plot is also a stright line in log log scale, strongly confirming the earlier remarks.
3. **Note that only in this special example given to us, the values of estimators are theoretically exactly same. In general these represent different quantities. But for this special example, they are equivalent, and we are using this to make comments on noise in estimation etc.**. Now we proceed to variance comparison between the different estimators, for a fixed N. One can immediately see from the plots that $\sigma_{\hat{gev}}$ has the least noise, followed by $\hat{\theta}$ and then $\sigma_{\hat{gev}}$.

a. The estimator $\sigma_{\hat{gpd}}$ is based on the excess value. For $\lambda = 0.2, d = 23$, a fraction $e^{-\lambda d} = 0.01$ of the total N x K observations are expected to be greater than d. For K = 20, this is equal to $\frac{N}{5}$ data points.

b. The estimator $\hat{\theta}$ is based on N observations of X_{nk}.

c. The estimator $\sigma_{\hat{gev}}$ is also based on N observations. But these N observations are based on the greatest value along the row. So there is this extra information available to us that these are the greatest in magnitude along a row, and hence despite also having the same number of datapoints available to us as $\hat{\theta}$, GEV estimator beats it because of this extra piece of information.

3.2 Part c

Asymptotic Distribution *Theorem*: Assume $X_1, X_2, ..., X_n$ are iid random variables with pdf $f(x; \theta_o)$ for $\theta_o \in \Omega$ such that all regularity conditions are obeyed. Suppose further that $0 < I(\theta_o) <$

∞. Then, any consistent sequence of solutions the MLE equation satisfies:

$$\sqrt{N}(\hat{\theta} - \theta_o) \xrightarrow{D} \mathcal{N}\,(0, \frac{1}{I(\theta_o)})$$

Here, $\mathcal{N}\,(0, \frac{1}{I(\theta_o)})$ is the asymptotic distribution of the MLE. We find the asymptotic distributions for the three distributions, namely the Exponential Distribution, Gumbel Distribution and Generalized Pareto Distribution.

Fisher Information

We first evaluate the Fisher Information for the Exponential Distribution, Gumbel Distribution and the Generalized Pareto Distribution. We know that $I(\theta) = \int_{-\infty}^{\infty} \frac{\partial log f_X(x;\theta)}{\partial \theta} \frac{\partial log f_X(x;\theta)}{\partial \theta} f_X(x;\theta), dx$

For the Exponential Distribution with the PDF defined as:

$$f_X(x;\theta) = \begin{cases} \frac{1}{\theta} exp(-\frac{x}{\theta}) & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{6}$$

The Fisher Information, $I(\theta)$ can be obtained in the closed form using the aforementioned equation, as follows:

$$I(\theta) = \frac{1}{\theta^2}$$

For $\theta = 5$, this comes out to be $I(\theta) = \frac{1}{25}$ Similarly, for the GPD, having the PDF defined as:

$$f_S(s;\sigma) = \begin{cases} \frac{1}{\sigma} exp(-\frac{s}{\sigma}) & s \geq 0 \\ 0 & s < 0 \end{cases} \tag{7}$$

The Fisher Information is given by:

$$I(\sigma) = \frac{1}{\sigma^2}$$

For $\sigma = 5$, this comes out to be $I(\sigma) = \frac{1}{25}$

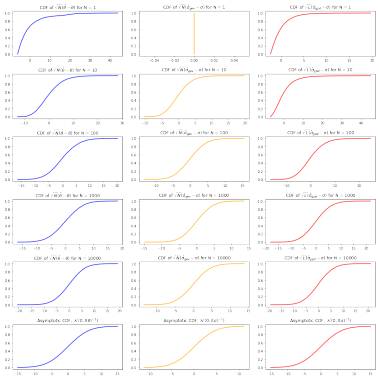
Now, it is not possible to easily evaluate a closed form expression for the Gumbel Distribution. Hence, we use scipy's integrate method to evaluate the above integral for the Gumbel distribution, whose pdf is given by:

$$f_Y(y;\sigma) = \begin{cases} \frac{1}{\sigma} exp\Big(\frac{-(y-\mu)}{\sigma}\Big) exp\Big(-exp\Big(\frac{-(y-\mu)}{\sigma}\Big)\Big) & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{8}$$

Numerically, the value of the integral, which in turn is equal to Fisher Information, $I(\sigma)$. For this case, $I(\sigma) = 0.0729$.

Estimated Distributions Now, we move on to plotting the cdf of the three distributions from the data samples obtained from parts a and b of this question. Code for the same is in the next code cell. The columns represents the CDF of $\sqrt{N}(\hat{\theta} - \theta)$, $\sqrt{N}(\hat{\sigma}_{gev} - \sigma)$ and $\sqrt{L}(\hat{\sigma}_{gpd} - \sigma)$ respectively for different values of N, namely $N = 1, 10, 100, 1000, 10000$. Additionally, at the bottom of each column, we have included the asymptotic distribution for each case as well to show convergence.

[11]:
```
#CDF Plot
```



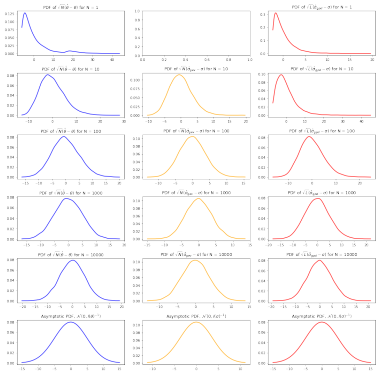
3.3 Observations about part c

1) The CDF curve for all the three cases take the typical S shape, which is characteristic of the Normal Distribution.
2) The CDF is centered around 0 for all the three distributions.
3) According to the theorem stated above, the sequences of MLE solutions should converge in distribution to Normal Distributions with mean = 0 and variance = I^{-1}, where I is the Fisher Information.
4) Comparing with the CDF of the asymptotic distribution, we find that as N increases, the CDF of the estimated distributions converge to the asymptotic distribution.

3.4 Part d

We perform a similar activity for the PDFs as well. *Note*: As per the equation for σ_{gev} described above, $\sigma_{gev} = 0$ for N = 1. Hence, if we run the MLE iterations in parts a and b above, the vector sigma_gev_matrix[0] does not get updated at all, hence remaining at zero. Therefore, this results in a singular matrix while evaluating the Gaussian KDE. Thus, the PDF for N = 1 for σ_{gev} cannot be evaluated using Gaussian KDE and without incorporating limits and has been left blank in order to facilitate readibility and alignment of the other graphs. This can be seen from the CDF of σ_{gev} at N = 1 from part c. It is basically a dirac-delta, which is known to be non-differentiable (and hence pdf cannot be obtained without using limits).

[12]:
```
#PDF plots
```



3.5 Observations about part d

1) The PDF curve for all the three cases take the typical bell curve shape, which is characteristic of the Normal Distribution.
2) The PDF is centered around 0 for all the three distributions.
3) The peak of the pdf is given by $\frac{1}{\sqrt{2\times\pi\times variance}}$. Since variance = I^{-1}, peak = $\frac{\sqrt{I}}{\sqrt{2\times\pi}}$. For the three cases, this comes out to be equal to 0.8 for the Exponential Distribution and GPD, and equal to 0.1077 for the Gumbel Distribution. These are the values that we also practically obtained.
4) Just as in the case of the CDF, according to the theorem stated above, the sequences of MLE solutions should converge in distribution to Normal Distributions with mean = 0 and variance = I^{-1}, where I is the Fisher Information.
5) Comparing with the PDF of the asymptotic distribution, we find that as N increases, the PDF of the estimated distributions converge to the asymptotic distribution.