Annual Progress Review

Thomas William Boughen

Newcastle University

School of Mathematics, Statistics and Physics

1 Extreme Value Theory

Extreme value theory is a field focused on studying properties at the tail ends of distributions where real world data may be scarce and hard to make inferences from. A lot of the standard theory assumes continuous distributions and that is what will be introduced first before looking at what has been done relating to discrete distributions.

1.1 Standard Theory

One approach for modelling the extreme values is to look at modelling the block maxima of independent and identically distributed random variables $X_1, X_2 \ldots$ that all have a common cumulative distribution function (CDF) F. The block maxima M_n being defined as $M_n = \max\{X_1, \ldots, X_n\}$ has its own CDF defined by:

$$\Pr(M_n \leq x) = F_n(x)$$

F is in the domain of attraction of an extreme value CDF G, if and only if the normalised version of M_n's CDF converges to a non degenerate G, that is, there exists some sequence of $a_n > 0$ and $b_n \in \mathbb{R}$ such that:

$$\Pr\left(\frac{M_n - b_n}{a_n} \leq x\right) = F^n(a_n x + b_n) \to G(x), \qquad \text{as } n \to \infty$$

If this holds, then F is in the domain of attraction of G which we will write as $F \in \mathcal{D}(G)$. The extreme value theorem states that is limit CDF G can be catagorised into one of three types:

- Gumbel: $\Lambda(x) = \exp\{-\exp(-x)\}, \quad x \in \mathbb{R}$
- Fréchet: $\Phi_a(x) = \exp\{-x^{-\alpha}\}, \quad x \geq 0, \alpha > 0$
- Weibull: $\Psi_\alpha(x) = \exp\{-x^{-a}\}, \quad x < 0, \alpha > 0$

Definition 1.1.1 (Generalised Extreme Value Distribution)**.** These three types of distribution can be combined into one single distribution called the Generalised Extreme Value (**GEV**) Distribution which has CDF:

$$G(x) = \exp\left\{-\left(1 + \frac{\xi(x-\mu)}{\sigma}\right)^{-1/\xi}\right\}$$

denoted $\text{GEV}(\mu, \sigma, \xi)$ for some $\mu \in \mathbb{R}, \sigma > 0, \xi \in \mathbb{R}$ and has support on $\{x \in \mathbb{R} : 1 + \xi(x-\mu)/\sigma > 0\}$ with each of the three types being obtained from changing the values of each of the parameters with $\xi = 0$ taken as the limit:

- Gumbel: $\text{GEV}(\mu, \sigma, 0)$
- Fréchet: $\text{GEV}(1, 1, 1/\alpha)$
- Weibull: $\text{GEV}(-1, -1, -1/\alpha)$

The most important parameter here is ξ which will be referred to as the shape parameter as it controls the tail behaviour of the distribution allowing it to occupy the three domains of attraction.

Definition 1.1.2 (Heavy Tails)**.** There are a few definitions that can be used to define a distribution that has heavy tails, one that will not be used here is that the tails of the distribution function are heavier than an exponential. Here, a distribution with CDF F will be said to have heavy tails if it is in the Fréchet domain of attraction with tail index α, or it is in the Gumbel domain of attraction.

Definition 1.1.3 (Generalised Pareto Distribution)**.** A related distribution called the Generalised Pareto (**GP**) Distribution is also often used to model the probability distribution of threshold excesses, it has the CDF:

$$H(x) = 1 - \left(1 + \frac{\xi x}{\sigma}\right)^{-1/\xi}$$

denoted $\text{GP}(\sigma, \xi)$ for some $\sigma > 0, \xi \in \mathbb{R}$ it has support on either $(0, \infty)$ when $\xi \geq 0$ or $(0, -\sigma/\xi)$ when $\xi < 0$.

This distribution of often used to model the conditional probability of iid random variables exceeding some cut-off u. However, like most of the theory above, it requires iid discrete random variable; in the case of networks and modelling the degrees of their vertices the focus is on discrete data so tools to aid in modelling discrete data are required.

1.2 Discrete Extremes

Since the focus of this report is discrete data, theory on discrete extremes will need to be examined starting with a discrete alternative to the GP distribution.

Definition 1.2.1 (Integrated Generalised Pareto Distribution)**.** Roughly following Rohrbeck et al. (2018), the integrated generalised pareto (**IGP**) distribution can be defined by considering modelling the random variable $Y = \lceil X \rceil$ for some continuous random variable X with support on the positive real line such that $X|X > u \sim \mathrm{GPD}(\sigma_0 + \xi u, \xi)$ for $\xi \in \mathbb{R}, u \in \mathbb{R}^+$. The probability mass function (PMF) of the IGP distribution can then be defined as:

For values $y = \lfloor u \rfloor, \lfloor u \rfloor + 1, ...$ and $\xi \in \mathbb{R}$ and $u, \sigma_0 \in \mathbb{R}^+$:

$$\begin{aligned}
\Pr(Y = y|Y > u) &= \Pr(X < y|X > \lfloor u \rfloor) - \Pr(X < y - 1|X > \lfloor u \rfloor) \\
&= \left(1 + \frac{\xi(y - \lfloor u \rfloor)}{\sigma_0 + \xi \lfloor u \rfloor}\right)_+^{-1/\xi} - \left(1 + \frac{\xi(y - 1 - \lfloor u \rfloor)}{\sigma_0 + \xi \lfloor u \rfloor}\right)_+^{-1/\xi}
\end{aligned}$$

By modelling the ceiling of a continuous random variable, it is also suggested that one could instead model the floor of a continuous random variable instead. Indeed, that is what will be done from here on out. Consider modelling the random variable $Y = \lfloor X \rfloor$, the PMF of the IGP then becomes:

For values $y = \lceil u \rceil, \lceil u \rceil + 1, ...$ and $\xi \in \mathbb{R}$ and $u, \sigma_0 \in \mathbb{R}^+$:

$$\Pr(Y = y|Y > u) = \left(1 + \frac{\xi(y + 1 - \lceil u \rceil)}{\sigma_0 + \xi \lceil u \rceil}\right)_+^{-1/\xi} - \left(1 + \frac{\xi(y - \lceil u \rceil)}{\sigma_0 + \xi \lceil u \rceil}\right)_+^{-1/\xi}$$

References

Rohrbeck, Christian, Emma F. Eastoe, Arnoldo Frigessi, and Jonathan A. Tawn. 2018. "Extreme value modelling of water-related insurance claims." *The Annals of Applied Statistics* 12 (1): 246–82. https://doi.org/10.1214/17-AOAS1081.