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 $X_1, X_2 \dots$ that all have a common cumulative distribution function (CDF) F. The block maxima M_n being defined as $M_n = \max\{X_1, \dots, X_n\}$ 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 GEV(μ, σ, ξ) 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: GEV$(\mu, \sigma, 0)$
- Fréchet: GEV$(1, 1, 1/\alpha)$
- Weibull: 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 GP(σ, ξ) 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 $X|X > u \sim \text{GPD}(\sigma_0 + \xi u, \xi)$ 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 $u, \sigma_0 \in \mathbb{R}^+$:

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

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 $u, \sigma_0 \in \mathbb{R}^+$: par

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

If this is to be used as a discrete alternative to the GP distribution, it needs to be verified that the parameter ξ affects the tail behaviour in the same way. The results from Shimura (2012) can be used to show that this distribution will indeed belong to the same domain of attraction as the GP distribution for any given value of ξ. Shimura (2012) also introduces an important quantity that will be useful in finding what domain of attraction a discrete distribution belongs to, namely:

$$\Omega(F, n) = \left(\log \frac{\bar{F}(n+1)}{\bar{F}(n+2)}\right)^{-1} - \left(\log \frac{\bar{F}(n)}{\bar{F}(n+1)}\right)^{-1}$$

where $\bar{F}$ is the survival function of a discrete random variable.

While a lot of distributions may be just be heavy tails, it will be useful to define what is means for a distribution to have super heavy tails. There are several definitions being used but here the one from [SUPER HEAVY ref] will be used.

Definition 1.2.2 (Super Heavy Tails)**.** A distribution with $f, F, \bar{F}$ as its pdf, cdf, and survival respectively is said to have super heavy tails if the survival function $\bar{F}$ is slowly varying. That is:

$$\lim_{x \to \infty} \frac{\bar{F}(cx)}{\bar{F}(x)} = 1, \qquad \forall c \in \mathbb{R}^+$$

It should be noted that this condition is sufficient but not necessary and does not capture all distributions with super heavy tails.

2 Networks

In the previous section an overview of extreme value theory was given, in this section context will be given as to what distributions are in the scope of this research and what they arise from.

2.1 Mathematical Definition

The networks being analysed will be treated as mathematical graphs that consist of two kinds of objects, vertices that usually represent the individual components in a complex system, and edges that represent the relationships between these components. The focus, for now, will be on undirected networks and therefore undirected graphs.

Definition 2.1.1 (Undirected Graph)**.** An undirected graph G consists of a set of vertices V and a set of edges between these vertices say E, this will be denoted as $G = (V, E)$ where V is usually a subset of $\mathbb{N}$ and E is a subset of the set of unordered pairs of V.

Definition 2.1.2 (Degree)**.** A vertex i has degree k_i, where k_i is the number of appearances of i in the set E. That is, the degree of a vertx is the number of edges that connect to it.

2.2 Analysis of Real World Networks

Before looking at what has and could be used to model the growth of real networks, it will be useful to analyse some examples. The property of focus here will be the degree distribution of real world networks, more specifically the tail behaviour. Figure 2.1 shows the survival function of the degree distribution (on a log-log plot) for four examples of real world networks that represent: game results between top tennis players, facebook friends of Harvard students and alumni, interactions between proteins in cells, and the dependencies of R packages.

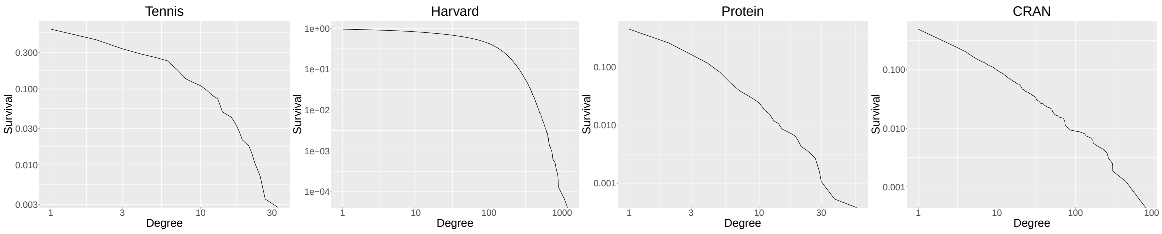

Figure 2.1: Survival function of the degrees.

The shapes of the curves in Figure 2.1 are all fairly different but, they all seem to start with what looks like one power law and then begins to deviate after a certain point indicating the while the bulk of the data follows the one power law the right tail does not necessarily exhibit the same behaviour. Seeing this and with the aim of investigating the tail-heaviness of networks degree distributions, fitting a model that uses a power law for the bulk of the data and the IGP distribution (Definition 1.2.1) for the right tail above a certain threshold should allow for estimation of the tail-index.

Power Law IGP Mixture Model (PL-IGP)

A traditional way of modelling continuous extreme values often involves using the GP distribution (Definition 1.1.3) to model the conditional distribution of the largest values however here the data is discrete and so instead the IGP will be used as an alternative. For the bulk of the data a truncated power law distribution will be used which is defined in Definition 2.2.1 below, and for the right tail the IGP will be used.

Definition 2.2.1 (Truncated Power Law)**.** The truncated power law distribution has two parameters $\alpha \in \mathbb{R}$,$v \in \mathbb{Z}^+$ and has support of $(1, v)$. The pdf of this distribution is:

$$f(x) = \begin{cases} \frac{x^{-(\alpha+1)}}{\sum_{i=1}^{v} i^{-(\alpha+1)}} & , x = 1, 2, \ldots, v \\ 0 & , \text{otherwise} \end{cases}$$

2.3 Network Generative Models

The way in which vertices and edges are added and removed from networks over time is of great interest as it can lead to inferences about the underlying mechanics of a complex system and allow for theorising what may happen to the network in the future. To that end, finding a model that is both simple enough and leads to properties that match real world networks is of utmost importance. Next, the degree distribution of various network models are analysed starting fairly simple and gradualling getting more complex.

Definition 2.3.1 (Uniform attachment model)**.** Consider an initial graph G_0 with vertex set $V_0 = \{1, 2, \dots, m_0\}$ and edge set $E_0 = \emptyset$. Starting at $t = 1$ repeat the steps below:

1. Add a vertex to the network, $V_t = V_{t-1} \cup \{m_0 + t\}$
2. Add $m \leq m_0$ edges to the network connecting the new vertex to those in already in the network. With the existing vertices to be chosen uniformly at random from V_{t-1} with replacement.

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.

Shimura, Takaaki. 2012. "Discretization of Distributions in the Maximum Domain of Attraction." *Extremes* 15 (September): 1–19. https://doi.org/10.1007/s10687-011-0137-7.