Extreme Values

Richard L. Smith
Department of Statistics and Operations Research,
University of North Carolina, Chapel Hill, USA

and

Ishay Weissman
Faculty of Industrial Engineering and Management,
Technion, Haifa, Israel.

Contents

Chapter 1

Introduction to Extreme Value Theory

1.1 Insurance Claims Example

To be added.

1.2 Running Times Example

To be added.

1.3 Overview of Univariate Extremes

To be added.

1.4 Hierarchical Models

To be added.

1.5 Applications to the Two Examples

To be added.

1.6 Extreme Precipitations on a Spatial Grid

To be added.

1.7 Introduction to Bivariate Extremes

To be added.

Chapter 2

Limit Theory for Univariate Extremes

2.1 The Three Types Theorem

The earliest papers on extreme value theory as we now know it were due to Fréchet [8] and Fisher and Tippett [7]. These papers explored limit distributions for sample extremes, and in particular, Fisher and Tippett identified what we now know as the "three types" of limiting distributions. Von Mises [19] recast the Fisher-Tippett results into what we now know as the Generalized Extreme Value distribution. The first rigorous proof that the three types are the only possible limit distributions for univariate extremes was given in the seminal paper of Gnedenko [9] who also characterized the domains of attraction of the three limit laws. The domain of attraction problem was completed by de Haan [11] in his seminal 1970 thesis. The following discussion is based primarily on the paper of de Haan [12] that included a simplified proof of the three types theorem and a simplified set of sufficient conditions for the domains of attraction.

Suppose $X_1, X_2, ...$ are a sequence of independent and identically distributed (IID) random variables with a common cumulative distribution function (CDF) F,

$$\Pr\{X_i \le x\} \quad = \quad F(x),\ x \in \mathbb{R},\ i = 1, 2, ...,$$

Let $M_n = \max\{X_1, X_2, ..., X_n\}$. Then

$$\Pr\{M_n \le x\} \quad = \quad F^n(x),\ x \in \mathbb{R}.$$

We are interested in limit laws of the form

$$\begin{aligned} \Pr\left\{\frac{M_n - b_n}{a_n} \le x\right\} \quad &= \quad F^n(a_n x + b_n) \\ &\xrightarrow{w} \quad G(x) \end{aligned} \tag{2.1}$$

where $a_n > 0,\ b_n \in \mathbb{R}$ are normalizing constants, G is a *nondegenerate* CDF (meaning there is at least one value of x for which $0 < G(x) < 1$ and $\xrightarrow{w}$ is weak convergence (also called convergence in distribution) which means that (2.1) holds at all continuity points of G. In this case, G is called an *extreme value distribution* and F is said to be in the *domain of maximum attraction* of G. The latter phrase will be simplified to *domain of attraction* where there is no ambiguity about the operation being conducted.

The key questions arising from these definitions are:

1. How can we characterize the extreme value distributions — are there simple necessary and sufficient conditions to determine whether any given G is an extreme value distribution?
2. For any given extreme value distribution G, how can we characterize the domain of attraction — necessary and sufficient conditions on F so that (2.1) holds for some sequences $a_n > 0,\ b_n \in \mathbb{R},\ n = 1,2,....$?

We should note that the same theory applies with only minor changes to sample minima, because of the relationship

$$\min\{X_1,...,X_n\} \quad = \quad -\max\{-X_1,...,-X_n\}. \tag{2.2}$$

Definition 2.1. Two CDFs G_1 and G_2 are said to be *of the same type* if there exist $a > 0,\ b \in \mathbb{R}$ such that

$$G_2(x) \quad = \quad G_1(ax+b),\ x \in \mathbb{R}.$$

We now state the main:

Theorem 2.1 (Extremal Types Theorem). If (2.1) holds, then G must be of the same type as one of

$$\Phi_\alpha(x) \quad = \quad \begin{cases} 0, & \text{if } x \le 0, \\ \exp(-x^{-\alpha}), & \text{if } x > 0, \end{cases} \tag{2.3}$$

$$\Psi_\alpha(x) \quad = \quad \begin{cases} \exp(-(-x)^{\alpha}), & \text{if } x < 0, \\ 1, & \text{if } x \ge 0, \end{cases} \tag{2.4}$$

$$\Lambda(x) \quad = \quad \exp\left(-e^{-x}\right),\ x \in \mathbb{R}. \tag{2.5}$$

In (2.3) and (2.4), $\alpha > 0$ is an arbitrary positive parameter.

Before going into the proof of this result, we show why equations (2.3)–(2.5) are equivalent to the Generalized Extreme Value (GEV) distribution

$$G(y) \quad = \quad \exp\left\{-\left(1+\xi\frac{y-\mu}{\psi}\right)_+^{-1/\xi}\right\} \tag{2.6}$$

where $\mu \in \mathbb{R},\ \psi > 0,\ \xi \in \mathbb{R}$. In (2.6), the symbol $(\cdot)_+$ is positive part, or $x_+ = x$ if $x \ge 0$, $x_+ = 0$ if $x < 0$.

We divide the analysis of (2.6) into three cases.

Case 1: $\xi > 0$. In this case, $1+\xi\frac{y-\mu}{\psi} < 0$ if and only if $y < y_* = \mu - \frac{\psi}{\xi}$ and for such y, $G(y) = e^{-\infty} = 0$. If the random variable Y has CDF G and we define $Z = \frac{\xi}{\psi}(Y - y_*)$ then for $z > 0$,

$$\begin{aligned} \Pr\{Z \le z\} &= \Pr\left(Y \le y_* + \frac{\psi z}{\xi}\right) \\ &= \exp\left[-\left\{1+\frac{\xi}{\psi}\left(\mu - \frac{\psi}{\xi} + \frac{\psi z}{\xi} - \mu\right)\right\}^{-1/\xi}\right] \\ &= \exp\left(-z^{-1/\xi}\right) \end{aligned}$$

which is of the form (2.3) with $\alpha = 1/\xi$.

Case 2: $\xi < 0$. In this case, $1 + \xi \frac{y-\mu}{\psi} < 0$ if and only if $y > y^* = \mu - \frac{\psi}{\xi}$ where $G(y) = 1$. After similar manipulations to Case 1, this reduces to (2.4) with $\alpha = -1/\xi$.

Case 3: $\xi = 0$. In this case we interpret (2.6) as the limit as $\xi \to 0$. Noting that $\lim_{\xi \to 0} \left(1 + \frac{z}{\xi}\right)^{-1/\xi} = e^{-z}$, (2.6) reduces to $\exp\left\{-\exp\left(-\frac{y-\mu}{\psi}\right)\right\}$ which is the same as (2.5) after a translation by μ and a scaling by ψ.

Now we turn to the proof of the Extremal Types Theorem. Our proof will very closely follow de Haan [12].

First, a key lemma due to Khinchine:

Lemma 2.1. Suppose $F_n,\ n = 1, 2, \ldots$ is a sequence of CDFs. Then the statements

$$F_n(a_n x + b_n) \quad \stackrel{w}{\to} \quad G(x), \tag{2.7}$$

$$F_n(\alpha_n x + \beta_n) \quad \stackrel{w}{\to} \quad G_*(x), \tag{2.8}$$

for nondegenerate CDFs $G,\ G_*$ and sequences $a_n > 0,\ b_n \in \mathbb{R},\ \alpha_n > 0,\ \beta_n \in \mathbb{R}$, are true if and only if there exist some $a > 0,\ b \in \mathbb{R}$ such that

$$\lim_{n\to\infty} \frac{\alpha_n}{a_n} \quad = \quad a \text{ and } \lim_{n\to\infty} \frac{\beta_n - b_n}{a_n} \ = \ b. \tag{2.9}$$

In that case,

$$G_*(x) \quad = \quad G(ax + b). \tag{2.10}$$

Proof. Let us begin by defining the inverse function for any CDF F:

$$F^{-1}(y) \quad = \quad \inf\{x : \ F(x) > y\}. \tag{2.11}$$

Our first claim is that (2.7) and (2.8) hold if and only if

$$\frac{F_n^{-1}(y) - b_n}{a_n} \quad \stackrel{w}{\to} \quad G^{-1}(y), \tag{2.12}$$

$$\frac{F_n^{-1}(y) - \beta_n}{\alpha_n} \quad \stackrel{w}{\to} \quad G_*^{-1}(y). \tag{2.13}$$

The simplest proof of this uses the Lévy metric $L(F,G)$ between two distribution functions F and G, defined by

$$L(F,G) \quad = \quad \inf\{\varepsilon : \ F(x-\varepsilon) - \varepsilon \le G(x) \le F(x+\varepsilon) + \varepsilon \text{ for all } x\}.$$

One property of L is that $L(F,G) = L(F^{-1}, G^{-1})$ [29]. If we define $F_n^*(x) = F_n(a_n x + b_n)$ then $F_n^{-1}(y) = a_n (F_n^*)^{-1}(y) + b_n$ and $F_n^* \stackrel{w}{\to} G$ if and only if $(F_n^*)^{-1} \stackrel{w}{\to} G^{-1}$ (because both statements are equivalent to $L(F_n^*, G) \to 0$). Hence (2.7) and (2.12) are equivalent, and similarly (2.8) and (2.13).

Since we are assuming G is nondegenerate, G^{-1} must take at least two distinct

values, say $G^{-1}(y_1) > G^{-1}(y_0)$. Similarly there exist y_1^* and y_0^* such that $G_*^{-1}(y_1^*) > G_*^{-1}(y_0^*)$. Take some number $z_1 \geq \max(y_1, y_1^*)$ and $z_0 \leq \max(y_0, y_0^*)$ such that z_0 and z_1 are continuity points of G^{-1} and G_*^{-1}. Applying (2.12) for y and z_0 we find

$$\frac{F_n^{-1}(y) - F_n^{-1}(z_0)}{a_n} \quad \stackrel{w}{\to} \quad G^{-1}(y) - G^{-1}(z_0) \tag{2.14}$$

and with (2.12) also

$$\frac{b_n - F_n^{-1}(z_0)}{a_n} \quad \stackrel{w}{\to} \quad -G^{-1}(z_0).$$

From (2.14) with $y = z_1$ we get

$$\frac{F_n^{-1}(z_1) - F_n^{-1}(z_0)}{a_n} \quad \stackrel{w}{\to} \quad G^{-1}(z_1) - G^{-1}(z_0) \; > \; 0.$$

Similarly, starting with (2.13) we deduce that the last two equations are also true with β_n, α_n replacing a_n, b_n and G_* replacing G. Hence (2.9) holds, and this implies (2.10). □

One consequence of Lemma 2.1 is that a relationship such as (2.1) only characterizes G up to type — if (2.1) is true for one G, then by redefinition of a_n and b_n we can make (2.1) true for any other G of the same type. Stated another way, if G and G_* are two CDFs of the same type, their domains of attraction are identical.

Lemma 2.2. A nondegenerate CDF G has non-empty domain of attraction if and only if, for every $s > 0$, there exist $A(s) > 0$ and $B(s) \in \mathbb{R}$ such that

$$G^s(A(s)x + B(s)) \quad = \quad G(x) \text{ for all } x \in \mathbb{R}. \tag{2.15}$$

Proof. If (2.15) holds, then automatically (2.1) holds with $F = G$. Therefore, we concentrate on proving that (2.1) implies (2.15).

For $s > 0$, write $\lfloor ns \rfloor$ for the integer part of ns. Then

$$F^{\lfloor ns \rfloor}\left(a_{\lfloor ns \rfloor}x + b_{\lfloor ns \rfloor}\right) \stackrel{w}{\to} G(x)$$

and hence

$$F^{n}\left(a_{\lfloor ns \rfloor}x + b_{\lfloor ns \rfloor}\right) \stackrel{w}{\to} G^{1/s}(x)$$

Lemma 2.1 implies that G and $G^{1/s}$ are of the same type, so (2.15) follows. □

Definition 2.2. A nondegenerate CDF G is said to be *max-stable* if, for any $n \in \mathbb{N}$, there exist constants $A_n > 0$ and $B_n \in \mathbb{R}$ such that

$$G^n(A_n x + B_n) \quad = \quad G(x).$$

To derive the Extremal Types Theorem, we need one additional result from analysis:

Lemma 2.3. Suppose $u : \mathbb{R} \to \mathbb{R}^+$ is monotone and satisfies

$$u(t+s) \quad = \quad u(t) + u(s)$$

for all real s, t. Then either $u(t) = 0$ for all t or $u(t) = e^{\rho t}$ for some real ρ.

Proof. Suppose $u(t_0) \neq 0$ for some $t_0 \neq 0$, for definiteness assume $t_0 > 0$. For all integers m and n we have $u(mt_0) = u(t_0)^m$ and $u\left(\frac{t_0}{n}\right) = u(t_0)^{1/n}$ so $u\left(\frac{mt_0}{n}\right) = u(t_0)^{m/n}$. But the set $\left\{\frac{m}{n}t_0 : \ m,n \in \mathbb{N}\right\}$ is dense in $\mathbb{R}^+$ and u is monotone, hence u is continuous and $u(tt_0) = u(t_0)^t$ for all $t > 0$. This is of the form $e^{\rho t}$ with $\rho = \log u(t_0)$. The proof for $t < 0$ follows from $u(-t) = u(t)/u(2t)$. □

Remark. This is also called the Cauchy functional equation and is true under weaker assumptions than monotonocity, for example, measurability suffices. See Theorem 1.1.7 of [1].

Proof of Theorem 2.1. We have seen that (2.1) implies that G obeys (2.15) for each $s > 0$ with some $A(s) > 0$ and $B(s) \in \mathbb{R}$. The challenge is therefore to show that any G obeying (2.15) is of the same type as one of (2.3)–(2.5).

For $0 < G(x) < 1$ we have

$$-\log[-\log\{G(A(s)x+B(s))\}] - \log s \quad = \quad -\log[-\log\{G(x)\}].$$

Let $U(\cdot)$ be the inverse of $-\log[-\log\{G(\cdot)\}]$. Then

$$\frac{U(x+\log s) - B(s)}{A(s)} \quad = \quad U(x)$$

for $s > 0,\ x \in \mathbb{R}$. Substracting the same relation fpor $x = 0$, we get

$$\frac{U(x+\log s) - U(\log s)}{A(s)} \quad = \quad U(x) - U(0).$$

Defining $A_1(y) = A(e^y),\ \tilde{U}(x) = U(x) - U(0),\ y = \log s$, we have

$$\tilde{U}(x+y) - \tilde{U}(y) \quad = \quad \tilde{U}(x)A_1(y). \tag{2.16}$$

Also writing (2.16) with x and y interchanged, and subtracting,

$$\tilde{U}(x)(1-A_1(y)) \quad = \quad \tilde{U}(y)(1-A_1(x)). \tag{2.17}$$

Case 1. Suppose $A_1(x) = 1$ for all x. THen from (2.16), $\tilde{U}(x+y) = \tilde{U}(x) + \tilde{U}(y)$ for all x and y. By Lemma 2.3 (applied to $e^{\tilde{U}(x)}$), we have $\tilde{U}(x) = \rho x$ for all x, for some $\rho > 0$. This is equivalent to the statement that G is of type (2.5).

Case 2. Suppose there is an x with $A_1(x) \neq 1$. We claim that this implies $A_1(y) \neq 1$ for all $y \neq 0$. Suppose, for contradiction, $A_1(y) = 0$. Then $\tilde{U}(y) = 0$ and hence, from (2.16), $\tilde{U}(x) = \tilde{U}(x+y)$ for all x. Since $\tilde{U}$ is monotone, that would imply that $\tilde{U}(x)$ is constant for all x, in which case G is degenerate, contrary to assumption. So $A_1(x) \neq 1$ for all $x \neq 0$. By (2.17), there exists some real $c_1 \neq 0$ such that $\tilde{U}(x) = c_1(1-A_1(x))$ for all x. Substituting into (2.16) and rearranging terms, we deduce

$$A_1(x+y) \quad = \quad A_1(x)A_1(y)$$

for all x and y. By Lemma 2.3, $A_1(x) = e^{\rho x}$ where $\rho \neq 0$ because we excluded the case $A_1(x) \equiv 1$. If $\rho > 0$ then G is of the same type as (2.3) with $\alpha = 1/\rho$; if $\rho < 0$ then G is of the same type as (2.4) with $\alpha = -1/\rho$. This concludes the proof.

2.2 Domains of Attraction

2.2.1 Von MIses Conditions

Now that we have characterized the possible extreme value distributions, the next step is to develop general conditions on a particular F such that (2.1) holds. More specifically, we would like to show how to find constants a_n and b_n for common distributions (e.g. normal, lognormal, beta, gamma, Pareto, etc.) so that $\frac{M_n - b_n}{a_n}$ converges in distribution, as well as determining the limit itself. Particularly useful in this regard are the *von Mises conditions*, due originally to von Mises [19], which are a set of simple sufficient though not necessary conditions for the domain of attraction. The theory given here may be regarded as a modern reinterpretation of von Mises.

We consider only absolutely continuous distributions. In practice nearly all continuous distributions are in the domain of attraction of some extreme value limit.

Suppose, then F is absolutely continuous, with density $f(x) = \frac{dF(x)}{dx}$ existing and positive on a range $x_* < x < x^*$ ($x_* \geq -\infty$, $x^* \leq +\infty$).

Define

$$\phi(x) \quad = \quad \frac{1-F(x)}{f(x)}, \;\; x_* < x < x^*, \tag{2.18}$$

and suppose that ϕ is continuously differentiable. We may write

$$1-F(x) \quad = \quad \exp\left\{ -\int_{x_*}^{x} \frac{dt}{\phi(t)} \right\}, \;\; x_* < x < x^*. \tag{2.19}$$

Consider the ratio

$$\begin{aligned} \frac{1-F(u+x\phi(u))}{1-F(u)} &= \exp\left\{ -\int_{u}^{u+x\phi(u)} \frac{dt}{\phi(t)} \right\} \\ &= \exp\left\{ -\int_{0}^{x} \frac{\phi(u)}{\phi(u+s\phi(u))} ds \right\}. \end{aligned} \tag{2.20}$$

When $x > 0$ (which we are not necessarily assuming), equation (2.20) is the conditional probability that $X - u > x\phi(u)$ given $X > u$, when X is a random variable with the distribution function F.

Now consider the ratio

$$\frac{\phi(u+s\phi(u))}{\phi(u)} \quad = \quad 1 + \int_0^s \phi'(u+w\phi(u)) dw.$$

By the Mean Value Theorem, this is $1 + s\phi'(u+\theta s\phi(u))$ where $\theta = \theta(s,u)$ is between 0 and 1. By considering

$$\int_0^x \left\{ \frac{\phi(u)}{\phi(u+s\phi(u))} - \frac{1}{1+s\phi'(u+t\phi(u))} \right\} ds$$

as a function of t between 0 and x, it is continuous (because ϕ' is), and takes on both

positive and negative values (unless ϕ' is constant). Hence it is 0 for at least one t. Thus there exists y between u and $u+x\phi(u)$ such that

$$\frac{1-F(u+x\phi(u))}{1-F(u)} \quad = \quad \exp\left\{-\frac{\log(1+x\phi'(y))}{\phi'(y)}\right\}. \tag{2.21}$$

(We interpret the right side of (2.21) as e^{-x} if $\phi'(y)=0$). Now suppose

$$\lim_{x\to x^*}\phi'(x) \quad = \quad \xi. \tag{2.22}$$

Suppose we fix x and let $u\to x^*$, $u+x\phi(u)\to x^*$. Then

$$\frac{1-F(u+x\phi(u))}{1-F(u)} \quad \to \quad \begin{cases}(1+\xi x)^{-1/\xi}, & \text{if } \xi\neq 0,\\ e^{-x}, & \text{if } \xi=0.\end{cases} \tag{2.23}$$

For $x>0$, we may think of u as a threshold value and (2.23) gives the limiting distribution of *excesses over the threshold*; this is called the *Generalized Pareto* family following Pickands [21], who showed that ((2.23) holds in general if and only if F is in the domain of attraction of an extreme value distribution with index ξ (see (2.6). The range of x is $0<x<\infty$ if $\xi\geq 0$ and $0<x<-1/\xi$ if $\xi<0$.

For the purposes of the present section, we define b_n by $F(b_n)=1-1/n$ and set $a_n=\phi(b_n)$. Then

$$\lim_{n\to\infty} n\{1-F(a_nx+b_n)\} = \begin{cases}(1+\xi x)^{-1/\xi}, & \text{if } \xi\neq 0,\\ e^{-x}, & \text{if } \xi=0,\end{cases}$$

and so

$$\lim_{n\to\infty} F^n(a_nx+b_n) = \begin{cases}\exp\{-(1+\xi x)^{-1/\xi}\}, & \text{if } \xi\neq 0,\\ \exp(-e^{-x}), & \text{if } \xi=0.\end{cases} \tag{2.24}$$

To summarize: if ϕ is defined by (2.18) and if (2.22) holds, then (2.24) holds with $b_n=F^{-1}(1-1/n)$ and $a_n=\phi(b_n)=1/\{nf(b_n)\}$.

This argument has skated over one point: the range of values of x for which (2.24) holds. If $x^*=+\infty$ then (2.22) implies

$$\lim_{x\to\infty}\frac{\phi(x)}{x} \quad = \quad \xi. \tag{2.25}$$

Hence, for any $x>-1/\xi$, we have that $u\to\infty$ implies $u+x\phi(u)\to\infty$ and (2.24) holds. If $x^*<\infty$, then $\phi(x)\to 0$ and (2.22) implies

$$\lim_{x\to x^*}\frac{\phi(x)}{x^*-x} \quad = \quad -\xi. \tag{2.26}$$

In this case (2.24) is valid on $-\infty<x<1/\xi$.

It may also be noted that, if $\xi>0$, then by (2.25) we have $a_n\sim\xi b_n$ and it does

not change the asymptotics if we define $a_n = \xi b_n$ instead of $a_n = \phi(b_n)$. In this case (2.24) becomes

$$\lim_{n\to\infty} F^n\{b_n(1+\xi x)\} \quad = \quad \exp\{-(1+\xi x)^{-1/\xi}\}.$$

If we now replace b_n by a_n, ξ by $\alpha = 1/\xi$ and $1+\xi x$ by x, we get $F^n(a_n x) \to \Phi_\alpha(x)$ as $n \to \infty$. This shows that it is possible to recover the conventional limit results from the theory given here, at least under the smoothness assumptions made about F.

Similarly, if $\xi < 0$, we may define $a_n = -\xi(x^* - b_n)$ when (2.24) becomes

$$\lim_{n\to\infty} F^n\{x^* - (1+\xi x)(x^* - b_n)\} = \exp\{-(1+\xi x)^{-1/\xi}\}$$

whenever $1+\xi x > 0$. Redefining $a_n = x^* - b_n$, $b_n = x^*$, $y = -(1+\xi x)$, $\alpha = -1/\xi$, we have

$$\lim_{n\to\infty} F^n(a_n y + b_n) \quad = \quad \exp\{-(-y)^\alpha\} \text{ for } y < 0.$$

2.2.2 Sample minima.

The theory for sample minima is a mirror image of the theory for sample maxima: just replace $F(x)$ by $1 - F(-x)$ everywhere. It is convenient, however, to have the main results for sample minima stated separately, so this will be done here.

Suppose F is a continuous distribution function with range (x_*, x^*), let $f(x) = dF(x)/dx$ and define

$$\phi(x) \quad = \quad \frac{F(x)}{f(x)}. \tag{2.27}$$

We now assume

$$\lim_{x\to x_*} \phi'(x) \quad = \quad -\xi. \tag{2.28}$$

Define b_n by $F(b_n) = 1/n$, $a_n = \phi(b_n)$. Then

$$\lim_{n\to\infty} \{1 - F(a_n x + b_n)\}^n = \begin{cases} \exp\{-(1-\xi x)^{-1/\xi}\}, & \text{if } \xi \neq 0, \\ \exp(-e^x), & \text{if } \xi = 0. \end{cases}$$

The range of x is $-\infty < x < 1/\xi$ for $\xi > 0$, $1/\xi < x < \infty$ for $\xi < 0$. If $\xi > 0$, then necessarily $x_* = -\infty$ and an alternative scheme is to set $a_n = -F^{-1}(1/n)$, $b_n = 0$, $\alpha = 1/\xi$ when

$$\lim_{n\to\infty} \{1 - F(a_n x + b_n)\}^n \quad = \quad \exp\{-(-x)^{-\alpha}\}, \quad x < 0.$$

If $\xi < 0$, then $x_* > -\infty$ and with $b_n = x_*$, a_n defined by $F(x_* + a_n) = 1/n$ and $\alpha = -1/\xi$ we have

$$\lim_{n\to\infty} \{1 - F(a_n x + b_n)\}^n \quad = \quad \exp(-x^\alpha), \quad x > 0.$$

Corresponding to (2.25) and (2.26) we have

$$\begin{aligned} \lim_{x\to-\infty}\frac{\phi(x)}{x} &= -\xi, \quad \xi>0, \\ \lim_{x\to x_*}\frac{\phi(x)}{x-x_*} &= -\xi, \quad \xi<0. \end{aligned} \tag{2.29}$$

2.2.3 Examples

1. *Pareto-type distributions.* Suppose $1-F(x)\sim cx^{-\alpha}$ as $x\to\infty$ ($c>0$, $\alpha>0$) and that this asymptotic relation remains valid under at least two differentiations, i.e. $f(x)\sim\alpha cx^{-\alpha-1}$, $f'(x)\sim-\alpha(\alpha+1)cx^{-\alpha-2}$. Then $\phi(x)\sim x/\alpha$, $\phi'(x)\to 1/\alpha$ so (2.22) and (2.25) are satisfed with $\xi=1/\alpha$. Defining a_n either by $F(a_n)=1-1/n$ or else $a_n=(nc)^{1/\alpha}$, we have $F^n(a_nx)\to\Phi_\alpha(x)$. Examples include the Pareto distributions of the first and second kind [15], the Cauchy, t and F distributions.
2. Suppose $F(x)\sim cx^{\alpha}$ as $x\downarrow 0$ (and $f(x)\sim\alpha cx^{\alpha-1}$, $f'(x)\sim\alpha(\alpha-1)cx^{\alpha-2}$) and consider the distribution of sample minima. Defining ϕ by (2.27), we have that (2.28) and (2.29) are satisfied with $\xi=-1/\alpha$. Defining a_n either by $F(a_n)=1/n$ or else $a_n=(nc)^{-1/\alpha}$, we have $\{1-F(a_nx)\}^n\to\exp(x^\alpha)$, $0<x<\infty$. Examples include the uniform and exponential distributions, and more generally anything in either the beta or gamma classes.
3. *Normal extremes.* Let $\Phi(x)=(2\pi)^{-1/2}\int_{-\infty}^{x}\exp(-t^2/2)dt$ denote the standard normal ditribution function. From the well-known expansion

$$1-\Phi(x) = (2\pi)^{-1/2}e^{-x^2/2}(x^{-1}-x^{-3}+3x^{-5}-15x^{-7}+....) \tag{2.30}$$

(see, for example [6], page 193) it follows that

$$\phi(x)=x^{-1}-x^{-3}+...$$

Note that, as a notational point, we are using ϕ in the same context as previously and *not* to denote the standard normal density as is very often done. In this case $\phi'\sim-x^{-2}$ which tends to 0, so we are in the Gumbel domain of attraction. Defining b_n by $\Phi(b_n)=1-1/n$, $a_n=\phi(b_n)$ or $a_n=1/b_n$, we have

$$\Phi^n(a_nx+b_n)\to\Lambda(x).$$

It is possible to replace b_n by the closed form expression

$$b_n = (2\log n)^{1/2}-\frac{1}{2}(2\log n)^{-1/2}\{\log\log n+\log(4\pi)\} \tag{2.31}$$

which can be derived as an asymptotic approximation from (2.30). This is not the best choice from the point of view of rate of convergence [14], but it can be useful in obtaining expressions for the growth rate of normal extremes. The first order of approximation is that the largest of n standard normals is asymptotic to $(2\log n)^{1/2}$ in probability.

4. *Lognormal extremes.* Let $F(x) = \Phi(\sigma^{-1}\log x)$ for $x > 0$, where Φ is the standard normal distibution function. For maxima, define ϕ by (2.18); then $\phi(x) \sim \sigma^2 x(\log x)^{-1}$ by (2.30), and $\phi'(x) \sim \sigma^2(\log x)^{-1} \to 0$. So we are in the domain of attraction of Gumbel; if we define B_n by $\Phi(B_n) = 1 - 1/n$ then suitable normalizing constants for the lognormal are $b_n = \exp(\sigma B_n)$, $a_n = \sigma^2 b_n(\log b_n)^{-1} = \sigma b_n/B_n$. For minima, define ϕ by (2.27), then $\phi(x) \sim -\sigma^2 x(\log x)^{-1}$ as $x \to 0$; now define $b_n = \exp(-\sigma B_n)$, $a_n = \sigma b_n/B_n$ to get (2.26) with $\xi = 0$. This is useful as an explicit example for which the range of the distribution is bounded below, but the limiting extreme value distribution for minima is still of the Gumbel form.
5. *Continuous distributions not in any domain of attraction.* These are hard to construct but they do exist! Here are two examples:
   (a) Any distribution function with slowly varying tail, for instance $F(x) = 1 - 1/\log x$ for $x > e$.
   (b) Consider $F_\delta(x) = 1 - x^{-1}\{1 + \delta\sin(\log x)\}$ valid for $x \geq$ some x_0, with $|\delta|$ small enough to make it a valid distribution function. Resnick [22] proved the curious result that F_δ is not in any domain of attraction, but the product $F_\delta F_{-\delta}$ is.

Background. Von Mises' conditions as they are usually stated are (2.24), (2.25) and (2.26) respectively for the cases $\xi = 0$, $\xi > 0$, $\xi < 0$. These are known to be sufficient but not necessary for the domain of attraction. The single condition (2.22) (for all ξ) is also sometimes referred to as von Mises' condition, though it is a little more restrictive than (2.25) or (2.26) when $\xi \neq 0$. The calculations in this section form useful background to the more detailed discussion of rates of convergence which will follow. For a short proof of the sufficiency of the von Mises conditions, see de Haan [12].

2.2.4 Penultimate approximations

Let us go back to (2.21). In the preceding section we replaced $\phi'(y)$ by ξ, defined by (2.22), and deduced (2.23). However, in (2.21) we know y is close to u, and a better approximation might result if we replace $\phi'(y)$ by $\phi'(u)$ rather than its "ultimate" limit ξ. Defining a_n and b_n as before, and

$$\xi_n \quad = \quad \phi'(b_n),$$

(2.24) is replaced by

$$\lim_{n\to\infty} |F^n(a_n x + b_n) - \exp\{-(1+\xi_n x)^{-1/\xi_n}\}| \quad = \quad 0 \qquad (2.32)$$

for each real x and hence uniformly over all x.

Whereas (2.24) represents the ultimate approximation in the sense that it gives a single extreme value distribution to which the extremes are eventually attracted, (2.32) defines a *sequence* of approximating distributions within the extreme value family. This has been termed the "penultimate approximation". The original motivation for this was given by Fisher and Tippett [7]. They proved that normal extremes

are attracted to Λ, but showed numerically that a much better approximation was within the Type III family, with $\xi < 0$. Recall that $\phi' < 0$ in the normal case.

Cohen [3] proved that, in the normal case, the penultimate approximation has a faster rate of convergence than the ultimate approximation. Cohen [2] and Gomes [10] showed that the same is true for a very wide class of distributions in the domain of attraction of Λ.

As an example, let us return to the case of normal extremes [3]. Using (2.30), we have

$$\begin{aligned}\frac{1-\Phi(u+x\phi(u))}{1-\Phi(u)} &= \left(1+\frac{x\phi(u)}{u}\right)^{-1}\cdot\exp\left[-\frac{(u+x\phi(u))^2}{2}+\frac{u^2}{2}\right]\cdot \\ &\quad\cdot\left\{1-\frac{1}{((u+x\phi(u))^2}+\frac{1}{u^2}+O\left(\frac{1}{u^4}\right)\right\}\end{aligned}$$

and hence

$$\begin{aligned}\log\left[\frac{1-\Phi(u+x\phi(u))}{1-\Phi(u)}\right] &= -\log\left(1+\frac{x\phi(u)}{u}\right)-x\phi(u)-\frac{1}{2}x^2\phi^2(u)+O\left(\frac{1}{u^4}\right) \\ &= -\frac{x\phi(u)}{u}+\frac{x^2\phi^2(u)}{u^2}-x\left(1-\frac{1}{u^2}\right)-\frac{x^2}{2u^2}+O\left(\frac{1}{u^4}\right) \\ &= -x-\frac{x^2}{2u^2}+O\left(\frac{1}{u^4}\right) \\ &= u^2\log\left(1-\frac{x}{u^2}\right)+O\left(\frac{1}{u^4}\right)\end{aligned}$$

where in the third term of the second line we used $u\phi(u) = 1-u^{-2}+O(u^{-4})$ and elsewhere simply $u\phi(u)\to 1$ as $u\to\infty$. Hence

$$\frac{1-\Phi(u+x\phi(u))}{1-\Phi(u)} = \left(1-\frac{x}{u^2}\right)^{u^2}+O\left(\frac{1}{u^4}\right) \tag{2.33}$$

$$= e^{-x}+O\left(\frac{1}{u^2}\right) \tag{2.34}$$

where (2.34) follows from (2.33) by the well-known result $\left(1-\frac{x}{n}\right)^n\to e^{-x}$ as $n\to\infty$.

These results may be interpreted in two ways. First, if we define b_n by $1-\Phi(b_n) = 1/n$ and $a_n = \phi(b_n)$, then (2.34) shows

$$n\{1-\Phi(a_nx+b_n)\} = e^{-x}+O\left(\frac{1}{b_n^2}\right)$$

and hence

$$\Phi^n(a_nx+b_n) = e^{-e^{-x}}+O\left(\frac{1}{b_n^2}\right)$$

pointwise for each x and ultimately uniformly over all x. This is the "ultimate approximation".

However, if we also define $\xi_n = -1/b_n^2$, (2.33) shows

$$n\{1-\Phi(a_n x+b_n)\} \quad = \quad (1+\xi_n)^{-1/\xi_n} + O\left(\frac{1}{b_n^4}\right)$$

and hence the "penultimate approximation" $\Phi^n(a_n x+b_n) \approx \exp\left\{-(1+\xi_n x)^{-1/\xi_n}\right\}$ also has error $O\left(\frac{1}{b_n^4}\right)$.

Since b_n also satisfies (2.31), we have finally that the ultimate approximation has error $O(\frac{1}{\log n})$ and the penultimate approximation has error $O(\frac{1}{\log^2 n})$. This result (with uniformity over all x) was first proved rigorously by Cohen [3].

The implication of this result is that, even though the Gumbel distribution (2.5) is the "ultimate" limit, the penultimate approximation with $\xi_n = -1/b_n^2$ is a better approximation in practice. Since $\xi_n < 0$, this is therefore of Weibull type (2.4), with $\alpha = -1/\xi_n$. From a statistical point of view, it's better to fit the three-parameter GEV even though this will approximate the Gumbel distribution for large n. Such results extend beyond the case of the normal distribution [2], and have contributed to the virtual disappearance of the Gumbel distribution from practical extreme value analysis.

2.3 Regular variation and domains of attraction

The previous section has shown a simple set of conditions, due originally to von Mises, that are adequate in the vast majority of cases for determining convergence of sample extremes to an extreme value limit. But for a full understanding of the theory, and as background to our eventual exploration of multivariate and spatial/temporal extremes, we need to introduce the key mathematical concept of regular variation.

There are by now numerous books on regular variation: the classic by Bingham, Goldie and Teugels [1] is the most comprehensive, but numerous books on extreme value theory, including those by Resnick [23, 24] and de Haan [11, 13] include full detailed treatments.

2.3.1 Regular Variation

Definition 2.3. A measurable function $h:\ \mathbb{R}_+ \to \mathbb{R}_+$ is regularly varying at infinity if there exists a function $\psi:\ \mathbb{R}_+ \to \mathbb{R}_+$ such that

$$\lim_{x\to\infty} \frac{h(tx)}{h(x)} \quad = \quad \psi(t) \text{ for all } t > 0. \tag{2.35}$$

If (2.35) holds, then by letting $x \to \infty$ in the relationship

$$\frac{h(tsx)}{h(x)} \quad = \quad \frac{h(tsx)}{h(sx)} \cdot \frac{h(sx)}{h(x)}$$

we deduce

$$\psi(ts) \quad = \quad \psi(t)\psi(s)$$

for all $s,t \in \mathbb{R}_+$, and hence by the Cauchy functional equation (Lemma 2.3), $\psi(t) = t^\rho$ for some $\rho \in \mathbb{R}$. When $\rho = 0$ the function is called *slowly varying*.

When (2.35) we write $h \in \mathrm{RV}_\rho$. Some basic properties are:

(a) If (2.35) holds for each $t > 0$, then the convergence is uniform on compact sets.

(b) The Karamata representation: if h satisfies (2.35), for x greater than some x_0 we have

$$h(x) \quad = \quad c(x) \exp\left\{ \int_{x_0}^{x} \frac{\rho(t)}{t} dt \right\}$$

where $c(x)$ is some measurable function that converges to a limit $c \in (0,\infty)$ are $x \to \infty$, and $\rho(x) \to \rho$ as $x \to \infty$.
A corollary of this result is that if $h \in \mathrm{RV}_\rho$, there exists some $h_* \in \mathrm{RV}_\rho$ which is infinitely differentiable and such that $h(x) \sim h_*(x)$ as $x \to \infty$ (in the sense that h/h^* tends to 1).

(c) Inverses of regularly varying functions: if $h \in \mathrm{RV}_\rho$ with $\rho > 0$ and if we define an inverse function $h^{-1}(y) \;=\; \inf\{t :\; h(t) > y\}$ then $h^{-1} \in \mathrm{RV}_{1/\rho}$.

2.3.2 Domains of Attraction

Now suppose the distribution function F is such that (2.1) holds for suitable $a_n > 0$, $b_n \in \mathbb{R}$ and distribution function G. We then say *F is in the domain of (maximum) attraction of G*, formally $F \in \mathscr{D}(G)$. We already know by the Three Types Theorem that G must be the same type as one of (2.3) through (2.5), or equivalently, of the same type as H_ξ, where $H_\xi(x) = \exp\left\{ -(1+\xi x)_+^{-1/\xi} \right\}$, some $\xi \in \mathbb{R}$. As everywhere, we interpret the case $\xi = 0$ as the limit $\xi \to 0$, in which case $H_0(x) = \exp(-e^{-x})$.

First, we state a lemma, which in some respects is an extension of Lemma 2.1:

Lemma 2.4. Suppose $\{F_n,\ n = 1,2,...,\}$ is a sequence of distribution functions and G is some nondegenerate limit. Let F_n^{-1} and G^{-1} be inverses of F_n and G. Then (2.7) holds, for some $a_n > 0$ and $b_n \in \mathbb{R}$m if and only if

$$\frac{F_n^{-1}(p) - F_n^{-1}(p_1)}{F_n^{-1}(p_2) - F_n^{-1}(p_1)} \quad \stackrel{w}{\to} \quad \frac{G^{-1}(p) - G^{-1}(p_1)}{G^{-1}(p_2) - G^{-1}(p_1)} \tag{2.36}$$

where $0 \le p_1 < p_2 \le 1$ are continuity points of G^{-1} and $p \in [0,1]$. Moreover, when (2.36) holds, we may take a_n, b_n to be defined by

$$a_n \;=\; \frac{F_n^{-1}(p_2) - F_n^{-1}(p_1)}{G^{-1}(p_2) - G^{-1}(p_1)}, \quad b_n \;=\; F_n^{-1}(p_1) - a_n G^{-1}(p_1).$$

Proof. We already saw in the proof of Lemma 2.1 that (2.7) is equivalent to

$$\frac{F_n^{-1}(p) - b_n}{a_n} \quad \stackrel{w}{\to} \quad G^{-1}(p). \tag{2.37}$$

The left side of (2.36) is

$$\frac{F_n^{-1}(p)-F_n^{-1}(p_1)}{a_n}+G^{-1}(p_1) \quad = \quad \frac{F_n^{-1}(p)-F_n^{-1}(p_1)}{F_n^{-1}(p_2)-F_n^{-1}(p_1)}\cdot\{G^{-1}(p_2)+G^{-1}(p_1)\}+G^{-1}(p_1).$$

Hence, (2.37) is equivalent to (2.36). □

Define

$$H_\xi(x) \quad = \quad \begin{cases} \exp\left\{-(1+\xi x)_+^{-1/\xi}\right\} & \text{if } \xi\neq 0, \\ \exp\left(-e^{-x}\right) & \text{if } \xi = 0. \end{cases} \tag{2.38}$$

and also let $h_\xi(t)=(t^\xi-1)/(e^\xi-1)$, or $\log t$ when $\xi=0$.

Theorem 2.2. $F\in\mathscr{D}(H_\xi)$ if and only if

$$\lim_{n\to\infty}\frac{F^{-1}\left(1-\frac{1}{nt}\right)-F^{-1}\left(1-\frac{1}{n}\right)}{F^{-1}\left(1-\frac{1}{ne}\right)-F^{-1}\left(1-\frac{1}{n}\right)} \quad = \quad h_\xi(t) \text{ for all } t>0. \tag{2.39}$$

If (2.39) holds, then if suffices to take

$$a_n \;=\; \frac{F^{-1}\left(1-\frac{1}{ne}\right)-F^{-1}\left(1-\frac{1}{n}\right)}{(e^\xi-1)/\xi}, \quad b_n \;=\; F^{-1}\left(1-\frac{1}{n}\right).$$

Proof. Let $H_\xi^{-1}(p)=((-\log p)^{-\xi}-1)/\xi$ if $\xi\neq 0$, or $-\log(-\log p)$ if $\xi=0$, and let $p_1=e^{-1},\ p_2=e^{-1/e},\ p=e^{-1/t}$. Then Lemma 2.4 implies that (2.6) holds (with $G=H_\xi$) if and only if

$$\lim_{n\to\infty}\frac{F_n^{-1}(e^{-1/t})-F_n^{-1}(e^{-1})}{F_n^{-1}(e^{-1/e})-F_n^{-1}(e^{-1})} \quad = \quad \frac{H_\xi^{-1}(e^{-1/t})-H_\xi^{-1}(e^{-1})}{H_\xi^{-1}(e^{-1/e})-H_\xi^{-1}(e^{-1})} \;=\; h_\xi(t) \tag{2.40}$$

where F_n is the same as F^n. Moreover, $F_n^{-1}(p)=F^{-1}(p^{1/n})$. If we define t_n by $\left(1-\frac{1}{nt_n}\right)^n=e^{-1/t}$ (this implies $t_n\to t$) then as $n\to\infty$, (2.40) is equivalent to (2.39). □

Let us now offer an alternative interpretation of the condition (2.39). Define U to be the inverse function of $1/(1-F)$: thus $U(y)=F^{-1}(1-1/y)$. Also, since F and hence U are monotone, there is no change in the condition if we make n a continuous variable in (2.39) and write x in place of n. With those changes, (2.39) is the same as

$$\lim_{x\to\infty}\frac{U(xt)-U(x)}{U(xe)-U(x)} \quad = \quad h_\xi(t) \text{ for all } t>0. \tag{2.41}$$

Now let us interpret some special cases of (2.41).

Case 1. Suppose $\xi>0$. Write $\alpha=1/\xi$. Rewriting (2.41) in the form $\frac{U(xt)/U(x)-1}{U(xe)/U(x)-1}\to(t^\xi-1)/(e^\xi-1)$, we see this is equivalent to $U(xt)/U(x)\to t^\xi$, in

other words, U is regularly varying with index ξ. Since U is the inverse of $1/(1-F)$, this implies that $1/(1-F)$ is regularly varying with index α. This in turn implies that

$$\lim_{x\to\infty} \frac{1-F(tx)}{1-F(x)} \quad = \quad x^{-\alpha}$$

which is the classical condition for the domain of attraction of Φ_α first proved by Gnedenko [9].

Case 2. Suppose $\xi < 0$. This corresponds to the limit Ψ_α with $\alpha = -1/\xi$. In this case the classical necessary and sufficient condition of Gnedenko [9] is that $x_* = \sup\{x:\ F(x) < 1\} < \infty$ and $1-F(x^* - 1/t) \in \mathrm{RV}_{-\alpha}$ as $t \to \infty$. This is equivalent to saying that

$$V(t) \quad = \quad \frac{1}{1-F(x_* - 1/t)} \ \in \ \mathrm{RV}_\alpha$$

and hence for the inverse function, $V^{-1} \in \mathrm{RV}_{1/\alpha}$.

Now, $U(x) = x_* - 1/t$ at a continuity point if and only if $1/\{1-F(x_* - 1/t)\} = x = V(t)$ and hence $t = V^{-1}(x)$. So $U(x) = x_* - 1/V^{-1}(x)$. Hence

$$\begin{aligned}
\lim_{x\to\infty} \frac{U(xt)-U(x)}{U(xe)-U(x)} &= \lim_{x\to\infty} \frac{1/V^{-1}(xt) - 1/V^{-1}(x)}{1/V^{-1}(xe) - 1/V^{-1}(x)} \\
&= \lim_{x\to\infty} \frac{V^{-1}(x)/V^{-1}(xt) - 1}{V^{-1}(x)/V^{-1}(xe) - 1} \\
&= \frac{t^{-1/\alpha} - 1}{e^{-1/\alpha} - 1}
\end{aligned}$$

which is (2.41).

2.4 Lecture of Feb 11

Start by assuming

$$\frac{U(tx)-U(x)}{a(x)} \quad = \quad \frac{t^\xi - 1}{\xi} + A(x)H(t) + o(A(x)) \tag{2.42}$$

Let $b_n = a(n),\ a_n = a(n)$,

$$\frac{U(nt)-b_n}{a_n} \quad = \quad \frac{t^\xi - 1}{\xi} + A(n)H(t) + o(A(n)).$$

Solving $y = \frac{U(nt)-b_n}{a_n}$ means $U(nt) = a_n y + b_n$ so $F(a_n y + b_n) = 1 - 1/(nt)$ so

$$n(1-F(a_n y + b_n)) \quad = \quad \frac{1}{t}. \tag{2.43}$$

This also means

$$\begin{aligned} n\log F(a_n y+b_n) &= n\log\{1-(1-F(a_n y+b_n))\} \\ &= n\log\left(1-\frac{1}{nt}\right) \\ &= -\frac{1}{t}+O\left(\frac{1}{n}\right) \end{aligned}$$

and hence

$$F^n(a_n y+b_n) = e^{-1/t}+O\left(\frac{1}{n}\right). \tag{2.44}$$

Our objective is to solve for t as a function of y. We start from

$$y = \frac{t^\xi-1}{\xi}+A(n)H(t)+o(A(n)).$$

First guess is to ignore $A(n)$ altogether: $t=(1+\xi y)^{1/\xi}$.

Second guess: write $t=(1+\xi y)^{1/\xi}(1+\varepsilon)$. In that case,

$$\frac{t^\xi-1}{\xi} = y+\varepsilon(1+\xi y)+o(\varepsilon)$$

so

$$y = y+\varepsilon(1+\xi y)+A(n)\cdot H((1+\xi y)^{-1/\xi})+...$$

and hence

$$\varepsilon \sim \frac{A(n)\cdot H((1+\xi y)^{-1/\xi})}{(1+\xi y)}.$$

Hence $t^{-1}=(1+\xi y)^{-1/\xi}(1-\varepsilon+o(\varepsilon))$ and so

$$\frac{1}{t} = (1+\xi y)^{-1/\xi}-\frac{A(n)\cdot H((1+\xi y)^{-1/\xi})}{(1+\xi y)}+o(A(n)). \tag{2.45}$$

Combining (2.45) with (2.43) or (2.44) gives our desired approximation. In most cases $A(n) >> 1/n$ so the $O(1/n)$ error in (2.44) does not affect the quality of the approximation.

Example 1. Suppose F satisfies

$$1-F(x) = cx^{-\alpha}\left(1+dx^{-\beta}+o(x^{-\beta})\right) \tag{2.46}$$

as $x\to\infty$, where $\alpha>0,\ \beta>0,\ c>0$ and $d\in\mathbb{R}$. In that case,

$$\begin{aligned} \frac{1}{1-F(x)} &= c^{-1}x^{\alpha}\left(1-dx^{-\beta}+o(x^{-\beta})\right) \\ &= y \text{ say.} \end{aligned}$$

Solve for x: the first-order solution is $x = (cy)^{1/\alpha}$ so we aim to improve on that by writing $x = (cy)^{1/\alpha}(1+\varepsilon)$. By writing

$$y = c^{-1}\cdot cy\cdot(1+\alpha\varepsilon+o(\varepsilon)\cdot\left\{1-d(cy)^{-\beta/\alpha}+o(y)^{-\beta/\alpha})\right\}$$

we deduce $\varepsilon \sim \frac{d}{\alpha}(cy)^{-\beta/\alpha}$ and hence

$$U(y) = (cy)^{1/\alpha}\left\{1+\frac{d}{\alpha}(cy)^{-\beta/\alpha}+o(y)^{-\beta/\alpha})\right\}.$$

After a little manipulation,

$$\frac{U(tx)-U(t)}{(cx)^{1/\alpha}/\alpha} = \alpha\left(t^{1/\alpha}-1\right)+d(cx)^{-\beta/\alpha}\left(t^{(1-\beta)/\alpha}-1\right)+o(x^{-\beta/\alpha}).$$

If we write $a(x) = (cx)^{1/\alpha},\ \xi = 1/\alpha,\ \rho = -\beta/\alpha$, we have

$$\frac{U(tx)-U(t)}{a(x)} = \frac{t^{\xi}-1}{\xi}+d(cx)^{\rho}(t^{\xi+\rho}-1)+o(x^{\rho})$$

which is exactly of the form (2.42).

Example 2: Normal Extremes. Suppose $F = \Phi$, the CDF of the standard normal distribution.

Fix some large y and $s \in \mathbb{R}_+$. Asymptotics will be as $y \to \infty$. Then

$$\begin{aligned}
F\left(y+\frac{s}{y}\right)-F(y) &= \frac{1}{\sqrt{2\pi}}\int_y^{y+s/y} e^{-t^2/2}dt \\
&= \frac{1}{y\sqrt{2\pi}}\int_0^s \exp\left\{-\frac{1}{2}\left(y+\frac{u}{y}\right)^2\right\}du \\
&= \frac{e^{-y^2/2}}{y\sqrt{2\pi}}\int_0^s e^{-u}\cdot\exp\left\{-\frac{1}{2}\left(\frac{u}{y}\right)^2\right\}du \\
&= \frac{e^{-y^2/2}}{y\sqrt{2\pi}}\int_0^s e^{-u}\cdot\left\{1-\frac{1}{2}\left(\frac{u}{y}\right)^2+o\left(\frac{1}{y^2}\right)\right\}du \\
&= \frac{e^{-y^2/2}}{y\sqrt{2\pi}}\left\{1-e^{-s}-\frac{1}{2y^2}(2-e^{-s}(2+2s+s^2))+o\left(\frac{1}{y^2}\right)\right\}.
\end{aligned}$$

Suppose we are given $1-F(y) = 1/x$ and $1-F(y+s/y) = 1/(tx)$ — our objective will be to determine s as a function of t with x and y fixed. Thus

$$\frac{1}{x}\left(1-\frac{1}{t}\right) = \frac{e^{-y^2/2}}{y\sqrt{2\pi}}\left\{1-e^{-s}-\frac{1}{2y^2}(2-e^{-s}(2+2s+s^2))+o\left(\frac{1}{y^2}\right)\right\}.$$

Substituting from (2.30),

$$\frac{e^{-y^2/2}}{y\sqrt{2\pi}}\left\{1-\frac{1}{y^2}+o\left(\frac{1}{y^2}\right)\right\}\left(1-\frac{1}{t}\right) = \frac{e^{-y^2/2}}{y\sqrt{2\pi}}\left\{1-e^{-s}-\frac{1}{2y^2}(2-e^{-s}(2+2s+s^2))+o\left(\frac{1}{y^2}\right)\right\}$$

and hence

$$\left(1-\frac{1}{t}\right) \quad = \quad \left\{1-e^{-s}+\frac{e^{-s}(2+2s+s^2)}{2y^2}+o\left(\frac{1}{y^2}\right)\right\}.$$

Solving for s, asymptotically with $y \to \infty$,

$$s \quad = \quad \log t\left\{1-\frac{2+2\log t+\log^2 t}{2y^2\log t}+o\left(\frac{1}{y^2}\right)\right\}.$$

If we write $U(x)=y$ and $U(tx)=y+\frac{s}{y}$, then

$$\begin{aligned} U(tx)-U(x) &= \frac{s}{y} \\ &= \frac{\log t}{U(x)}\left\{1-\frac{2+2\log t+\log^2 t}{2y^2\log t}+o\left(\frac{1}{y^2}\right)\right\}. \end{aligned}$$

Defining $a(x)=1/U(x)$,

$$\frac{U(tx)-U(x)}{a(x)} \quad = \quad \log t\left\{1-\frac{2+2\log t+\log^2 t}{2U(x)^2\log t}+o\left(\frac{1}{U(x)^2}\right)\right\}.$$

Then with $A(x)=\frac{1}{U(x)^2}$, $H(t)=-\left(1+\log t+\frac{1}{2}\log^2 t\right)$, we have

$$\frac{U(tx)-U(x)}{a(x)} \quad = \quad \log t+A(x)H(t)+o(A(x))$$

which is exactly of the form (2.42) (with $\xi=0$).

Chapter 3

Joint Distributions of Extremes and Point Processes

Our objective in this chapter will be to study the limiting behavior of extreme order statistics in more detail, including the joint distribution of several largest or smallest extremes. This will be achieved by embedding the process of extremes in a limiting point process, also known as the *extremal process*, and then applying results from point process theory. Other books that have made extensive use of point processes in extreme value theory include [18, 23, 24].

In this chapter we will particularly follow the approach due to Weissman [27, 28]. Other papers proving similar results are due to Pickands [20] and Leadbetter [16, 17]; the latter covered generalizations to dependent stationary sequences.

3.1 Asymptotic distribution of kth largest order statistic

Suppose we have a sequence of independent random variables $X_1,\ X_2,\ \ldots$ — not necessarily identically distributed, though in many cases they will be. Write the order statistics from the first n observations in the form $X_{1:n} \geq X_{2:n} \geq \ldots \geq X_{n:n}$. Natural questions to ask include

1. What is the asymptotic distribution of $X_{k:n}$?
2. What is the asymptotic joint distribution of $(X_{1:n}, X_{2:n}, \ldots, X_{k:n})$ for some given k?

For both these questions, we focus here on limit as $n \to \infty$ for fixed k. Alternative limiting forms (such such as $k = k_n \to \infty$ as $n \to \infty$ with $k_n/n \to p$, some $p \in (0,1]$) are also of interest but involve a different kind of theory from that considered here.

Let $I_n(x) = \sum_{i=1}^{n} I(X_i > x)$. Here $I(\cdot)$ is the indicator function (1 if the expression inside the parentheses is true, 0 otherwise). This immediately leads to the identity

$$\{X_{k:n} \leq x\} \quad = \quad \{I_n(x) < k\}\,.$$

In words: the k largest observation among $X_1, \ldots, X_n$ is $\leq x$ if and only if the number of observations that exceed x in the same sample is less than k.

Now suppose $X_1, \ldots, X_n$ all have the same distribution function F. We deduce

$$F_{kn}(x) \ = \ P\{X_{k:n} \leq x\} \quad = \quad \sum_{j=0}^{k-1} \binom{n}{j} \bar{F}^j(x) F^{n-j}(x)$$

where $\bar{F} = 1 - F$.

Now suppose we are in the domain of attraction of an extreme value distribution, so $\lim_{n\to\infty} F^n(a_n x + b_n) = G(x)$ for suitable $a_n,\ b_n,\ G$. This is equivalent to

$$n\bar{F}(a_n x + b_n) \quad \to \quad -\log G(x) \ = \ \Lambda(x) \text{ say.}$$

Note that we are (here and subsequently) using Λ as a general symbol for a Poisson generating measure; there is no connection with the Gumbel distribution (2.5). Let $a_n x + b_n = u_n,\ \Lambda(x) = \tau$, so

$$\lim_{n\to\infty} n\bar{F}(u_n) \ = \ \tau \text{ if and only if } \lim_{n\to\infty} F^n(u_n) \ = \ e^{-\tau}.$$

So

$$\lim_{n\to\infty} F_{kn}(u_n) \quad = \quad e^{-\tau} \sum_{i=0}^{k-1} \frac{\tau^i}{i!}$$

or alternatively

$$\lim_{n\to\infty} F_{kn}(a_n x + b_n) \quad = \quad G(x) \sum_{i=0}^{k-1} \frac{\Lambda(x)^i}{i!} \ = \ \psi_k(G(x))$$

where $\psi_k(x) = x\sum_{i=0}^{k-1}(-\log x)^i/i!$. This result was originally given by Smirnov [25].

Equivalently, one can write the result as follows:

$$I_n(a_n x + b_n) \xrightarrow{d} I(x) \tag{3.1}$$

where $I(x)$ is a Poisson random variable with mean $\Lambda(x)$. Equation (3.1) is just a restatement of the well-known convergence of a binomial distribution to a Poisson limit.

3.2 Point process viewpoint

So far, this is all for a single x (or equivalently, a single sequence $\{u_n\}$). A statement of the form (3.1) is much more powerful if interpreted in a stochastic process sense, treating x as a process parameter. Under suitable conditions, then, (3.1) may be reinterpreted as convergence to a *Poisson process* $I(x)$ with mean $\Lambda(x)$, for any $x \in \mathbb{R}$ for which $G(x) > 0$.

Let us make some definitions. Although we are concerned here with the case of a one-dimensional process, in general a Poisson process may be defined on any standard measurable space such as $\mathbb{R}^d$ for integer d, so we write $I(A)$ to denote the number of points in A where A is a Borel set in some general measure space $S \subseteq \mathbb{R}^d$. For the specific case when $S = \mathbb{R}$ and $A = (x, \infty)$ for some $x \in \mathbb{R}$, we may also write $I(x)$ in place of $I(A)$, consistent with (3.1). There are numerous definitions of a Poisson process: following the Appendix of [18], we may define $I(A),\ A \subseteq S$ to be a Poisson process with *intensity measure* Λ, if it satisfies the following conditions:

(a) For any set A with $\Lambda(A) < \infty$, $I(A)$ has a Poisson distribution with mean $\Lambda(A)$;

(b) $I(A_1)$, $I(A_2)$, ... $I(A_q)$ are independent random variables for any mutually disjoint sequence of sets $A_1, A_2, ..., A_q$.

If the measure Λ is absolutely continuous, the Poisson process also has the property of *no multiple points*, e.g. if $I(A) = q > 1$ for some A then it is possible to write $A = A_1 \cup A_2 \cup ... \cup A_q$ where $A_1, ..., A_q$ are disjoint and $I(A_j) = 1$ for each $j = 1, ..., q$. A point process with no multiple points is also called *simple*.

In order to establish limit statements such as (3.1) in a point process context, we need a notion of *convergence of point processes*. There are two standard definitions: *convergence in distribution* and *vague convergence*.

Definition 3.1 (convergence in distribution, following the Appendix of [18]). In I_n is a sequence of point processes and I is a limiting point process, each defined on a space S with Borel sets $\mathscr{S}$, then we say

$$I_n \xrightarrow{d} I$$

(I_n converges in distribution to I) if and only if

$$(I_n(A_1), I_n(A_2), \ldots, I_n(A_q)) \xrightarrow{d} (I(A_1), I(A_2), \ldots, I(A_q))$$

for each choice of q and bounded Borel sets $A_i \in \mathscr{S}$ such that $I(\partial A_i) = 0$ almost surely for each $i = 1, 2, \ldots, q$ (∂A_i denotes the boundary of the set A_i).

Definition 3.2 (vague convergence, following [24], pp. 48–49). Suppose S is a "nice space" (a locally compact topological space with a countable base, such as $\mathbb{R}^d$ for any d) and $\mathscr{S}$ is a sigma-field on S. A measure $\mu : \mathscr{S} \to [0, \infty]$ is an assignment of positive numbers to sets in $\mathscr{S}$ such that

1. $\mu(\phi) = 0$ and $\mu(A) \geq 0$ for all $A \in \mathscr{S}$;
2. σ-additivity: if $\{A_n,\ n \geq 1\}$ are mutually disjoint sets in $\mathscr{S}$ then $\mu\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i)$.

A measure μ is called *Radon* if $\mu(K) < \infty$ for any compact set K.
Next, let

$$M_+(S) \quad = \quad \{\mu :\ \mu \text{ is a nonnegative measure on } \mathscr{S} \text{ and } \mu \text{ is Radon}\}.$$

Define

$$C_K^+(S) \quad = \quad \{f :\ S \to \mathbb{R} :\ f \text{ is continuous with compact support}\}.$$

(Compact support means that $f = 0$ outside some compact set.)

If $\mu_n \in M_+(S)$ for $n \geq 0$, then μ_n converges vaguely to μ_0, written $\mu_n \xrightarrow{v} \mu_0$, if for all $f \in C_K^+(S)$,

$$\mu_n(f) \ = \ \int_S f(x)\mu_n(dx) \quad \to \quad \mu_0(f) \ = \ \int_S f(x)\mu_0(dx).$$

Vague convergence implies convergence in distribution, but the converse implication does not hold in general.

For the present chapter, convergence in distribution will suffice, but in later chapters involving multivariate and spatial extreme value theory, we may need the concept of vague convergence as well.

3.3 Poisson Convergence for Extreme Order Statistics from an Independent Sequence

This essentially follows Weissman [27].

We consider an *extremal process* generated by a sequence of independent random variables $\{X_i,\ i = 1,2,\ldots\}$, assumed to be defined on some common probability space $(\Omega,\mathscr{F},P)$. The case where the X_is are not only independent but also identically distributed (IID) is of particular interest, but the theory applies in this more general setting so we shall follow that for the genral presentation.

Assume, as in classical extreme value theory, that there are normalizing constants $a_n > 0$ and $b_n \in \mathbb{R}$ such that $\{\max(X_1,\ldots,X_n) - b_n\}/a_n$ converges in distribution, and define

$$X_{ni} \quad = \quad \frac{X_i - b_n}{a_n}.$$

For each n and k, define $m_n^k(t)$ to be the kth largest among $X_{n1},\ldots,X_{n[nt]}$. Here $[nt]$ is the integer part of nt and we are explicitly thinking of this as a function of the continuous time variable t (we define $m_n^k(t) = -\infty$ if $k > [nt]$).

Suppose there exists a process $m^k = \left\{m^k(t) :\ t > 0\right\}$ such that

$$m_n^k \quad \to \quad m^k$$

in the sense of convergence of finite-dimensional distributions. Let us define

$$I_n(t;x) \quad = \quad \sum_{i=1}^{[nt]} I(X_{ni>x}).$$

Note that the definition of $I_n(t;x)$ is different from the corresponding definition of $I_n(x)$ in Section 3.1 because we have renormalized the X_is before defining the process rather than after. This turns out to be a more convenient formulation for point process convergence.

Our objective, then, is to establish conditions sufficient to guarantee that $I_n(t;x)$ converges in distribution to a limiting point process $I(t;x)$ and, in such cases, to derive the structure of $I(t;x)$.

The limiting distribution function (when it exists) of $m_n^1(t)$ will be written G_t. In the IID case when $m_n^1(1) \xrightarrow{d} m^1(1) \sim G$, we will have $G_t = G^t$, but in general we do not require that structure.

More explicitly, for the IID case where $G = G_1$ exists, it may be written in the form of the GEV distribution and therefore

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

for some μ, σ, ξ, and although this is only a special case of the general theory, it will be useful to fix ideas.

In the general case, we assume that G_t is non-degenerate for all $t > 0$ and for $0 \le s < t$ define

$$G_{st}(x) \;=\; \frac{G_t(x)}{G_s(x)},\; G_s(x,y) \;=\; \frac{G_s(y)}{G_s(x)}\;(y \le x),$$

both quantities defined to be 0 when $G_s(x) = 0$. The support for G_t is assumed to be $({}_*x_t,\; x_{*t}\,)$ where $-\infty \le {}_*x_t < x_{*t} \le \infty$. We also write $Y \sim \mathscr{P}(\lambda)$ as a shorthand to denote that the random variable Y has a Poisson distribution with mean λ.

Define $T = \{(t,x) :\; t > 0, x >_* x_t\}$ and let $I = \{I(t;x) :\; (t,x) \in T\}$ be a point process on T, where $I(t;x)$ is the number of points in $(0,t] \times (x,\infty)$. We define $I(0,x) \equiv 0$.

Changing notation slightly, for a rectangle $E = (s,t] \times (x,y]$ with $0 \le s < t$ and $x < y \le \infty$, we define the increment of I around E to be

$$\begin{aligned} I(E) &= I(t;x) - I(s;x) - I(t;y) + I(s;y) \\ &= I(s,t;x,y) \text{ say.} \end{aligned}$$

If we also let $\Lambda(t;x) = -\log G_t(x)$, we can regard Λ as a measure on $[0,\infty) \times [{}_*x_t, \infty]$ where

$$\begin{aligned} \Lambda(E) &= \Lambda(t;x) - \Lambda(s;x) - \Lambda(t;y) + \Lambda(s;y) \\ &= \Lambda(s,t;x,y). \end{aligned}$$

This defines a measure on all Borel sets of T; in particular, again slightly abusing notation,

$$\Lambda((0,t] \times (x,\infty]) \;=\; \Lambda(t;x).$$

In the special case of the GEV,

$$\Lambda(t;x) \;=\; t\left(1 + \xi\frac{x-\mu}{\sigma}\right)^{-1/\xi}$$

defined on $t > 0$ and $x \in ({}_*x, x_*)$ where

$${}_*x \;=\; \begin{cases} \mu - \frac{\sigma}{\xi}, & \xi > 0, \\ -\infty, & \xi \le 0 \end{cases}$$

Chapter 4

Statistics Based on the Extreme Value Distributions

At this point in the book, we have seen four basic models used for extremes:

1. The extreme value distributions, derived mathematically through the Three Types Theorem, but most frequently represented in practice by the Generalized Extreme Value (GEV) distribution, $G(x;\mu,\sigma,\xi) = \exp\left[-\{1+\xi(x-\mu)/\sigma\}_+^{-1/\xi}\right]$;
2. The Generalized Pareto distribution (GPD), $G(x;\sigma,\xi) = 1-(1+\xi x/\sigma)_+^{-1/xi}$, $x>$ 0 which is used as a limiting distribution for exceedances over thresholds;
3. The limiting joint distribution for the k largest maxima from a sample, which we saw in Chapter 3 and is directly derived from the GEV;
4. The point process viewpoint, also derived in Chapter 3, which represents the two-dimensional process of exceedances over a threshold, together with the associated times of exceedances, as a point process that can be approximated as a Poisson process; in the homogeneous case, the intensity measure of the point process is also derived from the GEV distribution.

In practice, each of these models may be extended by bringing in covariates: for example, in the GEV case, instead of assuming the same GEV parameters for every observation, we may write $Y_i \sim GEV(\mu_i,\sigma_i,\xi_i)$ where each of μ_i,σ_i,ξ_i may depend on covariates; in the simplest case, $\mu_i = \sum_j \beta_j x_{ij}$ (with corresponding expressions, if desired, for either σ_i or for $\log\sigma_i$, and for ξ_i). As we proceed to develop practical examples, the reader will see many examples where relevant covariates are introduced into the analysis. More complicated questions — such as what to do if there is temporal dependence in the observations — will be deferred for the time being, but we return to questions like these later.

All of the models described so far are parametric models — even with covariates, we assume explicit parametric representations for the covariates — so they may in principle be estimated by well-established methods for parametric statistical models, of which the best known are maximum likelihood and Bayesian inference. We begin this chapter with a brief review of those methods, before turning to more specific issues associated with extreme value distributions.

4.1 Maximum Likelihood and Bayesian Statistics

To present the general theory, we assume $Y_1, ..., Y_n$ are independent random variables where Y_i has density $f_i(y_i;\theta)$ where f_i is a known density depending on unknown parameters $\theta \in \Theta$ for some parameter space $\Theta \subseteq \mathbb{R}^p$ for some p. For example, in the simplest case where Y_i has a GEV distribution with parameters (μ,σ,ξ) (the same for all i), the estimated parameter vector is $\theta = (\mu,\sigma,\xi)$ and the parameter space Θ may be represented as $\mathbb{R}\times\mathbb{R}_+\times\mathbb{R}$ though, as we shall see, there may be some advantage to restricting the range of one of more of these parameters (especially ξ).

The likelihood function for parameters θ based on a specific set of observations, $\mathbf{Y} = \begin{pmatrix} Y_1 & Y_2 & \dots & Y_n \end{pmatrix}$ may be written

$$L(\theta\ ;\ \mathbf{Y}) \quad = \quad \prod_{i=1}^{n} f_i(Y_i\ ;\ \theta). \tag{4.1}$$

We also write $\ell(\theta\ ;\ \mathbf{Y}) = -\log\mathbf{L}(\theta\ ;\ \mathbf{Y})$ to denote the negative log likelihood function. We can now define:

Definition 4.1: The *maximum likelihood estimator* of θ based on $\mathbf{Y}$ is the value $\hat{\theta}$ that maximizes $L(\theta;\mathbf{Y})$ or, equivalently, minimizes $\ell(\theta;\mathbf{Y})$.

Already, this definition runs into a small complication. The most common methods of finding a maximum likelihood estimator use calculus (analytically in simple cases, numerically in more complicated models such as the GEV) which means, in effect, they are trying to find estimates that solve the *likelihood equations*,

$$\frac{\partial\ell(\theta\ ;\ \mathbf{Y})}{\partial\theta_{\mathbf{j}}} \quad = \quad 0,\ j = 1,2,\dots,p. \tag{4.2}$$

However, in general there is no guarantee that a θ satisfying (4.2) will achieve the global maximum of L. In fact, in the GEV case, we can show that it does not (unless we restrict the range of ξ, or which more momentarily). Nevertheless, in practice, we usually define the maximum likelihood estimator (MLE) as a vector $\hat{\theta}$ that satisfies (4.2)

As an explicit example, for the GEV distribution without covariates we have, on the range where the distribution function G and its density g are defined,

$$G(y\ ;\ \mu,\sigma,\xi) \quad = \quad \exp\left\{-\left(1+\xi\frac{y-\mu}{\sigma}\right)^{-1/\xi}\right\}, \tag{4.3}$$

$$\begin{aligned} g(y\ ;\ \mu,\sigma,\xi) \quad &= \quad \frac{\partial G(y\ ;\ \mu,\sigma,\xi)}{\partial y} \\ &= \quad \frac{1}{\sigma}\left(1+\xi\frac{y-\mu}{\sigma}\right)^{-1/\xi-1}\exp\left\{-\left(1+\xi\frac{y-\mu}{\sigma}\right)^{-1/\xi}\right\}, \end{aligned} \tag{4.4}$$

and hence

$$\ell(\mu,\sigma,\xi\ ;\ Y_1,...,Y_n) \quad = \quad \sum_{i=1}^{n}\left[\log\sigma+\left(\frac{1}{\xi}+1\right)\log\left(1+\xi\frac{Y_i-\mu}{\sigma}\right)+\left(1+\xi\frac{Y_i-\mu}{\sigma}\right)^{-1/\xi}\right]. \tag{4.5}$$

The MLE chooses μ,σ,ξ to minimize ℓ, which in practice is found by optimization routines designed to solve the likelihood equations.

In the case of covariates, we replace μ,σ,ξ by μ_i,σ_i,ξ_i as appropriate; for example, if we represent time dependence by the function $\mu_i=\beta_0+\beta_1 t_i$ (t_i being the time of observation i) then the three-parameter $\theta=\begin{pmatrix}\mu & \sigma & \xi\end{pmatrix}$ is expanded into a four-parameter $\theta=\begin{pmatrix}\beta_0 & \beta_1 & \sigma & \xi\end{pmatrix}$ and we compute corresponding MLEs $\hat{\theta}=\begin{pmatrix}\hat{\beta}_0 & \hat{\beta}_1 & \hat{\sigma} & \hat{\xi}\end{pmatrix}$.

As already noted, it is assumed that ℓ is differentiable and that the MLE satisfies (4.2). In nearly all cases, we also evaluate (exactly or approximately) second-order derivatives. If

$$h_{jk} \quad = \quad \left.\frac{\partial^2\ell}{\partial\theta_j\partial\theta_k}\right|_{\theta=\hat{\theta}},\ j,k=1,\ldots,p,$$

and if H is the $p\times p$ matrix of $\{h_{jk}\}$, then H is referred to as the *Hessian matrix* or, specifically in the context of MLE, the *observed information matrix*. Alternatively, if $I(\theta)$ is the matrix of i_{jk}, defined as the expected value of $\frac{\partial^2\ell}{\partial\theta_j\partial\theta_k}$ when the true parameter vector is θ, then $I(\theta)$ is the *expected* or *Fisher information* matrix. In practice, of course, θ is unknown so we calculate $I(\hat{\theta})$ in place of $I(\theta)$. Regardless of whether $H(\hat{\theta})$ or $I(\hat{\theta})$ is used, for large samples and assuming certain regularity conditions, the inverse or $H^{-1}(\hat{\theta})$ or $I^{-1}(\hat{\theta})$ is an asymptotic approximation to the variance-covariance matrix of θ. A famous paper of Efron and Hinkley [5] argued that the observed information matrix is the better approximation in practice, and regardless of which one is theoretically superior, the observed information matrix is usually easier to calculate. Therefore, for most practical applications of maximum likelihood, the observed information is preferred.

The main alternative to the method of maximum likelihood is a Bayesian analysis. As will be seen, an argument can be made that Bayesian methods do a better job of representing the uncertainty of extreme value estimation, especially where related to prediction. The fundamental formula of Bayesian statistics is

$$\pi(\theta\mid\mathbf{Y}) \quad = \quad \frac{\prod_{i=1}^{n}f_i(Y_i\ ;\ \theta)\pi(\theta)}{\int_\Theta\prod_{i=1}^{n}f_i(Y_i\ ;\ \theta)\pi(\theta)d\theta}, \tag{4.6}$$

read as "the posterior density of θ given $\mathbf{Y}$", where we again assume that $\mathbf{Y}=\begin{pmatrix}Y_1 & Y_2 & \ldots & Y_n\end{pmatrix}$ where the Y_i are mutually independent and have densities $f_i(\cdot\ ;\ \theta)$ as in (4.1). The formula $\pi(\theta)$ is known as "the prior density of θ" and is usually chosen to be flat and diffuse (but typically proper) over the parameter space Θ. In practice the integral in the formula (4.3) cannot be evaluated analytically

and we resort to Monte Carlo simulation, for which the most popular algorithms are variants on the Gibbs sample and the Hastings-Metropolis algorithm; see [4] for a recent survey. For very large datasets, newer algorithms are becoming popular, such as variational Bayes [26].

In practice, the interest in fitting these models often lies not in the parameter estimates themselves, but as an intermediate step towards calculating other quantities that are of interest in connecting with predicting or characterizing extreme events. One particular measure of interest is the *N-year return value* (where in applications one may want to take N to be a relatively modest value such as 50 or 100, or something much larger, such as 500 or 1,000 or even 10,000, though the practical reality of estimating very extreme return values is questionable). This is often defined by defining y_N to be the level that is exceeded with probability $1/N$ in any given year. If the annual maximum distribution is given by (4.3), then we derive the formula

$$y_N \quad = \quad \mu + \sigma \frac{\left\{-\log\left(1-\frac{1}{N}\right)\right\}^{-\xi} - 1}{\xi}. \tag{4.7}$$

Typically, y_N is estimated by simply substituting the maximum likelihood estimators $\hat{\mu},\ \hat{\sigma},\ \hat{\xi}$, in place of $\mu,\ \sigma,\ \xi$, in 4.8). This is sometimes called the "plug-in approach". From a Bayesian point of view, a more natural alternative is the posterior mean

$$\hat{y}_N \quad = \quad \int y_N(\mu,\sigma,\xi)\pi(\mu,\sigma,\xi \mid \mathbf{Y}), \tag{4.8}$$

where the dependence of y_N on the parameters μ,σ,ξ is represented explicitly in the equation.

Returning to the general case of arbitrary densities $f_i(y_i\ ;\ \theta)$ for some parameter vector θ, we may be interested in some scalar quantity $h(\theta)$; (4.7) is the special case where h is the N-year return value from the GEV distribution. Based on the maximum likelihood estimators, this suggests using $\hat{h} = h(\hat{\theta})$ as an estimator, where $\hat{\theta}$ is the MLE. There are the three ways of computing a confidence interval for h:

1. The normal or *delta method* approach, where we approximate the standard error of $\hat{h}$ by $\sqrt{(\nabla h)^T V(\nabla h)}$, where ∇h represents the gradient of h with respect to the parameters θ (evaluated at $\hat{\theta}$) and $V(\theta)$ is an estimate of the variance-covariance matrix of θ, usually approximated with the inverse of either the observed or the Fisher information matrix. The endpoints of the confidence interval are then calculated using a normal approximation.
2. The *profile likelihood* approach is an approach where, for each candidate value h_0 say, we maximize the likelihood function (or minimize ℓ) with respect to θ under the constraint $h(\theta) = h_0$. The resulting minimized value of $-\log\ell$ is written $\ell^*(h_0)$. By definition, $\ell^*(h_0)$ is minimized when $h_0 = h(\hat{\theta})$, the overall MLE. An approximate $100(1-\alpha)\%$ confidence interval for h may be defined as the set of h_0 for which

$$2\left\{\ell^*(h_0) - \ell^*(h(\hat{\theta}))\right\} \quad \leq \quad \chi^2_{1-\alpha,1}$$

where $\chi^2_{1-\alpha,1}$ is defined as the $1-\alpha$ point of the χ^2_1 distribution. This method relies on the asymptotic χ^2_1 distribution fo the likelihood ratio statistic in this case. Even though this is also an approximation, it is generally considered a more accurate approach than the delta method; in particular, profile likelihood intervals are often highly asymmetric whereas the delta method necessarily leads to symmetric confidence intervals.

3. The third method is a *bootstrap*, which in turn may be either *parametric bootstrap* or *nonparametric bootstrap*. The parametric bootstrap assumes that the probability model is correct; for each $b = 1,\ldots,B$ (where B is the numer of bootstrap samples, and b is an index) generate a sample of size n from the assumed distribution, recalculate the MLE $\hat{\theta}_b$ say, and calculate an estimate $\hat{h}_b = h(\hat{\theta}_b)$. The $\alpha/2$ and $1-\alpha/2$ probability points of the sample $\left\{\hat{h}_b,\ b = 1,...,B\right\}$ then form a natural $100(1-\alpha)\%$ confidence interval for $h(\theta)$. The alternative noparametric bootstrap works the same way, except that the bootstrap samples are generated by resampling from the data rather than generating simulated samples from the underlying parametric distribution.

As an alternative to any of these methods, A Bayesian approach uses the posterior density of h,

$$h(\theta \mid \mathbf{Y}) \quad = \quad \frac{h(\theta)\prod_{i=1}^{n}\{f_i(Y_i\ ;\ \theta)\}\,\pi(\theta)}{\int_{\Theta} h(\theta)\prod_{i=1}^{n}\{f_i(Y_i\ ;\ \theta)\}\,\pi(\theta)d\theta}. \tag{4.9}$$

Prediction intervals or credible intervals for $h(\theta)$ may be calculated from (4.9), typically using the same Monte Carlo sample as was used to calculate the posterior density of θ.

4.2 Issues specific to the extreme value families

Let us look at the middle term in (4.5). Suppose $\xi < -1$. The condition $1+\xi(Y_i-\mu)/\sigma > 0$ must apply to every term, so with $\xi < 0$, this means $\mu > Y_{max}+\sigma/\xi$ where Y_{max} is the maximum among $Y_1,\ldots,Y_n$. As $\mu \downarrow Y_{max}+\sigma/\xi$, the log term tends to $-\infty$. Because $\frac{1}{\xi}+1 > 0$ in this instance, that means the whole expression (4.5) tends to $-\infty$. That creates an obvious problem with the definition of the MLE. Essentially the same issue arises with the other models used in extreme value theory, including the Generalized Pareto model and the models based on joint distributions of $k > 1$ order statistics.

In practice, this problem is solved in one of two ways. The most clean-cut solution is simply to define $\xi \in (-1,1)$ and so avoid the problem. The upper end of this interval is also a natural choice because of $\xi \geq 1$, the variable Y has infinite mean, which is also considered unrealistic for most (though not all) applications.

The alternative approach is just to ignore it. Numerical optimization techniques hardly ever fail because of the singularity and generally find solutions which exactly or approximately solve (4.2). However, in complicated models with many covariates, it is possible that the optimization will fail because of the singularity, so this needs to be borne in mind.

The second issue is *regularity of the maximum likelihood estimates*. Theoretical accounts of maximum likelihood estimation [insert references] emphasize the need for certain conditions, for example, the second-order derivatives of the log likelihood must exist and be continuous, and the elements defining the Fisher information matrix must be finite. Theoretical calculations of the Fisher information matrix have been given by [Jenkinson, Presscott and Walden] for the GEV distribution, [Smith] for the GPD, [Smith] for the k largest order statistics model based on the Gumbel family, and [Tawn] for the k largest order statistics model based on the GEV family. All of the models fail when $\xi \leq -\frac{1}{2}$, in the sense that at least one entry of the Fisher information matrix becomes infinite, and the asymptotic results of maximum likelihood fail in this case. Theoretical solutions for the case $\xi \leq -\frac{1}{2}$ were proposed by [Smith] but have never been used in practice asthis case very rarely arises. More practically, the results of [Smith, Bucher and Segers] confirm that for the cases where $\xi > -\frac{1}{2}$, the MLEs defined by (4.2) exist and satisfy the standard asymptotic properties of maximum likelihood estimation, in particular consistency and asymptotic normality, with asymptotic variance-covariance matrix given by the inverse of the Fisher information matrix.

Thus, in practice, we may apply the method of maximum likelihood (possibly with some restrictions on the range of ξ) and the estimators have the same asymptotic properties as in regular parametric problems. However, given some of the specific issues that arise in extreme value theory (especially, estimating the N-year return value Y_N for large N), there are still questions related to bias of the estimators or poor performance of the asymptotic approximations. Sone alternatives include *Generalized Maximum Likelihood Estimation* (GMLE) and the *L-moments method*.

Bibliography

[1] N. Bingham, C. Goldie, and J. Teugels. *Regular Variation*. Cambridge University Press, Cambridge, U.K., 1987.

[2] J.P. Cohen. Convergence rates for the ultimate and penultimate approximations in extreme value theory. *Advances in Applied Probability*, 14:833–854, 1982.

[3] J.P. Cohen. The penultimate form of appriximation to normal extremes. *Advances in Applied Probability*, 14:324–339, 1982.

[4] D.B. Dunson and J.E. Johndrow. The hastings algorithm at fifty. *Biometrika*, 107:1–23, 2020.

[5] B. Efron and D.V. Hinkley. Assessing the accuracy of the maximum likelihood estimator: Observed versus expected fisher information. *Biometrika*, 65:457–483, 1977.

[6] W. Feller. *An Introduction to Probability Theory and Its Applications, Volume 1 (Third edition)*. John Wiley & Sons, New York, 1968.

[7] R.A. Fisher and L.H.C. Tippett. Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proc. Cambridge Phil. Soc.*, 24:180–190, 1928.

[8] M Fréchet. Sur la loi de probabilité de l'écart maximum. *Ann. Soc. Math. Polon.*, 6:93–116, 1927.

[9] B. Gnedenko. Sur la distribution limite du terme maximum d'une série aléatoire. *Ann. Math.*, 44:423–453, 1943.

[10] M.I. Gomes. Penultimate limiting forms in extreme value theory. *Annals of the Institute of Statistical Mathematics*, 36 (1):71–85, 1984.

[11] L. de Haan. *On regular variation and its application to the weak convergence of sample extremes*. Mathematisch Centrum, Amsterdam, 1970.

[12] L. de Haan. Sample extremes: an elementary introduction. *Statistica Neerlandica*, 30:161–172, 1976.

[13] L. de Haan and A. Ferreira. *Extreme Value Theory: An Introduction*. Springer, New York, 2006.

[14] P. Hall. On the rate of convergence of normal extremes. *Journal of Applied Probability*, 16:433–439, 1979.

[15] N.L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous univariate distributions (Volume 1, 2nd edition)*. John Wiley & Sons, New York, 1994.

[16] M.R. Leadbetter. On extreme values in stationary sequences. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 28:289–303, 1974.

[17] M.R. Leadbetter. Weak convergence of high level exceedances by a staionary sequence. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 34:11–15, 1974.

[18] M.R. Leadbetter, G. Lindgren, and H. Rootzén. *Extremes and Related Properties of Random Sequences and Processes*. Springer-Verlag, New York, 1983.

[19] R. von Mises. La distribution de la plus grande de *n* valeurs. *Selected Papers II (Am. Math. Soc.)*, pages 271–294, 1936.

[20] J. Pickands III. The two-dimensional poisson process and extremal process. *Journal of Applied Probability*, 8:745–756, 1971.

[21] J. Pickands III. Statistical inference using extreme order statistics. *Annals of Statistics*, 3:119–131, 1975.

[22] S. Resnick. Tail equivalence and its applications. *Journal of Applied Probability*, 8:136–156, 1971.

[23] S.I. Resnick. *Extreme Values, Regular Variation abd Point Processes*. Springer-Verlag, New York, 1987.

[24] S.I. Resnick. *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer, New York, 2007.

[25] N.V. Smirnov. Limit distributions for the sums of a variational series. *American Mathematical Society Translations*, 67:1–67, 1952.

[26] Yixin Wang and David M. Blei. Frequentist consistency of variational bayes. *Journal of the American Statistical Association*, 114 (527):1147–1161, 2019.

[27] I. Weissman. Multivariate extremal processes genrated by independent non-identically distributed random variables. *Journal of Applied Probability*, 12:477–487, 1975.

[28] I. Weissman. Estimation of parameters and large quantiles based on the k largest observations. *Journal of the American Statistical Association*, 73 (364):812–815, 1978.

[29] V.M. Zolotarev. Lévy metric. *Encyclopedia of Mathematics*, https://www.encyclopediaofmath.org/index.php/L%C3%A9vy_metric, 2012.