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

1 Extreme Value Theory In Action **1**
1.1 Introduction 1
1.1.1 Climate and Weather Extremes 1
1.1.2 Insurance Risk of a Large Company 2
1.1.3 Women's Track Times 4
1.2 Overview of Univariate Extremes 6
1.2.1 The Three Types Theorem and the Generalized Extreme Value Distribution 6
1.2.2 Exceedances Over Thresholds 8
1.2.3 The *r*-largest Order Statistics Approach 10
1.2.4 Point Process Approach 11
1.2.4.1 GEV model 14
1.2.4.2 GPD model 14
1.2.4.3 *r*-largest order statistics model 15
1.3 Estimation 15
1.3.1 GEV Model 15
1.3.1.1 The delta method 19
1.3.1.2 Bayesian methods 20
1.3.1.3 Example: Jenkinson's data from Hartford, Connecticut 22
1.3.2 GPD and Poisson-GPD Models 25
1.3.3 The *r*-largest order statistics Model 26
1.3.3.1 Single Sample 26
1.3.3.2 Multiple Samples 28
1.3.4 Point Process Approach 31
1.4 Analysis of Data in Kelowna 31
1.5 Analysis of Insurance Data 36
1.6 Analysis of Women's Track Data 36
1.7 Software: the `extRemes` Package 36
1.8 Summary of Chapter 36
1.9 Exercises 36

2 Domains of Attraction, Rates of Convergence and Optimal Statistical Estimation **37**
2.1 The Theory of Gnedenko and de Haan 37

2.1.1 Convergence of threshold exceedances to the Generalized Pareto Distribution 40
2.2 Examples 40
2.2.1 The t distribution and extensions 40
2.2.2 The beta distribution and extensions 42
2.2.3 Normal distribution 43
2.2.4 Lognormal distribution 47
2.2.5 An example of a distribution with finite ω_F in the Gumbel domain of attraction 48
2.2.6 An example of a continuous distribution not in any domain of attraction 49
2.2.7 Discrete distributions 49
2.3 Reformulation in Terms of Inverse Functions 49
2.4 Second-order Approximations 51
2.4.1 Examples 52
2.5 Estimation theory based on second-order asymptotics 53
2.5.1 Side Section 1: A heuristic on biased estimation 55
2.5.2 Side section 2: Asymptotics of the Hill-Weissman Estimator 57
2.5.2.1 Extension to the GPD 60
2.5.2.2 Comparisons with the Hill-Weissman Estimator 62
2.5.2.3 Background References 62
2.5.3 Outline Derivation of Dombry-Ferreira result 62
2.6 Other topics to be added 64
2.6.1 Method of probability weighted moments 64
2.6.2 Corresponding results for threshold estimators 64
2.6.3 Estimating probabilities of extreme events 65
2.6.4 Adaptive choice of block size or threshold 65
2.6.5 Practical examples 65

3 Extremes in Dependent Sequences 67
3.1 Extremes in stationary sequences 67
3.2 The extremal index 69
3.3 Infinitely divisible random measures 74
3.4 Exceedances of a single level 76
3.5 The two-dimensional exceedance process 80
3.6 Markov chains 84
3.7 Computational Methods for the Extremal Index in Markov Chains and Extensions 92
3.7.1 Markov chains derived from bivariate extreme value distributions 92
3.7.2 Extension to k'th-order Markov chains 95
3.8 Models for Financial Time Series 101
3.9 Statistical Aspects 102
3.9.1 Parametric methods 102
3.9.2 Nonparametric methods 102

3.10 The Multivariate Extremal Index 102

4 Multivariate Extremes 103
4.1 Introduction to Multivariate Extreme Value Theory 103
4.2 The Pickands Representation 104
4.3 Nonparametric Estimation of Bivariate and Multivariate Extreme Value Distributions 108
4.4 Parametric Estimation of Bivariate and Multivariate Extreme Value Distributions 113
4.4.1 Asymptotic results for maximum likelihood estimation and testing 116
4.5 Threshold Methods for Multivariate Extremes 119
4.5.1 Estimation in the Coles-Tawn model 121
4.5.2 Alternative censored data approach 123
4.6 Asymptotic Dependence and Asymptotic Independence 126
4.6.1 Introduction: The coefficient of tail dependence 126
4.6.2 Extension to the full joint tail 128
4.6.3 New models based on hidden regular variation 129
4.6.4 Extensions to the case $p > 2$ 131
4.6.5 An application: Dependence among extreme weather events 131
4.7 Other Approaches to Multivariate Extremes 135
4.7.1 The conditional approach of Heffernan and Tawn 135
4.7.2 Combining AD and AI models: The approach of Wadsworth et al. 135
4.7.3 De Haan and de Ronde 136
4.7.4 General max-stable approach 136
4.7.5 Multivariate generalized Pareto distributions 137
4.7.6 High-dimensional multivariate extremes 137

5 Spatial Extremes 139
5.1 The Latent Process Approach 140
5.1.1 Background on Spatial Statistics 141
5.1.1.1 Intrinsic Stationarity and the Semivariogram 143
5.1.1.2 Lattice Models 144
5.1.1.3 Estimation of Gaussian Spatial Processes 144
5.1.1.4 Spatial Models with Measurement Error 145
5.1.2 Application to Precipitation Extremes Example 146
5.1.3 Results 149
5.1.4 Literature Review 151
5.1.5 Summary 159
5.2 Max-Stable Processes 159
5.2.1 Background on Poisson processes 159
5.2.2 Constructing a max-stable process 162
5.3 Probability Calculations for Max-Stable Processes 164
5.3.1 Brown-Resnick Process 164

5.3.2 Extremal t Process 167
5.3.3 Smith Process 167
5.3.4 Schlather Process 167
5.3.5 The Reich-Shaby Model 167
5.4 Inference for Max-Stable Processes 169
5.4.1 Method of composite likelihood 170
5.4.2 Progress towards exact maximum likelihood 171
5.5 Other Approaches to Spatial Extremes 171

Bibliography **173**

Chapter 1

Extreme Value Theory In Action

1.1 Introduction

Extreme Value Theory refers to the class of probabilistic and statistical techniques used to extremes in random sequences and processes. It has many applications including environmental and climate extremes, finance and insurance, reliability and strength of materials, and even the study of sports records. In this introductory chapter, we shall give some examples motivated by real data where there are natural questions concerning the probability of an extreme event that has occurred or might occur in the future, or how large an extreme event might be expected over some period of time. We shall introduce some of the main statistical methods that are used to study extremes, and show how they can be used to answer such questions. In later chapters, we shall develop the mathematical basis of this theory in much more detail, will study extensions such as extremes in dependent process, multivariate extremes and spatial extremes, and go into some of the applications in much greater depth.

1.1.1 Climate and Weather Extremes

In June and July, 2021, an extreme heatwave descended on western North America, particularly affecting the US states of Washington and Oregon, and the Canadian province of British Columbia. It resulted in, amongst others, a peak temperature of 49.6^{o}C (121.3^{o}F) in Lytton, British Columbia. Over 1,000 deaths were attributed to the heatwave and there were extensive consequences to wildfires, damage to the road and rail infrastructure, agriculture, and many other effects. Questions naturally arose to what extent climate change was responsible for these events, and a paper published online shortly afterwards [178] claimed that the event "was virtually impossible without human-caused climate change."

As an illustration of a "typical" city in this region, Fig. 1.1 shows annual maximum temperatures from 1984–2022 in the city of Kelowna, British Columbia. As is self-evident from the figure, the temperature for June 30, 2021, stands out very clearly, recording 44.6^{o}C, more than 5^{o}C higher than the second highest annual maximum in the series. Even without considering the impact of climate change, natural questions arising from this series include

1. What is the natural probability distribution of annual maximum temperatures in this or similar locations?

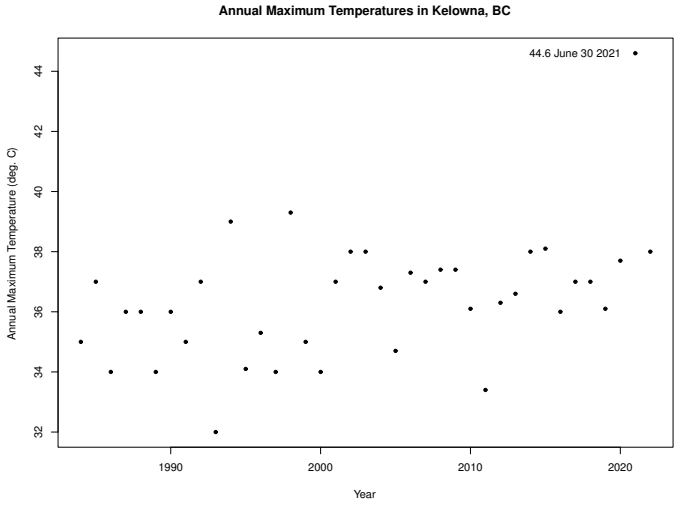

Figure 1.1 *Annual maximum temperatures in Kelowna, BC*

2. How extreme was the 2021 event? It is common to express extreme events in terms of *return values*, for example, the N-year return value is often defined colloquially at the event that occurs once in N years, though a more precise definition for climate calculations is that it is the event that occurs in an one year with probability $1/N$. Based on that, what would be the appropriate N to make 44.6°C the N-year return value?
3. Even before 2021, is there any evidence of an increasing trend in temperatures in Kelowna, and if so, are there large-scale climate indicators (such as global mean temperatures) that it can be related to?

In recent years, there have been many similar instances of extreme weather events; two more are shown in Fig. 1.2, depicting the heatwave that hit the southern half of the United Kingdom in July 2022, and the extreme rainfalls that followed Hurricane Harvey in the Houston area in August 2017.

1.1.2 Insurance Risk of a Large Company

This example is based on [230], and was also featured in [229].

The data are a 15-year record of inflation-adjusted insurance claims above a certain threshold by a multinational oil company. 425 claims were in the dataset, though some appeared to be multiple claims related to the same event and were grouped together for the analysis. The claims were grouped into seven "types", listed in Table 1.1.

In the units adopted for the originl publication, the total of all 393 claims was 2989.6, and the ten largest claims, in order, were 776.2, 268.0, 142.0, 131.0, 95.8,

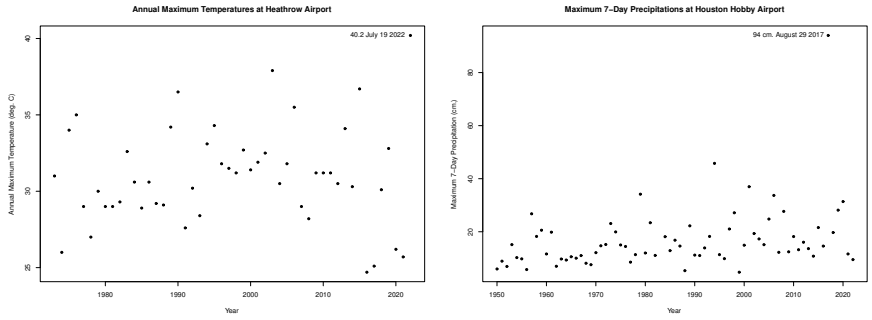

Figure 1.2 *Left: Annual maximum temperatures at Heathrow Airport, London, U.K. The peak of 40.2°C (104.4°F) occurred on July 19, 2022. Right: Annual maxima of 7-day rainfall totals during the hurricane season (July–November) at Houston Hobby airport, Texas, U.S.A. The largest total of 94 cm. (37 in.) was reached during Hurricane Harvey on August 29, 2017.*

Type	Description	Number	Mean
1	Fire	175	11.1
2	Liability	17	12.2
3	Offshore	40	9.4
4	Cargo	30	3.9
5	Hull	85	2.6
6	Onshore	44	2.7
7	Aviation	2	1.6

Table 1.1 *The seven types of insurance claim, with the total number of claims and the mean size of claim for each type*

56.8, 46.2, 45.2, 40.4, 30.7. Thus, the largest claim on its own accounted for 26% of the total, and the ten largest claims together for 55%. This is fairly typical of insurance data: a few of the very largest claims have by far the greatest impact, so any statistical modeling of such data must account for the most extreme values. However, this also raises the question of whether the most extreme claims should be treated as outliers and analyzed separately from the rest of the data.

Some plots of this dataset are shown in Fig. 1.3: plot (a) is in effect a scatterplot of the data (note the logarithmic scale on the vertical axis — the two largest claims are in fact substantially larger than the rest of the data); plots (b) and (c) are meant to illustrate the possibility of time trends (there is no visible evidence of any change in the overall claim rate in (b); the plot of claim amounts in (c) does show a jump in year 7, but this is largely explained by the two very large claims in that year); plot (d) is called a mean excess over threshold plot, also known in survival analysis as a mean residual life plot, which will be used later as a diagnostic in connection with the Generalized Pareto distribution. Summarizing, we can outine the following questions for discussion:

1. What is the distribution of insurance claims, focussing particularly on the very large claims?
2. Is there evidence of trend over time in either the claim times or the claim amounts?
3. Are there significantly different distributions among the seven claim types?
4. What is the likely distribution of future claims?

Each of these questions is important, but particularly, the likely distribution of future claims is relevant to risk assessment and is arguably the main question of interest to the company.

The original data on which this example are no longer accessible and were in any case confidential data, but to allow new analyses and comparisons, a simulated set of data has been constructed to mimic many of the properties of the original dataset.

1.1.3 Women's Track Times

1993 saw something sensational in the world of women's middle- and long-distance running. At an event in Beijing, several Chinese athletes broke world records by large margins; particular attention fell on the performances of Wang Junxia, who ran 8 minutes 6.11 seconds for the 3,000 meter event (more than 16 seconds faster than the world record prior to this event), and 29 minutes 31.78 seconds for 10,000 meters (41 seconds faster than the previous record). Rumors of drug use spread very rapidly, but no athlete failed a drug test and the only evidence available at the time was in the performances themselves. Subsequent papers by Robinson and Tawn [200] and Smith [220] added strength to the claim that these events were indeed vary unusual, but it was not until many years after the event that direct evidence of illegal drug use was made public.

Fig. 1.4 shows the ten best performances by different athletes in the women's 3,000 meter and 1,500 meter events, for the period 1974–1993. The data were recompiled for this publication from the World Athletics website (https://www.worldathletics.org/records/all-time-toplists) and differ slightly from the top-5 lists used in the original publications. In both cases, a large improvement can be seen in 1993, more dramatically so for the 3,000 meter event than for 1,500 meters. (The corresponding plot for 10,000 meters is not shown because detailed records are only available from 1984 onwards, which would make statistical analysis of the results more problematic.) As in our examples for the weather and insurance datasets discussed earlier, a key question is to characterize the probability distribution of extreme events, but in this case, the focus is on the joint distribution of the best 5 or the best 10 performances rather than the best one in each year. A minor difference from the previous examples is that, in this case, the focus is on minima rather than maxima, but minima are easily converted into maxima by simply reversing the signs, so we do not need a whole separate theory for this. However, as will be seen, the joint distribution for the best k events per year, where $k > 1$ (in this case, $k = 5$ or 10), is a direct extension of the case $k = 1$, so long as the events themselves can reasonably be treated as independent.

Given that the mere occurrence of a new world record would not, in itself, pro-

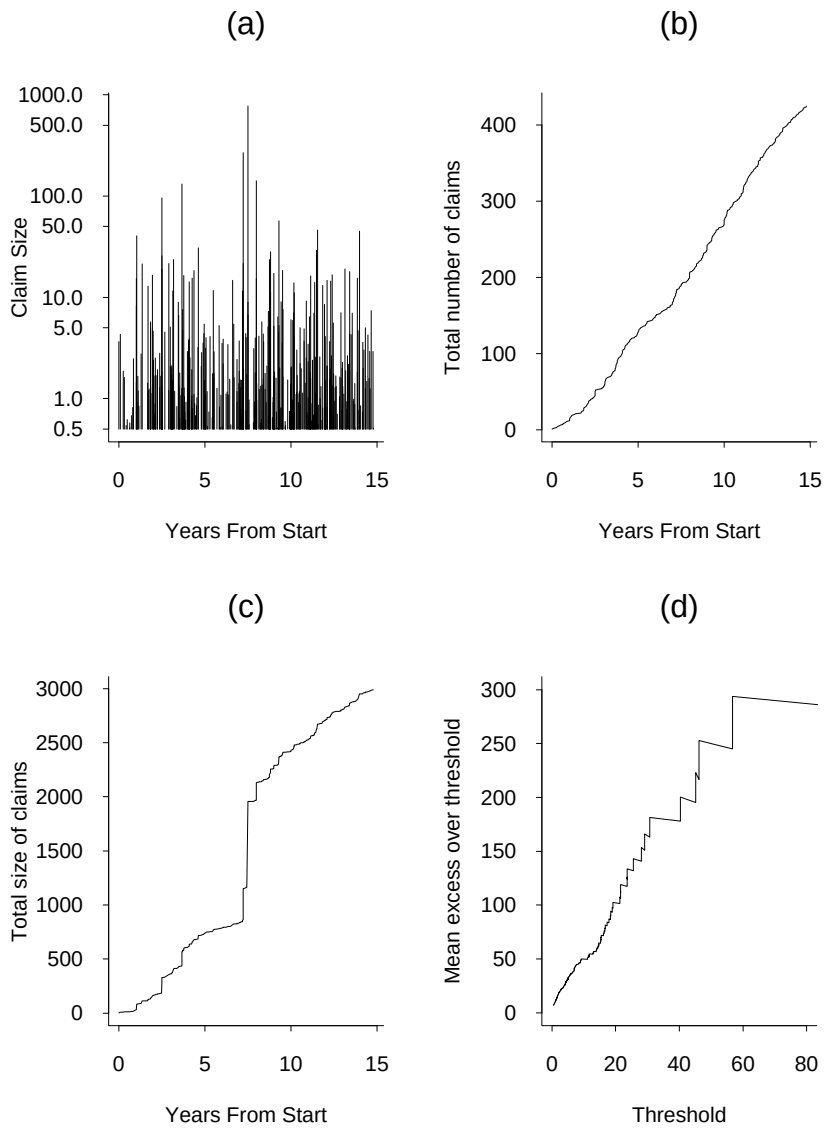

Figure 1.3 *Insurance data*

Figure 1.4 *Ten best times by different athletes in each year in the women's 3000 m. and 1500 m. events*

vide any reason to doubt its validity, the focus is more on the margin of the new world record. Smith [220] framed the question as "Given that a new world record occurred in 1993, what is the probability based on data prior to 1993 that the margin of improvement is equal to or greater than what was actually observed?" The answer to this question involved reformulating the question as one of predictive inference and solving it through a Bayesian analysis.

1.2 Overview of Univariate Extremes

1.2.1 The Three Types Theorem and the Generalized Extreme Value Distribution

The traditional starting point of extreme value theory is the *Three Types Theorem*, first stated by Fisher and Tippett [77] and later rederived (independently and more rigorously) by Gnedenko [87]. Suppose $X_1, X_2, \ldots,$ are a sequence of independent, identically distributed (i.i.d.) random variables whose common distribution function is F, that is, $F(x) = \Pr\{X_i \leq x\}$ for each i. Let M_n be the maximum of $X_1, \ldots, X_n$. An immediate result from the independence is

$$\Pr\{M_n \leq x\} \quad = \quad F^n(x). \tag{1.1}$$

The result (1.1) is not, in itself, very interesting because, for any x such that $0 < F(x) < 1$, $F^n(x) \to 0$ as $n \to \infty$. Another way to express that is to say $M_n \xrightarrow{p} \omega_F$ as $n \to \infty$, where $\xrightarrow{p}$ means convergence in probability and ω_F is the right-hand endpoint of the distribution defined by F, in other words, $\omega_F = \sup\{x : F(x) < 1\}$, which may be finite or infinite. However, in plain English, all this is saying is that as the sample size grows, the sample maxima get closer and closer to the right-hand endpoint of the distribution, whether that's finite or infinite, and this is not saying anything of fundamental importance. (As a side comment, the convergence of M_n to ω_F also holds in other modes of convergence, such as almost sure convergence and, if second moments are finite, L^2 convergence, but in this book, when we talk about convergence

of random variables, most of the time, this will be either convergence in probability or convergence in distribution.)

The theory becomes much more interesting if we *renormalize*, that is, allow for a scaling constant $a_n > 0$ and a location constant $b_n \in \mathbb{R}$, such that the renormalized sample maxima converge to a limiting distribution that is *nondegenerate*, in other words,

$$\Pr\left\{\frac{M_n - b_n}{a_n} \le x\right\} \quad = \quad F^n(a_n x + b_n) \;\to\; H(x) \tag{1.2}$$

where H is a nondegenerate distribution function, that is, the distribution function of some random variable that does not reduce to a deterministic constant (such as ω_F in our earlier discussion.)

The *Three Types Theorem* asserts that if nondegenerate H exists, it must be one of three types:

$$H(x) \quad = \quad \exp(-e^{-x}) \text{ for all } x \in \mathbb{R}, \tag{1.3}$$

$$H(x) \quad = \quad \begin{cases} 0 & \text{if } x < 0, \\ \exp(-x^{-\alpha}) & \text{if } x > 0, \end{cases} \tag{1.4}$$

$$H(x) \quad = \quad \begin{cases} \exp(-|x|^{\alpha}) & \text{if } x < 0, \\ 1 & \text{if } x > 0, \end{cases} \tag{1.5}$$

where in (1.4) and (1.5), $\alpha > 0$ is any positive constant.

We should clarify what we mean by "type" here. Two distribution functions H and H_1 are said to be *of the same type* if

$$H_1(x) \quad = \quad H(Ax + B) \text{ for all } x,$$

for some $A > 0$ and $B \in \mathbb{R}$. In other words, one distribution is derived from the other by a simple location-scale shift. It is a fundamental property of a convergence result like (1.2) that it can determine H only up to type, that is, if one nondegenerate H can arise as a limit in (1.2), so can any other distribution function of the form $H(Ax+B)$ for $A > 0$ and $B \in \mathbb{R}$. To see this, suppose (1.2) holds, and define $a'_n = a_n A,\ b'_n = a_n B + b_n$. Then

$$F^n(a'_n x + b'_n) \quad = \quad F^n(a_n A x + a_n B + b_n) \;=\; F^n(a_n(Ax+B) + b_n) \;\to\; H(Ax+B),$$

in other words, $H_1(x) = H(Ax+B)$ can arise as a limit by simply changing the normalizing constants a_n and b_n. As a consequence of this, the most we can hope to do, in determining the form of the limit H, is to define it up to transformations by type. Expressed another way, if any of (1.3)–(1.5) holds as a limiting results in (1.2), so does $H(Ax+B)$ for any $A > 0$ and $B \in \mathbb{R}$.

The distributions (1.3)–(1.5) are often called, respectively, the Gumbel, Fréchet and Weibull distributions (or types), following fundamental early works such as [93, 79, 253]. As a side note, Weibull was an engineer with a fundamental interest in strength of materials, and originally derived the distribution that bears his name from

the notion that the strength of a system of independent elements may be represented as the minimum of the strengths of the individual elements. In that form, it is often written in the form $F(x) = 0$ for $x < 0$, $1 - \exp\{-(x/\sigma)^\alpha\}$ for $x > 0$, where $\sigma > 0$ is a scaling constant and $\alpha > 0$ as in (1.5). However, this is just the mirror image of (1.5), and we have already pointed out that maxima may be transformed into minima by simply reversing the signs.

The three types may be combined into a single *Generalized Extreme Value* (GEV) distribution:

$$H(x) \quad = \quad \begin{cases} \exp\left\{-\left(1+\xi\frac{x-\mu}{\psi}\right)_+^{-1/\xi}\right\} & \text{if } \xi \neq 0, \\ \exp\left\{-\exp\left(-\frac{x-\mu}{\psi}\right)\right\} & \text{if } \xi = 0, \end{cases} \tag{1.6}$$

where μ is a location parameter, $\psi > 0$ is a scale parameter and ξ is a shape parameter, and we interpret $\left(1+\xi\frac{x-\mu}{\psi}\right)_+$ to be the larger of $\left(1+\xi\frac{x-\mu}{\psi}\right)$ and 0. The case $\xi = 0$ corresponds to the limit of the previous line as $\xi \to 0$ and represents the Gumbel distribution, while $\xi > 0$ to the Fréchet distribution with $\alpha = 1/\xi$, and $\xi < 0$ to the Weibull distribution with $\alpha = -1/\xi$. We often refer to $\xi > 0$ as the "long-tailed" case, where $1 - F(x) \propto x^{-1/\xi}$ as $x \to \infty$, $\xi = 0$ as the "exponential tail" case, and $\xi < 0$ as the "short-tailed" case, which has a finite right-hand endpoint at $\omega_H = \mu - \xi/\psi$. Note that since we have included a location parameter μ and a scale parameter ψ in the definition of (1.6), there is no need for a separate statement about all distributions of the same type, since all such cases are already included.

The GEV distribution was first proposed by von Mises [161] but was neglected for many years. Its revival owes partly to the work of the British meteorologist A.F. Jenkinson [127, 128], whose advocacy anticipated the modern treatment of this distribution with automated methods of estimation including the maximum likelihood and probability-weighted moments (PWM) methods which we shall discuss in some detail later.

1.2.2 Exceedances Over Thresholds

The second fundamental distributional result about extremes concerns exceedances over a high threshold. In many contexts, the distribution of very high (or low) values is the main object of interest; for instance, temperatures over 35°C, or insurance claims over \$10 million. A natural way to think of such values is to model the exceedances over a threshold; for instance, given that the daily rainfall at a site is over 10 cm., what is the distribution of the excess, i.e. the amount by which the rainfall on a given day exceeds 10 cm.? Mathematically, this is equivalent to considering the distribution of X conditionally on exceeding some high threshold u:

$$F_u(y) \quad = \quad \Pr(Y \leq u + y \mid Y > u) \; = \; \frac{F(u+y) - F(u)}{1 - F(u)}, \; y > 0,$$

where Y is a random variable with distribution function F. As $u \to \omega_F = \sup\{x : F(x) < 1\}$, often find a limit

$$\lim_{u \to \omega_F} F_u\left(\frac{y}{\sigma_u}\right) \quad = \quad G(y; 1, \xi) \tag{1.7}$$

where $\{\sigma_u\}$ is a sequence of scaling constants and G is the *Generalized Pareto Distribution* (GPD)

$$G(y; \sigma, \xi) = \begin{cases} 1 - \left(1 + \xi \frac{y}{\sigma}\right)_+^{-1/\xi} & \text{if } \xi \neq 0, \\ 1 - e^{-y/\sigma} & \text{if } \xi = 0. \end{cases} \tag{1.8}$$

Although (1.7) combined with (1.8) is the correct formal definition, in practice we often express the result more informally as

$$F_u(y) \quad \approx \quad G(y; \sigma_u, \xi) \text{ for } y > 0 \text{ and } u \text{ large.} \tag{1.9}$$

In (1.8), the interpretation of $(\ldots)_+$ is the same as in (1.6), i.e. the larger of the included term or 0.

The fact that (1.7) implies (1.8) is the threshold-exceedances equivalent of the Three Types Theorem. Pickands [182] proved that, for a given distribution function F, (1.7) and (1.8) hold if and only if F is in a domain of attraction for sample maxima, i.e. (1.2) holds for some non-degenerate H (which, by our previous discussion, is necessarily of GEV form). In fact, part of Pickands' result is that the ξ arising in (1.8) is the same ξ as in the equivalent GEV representation arising from (1.2). Balkema and de Haan [10] established the same result for the cases $\xi \geq 0$, and other results for what in survival analysis is known as the residual life distribution.

However, when thinking about extremes in terms of exceedances over a high threshold, the GPD is only part of the story. We must also consider the rate of exceedances over the threshold as well. In a context like climate change, we may well want to test whether the rate of high-threshold exceedances is increasing over time, which is a different question from whether the distribution of excess values over the threshold is increasing. The two may be tied together in the *Poisson-GPD model for extremes*, as follows:

1. The number, N, of exceedances of the level u in any one year has a Poisson distribution with mean λ,
2. Conditionally on $N \geq 1$, the excess values $Y_1, \ldots, Y_N$ are IID from the GPD.

Note that, with the inclusion of the parameter λ for the rate of exceedance, this model again has three parameters, same as the GEV. In fact, the two are closely related, as shown by the following argument.

Suppose $x > u$ and $\xi \neq 0$. The probability that the annual maximum of the

Poisson-GPD process is less than x is

$$\begin{aligned}\Pr\{\max_{1\le i\le N} Y_i \le x\} &= \Pr\{N=0\} + \sum_{n=1}^{\infty} \Pr\{N=n,\ Y_1 \le x,\ \ldots\ Y_n \le x\} \\ &= e^{-\lambda} + \sum_{n=1}^{\infty} \frac{\lambda^n e^{-\lambda}}{n!} \left\{1 - \left(1+\xi\frac{x-u}{\sigma}\right)_+^{-1/\xi}\right\}^n \\ &= \exp\left\{-\lambda\left(1+\xi\frac{x-u}{\sigma}\right)_+^{-1/\xi}\right\}. \end{aligned} \tag{1.10}$$

The right hand side of (1.10) is also of GEV form, equivalent to (1.6) with change of parameters.

1.2.3 The r-largest Order Statistics Approach

Suppose (1.2) holds with limit H of GEV form (1.6). Suppose, instead of just considering the maximum value from a sample of size n, we consider the joint distribution of the largest r order statistics, where $r \ge 1$ is fixed in the limit $n \to \infty$. If If $X_{n,1} \ge X_{n,2} \ge \ldots \ge X_{n,r}$ are the r largest order statistics from an IID sample of size n, and a_n and b_n are the same EVT normalizing constants as in (1.2), then

$$\left(\frac{X_{n,1}-b_n}{a_n}, \ldots, \frac{X_{n,r}-b_n}{a_n}\right)$$

converges in distribution to a limiting random vector $(Y_1, \ldots, Y_r)$, whose density is

$$h(y_1,\ldots,y_r) = \psi^{-r}\prod_{i=1}^{r}\left(1+\xi\frac{y_i-\mu}{\psi}\right)^{-1/\xi-1} \exp\left\{-\left(1+\xi\frac{y_r-\mu}{\psi}\right)^{-1/\xi}\right\} \tag{1.11}$$

provided only that $1+\xi\frac{y_i-\mu}{\psi} > 0$ for each of $i = 1,\ldots,r$. The equivalent form of (1.11) when $\xi = 0$ is given by

$$h(y_1,\ldots,y_r) = \psi^{-r}\prod_{i=1}^{r}\left\{e^{-(y_i-\mu)/\psi}\right\}\exp\left\{-e^{-(y_r-\mu)/\psi}\right\}. \tag{1.12}$$

Results equivalent to (1.11) or (1.12) were first developed in a series of papers on *extremal processes* by Dwass [62], Lamperti [137] and Weissman [254, 256, 255]; the first paper to propose this as a statistical model was by Weissman [257] and subsequently extended by Gomes [90], Smith and Weissman [233], Smith [219] and Tawn [239]. As a notational point, earlier authors used k rather than r to denote the number of order statistics, but Smith [219] wrote r to avoid confusion with the fact that k was, at that time, used by many authors as the shape parameter of the GEV distribution (equivalent to $-\xi$ in the notation of (1.6)). That notation has largely fallen out of use so the distinction seems irrelevant now.

From an applied viewpoint, this model seems natural when the data consist of the r or k largest or smallest values in each year, such as in our example of women's track times (Section 1.1.3), but it should be noted that the model only applies where the order statistics are derived from independent observations. In the women's track times example, this condition has been fulfilled by listing the best times in each year by different athletes.

1.2.4 Point Process Approach

The point process viewpoint was originally proposed by Pickands [181] and adopted for statistical estimation purposes in [225]. For a more up to date presentation, we refer particularly to Chapter 7 of Coles [34] and the review in [229].

Before describing the approach, we give some background about point processes and in particular the nonhomogenous Poisson process which is a key motivation for the methodology. Our explanation will closely follow previous expositions given by Coles [34] and Smith[229], but for a more detailed a mathematically rigorous treatment we refer to Resnick[196].

To explain this in a little more detail, we first elaborate on what the representation as a nonhomogeneous Poisson process actually means. Our explanation will closely follow previous expositions given in [34, 229], or [196] for a more thorough treatment of point processes in an extreme value context.

A *point process* on some sample space $\mathscr{S}$ is a stochastic process that places points at random locations in $\mathscr{S}$. Typically $\mathscr{S}$ is a complete separable metric space (in all our examples, $\mathscr{S}$ will be a subset of $\mathbb{R}^d$ for some d) and will be specific by intervalued-random variables $N(A)$, $A \in \mathscr{A}$ where $\mathscr{A}$ is a family of subsets of $\mathscr{S}$, for example, all the Borel sets. The point process is *simple* if it does not have multiple points, in other words, for any single-point set $A = \{x\}$ for some $x \in \mathscr{S}$, $N(A)$ is either 0 or 1.

A point process N is said to be a *nonhomogeneous Poisson process with intensity measure* Λ if it satisfies:

(i) N is simple,

(ii) $N(A_1), \ldots, N(A_k)$ are independent random variables for any sequence of *disjoint* sets $A_1, \ldots, A_k$ (disjoint means that the intersection of A_i and A_j is empty whenever $i \neq j$),

(iii) $\Pr\{N(A) = n\} = \frac{\Lambda(A)^n e^{-\Lambda(A)}}{n!}$ for $n = 0, 1, 2, \ldots$.

Property (iii) is, of course, equivalent to the statement that $N(A)$ has a Poisson distribution with mean $\Lambda(A)$, hence the name.

For statistical estimation purposes, we would like to define the joint density of the points in a nonhomogenous Poisson process. To define this, suppose first that Λ has a continuous density $\lambda(\cdot)$, in other words,

$$\Lambda(A) \quad = \quad \int_A \lambda(\mathbf{x}) d\mathbf{x}, \; A \in \mathscr{A}. \tag{1.13}$$

We use vector notation here because, in general, $\mathbf{x}$ is a vector in the d-dimensional set $\mathscr{S}$ for some $d \geq 1$.

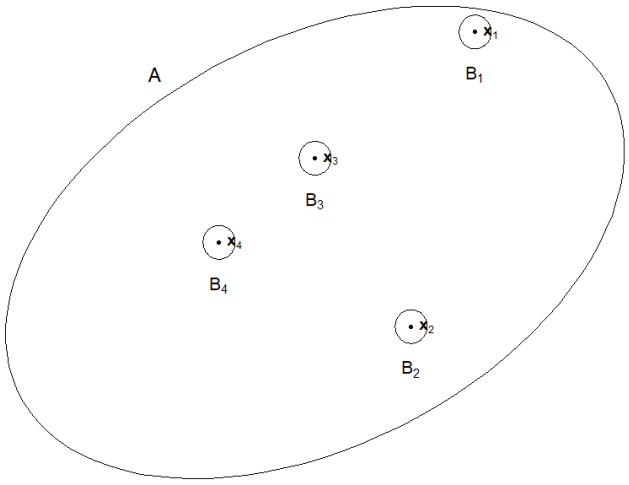

Figure 1.5 *Illustration of joint density calculation for a nonhomogeneous Poisson process.*

Now suppose there are a random number $N(A)$ points in some observed subset $A \subset \mathscr{S}$ and that they are at locations $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{N(A)}$. The *joint density* of these $N(A)$ points, at locations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{N(A)}$, is given by the expression

$$\left\{ \prod_{i=1}^{N(A)} \lambda(\mathbf{x}_i) \right\} \cdot e^{-\Lambda(A)}. \tag{1.14}$$

To explain and justify the formula (1.14), see Figure 1.5. In this case, we are assuming $N(A) = 4$, and that the points are located at $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$. What is the probability of this precise configuration? To answer this question, suppose we surround each point with a ball of some small radius δ; we assume δ is small enough so that the balls do not overlap (recall that we assumed the process does not have multiple points, so some such δ must exist). The probability of this precise configuration is

$$\begin{aligned} &\Lambda(B_1)e^{-\Lambda(B_1)} \cdot \Lambda(B_2)e^{-\Lambda(B_2)} \cdot \Lambda(B_3)e^{-\Lambda(B_3)} \cdot \Lambda(B_4)e^{-\Lambda(B_4)} \cdot e^{-\Lambda(A \setminus B_1 \setminus B_2 \setminus B_3 \setminus B_4)} \\ &\quad = \quad \Lambda(B_1) \cdot \Lambda(B_2) \cdot \Lambda(B_3) \cdot \Lambda(B_4) \cdot e^{-\Lambda(A)} \end{aligned}$$

However, for small δ, we can write $\Lambda(B_i) = \delta\lambda(\mathbf{x}_i)(1 + o(1))$ for $i = 1,2,3,4$, as $\delta \downarrow 0$. Therefore, the last expression may also be written

$$= \quad \delta^{N(A)} \prod_{i=1}^{N(A)} \lambda(\mathbf{x}_i) e^{-\Lambda(A)} (1 + o(1))$$

which is asymptotically proportional to (1.14) as $\delta \downarrow 0$. The factor $\delta^{N(A)}$ plays no role

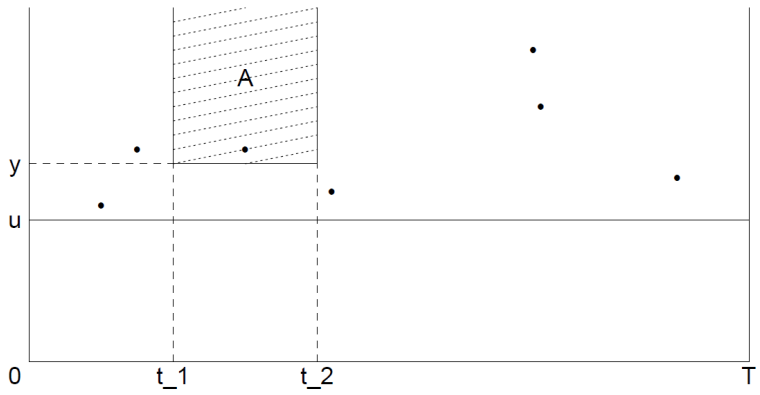

Figure 1.6 *Illustration of Point Process Model.*

in any statistical operation based on maximizing or integrating te likelihood function and may therefore be ignored.

Now we can explain the point process approach to sample extremes. The key idea is represented in Fig. 1.6. In this viewpoint, the exceedances over a high threshold u are represented by points on a two-dimensional diagram, where time is represented on the horizontal access and the level of the process on the vertical axis. In the limit as $u \to \omega_F$ and the time constant $T \to \infty$, this point process is approximated by a two dimensional point process, whose intensity measure is represented by the formula

$$\Lambda(A) \quad = \quad (t_2 - t_1)\left(1 + \xi\frac{y-\mu}{\psi}\right)_+^{-1/\xi}, \; y \geq u. \tag{1.15}$$

Here, A represents a set of the form $(t_2 - t_1) \times (y, \infty)$, i.e. A contains all the events within a time interval (t_1, t_2) that exceed some high level $y \geq u$. $\Lambda(A)$ is then the expected number of events within the set A. It can be shown that the GEV distribution for annual maxima, the GPD for high threshold exceedances, and the formula (1.11), can all be derived from (1.15), at least so long as we confine our attention to events above the threshold u.

Note that we may also express (1.15) in the form

$$\Lambda(A) \quad = \quad \int_{t_1}^{t_2}\int_{y}^{\infty}\frac{1}{\psi}\left(1 + \xi\frac{x-\mu}{\psi}\right)_+^{-1/\xi-1} dx.$$

Therefore, in this model, $\lambda(t,y) = \frac{1}{\psi}\left(1 + \xi\frac{y-\mu}{\psi}\right)^{-1/\xi-1}$ defined wherever $y \geq u$ and $1 + \xi\frac{y-\mu}{\psi} > 0$. Note that the expression for $\lambda(t,y)$ does not depend on t, since in this simplest form of the model, the process is assumed uniform along the time axis.

We will now show that each of our previous models — the GEV for sample maxima, the GPD for exceedances over a threshold, and the r-largest order statistics approach — can be derived from the above representation.

1.2.4.1 GEV model

Consider a time interval of length 1 in whatever time unit is being considered. In environmental applications, the most common unit of time is one year, so that we are talking about annual maxima, but other intervals of time can be ocnsidered. Since our discuss is purely theoretical for the time being, we are ignoring practical considerations such as seasonality and trend.

So, suppose in Figure 1.6, we have $t_2 = t_1 + 1$, and we are considering some $y > u$. Let M_{t_1,t_2} denote the maximum of the process over the time interval (t_1,t_2). Also define the set A to be $(t_1,t_2) \times (y,\infty)$. So,

$$\begin{aligned}
\Pr\{M_{t_1,t_2} \leq u\} &= \Pr\{N(A) = 0\} \\
&= e^{-\Lambda(A)} \\
&= \exp\left\{ -\left(1 + \xi \frac{y-\mu}{\psi}\right)_+^{-1/\xi} \right\}
\end{aligned}$$

which is exactly of GEV form. So the GEV distributon for annual maxima can be derived from the point process model.

1.2.4.2 GPD model

Suppose, within some very narrow time interval $(t,t+\delta)$, there is exactly one point above the threshold u. For any level $y > u$, the probability that this point exceeds y is

$$\Pr\{N(A_2) = 1 \mid N(A_1) = 1\}$$

where A_2 is the set $(t,t+\delta) \times (y,\infty)$ and A_1 is the set $(t,t+\delta) \times (u,\infty)$. Assume for simplicity that both $1 + \xi(y-\mu) > 0$ and $1 + \xi(u-\mu) > 0$. This conditional probability reduces to

$$\begin{aligned}
\frac{\Pr\{N(A_2) = 1 \text{ and } N(A_1) = 1\}}{\Pr\{N(A_1) = 1\}} &= \frac{\Pr\{N(A_2) = 1 \text{ and } N(A_1 \setminus A_2) = 0\}}{\Pr\{N(A_1) = 1\}} \\
&= \frac{\Lambda(A_2) e^{-\Lambda(A_2)} \cdot e^{-\Lambda(A_1 \setminus A_2)}}{\Lambda(A_1) e^{-\Lambda(A_1)}} \\
&= \frac{\Lambda(A_2)}{\Lambda(A_1)} \\
&\approx \frac{\delta\left(1 + \xi\frac{y-\mu}{\sigma}\right)^{-1/\xi}}{\delta\left(1 + \xi\frac{u-\mu}{\sigma}\right)^{-1/\xi}} \\
&= \left\{ 1 + \frac{\xi(y-u)}{\psi + \xi(u-\mu)} \right\}^{-1/\xi}
\end{aligned}$$

which is equivalent to the GPD form (1.9) if we define $\sigma_u = \psi + \xi(u-\mu)$.

1.2.4.3 r-largest order statistics model

Now suppose we are again considering a time interval of unit length, so $t_2 = t_1 + 1$, and let $y_1 > y_2 > \ldots > y_r$ be the r largest order statistics of the observations within this time interval. This is equivalent to saying that, if we define the set $A = (t_1, t_2) \times (y_r, \infty)$, there are exactly r points within the set A, and moreover, their positions on the vertical scale are at $y_1, \ldots, y_r$. By the same reasoning as led to (1.14), the joint density of these observations is given by

$$\prod_{i=1}^{r} \left\{ \frac{1}{\psi} \left(1 + \xi \frac{y_i - \mu}{\psi} \right)^{-1/\xi - 1} \right\} \cdot \exp \left\{ - \left(1 + \xi \frac{y_r - \mu}{\psi} \right)^{-1/\xi} \right\} \tag{1.16}$$

provided $1 + \xi \frac{y_i - \mu}{\psi} > 0$ for $i = 1, \ldots, r$. Ths is the same formula as (1.11).

Thus the GEV, GPD, r-largest and point process frameworks are all effectively equivalent, and which one should apply in any specific analysis is often a matter of taste, though it is sometimes constrained by the nature of the data available. One disadvantage which is often cited of threshold-exceedance approaches is that they require the specification of a specific threshold and how to do this is in many cases not entirely clear, but the same is true of the GEV and r-largest approaches as well, i.e. we have to define a suitable block size for computing maxima, and for the r-largest approach, we also have to specify r. In environmental applications, it is common to assume a time unit of one year, which is often justified by the fact that many environmental processes display some annual periodicity, so if we take annual maxima, then it becomes reasonable to assume that they are (approximately) independent with (approximately) the same distribution. However, this is more a matter of convenience than necessity. In Chapter 2, we shall consider in more detail how these approximations are justified and shows some theoretical results relevant to determining the optional block length or threshold.

1.3 Estimation

1.3.1 GEV Model

Finding appropriate distributions for extremes of natural processes is a problem that has been carefully studied for at least 100 years. Before the present-day interest in weather extremes, many of the most studies applications were to extreme in hydrological series, such as the annual maximum of a daily river flow series. The famous book by Gumbel [93] was motivated by these problems, but the approach is by now quite out of date. Briefly, Gumbel considered the three types of (1.3)–(1.5) as three separate families of distributions, and provided moments-based estimators for the location and scale parameters conditional on the type (and, in the case of (1.4) or (1.5), the parameter α). For the determination of type, and if needed an estimate of α, Gumbel relied on probability plotting techniques, essentially, fitting a curve through a normalized probability plot. This method had the advantage of being straightforward to apply without advanced computation, but it relied on visual judgment to

determine the appropriate curve, and in any case, equivalent computations may be made automatically with modern computers.

In contrast, the papers of Jenkinson [127, 128] were remarkably modern in their approach. Jenkinson [127] derived a formula equivalent to the GEV distribution, though he was apparently unaware of the precedent of von Mises [161]. This approach was extended by Jenkinson [128] who gave the first treatment of maximum likelihood for this problem.

In this connection, Jenkinson commented "The method of Maximum Likelihood ... is generally accepted as being the best for estimation of parameters; and although it requires a great deal of computation it must be considered essential when important decisions depend on the estimation, especially if the data are not a large sample and/or they are rather irregular." ([128], page 196). He went on to present an entirely hand-computation method of computing the maximum likelihood estimates, though he also proposed a *method of sextiles* that very closely approproximated the MLE. Briefly, his proposal was to calculate means in each of the sextiles of the distribution: w_1 for the mean of the lower one sixth of the order statistics, w_2 for the next sixth, and so on up to w_6 for the largest sixth. He then calculated the ratio $(w_2 - w_1)/(w_6 - w_5)$ — note that this quantity is location-scale invariant and, therefore, its distribution depends only on the GEV shape parameter, which we now call ξ but Jenkinson defined instead a quantity k, equivalent to $-\xi$ in our notation. Then, Jenkinson provided a table that translated the value for $(w_2 - w_1)/(w_6 - w_5)$ into an estimate of k that in many cases is very close to the MLE of k.

Value	Occurrences	Value	Occurrences	Value	Occurrences
12	1	19	11	25	5
14	2	20	9	26	5
15	4	21	21	27	3
16	4	22	6	28	1
17	3	23	8	29	1
18	4	24	3	30	1

Table 1.2 *Jenkinson's [128] table of annual maximum floods in Hartford, CT, 1843–1934*

Example. Table 1.2 is taken from [128]. Based on 92 years of annual maximum floods in Hartford, Connecticut, Jankinson calculated $\frac{w_2 - w_1}{w_6 - w_5} = 0.937$ and, based on his own tables, estimated ξ to be -0.26 (in Jenkinson's own notation, $k = 0.26$). He then showed how to improve that estimate by, in effect, a Newton-Raphson procedure to find the exact solutions of the likelihood equations (equivalent to equations (1.21) following). These were all hand calculations, apparently performed without any electronic assistance. After two iterations he claimed the final result as follows (again, translated to the notation of this chapter): $\hat{\mu} = 19.68,\ \hat{\psi} = 3.48,\ \hat{\xi} = -0.258$. He also stated, "The absolute maximum flood stage is estimated at 33.2 feet; and that for $T = 1000$ years is 30.9 feet." (He also quoted the $T = 100$ years value as 29.0 feet.)

For the present discussion, these values have been recalculated using the numer-

ical methods to be described in more detail below: $\hat{\mu} = 19.6809$, $\log\hat{\psi} = 1.2467$ (which translates to $\hat{\psi} = 3.4788$) and $\hat{\xi} = -0.2575$. The estimated endpoint is $\hat{\mu} - \hat{\psi}/\hat{\xi} = 33.193$. For the T-year return value, we solve $\exp\left\{-\left(1+\xi\frac{y-\mu}{\psi}\right)\right\} = 1-1/T$ which leads to

$$RV_T \quad = \quad \mu + \psi \cdot \frac{\{-\log(1-1/T)\}^{-\xi} - 1}{\xi}. \tag{1.17}$$

Substituting $\hat{\mu}, \hat{\psi}, \hat{\xi}$ for μ, ψ, ξ in (1.45) and setting $T = 1000$ and $T = 100$, we deduce $\widehat{RV_{1000}} = 30.9105$ and $\widehat{RV_{100}} = 29.0590$, fully consistent with Jenkinson's results.

To go into more detail about the equations for the MLE, based on the GEV model (1.6), we have the density

$$\begin{aligned} h(x\,;\,\mu,\psi,\xi) \quad &= \quad \frac{dH(x)}{dx} \\ &= \quad \begin{cases} \frac{1}{\psi}\left(1+\xi\frac{x-\mu}{\psi}\right)^{-1/\xi-1}\exp\left\{-\left(1+\xi\frac{x-\mu}{\psi}\right)_+^{-1/\xi}\right\} & \text{if } \xi \neq 0, \\ \frac{1}{\psi}\exp\left\{-\exp\left(-\frac{x-\mu}{\psi}\right) - \frac{x-\mu}{\psi}\right\} & \text{if } \xi = 0. \end{cases} \end{aligned} \tag{1.18}$$

We have expressed the function h as a function of μ, ψ and ξ, as well as x, to facilitate the transition to the following discussion of the method of maximum likelihood estimation (MLE).

From now on we consider only the case $\xi \neq 0$: even if we suspect that the true value of ξ is 0, there can still be advantages to assuming a GEV with $\xi \neq 0$ (see the discussion of penultimate approximations in Chapter 2) and in any case, the MLE never leads to $\hat{\xi}$ of exactly 0, though it may well be indistinguishable from 0 in the sense that a hypothesis test of $H_0 : \xi = 0$ against the alternative $H_1 : \xi \neq 0$ does not reject the null hypothesis. We shall see many examples like this later. Proceeding for now, however, assuming $\xi \neq 0$, for a sample $Y_1, \ldots, Y_n$, the negative log likelihood (NLLH) is given by

$$\begin{aligned} \ell(\mu,\psi,\xi \mid Y_1,\ldots,Y_n) \;&=\; \sum_{i=1}^{n}\{-\log h(Y_i)\,;\,\mu,\psi,\xi)\} \\ = \quad \sum_{i=1}^{n}&\left\{\log\psi + \left(\frac{1}{\xi}+1\right)\log\left(1+\xi\frac{Y_i-\mu}{\psi}\right) + \left(1+\xi\frac{Y_i-\mu}{\psi}\right)^{-1/\xi}\right\} \end{aligned} \tag{1.19}$$

the whole expression being defined only when $1+\xi\frac{Y_i-\mu}{\psi} > 0$ for each Y_i (otherwise ℓ is technically $+\infty$; in practice, numerical routines often set ℓ to a very large value, say 10^{10}, when the constraints are violated).

The likelihood equations based on (1.19) are given by

$$\frac{\partial\ell}{\partial\mu} \;=\; \frac{\partial\ell}{\partial\psi} \;=\; \frac{\partial\ell}{\partial\xi} \;=\; 0, \tag{1.20}$$

where algebraic expressions for the partial derivatives may be derived by routine calculus manipulations, though in practice, are more often approximated numerically. Technically, a solution of (1.21) defines only a *local* maximum of the likelihood function, that is not necessarily a global maximum. In fact, it can be shown that if $\xi < -1$, the global maximum of the likelihood function is $+\infty$ (and hence $\ell \to -\infty$), achieved as $1 + \xi \frac{Y_{\max} - \mu}{\psi} \to 0$, $Y_{\max}$ denoting the largest of $Y_1, \ldots, Y_n$. However, this is the extremely short-tailed version of the GEV and is generally considered unrealistic in practice. At the other end of the scale, $\xi > 1$ is the extremely long-tailed case, corresponding to a distribution with infinite mean, and this is also considered unrealistic in most practical applications (though not all). Therefore, in practice it is very common to restrict $|\xi| < 1$ and to use a numerical optimization routine to minimize ℓ, which is equivalent in practice to solving the equations (1.21). In R, the optimization functions `optim` or `nlm` have been found to work well in practice.

This discussion assumes that there can only ever been one solution to the equations (1.21), which has never been proved, though there are no known counterexamples. Another practical adjustment which is often made is to replace ψ by $\log \psi$, on the grounds that $\log \psi$ is unrestricted in range, but this does not affect the theoretical prioerties of the estimators and we shall ignore that here.

Once we have obtained the MLE, the logical next step is to estimate the *standard errors* of the estimates. A key step here is to calculate the *observed information matrix*, derived from the hessian of ℓ, i.e. the matrix

$$I(\hat{\mu}, \hat{\psi}, \hat{\xi}) = \begin{pmatrix} \frac{\partial^2 \ell}{\partial \mu^2} & \frac{\partial^2 \ell}{\partial \mu \partial \psi} & \frac{\partial^2 \ell}{\partial \mu \partial \xi} \\ \frac{\partial^2 \ell}{\partial \mu \partial \psi} & \frac{\partial^2 \ell}{\partial \psi^2} & \frac{\partial^2 \ell}{\partial \psi \partial \xi} \\ \frac{\partial^2 \ell}{\partial \mu \partial \xi} & \frac{\partial^2 \ell}{\partial \psi \partial \xi} & \frac{\partial^2 \ell}{\partial \xi^2} \end{pmatrix} \tag{1.21}$$

where all the second-order partial derivatives are evaluated at the maximum likelihood estimators $(\hat{\mu}, \hat{\psi}, \hat{\xi})$. Again, it is possible but extremely tedious to calcuate all the second-order derivatives by direct calculus methods, but it is common in practice to do this numerically as well, for example, using the `hessian=T` option with the numerical routines `optim` or `nlm`.

Once we have the observed information matrix $I(\hat{\mu}, \hat{\psi}, \hat{\xi})$, according to conventional maximum likelihood theory, its inverse $I^{-1}(\hat{\mu}, \hat{\psi}, \hat{\xi})$ is considered a good (asymptotic) approximation to the variance-covariance matrix of $(\hat{\mu}, \hat{\psi}, \hat{\xi})$. In particular, the square roots of the diagonal entries of $I^{-1}(\hat{\mu}, \hat{\psi}, \hat{\xi})$ are often displayed as the *standard errors* of $\hat{\mu}$, $\hat{\psi}$ and $\hat{\xi}$.

We can go further and calculate the *Fisher information matrix*, defined as

$$\mathscr{I}_n(\mu, \psi, \xi) = \begin{pmatrix} E\frac{\partial^2 \ell}{\partial \mu^2} & E\frac{\partial^2 \ell}{\partial \mu \partial \psi} & E\frac{\partial^2 \ell}{\partial \mu \partial \xi} \\ E\frac{\partial^2 \ell}{\partial \mu \partial \psi} & E\frac{\partial^2 \ell}{\partial \psi^2} & E\frac{\partial^2 \ell}{\partial \psi \partial \xi} \\ E\frac{\partial^2 \ell}{\partial \mu \partial \xi} & E\frac{\partial^2 \ell}{\partial \psi \partial \xi} & E\frac{\partial^2 \ell}{\partial \xi^2} \end{pmatrix} \tag{1.22}$$

where E denotes expected value, the partial derivatives are evaluated at the true values μ, ψ, ξ, and the expression is written $\mathscr{I}_n$ to denote also the dependence on sample

size n. Because the observations are IID, $\mathscr{I}_n = n\mathscr{I}$, for some matrix $\mathscr{I}$, which we very often we refer to as the Fisher information matrix.

The Fisher information matrix has a similar interpetation to the observed information matrix: in particular, $\mathscr{I}_n^{-1}$ is an approximation to the variance-covariance matrix of $(\hat{\mu}, \hat{\psi}, \hat{\xi})$, and its diagonal entries may also be quoted as standard errors of the MLEs. Ever since a famous paper of Efron and Hinkley [66], the approximation based on the observed information matrix is generally considered superior to the approximation based on the Fisher information matrix, though the latter remains important for theoretical calculations.

The first theoretical calculations of the Fisher information matrix for the GEV were due to Prescott and Walden [185, 186]. In particular, [185] showed that the Fisher information matrix $\mathscr{I}$ is finite when $\xi > -\frac{1}{2}$, and derived the matrix itself as

$$\mathscr{I}(\mu,\psi,\xi) = \begin{pmatrix} \frac{p}{\psi^2} & \frac{\Gamma(2+\xi)-p}{\psi^2\xi} & -\frac{1}{\psi\xi}\left(q-\frac{p}{\xi}\right) \\ \frac{\Gamma(2+\xi)-p}{\psi^2\xi} & \frac{1}{\psi^2\xi^2}\{1-2\Gamma(2+\xi)+p\} & -\frac{1}{\psi\xi^2}\left\{1-\gamma+\frac{1-\Gamma(2+\xi)}{\xi}-q+\frac{p}{\xi}\right\} \\ -\frac{1}{\psi\xi}\left(q-\frac{p}{\xi}\right) & -\frac{1}{\psi\xi^2}\left\{1-\gamma+\frac{1-\Gamma(2+\xi)}{\xi}-q+\frac{p}{\xi}\right\} & \frac{1}{\xi^2}\left\{\frac{\pi^2}{6}+\left(1-\gamma+\frac{1}{\xi}\right)^2-\frac{2q}{\xi}+\frac{p}{\xi^2}\right\} \end{pmatrix} \tag{1.23}$$

where $p = (1+\xi)^2\Gamma(1+2\xi)$, $q = \Gamma(2+\xi)\{\Psi(1+\xi)+1/\xi+1\}$, $\Gamma(x)$ is the gamma function, $\Psi(x) = \frac{d\log\Gamma(x)}{dx}$ is the digamma function, and $\gamma = 0.5772157..$ is Euler's constant.

A formal statement of the asymptotic normality of MLEs is given as

$$\sqrt{n}\left\{\begin{pmatrix}\hat{\mu}_n\\ \hat{\psi}_n\\ \hat{\xi}_n\end{pmatrix} - \begin{pmatrix}\mu\\ \psi\\ \xi\end{pmatrix}\right\} \quad \xrightarrow{d} \quad \mathscr{N}\left[\begin{pmatrix}0\\0\\0\end{pmatrix}, \mathscr{I}^{-1}\right], \tag{1.24}$$

a result originally given in [222] and refined by several recent authors [57, 24, 58, 265].

1.3.1.1 The delta method

Another question is how to calculate a standard error for a variable that is a nonlinear function of (μ,ψ,ξ), such as the endpoint ($\mu - \frac{\psi}{\xi}$ when $\xi < 0$) or RV_T as in (1.45). A common method calculating a standard error in this case is the *delta method*, described as follows.

To put it in its general context, suppose we rewrite the result (1.24) in the form

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \quad \xrightarrow{d} \quad \mathscr{N}\left[\mathbf{0}, \mathscr{I}^{-1}\right], \tag{1.25}$$

where $\boldsymbol{\theta} = \begin{pmatrix}\theta_1 & \theta_2 & \dots & \theta_d\end{pmatrix}^T$ is some d-dimensional (for finite d) vector of unknown parameters, $\hat{\boldsymbol{\theta}}_n$ is an estimator based on n observations, $\mathbf{0}$ is a vector of zeros, and $\mathscr{I}^{-1}$ is some limiting covariance matrix. In most cases of interest, $\hat{\boldsymbol{\theta}}_n$ will

be the MLE and $\mathscr{I}$ the Fisher information, but this is not necessary for the following result. Our previous result (1.24) is the special case of this for IID observations from the GEV where $\boldsymbol{\theta} = \begin{pmatrix} \mu & \psi & \xi \end{pmatrix}^T$ and $\mathscr{I}$ is given by (1.22).

Now suppose we are interested in some scalar quantity $g(\boldsymbol{\theta})$, a (typically nonlinear) function of the vector parameters $\boldsymbol{\theta}$. We assume the function g is continuously differentiable and write

$$\nabla g \quad = \quad \begin{pmatrix} \frac{\partial g}{\partial \theta_1} & \frac{\partial g}{\partial \theta_2} & \cdots & \frac{\partial g}{\partial \theta_d} \end{pmatrix}^T.$$

It is natural to estimate $g(\boldsymbol{\theta})$ by $g(\hat{\boldsymbol{\theta}}_n)$ — in other words, we estimate g by simply substituting our previous estimators $\hat{\boldsymbol{\theta}}$ for the unknown $\boldsymbol{\theta}$. The result is then

$$\sqrt{n}(g(\hat{\boldsymbol{\theta}}_n) - g(\boldsymbol{\theta})) \quad \xrightarrow{d} \quad \mathscr{N}\left[0, \nabla g(\boldsymbol{\theta}) \mathscr{I}^{-1} (\nabla g(\boldsymbol{\theta}))^T\right], \tag{1.26}$$

see, e.g. [247], Chapter 3..

In practice, we usually use the observed information matrix instead of the expected information matrix and do not explicitly write the $\sqrt{n}$ factor in (1.26), so the practical result is

$$Var\left\{g(\hat{\boldsymbol{\theta}}_n)\right\} \quad \approx \quad \nabla g(\hat{\boldsymbol{\theta}}_n) I_n^{-1} (\nabla g(\hat{\boldsymbol{\theta}}_n)^T. \tag{1.27}$$

As an example of the previous calculations, suppose we are interested in the function $g(\mu, \psi, \xi) = RV_T$ in the GEV case, given by (1.45) for a given value of the return time T. Here $g(\mu, \psi, \xi) = \mu + \psi \cdot \frac{\{-\log(1-1/T)\}^{-\xi} - 1}{\xi}$ and we can readily verify that ∇g is the vector

$$\begin{pmatrix} 1 & \frac{\{-\log(1-1/T)\}^{-\xi} - 1}{\xi} & \psi \left\{ - \frac{(-\log(1-1/T))^{-\xi} \log(-\log(1-1/T))}{\xi} - \frac{(-\log(1-1/T))^{-\xi} - 1}{\xi^2} \right\} \end{pmatrix}^T. \tag{1.28}$$

Substituting from (1.28) into (1.27) gives an approximation for the variance of $\widehat{RV_T}$, and the square root of this is usually called the standard error.

1.3.1.2 Bayesian methods

Another approach to the estimation of nonlinear functions of the form $g(\boldsymbol{\theta})$ is Bayesian, e.g. assume some prior density $\pi(\boldsymbol{\theta})$ for the parameter $\boldsymbol{\theta}$, then the posterior density $\pi(g(\boldsymbol{\theta}) \mid \mathbf{Y})$ based on a set of observations $\mathbf{Y} = \begin{pmatrix} Y_1 & Y_2 & \ldots & Y_n \end{pmatrix}$ is given by

$$\pi(g(\boldsymbol{\theta}) \mid \mathbf{Y}) \quad = \quad \frac{g(\boldsymbol{\theta}) L(\boldsymbol{\theta} \mid \mathbf{Y}) \pi(\boldsymbol{\theta})}{\int g(\boldsymbol{\theta}) L(\boldsymbol{\theta} \mid \mathbf{Y}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}}. \tag{1.29}$$

Here, $L(\boldsymbol{\theta} \mid \mathbf{Y})$ is the likelihood function, which n the case of IID pbservations is given by $\prod_{i=1}^n f(Y_i; \boldsymbol{\theta})$ or $e^{-\ell}$ in the notation of (1.19).

In practice, (1.29) is evaluated numerically, very commonly by some variant of

the Metropolis-Hastings algorithm [84]. We have found the adaptive Metropolis algorithm of [106] to be especially useful for problems of this nature.

R code for this method is given in *https://rls.sites.oasis.unc.edu/s834-2023/Data/AdaptMH.txt* (coding by the author, but very closely following the algorithm in [106]). This routine requires several parameters, as follows:

- `lh`: subroutine for the negative log likelihood function
- `par`: initial set of parameter estimates (in the following analyses, we use the MLE for this)
- `npar`: dimension of `par`
- `C0`: `npar` $\times$ `npar` matrix, initial guess of the posterior covariance matrix (we use the observed information matrix for this)
- `nsim`: total number of simulations
- `n1`: initial run with given `C0`
- `n2`: interval between subsequent updates
- `n3`: interval between updates of output posterior covariance matrix
- `eps`: some small positive number, used to insure the posterior covariance matrix does not become singular (we use 0.001);
- `scal`: scaling parameter, recommended value is 2.4 but user is allowed to change

There are two outputs:

- `parout`: output parameter matrix of dimension `n4` $\times$ `npar`, where `n4=floor(nsim/n3)`
- `acc`: acceptance rate, i.e. the proportion of Metropolis steps where the new paeameter vector is accepted. It is usually recommended that a value near 0.25 is optimal, though anywhere between 0.1 and 0.5 is acceptable. Outside that range, it is recommended to increase or decrease the step length, which may be achieved by adjusting `scal`.

For the examples that follow, we took `nsim` equal to 200,000, and `n2` equal to 10 (the values of `n1` and `n3` are less critical), so the output matrix `parout` has dimensions $20,000 \times 3$ (i.e. each row represents one sample from the posterior distribution of $(\mu, \log\psi, \xi)$. However, it is conventional to discard the first half of the sample as a "warm-up", so the final posterior density sample has dimensions $10,000 \times 3$. We used $\log\psi$ instead of ψ to improve numerical stability.

Another issue in Bayesian analysis is the choice of prior, i.e. the function $\pi(\boldsymbol{\theta})$ in (1.29). In the analyses that follow, the parameter vector is written $(\mu, \log\psi, \xi)$ and is assumed uniform over $(-\infty,\infty) \times (-\infty,\infty) \times (-1,1)$, i.e. it is "improper" (meaning the integral of the prior density is not defined) over an infinite region. The range of ξ is restricted to $(-1,1)$ because the log likelihood becomes numerically unstable outside that range. However there is no reason to suppose that this kind of choice is optimal — it is made here largely for convenience and simplicity. Another common choice is the Jeffreys prior, which assumes $\pi(\boldsymbol{\theta}) \propto |\mathscr{I}|^{-1/2}$ — here $\mathscr{I}$ is the Fisher information matrix and $|\cdot|$ denotes the determinant. This choice was advocated in a famous book by Jeffreys [126] but also has some disadvantages — one specific

issue for the GEV distribution is that since $\mathscr{I}$ is only defined on $\xi \in (-\frac{1}{2}, \frac{1}{2})$, the parameters must be confined to that range, which may be problematic under some circumstances. Another idea is the class of *reference priors*, originally introduced by Bernardo [18], which have been applied to the GEV distribution in a striking recent paper by Zhang and Shaby [268]. However their proposals also begin with the Fisher information matrix, so again the analysis must be confined to $\xi \in (-\frac{1}{2}, \frac{1}{2})$. A more pragmatic solution is to keep the prior distributions of μ and $\log \psi$ uniform over (∞, ∞), but to allow ξ to follow a beta prior density over the range $(-\frac{1}{2}, \frac{1}{2})$ or $(-1, 1)$. This idea, first proposed by Martins and Stedinger [154], does not have any mathematically optimal properties that we are aware of, but it provides a family of practical alternatives that could be used in a sensitivity analysis (or, alternatively, to speed up the convergence of the algorithm in cases where this may seem to be problematic). The following analysis only used the uniform prior but any of these alternatives could be tried.

1.3.1.3 Example: Jenkinson's data from Hartford, Connecticut

These methods were applied to Jenkinson's data [128] given in Table 1.2. Key points are the function `lh1` to evaluate the negative log likelihood, the R optimization routines `optim` and `nlm`, and the `AdaptMH` algorithm for Bayesian analysis, as just described. The R code is available at
`https://rls.sites.oasis.unc.edu/s834-2023/Data/HartfordGEV.txt`.

The negative log likelihood was minimized numerically and the standard errors computed by inverting the observed information matrix. For each parameter, the z-statistic is computed as the parameter estimate divided by the standard error, and the p-value represents the two-sided tail probability associated with the z-statistic, under the assumption of a standard normal distribution. These results are given in Table 1.3. Note that $\log \hat{\psi} = 1.2467$ implies that $\hat{\psi} = e^{1.2467} = 3.478$ with a standard error

Parameter	Estimate	S.E.	z-statistic	p-value
μ	19.6809	0.3967	49.6083	0
$\log \psi$	1.2467	0.0786	15.8514	0
ξ	-0.2575	0.0598	-4.3033	1.7×10^{-5}

Table 1.3 *Parameter estimates for GEV model for Hartford data*

$1.2467 \times 0.0786 = 0.2734$ by the delta method. Jenkinson [128] quoted estimates of $\hat{\mu} = 19.7,\ \hat{\psi} = 3.46,\ \hat{\xi} = -0.26$, in the present notation.

The p-values values may be interpreted as tests that the corresponding parameters are 0. In the case of μ and $\log \psi$, such a null hypothesis does not really make sense (river flows are certainly not centered at 0; setting $\log \psi = 0$, which corresponds to $\psi = 1$, is a plausible value but there is no reason to suppose that this is the true value of ψ). Therefore, the fact that we get a p-value of 0 in these two cases is neither surprising nor of any importance. The case of ξ is different: in this case $\xi = 0$ would be a plausible result, corresponding to the Gumbel distribution which was often used as a distribution for annual maxima until the use of the GEV distribution became

much more prevalent. Therefore, it is of interest and importance both that $\hat{\xi} < 0$ (implying a finite upper endpoint to the distribution) and that the result is statistically significant with a p-value of about 10^{-5}. The practical conclusion is that we can be highly confident that the distribution does have a finite endpoint.

For the endpoint of the distribution and RV_T for given T, we calculate an estimate by substitution and the standard error by the delta method as outlined in Section 1.3.1.1, which leads us to:

- The estimated endpoint is 33.193 with a standard error of 2.597.
- The estimated RV_{1000} is 30.910 with a standard error of 1.304.
- The estimated RV_{100} is 29.069 with a standard error of 0.832.

Jenkinson quoted 33.2, 30.9, 29.0 for the three estimates and 2.39, 1.21, 1.09 respectively for their standard errors. The point estimates agree very well with ours. The standard errors are a little different, but it should be noted that Jenkinson based his standard error estimates on the Fisher information whereas we have used observed information, which is generally considered to be a little more accurate. Beyond that, however, most analyses (including this one) that use the observed information matrix do not calculate it directly, but rely on the "hessian" output from an optimization routine; this is not a precise calculation, and some variation can be expected depending on which optimization routine is used and on the parameters that are used to initialize the routine.

If we calculate 95% confidence intervals by the standard formula of estimate plus or minus 1.96 times the standard error, we get (28.1,38.3) for the endpoint of the distribution, (28.4,33.5) for RV_{1000} (27.4,30.7) for RV_{100}. However, these results rely on the standard errors being a good approximation to the true standard deviations of the estimates (which we have just seen might be questionable) and on the distribution of the MLE being normal. These are all asymptotic results, and many theoretical and empirical studies have shown that the convergence of asymptotic results in the case of the GEV distribution is slow, not to mention that the GEV distribution itself may not be an accurate approximation to the annual maxima (more about this issue in Chapter 2).

The motivation for considering Bayesian analyses in this situation is that, while there is no guarantee that Bayesian analyses will perform better than MLEs by standard metrics such as the mean squared error of point estimates or the coverage probability of interval estimates, they do provide an alternative viewpoint that allows for the skewness of the distribution and the true shape of the log likelihood function. Asymptotic theory of MLEs relies on the quadratic shape of the log likelihood function near its maximum, whereas Bayesian methods integrate the likelihood function itself rather than a quadratic approximation to it.

With these considerations in mind, we conducted a Bayesian analysis of Jenkinson's data using the numerical techniques of Section 1.3.1.2. The posterior densities for the endpoint and for RV_{1000} and RV_{100} are shown in Figure 1.7. These show a clear right-skewness, contradicting the asymptotically normal shape which is implicit in the standard MLE calculations. We can also compute 95% credible intervals from the output of the adaptive Metropolis output (these correspond to the 2.5% and

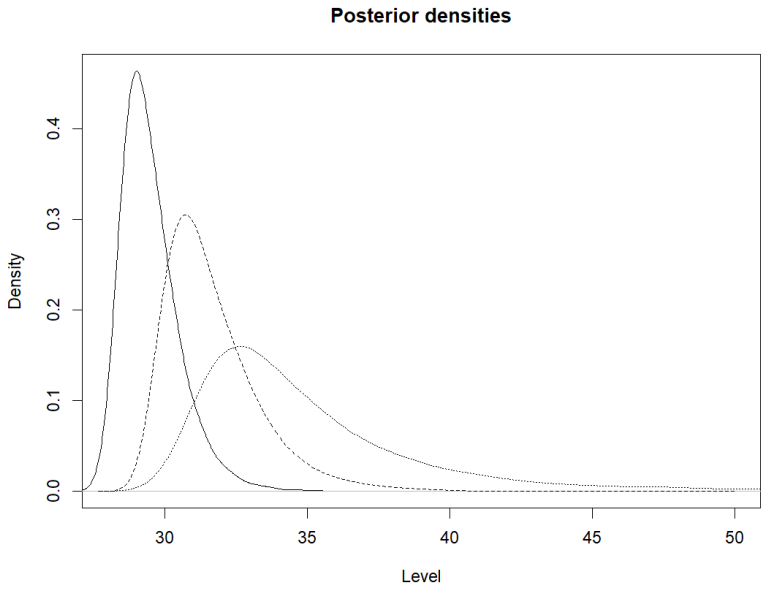

Figure 1.7 *Posterior densities for Hartford data. Solid curve: 100-year return value. Dashed curve: 1000-year return value. Dotted curve: endpoint of distribution* $\mu + \psi/\xi$.

97.5% quantiles of the empirical posterior distribution), as follows: (30.8, 48.6) for the endpoint of the distribution, (29.6, 33.0) for RV_{1000} and (28.1, 32.0) for RV_{100}.

All three intervals are skewed to the right compared with the corresponding ML-based confidence intervals, and especially, the right hand end of the interval is substantially greater than for the confidence interval in each case. This information could be important to an engineer trying to build a defensive system while allowing for uncertainty (and, just to emphasize a point that will be made in more detail later, none of this allows for the possibility of long-term trend in the distribution, such as we may well get as a result of climate change),.

The practical conclusion from all this is that we can use the method of maximum likelihood to fit the GEV distribution, and (as the detailed code shows) both the `optim` and `nlm` methods in R lead to the same estimates (which does not conclusively prove that the estmates are correct, but it is strong supportive evidence) but that the Bayesian analyses leads to alternative interval estimates for return values which bring out the right-skewed nature of the distribution. This does not prove that the Bayesian estimates are more accurate, still less that they give the correct "real world" values, but they do bring out how much uncertainty is involved in these estimates. This is part of the reason why alternative approaches to the whole problem, such as a threshold approach, have become more popular in recent decades.

1.3.2 GPD and Poisson-GPD Models

Suppose we have n observations $Y_1,\ldots,Y_n$ from the GPD model, $G(y\ ;\ \sigma,\xi) = 1-\left(1+\xi\frac{y}{\sigma}\right)_+^{-1/\xi}$. The density is $g(y\ ;\ \sigma,\xi) = \frac{dG(y\ ;\ \sigma,\xi)}{dy} = \frac{1}{\sigma}\left(1+\xi\frac{y}{\sigma}\right)_+^{-1/\xi-1}$ and hence the NLLH is given by

$$\ell_n(\sigma,\xi \mid Y_1,\ldots,Y_n) \quad = \quad n\log\sigma + \left(\frac{1}{\xi}+1\right)\sum\log\left(1+\xi\frac{Y_i}{\sigma}\right) \tag{1.30}$$

defined when $1+\xi\frac{Y_i}{\sigma} > 0$ for all i.

The previous calculations, in which we minimize ℓ by numerical optimization, or equivalently, solve the likelihood equations $\frac{\partial\ell}{\partial\sigma} = \frac{\partial\ell}{\partial\sigma} = 0$, remain valid for this model as well. Once again, we can define the Fisher information matrix to be $\frac{1}{n}$ times the matrix of expected second-order partial derivatives

$$\begin{pmatrix} \frac{\partial^2\ell_n}{\partial\sigma^2} & \frac{\partial^2\ell_n}{\partial\sigma\partial\xi} \\ \frac{\partial^2\ell_n}{\partial\sigma\partial\xi} & \frac{\partial^2\ell_n}{\partial\xi^2} \end{pmatrix}$$

which was calculated explicitly in the appendix of [224] as

$$\mathscr{I}_1 \quad = \quad \begin{pmatrix} \frac{1}{\sigma^2(1+2\xi)} & \frac{1}{\sigma(1+2\xi)(1+\xi)} \\ \frac{1}{\sigma(1+2\xi)(1+\xi)} & \frac{1}{(1+2\xi)(1+\xi)} \end{pmatrix} \tag{1.31}$$

provided $\xi > -\frac{1}{2}$; [224] also went into details about what happens when $\xi \le -\frac{1}{2}$. Again as with the GEV case, provided $\xi > -\frac{1}{2}$, the asymptotic covariance matrix of the MLEs is given by $n^{-1}\mathscr{I}_1^{-1}$ where an explicit expression for $\mathscr{I}^{-1}$ is (see [51])

$$\mathscr{I}_1^{-1} \quad = \quad (1+\xi)\begin{pmatrix} 2\sigma^2 & -\sigma \\ -\sigma & 1+\xi \end{pmatrix}. \tag{1.32}$$

(These formulas incorporate a change of notation from the original papers which, following what was originally Jenkinson's notation, wrote k in place of our $-\xi$.)

[51] noted that the minimization of (1.30) can be simplified, as follows: if we write $\sigma = \frac{\xi}{\tau}$, (1.30) may be rewritten as

$$\ell_n(\tau,\xi \mid Y_1,\ldots,Y_n) \quad = \quad n\log\xi - n\log\tau + \left(\frac{1}{\xi}+1\right)\sum\log(1+\tau Y_i) \tag{1.33}$$

whose minimum with respect to ξ, for fixed τ, can be calculated directly as

$$\hat{\xi}_\tau \quad = \quad \frac{1}{n}\sum_{i=1}^{n}\log(1+\tau Y_i) \tag{1.34}$$

again provided $1+\tau Y_i > 0$ for all i. Substituting (1.34) into (1.33) reduces the problem to a one-parameter optimization, for which simpler numerical algorithms exist.

However, this simplification of the MLE does not typically apply when the model is extended to include covariates, so the result is of limited utility.

The Poisson-GPD model is only a small extension of this. Suppose the number of exceedances in a given year is N, with a distribution which is assume Poisson with mean λ, and the exceedances $Y_1, \ldots, Y_N$ are independent GPD(σ, ξ) given N. The likelihood function for this model is

$$\frac{\lambda^N e^{-\lambda}}{N!} \cdot \prod_{i=1}^{N} g(Y_i \,;\, \sigma, \xi)$$

which immediately factorizes into the likelihood function for N (with MLE $\hat{\lambda} = N$) and the same likelihood as just discussed for the GPD component. These models become more complicated when covariates are introduced [51].

1.3.3 The r-largest order statistics Model

The asymptotic distribution of the r largest order statistics from a sample was already given in (1.11) and (1.15). The idea of using this model for estimation was first recognized in the 1970s; the paper by Weissman [257] was particulary influential in showing that this approach could lead to estimation and testing procedures distinct from those using the classical three types of extreme value limit distributions. Since then, the idea has been used for many more general models in extreme value theory. As we shall see, there is a close connection with threshold exceedances in the GPD, but the two approaches are sufficiently distinct to warrant separate treatment.

There are two ways to think about this problem. One is to treat the entire dataset as a single sample and pick out the r largest order statistics from that. The other is to subdivide the data into blocks, such as blocks of one year, and pick out the r largest order statistics in each block. For example, our previous example of women's track times is of this form, with $r = 5$ or 10 representing the number of extreme performances each year that are included in the sample. The two approaches require somewhat different approaches to estimation, so we treat them separately.

1.3.3.1 Single Sample

Consider the case of a single sample of r largest order statistics $Y_1 \geq Y_2 \geq \ldots \geq Y_r$ when the parent distribution is Gumbel, or the $\xi = 0$ subcase of the GEV distribution. The joint distribution (1.12) may then be treated as a likelihood function, and maximized to find estimators of the parameters μ and ψ. Writing the negative log likelihood as

$$\ell(\mu, \psi \mid Y_1, \ldots, Y_r) \quad = \quad r \log \psi + \sum_{i=1}^{r} \frac{Y_i - \mu}{\psi} + \exp\left(-\frac{Y_r - \mu}{\psi}\right). \qquad (1.35)$$

Setting $\frac{\partial \ell}{\partial \mu} = \frac{\partial \ell}{\partial \psi} = 0$ leads quickly to the estimators

$$\begin{aligned} \hat{\psi} &= \frac{1}{r} \sum_{i=1}^{r} (Y_i - Y_r), \\ \hat{\mu} &= Y_r + \hat{\psi} \log r, \end{aligned} \qquad (1.36)$$

a pleasing (and rather rare) case in extreme value theory where it is possible to obtain closed-form expressions for the maximum likelihood estimators. This result is due to Weissman [257], who also used the joint distribution (1.12) to investigate the statistical properties of the estimators.

Another case of a rather similar problem (in fact a simple transformation of the one just given) arises when the appropriate extreme value limit is of "two-parameter Fréchet" form, i.e. $H(x) = \exp\{-(x/\sigma)^{-\alpha}$ for $x > 0,\ \sigma > 0,\ \alpha > 0$. (We could also introduce a location parameter into this problem, i.e. replace x by $x - \mu$ for some unknown μ, but this has the same tail behavior and it is most common in practice just to assme $\mu = 0$.) In this case, the joint density formula corresponding to (1.11) is

$$h(y_1,\ldots,y_r) \quad = \quad \sigma^{-r}\prod_{i=1}^{r}\left\{\alpha\left(\frac{y_i}{\sigma}\right)^{-\alpha-1}\right\}\exp\left\{-\left(\frac{y_r}{\sigma}\right)^{-\alpha}\right\} \tag{1.37}$$

defined on the order statistics $y_1 \geq \ldots \geq y_r > 0$. Replacing the y_i's by sample values Y_i's, we define the negative log likelihood in this case to be

$$\ell(\sigma,\alpha|\ Y_1,\ldots,Y_r) = r\log\sigma - r\log\alpha + (\alpha+1)\sum_{i=1}^{r}\log\left(\frac{Y_i}{\sigma}\right) + \left(\frac{Y_r}{\sigma}\right)^{-\alpha}. \tag{1.38}$$

Minimization of (1.38) with respect to σ and α leads to the closed-form estimators

$$\begin{aligned} \hat{\alpha} \quad &= \quad \left\{\frac{1}{r}\sum_{i=1}^{r}(\log Y_i - \log Y_r)\right\}^{-1}, \\ \hat{\sigma} \quad &= \quad r^{1/\hat{\alpha}}Y_r \end{aligned} \tag{1.39}$$

which is in fact very similar to (1.36) (replace Y_i in (1.36) by $\log Y_i$ in (1.39).

There is even a third formulation that does not directly use extreme value theory at all, but assumes we have an "exact Pareto tail" above some threshold u: the full sample $X_1,\ldots,X_n$ are IID from a distribution function F that satisfies

$$F(x) \quad = \quad 1 - cx^{-\alpha},\ x \geq u, \tag{1.40}$$

where $c > 0,\ \alpha > 0$, and u is a known threshold, with $F(x)$ undefined for $x < u$. In this case, the natural formulation of the likelihood function is to assume order statistics $Y_1 \geq \ldots \geq Y_r > u$ in $[u,\infty)$, treating the observations below u as censored. In that case the joint density is

$$\prod_{i=1}^{r}\left(\alpha c Y_i^{-\alpha-1}\right)\left(1 - cu^{-\alpha}\right)^{n-r}.$$

The corresponding negative log likelihood in this case is

$$\ell(c,\alpha\mid Y_1,\ldots,Y_r,u) = -r\log\alpha - r\log c + (\alpha+1)\sum_{i=1}^{r}\log Y_i - (n-r)\log(1-cu^{-\alpha}). \tag{1.41}$$

Setting $\frac{d\ell}{dc} = \frac{d\ell}{d\alpha} = 0$ in this case produced the closed-form estimators

$$\begin{aligned} \hat{\alpha} &= \left\{ \frac{1}{r} \sum_{i=1}^{r} (\log Y_i - \log u) \right\}^{-1}, \\ \hat{c} &= \frac{r}{n} u^{\hat{\alpha}}. \end{aligned} \tag{1.42}$$

The estimators (1.42) were first derived by Hill [115] and in particular $\hat{\alpha}$ is often called Hill's estimator, but the close resemblence between $\hat{\alpha}$ in (1.42) and (1.39) (in the one case, conditioning on the threshold u, and in the other case, on Y_r) shows that they are really the same estimator, so henceforth we shall refer to either estimator as the *Hill-Weissman estimator*.

The corresponding case of a GEV distribution with all three parameters unknown does not lead to closed form estimators, but it is still possible to treat (1.11) as a likelihood function and find maximum likelihood estimators $\hat{\mu}$, $\hat{\psi}$, ξ numerically. The special case where this procedure is used to estimate the endpoint of the distribution was analyzed in detail by Smith and Weissman [233].

1.3.3.2 Multiple Samples

The "multiple samples" version of this analysis applies when the data are divided into blocks and we are taking into account the r largest order statistics within each block. In environmental and some other examples (including our track records dataset) the "block" is usually equated with one year of data. For convenience, in the following discussion we shall refer to the blocks as year, but the theory of course does not require that each block is exactly one year in length. This model is of greatest interest when there are additional covariates, so that the distributions are not the same in all blocks, but we first treat the simplest case where that possibility is ignored.

Suppose, then, we have observations in year t ordered as $Y_{t,1} \geq \ldots \geq Y_{t,r}$ and assume the distribution for annual maxima in year t is GEV with parameters μ, ψ, ξ. Assume the year index t ranges from 1 to T. Applying (1.11) to each of the T years of data, the negative log likelihood based on $\mathbf{Y} = \{Y_{t,i},\ 1 \leq i \leq r,\ 1 \leq t \leq T\}$ is

$$\ell(\mu, \psi, \xi; \mathbf{Y}) = Tr \log \psi + \left(\frac{1}{\xi} + 1 \right) \sum_{t=1}^{T} \sum_{i=1}^{r} \log \left(1 + \xi \frac{Y_{t,i} - \mu}{\psi} \right) + \sum_{t=1}^{T} \left(1 + \xi \frac{Y_{t,r} - \mu}{\psi} \right)^{-1/\xi}, \tag{1.43}$$

assuming $1 + \xi \frac{Y_{t,i} - \mu}{\psi} > 0$ for each t, i.

The extension of (1.43) for data with a trend is straightforward: suppose the (μ, ψ, ξ) parameters for year t are replaced by (μ_t, ψ_t, ξ_t) where each of μ_t, ψ_t, ξ_t is written as a function of parameters $\boldsymbol{\theta}$. We also allow for the possibility that r may vary from year to year (in some datasets, the number of available order statistics may not be the same every year) so we modify (1.43) to

$$\ell(\boldsymbol{\theta}; \mathbf{Y}) = \sum_{t=1}^{T} r_t \log \psi_t + \sum_{t=1}^{T} \left(\frac{1}{\xi_t} + 1 \right) \sum_{i=1}^{r_t} \log \left(1 + \xi_t \frac{Y_{t,i} - \mu_t}{\psi_t} \right) + \sum_{t=1}^{T} \left(1 + \xi_t \frac{Y_{t,r_t} - \mu_t}{\psi_t} \right)^{-1/\xi_t}, \tag{1.44}$$

assuming $1+\xi_t \frac{Y_{t,i}-\mu_t}{\psi_t} > 0$ for each t, i.

Estimation then proceeds using numerical optimization to find the minimum of (1.44) with respect to parameters $\boldsymbol{\theta}$. The calculation of standard errors, return values, and other statistics of interest, then follow by the same procedures as analyzed for the GEV case.

Example. Smith [219] studied a dataset consisting of the ten largest sea levels (in cm.) for each year in Venice, from 1931–1981 (except for 1935, when only six values were available). The dataset originated with Pirazzoli [184], and may serve as a reminder that concerns about rising sea levels long predated present-day concerns about global climate change. The dataset and accompanying R code may be downloaded from https://rls.sites.oasis.unc.edu/s834-2023/s834data.html.

We fit the model (1.44) with $r_t = 10$ except for 1935 when $r_t = 6$. We take $\psi_t = \psi$ and $\xi_t = \xi$ (independent of t) and cosnider four models for μ_t:

- Model 1: no trend, $\mu_t = \mu$ for all t;
- Model 2: linear trend, $\mu_t = \beta_0 + \frac{t}{10}\beta_1$ for all t;
- Model 3: quadratic trend, $\mu_t = \beta_0 + \frac{t}{10}\beta_1 + \left(\frac{t}{10}\right)^2 \beta_2$ for all t;
- Model 4: linear plus periodic trend, $\mu_t = \beta_0 + \frac{t}{10}\beta_1 + \beta_2 \cos\left(\frac{2\pi t}{P}\right) + \beta_3 \sin\left(\frac{2\pi t}{P}\right)$ for all t.

In Model 3, P is a specified period; [184] assumed $P = 18.62$ (years) to correspond to an astronomical tidal cycle but [219] found a much better fit based on $P = 11$, which is the period of the sunspot cycle. We try both models here.

As an example, Table 1.4 gives the detailed estimates for Model 2. This shows in particular that the trend (4.7453 cm. per decade, with a standard error of 0.3304) is very highly significant — the z-statistic is quoted as 14.36, which corresponds to a p-value of about 10^{-46}, definitive evidence of an increasing trend. Another feature is the value of ξ, here estimated to be -0.0678 with a standard error about 0.027 and p-value of 0.013. So $\hat{\xi}$ is quite close to 0 but still statistically significant.

Parameter	Estimate	S.E.	z-statistic	p-value
β_0	104.7639	1.2737	82.2492	0.0000
$\log\psi$	2.4678	0.0540	45.7179	0.0000
ξ	-0.0678	0.0274	-2.4748	0.0133
β_1	4.7453	0.3304	14.3612	0.0000

Table 1.4 *Parameter estimates for linear trend model for Venice sea level data*

If we compare the optimized negative log likelihood (NLLH) values for each of the models, we get the results in Table 1.5. Here, DF (degrees of freedom) correspond to the number of parameters in the model. The standard procedure for comparing nested models is via a likelihood ratio test (LRT). First, we compute the difference of log likelihod and multiply by 2: this quantity is known as the deviance. Then, we compute the p-value based on a chi-square distribution, where the DF of the chi-square test is the difference of the DF for the two models. Thus, in this case:

Model	DF	NLLH
Model 1	3	1142.8
Model 2	4	1090.2
Model 3	5	1090.0
Model 4 ($P = 18.62$)	6	1087.6
Model 4 ($P = 11$)	6	1079.7

Table 1.5 *Comparisons of NLLH for five models fitted*

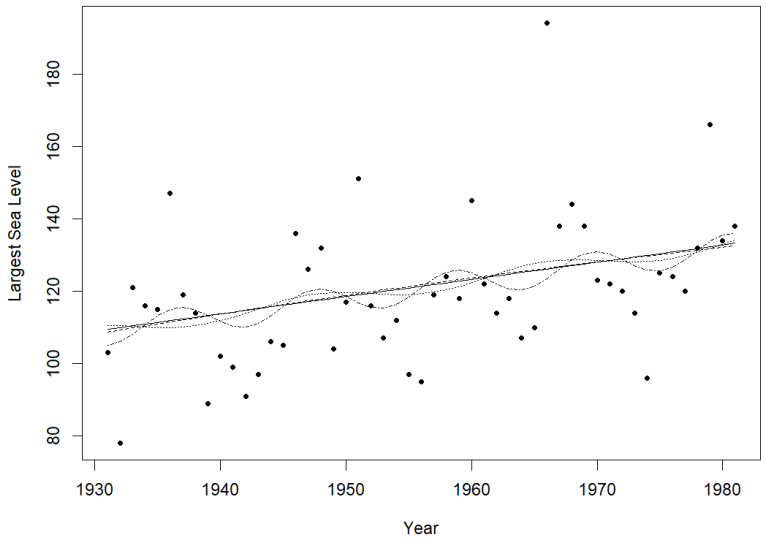

Figure 1.8 *Venice annual maxima 1931–1981 and median predicted values under four models.*

- Model 1 v. Model 2: deviance 105.2, DF=1, p-value 10^{-24}
- Model 2 v. Model 3: deviance 0.4, DF=1, p-value 0.82
- Model 2 v. Model 4 ($P = 18.62$): deviance 5.2, DF=2, p-value 0.07
- Model 2 v. Model 4 ($P = 11$): deviance 21, DF=2, p-value 3×10^{-5}

From this, it can be seen that Model 2 clearly improved on Model 1, so we should include the trend. The evidence of a periodic effect is less clear-cut, but Model 4 with $P = 11$ does show a very significant periodic trend in addition to the linear trend.

For a visual comparison, Figure 1.8 shows the annual maximum for each year, together with the median predicted value of the annual maximum under each of our four main models (we don't plot Model 1). The fits for Model 1 and Model 2 are almost the same, but Model 3 wth $P = 11$ shows a clear cyclic effect.

The original paper of [219] considered only the $\xi \to 0$ limiting case, the model based on the Gumbel distribution for annual maxima. In all other respects, the anal-

ysis almost exactly follows that here; Figure 1.8 is very similar to Figure 2 in [219]. The latter paper also discussed some other model fitting aspects including the sensitivity to the choice of r and probability plots to verify the fit of the model.

Tawn [239] extended the Gumbel model of [219] to the GEV case and also gave a theoretical expression for the Fisher information matrix in this model, thus extending the GEV calculation of [185] (i.e. from $r = 1$ to any fixed $r > 1$).

1.3.4 Point Process Approach

Suppose, for A as in 1.6, there are $N(A)$ points in A at locations $((t_1,x_1),\ldots,(t_{N(A)},x_{N(A)}))$. By (1.14), the joint density is given by

$$\prod_{i=1}^{N(A)} \left\{ \frac{1}{\psi}\left(1+\xi\frac{x_i-\mu}{\psi}\right)^{-1/\xi-1} \right\} \cdot \left(1+\xi\frac{y-\mu}{\psi}\right)_+^{-1/\xi} \tag{1.45}$$

provided $1+\xi\frac{x_i-\mu}{\psi} > 0$ for $i = 1,\ldots,N(A)$.

1.4 Analysis of Data in Kelowna

Consider the Kelowna data from Fig. 1.1. We are trying to characterize the distribution of the annual maximum temperatures with a view towards assessing how extreme was the 2021 event. In performing such an analysis, it makes sense to exclude the 2021 event itself — we are trying to assess how plausible the 2021 event was from the point of view of never having seen such an extreme event before. With this objective, it makes sense to fit a distribution to the data from 1984 to 2020 (37 years). Based on our earlier discussion of extreme value distributions, a natural model would appear to be the GEV distribution (1.6), where the parameters μ,ψ,ξ are unknown and estimated by maximum likelihood.

An initial maximum likelihood fit produces the estimates in Table 1.6.

Parameter	Estimate	S.E.	t-value	p-value
μ	35.7119	0.3098	115.2624	0
$\log\psi$	0.5477	0.1305	4.1962	2.7×10^{-5}
ξ	–0.4203	0.1007	–4.1718	3.0×10^{-5}

Table 1.6 *Fitting the initial GEV model to the Kelowna data, 1984–2021.*

The second column gives the maximum likelihood estimates; the third column the standard errors; the fourth column the t values (estimate divided by its standard error) and the fifth column the p-value, assuming a normal distribution for the estimate itself (which is certainly not an exact result, but good enough for the present discussion). The p-values for μ and $\log\psi$ are not especially meaningful since there is no a priori reason to expect either of these parameters to be 0, but the p-value for ξ is of interest because the case $\xi < 0$ is the "short-tailed" case while $\xi > 0$ would be "long-tailed". In this case we see $\hat{\xi} < 0$ and the associated p-value shows

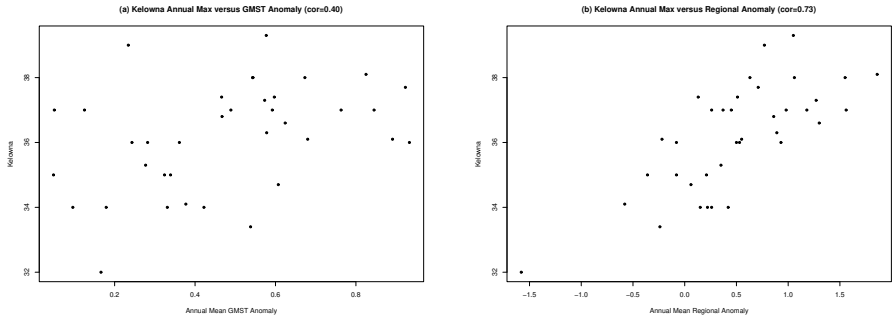

Figure 1.9 *Correlation of annual maximum temperatures in Kelowna with (a) anomalies from global mean summer temperature, (b) anomalies from a regional mean summer temperature based on 40–55°N latitude, 110–125°W longitude.*

a clear-cut rejection of the null hypothesis $\xi = 0$ (or $\xi \geq 0$). The distribution therefore has a finite upper endpoint, and the MLE fit leads to the estimated endpoint $\hat{\mu} - \frac{\psi}{\xi} = 35.7119 + \frac{e^{0.5477}}{0.4203} \approx 39.83$. This is less than the observed 2021 value of 44.6°C. Although the analysis given here is only for one station and not necessarily representative for the entire region, it is results of this nature than have given rise to statements suggesting that this heatwave was "virtually impossible without human-caused climate change" [178].

A more extensive version of the analysis would be to consider possible covariates. We know that global temperatures are rising; as a first step to considering the influence of climate change on extreme events, it is natural to consider how the fitted GEV distribution correlates with measure of global temperature rise. Indeed, this kind of analysis has been presented in [178] and similar references, though the version presented here is more limited because it only presents a single station (but the underlying themes are the same). Specifically, [178] considered global mean surface temperature (GMST) as a covariate. There are several publicly accessible datasets for global temperatures; the analysis considered here is from the Climate Research Unit of the University of East Anglia (known as the HadCRUT5 dataset). As with most datasets of this nature, the observations are represented as *anomalies*, which essentially means that stations included in the analysis are first standardized to the means from 1961–1990 before being aggregated spatially (the HadCRUT5 website gives detailed information about how the dataset was constructed and the reason for computing anomalies). Fig. 1.9(a) shows the Kelowna annual maxima plotted against the GMST summer (average of June, July, August monthly mean) anomalies. The correlation of 0.4 suggests that it would be worthwhile developing a more detailed model for the association between the two series.

One possible approach to this problem is to extend the model (1.6) to one of the

form

$$\Pr\{Y_t \leq y\} \quad = \quad \exp\left\{-\left(1+\xi_t\frac{x-\mu_t}{\psi_t}\right)_+^{-1/\xi_t}\right\} \tag{1.46}$$

where Y_t represents annual maximum temperature in year t and $\mu_t,\ \psi_t,\ \xi_t$ are the GEV parameters in year t. (We are not writing $\xi_t = 0$ as a separate case here, but if it were the case, it would reduce to the Gumbel distribution as in (1.6)). We write the model in this way to clarify the point that any combination of the three GEV parameters may be time dependent, but here, we use the simplest form of this model, with

$$\begin{aligned} \mu_t &= \beta_0 + \beta_1 x_t, \\ \log\psi_t &= \log\psi, \\ \xi_t &= \xi, \end{aligned} \tag{1.47}$$

i.e. μ_t depends linearly on the covariate x_t while ψ_t and ξ_t are independent of t. We represent this model in terms of $\log\psi_t$, rather than simply ψ_t, to ensure we fulfil the constraint $\psi_t > 0$, and this would be even more relevant if we had covariates in ψ_t as well (not included here, but will be in later analyses).

We therefore fit the model (1.47) with x_t the GMST anomaly, with the results in Table 1.7. It can be seen that the parameter β_1 is indeed statistically significant, with a p-value of 0.02, which confirms that global temperature means are associated with extreme temperatures in Kelowna (whether the association is strictly "causal" is a point we shall not address here), but the result is possibly not as convincing as we would like.

In this case we can estimate the maximum possible temperature for 2021 as $\hat\beta_0 + \hat\beta_1 x^* - \hat\psi/\hat\xi$ where x^* is the observed GMST anomaly for 2021 (0.762). This however leads to an estimate of 41.6°C, still well below the observed value of 44.6°C.

Parameter	Estimate	S.E.	t-value	p-value
β_0	34.3983	0.6427	53.5175	0.0000
$\log\psi$	0.4088	0.1292	3.1632	0.0016
ξ	–0.2879	0.1101	–2.6139	0.0090
β_1	2.6448	1.1503	2.2992	0.0215

Table 1.7 *Fitting the initial GEV model to the Kelowna data, 1984–2021.*

A second analysis of this form is illustrated by Fig. 1.9(b) and Table 1.8. Instead of GMST as the covariate x_t, we have used a regional mean, defined the same way as for the GMST but based on the region between latitudes 40–55°N and longitudes 110–125°W (the HadCRUT5 date are conveniently aggregated into $5^\circ \times 5^\circ$ grid cells so that this type of series is easy to calculate). Fig. 1.9(b) shows a much stronger correlation between this series and the Kelowna annual maxima, in this case, the sample correlation coefficient is 0.73. Table 1.8 also shows a much more convincing

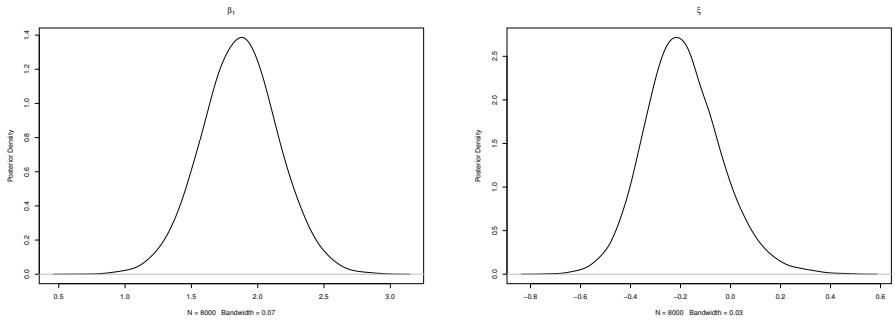

Figure 1.10 *Posterior Densities of β_1 and ξ, Model (1.47) with x_t the Regional Mean.*

relationship in the parameter β_1 (a two-sided p-value of 4×10^{-12} against the null hypothesis that this parameter is zero). This gives us some confidence that the association between these two variables, when combined with forwards or backwards projections of the regional mean, can be used to characterize long-term changes in the Kelowna anual maxima — a theme that will be developed elsewhere. In this case, the estimated "maximum possible temperature" for 2021 is 44.5°C, just below the observed 44.6°C. This takes us closer to being able to "explain" the 2021 event, but still does not give us a quantitative probability for the event itself.

Parameter	Estimate	S.E.	t-value	p-value
β_0	34.8340	0.2385	146.0587	0.0000
$\log \psi$	0.0381	0.1385	0.2751	0.7832
ξ	–0.2140	0.1463	–1.4381	0.1504
β_1	1.8431	0.2660	6.9277	4×10^{-12}

Table 1.8 *Fitting the initial GEV model to the Kelowna data, 1984–2021.*

An alternative approach to this whole analysis is Bayesian. With a prior distribution that is essentially uniform for $(\beta_0, \log \psi, \xi, \beta_1)$ over a large subset of the parameter space, we run a Markov chain Monte Carlo (MCMC) algorithm using Haario's [106] "adaptive Metropolis" procedure. In this context, the advantage of a Bayesian approach over the maximum likelihood method is better representation of the uncertainty of the estimates, especially with regard to predictive probabilities. As an illustration, Fig. 1.11 shows the posterior densities of β_1 and ξ. In the case of β_1, nearly all the posterior density lies between 1 and 3, confirming the conclusion from Table 1.8 that it is overwhelmingly likely that $\beta_1 > 0$. As for ξ, approximately 11% of the posterior density lies to the right of 0, confirming that, although the evidence still points towards $\xi < 0$, this is by no means certain.

These calculations do not directly address the probability of exceeding 44.6^oC, the temperature that was actually observed in 2021. Let P be the probability of exceeding 44.6^oC in 2021, based on Kelowna data up to 2020 and the regional mean

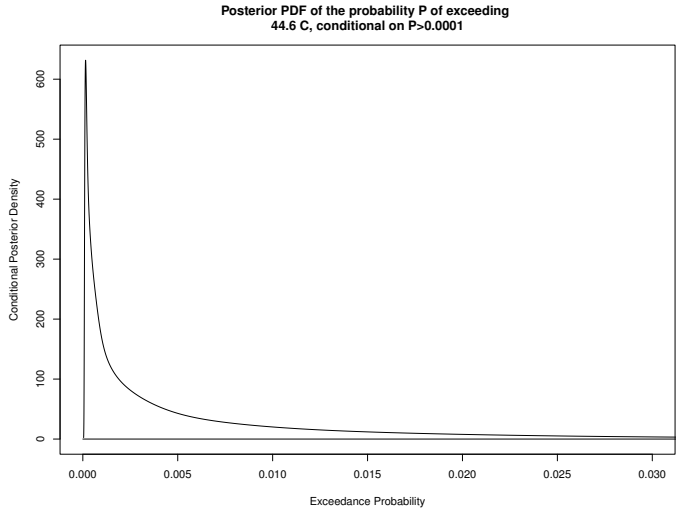

Figure 1.11 *Conditional Posterior Densities of Exceeding 44.6°C in 2021, Based on Regional Mean Summer Temperature.*

temperature for 2021. We may write $P = 1 - \exp\left[-\left(1+\xi\frac{y^*-\beta_0-\beta_1 x^*}{\psi}\right)_+^{-1/\xi}\right]$ where y^* and x^* are respectively the observed values for 2021 of the Kelowna temperature (44.6) and the regional mean. This P is a nonlinear function of the model parameters $(\beta_0, \log\psi, \xi, \beta_1)$ so its posterior density may be estimated from the MCMC run already computed. We have computed the posterior probability that $P > 0$ to be 0.56, and the posterior mean of P to be 0.0035. (Because of the randomness of the MCMC algorithm, these numbers may differ slightly from one run to another.) To summarize the interpretation of these numbers:

1. The event $P > 0$ is equivalent to the statement that the right-hand end of the distribution is greater than the observed 44.6 — in particular, this includes that part of the posterior distribution where $\xi \geq 0$. So, at least is this case where we are conditioning on the regional mean, the probability of this event is quite substantial. However, the probability that we actually observe such an extreme event is more relevant to the interpretation of extreme weather events than the event being within theoretical bounds, so the next calculation is more meaningful:
2. The posterior mean of P is arguably our "best guess" for the actual probability of interest. As is known from decision theory, the posterior mean minimizes the Bayes risk under squared error and a number of other loss functions. There are other Bayesian outputs (in particular, credible intervals) that capture the uncertainty of Bayesian estimates, but for problems of this nature, the lower bound of a Bayesian credible interval is nearly always 0 (in particular, if the posterior probability that $P = 0$ is > 0.025, as it is here, then a two-sided 95% credible interval will necessarily have lower bound 0).

3. The posterior mean of P being about 0.0035 corresponds very roughly to a return period of about 300 years, conditional on the regional mean. In contrast, using different data and statistical methods (but based on the same heatwave event), Philip et al. [178] commented, "In the most realistic statistical analysis the event is estimated to be about a 1 in 1000 year event in today's climate." The two estimates are based on different data and different assumptions about the analysis, but they are still of the same order of magnitude.
4. An alternative way to represent the uncertainty is through posterior density plots conditional on $P > \varepsilon$, where ε is a small positive number (we don't take $\varepsilon = 0$ because of the large spike in the posterior density as $P \to 0$). Fig. 1.11 shows such a plot with $\varepsilon = 0.0001$. Although there is still a large spike near 0, the plot shows a non-negligible posterior probability for values of P out to about 0.03 (in fact, the posterior probability that $P > 0.03$ is itself about 0.03, by no means negligible).

[Further material, possibly to be added later. In a Royal Statistical Society discussion paper that I only learned about after completing my initial calculations for Kelowna, Clarkson et al. [29] used a threshold exceedances approach including the GPD, and concluded that with this method, the problem of the estimated right-hand endpoint of the distribution being greater than the observed value no longer exists, supposedly because of the greater precision of GPD estimates because of them taking in more data. Again, this was for a different dataset than the one considered here, but based on the same heatwave event. I have not independently verified their conclusion and am somewhat skeptical that it is a general resolution of this issue, but it would be interesting to check this approach.]

1.5 Analysis of Insurance Data

1.6 Analysis of Women's Track Data

1.7 Software: the `extRemes` Package

1.8 Summary of Chapter

1.9 Exercises

1. Show that the three cases of (1.6), with $\xi = 0,\ \xi > 0,\ \xi < 0$, are equivalent to (1.3), (1.4) and (1.5), respectively, with suitable choices of $\mu,\ \psi$ and ξ.
2. Show that the argument leading to (1.10) also holds when $\xi = 0$, and show the equivalence of (1.10) and (1.6), i.e. exactly how are the parameters (λ, σ, ξ) from (1.10) related to the parameters (μ, ψ, ξ) from (1.6)?

   [Answer: ξ is the same while $\sigma = \psi + \xi(u-\mu), \lambda = \left(1 + \xi \frac{u-\mu}{\psi}\right)^{-1/\xi}$, assuming $\psi + \xi(u-\mu) > 0$.]
3. Prove the statements at the end of Section 1.2.4, concerning the equivalence of the point process approach to the three other approaches mentioned.

Chapter 2

Domains of Attraction, Rates of Convergence and Optimal Statistical Estimation

2.1 The Theory of Gnedenko and de Haan

In this chapter, we begin by exploring in more detail the implications of the limit relationship (1.2), which we restate here as

$$\Pr\left\{\frac{M_n - b_n}{a_n} \leq x\right\} \quad = \quad F^n(a_n x + b_n) \;\to\; H(x). \tag{2.1}$$

Here, $M_n = \max(X_1, \ldots, X_n)$, where $X_1, X_2, \ldots$ are IID random variables with common distribution function F, a_n and b_n are normalizing constants, and H is a non-degenerate limit. In Chapter 1, we asserted without detailed justification that the possible limits H are given by (1.3)–(1.5), or equivalently (1.6). Because (1.3)–(1.5) represent three different *types* of distributions (recall that two distribution functions are said to be of the same type if one can be obtained from the other by a location-scale transformation), this is known as the *Three Types Theorem*.

In this chapter, we explore in more detail the implications of these results. The original statement of the Three Types Theorem was given, without rigorous proof, by Fisher and Tippett [77] and, in the form (1.6), by von Mises [161]. A rigorous proof was first given by Gnedenko [87], who also derived necessary and sufficient conditions on F for a relation of the form (2.1) to be true. This is known as a *domain of attraction* problem: given a fixed H of the form (1.3)–(1.5) or (1.6), the domain of attraction of H is the class of all distribution functions F for which (2.1) holds for some $a_n > 0,\ b_n \in \mathbb{R}$.

The basic theory has been given in many previous books, for example [143, 194, 101], and our intention here is not to repeat detailed proofs that can be readily found elsewhere. Rather, the aim of this chapter is to explore some of the ramifications of these results, especially concerning rates of convergence, with the intention of giving greater insight into how extreme value approximations work in practice. Our objective is both probabilistic (e.g. deriving a rate of convergence in (2.1)) and statistical (providing detailed justifications for some of the statistical procedures in Chapter 1,

and considering optimality results such as the asymptotically best choice of threshold in threshold methods).

The first result is to establish a stability property for the possible limits H in (2.1). Suppose $F_n,\ n \geq 1$ and suppose there exist sequences $a_n > 0,\ b_n \in \mathbb{R}$ and also $\alpha_n > 0,\ \beta_n \in \mathbb{R}$ and distribution functions H_1 and H_2 such that

$$F_n(a_n x + b_n) \to H_1(x),\ F_n(\alpha_n x + \beta_n) \to H_2(x).$$

Then there exist constants $A > 0,\ B \in \mathbb{R}$ such that

$$\frac{\alpha_n}{a_n} \to A,\ \frac{\beta_n - b_n}{a_n} \to B$$

and then

$$H_2(x) \quad = \quad H_1(Ax + B). \tag{2.2}$$

This result is known as *Khinchine's Lemma* which we shall not prove here as it has been given numerous times by previous authors, for instance [71], page 253 or [143], page 7.

Now let's see what this implies for the special case $F_n = F^n$. Let $k > 1$ be a fixed integer and consider limits of F^{nk} as $n \to \infty$. By applying (2.1) twice, we get

$$F^{kn}(a_{kn} x + b_{kn}) \to H(x),\ F^{kn}(a_n x + b_n) \to H^k(x).$$

Hence there exist $A_k > 0$ and $B_k \in \mathbb{R}$ such that $\frac{a_n}{a_{nk}} \to A_k,\ \frac{b_n - b_{nk}}{a_{nk}} \to B_k$, and

$$H^k(x) \quad = \quad H(A_k x + B_k). \tag{2.3}$$

A distribution function H that satisfies (2.3) is said to be *max-stable*. This leads us to:

Theorem 2.1. If H is a max-stable distribution function, then H must be of the same type as one of (1.3)–(1.5), or equivalently, (1.6).

We shall not give the proof as this has been given many times in previous texts. Gnedenko's original paper [87] is still well worth reading (see [136] for an English translation) but a much simplified proof was given by de Haan [99]. A modern proof has been given by de Haan and Ferreira [101], Theorem 1.1.3.

We now state the main domain of attraction condition of Gnedenko [87]:

Theorem 2.2. Suppose (2.1) holds with $H(x)$ one of (1.3)–(1.5).

(i) If $H(x)$ is of the form (1.4) with given $\alpha > 0$, then a necessary and sufficient condition for (2.1) to hold for some $a_n > 0,\ b_n \in \mathbb{R}$ is that $\omega_F = \sup\{x :\ F(x) < 1\} = \infty$ and

$$\lim_{t \to \infty} \frac{1 - F(xt)}{1 - F(t)} \quad = \quad x^{-\alpha} \text{ for any } x > 0. \tag{2.4}$$

In this case we may, without loss of generality, define $b_n = 0,\ a_n = \inf\{x :\ 1 - F(x) \leq 1/n\}$.

(ii) If $H(x)$ is of the form (1.5) with given $\alpha > 0$, then a necessary and sufficient condition for (2.1) to hold for some $a_n > 0,\ b_n \in \mathbb{R}$ is that $\omega_F = \sup\{x:\ F(x) < 1\} < \infty$ and

$$\lim_{t\to 0} \frac{1-F(\omega_F - xt)}{1-F(\omega_F - t)} \quad = \quad x^{\alpha} \text{ for any } x > 0. \tag{2.5}$$

In this case we may define $b_n = \omega_F,\ a_n = \omega_F - \inf\{x:\ 1-F(x) \le 1/n\}$.

(iii) If $H(x)$ is of the form (1.3), then a necessary and sufficient condition for (2.1) to hold for some $a_n > 0,\ b_n \in \mathbb{R}$ is that there exists a function $\psi(t)$, defined for $t \le \omega_F$,

$$\lim_{t\to \omega_F} \frac{1-F(t+x\psi(t))}{1-F(t)} \quad = \quad e^{-x} \text{ for any } x > 0. \tag{2.6}$$

In this case ω_F may be finite or infinite and we may define $b_n = \inf\{x:\ 1-F(x) \le 1/n\},\ a_n = \psi(b_n)$.

A weakness of Gnedenko's result was that he did not give an explicit expression for the function $\psi(t)$. However, De Haan [98] showed that, when ψ exists, it may be given by

$$\psi(t) \quad = \quad \frac{\int_t^{\omega_F} \{1-F(s)\}ds}{1-F(t)}. \tag{2.7}$$

(It is part of the condition that the integral be finite.)

There have been numerous equivalent conditions given by authors such as Mejzler [158], Marcus and Pinsky [150], de Haan [98] and Pickands [183].

Although a full proof that the conditions of Theorem 2 are both necessary and sufficient is rather long-winded, we can rather quickly prove that the conditions are sufficient, which will serve to motivate the basic idea of how we test whether a distribution function F is in a domain of attraction.

In (i), the condition on a_n implies that $n\{1-F(a_n)\} \to 1$ as $n \to \infty$. This is immediate if F is continuous, because then we can find a_n so that $n\{1-F(a_n)\} = 1$. For F not everywhere continuous, we can still derive the same result by noting that, for any $\varepsilon > 0$, the condition $1-\varepsilon < n\{1-F(a_n)\} < 1+\varepsilon$ holds for all sufficiently large n; since ε is arbitrary, this can only be true if the limit is 1. Similarly in case (ii), $n\{1-F(\omega_F - a_n)\} \to 1$, and in case (iii), $n\{1-F(b_n)\} \to 1$. Then in case (i),

$$\begin{aligned} \lim_{n\to\infty} n\{1-F(a_n x)\} \quad &= \quad \lim_{n\to\infty} \left\{ \frac{1-F(a_n x)}{1-F(a_n)} \right\} \\ &= \quad x^{-\alpha}. \end{aligned}$$

Next, we note that because $1-F(a_n x) \to 0$ as $n \to \infty$, we must have $\frac{-\log F(a_n x)}{1-F(a_n x)} \to 1$, and hence

$$\lim_{n\to\infty} n \log F(a_n x) \quad = \quad -x^{-\alpha}.$$

Taking exponentials on both sides leads to

$$\lim_{n\to\infty} F^n(a_n x) \quad = \quad \exp(-x^{-\alpha})$$

whenever $x > 0$. Cases (ii) and (iii) are similar.

2.1.1 Convergence of threshold exceedances to the Generalized Pareto Distribution

We can also show that each of the conditions (2.4)–(2.6) implies convergence of threshold exceedances to the corresponding Generalized Pareto Distribution (GPD). Specifically, if X is a random variable whose distribution function is F, and if u is a high threshold, we want to show that

$$\Pr\{X - u \geq y\sigma_u \mid X > u\} \to (1+\xi y)^{-1/\xi} \tag{2.8}$$

as $u \to \omega_F$, for suitable ξ and σ_u, for any y such that $1+\xi y > 0$.

Consider first (2.4). We set $\sigma_u = \frac{u}{\alpha}$, then

$$\Pr\{X - u \geq y\sigma_u \mid X > u\} \quad = \quad \frac{1 - F\left(u + \frac{yu}{\alpha}\right)}{1 - F(u)} \;\to\; \left(1 + \frac{y}{\alpha}\right)^{-\alpha}.$$

This is of the form (2.8) with $\xi = \frac{1}{\alpha}$.

Next, assume (2.5). Suppose we have a threshold $u = \omega_F - t$ for some small t. Suppose we have $0 < y < \alpha$ and define $\sigma_u = \frac{t}{\alpha}$. Then

$$\Pr\{X > u + y\sigma_u \mid X > u\} \quad = \quad \frac{1 - F\left(u + \frac{yt}{\alpha}\right)}{1 - F(u)} \;=\; \frac{1 - F\left(\omega_F - t + \frac{yt}{\alpha}\right)}{1 - F(\omega_F - t)} \;\to \left(1 - \frac{y}{\alpha}\right)^{\alpha}.$$

This is of the form (2.8) with $\xi = -\frac{1}{\alpha}$.

Finally, assume (2.6). Let $\sigma_u = \psi(u)$, Then

$$\Pr\{X > u + y\sigma_u \mid X > u\} \quad = \quad \frac{1 - F\left(u + y\psi(u)\right)}{1 - F(u)} \;\to e^{-y},$$

which is the limiting form of (2.8) as $\xi \to 0$.

Therefore, in all three cases, we get the GPD as a limiting distribution for exceedances over a threshold.

2.2 Examples

2.2.1 The t distribution and extensions

The pdf of the t distribution with ν degrees of freedom is given by

$$f(t;\, \nu) \quad = \quad \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\,\Gamma(\nu/2)} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}, \; -\infty < t < \infty. \tag{2.9}$$

This is symmetric about $t = 0$, so it suffices to consider the limits as $t \to +\infty$. Expanding in Taylor series,

$$\begin{aligned} f(t;\,\nu) &= \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\Gamma(\nu/2)}\frac{t^{-(\nu+1)}}{\nu^{-(\nu+1)/2}}\left\{1-\frac{\nu(\nu+1)}{2t^2}+O\left(\frac{1}{t^4}\right)\right\} \\ &= \frac{\Gamma((\nu+1)/2)\nu^{\nu/2}}{\sqrt{\pi}\Gamma(\nu/2)}\left\{t^{-(\nu+1)}-\frac{\nu(\nu+1)}{2}t^{-(\nu+3)}+O(t^{-(\nu+5)})\right\} \end{aligned}$$

Integrating term by term,

$$1-F(t;\,\nu) = \frac{\Gamma((\nu+1)/2)\nu^{\nu/2-1}}{\sqrt{\pi}\Gamma(\nu/2)}t^{-\nu}\left\{1-\frac{\nu^2(\nu+1)}{2(\nu+2)}t^{-2}+O(t^{-4})\right\}. \tag{2.10}$$

To put (2.10) on a broader footing, consider an expansion of the form

$$1-F(t) = ct^{-\alpha}+dt^{-\alpha-\beta}+o(t^{-\alpha-\beta}) \tag{2.11}$$

where $\alpha,\ \beta,\ c$ and d are all constants, the first three being > 0. In the case of the t distribution, we have $\alpha = \nu,\ \beta = 2$.

If we define $a_n = (nc)^{1/\alpha}$, we have, for any $x > 0$,

$$n\{1-F(a_nx)\} \sim nc\,(a_nx)^{-\alpha} = x^{-\alpha}$$

from which it follows by arguments already given that, first, $n\log F(a_nx) \to -x^{-\alpha}$ and, second,

$$F^n(a_nx) = \exp(-x^{-\alpha}),\ x > 0, \tag{2.12}$$

which confirms the limit of the form (1.4).

Given the extra term in (2.11), we can go further. Still defining $a_n = (nc)^{1/\alpha}$, we have

$$n\{1-F(a_nx)\} = x^{-\alpha}+n^{-\beta/\alpha}c^{-1+\beta/\alpha}dx^{-\alpha-\beta}+o(n^{-\beta/\alpha}).$$

If $\beta < \alpha$ (more on this condition in a moment) the same expansion also holds for $-n\log\{F(a_nx)\}$ and so

$$F^n(a_nx) = \exp(-x^{-\alpha})\left\{1-n^{-\beta/\alpha}c^{-1+\beta/\alpha}dx^{-\alpha-\beta}+o(n^{-\beta/\alpha})\right\}, \tag{2.13}$$

valid for all $x > 0$. Equation (2.13) is our first instance of a *rate of convergence* result in extreme value theory, and raises numerous issues such as whether the rate of convergence is uniform and whether there is any possibility of achieving a faster rate of convergence by a different normalization. Both questions were addressed by [221], with an affirmative answer about uniformity and a negative answer, with one exception, to the question of a faster rate. From a practical point of view, the importance of a result like (2.13) is that it gives some concrete answers to how good the

extreme value distributions are as approximations, which, as we shall see, also plays into questions about statistical estimation.

The one exception in this discussion is when $\beta = 1$. Consider the case when b_n is a fixed constant b, and consider

$$\begin{aligned} n\{1-F(a_nx+b)\} &= c(a_nx+b)^{-\alpha} + d(a_nx+b)^{-\alpha-1} + o(a_n^{-\alpha-1}) \\ &= c(a_nx)^{-\alpha}\left(1-\frac{\alpha b}{a_nx}\right) + da_n^{-\alpha-1}x^{-\alpha-1} + o(a_n^{-\alpha-1}). \end{aligned}$$

If we set $b = d/(\alpha c)$, then central two terms cancel and we deduce

$$n\{1-F(a_nx+b)\} = c(a_nx)^{-\alpha} + o(a_n^{-\alpha-1}).$$

Again defining $a_n = (nc)^{1/\alpha}$, the same argument as led to (2.13) shows that

$$F^n(a_nx+b) = \exp(-x^{-\alpha})\left\{1+o(n^{-1/\alpha})\right\}, \tag{2.14}$$

in other words, the error rate is $o(n^{-1/\alpha})$ rather than $O(n^{-1/\alpha})$. However, the case $\beta = 1$ is the only case where an improvement in the rate of convergence is possible [221].

The reader may be wondering why we restricted to the case $0 < \beta < \alpha$. Note that in this case, the rate of convergence in (2.13) is $O(n^{-\beta/\alpha})$, which is slower than $O(1/n)$. However, the transformation from $1-F(a_nx)$ to $-\log\{F(a_nx)\}$ induces an additional error term of $O(1/n)$, which would mess up the result if $\beta \geq \alpha$. However, if the rate of convergence is as good as $O(1/n)$, this is already a very fast rate of convergence, so we should not worry about trying to improve it.

Remark. In (2.13), [221] used a different definition of a_n which leads to a slightly different formula. The present derivation is more direct since it does not require the full theory developed in [221].

2.2.2 The beta distribution and extensions

As in the previous section, we use the beta distribution as a motivating example but the intention is to explore a wider class of distributions.

Suppose

$$f(t;\, a,b) = \frac{1}{B(a,b)}t^{a-1}(1-t)^{b-1},\ 0<x<1, \tag{2.15}$$

where $a > 0,\ b > 0$ and $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.

This distribution has finite endpoint $\omega_F = 1$ and we can easily check that as $t \uparrow 1$,

$$1-F(t;\, a,b) = \frac{1}{B(a,b)}\left\{\frac{(1-t)^b}{b} - \frac{(a-1)(1-t)^{b+1}}{b+1} + O((1-t)^{b+2})\right\}.$$

This is a special case of the general formula

$$1-F(t) = c(\omega_F-t)^{\alpha} + d(\omega_F-t)^{\alpha+\beta} + o(\omega_F-t)^{\alpha+\beta},\ t\uparrow\omega_F<\infty, \tag{2.16}$$

where $\alpha,\ \beta$ and c are all > 0, so we take (2.16) as our starting point.

In this case, define $b_n = \omega_F,\ a_n = (nc)^{-1/\alpha}$, then for $x < 0$,

$$\begin{aligned} n\{1-F(a_n x+b_n)\} &= n\{1-F(\omega_F - a_n|x|)\} \\ &= nca_n^{\alpha} + nda_n^{\alpha+\beta}|x|^{\alpha+\beta} + \ldots \\ &= |x|^{\alpha} + n^{-\beta/\alpha}dc^{-1-\beta/\alpha}|x|^{\alpha+\beta} + o(n^{-\beta/\alpha}). \end{aligned}$$

By the same argument as in the previous section,

$$F^n(a_n x+b_n) \quad \to \quad \exp(-|x|^{\alpha})\ (x<0) \tag{2.17}$$

and if $0 < \beta < \alpha$,

$$F^n(a_n x+b_n) \quad = \quad \exp(-|x|^{\alpha}\left(1 - n^{-\beta/\alpha}dc^{-1-\beta/\alpha}|x|^{\alpha+\beta} + o(n^{-\beta/\alpha}\right).\ (x<0) \tag{2.18}$$

So (2.17) establishes that the limit distribution is (1.5), and (2.18) shows the rate of convergence (as in the previous section, if $\beta \geq \alpha$ the rate of convergence is $O(1/n)$).

Here again, [221] gave a slightly different version of the same result, as part of a more general theory. In this case, there is no improvement on the rate of convergence by a different choice of a_n and b_n.

Remark. The more usual context for this result is for sample minima in the case of a distribution where $F(t) \sim ct^{\alpha}$ as $x \downarrow 0$, which is a common assumption in reliability theory. If we assume $F(t) = ct^{\alpha} + dt^{\alpha+\beta} + o(t^{\alpha+\beta})$ for small t, we again define $a_n = (nc)^{-1/\alpha}$ when $nF(a_n x) = x^{\alpha} + n^{-\beta/\alpha}dc^{-1-\beta/\alpha}x^{\alpha+\beta} + o(n^{-\beta/\alpha})$ for $x > 0$. Writing $F(a_n x) \sim -\log(1-F(a_n x))$ we deduce that if $X_1, X_2, \ldots$ are IID with distribution function F,

$$\Pr\{\min(X_1,\ldots,X_n) \leq a_n x\} \quad = \quad 1-\{1-F(a_n x)\}^n \ \to 1-\exp(-x^{\alpha}).$$

This is the well-known Weibull distribution, often used for strength of materials and similar applications.

2.2.3 Normal distribution

At the time that Fisher and Tippett [77] wrote their foundational paper about extreme value theory, there was a much stronger belief than there is today that the normal distribution is ubiquitous in nature. Hence, the bulk of their paper, and much subequent theory, has been focused on extreme from normally distributed sample. It turns out that this case is, in fact, one of the most complex examples of extreme value theory.

To fix notation, we let $\phi(x) = 1/\sqrt{2\pi}e^{-x^2/2}$ be the density of a standard normal distribution, and $\Phi(x) = \int_{-\infty}^{x}\phi(t)dt$ the standard normal distribution function. The following theory relies critically on the expansion

$$1-\Phi(x) \quad = \quad \frac{\phi(x)}{x}\left(1 - \frac{1}{x^2} + \frac{3}{x^4} - \frac{15}{x^6} + \ldots\right) \tag{2.19}$$

([70], page 193).

Assume $\psi(t)$ is an arbitrary function of real variable t such that $\frac{\psi(t)}{t} \to 0$ as $t \to \infty$. Using (2.19), we calculate

$$\begin{aligned}\frac{1-\Phi(t+x\psi(t))}{1-\Phi(t)} &\sim \frac{\phi(t+x\psi(t))}{\phi(t)} \\ &= \exp\left\{-\frac{(t+x\psi(t))^2}{2}+\frac{t^2}{2}\right\} \\ &= \exp\left\{-tx\psi(t)-\frac{x^2\psi^2(t)}{2}\right\}.\end{aligned}$$

If we set $\psi(t)=\frac{1}{t}$, this converges to e^{-x} as $t\to\infty$. This is Gnedenko's condition (2.6), and therefore establishes that Φ is in the domain of attraction of the Gumbel distribution (1.3). Moreover, Gnedenko's theory also implies that we may define $a_n,\ b_n$ by $1-\Phi(b_n)=\frac{1}{n},\ a_n=\psi(b_n)=\frac{1}{b_n}$.

Application of (2.19) implies that

$$\frac{1}{n} = \frac{\phi(b_n)}{b_n}\left(1-\frac{1}{b_n^2}+\frac{3}{b_n^4}-\frac{15}{b_n^6}+\ldots\right) \tag{2.20}$$

We shall use the result (2.20) in two ways. First, we use it to derive an asymptotic approximation to b_n. A crude first guess would ignore b_n in the denominator and set $\phi(b_n)=\frac{1}{n}$, which would imply $b_n\sim\sqrt{2\log n}$. However, this would not achieve the desired result $1-\Phi(b_n)\sim\frac{1}{n}$ so we need to refine the approximation. We do that by writing

$$b_n = \sqrt{2\log n}+c_n$$

and trying to find a good approximation to c_n. The argument assumes (and will later verify) that $\frac{c_n}{\sqrt{2\log n}}\to 0$.

In this case, (2.20) implies

$$\begin{aligned}\frac{\phi(b_n)}{b_n} &\sim \frac{1}{\sqrt{2\pi}b_n}e^{-\frac{1}{2}(\sqrt{2\log n}-c_n)^2} \\ &= \frac{1}{\sqrt{2\pi}b_n}e^{-\frac{1}{2}(2\log n-2c_n\sqrt{2\log n}-c_n^2)}\end{aligned}$$

so b_n must satisfy

$$\begin{aligned}\frac{1}{n} &\sim \frac{1}{\sqrt{2\pi}b_n}e^{-\frac{1}{2}(\sqrt{2\log n}-c_n)^2} \\ &= \frac{1}{\sqrt{4\pi\log n}}\cdot\frac{1}{n}\cdot e^{-c_n\sqrt{2\log n}+\frac{1}{2}c_n^2)}\end{aligned}$$

If we ignore the term $\frac{1}{2}c_n^2$, we can solve directly for $c_n=-\frac{1}{2\sqrt{2\log n}}(\log 4\pi+\log\log n)$

(which incidentally does prove that $\frac{c_n}{\sqrt{2\log n}} \to 0$) so we deduce the approximation

$$b_n = \sqrt{2\log n} - \frac{1}{2\sqrt{2\log n}}(\log 4\pi + \log\log n) + o\left(\frac{\log\log n}{\sqrt{2\log n}}\right). \quad (2.21)$$

In fact, if we ignore the term $o\left(\frac{\log\log n}{\sqrt{2\log n}}\right)$, then b_n defined by (2.21), together with $a_n = \frac{1}{b_n}$, are often taken as *the* normalizing constants for extreme value theory for the normal distribution, but Hall [108] showed that this is not the optimal choice for a_n and b_n.

We therefore take the alternative approach recommended by Hall, and use (2.20) to motivate the alternative definition of b_n as the value that satisfies

$$\frac{\phi(b_n)}{b_n} = \frac{1}{n}. \quad (2.22)$$

Of course, with modern computers, an effectively exact solution to (2.22) is quickly obtained for any likely value of n, a statement that was not true at the time of Fisher and Tippett. So it's reasonable to take (2.22) as the modern definition of b_n.

Now let's proceed further with an expansion based on (2.19). We have,

$$1-\Phi\left(b_n+\frac{x}{b_n}\right) = \left(b_n+\frac{x}{b_n}\right)^{-1}\phi\left(b_n+\frac{x}{b_n}\right)\left\{1-\left(b_n+\frac{x}{b_n}\right)^{-2}+3\left(b_n+\frac{x}{b_n}\right)^{-4}-\ldots\right\}$$

Expanding each of the terms as far as $O\left(\frac{1}{b_n^4}\right)$,

$$\begin{aligned}\left(b_n+\frac{x}{b_n}\right)^{-1} &= \frac{1}{b_n}\left(1-\frac{x}{b_n^2}+\frac{x^2}{b_n^4}+\ldots\right),\\ \phi\left(b_n+\frac{x}{b_n}\right) &= \phi(b_n)\,e^{-x}\left(1-\frac{x^2}{2b_n^2}+\frac{x^4}{8b_n^4}\right),\\ 1-b_n^{-2}\left(b_n+\frac{x}{b_n}\right)^{-2}+3b_n^{-4} &= 1-\frac{1}{b_n^2}+\frac{2x+3}{b_n^4}+\ldots\end{aligned}$$

so

$$1-\Phi\left(b_n+\frac{x}{b_n}\right) = \frac{\phi(b_n)}{b_n}e^{-x}\left\{1-\frac{1}{b_n^2}\left(1+x+\frac{x^2}{2}\right)+O\left(\frac{1}{b_n^4}\right)\right\}.$$

But $\frac{\phi(b_n)}{b_n} = \frac{1}{n}$ by definition, so if we also define $a_n = \frac{1}{b_n}$ we have

$$n\{1-\Phi(a_nx+b_n)\} = e^{-x}\left\{1-\frac{1}{b_n^2}\left(1+x+\frac{x^2}{2}\right)+O\left(\frac{1}{b_n^4}\right)\right\}. \quad (2.23)$$

As with our previous examples, we can replace $\{1-\Phi(a_nx+b_n)\}$ by $-\log\{\Phi(a_nx+b_n)\}$ with an error of $O\left(\frac{1}{n}\right)$, and hence we deduce from (2.23) that $\Phi^n(a_nx+b_n) \to$

$\exp(-e^{-x})$ with an error of $O\left(\frac{1}{b_n^2}\right) = O\left(\frac{1}{\log n}\right)$ since $b_n \sim \sqrt{2\log n}$ as $n \to \infty$. However, this is a very slow rate of convergence, and many practical examples over the years have shown that this is not a good approximation to the distribution of maxima of normally distributed random variables.

This discussion raises the question of whether we could use the explicit form of the $O\left(\frac{1}{b_n^2}\right)$ term in (2.23) to get a better approximation. We could, for example, incorporate that term directly into the approximation, but there is a better way.

The idea is to investigate whether we could rewrite the first two terms in the right side of (2.23)

$$e^{-x}\left\{1 - \frac{1}{b_n^2}\left(1 + x + \frac{x^2}{2}\right)\right\},$$

in the form appropriate for a GEV limit

$$\left(1 + \xi_n \frac{x - \mu_n}{\psi_n}\right)^{-1/\xi_n},$$

still with error $O\left(\frac{1}{b_n^4}\right)$, where we have written the GEV parameters $\mu_n,\ \psi_n,\ \xi_n$ as functions of n since we know we need $\mu_n \to 0,\ \psi_n \to 1,\ \xi_n \to 0$ in order to achieve the correct limit e^{-x} as $n \to \infty$.

At first sight, this idea might seem crazy, since the last two expressions are of completely different functional forms. However, if we take logarithms of both sides and equate

$$\begin{aligned} -x - \frac{1}{b_n^2}\left(1 + x + \frac{x^2}{2}\right) &\approx -\frac{1}{\xi_n}\log\left(1 + \xi_n \frac{x - \mu_n}{\psi_n}\right) \\ &\approx -\frac{x - \mu_n}{\psi_n} + \frac{\xi_n}{2}\left(\frac{x - \mu_n}{\psi_n}\right)^2, \end{aligned}$$

we can see that both sides of the approximation are quadratic in x, so if we can find $\mu_n,\ \psi_n,\ \xi_n$ to equate the coefficients of 1, x and x^2 (with error of $o(1/b_n^2)$), we will indeed have the result we want. The reader can quickly check that this is true if we set

$$\mu_n = -\frac{1}{b_n^2},\ \psi_n = 1 - \frac{1}{b_n^2},\ \xi_n = -\frac{1}{b_n^2} \tag{2.24}$$

and, in that case, the error of the approximation is $O\left(\frac{1}{b_n^4}\right) = O\left(\frac{1}{\log^2 n}\right)$.

We therefore deduce that, with $\mu_n,\ \psi_n,\ \xi_n$ defined by (2.24),

$$\Phi^n(a_n x + b_n) = \exp\left(1 + \xi_n \frac{x - \mu_n}{\psi_n}\right) + O\left(\frac{1}{\log^2 n}\right). \tag{2.25}$$

These arguments have been presented in a heuristic way but they can be made

rigorous. For the $O\left(\frac{1}{\log n}\right)$ error in the Gumbel approximation, Hall [108] showed that the convergence is uniform over x and he derived a numerical upper bound for the error. He also showed that the rate of convergence cannot be improved by a different choice of a_n and b_n. The corresponding results for (2.24) were derived by Cohen [31]. Numerical calculations confirm that, not only is (2.24) a superior approximation asymptotically, but it produces a far better approximation in practice.

Equation (2.25) is known as the *penultimate approximation* to normal extremes. Remarkably, the idea (and the name) goes back to Fisher and Tippett [77]. The notation and detailed calculation were different, but they recognized and stated that the Weibull-type of approximation is a better fit to the distribution of normal extremes than the Gumbel model. The practical implications for statistics are twofold:

1. Even though the Gumbel limit ($\xi = 0$ in the GEV) may be the "ultimate" limiting approximation, the GEV with $\xi \neq 0$ still fits the data better. Therefore, statisticians should just fit the GEV, and not try to formally distinguish the different model classes.
2. In the case of normal extremes, the best-fitting GEV has finite upper endpoint, even though the distribution being approximated is unbounded. This has implications for datasets like the temperature extreme datasets discussed in Chapter 1 — even though the statistical modeling may point towards a distribution with finite upper endpoint, this is not necessarily the correct conclusion.

2.2.4 Lognormal distribution

Assume X is lognormal, i.e. $\log X \sim \mathcal{N}[\mu, \sigma^2]$ for some μ and σ^2. There is again no loss of generality in assuming $\mu = 0$ but we allow σ^2 to be flexible because the value of σ^2 does affect the shape of the distribution. We therefore write

$$F(x) \quad = \quad \Phi(\delta \log x) \tag{2.26}$$

where $\delta = 1/\sigma$ and Φ is again the standard normal distribution function.

Suppose $t \to \infty$ and define $\psi(t) = \frac{t}{\delta^2 \log t}$. Starting from (2.19), by Taylor expansion and considerable algebraic manipulation, we deduce

$$\log\left\{\frac{1-F(t+x\psi(t))}{1-F(t)}\right\} \quad = \quad -x + \frac{x^2}{2\delta^2 \log t} - \frac{x}{\delta^2 \log^2 t} - \frac{x^2}{2\delta^2 \log^2 t} - \frac{x^3}{3\delta^4 \log^2 t} + O\left(\frac{1}{\log^3 t}\right). \tag{2.27}$$

Define b_n by $n\{1-F(b_n)\} = n\{1-\Phi(\delta \log b_n)\}$ and $a_n = \psi(b_n)$. Then $n\{1-F(a_n x + b_n)\} \to e^{-x}$ so we have convergence to the Gumbel limit $F^n(a_n x + b_n) \to \exp(-e^{-x})$. The rate of convergence is $O\left(\frac{1}{\log b_n}\right)$. However, since we previously showed that the solution of $n\{1-\Phi(b_n)\} = 1$ satisfies $b_n \sim \sqrt{2\log n}$, it follows that the solution of $n\{1-\Phi(\delta \log b_n)\} = 1$ satisfies $\delta \log b_n \sim \sqrt{2\log n}$. Therefore, as a function of n, the error in the Gumbel approximation is $O\left(\frac{1}{\sqrt{\log n}}\right)$.

As with the previous example, we attempt to improve on this via a penultimate approximation. Rewriting (2.27) in the form

$$\log[n\{1-F(a_nx+b_n)\}] = -x+\frac{x^2}{2\delta^2\log b_n}-\frac{x}{\delta^2\log^2 b_n}-\frac{x^2}{2\delta^2\log^2 b_n}-\frac{x^3}{3\delta^4\log^2 b_n}+O\left(\frac{1}{\log^3 b_n}\right), \tag{2.28}$$

we aim to approximate the left hand side of (2.28) by an expression of the form $-\frac{1}{\xi_n}\log\left(1+\xi_n\frac{x-\mu_n}{\psi_n}\right)$: by the same argument that we have now used several times, this will lead to a GEV approximation for $F^n(a_nx+b_n)$. Taking $\mu_n=0,\ \psi_n=\left(1+\frac{1}{\delta^2\log^2 b_n}\right)^{-1},\ \xi_n=\frac{1}{\delta^2\log b_n}-\frac{1}{\delta^2\log^2 b_n}$, we deduce

$$\begin{aligned}-\frac{1}{\xi_n}\log\left(1+\frac{\xi_nx}{\psi_n}\right) &= -\frac{x}{\psi_n}+\frac{\xi_nx^2}{2\psi_n^2}-\frac{\xi_n^2x^3}{3\psi_n^3}+O\left(\frac{1}{\log^3 b_n}\right)\\ &= -x-\frac{x}{\delta^2\log^2 b_n}+\frac{x^2}{2\delta^2\log b_n}-\frac{x^2}{2\delta^2\log^2 b_n}-\frac{x^3}{3\delta^4\log^2 b_n}+O\left(\frac{1}{\log^3 b_n}\right)\end{aligned}$$

whereas for the right side of (2.28), the same operation yields

$$-x+\frac{x^2}{2\delta^2\log b_n}-\frac{x}{\delta^2\log^2 b_n}-\frac{x^2}{2\delta^2\log^2 b_n}-\frac{x^3}{3\delta^4\log^2 b_n}+O\left(\frac{1}{\log^3 b_n}\right).$$

so the two expressions agree to $O\left(\frac{1}{\log^3 b_n}\right)$. Exponentiating back,

$$n\{1-F(a_nx+b_n)\} = \left(1+\frac{\xi_nx}{\psi_n}\right)^{-1/\xi_n}+O\left(\frac{1}{\log^3 b_n}\right)$$

and hence

$$F^n(a_nx+b_n) = \exp\left\{-\left(1+\frac{\xi_nx}{\psi_n}\right)^{-1/\xi_n}\right\}+O\left(\frac{1}{\log^3 b_n}\right). \tag{2.29}$$

Thus, in this case, the penultimate approximation not only kills the $O\left(\frac{1}{\log b_n}\right)$ term but the $O\left(\frac{1}{\log^2 b_n}\right)$, making the final rate of convergence $O\left\{(\log n)^{-3/2}\right\}$. This remarkable fact was discovered by Cohen [30], who also proved that the rate is uniform in $x\in\mathbb{R}$.

2.2.5 *An example of a distribution with finite ω_F in the Gumbel domain of attraction*

Consider $1-F(t)=\exp(1/t)$ for $t<0$. Then $F(t)\to 1$ as $t\uparrow 0$ so $\omega_F=0$. Define $\psi(t)$ for $t<0$ so that $\frac{\psi(t)}{t}\to 0$ as $t\uparrow 0$. Then

$$\begin{aligned}\frac{1-F(t+x\psi(t))}{1-F(t)} &= \exp\left\{\frac{1}{t+x\psi(t)}-\frac{1}{t}\right\}\\ &= \exp\left\{\frac{-x\psi(t)}{t(t+x\psi(t))}\right\}.\end{aligned}$$

If we define $\psi(t) = t^2$, then the exponent $\to x$ as $t \uparrow 0$. By Gnedenko's condition, F is in the domain of attraction of the Gumbel distribution.

Earlier, we argued that by means of the penultimate approximation, normal extremes (which have infinite range) could be approximated by a GEV with finite range. This shows the opposite behavior: that the extremes from a distribution with finite range could still tend to the Gumbel distribution as $n \to \infty$. Put another way, fitting a Gumbel distribution to data does not preclude the possibility that the true distribution may have finite range. This may be relevant, for example, in the case of rainfall extremes, where extreme value theory often produces a distribution of infinite upper endpoint but meteorologists argue that there is a "probable maximum precipitation", reflecting the physical fact that there is an upper limit to the amount of moisture that the atmosphere can hold. The two statements are not contradictory.

2.2.6 An example of a continuous distribution not in any domain of attraction

Consider $1 - F(x) = \frac{1}{\log x}$ for $x > 1$. Since this distribution has $\omega_F = \infty$, the limit (1.5) is ruled out. For (1.4), Gnedenko's condition (2.4) would imply $\frac{1-F(tx)}{1-F(t)} \to x^\alpha$ as $t \to \infty$ for some $\alpha > 0$. But $\frac{1-F(tx)}{1-F(t)} \to 1$ so this cannot hold (the case $\alpha = 0$ is not allowed). Likewise, de Haan's condition (2.7) would require that $\int_t^\infty (1 - F(s))ds = \int_t^\infty \frac{1}{\log s} ds$ be finite, but it obviously is not. Therefore, this distribution cannot be in any domain of attraction.

2.2.7 Discrete distributions

Another class of distributions where classical extreme value theory typically does not hold is for discrete distributions such as geometric and Poisson. Essentially, these distributions are too lumpy for smooth limits to exist, but there other things one can do. Anderson [3, 4] classified the limit behavior in a number of such cases. These results have found applications in probability problems such as the longest run of heads in a sequence of coin tosses, or the longest run of matching genes in tissue samples [46, 91, 7].

Other methods are based on the Stein-Chen (or Chen-Stein) method of Poisson approximation, see e.g. [12, 5, 6, 8, 203].

2.3 Reformulation in Terms of Inverse Functions

The crux of de Haan's conditions for convergence of extreme value distribution is to reformulate the problem in terms of *inverse functions*. Here, we are essentially following de Haan and Ferreira [101].

Let f be a non-decreasing function on $\mathbb{R}$. Define $f^{\leftarrow}(x) = \inf\{y : f(y) \geq x\}$. Then $f^{\leftarrow}(x)$ is the left-continuous inverse of f (if $f(x)$ is a constant y on an interval $a < x < b$, then $f^{\leftarrow}(y) = a$).

Following [101], we define

$$U(t) \;=\; \left(\frac{1}{1-F}\right)^{\leftarrow}(t) \;=\; \inf\left\{y:\; F(y) \geq 1-\frac{1}{t}\right\}. \tag{2.30}$$

Note that in many of our extreme value theory examples so far, either a_n or b_n is defined to be $U(n)$.

Much of the theory is concerned with relations of the form

$$\lim_{t\to\infty} \frac{U(tx)-U(t)}{a(t)} \;=\; \frac{x^{\xi}-1}{\xi} \text{ for any } x>0, \tag{2.31}$$

where the limit $\xi \to 0$ is defined to be $\log x$.

Example 1. Suppose $1-F(x) \sim cx^{-\alpha}$ as $x \to \infty$. Then $1-F(x) = \frac{1}{n}$ corresponds to $x \sim (nc)^{1/\alpha}$; for present discussion we define it so that $U(t) = (tc)^{1/\alpha}$ exactly. Define $a(t) = \alpha^{-1}(ct)^{1/\alpha}$. Then

$$\frac{U(tx)-U(t)}{a(t)} \;=\; \frac{(txc)^{1/\alpha}-(tc)^{1/\alpha}}{\alpha^{-1}(ct)^{1/\alpha}} \;=\; \frac{x^{1/\alpha}-1}{\alpha^{-1}}$$

of form (2.31) with $\xi = \frac{1}{\alpha}$.

Example 2. Suppose $1-F(x) \sim c|x|^{\alpha}$ as $x \uparrow 0$. Then $1-F(x) = \frac{1}{n}$ corresponds to $x \sim -(nc)^{-1/\alpha}$; again we assume that the asymptotic relation is exact so $U(t) = (tc)^{-1/\alpha}$. Define $a(t) = \alpha^{-1}(ct)^{-1/\alpha}$. Then

$$\frac{U(tx)-U(t)}{a(t)} \;=\; \frac{-(txc)^{-1/\alpha}+(tc)^{-1/\alpha}}{\alpha^{-1}(ct)^{-1/\alpha}} \;=\; \frac{x^{-1/\alpha}-1}{(-\alpha^{-1})}$$

of form (2.31) with $\xi = -\frac{1}{\alpha}$.

Example 3. Suppose $F(x) = \Phi(x)$, the standard normal distribution. Define $U(t)$ by $(1-\Phi(U(t)) = \frac{1}{t}$ for all $t>0$. This corresponds to one of the definitions of b_n in Section 2.2.3, so in particular $U(n) = b_n$. But we already saw that $b_n \sim \sqrt{2\log n}$, hence $U(t) \sim \sqrt{2\log n}$. But in that case,

$$\begin{aligned} U(tx)-U(t) &\approx \sqrt{2\log(tx)} - \sqrt{2\log(t)} \\ &= \sqrt{2\log(t)}\left\{\left(1+\frac{\log x}{\log t}\right)^{1/2}-1\right\} \\ &\sim \sqrt{2\log(t)}\,\frac{\log x}{2\log t} \;=\; \frac{\log x}{\sqrt{2\log t}}. \end{aligned}$$

So with $a(t) = \frac{1}{\sqrt{2\log t}}$,

$$\frac{U(tx)-U(t)}{a(t)} \;\to\; \log x,$$

the $\xi = 0$ case of (2.31).

Thus, at least in these extremely simple cases, all three domains of attraction are covered by (2.31).

The importance of (2.31) is given by the following:

Theorem 2.3. Suppose F is a distribution function and that $U(t)$ is defined by (2.30). Then a necessary and sufficent condition for the existence of constants $a_n > 0$ and $b_n \in \mathbb{R}$ such that $F^n(a_n x + b_n) \to (1+\xi x)_+^{-1/\xi}$ as $n \to \infty$, is that (2.31) holds for a suitable function $a(t)$.

We shall not give the proof as this is one of many equivalent statements of Gnedenko's theorem, as reformulated by de Haan. One reference for this specific result is Theorem 1.1.6 of [105].

Suppose (2.31) holds. Consider the limit two ways of

$$\begin{aligned} \frac{U(txy)-U(tx)}{a(t)} &\to \frac{(xy)^\xi - x^\xi}{\xi} \\ \frac{U(txy)-U(tx)}{a(tx)} &\to \frac{y^\xi - 1}{\xi.} \end{aligned}$$

This is possible only if

$$\lim_{t\to\infty} \frac{a(tx)}{a(t)} = x^\xi \text{ for all } x > 0. \tag{2.32}$$

all three limits being as $y \to \infty$. Thus the function $a(t)$ is *regularly varying* with index ξ.

Definition. Let $a(t)$ be a positive function defined on $t > 0$. Then $a(\cdot)$ is said to be regularly varying of index ξ if (2.32) is valid for all $x > 0$.

The case $\xi = 0$ is called *slowly varying*.

2.4 Second-order Approximations

This relies on a remarkable theorem of de Haan and Stadtmüller [105]. Our treatment here relies heavily on Appendix B.3 of [101].

Suppose (2.31) holds. To derive a comprehensive theory for rates of convergence, we need to consider how to extend (2.31) to include second-order terms. Accordingly, consider limiting results of the form

$$\lim_{t\to\infty} \frac{\frac{U(tx)-U(t)}{a(t)} - \frac{x^\xi - 1}{\xi}}{A(t)} = H(x), \tag{2.33}$$

valid for all $x > 0$, where $A(t)$ is some rate function with $A(t) \to 0$ as $t \to \infty$, and $H(x)$ is some non-zero limit function. The problem is essentially this: if we assume that a limit of the form (2.33) exists, what can we deduce about the functions $A(t)$ and $H(x)$?

The first step is to exclude limits of the form $H(x) = c\frac{x^\xi - 1}{\xi}$ for some $c \neq 0$. In that case, we can rearrange (2.33) to give

$$\lim_{t\to\infty} \frac{\frac{U(tx)-U(t)}{a(t)(1+cA(t))} - \frac{x^\xi - 1}{\xi}}{A(t)} \quad = \quad 0,$$

but this clearly does not provide useful information about the rate of convergence in (2.31). Therefore, we exclude cases in which the limit function H is of the form $c\frac{x^\xi - 1}{\xi}$ for some $c \in \mathbb{R}$.

With this, the main theorem in [105] states:

Suppose U is a measurable function and functions $a(t) > 0$, $A(t) > 0$ hold such that the limit (2.33) holds for all $x > 0$ where $H(x)$ is not a multiple of $\frac{x^\xi - 1}{\xi}$. Then there exist real constants c_1, c_2 and $\rho \leq 0$ such that

$$\lim_{t\to\infty} \frac{\frac{U(tx)-U(t)}{a(t)} - \frac{x^\xi - 1}{\xi}}{A(t)} \quad = \quad c_1 \int_1^x s^{\xi-1} \int_1^s u^{\rho-1} du ds + c_2 \int_1^x s^{\xi+\rho-1} ds. \tag{2.34}$$

Moreover, if (2.34) holds, we also have,

$$\lim_{t\to\infty} \frac{\frac{a(tx)}{a(t)} - x^\xi}{A(t)} \quad = \quad c_1 x^\xi \frac{x^\rho - 1}{\rho}, \tag{2.35}$$

$$\lim_{t\to\infty} \frac{A(tx)}{A(t)} \quad = \quad x^\rho. \tag{2.36}$$

Moreover, $c_1 \neq 0$ if $\rho = 0$.

Remark B.3.5 of [101] notes that if (2.34) holds, we can also redefine the functions $a(t)$ and $A(t)$ so that $c_1 = 1$ and $c_2 = 0$. In that case, we get the specific form

$$\begin{aligned} H(t) \quad &= \quad \int_1^x s^{\xi-1} \int_1^s u^{\rho-1} du ds \\ &= \quad \begin{cases} \frac{1}{\rho}\left(\frac{x^{\xi+\rho}-1}{\xi+\rho} - \frac{x^\xi - 1}{\xi}\right) \text{ if } \rho < 0, \xi \neq 0, \\ \frac{1}{\xi}\left(x^\xi \log x - \frac{x^\xi - 1}{\xi}\right) \text{ if } \rho = 0,\ \xi \neq 0, \\ \frac{1}{\rho}\left(\frac{x^\rho - 1}{\rho} - \log x\right) \text{ if } \rho < 0,\ \xi = 0, \\ \frac{1}{2}(\log x)^2 \text{ if } \rho = \xi = 0. \end{cases} \end{aligned} \tag{2.37}$$

2.4.1 Examples

Let us first consider the case where $1 - F$ has the expansion (2.11).

In this case both F and its inverse are continuous, so $U(t) = y$ if $1 - F(y) = t^{-1}$. Applying (2.11), we need

$$t^{-1} \quad = \quad cy^{-\alpha} + dy^{-\alpha-\beta} + o(y^{-\alpha-\beta}).$$

It can be readily checked that this implies

$$y \;=\; (ct)^{1/\alpha}\left\{1+\frac{d}{\alpha}c^{-1-\beta/\alpha}t^{-\beta/\alpha}+o(t^{-\beta/\alpha})\right\}. \tag{2.38}$$

Therefore, $U(t)$ satisfies the right hand side of (2.38).

Hence,

$$\begin{aligned} U(tx)-U(t) &= (cxt)^{1/\alpha}\left\{1+\frac{d}{\alpha}c^{-1-\beta/\alpha}(xt)^{-\beta/\alpha}\right\}-(ct)^{1/\alpha}\left\{1+\frac{d}{\alpha}c^{-1-\beta/\alpha}t^{-\beta/\alpha}\right\}+o(t^{1/\alpha-\beta/\alpha}) \\ &= (ct)^{1/\alpha}(x^{1/\alpha}-1)+\frac{d}{\alpha}c^{1/\alpha-1-\beta/\alpha}t^{1/\alpha-\beta/\alpha}(x^{1/\alpha-\beta/\alpha}-1)+o(t^{1/\alpha-\beta/\alpha}). \end{aligned} \tag{2.39}$$

If we define $a(t)=\alpha^{-1}(ct)^{1/\alpha}$, we get

$$\frac{U(tx)-U(t)}{a(t)}-\frac{x^{1/\alpha}-1}{1/\alpha} \;=\; dc^{-1-\beta/\alpha}t^{-\beta/\alpha}(x^{1/\alpha-\beta/\alpha}-1)+o(t^{-\beta/\alpha}),$$

which, however, does not give the form of limit function we are aiming at.

Therefore, we return to (2.39) and rewrite

$$\begin{aligned} U(tx)-U(t) &= \left\{(ct)^{1/\alpha}+\frac{(1-\beta)d}{\alpha}c^{1/\alpha-1-\beta/\alpha}t^{1/\alpha-\beta/\alpha}\right\}(x^{1/\alpha}-1) \\ &+\frac{(1-\beta)d}{\alpha^2}c^{1/\alpha-1-\beta/\alpha}t^{1/\alpha-\beta/\alpha}\cdot\frac{\alpha}{1-\beta}\left\{x^{1/\alpha-\beta/\alpha}-1-(1-\beta)(x^{1/\alpha}-1)\right\}+o(t^{1/\alpha-\beta/\alpha}). \end{aligned}$$

Now define $a(t)=\alpha^{-1}\left\{(ct)^{1/\alpha}+\frac{\beta d}{\alpha}c^{1/\alpha-1-\beta/\alpha}t^{1/\alpha-\beta/\alpha}\right\}$, $A(t)=-\frac{(1-\beta)d}{\beta}c^{-1-\beta/\alpha}t^{-\beta/\alpha}$, then

$$\lim_{t\to\infty}\frac{\frac{U(tx)-U(t)}{a(t)}-\frac{x^{1/\alpha}-1}{(1/\alpha)}}{A(t)} \;=\; -\frac{\alpha}{\beta}\left(\frac{x^{1/\alpha-\beta/\alpha}-1}{1/\alpha-\beta/\alpha}-\frac{x^{1/\alpha}-1}{1/\alpha}\right).$$

This is precisely of the form (2.37) with $\xi=\frac{1}{\alpha}$, $\rho=-\beta/\alpha$.

2.5 Estimation theory based on second-order asymptotics

We focus here on a paper by Dombry and Ferreira [58], but this is just one of a series of papers going back to the 1980s [224, 61, 56, 74, 57, 174].

Consider an IID random sequence $\{X_i,\ i=1,2,\ldots\}$ where the common distribution function is F. Suppose the observations are grouped into blocks of length m, and let $M_{k,m}=\max\{X_i:(k-1)m+1,\ldots,km\}$ be the maximum of the k'th block. We assume F is in the domain of attraction of the GEV, so that

$$\Pr\left\{\frac{M_{k,m}-b_m}{a_m}\le x\right\} \;=\; F^m(a_mx+b_m) \;\to\; G_{\xi_0}(x) \;=\; \exp\left\{-(1+\xi_0x))_+^{-1/\xi_0}\right\}. \tag{2.40}$$

for some "true value" ξ_0 which we write that way to distinguish it from the unknown parameter ξ in the following likelihood analysis. We define $g_{\xi_0}(x) = \frac{dG_{\xi_0}(x)}{dx} = (1+\xi_0 x))^{-1/\xi_0-1} \exp\left\{-(1+\xi_0 x))^{-1/\xi_0}\right\}$ defined whenever $1+\xi_0 x > 0$ to be the density of G_{ξ_0} and let

$$\ell(\mu,\psi,\xi;x) \quad = \quad \log\psi + \log g_\xi\left(\frac{x-\mu}{\psi}\right) \tag{2.41}$$

be the log density for arbitrary ξ when the distribution is extended to include a location and scale parameter. The idea is that we treat the block maxima $M_{i,m}$ for $1 \le i \le k$ as if their exact distribution was GEV with parameters $\boldsymbol{\theta} = (\mu,\psi,\xi)$ though we know that for finite m this is only an approximation. Define the log likelihood

$$L_{k,m}(\boldsymbol{\theta}) \quad = \quad \sum_{i=1}^{k} \ell(\boldsymbol{\theta}, M_{i,m}) \tag{2.42}$$

In the following, we shall consider a sequence of sample sizes and block lengths $k_n,\ m_n$ where both k_n and M_n are indexed by n. We define $\hat{\boldsymbol{\theta}}_n = (\hat{\mu}_n, \hat{\psi}_n, \hat{\xi}_n)$ to be a local maximizer of the log likelihood function, or just the MLE for short, if it satisfies the likelihood equations

$$\frac{\partial L_{k,m}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \quad = \quad 0 \tag{2.43}$$

and if the hessian matrix $\frac{\partial^2 L_{k,m}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}$ is positive definite at $\hat{\boldsymbol{\theta}}_n$.

Dombry and Ferreira differ slightly from the notation of the previous section by defining $V = (-1/\log F)^{\leftarrow}$ (instead of $U = (1/(1-F))^{\leftarrow}$ as previously, though in most cases the two definitions will lead to the same asymptotics). In that context they assume, first, that there exists a_m such that

$$\lim_{m\to\infty} \frac{V(mx)-V(m)}{a_m} \quad = \quad \frac{x^{\xi_0}-1}{\xi_0} \tag{2.44}$$

and, second, that for some positive function $a(t)$ as $t\to\infty$ and some positive or negative function $A(t)$ as $t\to\infty$ with $\lim_{t\to\infty} A(t) = 0$,

$$\lim_{t\to\infty} \frac{\frac{V(tx)-V(t)}{a(t)} - \frac{x^{\xi_0}-1}{\xi_0}}{A(t)} \quad = \quad \int_1^x \int_1^s s^{\xi_0-1} u^{\rho-1} du ds \;=\; H_{\xi_0,\rho}(x),\ x>0, \tag{2.45}$$

where $\xi_0 > -\frac{1}{2},\ \rho \le 0$, the function A is regularly varying with index ρ, and $H_{\xi_0,\rho}$ is given by (2.37) with $\xi = \xi_0$. As noted previously, in any case where a limit of the form (2.45) exists, we can without loss of generality, redefining the functions $a(t)$ and $A(t)$ is necessary, assume that the right hand side is $H_{\xi_0,\rho}(x)$ for suitable $\rho \le 0$.

Dombry and Ferreira consider limiting cases as $k = k_n \to \infty,\ m = m_n \to \infty$ where

$$\lim_{n\to\infty} \sqrt{k_n} A(m_n) \quad = \quad \lambda \in \mathbb{R}. \tag{2.46}$$

They define $\boldsymbol{\theta}_0 = (0,1,\xi_0)$ and then

$$\begin{aligned} Q_{\xi_0}(s) &= \frac{(-\log s)^{-\xi_0}-1}{\xi_0},\ s\in(0,1) \\ \mathbf{b}(\xi_0,\rho) &= \int_0^1 \frac{\partial^2 \ell}{\partial x \partial \boldsymbol{\theta}}(\boldsymbol{\theta}_0, Q_{\xi_0}(s)) H_{\xi_0,\rho}\left(\frac{1}{-\log s}\right) ds, \\ I_{\xi_0} &= -\int_0^1 \frac{\partial^2 \ell}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}(\boldsymbol{\theta}_0, Q_{\xi_0}(s)) ds. \end{aligned}$$

Note that I_{ξ_0} is the Fisher information for the GEV evaluated at $\boldsymbol{\theta}_0$; this is the same matrix as was shown in Chapter 1 following [185].

With these preliminaries, Theorem 2.2 of [58] states:

(a) There exists a sequence of estimators $\hat{\boldsymbol{\theta}}_n = \hat{\mu}_n, \hat{\psi}_n, \hat{\xi}_n$ such that

$$\begin{aligned} \lim_{n\to\infty} \Pr\left\{\hat{\boldsymbol{\theta}}_n \text{ is a MLE}\right\} &= 1, \\ \sqrt{k_n}\left(\frac{\hat{\mu}_n - b_{m_n}}{a_{m_n}}, \frac{\hat{\psi}_n}{a_{m_n}} - 1, \hat{\xi}_n - \xi_0\right) &\xrightarrow{d} \mathcal{N}\left(\lambda I_{\xi_0}^{-1}\mathbf{b}, I_{\xi_0}^{-1}\right). \end{aligned}$$

(b) If $\hat{\boldsymbol{\theta}}_n^i = (\hat{\mu}_n^i, \hat{\psi}_n^i, \hat{\xi}_n^i),\ i = 1,2$ are two sequences of estimators satisfying

$$\begin{aligned} \lim_{n\to\infty} \Pr\left\{\hat{\boldsymbol{\theta}}_n^i \text{ is a MLE}\right\} &= 1, \\ \lim_{n\to\infty} \Pr\left\{\sqrt{k_n}\left(\frac{\hat{\mu}_n^i - b_{m_n}}{a_{m_n}}, \frac{\hat{\psi}_n^i}{a_{m_n}} - 1, \hat{\xi}_n^i - \xi_0\right) \in H_n\right\} &= 1, \end{aligned}$$

where H_n is a ball in $\mathbb{R}^3$ of center 0 and radius r_n, where $r_n = O(k_n^\delta),\ 0 < \delta < \min\left(\frac{1}{2}, \xi_0 + \frac{1}{2}\right)$ as $n \to \infty$, then

$$\lim_{n\to\infty} \Pr\left\{\hat{\boldsymbol{\theta}}_n^1 = \hat{\boldsymbol{\theta}}_n^2\right\} = 1.$$

2.5.1 Side Section 1: A heuristic on biased estimation

Suppose we have a sequence of experiments indexed by n, where in the nth experiment there are k_n observations $X_1, \ldots, X_{k_n}$ whose true joint density is g_n, but for reasons of convenience or because we don't know how to exactly calculate g_n, we replace g_n by a known joint density f_n indexed by a parameter vector $\boldsymbol{\theta}_n$. The examples of interest to us include the X_i's being either block maxima or exceedances over a threshold and their density f_n being approximated by a GEV or GPD density. We will always want $f_n - g_n \to 0$ under some suitable metric (e.g. total variation norm or Hellinger distance) but we won't worry about precise modes of convergence for the moment — that can come later.

Suppose we estimate $\boldsymbol{\theta}$ by defining a set of equations

$$\sum_{i=1}^{k_n} \mathbf{T}(X_i; \boldsymbol{\theta}) = 0$$

where $\mathbf{T}(X_i;\boldsymbol{\theta})$ is a vector of the same length as $\boldsymbol{\theta}$ that form a set of *unbiased estimating equations* in the sense that

$$\mathrm{E}\{\mathbf{T}(X_i;\boldsymbol{\theta})\} \quad = \quad 0 \text{ when } X_i \sim f_n(\cdot\,;\,\boldsymbol{\theta}).$$

The classical case is when $\mathbf{T}$ is the vector of first-order derivatives of the log likelihood but we are writing the formula in this alternative format to allow for other possible estimators (in particular, in the case of extreme value theory, probability weighted moments estimators or PWMs, which are a popular alternative to maximum likelihood estimation).

We also define a matrix $W(X_i)$ with entries $w_{rs}(X_i) = \frac{\partial T_r(X_i)}{\partial \theta_s}$ where T_r is the rth component of T and θ_s is the sth component of $\boldsymbol{\theta}$. In standard maximum likelihood theory, W is the hessian matrix of the log likelihood function (for a single observation), also known as the observed information matrix, and the expectation of W when $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ is I_0, the Fisher information matrix assuming the model f_n is correct with parameter vector $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

Assuming suitable regularity conditions,

$$\begin{aligned} \mathbf{0} &= \sum_{i=1}^{k_n} \mathbf{T}(X_i;\hat{\boldsymbol{\theta}}_n) \\ &\approx \sum_{i=1}^{k_n} \mathbf{T}(X_i;\boldsymbol{\theta}_0) + W(X_i;\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \end{aligned}$$

and hence

$$\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 \quad \approx \quad -\left\{\sum_{i=1}^{k_n} W(X_i;\boldsymbol{\theta}_0)\right\}^{-1} \left\{\sum_{i=1}^{k_n} \mathbf{T}(X_i;\boldsymbol{\theta}_0)\right\}. \tag{2.47}$$

If we assume

(i) The mean of $W(X_i;\boldsymbol{\theta}_0)$ is J_0 for each i,

(ii) The covariance matrix of $\mathbf{T}(X_i;\boldsymbol{\theta}_0)$ is C_0 for each i,

and assume f_n is the true density, we will have

$$\sqrt{k_n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \quad \xrightarrow{d} \quad \mathcal{N}(\mathbf{0}, J_0^{-1} C_0 J_0^{-1}). \tag{2.48}$$

Formula (2.48) is widely known as the *information sandwich formula*. When estimation is by maximum likelihood, J_0 and C_0 both reduce to I_0, the Fisher information matrix, and (2.48) is the standard asymptotic distribution for maximum likelihood estimators.

Now, however, suppose the true density is g_n rather than f_n. *Typically*, the following is true: the covariance matrix of $\sum_{i=1}^{k_n} \mathbf{T}(X_i;\boldsymbol{\theta}_0)$ and the mean of $\sum_{i=1}^{k_n} W(X_i;\boldsymbol{\theta}_0)$ are still asymptotic to $k_n C_0(\boldsymbol{\theta})$ and $k_n J_0(\boldsymbol{\theta})$ respectively, but the mean of $\sum_{i=1}^{k_n} \mathbf{T}(X_i;\boldsymbol{\theta}_0)$ is non-zero. To be precise the mean is $\mathbf{b}_n$. In that case, the CLT for

$\sum_{i=1}^{k_n} \mathbf{T}(X_i;\boldsymbol{\theta}_0)$ takes the form

$$k_n{}^{-1/2}\sum_{i=1}^{k_n}\mathbf{T}(X_i;\boldsymbol{\theta}_0) \quad\sim\quad \mathcal{N}[k_n{}^{-1/2}\mathbf{b}_n, C_0(\boldsymbol{\theta}_0)](1+o_p(1))$$

and the final result for $\hat{\boldsymbol{\theta}}_n$ becomes

$$\sqrt{k_n}(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0)+k_n{}^{-1/2}J_0^{-1}\mathbf{b}_n \quad\xrightarrow{d}\quad \mathcal{N}[0,J_0^{-1}C_0J_0^{-1}]. \tag{2.49}$$

Note that there are different special cases of this result depending on the asymptotic behavior of $k_n{}^{-1/2}\mathbf{b}_n$. If $k_n{}^{-1/2}\mathbf{b}_n \to 0$ then the asymptotic bias of $\hat{\boldsymbol{\theta}}_n$ is negligible compared with its statistical variability as represented by the Fisher information matrix. In effect, this means we can ignore the discrepancy between f_n and g_n. Conversely, if $n^{-1/2}\mathbf{b}_n \to \infty$ in at least one component, the bias dominates the variance, which has the practical interpretation that we can't really use the standard results in this case. However if $k_n{}^{-1/2}\mathbf{b}_n \to \mathbf{c}$ for some vector $\mathbf{c}$ whose components are finite and not all zero, we can rewrite the result (2.49) as

$$\sqrt{k_n}(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0) \quad\xrightarrow{d}\quad \mathcal{N}[-J_0^{-1}\mathbf{c},J_0^{-1}C_0J_0^{-1}]. \tag{2.50}$$

This is a true case of "bias-variance tradeoff" which can be the basis for various decision processes, such as the choice of a threshold in a peaks over threshold analysis (the ultimate objective of [224]).

2.5.2 Side section 2: Asymptotics of the Hill-Weissman Estimator

In this section we consider the special case of extreme value theory based on the Type I or Fréchet limit. Gnedenko [87] showed that a limit of the form

$$F^n(a_nx) \quad\to\quad \Phi_\alpha(x) \;=\; \exp\left(-cx^{-\alpha}\right),\; x\geq 0,\; \alpha>0,\; c>0, \tag{2.51}$$

holds if $1-F(x)$ is regularly varying with index α, and in that case a_n may without loss of generality be taken as the solution of $F(a_n)=1-1/n$, and $c=1$. Note that in this case, there is no location parameter to the distribution ($b_n=0$), but for statistical purposes, it makes sense to retain c as well as α as an unknown parameter.

In this case, Weissman's representation [255, 257] for the asymptotic joint distribution of the k largest order statistics $m_1 \geq m_2 \geq \ldots m_k$ reduces to

$$L(\alpha,c\mid m_1,...,m_k) \quad=\quad \prod_{i=1}^{k}\left(c\alpha m_i^{-\alpha-1}\right)\cdot\exp\left(-cm_k^{-\alpha}\right). \tag{2.52}$$

where the notation is intended to indicate that we are thinking of (2.52) as a likelihood function for the parameters α and c. The dependence on $m_1,...,m_k$ will be omitted in many of the formulas. Taking logarithms, we want to minimize

$$\ell(\alpha,c) \;=\; -\log L(\alpha,c) \;=\; -k\log\alpha-k\log c+(\alpha-1)\sum_{i=1}^{k}\log m_i+cm_k^{-\alpha}.$$

It is quickly established that this expression is minimized when $\alpha = \hat{\alpha},\ c = \hat{c}$ where

$$\hat{\alpha} \ = \ \left(\frac{1}{k} \sum_{i=1}^{k} \log \frac{m_i}{m_k} \right)^{-1}, \ \hat{c} \ = \ k m_k^{\hat{\alpha}}. \tag{2.53}$$

Note, in particular, the simple direct formula for the estimator of α. The derivation is the same as that in [255], but that paper did it for the equivalent case where the limit distribution is Gumbel (the Fréchet model is turned into the Gumbel model by taking logarithms of the observations).

An alternative, even simpler, derivation of an equivalent result was given by Hill [115]. Hill assumed, in effect, that the relationship $1 - F(x) = cx^{-\alpha}$ is exact for $x \geq u$, for some known threshold u, but that $F(x)$ is unspecified for $x < u$. If data $X_1, ..., X_n$ are ordered so that $X_1 \geq X_2 \geq \ldots X_k > u \geq X_{k+1} \geq \ldots X_n$ then the likelihood function is

$$L(\alpha, c \mid X_1, ..., X_n) \quad = \quad \prod_{i=1}^{k} \left(\alpha c X_i^{-\alpha-1} \right) \cdot \left(1 - c u^{-\alpha} \right)^{n-k}$$

Taking logarithms and minimizing with respect to first c and then α leads to

$$\hat{\alpha} \ = \ \left(\frac{1}{k} \sum_{i=1}^{k} \log \frac{X_i}{u} \right)^{-1}, \ \hat{c} \ = \ \frac{k}{n} u^{\hat{\alpha}}. \tag{2.54}$$

Note, in particular, the similarity of the two estimators of α: in effect, the role of the threshold u in (2.54) is replaced by the kth largest order statistic in (2.53). (The different estimators of c arise because of different definitions: (2.53) uses the limit distribution for sample maxima whereas (2.54) assumes the same functional form directly for the individual observations. The two definitions differ by a factor of n, which is reflected in the estimates.)

The estimator $\hat{\alpha}$ in (2.54) is widely known as *Hill's estimator* but in the present section, to emphasize the close similarity with Weissman's [257] result, we shall call it the *Hill-Weissman estimator*.

In order to develop some asymptotics for this estimator, we assume an expansion of the form

$$1 - F(x) \quad = \quad cx^{-\alpha} \left\{ 1 + dx^{-\beta} + o(x^{-\beta}) \right\}, \ x \to \infty. \tag{2.55}$$

In general, the assumption (2.55) may be replaced by an assumption of *second-order regular variation* which allows the terms with $x^{-\alpha}$ and $x^{-\beta}$ to be replaced by general regularly varying functions; see in particular [89] for a survey of this theory and its applications (including the present one). This, in turn, is a special case of the general sceond-order regular variation theory of [105], For the present discussion, we mke the simpler assumption (2.55) which is sufficient for most practical applications, and easier to manipulate.

Our focus will be on the condition distribution of X given $X > u$, for some high

threshold u. Let $Y_u = X/u$. Then the conditional probability $P\{Y_u > y \mid Y_u > 1\}$ is represented as

$$\frac{1-F(uy)}{1-F(u)} \quad = \quad y^{-\alpha}\left\{1+du^{-\beta}(y^{-\beta}-1)+o(u^{-\beta})\right\}$$

so, assuming it is valid to differentiate term by term, we calculate the density as

$$f_{Y_u}(y) \quad = \quad \alpha y^{-\alpha-1}+du^{-\beta}\left\{(\alpha+\beta)y^{-\alpha-\beta-1}-\alpha y^{-\alpha-1}\right\}+o(u^{-\beta}).$$

We note integrals of the form

$$\int_1^\infty (\log y)^k y^{-\alpha-1}dy \quad = \quad \alpha^{-k-1}k!$$

where we shall mainly be interested in the cases $k=1$ and 2 but for non-integer k the same formula holds with $k!$ replaced by $\Gamma(k+1)$. We therefore deduce

$$\mathrm{E}(\log Y_u)^k \quad = \quad \alpha^{-k}k!+du^{-\beta}k!\left\{(\alpha+\beta)^{-k}-\alpha^{-k}\right\}=o(u^{-\beta}). \qquad (2.56)$$

Now let's consider the bias and variance of $\frac{1}{\hat{\alpha}} = \frac{1}{k}\sum_{i=1}^k \log\frac{X_i}{u}$ as an estimator of $\frac{1}{\alpha}$, where k is the number of exceedances of u. Since $\mathrm{E}(\log Y_u) = \frac{1}{\alpha} - du^{-\beta}\frac{\beta}{\alpha(\alpha+\beta)} + o(u^{-\beta})$, we deduce

$$\text{Bias of } \frac{1}{\hat{\alpha}} \quad \approx \quad -du^{-\beta}\frac{\beta}{\alpha(\alpha+\beta)}.$$

However, we also have from the $k=1$ and $k=2$ cases of (2.56) that $\mathrm{Var}(\log Y_u) \to \frac{1}{\alpha^2}$ as $u\to\infty$ and hence the variance of $\frac{1}{\alpha}$ is asymptotically $\frac{1}{k\alpha^2}$. However if the whole sample is of size n, and k is the random number of exceedances of u, we have $k \sim ncu^{-\alpha}$. Therefore, in large samples we have

$$\text{Variance of } \frac{1}{\hat{\alpha}} \quad \approx \quad \frac{1}{\alpha^2 ncu^{-\alpha}}.$$

Combining the espressions for bias and variance, and writing mean squared error (MSE) for the sum of squared bias and variance, we deduce

$$\text{MSE of } \frac{1}{\hat{\alpha}} \quad \approx \quad \frac{Au^\alpha}{n}+B^2u^{-2\beta}$$

where $A = \frac{1}{\alpha^2 c}$ and $B = \frac{d\beta}{\alpha(\alpha+\beta)}$.

This asymptotic MSE is minimized with

$$u \quad = \quad \left(\frac{2\beta B^2 n}{\alpha A}\right)^{1/(\alpha+2\beta)}$$

which in turn leads to an asymptotic MSE of

$$MSE \quad = \quad \frac{B^2(\alpha+2\beta)}{\alpha}\left(\frac{2\beta B^2 n}{\alpha A}\right)^{-2\beta/(\alpha+2\beta)}.$$

The most important consequence of this is that the MSE is of $O\left(n^{-2\beta/(\alpha+2\beta)}\right)$ as $n \to \infty$, which could be arbitrarily slow for very small β but is of $O(n^{-1})$ as $\beta \to \infty$ — this makes sense, because in thaty limit the $cx^{-\alpha}$ result is exact and we are back in the original case considered by Hill.

2.5.2.1 Extension to the GPD

The above calculation was relatively straightforward because of the explicit closed form of the estimator. In most cases of interest (for example, estimating the two-parameter GPD or the three-parameter GEV distribution), there is no closed form estimator and the MLE is obtained by solving the likelihood equations. In such case, we may in principle proceed as follows. Suppose the negative log likelihood function based on n observations is $\ell_n(\theta)$ for some multidimensional parameter θ whose true value we shall write θ_0. Also write $\hat{\theta}_n$ for the MLE. The Taylor expansion

$$\nabla\ell_n(\hat{\theta}_n) - \nabla\ell_n(\theta_0) \quad \approx \quad \left(\hat{\theta}_n - \theta_0\right)^T \nabla^2\ell_n(\theta_0)$$

leads to the approximation

$$\hat{\theta}_n - \theta_0 \quad \approx \quad -\left(\nabla^2\ell_n(\theta_0)\right)^{-1}\nabla\ell_n(\theta_0).$$

Now suppose that as $n \to \infty$, $n^{-1}\nabla^2\ell_n(\theta_0) \xrightarrow{p} J$ (the Fisher information matrix) and $n^{-1}\nabla\ell_n(\theta_0) \xrightarrow{p} \mathbf{b}$ (bias due to model misspecification; if the model is correctly specified, $\mathbf{b} = \mathbf{0}$). Then for $\hat{\theta}_n$ we have, for large n,

$$\text{Bias } \approx J^{-1}\mathbf{b}, \text{ Covariance Matrix } \approx n^{-1}J^{-1}. \tag{2.57}$$

Now let's apply this to the case of the GPD, again under the assumption that the true distribution satisfies (2.55). Note that in the case where $1-F(x) = cx^{-\alpha}$ is exact, we have

$$\frac{1-F(u+y)}{1-F(u)} \quad = \quad \left(1+\frac{y}{u}\right)^{-\alpha} \quad = \quad \left(1+\xi\frac{y}{\sigma}\right)^{-1/\xi}$$

so the two forms are identical if $\sigma = \frac{u}{\alpha}$, $\xi = \frac{1}{\alpha}$. From now on, we treat these as the "true" GPD parameter values in this case.

In this model, the Fisher information matrix [224] is

$$J \quad = \quad \begin{pmatrix} \frac{1}{\sigma^2(1+2\xi)} & \frac{1}{\sigma(1+\xi)(1+2\xi)} \\ \frac{1}{\sigma(1+\xi)(1+2\xi)} & \frac{2}{(1+\xi)(1+2\xi)} \end{pmatrix}$$

provided $1+2\xi > 0$, and hence

$$J^{-1} = (1+\xi)\begin{pmatrix} 2\sigma^2 & -\sigma \\ -\sigma & (1+\xi) \end{pmatrix}$$

Now let's compute the **b** term in (2.57). The log likelihood for a single observation is

$$\ell(\sigma,\xi) = \log\sigma + \left(\frac{1}{\xi}+1\right)\log\left(1+\xi\frac{y}{\sigma}\right).$$

Hence,

$$\begin{aligned} \sigma\frac{\partial\ell}{\partial\sigma} &= -\frac{1}{\xi}+\left(\frac{1}{\xi}+1\right)\left(1+\xi\frac{y}{\sigma}\right)^{-1}, \\ \frac{\partial\ell}{\partial\xi} &= -\frac{1}{\xi^2}\log\left(1+\xi\frac{y}{\sigma}\right)+\frac{1}{\xi}\left(\frac{1}{\xi}+1\right)\left\{1-\left(1+\xi\frac{y}{\sigma}\right)^{-1}\right\}. \end{aligned}$$

To calculate **b**, we need to find expressions for the expected values of these terms.

To recast in the notation of Section 2.5.2, we first make the substitutions $\sigma = \frac{u}{\alpha}$, $\xi = \frac{1}{\alpha}$, and also that if y denotes the excess over the threshold u, then $y = u(Y_u - 1)$ and so $1+\xi\frac{y}{\sigma} = Y_u$. Also, by the same reasoning as led to (2.56)

$$\mathrm{E}\left(Y_u^{-1}\right) = \frac{\alpha}{\alpha+1} + du^{-\beta}\cdot\frac{\beta}{(\alpha+1)(\alpha+\beta+1)} + o(u^{-\beta}).$$

We now calculate the expectations of $\sigma\frac{\partial\ell}{\partial\sigma}$ and $\frac{\partial\ell}{\partial\xi}$, respectively, to be

$$-\alpha+(\alpha+1)\left\{\frac{\alpha}{\alpha+1}+du^{-\beta}\cdot\frac{\beta}{(\alpha+1)(\alpha+\beta+1)}+o(u^{-\beta})\right\} = du^{-\beta}\cdot\frac{\beta}{\alpha+\beta+1}+o(u^{-\beta})$$

and

$$\begin{aligned} &-\alpha^2\left\{\frac{1}{\alpha}-du^{-\beta}\frac{\beta}{\alpha(\alpha+\beta)}\right\}+\alpha(\alpha+1)\left\{\frac{1}{\alpha+1}-du^{-\beta}\frac{\beta}{(\alpha+1)(\alpha+\beta+1)}\right\}+o(u^{-\beta}) \\ = \;& du^{-\beta}\cdot\frac{\alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)}+o(u^{-\beta}). \end{aligned}$$

Therefore, we conclude

$$\begin{aligned} \mathbf{b} &\sim du^{-\beta}\begin{pmatrix} \frac{1}{\sigma}\frac{\beta}{\alpha+\beta+1} \\ \frac{\alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)} \end{pmatrix}, \\ J^{-1}\mathbf{b} &\sim du^{-\beta}\frac{(\alpha+1)\beta}{\alpha(\alpha+\beta)(\alpha+\beta+1)}\begin{pmatrix} \sigma(\alpha+2\beta) \\ 1-\beta \end{pmatrix}. \end{aligned}$$

Focussing on the second entries in these vectors, we deduce that $\hat{\xi}$ has asymptotic bias

$$du^{-\beta}\frac{(\alpha+1)\beta(1-\beta)}{\alpha(\alpha+\beta)(\alpha+\beta+1)}$$

and asymptotic variance (based on $k \approx ncu^{-\alpha}$ exceedances of the threshold

$$\frac{1}{k}\left(\frac{\alpha+1}{\alpha}\right)^2 \sim \frac{(\alpha+1)^2}{\alpha^2 ncu^{-\alpha}}.$$

2.5.2.2 Comparisons with the Hill-Weissman Estimator

For the Hill-Weissman estimator, we deduced that the bias was asymptotically $Bu^{-\beta}$, variance Au^{α}/n, with $B = -d\beta/(\alpha(\alpha+\beta))$, $A = 1/(\alpha^2 c)$.

For the GPD estimator, we get asymptotic bias $B'u^{-\beta}$, asymptotic variance $A'u^{\alpha}/n$, where $B' = d\beta(1-\beta)(\alpha+1)/(\alpha(\alpha+\beta)(\alpha+\beta+1))$.

The optimal MSE is proportional to

$$|B|^{2\alpha/(\alpha+2\beta)} A^{2\beta/(\alpha+2\beta)}$$

Therefore, the ratio of the optimal MSE for the GPD estimator to that of the Hill-Weissman estimator is

$$\left|\frac{B'}{B}\right|^{2\alpha/(\alpha+2\beta)} \left|\frac{A'}{A}\right|^{2\beta/(\alpha+2\beta)} = \left|\frac{(1-\beta)(\alpha+1)}{\alpha(\alpha+\beta)(\alpha+\beta+1)}\right|^{2\alpha/(\alpha+2\beta)} |\alpha+1|^{4\beta/(\alpha+2\beta)}$$

See Figure 5.1.

2.5.2.3 Background References

The Hill estimator was introduced in [115] and the Weissman estimator, in its original form, in [257]. Asymptotic properties of the Hill estimator were obtained by [111, 107, 89] Optimality of the derived rate of convergence was proved by [109], and an adaptive estimator to achieve the optimal threshold was given by [110]. Many variants on the method exists, for example, [44] used a kernel-weighted version. The comparison of the two estimators was first derived in [224]. Many other authors have contributed to the theory and a more complete bibliography will be given later.

2.5.3 Outline Derivation of Dombry-Ferreira result

Health warning: This is not the proof. For that, we refer to the original paper [58]. The intention here is to motivate the result, and to show how it follows logically from the asymptotic approximations we have been developing in this chapter.

First, let us assume that the relationship (2.44) is exact, i.e. the left and right hand sides are identical for every m. Since $M_{i,m}$ has the distribution function F^m, by the probability integral transformation we can write $F^m(M_{i,m}) = S$ where S is uniform on (0,1). In that case $-\frac{1}{\log F(M_{i,m})} = \frac{m}{-\log S}$. But $-\frac{1}{\log F(\cdot)}$ was defined to be the inverse of V, so $M_{i,m} = V\left(\frac{m}{-\log S}\right)$. We also define $b_m = V(m),\ a_m = a(m)$. If we assume (2.44 is exact, then

$$\frac{M_{i,m} - b_m}{a_m} = \frac{\left(-\frac{1}{\log S}\right)^{\xi_0} - 1}{\xi_0}.$$

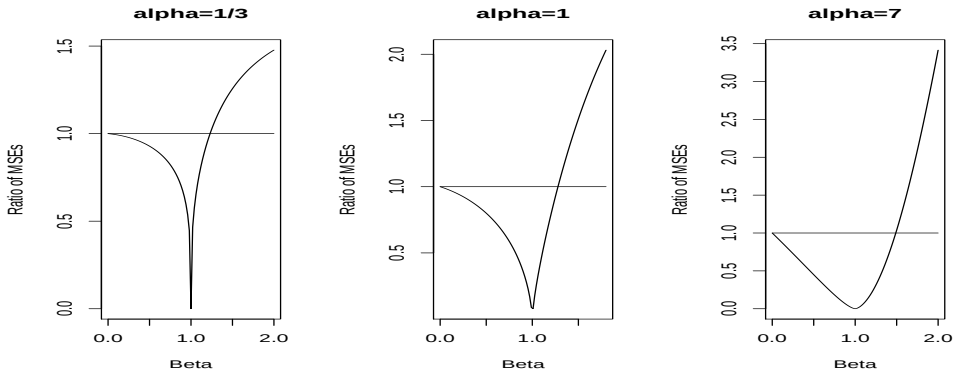

Figure 2.1 *Ratio of optimal mean squared error for the GPD estimator to that of the Hill-Weissman estimator, for a variety of values of* α *and* β.

But the right hand side has the GEV distribution:

$$\Pr\left\{\frac{\left(-\frac{1}{\log S}\right)^{\xi_0}-1}{\xi_0}\le y\right\} \;=\; \Pr\left\{S\le e^{-(1+\xi y)^{-1/\xi}}\right\} \;=\; e^{-(1+\xi y)^{-1/\xi}} \text{(provided } 1+\xi y>0).$$

Now, however, suppose (2.45) holds instead of (2.44) being exact. In that case, we can write

$$\frac{M_{i,m}-b_m}{a_m} \;=\; \frac{\left(-\frac{1}{\log S}\right)^{\xi_0}-1}{\xi_0}+A(b_m)H_{\xi_0,\rho}\left(\frac{1}{-\log S}\right)+o_p(A(b_m)).$$

Suppose we want to find the expectation of $h\left(\frac{M_{i,m}-b_m}{a_m}\right)$, where h is some nonlinear continuously differentiable function. We proceed formally, assuming limiting operations are valid without rigorous proof. By Taylor expansion, we write

$$h\left(\frac{M_{i,m}-b_m}{a_m}\right) \;=\; h\left\{\frac{\left(-\frac{1}{\log S}\right)^{\xi_0}-1}{\xi_0}\right\}+A(b_m)H_{\xi_0,\rho}\left(\frac{1}{-\log S}\right)h'\left\{\frac{\left(-\frac{1}{\log S}\right)^{\xi_0}-1}{\xi_0}\right\}+o_p(A(b_m)).$$

Taking expectations term by term

$$\mathrm{E}\left\{h\left(\frac{M_{i,m}-b_m}{a_m}\right)\right\} \;=\; \int_0^1 h\left\{\frac{\left(-\frac{1}{\log s}\right)^{\xi_0}-1}{\xi_0}\right\}ds+\int_0^1 A(b_m)H_{\xi_0,\rho}\left(\frac{1}{-\log s}\right)h'\left\{\frac{\left(-\frac{1}{\log s}\right)^{\xi_0}-1}{\xi_0}\right\}ds$$
$$+o_p(A(b_m)).$$

Now suppose the function h is any of $\frac{d\ell}{d\mu}$, $\frac{d\ell}{d\psi}$, $\frac{d\ell}{d\xi}$, where ℓ is given by (2.41). Because

h is a derivative of the log likelihood of the GEV model, $\int_0^1 h \left\{ \frac{\left(-\frac{1}{\log s}\right)^{\xi_0} - 1}{\xi_0} \right\} ds = 0$ and we are left with

$$\mathrm{E}\left\{ h\left(\frac{M_{i,m} - b_m}{a_m} \right) \right\} \quad \sim \quad A(b_m) \int_0^1 H_{\xi_0,\rho}\left(\frac{1}{-\log s} \right) \frac{\partial h}{\partial x} \left\{ \frac{\left(-\frac{1}{\log s}\right)^{\xi_0} - 1}{\xi_0} \right\} ds.$$

Representing $\ell(\mu, \psi, \xi; x) = \ell(\boldsymbol{\theta}, x)$ where $\boldsymbol{\theta} = (\ \theta_1 \quad \theta_2 \quad \theta_3$ and $\theta_1 = \mu,\ \theta_2 = \psi,\ \theta_3 = \xi$, we therefore have

$$\mathrm{E}\left\{ \frac{\partial \ell}{\partial \boldsymbol{\theta}} \left(\frac{M_{i,m} - b_m}{a_m} \right) \right\} \quad \sim \quad A(b_m) \int_0^1 H_{\xi_0,\rho}\left(\frac{1}{-\log s} \right) \frac{\partial^2 \ell}{\partial x \partial \boldsymbol{\theta}} \left\{ \frac{\left(-\frac{1}{\log s}\right)^{\xi_0} - 1}{\xi_0} \right\} ds. \tag{2.58}$$

The right hand side of (2.58) is $A(m_n)$ multiplied by $\mathbf{b}(\xi_0, \rho)$ in the notation of Dombry–Ferreira.

Equation (2.58) applies to just a single value of the likelihood function, whereas the formula (2.42) represents the sum of $k = k_n$ similar terms. In the notation of Section 2.5.1, we have $\mathbf{b}_n = k_n A(m_n)\mathbf{b}(\xi_0, \rho)$ and hence $k_n^{-1/2}\mathbf{b}_n = \sqrt{k_n} A(m_n)\mathbf{b}(\xi_0, \rho) \to \lambda \mathbf{b}(\xi_0, \rho)$. Since this case the method of estimation under the GEV model is maximum likelihood, in this case the matrices J_0 and C_0 of Section 2.5.1 are both I_0, the Fisher information matrix for the limiting GEV distribution. Thus, (2.50) implies

$$\sqrt{k_n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 \right) \quad \xrightarrow{d} \quad \mathcal{N}\left\{ \lambda I_0^{-1} \mathbf{b}(\xi_0, \rho), I_0^{-1} \right\}$$

The Dombry-Ferreira result differs from this because it assumes the GEV maximum likelihood estimation procedure is applied directly to the block maxima $M_{i,m}$, rather than the normalized maxima $\frac{M_{i,m} - b_m}{a_m}$ as we have written here. Nevertheless, this argument should serve to motivate their result and to define a general context to derive similar results under different variations of the basic model and estimation procedure.

2.6 Other topics to be added

2.6.1 Method of probability weighted moments

An alternative to the maximum likelihood method that achieved popularity after a famous paper of Hosking, Wallis and Wood [118], but theoretically do not perform as well as maximum likelihood estimators [56, 72].

2.6.2 Corresponding results for threshold estimators

Cite paper of Smith [224]; show how results may be reinterpreted in terms of the de Haan-Stadtmüller representation

2.6.3 Estimating probabilities of extreme events

In most cases the real interest is not in estimating the GEV/GPD parameters but in applying them to estimate probabilities of extreme events or, equivalently, return values corresponding to given extreme probabilities. We will review results where, in addition to the kinds of asymptotics ocnsidered here, we also have tail probabilities $p_n \to 0$ as $n \to \infty$.

2.6.4 Adaptive choice of block size or threshold

2.6.5 Practical examples

e.g. [2]

Chapter 3

Extremes in Dependent Sequences

3.1 Extremes in stationary sequences

The random sequence $\{X_n\}$, where the index n ranges over all non-negative integers or over all integers, is called *stationary* if, for any n, $k \geq 1$ and $x_1, ..., x_k$,

$$\Pr\{X_{n+1} \leq x, ..., X_{n+k} \leq x_k\} \quad = \quad \Pr\{X_1 \leq x, ..., X_k \leq x_k\}. \tag{3.1}$$

It follows that, if all the means and covariances of the process exist, then they satisfy

$$\mathrm{E}\{X_n\} = \mu, \ \mathrm{Cov}\{X_n, X_{n+k}\} = \rho_k \ \text{ independent of } n. \tag{3.2}$$

A process that satisfies (3.2) without (3.1) is called *weakly* or *second-order* stationary. In much of time series analysis and linear prediction theory, second-order stationarity suffices, but in extreme value theory, except for the special case of Gaussian processes (when the two concepts coincide), we do need full stationarity of the process. Therefore, from now on, by stationarity we shall mean (3.1).

A stationary sequence is *strong mixing* if there exists a function $g(k), \ k = 1, 2, \ldots$, such that $g(k) \to 0$ as $k \to \infty$ and

$$|\Pr\{AB\} - \Pr\{A\}\Pr\{B\}| \leq g(k) \tag{3.3}$$

whenever $A \in \mathscr{F}_{-\infty}^{n}$ $B \in \mathscr{F}_{n+k+1}^{\infty}$ for some n. Here $\mathscr{F}_{m}^{n}$ for $-\infty \leq m \leq n \leq \infty$ denotes the σ-algebra generated by $\{X_j, \ m \leq j \leq n\}$, i.e. the set of all events determined by the random variables $\{X_j, \ m \leq j \leq n\}$. If there exists $m \geq 0$ such that $g(k) = 0$ for all $k > m$, then the sequence is said to be *m-dependent*.

One useful concept in talking about a stationary sequence is that of an *associated independent sequence*, i.e. a sequence $\{\hat{X}_n\}$ of independent random variables with the same marginal distribution as $\{X_n\}$. Let

$$M_n = \max\{X_1, \ldots, X_n\}, \ \hat{M}_n = \max\{\hat{X}_1, \ldots, \hat{X}_n\}.$$

Much of the theory concerns asymptotic results of the form

$$\Pr\{M_n \leq u_n\} - \Pr\{\hat{M}_n \leq u_n\} \to 0 \ \text{ as } n \to \infty \tag{3.4}$$

for suitable increasing $\{u_n\}$. Early results of this form were those of Watson [252] for

m-dependent sequences and Loynes [149] for strong mixing sequences, in each case under an additional condition which essentially ensures that high-level exceedances of the process occur in isolation rather than as dependent clusters. In the case of Gaussian sequences, Berman [17] proved (3.4) under the condition $r_n \log n \to 0$. This is not strictly speaking a necessary condition — indeed, Berman also gave the alternative sufficient condition $\sum r_n^2 < \infty$ — but it is almost necessary in the sense that if $r_n \log n \to c \neq 0$ then a different limit arises. Moreover, in the case of a Gaussian sequence, this is much weaker condition than strong mixing, so it became clear that Loynes' conditions were not the best possible. The problem of unifying these different conditions was essentially solved by Leadbetter [140, 141] who showed that a significantly weaker form of mixing condition covered the results of both Loynes and Berman. The description which follows is based on [143].

For any sequence of real values $\{u_n\}$, *Condition* $D(u_n)$ is said to hold if, for any integers $i_1 < ... < i_p < j_1 < ..., j_q$ with $j_1 - i_p > \ell$, we have

$$\begin{aligned} &|\Pr\{X_{i_1} \leq u_n, ..., X_{i_p} \leq u_n, X_{j_1} \leq u_n, ..., X_{j_q} \leq u_n\} - \\ &\Pr\{X_{i_1} \leq u_n, ..., X_{i_p} \leq u_n\} \Pr\{X_{j_1} \leq u_n, ..., X_{j_q} \leq u_n\}| \ \leq \ \alpha_{n,\ell} \end{aligned} \tag{3.5}$$

where $\alpha_{n,\ell_n} \to 0$ for some sequence $\ell_n = o(n)$. Without loss of generality, $\alpha_{n,\ell}$ may be taken non-increasing in ℓ for each n, and the condition $\alpha_{n,\ell_n} \to 0$ may be replaced by

$$\alpha_{n,\lfloor n\lambda \rfloor} \to 0 \text{ for each } \lambda > 0.$$

Here $\lfloor \cdot \rfloor$ denotes integer part.

The main result about condition $D(u_n)$ is that it suffices for the Extremal Types theorem: if $a_n > 0,\ b_n$ are such that

$$\Pr\{M_n \leq a_n x + b_n\} \to G(x)$$

where G is non-degenerate, and if $D(u_n)$ holds for each sequence of the form $u_n = a_n x + b_n$ for fixed x, then G is one of the three extreme value types described in Chapter 1

Under a further condition denoted $D'(u_n)$, defined by

$$\limsup_{n\to\infty} n \sum_{j=2}^{\lfloor n/r \rfloor} \Pr\{X_1 > u_n, X_j > u_n\} \to 0 \text{ as } r \to \infty \tag{3.6}$$

it follows that (3.4) holds and hence that the limiting distribution is the same as in the independent case. An alternative way of stating the result is that if both $D(u_n)$ and $D'(u_n)$ hold for a particular sequence $\{u_n\}$, then for $0 \leq \tau < \infty$,

$$\Pr\{M_n \leq u_n\} \to e^{-\tau} \text{ if and only if } n \Pr\{X_1 > u_n\} \to \tau. \tag{3.7}$$

This result is exactly of the form (3.4). The condition $D'(u_n)$ represents precisely

what is meant by saying that high-level exceedances do not cluster, a concept which, as has already been mentioned, was present in earlier work of Watson and Loynes.

Putting all these results together, we have:

Theorem. If $D(u_n)$ and $D'(u_n)$ hold whenever $u_n = a_n x + b_n$ for fixed x, and if G is a nondegenerate distribution function, then

$$\Pr\{M_n \leq a_n x + b_n\} \to G(x) \text{ if and only if } \Pr\{\hat{M}_n \leq a_n x + b_n\} \to G(x).$$

For proofs we refer to Leadbetter, Lindgren and Rootzén [143], Chapter 3.

One direct application of these results is to Gaussian sequences. If $\{X_n\}$ is a stationary Gaussian sequence satisfying Berman's condition $r_n \log n \to 0$, and if $\{u_n\}$ is such that $n \Pr\{X_1 > u_n\}$ is bounded. then both $D(u_n)$ and $D'(u_n)$ hold and so M_n and $\hat{M}_n$ have the same limiting distribution. This is described in Chapter 4 of [143].

To summarise, if $D(u_n)$ holds for suitable sequences $\{u_n\}$, then the only possible limit laws are the classical extreme value distributions, but M_n and $\hat{M}_n$ do not necessarily converge to the same limit under the same normalisation. Under the additional condition $D'(u_n)$, the two limits are the same and the asymptotic extreme value theory is the same as if the two sequences were independent. Neither of the conditions $D(u_n)$ or $D'(u_n)$ is universal. For example, exchangeable sequences of random variables typically have a component that persists through the whole series, thereby violating $D(u_n)$. The book by Galambos [82] contained a thorough review of this class of processes. Our own point of view, however, is that $D(u_n)$ is quite a mild condition which we would expect to be satisfied for the great majority of natural processes, whereas $D'(u_n)$ is much more questionable. The bulk of the rest of the chapter will therefore be devoted to cases in which $D(u_n)$ holds but $D'(u_n)$ does not. The key concept in handling these cases is what has come to be known as the *extremal index*.

3.2 The extremal index

The intuitive concept is that if high-level exceedances occur in clusters, then it is only the cluster maxima that have any bearing on the extreme value distribution. Consequently, the calculations must be scaled by a factor which depends on the mean number of exceedances in a cluster. Heuristic arguments of this form go back at least as far as Cartwright [26], and were made more precise by Newell [165] and Loynes [149]. For example, Loynes showed that if $\{X_n\}$ is strong mixing and, for each τ, $u_n(\tau)$ is defined by

$$\Pr\{X_1 > u_n(\tau)\} \leq \tau/n \leq \Pr\{X_1 \geq u_n(\tau)\} \tag{3.8}$$

then the only possible non-degenerate limits of $\Pr\{M_n \leq u_n(\tau)\}$, $0 < \tau < \infty$ are of the form $e^{-\theta\tau}$ with $0 < \theta \leq 1$. O'Brien [169] showed that all such values of θ are possible, and O'Brien [170] extended the theory to cover the case $\theta = 0$. Later Davis [47] and ultimately Leadbetter [142] extended the theory to the case where the only mixing condition assumed is $D(u_n)$, and Leadbetter coined the term *extremal index* for the parameter θ.

The main results are as follows [142]:

Theorem 1. Suppose (3.8) holds, and that $D(u_n(\tau_0))$ holds for some $\tau_0 > 0$. Then there exist constants θ and θ' such that

$$\begin{aligned} \limsup_{n\to\infty} \Pr\{M_n \le u_n(\tau)\} &= e^{-\theta\tau}, \\ \liminf_{n\to\infty} \Pr\{M_n \le u_n(\tau)\} &= e^{-\theta'\tau}, \end{aligned}$$

for all $\tau \le \tau_0$. Hence, if $\Pr\{M_n \le u_n(\tau)\}$ converges for at least one $\tau \le \tau_0$, then it does so for all $\tau \le \tau_0$ and $\theta = \theta'$. In that case, the extremal index exists and is θ.

Theorem 2. Suppose $\{X_n\}$ has extremal index θ. Let $\{v_n\}$ denote any sequence of constants and let $0 \le \rho \le 1$. Then

(i) If $\theta > 0$, then

$$\Pr\{\hat{M}_n \le v_n\} \to \rho \text{ if and only if } \Pr\{M_n \le v_n\} \to \rho^\theta.$$

(ii) If $\theta = 0$, then

$$\liminf_{n\to\infty} \Pr\{\hat{M}_n \le v_n\} > 0 \Rightarrow \Pr\{M_n \le v_n\} \to 1$$

and

$$\limsup_{n\to\infty} \Pr\{M_n \le v_n\} < 1 \Rightarrow \Pr\{\hat{M}_n \le v_n\} \to 0.$$

O'Brien [172] gave an alternative approach which showed that θ may also be defined as

$$\theta = \lim_n \Pr\{\max(X_2, ..., X_{p_n}) \le u_n | X_1 > u_n\} \tag{3.9}$$

for suitable sequences $\{u_n\}$, $\{p_n\}$. The conditions for this are as follows:

(i) the sequence $\{u_n\}$ must satisfy one of the conditions

$$\liminf F^n(u_n) > 0$$

or

$$\liminf \Pr\{\max(X_2, ..., X_{p_n}) \le u_n | X_1 > u_n\} > 0;$$

(ii) the process must satisfy $D(u_n)$ (O'Brien actually used a slightly weaker form of this which encompasses certain periodic processes);

(iii) $p_n = o(n)$, $\ell_n = o(p_n)$ and $n\alpha_{n,\ell_n} = o(p_n)$ where ℓ_n, α_{n,ℓ_n} are as in (3.5).

Rootzén [201] obtained a similar result when $\theta > 0$: if $\{u_n\}$ is such that $n\Pr\{X_1 > u_n\} \to \tau > 0$ and if $D(u_n)$ holds, then a necessary and sufficient condition for $\{X_n\}$ to have extremal index θ is that

$$\lim_{\varepsilon\downarrow 0} \limsup_{n\to\infty} |\Pr\{\max(X_2, ..., X_{\lfloor n\varepsilon \rfloor}) \le u_n | X_1 > u_n\} - \theta| = 0.$$

We now give three examples for which it is possible to calculate the extremal index exactly.

Example 1.

Let $\{Z_n,\ -\infty < n < \infty\}$ denote a doubly infinite sequence of independent random variables with common distribution function $\Pr\{Z_n \le z\} = \exp(-z^{-\alpha}),\ z > 0$, and let $\{c_j,\ j \ge 0\}$ denote an increasing sequence of positive numbers with $c_0 = 1$ and $c_j \to \infty$ as $j \to \infty$. Define

$$X_n = \max_{j \ge 0} Z_{n-j}/c_j. \tag{3.10}$$

It is readily checked that the distribution function of X_n is

$$\Pr\{X_n \le x\} = \exp(-Kx^{-\alpha}) \tag{3.11}$$

where $K = \sum_n c_n^{-\alpha}$, which we assume to be finite. If $a_n = (nK)^{1/\alpha}$ then

$$n\Pr\{X_1 > a_n x\} \sim nK(a_n x)^{-\alpha} = x^{-\alpha}$$

so $\Pr\{\hat{M}_n \le a_n x\} \to \exp(-x^{-\alpha})$. Now, the event $\{M_n \le x\}$ may be written as $A \cap B$ where

$$A = \cap_{j=1}^{n}\{Z_j \le x\}, \quad B = \cap_{j=-\infty}^{0}\{Z_j \le c_{1-j}x\}.$$

By (3.11), $\Pr\{B\} \to 1$ as $x \to \infty$ while $P\{A\} \to \exp(-nx^{-\alpha})$. Hence with $\{a_n\}$ as above,

$$\Pr\{M_n \le a_n x\} \sim \exp\{-n(a_n x)^{-\alpha}\} = \exp(-K^{-1}x^{-\alpha}).$$

Thus the extremal index is K^{-1}, which can be any value in $(0, 1]$.

Example 2. This example is similar in structure to the previous one, but results in $\theta = 0$. Suppose $\{Z_n\}$ is as in Example 1, and let $\{c_j,\ j \ge 0\}$ denote an increasing sequence with $c_0 = 0,\ c_j \to \infty$ as $j \to \infty$. Let

$$X_n = \max_{j \ge 0}\{Z_{n-j} - c_j\}.$$

Then

$$\Pr\{X_n \le x\} = \exp\{-\sum_j (x + c_j)^{-\alpha}\}.$$

Suppose further $c_j \sim Kj^{1/\beta}$ as $j \to \infty$, where $K > 0$ and $0 < \beta < \alpha$. An argument given below shows that

$$\sum_j (x + c_j)^{-\alpha} \sim \frac{K^{-\beta}\Gamma(\alpha - \beta)\Gamma(\beta + 1)}{\Gamma(\alpha)} x^{\beta - \alpha} \text{ as } x \to \infty. \tag{3.12}$$

Hence with suitable $\hat{a}_n = O\{n^{1/(\alpha+\beta)}\}$ we have

$$\Pr\{\hat{M}_n \le \hat{a}_n x\} \to \exp(-x^{\beta-\alpha}).$$

But $\Pr\{M_n \le x\} \sim \Pr\{\max(Z_1,...,Z_n) \le x\}$ as in Example 1, so with $a_n = n^{1/\alpha}$,

$$\Pr\{M_n \le a_n x\} \to \exp(-x^{-\alpha}).$$

This is a case where M_n and $\hat{M}_n$ converge to different limits at different rates of convergence, and $\theta = 0$. This example is due to L. de Haan (see [142]).

Proof of (3.12). Extend $\{c_j\}$ to a monotonically increasing function $c(t)$, $0 \le t < \infty$ with $c(j) = c_j$ for each integer j. Let $K_1 = \inf_{t>0}\{t^{-1/\beta}c(t)\}$ (positive and finite) and consider

$$\begin{aligned}\int_0^\infty \{x+c(t)\}^{-\alpha}dt &= K^{-\beta}x^\beta \int_0^\infty \{x+c(K^{-\beta}x^\beta t)\}^{-\alpha}dt \\ &= K^{-\beta}x^{\beta-\alpha}\int_0^\infty \left\{1+\frac{c(K^{-\beta}x^\beta t)}{x}\right\}^{-\alpha} dt \\ &\sim K^{-\beta}x^{\beta-\alpha}\int_0^\infty (1+t^{1/\beta})^{-\alpha}dt\end{aligned}$$

by the dominated convergence theorem (the integrand is dominated by the expression $(1+K_1K^{-1}t^{1/\beta})^{-\alpha}$). The integral is bounded above and below by the sums

$$\sum_0^\infty (x+c_j)^{-\alpha}, \quad \sum_1^\infty (x+c_j)^{-\alpha},$$

and hence is asymptotic to either. Finally, the substitution $t = u^{-\beta}(1-u)^\beta$ shows that

$$\int_0^\infty (1+t^{1/\beta})^{-\alpha}dt = \frac{\Gamma(\alpha-\beta)\Gamma(\beta+1)}{\Gamma(\alpha)}$$

from which (3.12) follows.

Example 3.

This is motivated by examples in [55, 172], and shows an alternative mechanism which results in $\theta = 0$. Suppose $\{W_n\}$ is the Markov chain with state space $\{1,2,3,....\}$ and transition probabilities

$$\Pr\{W_{n+1} = j | W_n = i\} = \begin{cases} \{i/(i+1)\}^\beta, \text{ if } j = i+1, \\ 1-\{i/(i+1)\}^\beta, \text{ if } j = 1, \\ 0, \text{ otherwise,}\end{cases}$$

where $1 < \beta < 2$. This is positive recurrent with stationary distribution

$$\Pr\{W_n = i\} = ci^{-\beta}, \quad i \ge 1,$$

where $c = 1/\sum_i i^{-\beta}$.

Now suppose $\{X_n\}$ is another sequence of random variables, conditionally independent given $\{W_n\}$, where

$$\Pr\{X_n \le x | W_j, \text{ all } j\} = \Pr\{X_n \le x | W_n = m\} = \exp(-mx^{-\alpha}) \quad (x > 0, \alpha > 0). \tag{3.13}$$

Direct calculations given below show that

$$\Pr\{X_n > y\} \sim c_1 y^{\alpha - \alpha\beta} \text{ as } y \to \infty, \text{ where } c_1 = c \int_0^\infty z^{-\beta}(1 - e^{-z})dz, \tag{3.14}$$

$$\lim_{y\to\infty} \Pr\{W_n \le y^\alpha w | X_n = y\} = \int_0^w \frac{z^{1-\beta} e^{-z}}{\Gamma(2-\beta)} dz, \tag{3.15}$$

and

$$\Pr\{\max(X_2, ..., X_{p+1}) \le y | X_1 = y\} \to (p+1)^{\beta-2} \text{ as } y \to \infty. \tag{3.16}$$

The probability in (3.16) may be made arbitrarily small by taking p sufficiently large. Moreover, replacing the conditioning event $X_1 = y$ by $X_1 \ge y$ makes it even smaller. Hence by O'Brien's condition (3.9), the extremal index is 0 for this process.

Although this example has been artifically constructed, it is intended to illustrate a general scenario in which extremal index 0 may arise, namely one in which there is a slowly moving "background noise" process $\{W_n\}$ which highly influences the values of $\{X_n\}$.

Proofs of (3.14)-(3.16). For fixed $w_1,\ w_2,\ x_1 > 0$ and x_2, as $y \to \infty$,

$$\begin{aligned}
& \Pr\{w_1 y^\alpha \le W_n \le w_2 y^\alpha,\ \ x_1 y \le X_n \le x_2 y\} \\
= & \sum_{j=\lfloor w_1 y^\alpha \rfloor}^{j=\lfloor w_2 y^\alpha \rfloor} c j^{-\beta} [\exp\{-j(x_2 y)^{-\alpha}\} - \exp\{-j(x_1 y)^{-\alpha}\}] \\
\sim & \ c \int_{w_1 y^\alpha}^{w_2 y^\alpha} z^{-\beta} [\exp\{-z(x_2 y)^{-\alpha}\} - \exp\{-z(x_1 y)^{-\alpha}\}] dz \\
= & \ c y^{\alpha - \alpha\beta} \int_{w_1}^{w_2} z^{-\beta} [\exp\{-z(x_2)^{-\alpha}\} - \exp\{-z(x_1)^{-\alpha}\}] dz.
\end{aligned} \tag{3.17}$$

Setting $w_1 = 0,\ w_2 = \infty,\ x_1 = 1,\ x_2 = \infty$ gives (3.14). Dividing (3.18) by the same expression with $w_1 = 0,\ w_2 = \infty$ and letting $x_2 \to 1, x_1 \to 1$ we have

$$\Pr\{w_1 y^\alpha \le W_n \le w_2 y^\alpha | X_n = y\} \to \int_{w_1}^{w_2} \frac{z^{1-\beta} e^{-z}}{\Gamma(2-\beta)} dz$$

which is (3.15). Finally, given $X_1 = y$ (large) we have $W_1 \approx y^\alpha Z$ with Z having a Gamma($2-\beta$,1) distribution. For fixed p, the probability that $W_j = W_1 + j - 1$ for all $j = 2, ..., p+1$ tends to 1 as $y \to \infty$. Hence

$$\Pr\{\max(X_2, ..., X_{p+1}) \le y | Z\} \sim [\exp\{-(y^\alpha Z) y^{-\alpha}\}^p] = e^{-pZ}$$

and integrating with respect to the distribution of Z gives (3.16).

3.3 Infinitely divisible random measures

In studying the point processes generated by exceedances in stationary sequences, an important concept is that of an *infinitely divisible* point process or, more generally, random measure. We give here a brief non-rigorous account of the basic representation theorem for such processes and refer to a text such as Kallenberg [133] for full details.

Suppose η is a random measure on a set S. If f is a non-negative measureable function on S, we write

$$\eta(f) = \int_S f(x)\eta(dx) \tag{3.18}$$

and define the *Laplace transform* of η by

$$L_\eta(f) = \mathrm{E}\{e^{-\eta(f)}\}.$$

In the case that η is a counting measure, or in other words a point process with points at $\{T_i\}$ say, (3.18) is equivalent to

$$\eta(f) = \sum_i f(T_i).$$

Here are two examples:

(i) Let η be a non-homogeneous Poisson process on a Euclidean space with intensity $\nu(x),\ x \in S$, with respect to Lebesgue measure. Suppose the support of f is contained in a set A (i.e. $f(x) = 0$ outside A) where

$$\mu = \int_A \nu(x)dx < \infty.$$

The process ν, restricted to A, may be constructed as follows: first let N, the total number of points in A, have a Poisson distribution with mean μ, then let $T_1, ..., T_N$ be conditionally independent given N, each with density $\nu(x)/\mu,\ x \in A$. Thus

$$\mathrm{E}\left\{\exp\left(-\sum_{i=1}^N f(T_i)\right)|N\right\} = \left\{\mu^{-1}\int_A e^{-f(x)}\nu(x)dx\right\}^N$$

and hence, since $\mathrm{E}\{z^N\} = e^{-\mu(1-z)}$ for any $z > 0$, we have

$$L_\eta(f) = \exp\left[-\int_A \left\{1 - e^{-f(x)}\right\}\nu(x)dx\right] \tag{3.19}$$

$$= \exp\left[-\int_S \left\{1 - e^{-f(x)}\right\}\nu(x)dx\right] \tag{3.20}$$

The final formula does not depend on A and hence, by monotone convergence arguments, remains valid without the initial restriction of f to a set A on which $\mu < \infty$.

(ii) Let $\{T_i\}$ denote a homogeneous Poisson process on S, with intensity ν, and let $\{Y_i\}$ denote an independent sequence of independent, identically distributed non-negative random variables with Laplace transform $\phi(t) = \mathrm{E}\{\exp(-tY_i)\}$. Let η denote the compound Poisson process that puts mass Y_i at each T_i, so that

$$\eta(f) = \sum_i Y_i f(T_i).$$

To calculate the Laplace transform in this case, suppose f is restricted to a set A with Lebesgue measure $\mu < \infty$. If $T_1, ..., T_N$ are the points of the Poisson process in A, then N has a Poisson distribution with mean $\mu\nu$ and, conditionally on N, $T_1, ..., T_N$ are independent uniform over A. So

$$\begin{aligned} \mathrm{E}\{\exp(-\eta(f))|T_1, ..., T_N\} &= \prod_{i=1}^{N} \phi\{f(T_i)\}, \\ \mathrm{E}\{\exp(-\eta(f))|N\} &= \left[\mu^{-1} \int_A \phi\{f(x)\} dx\right]^N \end{aligned}$$

and hence

$$L_\eta(f) = \exp\left[-\nu \int_S \{1 - \phi(f(x))\}\, dx\right]. \tag{3.21}$$

Again, the fact that this formula is independent of A shows that it may be extended to cases in which the support of f is not restricted to a bounded set.

These two examples are both examples of *infinitely divisible* random measures. A random measure η is infinitely divisible if, for any positive integer n, it can be written $\eta = \eta_1 + ... + \eta_n$ where $\eta_1, ..., \eta_n$ are i.i.d. random measures. Equivalently, η is infinitely divisible if, for each n, $\{L_\eta(f)\}^{1/n}$ is the Laplace transform of a random measure. It is obvious that both (3.20) and (3.21) have this property.

In general, if we let $\mathscr{M}$ denote the class of all locally finite measures on S, then any infinitely divisible random measure on S has the Laplace transform

$$-\log L_\eta(f) = \alpha(f) + \int_{\mathscr{M}-\{0\}} [1 - \exp\{-\mu(f)\}] \lambda(d\mu) \tag{3.22}$$

where $\alpha \in \mathscr{M}$, 0 denote the null measure that assigns mass 0 to every set, and λ is an arbitrary measure on $\mathscr{M}$ satisfying

$$\int_{\mathscr{M}-\{0\}} [1 - \exp\{-\mu(I_B)\}] \lambda(d\mu) < \infty \text{ for each bounded set } B \tag{3.23}$$

where I_B denotes the indicator function of the set B. Proof of this formula is given by Kallenberg [133], Theorem 6.1.

To see that (3.20) and (3.21) are both of the form (3.22):

(i) Let λ be concentrated on the Dirac (delta) measures on S and identify $\mu(f)$ with $f(x)$, $\lambda(d\mu)$ with $\nu(x)dx$ when $\mu = \delta_x$ for some $x \in S$. Then (3.22) with $\alpha = 0$ reduces to (3.20).

(ii) First suppose the Lebesgue measure $|S|$ is finite. Define a random element μ of M by first choosing T uniformly distributed on S, then putting mass Y on T where Y is chosen independently with the same distribution as the $\{Y_i\}$. Let λ denote $\nu|S|$ times the probability measure of this random element. Let $\mu(f) = Yf(T)$. Then

$$\int_{\mathscr{M}-\{0\}} [1-\exp\{-\mu(f)\}]\lambda(d\mu) = \nu|S|\mathrm{E}\{e^{-Yf(T)}\} = \nu\int_S \phi\{f(x)\}dx.$$

With this identification (and $\alpha = 0$), (3.22) reduces to (3.21). The case where S is unbounded may be handled by writing $S = \cup S_n$, where each S_n is bounded, and taking limits in the above representation.

The general form of (3.22) shows that any infinitely divisible random measure may be written as the sum of a non-random α (it is obvious that any non-random measure is infinitely divisible) and a form of generalised compound Poisson process in which random measures on S are chosen corresponding to an intensity measure λ.

Another interpretation is given by Kallenberg (1983), Section 6.2. Suppose η is given by (3.23) and write $\mathscr{M} = \cup\mathscr{M}_n$ where the $\mathscr{M}_n$ are disjoint and, for each n, $\lambda(\mathscr{M}_n) < \infty$. For each n, let N_n have a Poisson distribution with mean $\lambda(M_n)$ and, conditionally on N_n, let $\{\xi_{nj}\}$ $(1 \le j \le N_n)$ be independent random members of $\mathscr{M}_n$ with distribution $\lambda(d\mu)/\lambda(\mathscr{M}_n)$, $\mu \in \mathscr{M}_n$. Then η has the representation

$$\eta = \alpha + \sum_n \sum_{j=1}^{N_n} \xi_{nj}.$$

3.4 Exceedances of a single level

Let $\{X_n\}$ denote a stationary sequence and let $M_n = \max(X_1,...,X_n)$. For an increasing sequence of thresholds $\{u_n\}$, define $\{N_n\}$ to be the point process on $[0,1]$ which puts a point at j/n if $X_j > u_n$, $1 \le j \le n$. In this section we describe the limiting behaviour of the point processes $\{N_n\}$.

If $n\Pr\{X_1 > u_n\} \to \tau$ $(0 < \tau < \infty)$, and if $D(u_n)$ and $D'(u_n)$ hold as defined by equations (3.5) and (3.6), then N_n converges weakly to a homogeneous Poisson process with intensity τ ([143], Theorem 5.2.1). Under a slightly stronger condition, both the definition of N_n and the Poisson convergence may be extended from $[0,1]$ to $[0,\infty)$.

Now suppose $D(u_n)$ holds but $D'(u_n)$ does not. In describing this situation we follow Hsing, Hüsler and Leadbetter [123]. They assume a slightly stronger form of $D(u_n)$, as follows: given the sequence $\{u_n\}$, define $\mathscr{B}(i,j)$ to be the σ-field generated by the events $\{X_m \le u_n,\ i \le m \le j\}$, and let

$$\alpha(n,\ell) = \max\{|\Pr\{AB\} - \Pr\{A\}\Pr\{B\}| : A \in \mathscr{B}(1,k),\ B \in \mathscr{B}(k+\ell,n),\ 1 \le k \le n-\ell\} \quad (3.24)$$

for $n \ge 1$, $\ell \ge 1$. *Condition* $\Delta(u_n)$ is said to hold if there exist ℓ_n, $n \ge 1$ such that

$$\ell_n = o(n), \quad \alpha(n,\ell_n) \to 0. \quad (3.25)$$

This is nominally a stronger condition than $D(u_n)$, but it is not clear whether there actually exist processes satisfying $D(u_n)$ but not $\Delta(u_n)$.

Hsing, Hüsler and Leadbetter showed that if $a_n N_n$ converges to a limiting random measure N, for some sequence of positive constants $\{a_n\}$, then N necessarily has the properties

(i) N is stationary — in particular, if $A = (a,b)$ and $B = (a+h, b+h)$ for some $0 \le a \le b \le b+h \le 1$, then $N(A)$ and $N(B)$ have the same distribution,

(ii) N has no fixed atoms, i.e. $N(\{x\}) = 0$ with probability 1 for each fixed x,

(iii) N is infinitely divisible.

By combining properties (i) and (ii) with the general representation (3.22), they deduced the representation

$$-\log L_N(f) = \alpha \int_0^1 f(x)dx + \int_0^1 \int_0^\infty \{1 - e^{-yf(x)}\}\nu(dy)dx \tag{3.26}$$

where $\alpha \ge 0$ and ν is a measure on $(0,\infty)$ satisfying

$$\int_0^\infty (1 - e^{-y})\nu(dy) < \infty.$$

Thus, N consists of αm (m = Lebesgue measure) plus a sequence of point masses $Y_i > 0$ at points $T_i \in [0,1]$, where $\{(T_i, Y_i), i = 1,2,...,\}$ for a Poisson process on $[0,1] \times (0,\infty)$ with intensity measure $m \times \nu$. In the case where ν is a finite measure this reduces to a compound Poisson process as described in example (ii) of Section 3.3.

Here are some examples to illustrate the application of this result.

1. Consider example 1 of Section 3.2. Let $u_n = (n/\tau)^{1/\alpha}$ for some fixed τ, so that $\Pr\{M_n \le u_n\} \to e^{-\tau}$. Suppose $Z_j > u_n$. Now

$$\Pr\{Z_j > u_n y | Z_j > u_n\} \to y^{-\alpha} \text{ as } u_n \to \infty \text{ for } y > 1.$$

Hence for fixed $r \ge 1$,

$$\Pr\{X_{j+r} > u_n | Z_j > u_n\} \sim \Pr\{Z_j/c_r > u_n | Z_j > u_n\} \to c_r^{-\alpha}.$$

Consequently, each exceedance by the process $\{Z_j\}$ over u_n gives rise independently to a random number R of exceedances by $\{X_j\}$ over u_n, where as $n \to \infty$,

$$\Pr\{R = r\} \to \pi(r) = c_{r-1}^{-\alpha} - c_r^{-\alpha} \quad (r \ge 1).$$

Thus N_n converges to a limit N which is a compound Poisson process with Laplace transform

$$\begin{aligned} -\log L_N(f) &= \tau \int_0^1 [1 - \phi\{f(x)\}]dx, \\ \phi(t) &= \sum_{r=1}^\infty e^{-rt}\pi(r). \end{aligned}$$

The mean cluster size in the limit is

$$\sum_{r=0}^{\infty} \Pr\{R > r\} = \sum_{r=0}^{\infty} c_r^{-\alpha} = K$$

where, as shown in Section 3.2, K is the reciprocal of the extremal index. Note that any distribution of cluster size with finite mean can arise through this scheme.

2. Consider example 2 of Section 3.2 with $u_n = (n/\tau)^{1/\alpha}$ again. In this case

$$\Pr\{X_{j+r} > u_n | Z_j > u_n\} \sim \Pr\{Z_j > u_n + c_r | Z_j > u_n\} \sim \left(1 + \frac{c_r}{u_n}\right)^{-\alpha}$$

which tends to 1 as $u_n \to \infty$ for any fixed r. Thus the clusters are asymptotically of infinite size. To obtain a non-degenerate result we must renormalise: if $r = y(u_n/K)^{\beta}$ for fixed $y > 0$ then

$$\Pr\{X_{j+r} > u_n | Z_j > u_n\} \to (1 + y^{1/\beta})^{-\alpha}.$$

Thus N_n is approximately a Poisson process of clusters having intensity τ, where each cluster size is of the form $Y(u_n/K)^{\beta}$ with $\Pr\{Y > y\} = (1 + y^{1/\beta})^{-\alpha}$. The mean cluster size is $O(u_n^{\beta}) = O(n^{\beta/\alpha})$, which is $o(n)$ since we assumed $\beta < \alpha$.

To obtain a limiting process in this case, let $a_n = K^{\beta} u_n^{-\beta}$. Then $a_n N_n \to N$ where N is the compound Poisson process $\{T_i, N_i\}$, with $\{T_i\}$ a homogeneous Poisson process on $[0,1]$ with intensity τ, and $\{Y_i\}$ independent with common distribution function $1 - (1 + y^{1/\beta})^{-\alpha}$. This is a case where the limiting measure is not a point process.

3. Let $\{X_n\}$ be any ergodic stationary process $u_n \equiv u$ a fixed threshold and suppose $\alpha = \Pr\{X_1 > u\}$. Then N_n/n converges to α times Lebesgue measure. This trivial example shows that the case $\alpha > 0$ in (3.26) cannot be dispensed with.

If $\{N_n\}$ converges without any renormalisation, i.e. if we may take $a_n \equiv 1$, then N is necessarily a point process and the Laplace transform is given by

$$L_N(f) = \exp\left[-\nu \int_0^1 \{1 - \phi(f(x))\} dx\right] \tag{3.27}$$

where $\nu > 0$ and ϕ is of the form

$$\phi(t) = \sum_{j=1}^{\infty} e^{-tj} \pi(j), \tag{3.28}$$

where $\pi(\cdot)$ is a probability distribution on the positive integers. This corresponds to a clustered point process in which the clusters form a homogeneous Poisson process with intensity ν and the independent cluster sizes have distribution π. In this case, some more specific results due to [123] were as follows:

(i) If (3.25) is strengthened to

$$k_n \ell_n = o(n), \quad k_n \alpha(n, \ell_n) \to 0$$

for some sequence $k_n \to \infty$, and define $r_n = \lfloor n/k_n \rfloor$,

$$\pi_n(j) = \Pr\left\{ \sum_{i=1}^{r_n} I(X_i > u_n) = j \Big| \sum_{i=1}^{r_n} I(X_i > u_n) > 0 \right\}$$

for $j = 1, 2, \ldots$ If $\Delta(u_n)$ holds and N_n converges to N, then the distribution $\pi(\cdot)$ in (3.29) is defined by

$$\pi(j) = \lim_{n\to\infty} \pi_n(j), \quad j = 1, 2, \ldots \tag{3.29}$$

Conversely, if $\Pr\{M_n \le u_n\} \to e^{-\tau}$ and both $\Delta(u_n)$ and (3.29) hold, then N_n converges to a point process N with Laplace transform (3.27).

(ii) Suppose $\{u_n(\tau),\ \tau > 0,\ n \ge 1\}$ is a sequence of thresholds such that $n\Pr\{X_1 > u_n(\tau)\} \to \tau$ for each τ, and $\Delta(u_n(\tau))$ holds for each τ. If $N_n(\tau)$ converges for at least one $\tau > 0$, then it does so for all τ and the limiting process has Laplace transform

$$L_N(f) = \exp\left[-\theta\tau \int_0^1 \{1 - \phi(f(x))\} dx \right]$$

where both θ and $\phi(t) = \sum_{j=1}^{\infty} e^{-tj} \pi(j)$ are independent of τ. Here θ is the extremal index, which also satisfies

$$\theta^{-1} = \lim_{n\to\infty} \sum_{j=1}^{\infty} j\pi_n(j). \tag{3.30}$$

If $\sum_{j=1}^{\infty} j\pi_n(j) \to \sum_{j=1}^{\infty} j\pi(j)$ then we also have

$$\theta^{-1} = \sum_{j=1}^{\infty} j\pi(j). \tag{3.31}$$

giving a direct interpretation of the extremal index as the reciprocal of the mean cluster size in the limiting point process. It is possible, however, for (3.30) to hold without (3.31); Smith (1988a) had an explicit counterexample.

An application of these results is to the limiting distribution of the k'th largest order statistic. Suppose the preceding results holds and there exist $a_n,\ b_n$ such that

$$\Pr\{M_n \le a_n x + b_n\} \to G(x) \tag{3.32}$$

with nondegenerate G. Then if $M_n^{(k)}$ denotes the k'th largest order statistic in $\{X_1, \ldots, X_n\}$ for fixed $k \ge 1$, we have

$$\Pr\{M_n^{(k)} \le a_n x + b_n\} \to G(x) \left[1 + \sum_{j=1}^{k-1} \sum_{i=j}^{k-1} \frac{\{-\log G(x)\}^j}{j!} \pi^{*j}(i) \right] \tag{3.33}$$

where π^{*j} denotes the j-fold convolution of π. To see this, if $u_n = a_n x + b_n$ and $\nu = -\log G(x)$, then $\Pr\{M_n \le u_n\} \to e^{-\nu}$, hence N is compound Poisson with cluster intensity ν and cluster size distribution π. We want to calculate

$$\lim_{n\to\infty} \Pr\{M_n^{(k)} \le u_n\} = \Pr\{N(0,1) \le k-1\}.$$

Let J denote the number of clusters. Then

$$\Pr\{N(0,1) \le k-1 | J\} = \begin{cases} 1, \text{ if } J = 0, \\ \sum_{i=j}^{k-1} \pi^{*j}(i), \text{ if } J = j,\ 0 < j < k, \\ 0, \text{ if } J = j \ge k. \end{cases}$$

Hence

$$\Pr\{N(0,1) \le k-1\} = e^{-\nu} \left\{ 1 + \sum_{j=1}^{k-1} \sum_{i=j}^{k-1} \frac{\nu^j \pi^{*j}(i)}{j!} \right\}$$

which is (3.33). If $\theta = 1$ then $\pi(1) = 1$ and (3.33) reduces to

$$\Pr\{M_n^{(k)} \le a_n x + b_n\} \to G(x) \left[1 + \sum_{j=1}^{k-1} \frac{\{-\log G(x)\}^j}{j!} \right], \tag{3.34}$$

exactly as in the independent case.

Some parallel results under rather different assumptions were obtained by Dziubdziela [64, 65].

3.5 The two-dimensional exceedance process

We now consider the two-dimensional point process generated by the high-level exceedances os a stationary sequence $\{X_n\}$. Suppose $u_n(\tau)$ is continuous and strictly decreasing in τ for each n, and satisfies

$$\Pr\{M_n \le u_n(\tau)\} \to e^{-\tau} \text{ as } n \to \infty \text{ for each } \tau > 0. \tag{3.35}$$

Let $u_n^{-1}(\cdot)$ denote the inverse function of $u_n(\cdot)$. For each n, define a point process N_n on the plane by putting a point at each $(j/n, u_n^{-1}(X_j))$, $1 \le j \le n$. In the independent case, Pickands [181] and Resnick [192] showed that N_n converges in distribution to a homogeneous Poisson process N on $\mathscr{R} \times (0,\infty)$ (section 1.8). This was extended by Adler [1] and by Leadbetter, Lindgren and Rootzén [143], Section 5.7, to processes satisfying a version of condition $D(u_n(\tau))$, together with $D'(u_n(\tau))$ for each τ. Adler showed that this covers stationary Gaussian sequences satisfying Berman's condition $\rho_n \log n \to 0$. As in the previous section, we shall concentrate on the case when D' does not hold. This has been developed by Hsing [121, 122] following earlier work by Mori [162]. The mean number of points of N_n in any bounded subset of the plane is finite (and bounded as $n \to \infty$), so it makes sense to look for a limiting point process N. Our interest is in characterising this limit when it exists.

We know from the previous section that the projection of this two-dimensional process onto the time axis is a one-dimensional clustered Poisson process, so the two-dimensional process must consist of columns of points with a common time coordinate. The case where each column contains exactly one point is the case of extremal index 1, when N is homogenous Poisson. In general, however, N is not Poisson.

Hsing [121] approached this problem by first listing the invariance properties thet N must have, and then characterising all processes N having these properties. Mori [162] obtained the same characterization of the limit, under more restrictive assumptions on the original process. Hsing's assumptions are essentially an extension of condition $\Delta(u_n)$ (section 3.4) so that it applies simultaneously to all $u_n(\tau),\ 0 < \tau < \infty$.

Under such assumptions, Mori and Hsing showed that the limiting process must be infinitely divisible and have points of the form $(S_i,\ T_iY_{ij})\ \ (i \geq 1,\ j \geq 1)$ where $\{(S_i,\ T_i),\ i \geq 1\}$ are the points of a two-dimensional homogeneous Poisson process of intensity 1 and, for each i, $(Y_{ij},\ j = 1,...,K_i)$ is a point process on $[1,\infty)$ with $Y_{i1} = 1$ and a random number of points K_i. The processes $\{Y_i\}$ are independent for each i, and identically distributed.

Thus, the columns of points in Fig. 3.1 are characterised by the property that the points at the base of each column form a homogeneous Poisson process, and the ratio of each column to the height of its base form independent, identically distributed point processes on $[1,\infty)$. Mori showed by example that any process having this structure may arise as the limiting exceedance process of a strong mixing process.

To illustrate these concepts, consider again Example 1 of section 3.2. It suffices for (3.35) to take $u_n(\tau) = (n/\tau)^{1/\alpha}$ and hence $u_n(x)^{-1} = nx^{-\alpha}$. Suppose m is such that $X_m = Z_m = z$, where z is very large. Then, for fixed $r/ge0$ we will have

$$\Pr\{X_{m+r} = c_r^{-1}Z_m \mid Z_m = z\} \to 1 \text{ as } z \to \infty$$

and hence $u_n^{-1}(X_{m+r})/u_n^{-1}(X_m) = c_r^{\alpha}$ with conditional probability tending to 1 as $z \to \infty$ for any fixed r. The processes $\{Y_{ij}$ will simply be given by

$$Y_{ij} = c_{j-1}^{\alpha},\ \text{ all } i,\ j = 1,2,\ldots,$$

In this case, therefore, the Y_{ij} turn out to be non-random. Each column consists of infinitely many points, though of course there are only a finite number of points in any finite region. Mori [162] gave a rather similar construction in which the constants $c_r,\ r \geq 1$ are replaced by random variables, and used this to show that any distribution of $\{Y_{ij},\ j \geq 1\}$ with $Y_{i1} = 1$ can arise in the limiting process from a construction of this nature.

Hsing [120] gave a more detailed treatment of our Example 1, expanding it to the case where the common distribution of Z_i has any regularly varying tail.

Example 2 of section 3.2 illustrates a case in which the Mori-Hsing theory fails, but a limiting random measure nevertheless exists under a different renormalization.

Again let $u_n^{-1}(x) = nx^{-\alpha}$ and fix m such that $X_m = Z_m = z$, where z is large. In the limiting process this will correspond to a point (S_i, T_i) where $S_i = m/n,\ T_i = nZ_m^{-\alpha}$. For a large number of $r \geq 1$ we will have

$$\Pr\{X_{m+r} = Z_m - c_r \mid Z_m = z\} \to 1 \text{ as } z \to \infty$$

so each such value will give rise to a point $((m+r)/n,\ n(Z_m - c_r)^{-\alpha}$ which also converges to (S_i, T_i) in the limit. Thus, N_n looks like a clustered point process with cluster sizes tending to infinity. However, for fixed $y \geq 1$ we have

$$\begin{aligned} \sum_{r=0}^{\infty} I\{n(Z_m - c_r)^{-\alpha} \leq T_i y\} &\sim \left(\frac{Z_m}{K}\right)^{\beta} (1 - y^{-1/\alpha})^{\beta} \\ &= n^{\beta/\alpha} T_i^{-\beta/\alpha} K^{\beta} (1 - y^{-1/\alpha})^{\beta} \end{aligned} \tag{3.36}$$

so $n^{-\beta/\alpha} N_n$ converges to a random measure with Laplace transform

$$-\log L_n(f) = \int_{-\infty}^{\infty} \int_0^{\infty} \left[1 - \exp\left\{-\int_0^{\infty} f(s,ty) t^{-\beta/\alpha} K^{-\beta} d((1 - y^{-1/\alpha})\beta\right\}\right] ds dt \tag{3.37}$$

Essentially, (E3.35) is of the form (E3.19) in which $\gamma = 1$, $S = \mathscr{R} \times (\iota, \infty)$ and

$$\phi(f(s,t)) = \exp\left\{-\int_0^{\infty} f(s,ty) t^{-\beta/\alpha} K^{-\beta} d((1 - y^{-1/\alpha})\beta\right\}$$

is the Laplace transform of a single column of the process in the renormalized limit.

This example illustrated the general point that, when the extremal index is 0, a more general theory involving convergence of the rescaled point process to a random measure is required. The presence of $T_i^{-\beta/\alpha}$ in (E3.34), and of $t^{-\beta/\alpha}$ in (E3.35), shows that this random measure has different scaling properties from the point process N considered by Mori and Hsing.

One area of application of these results is to the limiting joint distribution of k largest order statistics, for fixed $k \geq 1$ as $n \to \infty$. Let $M_n^{(1)} \geq ... \geq M_n^{(k)}$ denote the k largest order statistics from $X_1, ..., X_n$ and suppose $a_n,\ b_n$ are such that

$$\Pr\{M_n^{(1)} \leq a_n x + b_n\} \to G(x)$$

where G is one the of extreme value distributions. In this case we may define $u_n^{-1}(x) = \Psi((x - b_n)/a_n)$ where $\Psi(y) = -\log G(y)$. For,

$$\begin{aligned} \Pr\{u_n^{-1}(M_n) \leq \tau\} &= \Pr\{M_n \leq a_n \Psi^{-1}(T) + b_n\} \\ &\to G(\Psi^{-1}(T)) = e^{-\tau} \end{aligned}$$

satisfying (3.35). Note that, since G is an extreme value distribution, Ψ is strictly increasing and hence there is no problem about the existence of Ψ^{-1}.

Suppose the point process N_n, which puts mass 1 at $(j/n,\ \Psi((X_j - b_n)/a_n))$ for each j between 1 and n, converges to a limiting point process N. Let us enumerate

the points of N as $(U_i,\ V_i)$ for $i \geq 1$, where $V_1 \leq V_2 \leq V_3 \leq ...$ Then the limiting distribution of

$$\left[\frac{M_n^{(1)} - b_n}{a_n}, ..., \frac{M_n^{(k)} - b_n}{a_n}\right]$$

is the same as that of

$$\left[\Psi^{-1}(V_1), ..., \Psi^{-1}(V_k)\right].$$

If $\{X_n\}$ satisfies D_n and D_n', then N is homogeneous Poisson and the limiting joint distribution is the same as for the independent case. This result, under a strong mixing assumption and without the point process interpretation, was first obtained by Welsch [258].

In the more general case in which N is the process described by Mori and Hsing, one would in principle obtain the full class of joint distributions. In general this is rather complicated, however, so we content ourselves with the case $k = 2$. In this case Welsch [259] showed that, if the two limits

$$H(x,y) = \lim \Pr\{M_n^{(1)} \leq a_n x + b_n, M_n^{(2)} \leq a_n x + b_n\}, \quad G(x) = \lim \Pr\{M_n^{(1)} \leq a_n x + b_n\}$$

both exist, then H is related to G by

$$H(x,y) = \begin{cases} G(x),\ y \geq x, \\ G(y)\{1 - \rho(\log G(x)/\log G(y))\log G(y)\},\ y < x \end{cases} \tag{3.38}$$

where ρ is a concave, non-increasing function satisfying $0 \leq \rho(s) \leq 1 - s$ for $0 \leq s \leq 1$. The independent (or D') case corresponds to $\rho(s) = 1 - s$. Mori (1976) showed that any ρ satisfying Welsch's conditions can occur, and also gave an example in which G exists but the limiting joint distribution of $(M_n^{(i)} - b_n)/a_n$, $i = 1, 2$, does not.

To derive (3.38), suppose $N_n \to N$ and write $\tau_1 = \Psi(x),\ \tau_2 = \Psi(y)$. We consider only the case $y < x$, i.e. $\tau_1 < \tau_2$, when (E3.36) is equivalent to the statement

$$\Pr\{V_1 > \tau_1,\ V_2 > \tau_2\} = e^{-\tau_2}\{1 + \rho(\tau_1/\tau_2)\tau_2\}. \tag{3.39}$$

Suppose, following Mori and Hsing, the point process N consists of points $(S_i,\ T_i Y_{ij})$ with $(S_i,\ T_i)$ homogeneous Poisson and $\{Y_{ij},\ j \geq 1\}$ independent and identically distributed for each i, with $Y_{i1} = 1$. Let $\tilde{G}$ denote the distribution function of Y_{i2}. The event $(V_1 > \tau_1,\ V_2 > \tau_2)$ can arise in one of two ways. Either there is no point in $(0,1) \times (0, \tau_2)$ with probability $e^{-\tau_2}$, or there is exactly one point at (s,t) $(t > \tau_1)$, with probability $e^{-\tau_2}\{1 - \tilde{G}(\tau_2/t)\}dsdt$. Consequently

$$\Pr\{V_1 > \tau_1,\ V_2 > \tau_2\} = e^{-\tau_2}\left[1 + \int_{\tau_1}^{\tau_2}\left\{1 - \tilde{G}\left(\frac{\tau_2}{t}\right)\right\}dt\right]$$

which is of the form (3.39) with

$$\rho(u) = \int_1^u \left\{ 1 - \tilde{G}\left(\frac{1}{s}\right) \right\} ds. \tag{3.40}$$

It is easy to verify that ρ has the required properties, and moreover that any such ρ arises in the form (3.40) for some (possibly improper) $\tilde{G}$.

This derivation is due to Mori [162], and has been generalised by Hsing [122] to cover all cases in which $((M_n^{(1)} - b_n)/a_n,\ (M_n^{(k)} - b_n)/a_n)$ converges to a bivariate limiting distribution for some fixed $k > 1$.

3.6 Markov chains

Many interesting processes can be formulated either as Markov chains or as (deterministic or random) functions of Markov chains. Extremes of discrete-state Markov chains were studied by Anderson [3]. In this case the discreteness of the process means that, in most cases, a limiting distribution for the extreme values does not exist, but bounds and approximations may still be obtained. A generalisation is to consider *chain-dependent processes* in which $\{S_n\}$ is a Markov chain and $\{X_n\}$ is a sequence of random variables, conditionally independent given $\{S_n\}$, with a distribution of the form

$$\Pr\{X_n \le x \mid S_m,\ -\infty < m < \infty\} = \Pr\{X_n \le x \mid S_n = i\} = H_i(x).$$

Extremes in such processes were studies by Resnick and Neuts [193] for the case when the state space of S_n is finite, and extended by Denzel and O'Brien [55] to countably infinite state spaces.

O'Brien [172] and Rootzén [201] studied extremes of Markov chains, and functions of Markov chains, under the assumption that the chain is regenerative in a certain sense.

The theory of Markov chains on general state spaces has been developed in a number of books such as Nummelin [167] and Meyn and Tweedie [159]. We follow here Section VI.3 of Asmussen [9].

A stochastic sequence $\{X_n\}$ defined on a general set E is called a *Markov chain* if, for any measurable subset F of E,

$$\Pr\{X_{n+1} \in F \mid X_j,\ j \le n\} = P(X_n, F)$$

where, for each $x \in E$, $P(x, \cdot)$ is a probability measure on E. The function P is called the *transition kernel* of the process. It is easily extended to an r'th-order transition kernel

$$P^r(x, F) = \Pr\{X_{n+r} \in F \mid X_n = x\}.$$

If A is an event depending on the whole sequence $\{X_n,\ n \ge 0\}$, we write $P_x(A)$ for the probability of the event A given $X_0 = x$.

We may also define the hitting time of a set: if R is a measurable subset of E then

$$\tau(R) = \inf\{n \geq 1 : \ X_n \in R\}.$$

In the case when $X_n \notin R$ for all $n \geq 1$, we write $\tau_R = +\infty$. The set R is called *recurrent* if

$$P_x\{\tau_R < \infty\} = 1 \text{ for all } x \in E.$$

In other words, R is recurrent if it is certain to be visited by the process regardless of its initial state.

A related concept is that of a *regeneration set*. Asmussen defines R to be a regeneration set if it is recurrent and there exists $\varepsilon.0$, $r \geq 1$ and a probability measure λ on E such that

$$P^r\{x,B) \geq \varepsilon\lambda(B) \text{ for all } x \in R, \text{ measurable } B \subseteq E.$$

Finally, the chain is called *Harris recurrent* (or just a *Harris chain*) if there exists a regeneration set.

Any discrete-state recurrent chain is Harris, because we may take $R = \{x\}$ for any recurrent state x and define $\lambda(B) = P^r(x,B)$. However, the application is much wider than that, and covers many continuous-state processes.

A stochastic sequence $\{X_n\}$ is called *regenerative* if there exist integer-valued variables $0 < T_0 < T_1 < ...$ breaking up the process into cycles

$$C_i = \{X_n \ : \ T_{i-1} < n \leq T_i\}, \quad i \geq 1.$$

It is required that the post-T_k process

$$\{(C_i, T_i), \ i > k\}$$

be independent of $T_1, ..., T_k$, with the same distribution for each k.

A discrete-state recurrent Markov chain is trivially regenerative: define T_k to be the $(k+1)$'st time the chain hits a fixed recurrent state x. The $\{C_i\}$ are then independent.

For a Harris chain, the main result is that the process is regenerative. The $\{C_i\}$ are not necessarily independent, but they may be taken to be 1–dependent, i.e. if A is an event depending only on $\{C_j, \ j \leq i-1\}$ and B is an event depending only on $\{C_j, \ j \geq i+1\}$ then A and B are independent. An informal proof of this is as follows. Let R be a regenerative set, and let the process run until $X_n \in R$. Let J_n be independent of the past with $\Pr\{J_n = 1\} = 1 - \Pr\{J_n = 0\} = \varepsilon$. If $J_n = 1$, let X_{n+r} be chosen according to the measure λ. If $J_n = 0$, choose X_{n+r} according to the measure $\{P^r(X_n, \cdot) - \varepsilon\lambda(\cdot)\}/(1-\varepsilon)$. Then let $X_{n+1}, ..., X_{n+r-1}$ be chosen from their joint conditional distribution given X_n and X_{n+r}.

It is obvious that this construction preserves the structure of $\{X_n\}$. Moreover, the time points $\{n : \ X_n \in R, \ J_n = 1\}$ constitute regeneration points (the T_i's). If $r = 1$ then the cycles C_i of values between regeneration points are independent, but if $r > 1$ then

the first $r-1$ values of each cycle depend on the last value of the previous cycle. For this reason, the cycles cannot in general be independent, but they are 1-dependent.

A Harris chain is called *aperiodic* if the common distribution of $T_i - T_{i-1},\ i = 1,2,...,$ is aperiodic. This property does not depend on the specific choice of $R,\ \varepsilon$ or λ. A measure ν on E is called *stationary* for P if

$$\nu(F) = \int \nu(dx)P(x,F) \ \text{ for all measurable } F \subseteq E.$$

Any Harris chain has a stationary measure which is unique up to a multiplicative constant. If $\nu(E) < \infty$ the chain is *positive recurrent*. In this case ν may be normalised to be a probability measure, and we write π in place of ν. An aperiodic, positive recurrent Harris chain is called *Harris ergodic*. For a Harris ergodic chain we have

$$P^n(x,F) \to \pi(F) \ \text{ for all } x \in E, \ \text{ measurable} F \subseteq E. \tag{3.41}$$

We now consider extreme value theory for Harris chains, and for functions defined on Harris chains. One obvious question is whether such processes satisfy Condition D. This question would be easy to answer if it were possible to make statements of the form

$$\Pr\{X_{n+1} \le x \mid X_j \le x,\ j = 1,...,n\} = \Pr\{X_{n+1} \le x \mid X_n \le x\}. \tag{3.42}$$

However (3.42) is in general false — this point has caused some confusion in the literature. The correct statement is due to O'Brien [172]. If the chain is Harris ergodic and in its stationary distribution, then it is strong mixing (recall (3.3)) with mixing function

$$g(k) = \int_E ||P^k(x,\cdot) - \pi(\cdot)||\pi(dx) \tag{3.43}$$

which tends to 0 by (3.41). Here $||\cdot||$ denote total variation of a measure, i.e.

$$||\mu|| = \int_E |\mu(dx)| = \sup_{f:\ |f|\le 1} \int |f(x)|\mu(dx).$$

To see (3.43), suppose A and B are events depending on $\{X_j,\ j \le n\}$, $\{X_j,\ j \ge n+k\}$ respectively. Then A and B are conditionally independent given X_n and X_{n+k}. Moreover, the marginal distributions of X_n and X_{n+k} are each π (by stationarity), while the joint distribution is πP^k. Hence

$$\begin{aligned} \Pr(AB) &= \int\int_{E\times E} \Pr\{A \mid X_n = x\} \Pr\{B \mid X_{n+k} = y\} \pi(dx) P^k(x,dy), \\ \Pr(A)\Pr(B) &= \int\int_{E\times E} \Pr\{A \mid X_n = x\} \Pr\{B \mid X_{n+k} = y\} \pi(dx)\pi(dy), \end{aligned}$$

so $|\Pr(AB) - \Pr(A)\Pr(B)|$ is bounded by the total variation of $\pi(dx)\{P^k(x,dy) - \pi(dy)$, which is (3.43).

In the case of a periodic chain, strong mixing and condition D do not hold, but

O'Brien [172] defined a version of these conditions which is satisfied by periodic, postivie recurrent Harris chains, and which also suffices for the usual extreme value results.

Another question is what happens if the Markov chain starts from an arbitrary initial state instead of its stationary distribution. Suppose $M_n = \max\{f(X_1), ..., f(X_n)\}$ when the process is started from an arbitrary initial state, and M'_n is the same thing for the stationary process. O'Brien showed that

$$\Pr\{M_n \le u_n\} - \Pr\{M'_n \le u_n\} \to 0$$

whenever $\{u_n\}$ is such that

$$\Pr\{f(X_k)\} \le u_n\} \to 0, \text{ for each } k \ge 1.$$

Rootzén [201] took a different approach to the study of Harris chains, breaking them up into regenerative cycles, as device first used by Berman [16] and Anderson [3]. Suppose $\{X_k\}$ is a stationary regenerative process with mean cycle length $\mu < \infty$. This includes positive recurrent Harris chains as well as processes defined on such chains. Let Z_i denote the i'th cycle maximum

$$Z_i = \max\{X_n \ : \ T_{i-1} < n \le T_i\} \quad (i \ge 1).$$

If there are $S(n)$ cycles amongst the first n observations (so that $T_{S(n)} \le n < T_{S(n)+1}$) then M_n may be approximated by

$$\max\{Z_i \ : \ 1 \le i \le S(n)\}$$

and this in turn may be approximated by the maximum over n/μ values of Z_i. Recall $S(n)/n \to 1/\mu$ almost surely, by standard renewal theory. If the Z_i are independent we have

$$\Pr\{M_n \le u_n\} \to e^{-\tau} \text{ if and only if } \frac{n}{\mu}\Pr\{Z_1 > x\} \to \tau. \tag{3.44}$$

In this case the extremal index of the process is given by

$$\theta = \frac{1}{\mu}\lim_{x \to x^*}\frac{\Pr\{Z_1 > x\}}{\Pr\{X_1 > x\}} \tag{3.45}$$

provided the limit exists.

Analogues of the results in Section 3.4 also hold in such processes. Suppose $Y(u)$ denotes the number of exceedances for each level u during a single cycle and suppose

$$\pi(j) = \lim_{u \to x^*}\Pr\{Y(u) = j \mid Y(u) > 0\} \quad (j = 1, 2, ...,)$$

exists. Define $\phi(t) = \sum_j e^{-tj}\pi(j)$. Let $\{u_n\}$ be such that $\Pr\{M_n \le u_n\} \to e-\nu$ and let N_n denote the point process which puts a point at $\frac{j}{n}$ whenever $X_j > u_n$, $1 \le j \le n$, Then N_n converges on $[0,1]$ to a limiting point process N with Laplace transform (3.27).

In the case that the Z_i are 1-dependent instead of independent, the results are more complicated. For example, the extremal index is now defined by

$$\theta = \frac{1}{\mu} \lim_{x \to x^*} \frac{\Pr\{Z_1 > x,\ Z_2 \le x\}}{\Pr\{X_1 > x\}} \tag{3.46}$$

instead of (3.45).

As an illustration of these concepts, let us again consider example 3 of section 3.2. The process is obviously independent regenerative, with a new cycle beginning whenever $W_n = 1$. Let M denote the length of a cycle, Z the maximum value of X over one cycle. Also write $\mu = \mathrm{E}\{M\} < \infty$. Then

$$\begin{aligned}
\Pr\{M > m\} &= \prod_{i=1}^{m} \left(\frac{i}{i+1}\right)^{\beta} = (m+1)^{-\beta}, \\
\Pr\{M = m\} &= m^{-\beta} - (m+1)^{-\beta} \sim \beta m^{-\beta-1} \quad (m \to \infty), \\
\Pr\{Z \le z \mid M = m\} &= \exp\left(-\sum_{j=1}^{m} j z^{-\alpha}\right) = \exp\left\{-\frac{m(m+1)}{2} x^{-\alpha}\right\}.
\end{aligned}$$

Also define $V_z = z^{-\alpha} M^2/2$. Calculations given below establish that:–

(i) The tail distribution of Z is given by

$$\Pr\{Z > z\} \sim \beta 2^{-\beta/2-1} z^{-\alpha\beta/2} \int_0^{\infty} v^{-\beta/2-1}(1-e^{-v})dv. \tag{3.47}$$

(ii) The conditional density of V_z given $Z > z$ converges as $z \to \infty$ to

$$K_1 v^{-\beta/2-1}(1-e^{-v}) \quad (0 < v < \infty) \tag{3.48}$$

where K_1 is a normalising constant.

(iii) The conditional density of V_z given $Z = z$ converges as $z \to \infty$ to

$$\frac{v^{-\beta/2} e^{-v}}{\Gamma(1-\beta/2)} \quad (0 < v < \infty). \tag{3.49}$$

(iv) Fixing z, $V_z = v$ and hence M, let $\eta_z(v)$ denote a point process with points at $X_j^{-\alpha\beta/2} z^{\alpha\beta/2}$, $1 \le j \le M$, where $(X_1, ..., X_m)$ denotes a single cycle. As $z \to \infty$ with v fixed, $\eta_z(v)$ converges to a Poisson process with intensity

$$\lambda(y; v) = \frac{2v}{\beta} y^{2/\beta - 1}, \quad 0 < y < \infty. \tag{3.50}$$

Rewrite (3.47) as $\Pr\{Z > z\} \sim K_0 z^{-\alpha\beta/2}$, then define

$$u_n(\tau) = \left(\frac{nK_0}{\mu\tau}\right)^{2/\alpha\beta}.$$

Thus

$$\frac{n}{\mu}\Pr\{Z > u_n(\tau) \to \tau$$

and so

$$\Pr\{M_n \le u_n(\tau)\} \to e^{-\tau}. \tag{3.51}$$

Alternatively, let $a_n = u_n(1)$ and write

$$\Pr\{M_n \le a_n x\} \to \exp(-x^{-\alpha\beta/2}).$$

Note that $M_n = O_p(n^{2/\alpha\beta})$ whereas (Section 3.2) we have $\hat{M}_n = O_p(n^{1/(\alpha\beta-\alpha)})$. Since

$$\frac{2}{\alpha\beta} < \frac{1}{\alpha\beta-\alpha}$$

it follows that $M_n << \hat{M}_n$, providing direct proof that the extremal index is 0.

Now consider the point process of exceedances of a single level, as in Section 3.4. Let N_n bethe point process that puts a point at $\frac{j}{n}$ whenever $X_j > u_n(\tau),\ 1 \le j \le n$. We apply (ii) and (iv) with $z = u_n(\tau)$. Given $V_z = v$, the number of exceedances in a single cycle converges to a Poisson distribution with mean

$$\int_0^1 \lambda(y,v)dy = v.$$

Hence the number of exceedances in a single cycle, conditioned on being at least 1, has distribution

$$\pi(j) = \int_0^\infty \frac{v^j e^{-v}}{(1-e^{-v})j!} \cdot K_1 v^{-\beta/2-1}(1-e^{-v})dv \tag{3.52}$$

$$= \frac{K_1\Gamma(j-\beta/2)}{j!}, \quad j = 1,2,\ldots \tag{3.53}$$

This is a proper distribution but with infinite mean.

In this case the point process N_n converges to a compound Poisson process N, in which the clusters form a Poisson process of unit intensity on $(0,1)$, and the cluster sizes are independently distributed according to (3.53).

Now consider the two-dimensional point process of Section 3.5. Since

$$u_n^{-1}(x) = nK_0\mu^{-1}x^{-\alpha\beta/2},$$

we define the process N_n on $(0,1)\times(0,\infty)$ by putting a point at each of

$$\left(\frac{k}{n}, \frac{nK_0}{\mu}X_k^{-\alpha\beta/2}\right), \quad k = 1,2,\ldots,n.$$

We can also define a process $\tilde{N}_n$, which consists of those points of N_n which correspond to cycle maxima.

Define N and $\tilde{N}$ to be the limits of N_n and $\tilde{N}_n$ as $n \to \infty$. By standard arguments, $\tilde{N}$ will be a Poisson process with unit intensity. Let the points of $\tilde{N}$ be (S_i, T_i), $i = 1, 2, \ldots$ As explained in Section 3.5, the process N consists of points $(S_i,\ T_i Y_{ij})$ where, for each i, we have that $(Y_{ij},\ j \geq 1)$ is an independent point process on $[1, \infty)$ with an atom at 1. Intuitively, the points $T_i Y_{ij}$ $(j \geq 1)$ correspond to all the points in the cycle that gave rise to T_i. This is a cycle in which the maximum is $Z_1 = z_i = (nK_0/\mu T_i)^{2/\alpha\beta}$. The Y_{ij}'s correspond to values of $(X_j/z)^{-\alpha\beta/2}$ where X_j's are other points in the cycle that gave rise to T_i. Let M_i denote the length of this cycle, and define $V_i = z_i^{-\alpha} M_i^2/2$. The density of V_i is given asymptotically by (3.49).

The conditional distribution of $\{Y_{ij},\ j \geq 1\}$ given Z_i and V_i is given by property (iv), modified by the knowledge that min(Y_{ij})=1. Thus, $Y_{i1} = 1$ and $Y_{ij},\ j \geq 2$ are the points of a Poisson process on $(1, \infty)$ with intensity given by (3.51) in this range. So, if f is a non-negative function on $(0,1) \times (0, \infty)$, by (3.20) we have

$$\begin{aligned} & \mathrm{E}\left\{ \exp\left(-\sum_i f(S_i, T_i Y_{ij}) \right) \mid S_i = s,\ T_i = t,\ V_i = v \right\} \\ = \quad & \exp\left[-f(s,t) - v \int_1^\infty \left\{ 1 - e^{-f(s,ty)} \right\} \frac{2}{\beta} y^{2/\beta - 1} dy \right]. \end{aligned}$$

Since $\mathrm{E}\{\exp(-\theta V_i)\} = (1+\theta)^{\beta/2-1}$ from (3.49), we have

$$\begin{aligned} & \mathrm{E}\left\{ \exp\left(\sum_j f(S_i, T_i Y_{ij}) \right) \mid S_i = s,\ T_i = t \right\} \\ = \quad & e^{-f(s,t)} \left\{ 2 + \int_1^\infty e^{-f(s,ty)} \frac{2}{\beta} y^{\beta/2-1} dy \right\}^{\beta/2-1} \\ = \quad & e^{-g(s,t)} \quad \text{(say)}. \end{aligned} \tag{3.54}$$

Finally, for the Laplace transform of the whole process we have

$$\begin{aligned} & \mathrm{E}\left\{ \exp\left(\sum_i \sum_j f(S_i, T_i Y_{ij}) \right) \right\} \\ = \quad & \mathrm{E}\left\{ \exp\left(-\sum_i g(S_i, T_i) \right) \right\} \\ = \quad & \exp\left\{ -\int_0^\infty \int_0^1 \left(1 - e^{-g(s,t)} \right) ds dt \right\}, \end{aligned}$$

with g related to f as in (3.54). This is the Laplace transform of an infinitely divisible process.

For this example, in contrast to example 2 of sections 3.4 and 3.5, the point processes N_n converge to a limiting point process N without any renormalisation. The process N has only finitely many points in any finite set with probability one, but, because $\mathrm{E}(V_i) = +\infty$, the expected number of points in any finite set is ∞. This

shows that both kinds of behaviour are possible when the extremal index is 0, i.e. N_n may converge to a limiting point process or $a_n N_n$ (for some sequence $a_n \to \infty$) may converge to a more general random measure.

Proof of the statements (i)–(iv).

First calculate, for $0 \le w_1 < w_2 < \infty$,

$$\begin{aligned}
& \Pr\{w_2^{-1/\alpha} < Z < w_1^{-1/\alpha},\ v_1 < V_z < v_2\} \\
= & \sum_{\lfloor\sqrt{v_1 z^\alpha}\rfloor}^{\lceil\sqrt{v_2 z^\alpha}\rceil} \Pr\{M = m\} \left[\exp\left\{-\frac{w_1 m(m+1)}{2} z^{-\alpha}\right\} - \exp\left\{-\frac{w_2 m(m+1)}{2} z^{-\alpha}\right\}\right] \\
\sim & \int_{\sqrt{v_1 z^\alpha}}^{\sqrt{v_2 z^\alpha}} \beta m^{-\beta-1} \left[\exp\left\{-\frac{w_1 m^2}{2} z^{-\alpha}\right\} - \exp\left\{-\frac{w_2 m^2}{2} z^{-\alpha}\right\}\right] dm \\
= & \beta 2^{-\beta/2-1} z^{-\alpha\beta/2} \int_{v_1}^{v_2} v^{-\beta/2-1} \{exp(-vw_1) - exp(-vw_2)\}\, dv. \qquad (3.55)
\end{aligned}$$

Setting $w_1 = 0,\ w_2 = 1$ in (3.55) gives

$$\Pr\{Z > z,\ v_1 < V_z < v_2\} \sim \beta 2^{\beta/2-1} z^{-\alpha\beta/2} \int_{v_1}^{v_2} v^{-\beta/2-1}(1 - e^{-v}) dv. \qquad (3.56)$$

Setting $v_1 = 0,\ v_2 = \infty$ gives (3.47). Dividing (3.56) by (3.47) gives the asymptotic conditional distribution of V_z given $Z > z$, which has density (3.48). Taking the limit as $w_1 \to 1, w_2 \to 1$ in (3.55) so as to obtain a conditional density of V given $Z = z$, we get (3.49).

Finally, we must verify statement (iv). Consider a single cycle $X_1, ..., X_n$ and let I_j be the indicator function of the event $y_1 < X_1^{-\alpha\beta/2} z^{\alpha\beta/2} < y_2$. Then

$$\begin{aligned}
\lambda_j(v) &= \mathrm{E}\{I_j \mid V_z = v\} \\
&\sim \begin{cases} jz^\alpha \left(y_2^{2/\beta} - y_1^{2/\beta}\right) \text{ if } j \le M = (2vz^\alpha)^{1/2}, \\ 0 \text{ otherwise.} \end{cases}
\end{aligned}$$

Thus

$$\sum_{j=1}^{m} \lambda_j(v) \sim v\left(y_2^{2/\beta} - y_1^{2/\beta}\right), \qquad (3.57)$$

$$\sum_{j=1}^{m} \lambda_j^2(v) \sim O\left(M^3 z^{-2\alpha}\right) = O\left(z^{-\alpha/2}\right) \to 0. \qquad (3.58)$$

By (3.57), the expected number of points of $\eta_z(v)$ in (y_1, y_2) is consistent with (3.50). By (3.58) and well-known results about convergence to Poisson distributions, the number of points of $\eta_z(v0$ in (y_1, y_2) is asmptotically Poisson and therefore, in particular,

$$\begin{aligned}
\Pr\{\eta_z(v) \text{ has no points in } (y_1, y_2)\} &\to \exp\left(-v \int_{y_1}^{y_2} \frac{2}{\beta} y^{2/\beta-1} dy\right) \\
&= \exp\left\{-v\left(y_2^{2/\beta} - y_1^{2/\beta}\right)\right\}.
\end{aligned}$$

This property is easily extended from a single interval (y_1, y_2) to a union of dispoint intervals.

Thus suffices, by [133] or results quoted in the Appendix of [143], for $\eta_z(v)$ to converge to a Poisson process, as required.

3.7 Computational Methods for the Extremal Index in Markov Chains and Extensions

So far, our discussion of Markov chains has been concerned with fairly abstract properties, inllustrated by a few artificial examples. In this section, we describe some techniques that have been used for markov chains and extensions (k'th-order or higher-dimensional Markov chains) where some direct but non-trivial calculations are possible.

3.7.1 Markov chains derived from bivariate extreme value distributions

A Markov chain is specified by the conditional probabilities $\Pr\{X_{n+1} \leq y \mid X_n = x\}$ for any pair (x, y) in the state space. If the Markov chain is stationary and the marginal distribution of X_n is known, this is equivalent to specifying the joint distributions of (X_n, X_{n+1}). To develop an extreme value theory for such cases, a natural class of models for the joint distribution are the *bivariate extreme value distributions*. These will be developed in more detail in Chapter **??**, but we will state here the specific results that are needed. The results of htis section follow Smith [227].

As a specific example, consider the class of models

$$\Pr\{X_n \leq x, X_{n+1} \leq y\} = \exp\left\{-\left(e^{-rx} + e^{-ry}\right)^{1/r}\right\} \tag{3.59}$$

for $1 \leq r < \infty$.

The marginal distributions of this model are $\Pr\{X_n \leq x\} = \exp(-e^{-x})$, the standard Gumbel distribution from Chapter 1, but any Markov chain with GEV marginals can be transformed to this through a one-dmensional transformation of each variable. Therefore, there is no loss of generality in assuming Gumbal marginals. As for the joint distribution in (3.59), this is an example of the *logistic* family of bivariate extreme value distributions, which is one of a wide class of bivariate extreme value distributions that will be developed in more detail in Chapter 4. As will be seen, the calculations given here for the specific model (3.59) in fact apply to a very wide class of models derived from bivariate extreme value distributions. As for the specific model (3.59), it is readily checked that the case $r = 1$ corresponds to independence of X_n and X_{n+1}, while the limit $r \to \infty$ is the limit where $X_n = X_{n+1}$ with probability 1 (but in that case, all the sample maxima are exactly equal to X_1, so we omit this trivial case from consideration).

From (3.59), it is possible to derive

$$\begin{aligned} \Pr\{X_{n+1} \leq x+z \mid X_n = x\} &= \exp\left[-e^{-x}\left\{1-(1+e^{-rz})^{1/r}\right\}\right](1+e^{-rz})^{1/r-1} \\ &\to (1+e^{-rz})^{1/r-1} \text{ as } x \to \infty. \end{aligned} \tag{3.60}$$

The interpretation of (3.60) is that the jumps of the process, given a starting value $X_n = x$, are almost independent of x as $x \to \infty$. In other words, asymptotically at high levels, the process behaves like a *random walk*.

The density of the jumps in this random walk can be derived from (3.60), and for $r > 1$ is given by

$$h(z) \quad = \quad (r-1)e^{-rz}(1+e^{-rz})^{1/r-2}, \ -\infty < z < \infty, \tag{3.61}$$

a distribution which can be shown to have negative mean.

In fact, there is a general result based on Resnick [194], Proposition 5.15, that if (X_n, X_{n+1}) follow a bivariate extreme value distribution with standard Gumbel margins, then

$$\frac{1 - \Pr\{X_n \le u + x_1, \ X_{n+1} \le u + x_2\}}{1 - \Pr\{X_n \le u, \ X_{n+1} \le u\}}$$

converges to a limit as $u \to \infty$. From this, it can be shown that for any $\delta > 0$ and $z \in \mathbb{R}$, the limit

$$\lim_{u \to \infty} \Pr\{X_{n+1} > u + z \mid u - \delta < X_1 < u + \delta\}$$

exists, though it may be 0 for all z (this corresponds to independence, which is of course an important special case). The results of [227] essentially hold if there is a density equivalent of this result: if $q(x,y)$ denotes the transition density of the Markov chain (the density of X_{n+1}, evaluated at y, given $X_n = x$), then we assume that the limiting density

$$h(z) \quad = \quad \lim_{u \to \infty} q(u, u+z) \tag{3.62}$$

exists for $z \in \mathbb{R}$.

In fact [227] assumed something a bit more general than (3.62), the existence of a c.d.f. $H(z)$ on $-\infty \le z < \infty$ whose derivative $h(z)$ is given by (3.62) on $-\infty < z < \infty$. The point of this is to admit the possibility that the jump distribution has mass at $-\infty$ — in particular, in the independent case, all the mass is at $-\infty$.

Next, we consider the definition of the extremal index in this case. The definition is based on (3.9), but in an older form first proposed by [171]. In place of (3.9), we define θ by

$$\theta \quad = \quad \lim_{p \to \infty} \lim_{u \to \infty} \Pr\{X_i \le u, \ 2 \le i \le p \mid X_1 > u\}. \tag{3.63}$$

It can be shown that the definitions (3.9) and (3.63) are equivalent if

$$\lim_{p \to \infty} \lim_{n \to \infty} \sum_{k=p}^{p_n} \Pr\{X_k > u_n \mid X_1 > u_n\} \quad = \quad 0, \tag{3.64}$$

and henceforth we assume (3.64).

From this point, [227] argued as follows. Define a function

$$\begin{aligned}\phi(p,x,u) &= \Pr\left\{X_2 < u,\ldots,X_{p-1} < u, X_p > u \mid X_1 = u-x\right\} \\ &= \int_0^\infty \ldots \int_0^\infty \int_{-\infty}^0 q(u-x,u-x_2)\ldots q(u-x_{p-1},u-x_p)dx_p\ldots dx_2 \\ &\to \int_0^\infty \ldots \int_0^\infty \int_{-\infty}^0 h(x-x_2)\ldots h(x_p-x_{p-1})dx_p\ldots dx_2 \\ &= \phi(p,x) \text{ say,}\end{aligned}$$

where the convergence assumes that it is valid to take limits under the integral sign.

However, equation (3.64) implies

$$1-\theta = \lim_{p\to\infty}\lim_{u\to\infty}\sum_{j=2}^{p}\int_{-\infty}^{0}\phi(j,x,u)\frac{f(u-x)}{1-F(u)}dx$$

where f is the standard Gumbel density and we can easily calculate $\frac{f(u-x)}{1-F(u)} \to e^x$ as $u \to \infty$. If we can again justify the interchange of limits and integration, we then have

$$1-\theta = \lim_{p\to\infty}\sum_{j=2}^{p}\int_{-\infty}^{0}\phi(j,x)e^x dx. \tag{3.65}$$

[227] shows that these limiting operations are valid under two conditions, in addition to (3.62 and (3.64):

(i) there exists u^* such that, for all M, $q(u,u+y)$ is uniformly bounded over $u \geq u^*,\ y \geq -M$,

(ii) $\lim_{M\to\infty}\lim_{u\to\infty}\sup_{x\leq u-M}\Pr\{X_2 > u \mid X_1 = x\} = 0$.

Although the above argument shows theoretically how the etxremal index may be calculated, it is not the most convenient practical solution. [227] defined a sequence of functions $Q_p(x),\ p \geq 1, x \in \mathbb{R}$ by $Q_1 \equiv 1$ and

$$Q_p(x) = \int_0^\infty Q_{p-1}(y)H(x-dy) \tag{3.66}$$

writing the integral in Stieltjes form to allow for the possibility $H(-\infty) > 0$. As $p \to \infty,\ Q_p \to Q$ where Q satisfies the Wiener-Hopf equation

$$Q(x) = \int_0^\infty Q(y)H(x-dy) \tag{3.67}$$

subject to the normalizing condition $\lim_{x\to\infty}Q(x) = 1$. The extremal intex is then given by

$$\theta = \int_{-\infty}^{0} e^x Q(x)dx. \tag{3.68}$$

[227] essentially evaluated θ by iterating (3.66) to convergence, and then evaluating (3.68), using numerical integration on up to 2^{14} sampling points.

Hooghiemstra and Meester [117] presented an alternative method based on Grübel's [92] algorithm for solving the Wiener-Hopf equation arising from a G/G/1 queue. Their algorithm is undoubtedly the most efficient solution to this problem, but the method does not appear to extend to more complicated models such as those in Section 3.7.2.

Numerical example.

Suppose Z_0 has an exponential distribution with mean 1 and $Z_1 - Z_0, Z_2 - Z_1, \ldots$ are IID with density h given by (3.61). We approximate the extremal index by $\Pr\{Z_1 < 0, \ldots, Z_p < 0\}$ for some large p. In practice, we have found that p sometimes needs to be as large as 250 (larger as r grows) to ensure an accurate approximation. The following R program evaluates this quantity by simulation:

```
z=-log(runif(nsim))-log(runif(nsim)^(r/(1-r))-1)/r
z=cbind(z,z-log(runif(nsim)^(r/(1-r))-1)/r)
for(j in 3:p)(z=cbind(z,z[,j-1]-log(runif(nsim)^(r/(1-r))-1)/r))
th=mean(apply(z, 1, max, na.rm=TRUE)<0)
print(c(th,sqrt((th*(1-th))/nsim)))
```

The user must simply the values of r, p and the number of simulations *nsim*. The last line gives the estimated value of θ and its simulation-based standard error. Some sample results are in Table 3.1.

r	2	3	4	5
θ	0.3285	0.1581	0.0924	0.0604

Table 3.1 *Extremal index for the Markov chain defined by the logistic dependence model.*

In the case $r = 5$, [227] claimed the value $\theta = 0.0616$ but [117] stated $\theta = 0.0604$. The present calculations (based on approximatley 6 million simulations) agreed with those of [117].

Addendum: An amusing counterexample.

The following example (due to an anonymous referee to [227]) shows that technical conditions like (i) and (ii), while troublesome to check, cannot be ignored. Consider a sequence $\{Z_n\}$ of independent Bernoulli r.v.s where $\Pr\{Z_n = 0\} = \Pr\{Z_n = 1\} = \frac{1}{2}$. Let X_1 be standard Gumbel; for $n \geq 1$, if $Z_n = 1$ then $X_{n+1} = -X_n$, if $Z_n = 0$ then X_{n+1} is again distributed as standard Gumbel independently of all prior variables. For any given value of X_n, there is probability $\frac{1}{4}$ that $X_{n+2} = X_n$; it follows immediately from O'Brien's definition that the extremal index is $\frac{3}{4}$. But $\lim_{u\to\infty} \Pr\{X_{n+1} > u + z \mid X_n = u\} = 0$ for any $z \in \mathbb{R}$, so an attempt to apply the random walk argument would yield an extremal index of 0.

3.7.2 Extension to k'th-order Markov chains

A k'th-order Markov chain is an extension of the concept of a Markov chain in which the conditional distribution of an observation, given the past, is dependent only on

the most recent k values of the process. The case $k = 1$ corresponds to the standard definition of a Markov chain. This assumption may be written as

$$\Pr\left\{X_{n+1} \le x \mid X_{n-j},\ j \ge 0\right\} \quad = \quad \Pr\left\{X_{n+1} \le x \mid X_{n-j},\ j = 0,\ldots,k-1\right\} \quad (3.69)$$

defined on a state space $X_n \in \mathscr{S}$ (which, for the cases considered here, is usually $\mathbb{R}$).

An equivalent definition is that the process $Y_n = \begin{pmatrix} X_{n-k+1} & X_{n-k+2} & \ldots & X_n \end{pmatrix}$ is a Markov chain on the state space $\mathscr{S}^k$. Thus, there is a direct connection between k'th-order Markov chains and Markov chains in a higher-dimensional space.

Extensions of the theory in [227] were presented by various authors. We highlight here the contributions of Perfekt [176, 177] and Yun [263].

The key idea behind [227] is that, for a Markov chain with Gumbel marginal distributions and pairwise distributions defined by a two-dimensional extreme value distribution, the distribution of successive values of $X_{n+1} - X_n$, given a starting value at some high level u, is asymptotically independent of u as $u \to \infty$. If, instead, the process is kth-order Markov, we might expect the same to be true of the k-dimensional differences $X_{n+1} - X_n, X_{n+2} - X_{n+1}, \ldots, X_{n+k} - X_{n+k-1}$. This is indeed the correct intuition but the paper [263] made other extensions as well, including allowing the limiting marginal distributions to be of Generalized Extreme Value form with ξ not necessarily equal to 0, and also assuming the $(k+1)$-dimensional joint distributions are in the domain of attraction of a multivariate extreme value distribution (MEVD), rather than assuming they are of exactly MGEV form. It should be noted that similar generalizations had earlier been made in [176, 177].

[263] assumed that the process is stationary with marginal distributions, say $F_1(x) = \Pr\{X_n \le x\}$, satisfying

$$\lim_{u \uparrow \omega_{F_1}} \frac{1 - F_1(u + g(u)x)}{1 - F_1(u)} \quad = \quad (1 + \xi x)^{-1/\xi} \text{ whenever } 1 + \xi x > 0 \quad (3.70)$$

where $g(u)$ satisfies the same regularity conditions as in classical one-dimensional extreme value theory; in particular, $g(u) > 0$. The corresponding pth-order property, based on results in [152], is

$$\lim_{u \uparrow \omega_{F_1}} \frac{1 - F_p(u + g(u)\mathbf{x})}{1 - F_1(u)} \quad = \quad -\log G_p(\mathbf{x}) \text{ whenever } 1 + \xi \mathbf{x} > \mathbf{0} \quad (3.71)$$

where G_p is a p-dimensional MEVD. Here, (3.71) is assumed for $p = 1, \ldots, k+1$. In addition, we assume that there are corresponding results for densities in the form of

$$\lim_{u \uparrow \omega_{F_1}} \frac{g(u) f_1(u + g(u)x_1)}{1 - F_1(u)} \quad = \quad (1 + \xi x_1)^{-1/\xi - 1} \text{ whenever } 1 + \xi x > 0, \quad (3.72)$$

$$\ell_j(x_{j+1}; \mathbf{x}_j) \quad = \quad \lim_{u \uparrow \omega_{F_1}} g(u) f_{j+1}(u + g(u)x_{j+1} \mid u + g(u)\mathbf{x}_j) \text{ exists whenever } 1 + \xi \mathbf{x} > \mathbf{0}, \quad (3.73)$$

where, in (3.73), $f_j(\cdot \mid \cdot)$ denotes the conditional density of any value of the process given the j previous values.

Based on these results, [263] established:

Lemma. If f_{k+1} is continuous, then for each $j = 1,\ldots,k$, $(1+\xi x_{j+1})\ell_j(x_{j+1};\mathbf{x}_j)$ must be a function of

$$\nabla\mathbf{x}_{j+1} \quad = \quad \left(\frac{1}{\xi}\left(\frac{1+\xi x_2}{1+\xi x_1}\right),\ldots,\frac{1}{\xi}\left(\frac{1+\xi x_{j+1}}{1+\xi x_j}\right)\right). \tag{3.74}$$

Note that, in the case $\xi = 0$, (3.74) is a function of $x_2 - x_1,\ldots,x_{j+1} - x_j$, thus establishing a direct connection to the results of [227] in this case.

Based on (3.74), [263] defined a function $h_j(y_j;y_1,\ldots,y_{j-1})$ on $\mathbb{R}^j$ by the property

$$h_j\left(\frac{1}{\xi}\log\left(\frac{1+\xi x_{j+1}}{1+\xi x_j}\right);\nabla\mathbf{x}_j\right) \quad = \quad (1+\xi x_{j+1})\ell_j(x_{j+1};\mathbf{x}_j) \text{ where } 1+\xi\mathbf{x}_{j+1} > 0.$$

Yun also defined the corresponding cumulative distribution function by

$$H_j(y;y_1,\ldots,y_{j-1}) \quad = \quad 1 - \int_y^\infty h_j(t;y_1,\ldots,y_{j-1})dt,\ y \in \{-\infty\}\cup\mathbb{R}.$$

In effect, $H_1,\ldots,H_k$ are the conditional distribution functions of an embedded $(k-1)$th-order Markov chain which directly generalizes the random walk of [227]. The notation, including the case $y = -\infty$, is intended to accommodate the case that was earlier seen in [227], that the process may jump to $-\infty$ which then becomes an absorbing state of the embedded Markov chain.

We generate a sequence $Y_1 \sim H_1(\cdot)$; $Y_j \mid (Y_1,\ldots,Y_{j-1}) \sim H_j(\cdot;Y_1,\ldots,Y_{j-1})$ for $j = 2,\ldots,k$, $Y_j \mid (Y_{j-k+1},\ldots,Y_{j-1}) \sim H_k(\cdot;Y_{j-k+1},\ldots,Y_{j-1})$ for $j > k$, where at any stage, if $Y_j = -\infty$ then all subsequent $Y_{j'},\ j' > j$ are $-\infty$ as well. Finally, define

$$Z_n \quad = \quad Y_1 + \ldots + Y_n,\ n = 1,2,\ldots \tag{3.75}$$

At this point, the argument splits into two cases. The simpler case is that of "no infinite jumps", i.e. $H_j(-\infty;\mathbf{x}_{j-1}) = 0$ for every $j = 1,\ldots,k$ and $\mathbf{x}_{j-1}$. As in the case of a simple Markov chain, if (3.64) holds than we can define the extremal index by (3.63). Assuming conditions (3.70–3.73), (3.64) and the no infinite jumps condition, Lemma 2.2 of [263] shows that the extremal index is given by

$$\theta \quad = \quad \lim_{p\to\infty} \Pr\left\{Z_1 \le -T,\ldots,Z_p \le -T\right\} \tag{3.76}$$

where T is an exponential random variable of mean 1, independent of $\{Z_n,\ n \ge 1\}$.

The more complex case is where jumps to $-\infty$ are allowed. In that case, Yun makes additional assumptions:

Assumption B. Suppose there exists $u^* < \omega_{F_1}$ such that the class

$$\left\{\frac{g(u)f_1(u+g(u)x_1)}{1-F(u)} :\ u^* \le u < \omega_{F_1}\right\}$$

of functions of x_1 is locally uniformly integrable over $(x_{\Omega_\xi}, x^*_{\Omega_\xi})$, and that, for each $j = 1,\ldots,k$ and for every fixed $\mathbf{x}_j$ with $1+\xi\mathbf{x}_j > \mathbf{0}$, there exists a $u^*_j(\mathbf{x}_j) < \omega_{F_1}$ such that the class

$$\left\{g(u)f_{j+1}(u+g(u)x_{j+1} \mid u+g(u)\mathbf{x}_j) :\ uj^*(\mathbf{x}_j) \le u < \omega_{F_1}\right\}$$

of functions of x_{j+1} is locally uniformly integrable over $(x_{\Omega_\xi}, x^*_{\Omega_\xi})$. Here, x_{Ω_ξ} and $x^*_{\Omega_\xi}$ are the endpoints of the distribution $\Omega_\xi(x) = \exp\{-(1+\xi x)^{-1/\xi}\}$.

Yun remarks that (3.72), (3.73) and Assumption B hold automatically when F_{k+1} is exactly a MEVD.

With all these preliminaries, Yun's main theorem (his Theorem 3.1) is as follows:

Theorem. Let $\{X_n,\ n \ge 1\}$ be a kth-order stationary Markov chain, and let F_{k+1} be the d.f. of $(X_1,\ldots,X_{k+1})$ having continuous p.d.f. f_{k+1} such that F_{k+1} is in the domain of attraction of an MEVD G_{k+1} with auxiliary function g from (3.71), where G_{k+1} has equal univariate margins Ω_ξ for some $\xi \in \mathbb{R}$. Suppose (3.72), (3.73) and Assumption B hold and that

$$\lim_{M\to\infty} \limsup_{u\uparrow\omega_{F_1}} \sup \left[\Pr\{X_{k+1} > u \mid \mathbf{X}_k = \mathbf{x}_k\} :\ \min_{1\le i\le k} x_i \le u - g(u)(1-M^{-\xi})/\xi\right] \quad = \quad 0. \tag{3.77}$$

Let $\{u_n(\tau),\ \tau > 0\}$ satisfy $n\{1-F_1(u_n(\tau))\} \to \tau$. Assume that $D(u_n(\tau))$ holds for each $\tau > 0$ and that, for some $\tau_0 > 0$, (3.64) holds with $u_n = u_n(\tau_0)$ and $p_n = o(n)$, $n\alpha_{n,\ell_n} = o(p_n)$, $\ell_n = o(p_n))$ where the function $\alpha_{n,\ell}$ is as in (3.5). Then, $\{X_n\}$ has extremal index θ given by

$$\theta \quad = \quad \log(G_k(\mathbf{0})/G_{k+1}(\mathbf{0})) - \Pr\left\{\max_{1\le i\le k} Z_i \le -T,\ \sum_{i\ge k+1} Z_i > -T\right\} \tag{3.78}$$

where $\{Z_n,\ n \ge 1\}$ is the kth-order Markov chain defined by (3.75) and T is an exponentially distributed random variable with mean 1 which is independent of $\{Z_n\}$.

However, Yun also showed that if the condition (3.77) is strengthened to

$$\lim_{M\to\infty} \limsup_{u\uparrow\omega_{F_1}} \sup \left[\Pr\left\{X_{j+1} > u \mid \mathbf{X}_j = \mathbf{x}_j\right\} :\ \min_{1\le i\le k} x_i \le u - g(u)(1-M^{-\xi})/\xi\right] \quad = \quad 0 \tag{3.79}$$

for each $j = 2,\ldots,k$, then the earlier formula (3.76) is also valid in this case. It is unclear, to this writer, exactly what feature of the problem makes (3.79) harder to prove than (3.77).

Examples

$(k+1)$*-dimensional logistic model.* Our first example is the direct generalization

of (3.59) to the kth-order case. The corresponding joint distribution for $k+1$ variables is

$$\Pr\{X_1 \le x_1, \ldots, X_{k+1} \le x_{k+1}\} \quad = \quad \exp\left\{ -\left(\sum_{i=1}^{k+1} e^{-rx_i} \right)^{1/r} \right\} \tag{3.80}$$

for $1 \le r < \infty$, and this is a known example of a multivariate extreme value distribution, see e.g. [238]. In this case, [263] showed that the transition probabilities for the embedded chain are given by

$$H_j(y_j; \mathbf{y}_{j-1}) \quad = \quad \left[1 + \frac{\exp(-ry_j)}{1 + \sum_{s=1}^{j-1} \exp\{r(y_s + \ldots + y_{j-1})\}} \right]^{(1/r)-j}, \ \mathbf{y}_j \in \mathbb{R}^j, \ 1 \le j \le k. \tag{3.81}$$

This is a case where the limiting process (3.75) does not have jumps to $-\infty$, therefore the formula (3.76) is valid. Yun [263] gave an explicit simulation algorithm for θ which we present as the following R code:

```
z=matrix(nrow=nsim,ncol=p)
u=runif(nsim)
z[,1]=-log(u^(r/(1-r))-1)/r
if(k>1){
for(j in 2:k){
u=runif(nsim)
z[,j]=-(log(u^(r/(1-r*j))-1)+log(1+as.matrix(exp(-r*z[,1:(j-1)]))%*%rep(1,j-1)))/r
}}
for(j in (k+1):p){
u=runif(nsim)
z[,j]=-(log(u^(r/(1-r*k))-1)+log(as.matrix(exp(-r*z[,(j-k):(j-1)]))%*%rep(1,k)))/r
}
T=-log(runif(nsim))
theta=mean(apply(z, 1, max, na.rm=TRUE)+T<0)
```

Once again, the user is responsible for supplying the values of k, r, p and the number of simulations *nsim*.

As an example, this code was used to compute Table 3.2. The numbers differ slightly from those in Table 1 of [263].

Mixture model.

This example, also taken from [263], illustrates a case where there are jumps to $-\infty$ and the more complicated formula (3.78) is apparently needed.

The model is

$$F_{k+1}(\mathbf{x}_{k+1}) \quad = \quad \exp\left[-\alpha \left\{ \sum_{s=1}^{k+1} (1+\xi x_s)^{-r/\xi} \right\}^{1/r} - (1-\alpha) \sum_{s=1}^{k+1} (1+\xi x_s)^{-1/\xi} \right], \ 1 + \xi \mathbf{x} > 0, \tag{3.82}$$

	k					
$1/r$	1	2	3	4	5	10
0.1	0.017	0.007	0.005	0.004	0.003	0.002
0.2	0.06	0.024	0.015	0.011	0.01	0.006
0.3	0.13	0.057	0.036	0.027	0.022	0.015
0.4	0.22	0.108	0.071	0.054	0.045	0.031
0.5	0.329	0.184	0.128	0.1	0.084	0.06
0.6	0.45	0.285	0.213	0.174	0.151	0.111
0.7	0.581	0.415	0.333	0.284	0.254	0.199
0.8	0.719	0.577	0.497	0.447	0.414	0.347
0.9	0.86	0.771	0.716	0.679	0.652	0.594

Table 3.2 *Extremal index in the higher-order logistic model.*

where $0 < \alpha < 1$, $r > 1$ and $\xi \in \mathbb{R}$. In this case, $H_1(y_1) = 1 - \alpha + \alpha\{1 + exp(-ry_1)\}^{1/r-1}$, $y \geq -\infty$, and (3.81) holds for all $j > 1$. In this case, the extremal index is given by (3.78); writing $\theta = \lim_{p\to\infty} \theta_p$, [263] gives the specific formula

$$\theta_p = \alpha\{(k+1)^{1/r} - k^{1/r}\} + 1 - \alpha - \Pr\left\{\max_{1\leq i\leq k} Z_i \leq -T, \max_{k+1\leq i\leq p} Z_i > -T\right\} \tag{3.83}$$

where T is an independent unit exponential random variable as before.

The following R code is based on the algorithm presented in [263]. As before, the user is left to supply the values of k, p, r, α (`alf`) and *nsim*. Note that the cases $k = 1$, $k > 1$ are treated differently.

```
z=matrix(nrow=nsim,ncol=p)
u=runif(nsim);z[,1]=-log(u^(r/(1-r))-1)/r
if(k>1){
for(j in 2:k){u=runif(nsim)
z[,j]=-(log(u^(r/(1-r*j))-1)+log(1+as.matrix(exp(-r*z[,1:(j-1)]))%*%rep(1,j-1)))/r}
for(j in (k+1):p){u=runif(nsim)
z[,j]=-(log(u^(r/(1-r*k))-1)+log(as.matrix(exp(-r*z[,(j-k):(j-1)]))%*%rep(1,k)))/r}
# additional randomization: with probability 1-alpha, entire row of z is -inf
u=runif(nsim);z[u>alf,]=-10e10}
if(k==1){
u=runif(nsim);z[,1]=-log(u^(1/(1-r))-1)/r
u=runif(nsim);z[u>alf,1]=-10e10
for(j in 2:p){u=runif(nsim)
z[,j]=z[,j-1]-log(u^(1/(1-r))-1)/r
u=runif(nsim);z[u>alf,j]=-10e10}
}
T=-log(runif(nsim));if(k==1)u1=z[,1]+T<0
if(k>1)u1=apply(z[,1:k], 1, max, na.rm=TRUE)+T<0
u2=apply(z[,(k+1):p], 1, max, na.rm=TRUE)+T>0
thetap=alf*((k+1)^(1/r)-k^(1/r))+1-alf-mean(u1&u2)
```

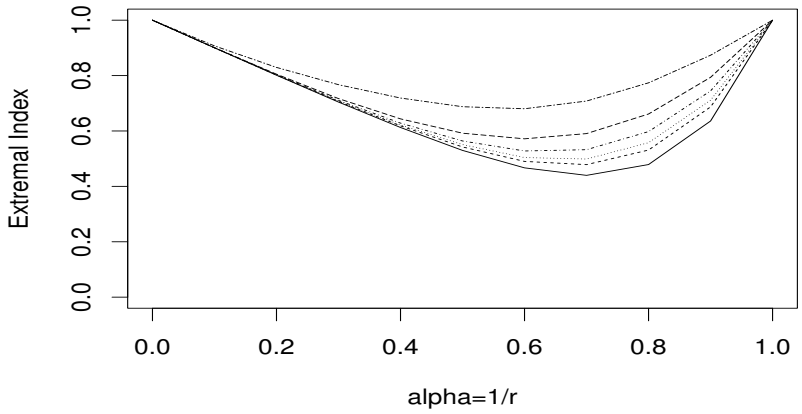

Figure 3.1 *Extremal index for the kth-order mixture model with* $\alpha = 1/r$. *Top to bottom:* $k = 1, 2, 3, 4, 5, 10$. *Adapted from [263].*

For the case $\alpha = 1/r$ and taking $p = 100$, Figure 3.1 shows the simulated extremal index. This was calculated to mimic Figure 2 of [263]; the two figures indeed look very similar.

As far as this writer can tell, formula (3.76) leads to the same numerical results as (3.83), except when $k = 1$.

Conclusions

Sections 3.7.1 and 3.7.2 are intended to illustrate two non-trivial classes of models for which the extremal index may be calculated exactly (to the extent that any simulation result may be considered exact in practice). These models are potentially rich enough to be considered viable dependence models for real time series, such as daily measurements of temperature or air pollution. As a result, they provide a realistic basis for computing extreme value distributions in such datasets.

3.8 Models for Financial Time Series

The autoregressive conditional heterscedastic (ARCH) model was introduced by Engle [67] and extended to the generalized autoregressive conditional heterscedastic (GARCH) model by Bollerslev [21]. They are among the most widely used models for econometric time series such as stock prices and currency exchange rates, allowing for wide fluctations in the variability of log-returns (volatility) while also being faithful to short-term properties of financial time series such as the no-arbitrage condition (that essentially says that there cannot be a combination of trades that would guarantee financal gain). If the price of a commodity on day t is written P_t, then

the log return (which we shall also just call the return) is $X_t = \log \frac{P_t}{P_{t-1}}$. ARCH and GARCH models start with a relationship of the form

$$X_t = \sigma_t Z_t \tag{3.84}$$

where σ_t is a stochastic process (the volatility) and Z_t are typically modeled as independent random variables with a distribution symmetric around 0 and $\mathrm{E}\{Z_t^2\} = 1$ — very often, Z_t is assumed to be a standard normal distribution, $Z_t \sim \mathcal{N}[0,1]$.

One specific and very widely studied example is the GARCH(1,1) model which, following the notation of Mikosch and Stărică [160], is defined by

$$\sigma_t^2 = \alpha_0 + \beta_1 \sigma_{t-1}^2 + \alpha_1 X_{t-1}^2 = \alpha_0 + \sigma_{t-1}^2(\beta_1 + \alpha_1 Z_{t-1}^2) \tag{3.85}$$

with $\alpha_0,\ \alpha_1,\ \beta_1$ all ≥ 0. In many cases, it is found empirically that $\alpha_1 + \beta_1 \approx 1$.

The special case $\beta_1 = 0$ leads to the model

$$X_t^2 = \alpha_0 Z_t^2 + \alpha_1 Z_t^2 X_{t-1}^2 \tag{3.86}$$

which was the original ARCH model of [67], generally regarded as less realistic for financial time series but nevertheless a key model for the development of the theory.

also see [139]

Citations:

[135, 248, 88] for background; [103, 176] for applications to extreme value theory. Random difference equations arise in econometric models (e.g. ARCH, GARCH) so this theory is relevant to understanding the behavior of financial time series.

3.9 Statistical Aspects

The discussion in this chapter raises natural questions about how one would estimate the extremal index in practice. Broadly, there are two approaches: parametric and nonparametric. Parametric methods would take a specific model for dependence, such as the models of Sections 3.7.1 or 3.7.2, fit it to data, and then use the theoretical formulas (via simulation) to estimate θ. Nonparametric methods are based on identifying clusters of high-level exceedances; the reciprocal of the mean cluster size is then taken as the estimator of θ.

3.9.1 Parametric methods

Principal reference: [231, 148]

3.9.2 Nonparametric methods

Focus, particularly, on the methods of [234, 75, 199, 73]

3.10 The Multivariate Extremal Index

[163, 164, 235, 269, 153, 198]

Chapter 4

Multivariate Extremes

4.1 Introduction to Multivariate Extreme Value Theory

The first attempts to extend extreme value theory to two-dimensional cases were due to Geffroy [83], Tiago de Oliveira [242] and Sibuya [217]. The first practical examples were in papers by Gumbel [94, 95], Gumbel and Goldstein [96], Gumbel and Mustafi [97]. Characterizations in $p > 2$ dimensions were first given by Pickands [180, 179], de Haan and Resnick [102] and Deheuvels [54].

A key difference from the one-dimensional case is the absence of any finite-parameter family to represent dependence. Therefore, statistical methods have taken on two forms: (a) parametric methods based on parametric subfamilies of bivariate or multivariate extreme value distributions, (b) nonparametric methods.

However, another distinction that makes multivariate extreme value theory much more complicated than the univariate case is that there essentially three distinct theories of multivariate extremes. The early theories were all concerned with what we now call the *asymptotically dependent* case of multivariate extremes, i.e. distributions where the dependence between components persists even at very high thresholds. A second theory, introduced by Ledford and Tawn [144, 147] is for *asymptotically independent* families, i.e. distributions that tend to independence at very high thresholds but for which the dependence at moderate thresholds is high enough to account for in the statistical models adopted. A third theory, introduced by Heffernan and Tawn [113], is the *conditional* approach, where at least one component is extreme but the others may not be. Recently, new models have been introduced that combine these different approaches. However, for at least the first part of this chapter, we introduced the classical theory, now known as the asymptotically dependent case.

Suppose we have independent, identically distributed (IID) vectors $\mathbf{Y}_i = \begin{pmatrix} Y_{i1} & \dots & Y_{ip} \end{pmatrix}$ for $i = 1, 2, \dots$ Let $M_{nj} = \max\left(Y_{1j}, \dots, Y_{nj}\right)$ for $1 \leq j \leq p$, and define the vector of sample maxima, $\mathbf{M}_n = \begin{pmatrix} M_{n1} & \dots & M_{np} \end{pmatrix}$. The classical theory looks for normalizing constants $\mathbf{a}_n = \begin{pmatrix} a_{n1} & \dots & a_{np} \end{pmatrix}$, $\mathbf{b}_n = \begin{pmatrix} b_{n1} & \dots & b_{np} \end{pmatrix}$ and a nondegenerate p-dimensional limiting distribution function G such that

$$\frac{\mathbf{M}_n - \mathbf{b}_n}{\mathbf{a}_n} \xrightarrow{d} G$$

in the sense that

$$\lim_{n\to\infty} \Pr\left\{ \frac{M_{nj} - b_{nj}}{a_{nj}} \le x_j,\ j = 1,\ldots,p \right\} \quad = \quad G(x_1,\ldots,x_p) \text{ for all } \begin{pmatrix} x_1 & \ldots & x_p \end{pmatrix} \in \mathbb{R}^p. \tag{4.1}$$

Here, $b_{nj} \in \mathbb{R}$ and $a_{nj} > 0$ for each n and j. The resulting G is called a *multivariate extreme value distribution*.

One comment we can make at once: if (4.1) holds, then each of the marginals of G is GEV. That's because if we just isolate the jth component, for some j between 1 and p, we again have that $\Pr\left\{(M_{nj} - b_{nj})/a_{nj} \le x_j\right\}$ converges to a nondegenerate limit, which must therefore be GEV. But if we know the marginal distributions, a famous result known as Sklar's theorem [218] states that we can reduce the problem to that of finding a *copula*,

$$G(x_1,\ldots,x_p) \quad = \quad D_G(G_1(x_1),\ldots,G_p(x_p)) \tag{4.2}$$

where $G_1,\ldots,G_p$ are the marginal distributions of G. This is equivalent to using the probability integral transformation to transform each of the marginal distributions to uniform on $[0,1]$, then D_G is the joint CDF of the resulting distribution.

In terms of copulas, a necessary and sufficient condition for G to be a multivariate extreme value distribution is

$$D_G(u_1,\ldots,u_p) \quad = \quad D_G^k(u_1^{1/k},\ldots,u_p^{1/k}) \tag{4.3}$$

for any integer $k \ge 1$. (4.3) is equivalent to:

$$G^k(\mathbf{x}) \quad = \quad G(\mathbf{A}_k\mathbf{x} + \mathbf{B}_k) \tag{4.4}$$

for any $\mathbf{x} \in \mathbb{R}^p$, where $\mathbf{A}_k$ is a vector of positive constants, $\mathbf{B}_k$ is a vector of real constants, and all operation are interpreted pointwise.

Therefore, a large part of the literature on multivariate extreme value distributions has been concerned with finding joint distribution functions, or equivalently copulas, that satisfy (4.3) or (4.4).

4.2 The Pickands Representation

One of the first representations of multivariate extreme value distributions was due to Pickands [180]. As originally presented by Pickands, it was stated as a condition for a p-dimensional distribution with unit exponential margins ($\Pr\{X_j > x\} = e^{-x}$ for $x \ge 0,\ j = 1,\ldots,p$) to be the survivor function of a multivarate extreme value distribution for minima, and since this is an interesting case in its own right, we follow that usage here.

One point to make right away is that if $\Pr\{X_j > x\} = e^{-x}$, then $\Pr^n\{X_j > x\} = e^{-nx} = \Pr\{X_j > nx\}$, so the unit exponential distribution is min-stable with $b_n = 0,\ a_n = 1/n$.

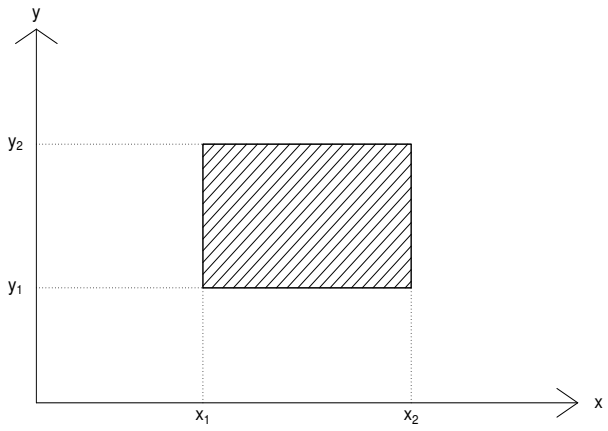

Figure 4.1 *Illustration of the non-negativity condition.*

With this preliminary, we seek a p-dimensional survivor function

$$S(x_1,\ldots,x_p) \quad = \quad \Pr\left\{X_1 > x_1,\ldots,X_p > x_p\right\}$$

such that

(i)

$$S(0,\ldots,0,x,0,\ldots,0) \quad = \quad e^{-x} \text{ for all } j \tag{4.5}$$

(in other words, the vector $\mathbf{x} = \begin{pmatrix} x_1 & \ldots & x_p \end{pmatrix}$ contains x in the jth position with all other entries 0),

(ii)

$$S^a(x_1,\ldots,x_p) \quad = \quad S(ax_1,\ldots,ax_p) \tag{4.6}$$

(log S is homogeneous of order 1),

(iii) *Non-negativity condition*: S is a valid survivor function in the sense of giving non-negative mass to any non-empty subset of $\mathbb{R}_+^p$. For example, in two dimensions this would require that for any rectangle $[x_1,x_2]\times[y_1,y_2]$ with $x_1 < x_2,\ y_1 < y_2$, $S(x_1,y_1) - S(x_2,y_1) - S(x_1,y_2) + S(x_2,y_2) > 0$. See Figure 4.1.

To get some idea how this might work, let's try an example. Consider the model

$$X_j \quad = \quad \min_{i=1,\ldots,q} \frac{Z_i}{c_{ij}},\ j = 1,\ldots,p \tag{4.7}$$

where $Z_1,\ldots,Z_q$ are independent identically distributed (IID) exponential random variables with mean 1 ($\Pr(Z_i > z) = e^{-z}$ for $0 \le z < \infty$) and $\{c_{ij},\ i = 1,\ldots,q\ j = 1,\ldots,p\}$ are constants satisfying

$$0 \le c_{ij} \le 1 \text{ and } \sum_i c_{ij} = 1 \text{ for all } j. \tag{4.8}$$

We calculate:

$$\begin{aligned}
\Pr\left\{X_j > x_j \text{ for all } j\right\} &= \Pr\left\{\frac{Z_i}{c_{ij}} > x_j \text{ for all } i,j\right\} \\
&= \Pr\left\{Z_i > \max_j c_{ij}x_j \text{ for all } i\right\} \\
&= \exp\left(-\sum_{i=1}^{q} \max_{1\le j\le p} c_{ij}x_j\right). \qquad (4.9)
\end{aligned}$$

From (4.9), we immediately deduce that

$$\Pr^n\left\{X_1 > x_1,\ldots,X_p > x_p\right\} = \Pr\left\{X_1 > nx_1,\ldots,X_p > nx_p\right\}$$

for any $x_1,\ldots,x_j \in (0,\infty)$, which proves that the joint distribution is min-stable with normalizing constants $a_{nj} = 1/n,\ b_{nj} = 0$.

We write (4.9) in the form

$$S(x_1,\ldots,x_p) = \exp\left\{-\int_{\mathscr{S}_p} \max_j(w_jx_j)dH(\mathbf{w})\right\} \tag{4.10}$$

where $\mathscr{S}_p$ is the p-dimensional simplex

$$\mathscr{S}_p = \left\{\mathbf{w} = \begin{pmatrix} w_1 & \ldots & w_p \end{pmatrix} :\ w_j \ge 0,\ \sum_{j=1}^{p} w_j = 1\right\},$$

and H is a (discrete) non-negative measure on $\mathscr{S}_p$.

To see that (4.9) can be rewritten in the form (4.10), let H be the measure on $\mathscr{S}_p$ that puts mass $\sum_j c_{ij}$ on each point $\frac{1}{\sum_j c_{ij}}\begin{pmatrix} c_{i1} & \ldots & c_{ip} \end{pmatrix}$, $i = 1,\ldots,q$. Then

$$\begin{aligned}
\int_{\mathscr{S}_p} \max_j(w_jx_j)dH(\mathbf{w}) &= \sum_i\left(\sum_j c_{ij}\right)\cdot\max_j\left(\frac{c_{ij}x_j}{\sum_j c_{ij}}\right) \\
&= \sum_i \max_j c_{ij}x_j.
\end{aligned}$$

We also have, for each $j = 1,\ldots,p$,

$$\int_{\mathscr{S}_p} w_j dH(\mathbf{w}) = \sum_i\left(\sum_j c_{ij}\right)\cdot\frac{c_{ij}}{\sum_j c_{ij}} = \sum_i c_{ij} = 1,$$

which establishes the desired equivalence.

Pickands' Theorem states, in effect, that *any* min-stable distribution satisfying the earlier conditions (i), (ii) and (iii) satisfied (4.10) for *some* non-negative measure H with

$$\int_{\mathscr{S}_p} w_j H(d\mathbf{w}) \quad = \quad 1,\ j = 1,\ldots,p. \tag{4.11}$$

This is the *Pickands Representation*, first presented by Pickands in an unpublished paper from 1976.

De Haan and Resnick [102] gave a mathematically equivalent representation using different notation and methods. Yet another version of the result was due to Deheuvels [54].

Another way of representing (4.10) is through the formula

$$S(x_1,\ldots,x_p) \quad = \quad \exp\left\{\sum_j x_j A\left(\frac{x_1}{\sum_j x_j},\ldots,\frac{x_p}{\sum_j x_j}\right)\right\} \tag{4.12}$$

where A is a function on $\mathscr{S}_p$ called the *dependence function*.

In the special case $p = 2$, this may be rewritten in the form

$$S(x,y) \quad = \quad \exp\left\{-(x+y)A\left(\frac{y}{x+y}\right)\right\} \tag{4.13}$$

where

$$A(w) \quad = \quad \int_0^1 \max\{u(1-w),(1-u)w\}dH(u)$$

where H is non-decreasing on $[0,1]$ satisfying $\int_0^1 udH(u) = \int_0^1 (1-u)dH(u) = 1$.

In the case that H is differentiable, say $dH(u) = h(u)du$, we can give an alternative statement of the relation between the functions A and h. Noting that $u = w$ is the crossover point when $u(1-w) = w(1-u)$, we rewrite the last equation as

$$A(w) \quad = \quad w\int_0^w (1-u)h(u)du + (1-w)\int_w^1 uh(u)du$$

Hence

$$\begin{aligned} A'(w) &= \int_0^w (1-u)h(u)du + w(1-w)h(w) - \int_w^1 uh(u)du - (1-w)wh(w) \\ &= \int_0^w (1-u)h(u)du - \int_w^1 uh(u)du, \\ A''(w) &= (1-w)h(w) + wh(w) \ = \ h(w). \end{aligned} \tag{4.14}$$

In general, $A(w)$ may be any convex function defined on $0 \le w \le 1$ lying within the triangle bounded by the points $(0,1),\ (1,1),\ (0.5,0.5)$. See Figure 4.2 for an explicit example and illustration of the general concept.

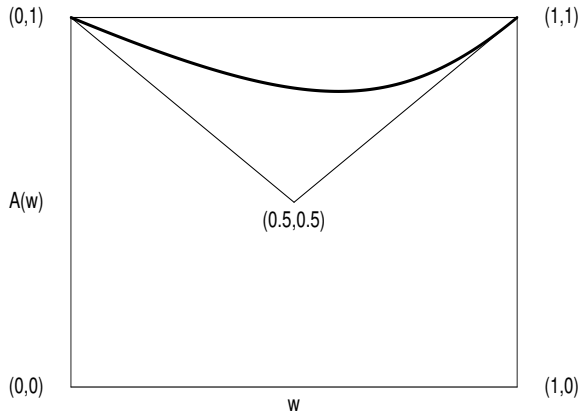

Figure 4.2 *A possible function $A(w)$, shown by the thick black line. This particular function is the asymmetric logistic function with parameters $\theta = 1,\ \phi = 0.5,\ r = 2$.*

4.3 Nonparametric Estimation of Bivariate and Multivariate Extreme Value Distributions

The previous section has shown that we can characterize multivariate extreme value distributions in terms of the spectral measure H or, equivalently, the Pickands dependence function A. Although these quantities are defined for any dimension p, they take a particularly simple form when $p = 2$, so we concentrate on that case, at least for our early discussion.

Unlike the one-dimensional case, there is no single finite-parameter family that covers all the multivariate or even bivariate extreme value distributions. The literature has therefore split into two approaches: either choose a paramateric subfamily and estimates its parameters, either by maximum likelihood or some other method — the issues surrounding maximum likelihood will be discussed in Section 4.4. Or we could do it nonparametrically, which is what we describe here.

Very often, a good approach is to use a nonparameteric method to determine the general shape of the dependence function, and then to select a parametric model for more detailed inference. The ideas in this section could be viewed as a way to tackle the first half of that program.

The first general-purpose nonparametric procedure was due to Pickands [179]. Although the idea of the method works for any dimension p, it is much simpler to explain in the bivariate case $p = 2$ so, initially, we restrict to that case.

Our starting point is the formula (4.13). Consider the random variable

$\min\left(\frac{X}{1-w}, \frac{Y}{w}\right)$ for some fixed $w \in (0,1)$. Then

$$\begin{aligned} \Pr\left\{\min\left(\frac{X}{1-w}, \frac{Y}{w}\right) > z\right\} &= \Pr\{X > z(1-w), Y > zw\} \\ &= \exp\{-zA(w)\}. \end{aligned}$$

In other words, $\min\left(\frac{X}{1-w}, \frac{Y}{w}\right)$ has an exponential distribution with mean $1/A(w)$, for any fixed w.

Now suppose we have a sample, $(X_i, Y_i),\ i = 1, \ldots, n$. This discussion suggests the estimator

$$A_n(w) = n\left\{\sum_{i=1}^{n} \min\left(\frac{X_i}{1-w}, \frac{Y_i}{w}\right)\right\}^{-1}. \tag{4.15}$$

Then $1/A_n(w)$ is an unbiased estimator of $1/A(w)$, and by the standard strong law of large numbers (SLLN) we will have $1/A_n(w) \stackrel{a.s.}{\to} 1/A(w)$ and hence also $A_n(w) \stackrel{a.s.}{\to} A(w)$. However, $A_n(w)$ is neither convex nor differentiable as a function of w.

Pickands' solution of this problem was to replace $A_n(w)$ by its greatest convex minorant (gcm). We note that for each i, $\min\left(\frac{X_i}{1-w}, \frac{Y_i}{w}\right)$ has discontinuous derivative at the "crossover point" $w = \frac{Y_i}{X_i+Y_i}$. Pickands recommended evaluating $A_n(w)$ at each crossover point, and then connecting by a straight line each of the points in the lower convex hull of that set. Figure 4.3 illustrates the idea.

One point we should note about Figure 4.3. The raw data $(X_i, Y_i),\ i = 1, \ldots, n$. simulated from a bivariate exponential distribution, need not have sample means 1, even though the populations means are 1. However, if the population means of X and Y were treated as unknown, a very natural approach would be to estimate the population means by the sample means $\bar{X}_n$ and $\bar{Y}_n$. In that case, it would be natural to replace X_i and Y_i by $X_i/\bar{X}_n$ and $Y_i/\bar{Y}_n$ respectively. This doesn't affect any of the asymptotic results (we obviously have $\bar{X}_n \stackrel{a.s.}{\to} 1$ and $\bar{Y}_n \stackrel{a.s.}{\to} 1$ as $n \to \infty$) but it does affect the practical properties of the estimator for small n — for example, this is necessary to guarantee that $A_n(0) = A_n(1) = 1$. This is what we have done in Figure 4.3, and would recommend for any practical application of the procedure.

The resulting procedure (without normalizing by $\bar{X}$ and $\bar{Y}$) is written $\tilde{A}_n(w)$ by Pickands. Pickands' main result was the almost sure consistency in this case:

$$\Pr\left\{\lim_{n\to\infty} \sum_{0 \le w \le 1} \left|\tilde{A}_n(w) - A(w)\right| \to 0\right\} = 1. \tag{4.16}$$

The result (4.16) essentially guarantees that $\tilde{A}_n$ will have good properties in large samples, though it does not say anything about the rate of convergence.

Pickands also discussed the extension to multivariate ($p > 2$) cases, remarking that this was "straightforward" though not providing much detail. Essentially, the idea is that the estimator (4.15) can be written down for the multivariate case, for

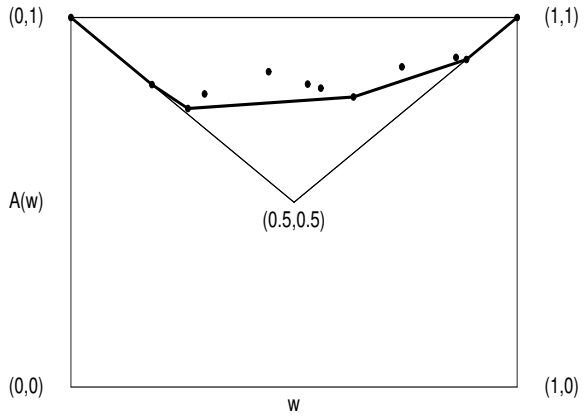

Figure 4.3 *Illustration of Pickands' estimator on a simulated dataset.*

example in the form

$$A_n(\mathbf{w}) \quad = \quad n\left\{\sum_{i=1}^{n}\min\left(\frac{X_{i1}}{w_1},\ldots,\frac{X_{ip}}{w_1}\right)\right\}^{-1}$$

where $\mathbf{w} = \begin{pmatrix} w_1 & \ldots & w_p \end{pmatrix}$ is a member of $\mathscr{S}_p$ with each $w_j \geq 0$ and $\sum_j w_j = 1$, and $\begin{pmatrix} X_{i1} & \ldots & X_{ip} \end{pmatrix}$ for $i = 1,\ldots,n$ is a sample of independent vectors from a p-dimensional multivariate extreme value distribution satisfying (4.12). In that case, $A_n(\mathbf{w})$ is a consistent estimator of $A(\mathbf{w})$ for each $\mathbf{w}$, but as in the two-dimensional case, does not have any convexity and differentability properties. Pickands recommended, in effect, evaluating $A_n(\mathbf{w})$ at each p-dimensional crossover point, and defining $\tilde{A}_n(\mathbf{w})$ to be the greatest convex minorant of that. In the three-dimensional case this amounts to a union of supporting hyperplanes.

Deheuvels [53] further developed the mathematical properties of Pickands' estimator in the two-dimensional case. As described by Deheuvels, these are "an application of general results on sums of random variables taking values in a Banach space". Specifically, he calculated the covariance of $1/A_n(w)$ and $1/A_n(v)$ for $w,v \in (0,1)$, and showed that the resulting process $\sqrt{n}\{1/A_n(w) - 1/A(w)\}$ converges in $C(0,1)$ to a limiting Gaussian process with the same covariance function. Here, $C(0,1)$ is the space of continuous functions on (0,1), and convergence refers to weak convergence in the sup-norm topology (if f and g are two continuous functions on $(0,1)$, then $||f-g|| = \sum_{0\leq u\leq 1}|f(u)-g(u)|$). The reason for expressing the result in terms of

$1/A_n(w)$ rather than $A_n(w)$ directly is presumably to simplify the mathematical form of the result, though equivalent results presumably hold for the convergence of $A_n(w)$ to $A(w)$ in $C(0,1)$. This sort of result is important for understanding the properties of estimators or other statistics that depend on the entire function $A_n(w)$ rather than just a single value or finite set of values of w. Deheuvels also proved almost sure convergence results of iterated logarithm form, and proposed an alternative method of normalizing when $\bar{X}$ and $\bar{Y}$ are not equal to 1.

Smith [223] and Yuen [262] presented alternative estimators based on Pickands' results but designed to handle the convexity question in a different way. Both methods are summarized more succinctly in [232]. Recalling (4.14), the idea is to numerically differentiate $A_n(w)$ to obtain a direct estimate of the density $h(w)$ for any fixed $w \in (0,1)$; by taking the parameter λ (defined below) sufficiently large, we can ensure that $h_n(w;\lambda) > 0$ and thus the convexity of our estimated of $A(w)$.

[223] started again from the estimator $A_n(w)$ in (4.15) and then defined

$$h_n(w;\lambda) \quad = \quad \frac{A_n(w+\lambda)+A_n(w-\lambda)-2A_n(w)}{\lambda^2} \tag{4.17}$$

defined on $\lambda < w < 1-\lambda$; a slight modification (which does not affect the asymptotic results) is needed when $w < \lambda$ or $w > 1-\lambda$. In effect, formula (4.17) is a crude but direct estimator of the second derivative of A. By direct manipulation, [222] developed the asymptotic results

$$\text{Bias of } h_n(w) \sim \frac{\lambda^2 h''(w)}{12}, \text{ Variance of } h_n(w) \sim \frac{C(w)}{n\lambda},$$

where

$$C(w) \quad = \quad \frac{12A^2(w)+12(1-2w)A(w)A'(w)-12w(1-w)A'(w)^2+4w(1-w)A(w)A''(w)}{3w^2(1-w)^2}$$

for $w \in (0,1)$ and assuming all the needed derivatives exist.

If we treat λ as a variable parameter (to be chosen by the user) then the structure of these results is that the mean squared error (squared bias plus variance) of $h_n(w;\lambda)$ is asymptotically of the form $B\lambda^4 + \frac{D}{n\lambda}$ for constants B and D. This is minimized by setting $\lambda = \left(\frac{D}{4Bn}\right)^{1/5}$ and leads to an asymptotic mean squared error of $5B^{1/5}D^{4/5}(4n)^{-4/5}$; note, in particular, the fact that $\lambda = O(n^{-1/5})$ for an asymptotic mean squared error of $O(n^{-4/5})$. In practice, we cannot just plug in these formulas because the constants B and D depend on derivatives of the unknown $A(w)$ function in a rather complicated way, but they give theoretical results for the structure of the optimal solution. These issues are familiar from the theory of kernel density estimation where λ is called the bandwidth and similar questions of optimal choice of bandwidth have long been known. The papers [223, 232] did not address the issues of practical implementation but with modern computational facilities it would be feasible to use cross-validation or some closely related technique.

Yuen [262, 232] made an extension even more closely related to kernel density

estimation by defining

$$A_{n1}(w;\lambda) \quad = \quad \frac{1}{\lambda}\int_0^1 A_n(u)K\left(\frac{w-u}{\lambda}\right)du \tag{4.18}$$

where we choose λ large enough to make (4.18) a convex function and K is a fixed kernel function; this can be any function $K(x) > 0$ for all $x \in \mathbb{R}$ with $\int_{-\infty}^{\infty} K(x)dx = 1$; for instance, the standard normal density could be a suitable K though in practice, density estimation experts often prefer the Epanechnikov kernel $K(x) = \frac{3}{4}(1-x^2)$ for $|x| < 1$ and 0 for $|x| > 1$; this has certain optimality properties.

With those definitions, Yuen showed that

$$\text{Bias of } h_n(w) \sim \frac{\lambda^2 h''(w)}{2}\int_{-\infty}^{\infty} x^2K(x)dx, \text{ Variance of } h_n(w) \sim \frac{3C(w)}{2n\lambda}\int_{-\infty}^{\infty} K^2(x)dx.$$

The mean squared error again has the structure $B\lambda^4 + \frac{D}{n\lambda}$ so the same asymptotic results apply as for the estimator of [223]; in fact it appears that the optimal asymptotic mean squared errors for the two estimators are very similar

A different nonparametric approach to the problem of bivariate extreme copula estimation was taken in Capéraà, Fougères and Genest [25]. Since their focus was on the copula, they assumed the marginal distributions to be uniform on $[0,1]$, so they wrote the sample values as (U_i, V_i), though this would be equivalent to our previous model if we wrote $X_i = -\log U_i,\ Y_i = -\log V_i$. In other respects, the model is the same as in the earlier papers [179, 232]: they assume the pairs $(U_i, V_i),\ i = 1,\ldots,n$ are independent identically distributed random vectors in $\mathbb{R}^2$ with joint distribution defined by the Pickands dependence function. In their notation, they defined the crossover points as $Z_i = \frac{\log U_i}{\log(U_i V_i)}$ and the order statistics $Z_{(1)},\ldots,Z_{(n)}$ and then

$$\begin{aligned} Q_i &= \left\{\prod_{i=1}^{n}\frac{Z_{(k)}}{1-Z_{(k)}}\right\}^{1/n},\ i = 1,\ldots,n, \\ A_n(t) &= \begin{cases} (1-t)Q_n^{1-p(t)}, & 0 \le t \le Z_{(1)}, \\ t^{i/n}(1-t)^{1-i/n}Q_n^{1-p(t)}Q_i^{-1}, & Z_{(i)} \le t \le Z_{(i+1)},\ i = 1,\ldots,n-1, \\ tQ_n^{-p(t)}, & Z_{(n)} \le t \le 1. \end{cases} \end{aligned}$$

Here, $p(t)$ is described as a weight function (fixed by the user), that they characterized as an arbitrary bounded function on $[0,1]$ with $p(0) = 1 - p(1) = 1$. They stated a formula for the weight function that minimizes the variance of the estimator, but the formula is complicated and depends on functions that are themselves unknown. Instead, they recommended $p(t) = 1-t$ as a simple practical choice, which they used for their simulation results.

The main result stated by [25] is the following:

Proposition. Suppose p is a bounded function on $[0,1]$. The estimator $A_n(t),\ 0 \le t \le 1$ is an asymptotically unbiased estimator of $A(t)$ which is uniformly strongly consistent.

The authors ran simulations to compare their estimator with the nonparametric estimators of Pickands [179] and Deheuvels [53], and also maximum likelihood estimators using three of the models analyzed by Tawn [237]. They stated that (unspecified) numerical problems prevented a direct comparison with the kernel method of [232]. In comparisons, they claimed that their method performed better than either Pickands' or Deheuvels' nonparametric estimators; their method performed less well than maximum likelihood when the assumed model was correct, as expected, but very often better than maximum likelihood when the assumed model was incorrect, which is not necessarily as expected since there are many instances in the theory of statistics when a maximum likelihood estimator continues to perform well when the model is slightly misspecified. Based on these results, they concluded that their nonparametric estimator could be used as a preliminary estimator to help select a parametric model. This conclusion should be compared with the similar conclusions reached in [232], which were based on a real-data analysis of sea level heights in eastern England rather than simulations.

Another nonparametric estimator that satisfies the required constraints on $A(w)$ was given by Tiago de Oliveira [244].

4.4 Parametric Estimation of Bivariate and Multivariate Extreme Value Distributions

The alternative approach (to the nonparametric estimators described in the previous section) is to find families of bivariate distributions that satisfy the Pickands conditions and then proceed to some parametric method of estimation, such as maximum likelihood, to estimate the parameters. However, early literature on this topic did not favor maximum likelihood, in part because of computational issues in the days before the use of computers in statistics became commonplace, but also because of some non-regularity issues in the estimation, specifically, that the Fisher information may not exist for certain values of the parameters. These issues were largely resolved by Tawn [237, 238]. In this section, we first give a partial list of parametric models that have been adopted, and then describe Tawn's results on maximum likelihood estimation. A review of older methods was given by Tiago de Oliveira [243]).

We focus on *differentiable models*, i.e. models for which a joint density exists, since if this condition is not satisfied, estimation by maximum likelihood makes no sense. This rules out, for example, the Marshall-Olkin distribution [151],

$$S(x,y) \quad = \quad \exp\{-\lambda_1 x - \lambda_2 y + \lambda_{12}\max(x,y)\}\,,\ x>0,\ y>0,\ \lambda_1>0,\ \lambda_2>0,\ \lambda_{12}>0,$$

(often called the "shock model" of reliability theory, since it corresponds to a shock that destroys both components simultaneously, i.e. $X=Y$ with positive probability). However, such a model is not generally realistic for applications in areas such as environment and economics.

Some of the common differentiable models for bivariate extremes are:

(a) *Mixed Model* [97]: $A(w)=\theta w^2-\theta w+1$. This satisfies the desired conditions if

$0 \le \theta \le 1$, and leads to the joint survivor function (with exponential marginals)

$$S(x,y) \quad = \quad \exp\left(-x-y+\frac{\theta xy}{x+y}\right),\ x>0,\ y>0.$$

(b) *Logistic Model* [94, 97]: $A(w) = \{(1-w)^r + w^r\}^{1/r}$, which is convex provided $r \ge 1$. The joint survivor function is

$$S(x,y) \quad = \quad \exp\left\{-(x^r+y^r)^{1/r}\right\},\ x>0,\ y>0.$$

Note that $r=1$ corresponds to X and Y independent ($A(w) \equiv 1$) while the limit $r \to \infty$ is the "complete dependence" case when $X = Y$ with probability 1, which also corresponds to the lower boundary of the triangle in Figure 4.2.

(c) *Asymmetric Mixed Model*. This is one of the models introduced for the first time by Tawn [237]. The dependence function is $A(w) = \phi w^3 + \theta w^2 - (\theta+\phi)w + 1,\ \theta \ge 0,\ \theta+\phi \le 1,\ \theta+2\phi \le 1,\ \theta+3\phi \ge 0$.

(d) *Asymmetric Logistic Model*. Also introduced by Tawn [237]. $A(w) = \{\theta^r(1-w)^r + \phi^r w^r\}^{1/r} + (\theta-\phi)w + 1 - \theta$ where $0 \le \theta \le 1,\ 0 \le \phi \le 1,\ r \ge 1$.

(e) *Inverted Logistic Model* (Joe [130]). $A(w) = 1 - \{\phi_1 w^{-\tau} + \phi_2(1-w)^{-\tau}\}^{-1/\tau}$, $0 \le \phi_1 \le 1,\ 0 \le \phi_2 \le 1, \tau \ge 0$. (Check convexity.)

(f) *Hüsler-Reiss Model* ([125]). $A(w) = (1-w)\Phi\left(\frac{a}{2} + \frac{1}{a}\log\frac{1-w}{w}\right) + w\Phi\left(\frac{a}{2} + \frac{1}{a}\log\frac{w}{1-w}\right)$ where $0 \le a \le \infty$ and $\Phi(\cdot)$ is the standard normal CDF. This was derived as a limiting distribution for bivariate normal extremes as the correlation coefficient tends to 1. The limits $a \to 0$ and $a \to \infty$ correspond respectively to the completely dependent and independent cases.

(g) *Bilogistic Model* ([131]). $A(w) = \int_0^1 \max\{(1-\alpha)(1-w)u^{-\alpha}, (1-\beta)w(1-u)^{-\beta}du\}$ for $\alpha, \beta \in [0,1]$. This was introduced as another asymmetric form of the logistic model that is possibly a little more intuitive than (d) — when $\alpha = \beta$, the model reduces to (b). The integral is in practice evaluated with a straightforward linear interpolation.

We also list some of the simpler multivariate models for general $p \ge 2$. For more general results, see Joe [130, 132]. In these cases, $\mathbf{w} = \begin{pmatrix} w_1 & \dots & w_p \end{pmatrix}$ is a general element of the simplex $\mathcal{S}_p$.

(h) *Multivariate Logistic Model*, due to Gumbel ([95]), $A(\mathbf{w}) = \left(\sum_{j=1}^p w_j^r\right)^{1/r}$, $r \ge 1$. This was the original and, for many years, the only widely recognized model for p-dimensional extreme values with $p > 2$, but its complete symmetry in the p variables is, for most practical purposes, a significant disadvantage, since it is unrealistic for most practical applications.

(i) *McFadden's Discrete Choice Models* ([157]),

$$A(\mathbf{w}) = \sum_{m=1}^{M} a_m \left(\sum_{i \in C_m} w_i^{r_m} \right)^{1/r_m}, \tag{4.19}$$

$$A(\mathbf{w}) = \sum_{m=1}^{M} a_m \left\{ \sum_{q \in D_m} \left(\sum_{i \in C_q} w_i^{t_q} \right)^{r_m/t_q} \right\}^{1/r_m}, \tag{4.20}$$

where in both cases $\cup_m C_m = \{1, \ldots, p\}$, $r_m \geq 1$, $a_m \geq 0$ and for (4.19)), $t_q \geq r_m$ for $q \in D_m$ and D_m is an arbitrary subset of $\{1, \ldots, p\}$.

(j) *Tawn's Extensions of McFadden's Models* ([238]). Motivated by McFadden's models and by an argument that combines a rainfall-storms interpretation of extreme events with multivariate survival models derived from stable laws [119, 43], Tawn proposed two more general models, the first

$$A(\mathbf{w}) = \sum_{C \in S} \left\{ \sum_{i \in C} (\theta_{i,C} w_i)^{r_C} \right\}^{1/r_C}, \tag{4.21}$$

a model with $2^{p-1}(p+2) - (2p+1)$ parameters, where each C is a non-empty subset of $\{1, \ldots, p\}$, S is the class of all non-empty subsets of $\{1, \ldots, p\}$, $r_C \geq 1$, $0 \leq \theta_{i,C} \leq 1$ and $\sum_{C \in \{1,\ldots,p\}} \sum_{i \in C} \theta_{i,C} = 1$, and the second

$$A(\mathbf{w}) = \sum_{C \in S} \left[\sum_{D \in C^*} \left\{ \sum_{i \in C \setminus D} (\phi_{i,D,C} w_i)^{r_C} + \left(\sum_{i \in D} (\phi_{i,D,C} w_i)^{r_C r_{D,C}} \right)^{1/r_{D,C}} \right\} \right]^{1/r_C} \tag{4.22}$$

where each $C \in S$, C^* is the class of nonempty subsets of C, $r_C \geq 1$, $r_{D,C} \geq 1$ and $\phi_{i,D,C} = \tau_{D,C} / \left\{ \sum_{C \in S_{(i)}} \left(\sum_{D \in C^*} \tau_{D,C}^{r_C} \right)^{1/r_C} \right\}$ where $\tau_{D,C} \geq 0$ and $S_{(i)}$ is the subclass of S which contains all nonempty subsets that include i. In particular, McFadden's model (4.19) is a special case of (4.21), and (4.20) is a special case of (4.22).

(k) *Tilted Dirichlet family*. This model is quite differently motivated, first proposed by Coles and Tawn [32]. The Dirichlet family whose density is given by

$$h^*(\mathbf{w}) = \frac{\Gamma(\sum_j \alpha_j)}{\prod_{j=1}^{p} \Gamma(\alpha_j)} \cdot \prod_{j=1}^{p} w_j^{\alpha_j - 1}, \tag{4.23}$$

with $\alpha_1 > 0, \ldots, \alpha_p > 0$, is possibly the best known of all probability distributions over a simplex; however, as it stands it is not a candidate to be the generating model of a MEVD because the quantity

$$m_j = \int_{\mathscr{S}_p} u_j h^*(\mathbf{w}) d\mathbf{w} = \frac{\alpha_j}{\sum_k \alpha_k} \neq 1,$$

violating a required moment condition for a MEVD. Instead, Coles and Tawn proposed

$$h(\mathbf{w}) \quad = \quad \prod_{j=1}^{p} \left\{ \frac{\alpha_j}{\Gamma(\alpha_j)} \right\} \cdot \frac{\Gamma(\sum_j \alpha_j + 1)}{(\sum_j \alpha_j w_j)^{p+1}} \cdot \prod_{j=1}^{p} \left(\frac{\alpha_j w_j}{\sum_k \alpha_k w_k} \right)^{\alpha_j - 1}. \quad (4.24)$$

They showed that $h(\mathbf{w})$ is indeed a probability density function over $\mathscr{S}_p$ and that $\int_{\mathscr{S}_p} w_j h(\mathbf{w}) d\mathbf{w} = 1$ and hence, writing $h(\mathbf{w})d\mathbf{w}$ instead of $dH(\mathbf{w})$ in (4.10), the resulting $S(x_1, \ldots, x_p)$ is indeed the survivor function of a min-stable distribution with unit exponential marginals.

4.4.1 Asymptotic results for maximum likelihood estimation and testing

The development of asymptotic results for these models was a major theme of the two papers of Tawn [237, 238]. For properties of maximum likelihood in general, we refer back to Chapter 1. In particular, maximum likelihood estimation is considered "regular" if the traditional asymptotic properties — consistency, asymptotic efficiency and asymptotic normality — are satisfied, and these conditions are typically satisfied if the Fisher information matrix is finite. [Note from the author: somewhere in this book there needs to be a self-contained introduction to maximum likelihood where the traditional regularity conditions are written out in detail. Possibly use the book by van der Vaart [247] as a source reference for this. I haven't yet written such a section, but it belongs in Chapter 1, not here.]

For bivariate and multivariate extreme value distributions, a very typical behavior is the following: in the interior of the parameter space, the Fisher information is satisfied and all the usual results for maximum likelihood estimation hold good. However, on the boundary, the Fisher information is very often infinite and some specialized results are needed. Defining these results was the main contribution of Tawn [237, 238].

As an example, consider the joint probability distribution function for the logistic dependence model with unit exponential marginal distributions,

$$G(x, y; r) \quad = \quad 1 - \exp\left\{ -(x^r + y^r)^{1/r} \right\}, \; x > 0, \; y > 0, \; r \geq 1. \quad (4.25)$$

The boundary point $r = 1$ corresponds to independence, so it is important to know how to test the null hypothesis $H_0 : \; r = 1$ against the alternative $H_1 : \; r > 1$, or to know how the maximum likelihood estimator behaves when the true value of r is 1.

The joint density is given by

$$g(x, y; r) \quad = \quad \frac{\partial^2 G}{\partial x \partial y} \; = \; (xy)^{r-1} (x^r + y^r)^{-2+1/r} \left\{ (x^r + y^r)^{1/r} + r - 1 \right\} \exp\left\{ -(x^r + y^r)^{1/r} \right\}.$$

Suppose we have observations (x_i, y_i), $i = 1, \ldots, n$. We define the log likelihood and

the score statistic,

$$\begin{aligned} \ell_n(r) &= \sum_{i=1}^{n} \log g(x_i, y_i; r), \\ U_n(r) &= \frac{d\ell_n}{dr}(r). \end{aligned}$$

For the case $r = 1$, we have $U_n(1) = \sum_{i=1}^{n} u(x_i, y_i)$ where

$$u(x,y) = \log(xy) + (x+y-2)\log(x+y) - x\log x - y\log y + (x+y)^{-1}.$$

Substituting $x = X,\ y = Y$ where X and Y are independent exponential random variables with unit mean, we find that

$$\mathrm{E}\{u(X,Y)\} = 0 \text{ but } \mathrm{Var}\{u(X,Y)\} = \infty.$$

For $r > 1$, the variance of $U_n(r)$ is finite and we can apply maximum likelihood estimation in the usual way, but the case $r = 1$ requires special treatment.

In this case, results from stable law theory (e.g [71]) show that the limiting distribution of $U_n(1)$ is still asymptotially normal but with a non-standard normalization, specifically

$$\sqrt{\frac{2}{n\log n}} U_n(1) \xrightarrow{d} \mathcal{N}[0,1]. \tag{4.26}$$

The result (4.26) can be used as a test statistic for the null hypothesis of independence against the alternative of a logistic dependence model with $r > 1$.

Tawn also considered the behavior of the maximum likelihood estimator itself. Suppose $\bar{r}_n$ denotes the MLE of r based on n observations, i.e. the value of r that maximizes $\ell_n(r)$ in $r \geq 1$. The notation is intended to distinguish from the case where the marginal parameters are unknown, considered later, where we write $\hat{r}_n$. There is a non-trivial probability that this maximum is attained at $r = 1$ and in that case we write $\bar{r}_n = 1$. Tawn showed that

$$\begin{aligned} \sqrt{\frac{n\log n}{2}}(\bar{r}_n - 1) &\xrightarrow{d} S \text{ where} \\ \Pr\{S \leq s\} &= \begin{cases} 0 & \text{if } s < 0, \\ \Phi(s) & \text{if } s \geq 0, \end{cases} \end{aligned} \tag{4.27}$$

where $\Phi(\cdot)$ is the standard normal CDF. In words, the limiting random variable S is 0 with probability $\frac{1}{2}$ (corresponding to the case $\bar{r}_n = 1$), but otherwise has a standard normal distribution on the positive half-line.

For the mixed model, Tawn showed that very similar results hold in the limiting case $\theta = 0$ that corresponds to independence. The asymmmetric mixed model and the asymmmetric logistc model are more complicated because of the multiple parameters, and full asymptotic results have never been obtained for these models, but it

is anticipated that they have similar behavior, with nonstandard asymptotic results at the boundary. From a modern perspective, a simulation or bootstrap-based procedure could be recommended as a practical way to perform estimation and testing in these models.

Tawn also considered the model in which the marginal parameters are unknown — as would typically be the case in practice, where we have to fit some model to both marginal distributions (for example, the generalized extreme value distribution) and then transform both marginal distributions to unit exponential distribution prior to assuming (4.25) or the equivalent for one of the other bivariate extreme models. Tawn's results in this case were influenced by earlier results [222] that showed that in certain nonregular cases — where one set of parameters has regular behavior with the usual asymptotic results, while another parameter has nonregular behavior — the regular and nonregular parameters often have independent limiting behavior. Tawn showed that a similar result holds for certain bivariate extreme models, as follows.

The model in this case is that there is a finite-dimensional set of marginal parameters, written $\boldsymbol{\phi} = \begin{pmatrix} \phi_1 & \dots & \phi_q \end{pmatrix}$, as well as a scalar dependence parameter θ (in the logistic model, we write θ in place of r). We also assume there is a critical value θ_0 where the estimation for θ becomes nonregular.

We assume that in the case where θ is known, the MLE for $\boldsymbol{\phi}$ exists, which we write as $\bar{\boldsymbol{\phi}}_n$ to indicate the dependence on sample size n. (If we needed also to distinguish the individual components of $\boldsymbol{\phi}$, we would write $\bar{\phi}_{n,j}$, for example, to indicate the jth component of $\bar{\boldsymbol{\phi}}_n$, but that will not be needed for the discussion here.) We assume that standard asymptotic results apply to $\bar{\boldsymbol{\phi}}_n$,

$$\sqrt{n}\,(\bar{\boldsymbol{\phi}}_n - \boldsymbol{\phi}) \quad \xrightarrow{d} \quad \mathcal{N}_q[\mathbf{0}, M^{-1}]$$

where M is the Fisher information matrix for $\boldsymbol{\phi}$, assumed to be strictly positive definite.

For the case θ is unknown, we assume the existence of a joint MLE $(\hat{\theta}_n, \hat{\boldsymbol{\phi}}_n)$, using different notation to indicate that this is now the joint MLE where both θ and $\boldsymbol{\phi}$ are unknown. As was the case previously for $\bar{\theta}_n$, we assume $\hat{\theta}_n = \theta_0$ if the maximum is attained on the boundary. There are now two cases:

(i) If $\theta > \theta_0$, the problem is regular, and we use the standard ($(q+1)$-dimensional) Fisher information matrix to deduce the joint asymptotic distribution of $(\hat{\theta}_n, \hat{\boldsymbol{\phi}}_n)$. In general it will not be the case that $\hat{\theta}_n$ and $\hat{\boldsymbol{\phi}}_n$ are asymptotically independent.

(ii) If $\theta = \theta_0$ (e.g. the independent case $\theta = 0$ for the mixed model, or $r = 1$ for the logistic model) then there is an asymptotic result of the form

$$\left(\sqrt{cn\log n}(\hat{\theta}_n - \theta_0), \sqrt{n}(\hat{\boldsymbol{\phi}}_n - \boldsymbol{\phi})\right) \quad \xrightarrow{d} \quad (S, Z_1, \dots, Z_q), \tag{4.28}$$

where $c > 0$, S has the same half-normal distribution as in (4.27), $\mathbf{Z} = \begin{pmatrix} Z_1 & \dots & Z_q \end{pmatrix}$ is $\mathcal{N}_q[\mathbf{0}, M^{-1}]$, and S and $\mathbf{Z}$ are *independent*.

Tawn also showed the following interesting side results:

$$\left(\bar{\theta}_n - \theta_0\right)\left\{-\frac{\partial^2 \ell_n(\bar{\theta}_n, \boldsymbol{\phi}_0)}{\partial \theta^2}\right\}^{1/2} \xrightarrow{d} S,$$

$$\left(\hat{\theta}_n - \theta_0\right)\left\{-\frac{\partial^2 \ell_n(\hat{\theta}_n, \hat{\boldsymbol{\phi}}_n)}{\partial \theta^2}\right\}^{1/2} \xrightarrow{d} S,$$

where $\boldsymbol{\phi}_0$ denotes the true value of $\boldsymbol{\phi}$, $\ell_n(\theta, \boldsymbol{\phi})$ now denotes the joint likelihood of θ and $\boldsymbol{\phi}$, and S is again the half-normal random variable in (4.27). The implication is that we still get the correct asymptotic normal distribution for either $\bar{\theta}_n$ or $\hat{\theta}_n$ if we normalize by the observed information of θ, which is what the terms $-\frac{\partial^2 \ell_n(\bar{\theta}_n, \boldsymbol{\phi}_0)}{\partial \theta^2}$ and $-\frac{\partial^2 \ell_n(\hat{\theta}_n, \hat{\boldsymbol{\phi}}_n)}{\partial \theta^2}$ represent, even though the Fisher information does not exist for this parameter. In regular estimation problems, a famous paper of Efron and Hinkley [66] argued that it is better to normalize using the observed information rather than the Fisher information, but the proof relied on complicated arguments involving ancillary statistics and conditional inference. In this case, using Fisher information is not an option but observed information still gives the right answers.

Further results on nonregular estimation in cases where $p > 2$ were stated in [238]. For example, for testing $r = 1$ in the multivariate logistic dependence model (h), or in one special case of the model (4.22), the asymptotic distribution of the score statistic can be shown to be of stable law (non-normal) form, but a complete enumeration of all the nonregular estimation results for these models has not been attempted.

A further comment on the joint estimation of θ and $\boldsymbol{\phi}$ is that, while the asymptotic results are the same whether the parameters are estimated together $(\hat{\theta}_n, \hat{\boldsymbol{\phi}}_n)$ or separately $(\bar{\theta}_n, \bar{\boldsymbol{\phi}}_n)$, in practice, there may be considerable advantages to the joint estimation approach; see in particular Shi [215] and an earlier preprint on the same theme [216].

4.5 Threshold Methods for Multivariate Extremes

All the analyses so far have assumed a block maximum approach, in other words, for p-dimensional we divide the data into blocks and compute the maximum or minimum of each of the p components in each year. The limiting distribution, as the block length n tends to infinity, is then the p-dimensional multivariate extreme value distribution defined by (4.1) or, in its equivalent form of a min-stable distribution with unit exponential marginal distributions, by (4.10) or (4.12). In environmental applications, the block length is usually fixed at one year and the analysis then proceeds on the assumption that n is large enough for the limiting distributions to be valid, though as in the univariate case, there is a legitimate question whether this is adequate or whether it would be better to choose a larger block size (e.g. two-year or five-year maxima; occasionally but more rarely, analysts use block lengths that are shorter than one year).

The alternative approach, well studied already in the case of univariate extremes,

is a *threshold approach* in which the analysis is essentially confined to exceedances over a high threshold. This naturally brings us into considerations of *multivariate regular variation*, of which the books by Resnick [194, 196] are the most comprehensive reference. In this section we do not attempt a full presentation of this theory, but we summarize the main results that are relevant for a threshold approach. The earliest attempt to do this was by Coles and Tawn [32], and for our initial discussion, we summarize their approach.

As noted in (4.2), there is no loss of generality in transforming the marginal distributions to uniform on $[0,1]$ and focusing attention on the *copula* which defines the joint distributions in this case. In practice, we can equally well transform to some other marginal distribution, for example, if U is uniform on $[0,1]$, then $-\log U$ has a unit exponential distribution or $-1/\log U$ has a unit Fréchet distribution with distribution function $e^{-1/x}$ for $0 < x < \infty$. In the previous sections we found it convenient to use unit exponential margins and the Pickands representation; for the present analysis, it is more convenient to assume unit Fréchet margins and a slightly different formulation of multivariate extreme value distributions due to de Haan and Resnick [102].

With these preliminaries, Coles and Tawn assumed a sequence $\mathbf{X}_i = \begin{pmatrix} X_{i1} & \dots & X_{ip} \end{pmatrix}$, $i = 1,2,\dots$ of IID random vectors in $\mathbb{R}^p_+$ with unit Fréchet margins: $\Pr\{X_{ij} \le x\} = e^{-1/x}$ on $0 < x < \infty$ for $j = 1,\dots,p,\ i = 1,2,\dots$ They then defined a point process

$$P_n \quad = \quad \left\{ n^{-1}\mathbf{X}_i : \ i = 1,\dots,n \right\}. \tag{4.29}$$

Then, by results in [102, 194], P_n converges in distribution to a limiting point process P on $\mathbb{R}_+ \backslash \{\mathbf{0}\}$. For a more detailed description of modes of convergence in point processes, we refer to [196].

To define the nonhomogeneous intensity measure of P, we first define new coordinates:

$$r_i \quad = \quad \frac{1}{n}\sum_{j=1}^{p} X_{ij},\ w_{ij} \ = \ \frac{X_{ij}}{nr_i},\ \mathbf{w}_i = \begin{pmatrix} w_{i1} & \dots & w_{ip} \end{pmatrix} \in \mathscr{S}_p. \tag{4.30}$$

With this transformation, the intensity measure of P on $\mathbb{R}^p \backslash \mathbf{0}$ is

$$\mu(dr \times d\mathbf{w}) \quad = \quad \frac{dr}{r} \cdot dH(\mathbf{w}) \tag{4.31}$$

where H is a measure on $\mathscr{S}_p$ satisfying

$$\int_{\mathscr{S}_p} w_j dH(\mathbf{w}) \quad = \quad 1 \text{ for all } j = 1,\dots,p. \tag{4.32}$$

The corresponding limit distribution for block maxima is as follows. If $M_{nj} = \max\{X_{1j},\dots,X_{nj}\}$ for each $j = 1,\dots,p$, then

$$\Pr\left\{ \frac{1}{n} M_{nj} \le x_j,\ j = 1,\dots,p \right\} \quad \to \quad G(\mathbf{x}) \ = \ e^{-V(\mathbf{x})}$$

where

$$V(\mathbf{x}) \quad = \quad \int_{\mathscr{S}_p} \max_{j=1,\ldots,p} \left(\frac{w_j}{x_j} \right) dH(\mathbf{w}) \tag{4.33}$$

where $V(\mathbf{x})$ is called the *exponent measure*. This is clearly related to Pickands' dependence measure (compare (4.10) and (4.12) with (4.33)) but they are not the same thing, and for present purposes, modeling in terms of the exponent measure is more convenient for the proposed approach. However, one comment we could make is that each of the models in Section 4.4 can be rewritten in terms of the exponent measure rather than Pickands' dependence function, and the paper [32] gives a number of examples of that.

One complication about this theory in the case $p > 2$ is that even when the exponent measure V is differentiable (which we assume), it is possible that some of the measure H is concentrated on lower-dimensional boundaries of the simplex $\mathscr{S}_p$. To make this precise, Coles and Tawn defined subsets c of the form $\{i_1, \ldots, i_j\}$ where $1 \leq j \leq p$ and each of $i_1, \ldots, i_j$ is one of $1, \ldots, p$. They defined a subset $\mathscr{S}_{j,c} = \{\mathbf{w} \in \mathscr{S}_p : \ w_k = 0 \text{ for each } k \notin c\}$. The measure H is assumed to have density $h_{j,c}$ on $\mathscr{S}_{j,c}$. As described by Coles and Tawn, "the density $h_{j,c}$ describes the dependence structure for events which are extreme only in the components $c = \{i_1, \ldots, i_j\}$". They then stated the main theorem linking V to these densities on sub-simplexes of $\mathscr{S}_p$:

Theorem. For each $c\{i_1, \ldots, i_m\}$,

$$\frac{\partial^m V}{\partial x_{i_1} \ldots \partial x_{i_m}} \quad = \quad -\left(\sum_{j=1}^{m} x_{i_j} \right)^{-(m+1)} h_{m,c} \left(\frac{x_{i_1}}{\sum x_{i_j}}, \ldots, \frac{x_{i_m}}{\sum x_{i_j}} \right) \tag{4.34}$$

defined on any $\mathbf{x} \in \mathbb{R}_+^p$ for which $x_r = 0$ for all $r \notin c$. Note that when translated back to Pickands' dependence measure, this formula is the extension of (4.14) to the fully multivariate case.

The second characterization result given by Coles and Tawn is effectively the transformation result linking (4.23) and (4.24). They show that if h^* is any positive function on $\mathscr{S}_p$ with $m_j = \int_{\mathscr{S}_p} u_j dh^*(\mathbf{u}) < \infty$, then the measure H with density

$$h(\mathbf{w}) \quad = \quad \left(\sum m_k w_k \right)^{-(p+1)} \prod_{j=1}^{p} m_j h^* \left(\frac{m_1 w_1}{\sum m_k w_k}, \ldots, \frac{m_p w_p}{\sum m_k w_k} \right) \tag{4.35}$$

is a valid intensity measure on $\mathscr{S}_p$ satisfying $\int_{\mathscr{S}_p} w_j dh(\mathbf{w}) = 1$, as in (4.33). The tilted Dirichlet model (4.24) is an obvious special case of this, in fact, the only special case that immediately comes to mind (though Coles and Tawn do remark that a generalization of this result may be used to generate models with mass on the boundaries of $\mathscr{S}_p$ as well as the interior).

4.5.1 Estimation in the Coles-Tawn model

The models and methods used by Coles and Tawn were essentially parametric: since any of the previous parametric models for multivariate extremes may be reparame-

terized in the form of the exponent measure V and hence the measure H, we may reduce them to parametric estimators based on the limiting point process. There are two cases:

(i) *Marginal distributions known.* In this case there is no loss of generality in assuming the marginal transformations are unit Fréchet (after a transformation, if needed). Assume the process is observed in an open set $A \in \mathbb{R}_+^p \setminus \{\mathbf{0}\}$. Also assume the exponent measure V (hence the measure H, and hence further the measure μ of (4.31)) are functions of a parameter vector $\boldsymbol{\theta}$. In that case, and using the limiting Poisson process as if it were the true model for the observations $\{n^{-1}\mathbf{X}_i,\ i = 1,\ldots,n\}$, the likelihood function for $\boldsymbol{\theta}$ is of the form

$$L_A(\boldsymbol{\theta}; \{n^{-1}\mathbf{X}_i,\ i = 1,\ldots,n\}) \quad = \quad \exp\{-\mu(A)\} \cdot \prod_{i=1}^{n_A} \mu\left(dr_i \times d\mathbf{w}_i\right). \tag{4.36}$$

Here, n_A denotes the number of observations in A and (possibly after reordering) r_i and $\mathbf{w}_i$ denote the radial and angular components of the ith observation in A. In practice, Coles and Tawn suggested taking $A = \mathbb{R}_+^p \setminus \left\{(0, v_1) \times \ldots \times (0, v_p)\right\}$ — in other words, a threshold v_j for the jth component of $\mathbf{X}$ where the set A consists of all observations $\mathbf{X}_i$ any one of whose coordinates is over the threshold for that coordinate.

(ii) *Marginal distributions unknown.* Suppose the original observations are denoted $\mathbf{Y}_i$ rather than $\mathbf{X}_i$, but after estimating the marginal distributions, each $\mathbf{Y}_i$ is transformed to $\mathbf{X}_i$ with unit Fréchet marginal distibutions. For the jth coordinate, we even write $X_j = X_j(Y_j)$ to make explicit that each X_j is determined by the corresponding Y_j. Since it is common to model high-level threshold exceedances of a univariate process using the generalized Pareto distribution (GPD) [51], they assumed

$$\Pr\left\{Y_j > y\right\} \quad = \quad p_j \left(1 + \xi_j \frac{y - u_j}{\sigma_j}\right)^{-1/\xi_j}$$

for $y \geq u_j$, where u_j is a threshold for the jth coordinate, $p_j = \Pr\{Y_j > u_j\}$, and σ_j and ξ_j are the scale and shape parameters for the GPD (in keeping with the notation that was widely used at the time, they used k_j in place of our $-\xi_j$). This immediately defines the probability integral transformation from Y_j to X_j when $Y_j > u_j$, but since the joint likelihood requires that only one of the p coordinates be above its corresponding threshold, we also need to define $X_j(Y_j)$ when $Y_j \leq u_j$. For this, Coles and Tawn used an empirical (rank-based) estimate of the marginal distribution. Hence, they defined

$$X_j(Y_j) \quad = \quad \begin{cases} \left[-\log\left\{1 - p_j\left(1 + \xi_j \frac{Y_j - u_j}{\sigma_j}\right)^{-1/\xi_j}\right\}\right]^{-1}, & Y_j > u_j, \\ \left\{-\log \frac{R(Y_j)}{n+1}\right\}^{-1}, & Y_j \leq u_j, \end{cases} \tag{4.37}$$

where $R(Y_j)$ denotes the rank of Y_j among the sample $Y_{ij},\ i = 1,\ldots,n$. They then

defined $v_j = n^{-1}X_j(u_j)$ and adapted the earlier formula (4.36) to define an expanded likelihood function,

$$L_A(\boldsymbol{\theta}, \sigma_1, \ldots, \sigma_p, \xi_1, \ldots, \xi_p; \{\mathbf{Y}_i,\ i = 1, \ldots, n\}) \quad = \quad \exp\{-V(\mathbf{v})\} \cdot \prod_{i=1}^{n_A} \left[h(\mathbf{w}_i)(nr_i)^{-(p+1)} \cdot \prod_{j=1,\ldots,p: X_{ij} > nv_j} \left\{ \sigma_j^{-1} p_j^{-\xi_j} X_{ij}^2 e^{1/X_{ij}} \left(1 - e^{-1/X_{ij}}\right)^{1+\xi_j} \right\} \right] \tag{4.38}$$

where $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})$ are defined by the transformations (4.37).

As in the earlier papers of Tawn, Coles and Tawn advocated maximum likelihood as the preferred method of estimation, though they noted that in certain cases (such as $r = 1$ in the logistic dependence model) the problem may become nonregular; although they did not describe a procedure to deal with this kind of difficulty, the presumed advice is to be cautious using this likelihood when one or more of the parameters appears to lie along a boundary of the parameter space.

As an application, Coles and Tawn analyzed a dataset of hourly sea surge levels at three sites on the east cost of England. They considered six dependence models and used the maximized log likelihood to decide which model fit best. They also argued that the real-data and simulation results point towards the superiority of estimating the marginal and dependence parameters simultaneously rather than separating the two estimation procedures.

4.5.2 Alternative censored data approach

An alternative approach was introduced in a series of papers [228, 145, 231] that aimed to construct a likelihood based directly on limiting approximations to the joint distribution function of high-level exceedances. This approach does not seem to have a widely recognized name but we call it here the *censored data approach* since it borrows idea familiar from the treatment of censored data in survival analysis.

The starting point is an alternative formula for the limiting joint distribution for multivariate extremes. We start with a characterization of the domain of attraction of a multivariate extreme value distribution given in Proposition 5.15 of Resnick [194]. Suppose we have a p-dimensional random variable $\mathbf{Y} = \begin{pmatrix} Y_1 & \ldots & Y_p \end{pmatrix}$ with distribution function $F(y_1, \ldots, y_p)$ and let F_j denote the marginal distribution function of Y_j. Define $Z_j = -1/\log F_j(Y_j)$ for each j and let $F_*(z_1, \ldots, z_p) = F(F_1^{\leftarrow}(e^{-1/z_1}), \ldots, F_p^{\leftarrow}(e^{-1/z_p}))$ where $F_j^{\leftarrow}$ is the inverse function of F_j. This corresponds to each of the marginal distributions of F being transformed to unit Fréchet form, i.e. $\Pr\{Z_j \leq z\} = e^{-1/z}$. The condition for F_* to be in the domain of attraction of a multivariate distribution function G_* with unit Fréchet marginal distributions is then given as

$$\lim_{t\to\infty} \frac{\log F_*(tz_1, \ldots, tz_p)}{\log F_*(t, \ldots, t)} \quad = \quad \lim_{t\to\infty} \frac{1 - F_*(tz_1, \ldots, tz_p)}{1 - F_*(t, \ldots, t)} \quad = \quad \frac{\log G_*(z_1, \ldots, z_p)}{\log G_*(1, \ldots, 1)}. \tag{4.39}$$

(The second equality in (4.39) comes from Resnick; the first follows from the familiar result that $\frac{-\log F}{1-F} \to 1$ as $F \to 1$.) However, we have already seen in (4.33) that $G_*(z_1,\ldots,z_p) = \exp\{-V(z_1,\ldots,z_p)\}$. This suggests an approach, directly generalizing the use of the generalized Pareto distribution (GPD) in univariate extreme value analysis [51], that treats either of the limiting relations in (4.39) as an identity for sufficiently large t.

Specifically, we assume that either $1-F_j(x)$ or $-\log F_j(x)$ is of the form $\lambda_j\{1+\xi_j(x-u_j)/\sigma_j\}_+^{-1/\xi_j}$ for $x \geq u_j$, where u_j is the arbitrarily chosen threshold for variable j. When combined with (4.39), this leads to one of the approximations

$$F(x_1,\ldots,x_p) \;=\; 1 - V\left\{\lambda_1^{-1}\left(1+\xi_1\frac{x_1-u_1}{\sigma_1}\right)_+^{1/\xi_1},\ldots,\lambda_p^{-1}\left(1+\xi_p\frac{x_p-u_p}{\sigma_p}\right)_+^{1/\xi_p}\right\}, \tag{4.40}$$

or

$$F(x_1,\ldots,x_p) \;=\; \exp\left[-V\left\{\left(-\log\left(1-\lambda_1\left(1+\xi_1\frac{x_1-u_1}{\sigma_1}\right)_+^{-1/\xi_1}\right)\right)^{-1},\ldots,\right.\right.$$
$$\left.\left.\left(-\log\left(1-\lambda_p\left(1+\xi_p\frac{x_p-u_p}{\sigma_p}\right)_+^{-1/\xi_p}\right)\right)^{-1}\right\}\right], \tag{4.41}$$

where in either case the formula is assumed valid whenever $x_j \geq u_j$ for $j = 1,\ldots,p$. Both formulas (4.40) and (4.41) were initially stated in [228] but the motivation behind (4.41), as an alternative to (4.40), was due to Ledford and Tawn and further developed in [145]; in particular, they argued that (4.41) works better than (4.40) in cases where the variables are nearly independent.

The reason for calling this a "censored data" approach is that the likelihood based on either (4.40) or (4.41) has to take into account the fact that we are assuming no model for the data when $x_j < u_j$; thus, we have to treat any component below the threshold as effectively censored at the threshold. Thus, following [145], the likelihood contribution due to a typical observation $\begin{pmatrix} y_1 & \ldots & y_p \end{pmatrix}$ in which the components $j_1,\ldots,j_m$ exceed their respective thresholds is given by

$$\left.\frac{\partial^m F(x_1,\ldots,x_p)}{\partial x_{j_1}\ldots\partial x_{j_m}}\right|_{x_j=\max(u_j,y_j),\ j=1,\ldots,p}. \tag{4.42}$$

As an example of how this works out in practice, and to make clear the connection with censored data, consider what happens when $p = 2$ and we have n independent pairs of observations $\{(y_{i1},y_{i2}),\ i = 1,\ldots,n\}$. The contribution to the likelihood from (y_{i1},y_{i2}) is

$$L_i \;=\; \begin{cases} F(u_1,u_2) & \text{if } y_{i1} \leq u_1, y_{i2} \leq u_2, \\ \frac{\partial F}{\partial x_2}(u_1,y_{i2}) & \text{if } y_{i1} \leq u_1, y_{i2} > u_2, \\ \frac{\partial F}{\partial x_1}(y_{i1},u_2) & \text{if } y_{i1} > u_1, y_{i2} \leq u_2, \\ \frac{\partial^2 F}{\partial x_1 \partial x_2}(y_{i1},y_{i2}) & \text{if } y_{i1} > u_1, y_{i2} > u_2. \end{cases} \tag{4.43}$$

The full likelihood L is then $\prod_{i=1}^{n} L_i$. The first three terms in (4.43) are in effect censored data terms because we don't make any assumption for the distribution of (y_{i1}, y_{i2}) below the threshold: in effect, we treat the observations as censored. This avoids one slightly awkward feature of the Coles-Tawn method, the need to introduce an empirical distribution function in (4.37).

On the face of it, this approach gives a clear-cut method of threshold-based statistical inference in multivariate extreme value distributions. First, we adopt a parametric model for $V(z_1, \ldots, z_p)$, of which we saw numerous examples in Section 4.4. Second, we assume either of the representations (4.40) or (4.41) which incorporate the marginal distributions as represented by the GPD parameters $\lambda_j, \sigma_j, \xi_j$. Then, the individual likelihood terms are defined by (4.42), which reduced to (4.43) in the bivariate case $p = 2$.

However, it turns out that there are a number of issues with this approach. One of them is a technical issue that we have discussed previously, the nonregularity of the MLE at the boundary points of the model, such as $\theta = 0$ in the mixed model or $r = 1$ in the logistic model (examples (a) and (b) of our catalog of models in Section 4.4). Imitating earlier arguments from [237, 238], Ledford and Tawn [145] showed that for these two models, the asymptotic distribution of the score statistic is normal but with a nonstandard normalization, and they imply that simliar but more complicated results must hold for other models such as the asymmmetric moxed or logistic models.

There are also issues with the combinatorial explosion of the number of terms in the likelihood function as p grows. Equation (4.43) splits the likelihood into 4 possible terms depending on which components are above or below the threshold. Evidently, the corresponding split for a p-dimensional multivariate extreme value distribution will contain 2^p terms, which grows rapidly as p grows.

However, the computational issue does not end there. Consider the formula $F_*(z_1, \ldots, z_p) = \exp\{-V(z_1, \ldots, z_p)\}$. Recall that all of our parametric formulas for multivariate extreme value models have relied on the specification of V rather than directly F_*. Therefore, the derivatives of F_* (and hence F itself) must be expressed in terms of V. We may calculate

$$\begin{aligned}
\frac{\partial F_*}{\partial z_i} &= -e^{-V} \frac{\partial V}{\partial z_i}, \\
\frac{\partial^2 F_*}{\partial z_i \partial z_j} &= e^{-V} \left(\frac{\partial V}{\partial z_i} \frac{\partial V}{\partial z_j} - \frac{\partial^2 V}{\partial z_i \partial z_j} \right), \\
\frac{\partial^3 F_*}{\partial z_i \partial z_j \partial z_k} &= e^{-V} \left(-\frac{\partial V}{\partial z_i} \frac{\partial V}{\partial z_j} \frac{\partial V}{\partial z_k} + \frac{\partial^2 V}{\partial z_i z_j} \frac{\partial V}{\partial z_k} + \frac{\partial^2 V}{\partial z_i z_k} \frac{\partial V}{\partial z_j} + \frac{\partial^2 V}{\partial z_j z_k} \frac{\partial V}{\partial z_i} - \frac{\partial^3 V}{\partial z_i \partial z_j \partial z_k} \right)
\end{aligned}$$

and so on. The number of terms in this expansion is the pth Bell number, of which the next few (for $p = 4, \ldots, 8$) are 15, 52, 203, 877 and 4140. Clearly, this also quickly gets out of control. The combinatorial problem may be simplified by using computer algebra, but this would not reduce the complexity of computing an exact likelihood for large p.

However, Ledford and Tawn were also the first authors to draw attention to a

more fundamental issue with all the models for multivariate extremes considered so far, that they only deal with cases of *asymptotic dependence* and neglect the equally important case of *asymptotic independence*. We turn to this question now.

4.6 Asymptotic Dependence and Asymptotic Independence

4.6.1 Introduction: The coefficient of tail dependence

Let us go back to (4.13), which we rewrite for present purposes (with unit Fréchet margins) in the form

$$\Pr\{X \le x, Y \le y\} \quad = \quad \exp\left\{-\left(x^{-1}+y^{-1}\right) A\left(\frac{x}{x+y}\right)\right\}. \tag{4.44}$$

In particular, if we set $x = y$, we get

$$\Pr\{X \le x, Y \le x\} \quad = \quad \exp\left\{-\frac{2A(1/2)}{x}\right\}.$$

Combining this with the unit Fréchet marginal distribution for both X and Y, as $x \to \infty$ we get

$$\begin{aligned}
\Pr\{X \le x \text{ or } Y \le x\} &= \Pr\{X \le x\} + \Pr\{Y \le x\} - \Pr\{X \le x, Y \le x\} \\
&= 1 - \frac{1}{x} + 1 - \frac{1}{x} - 1 + \frac{2A(1/2}{x} + O\left(\frac{1}{x^2}\right) \\
&= 1 - \frac{2(1-A(1/2))}{x} + O\left(\frac{1}{x^2}\right).
\end{aligned}$$

Hence

$$\Pr\{X > x, Y > x\} \quad \sim \quad \begin{cases} \frac{2(1-A(1/2))}{x} & \text{if } A(1/2) < 1, \\ O\left(\frac{1}{x^2}\right) & \text{if } A(1/2) = 1. \end{cases} \tag{4.45}$$

One immediate consequence of (4.45) is the formula

$$\chi \quad = \quad \lim_{x \to \infty} \Pr\{Y > x \mid X > x\} \; = \; 2\left\{1 - A\left(\frac{1}{2}\right)\right\} \tag{4.46}$$

which is > 0 if $A(1/2) < 1$ but $\chi = 0$ if $A(1/2) = 1$.

If $\chi = 0$ then we say X and Y are *asymptotically independent*; otherwise, they are *asymptotically dependent*. For the case where X and Y have a bivariate extreme value distribution, (4.46) shows that they are asymptotically dependent except for the case $A(1/2) = 1$, but in that case, the convexity of A, combined with $A(0) = A(1) = 1$, implies that $A(w) = 1$ for every $w \in [0,1]$ and X and Y are exactly independent.

The difficulty with this concept is that there are many important, relevant examples of bivariate distributions that are asymptotically independent without being

exactly independent, including the best known of all bivariate distributions, the bivariate normal distribution. Ledford and Tawn cited the result that when X and Y are bivariate normal with means 0, variances 1 and correlation coefficient $\rho < 1$,

$$\Pr\{X > x, Y > x\} \quad = \quad (1+\rho)^{3/2}(1-\rho)^{-1/2}(4\pi)^{-\rho/(1+\rho)}x^{-2/(1+\rho)}(\log x)^{-\rho/(1+\rho)}. \tag{4.47}$$

Thus $\chi = 0$ in this case, but the asymptotic result (4.47) shows that there is a non-trivial tail dependence between X and Y in practice. To emphasize the point, they proved that if $(X_i, Y_i),\ i = 1, \ldots, n$ were independent bivariate normal vectors with $\rho < 1$, the score statistic test discussed in Section 4.5.2 would reject independence with probability 1 as $n \to \infty$, even though X and Y are asymptotically independent according to (4.47) and it has long been known that the joint distribution of the extremes of a bivariate normal distribution, suitably normalized, converge to a pair of independent Gumbel distributions [217]. Ledford and Tawn gave several other examples of bivariate families where similar behavior occurs.

Clearly, some new concept is needed, and returning to the case of unit Fréchet marginal distribution, Ledford and Tawn proposed the model

$$\Pr\{X > x, Y > x\} \quad \sim \quad \mathscr{L}(x)x^{-1/\eta} \tag{4.48}$$

where $\mathscr{L}$ is a slowly varying function and η is a new parameter satisfying $0 < \eta \leq 1$ which they called the *coefficient of tail dependence*. Recall that a function is slowly varying if $\mathscr{L}(tx)/\mathscr{L}(t) \to 1$ as $t \to \infty$ for all $x > 0$. Thus the definition (4.48) includes cases where $\mathscr{L}(x)$ tends to a constant as well as examples (including the bivariate normal) where it is proportional to a power of $\log x$ or some similar slowly varying behavior. Indeed, in the case that X and Y are derived through marginal transformation of a joint bivariate normal distribution with correlation coefficient $\rho \in (-1, 1)$, we have $\eta = \frac{1+\rho}{2}$, by (4.47). In their original paper, Ledford and Tawn [145] restricted η to the range $[\frac{1}{2}, 1]$, but this is unnecessary since [146] and all subsequent papers assumed $\eta \in (0, 1]$ as given here.

Ledford and Tawn defined a number of special cases of (4.48). The case where $\eta = 1$ and $\mathscr{L}(u) \nrightarrow 0$ corresponds to the traditional definition of bivariate extreme value distributions that we have been using up to this point. The case $\frac{1}{2} < \eta < 1$ is called *positive association* which applies, in particular, for a bivariate normal distribution with $0 < \rho < 1$. The third case of interest is the *near independence* case when $\eta = \frac{1}{2}$ and $\mathscr{L}(x) \geq 1$. The fourth case when $0 < \eta < \frac{1}{2}$ corresponds to negative association between X and Y.

To estimate η, Ledford and Tawn defined a random variable $T = \min(X, Y)$ and noted that (4.48) immediately implies $\Pr\{T > x\} \sim \mathscr{L}(x)x^{-1/\eta}$. Based on that, they showed that

$$\Pr\{T > u + t \mid T > u\} \quad \sim \quad \left(1 + \frac{t}{u}\right)^{-1/\eta}$$

which corresponds to a GPD pver threshold u where $\xi = \eta$. Therefore, they proposed fitting a GPD or, equivalently, adopting the point process approach of [225], and

estimating η by $\hat{\xi}$, the MLE of ξ in the GPD or point process approach. This leaves open the question of how to choose the threshold but this is similar to any problem that involves exceedances over a threshold.

The definition of χ in (4.46) was introduced by Coles, Heffernan and Tawn [35] but does not have a fixed name. Serinaldi [214] defined the same concept with the notation λ_U which he called the *upper tail dependence coefficient*, but consistent with the theory that was started by Ledford and Tawn and developed by numerous subsequent authors, we reserve the term *coefficient of tail dependence* for the parameter η in (4.48.

4.6.2 Extension to the full joint tail

The follow-up paper by Ledford and Tawn [146] considered the extension of the preceding approach to the full joint tail where we want to estimate $\Pr\{X > x, Y > y\}$ for all large values of x and y (i.e. not restricting $x = y$). The model (4.48) does not lead immediately to a specific form of joint tail but Ledford and Tawn proposed

$$\Pr\{X > x, Y > y\} \quad = \quad \mathscr{L}(x,y)x^{-c_1}y^{-c_2} \tag{4.49}$$

where $c_1 + c_2 = 1/\eta$ and $\mathscr{L}$ is a bivariate slowly varying function satisfying the property that

$$g(x,y) \quad = \quad \lim_{t\to\infty}\left\{\frac{\mathscr{L}(tx,ty)}{\mathscr{L}(t,t)}\right\}$$

exists for all $x, y > 0$. Here $g(cx,cy) = g(x,y)$ for all $c > 0$ which implies that $g(x,y) = g_*\left(\frac{x}{x+y}\right)$ for some function g_*. In fact, they proposed replacing $\mathscr{L}(x,y)$ in (4.49) by $Kg_*(w)$ where $K > 0$ is a constant and $w = \frac{x}{x+y}$.

The difficulty with this approach is the absence of a well-defined procedure for defining the function g_*. In principle, a nonparametric technique might be preferred, but Ledford and Tawn noted the difficulty in defining such an estimator and proposed instead the model $\mathscr{L}(x,y) = \mathscr{L}_*(x/(x+y))$ where

$$\mathscr{L}_*(w) \quad = \quad a_0 + a_1\{w(1-w)\}^{-1/2}[1 - V\{(1-w)^{-1}w^{-1}\}] \tag{4.50}$$

where V defined by (4.33) could be any of the parametric models in Section 4.4. As a specific example, they took

$$V(x,y) \quad = \quad (x^{-1/\alpha} + y^{-1/\alpha})^{\alpha} \tag{4.51}$$

with $\alpha \in (0,1]$, equivalent to the logistic dependence model of Section 4.4 with $r = 1/\alpha$.

If we assume that equations (4.49)–(4.51) hold exactly for x and y sufficiently large, combined with the transformation of the marginal distributions to unit Fréchet, this defines a model for high-threshold exceedances of a bivariate pair which may be estimated by the technique of Section 4.5.2.

In the latter parts of their paper, Ledford and Tawn discussed a number of diagnostic procedures, including the choice of thresholds, a nonparametric diagnostic for estimating $c_1 - c_2$ (since $c_1 + c_2 = 1/\eta$ and we already have an estimator of η, this would allow us to estimate both c_1 and c_2 without relying on a specific model for g_*) and the possibility of higher-order expansions to refine the fit of the model. They also consider a number of possible submodels (for example, forcing $a_0 = 1$ or constraining η to be either $\frac{1}{2}$ or 1) noting, as expected from previous results, that test statistics for choosing among these models generally have non-standard asymptotic properties. In they end, they conclude that "η largely governs the extrapolation properties of the joint tail and is therefore the parameter of greatest importance for statistical applications."

4.6.3 New models based on hidden regular variation

The problem up to this point is that while (4.49) defines a general class of models for the joint tail consistent with (4.48), there is no clear-cut procedure for defining the function $\mathscr{L}(x,y)$, and the specification (4.50) seems arbitrary. Ramos and Ledford [188] proposed a new approach to this by building on the theory of *hidden regular variation*, developed by Resnick [195] and extended by Maulik and Resnick [156].

The starting point for the Ramos-Ledford theory is the equation

$$\bar{F}_{XY}(x,y) \quad = \quad \Pr\{X > x, Y > y\} \; = \; \frac{\mathscr{L}(x,y)}{(xy)^{1/(2\eta)}} \tag{4.52}$$

where $\eta \in (0,1]$ and $\mathscr{L}$ is bivariate slowly varying in the sense that there is a limit function g such that

$$g(x,y) \quad = \quad \lim_{u\to\infty} \left\{ \frac{\mathscr{L}(ux,uy)}{\mathscr{L}(u,u)} \right\}. \tag{4.53}$$

If (4.53) holds then we must have $g(cx,cy) = g(x,y)$ for any $c > 0$ and hence $g(x,y) = g_*(x/(x+y))$ for some g_*. So far, the formulation is the same as (4.49) except that the authors assume without comment that $c_1 = c_2 = 1/(2\eta)$.

Ramos and Ledford defined random variables $(S,T) \in [1,\infty)^2$ to be the weak limits of $(X/u, Y/u)$ given $X > u, Y > u$. Thus

$$\bar{F}_{ST}(s,t) \quad = \quad \Pr\{S > s, T > t\} \; = \; \lim_{u\to\infty} \left\{ \frac{\Pr(X > us, Y > ut)}{\Pr(X > u, Y > u)} \right\} \; = \; \frac{g_*(s/(s+t))}{(st)^{1/(2\eta)}}. \tag{4.54}$$

Define the change of variables $R = S + T, W = S/R$. Then, following [195, 156], the joint distribution of (R,W) factorizes as $\mu_{RW}(dr,dw) = r^{-(1+1/\eta)} dH_\eta(w)$ where H_η is some non-negative measure on $[0,1]$. For given w define $r^* = \max(s/w, t/(1-w))$,

then

$$\begin{aligned}\frac{g_*(s/(s+t))}{(st)^{1/(2\eta)}} &= \bar{F}(s,t) = \int_0^1 \int_{r^*}^{\infty} r^{-(1+1/\eta)} dr dH_\eta(w) \\ &= \eta \int_0^1 \min\left(\frac{w}{s}, \frac{1-w}{t}\right)^{1/\eta} dH_\eta(w) \\ &= \eta \int_0^{s/(s+t)} \left(\frac{w}{s}\right)^{1/\eta} dH_\eta(w) + \eta \int_{s/(s+t)}^1 \left(\frac{1-w}{t}\right)^{1/\eta} dH_\eta(w).\end{aligned} \tag{4.55}$$

Equation (4.55) is reminiscent of similar formulas for classical bivariate extreme value distributions (the $p = 2$ case of (4.33)) but is clearly a different formula with different consequences. Just as (4.33) was subject to the constraint (4.32), so there is a constraint (similarly motivated, but quite different in form) associated with (4.55). Specifically, (4.53) implies that $g(1,1) = g_*(1/2) = 1$ so, substituting $s = t$ in (4.55), we find that

$$\eta^{-1} = \int_0^{1/2} w^{1/\eta} dH_\eta(w) + \int_{1/2}^1 (1-w)^{1/\eta} dH_\eta(w). \tag{4.56}$$

Setting $s = rw,\ t = r(1-w)$ (where r is arbitrary) in (4.55) leads to the formula

$$g_*(w) = \eta \left(\frac{1-w}{w}\right)^{1/(2\eta)} \int_0^w z^{1/\eta} dH_\eta(z) + \eta \left(\frac{w}{1-w}\right)^{1/(2\eta)} \int_w^1 (1-z)^{1/\eta} dH_\eta(z) \tag{4.57}$$

where H_η is a non-negative measure on $[0,1]$, arbitrary except for the constraint (4.56).

As an example, Ramos and Ledford considered the model

$$h_\eta(w) = \frac{dH_\eta(w)}{dw} = \frac{\eta - \alpha}{\alpha \eta^2 N_\rho} \left\{ (\rho w)^{-1/\alpha} + \left(\frac{1-w}{\rho}\right)^{-1/\alpha} \right\}^{\alpha/\eta - 2} \{w(1-w)\}^{-(1+1/\alpha)} \tag{4.58}$$

where $N_\rho = \rho^{-1/\eta} + \rho^{1/\eta} - (\rho^{-1/\alpha} + \rho^{1/\alpha})^{\alpha/\eta}$ with $\eta \in (0,1],\ 0 < \alpha \leq 1,\ \rho > 0$. They described this model as a modified version of Tawn's [237] asymmetric logistic model. Based on (4.58), they defined the joint survivor model for (X,Y) as

$$\bar{F}_{XY}(x,y) = \frac{\lambda u^{1/\eta}}{N_\rho} \left[(\rho x)^{-1/\eta} + \left(\frac{y}{\rho}\right)^{-1/\eta} - \left\{ (\rho x)^{-1/\alpha} + \left(\frac{y}{\rho}\right)^{-1/\alpha} \right\}^{\alpha/\eta} \right] \tag{4.59}$$

valid in $x \geq u,\ y \geq u$, where u is some chosen high threshold and $\lambda = \Pr\{X > u, Y > u\}$.

The model (4.59) represents an interesting extension to the classical bivariate extreme models that include both asymptotically dependent and asymptotically independent cases. If we substitute $y = x$ we see immediately that $\bar{F}_{XY}(x,x) = O\left(x^{-1/\eta}\right)$. However, the marginal distribution of either X or Y depends on the relative sizes of α and η: if $\eta \geq \alpha$ then $\Pr\{X > x\} = O\left(x^{-1/\eta}\right)$ but if $\eta < \alpha$ then $\Pr\{X > x\} = O\left(x^{-1/\alpha}\right)$.

Thus if we define $\chi = \lim_{x\to\infty} \Pr\{Y > x | X > x\}$ as in (4.46) and define asymptotic dependence as $\chi > 0$, we find that the model (4.59) is asymptotically dependent if $\eta \geq \alpha$ but asymptotically independent if $\eta < \alpha$ (in particular, asymptotic dependence does not apply only to the case $\eta = 1$).

4.6.4 Extensions to the case $p > 2$

References: [187, 189]

4.6.5 An application: Dependence among extreme weather events

The analysis in this section is based on a contribution to a National Research Council report [40].

Many studies in recent years have documented the increased frequency and severity of extreme weather events, which is commonly believed to be a direct consequence of human-induced climate change. However, less attention has been given to the possibility of seemingly unrelated extreme weather events possibly having a similar climate cause. This could be of particular concern if extreme events in different parts of the world occur in rapid succession, because of the limited human and financial resources available to recover from such events. Bivariate extreme value theory provides a possible tool for analyzing the dependence of extreme weather events.

Example 1. Herweijer and Seager [114] argued that the persistence of drought patterns in various parts of the world may be explained in terms of sea surface temperature patterns. One of their examples (Figure 3 of their paper) demonstrated that precipitation patterns in the south-west United States are highly correlated with those of a region of South America including parts of Uruguay and Argentina. As an illustration of this, we have computed annual precipitation means corresponding to the same regions that they defined, and we show a scatterplot of the data in the left-hand panel of Figure 4.4. The two variables are clearly correlated ($r = 0.38$; $p < 0.0001$). The correlation coefficient is lower than that found by Herweijer and Seager ($r = 0.57$), but this is explained by their use of a six-year moving average filter, which naturally increases the correlation. However, the feature of interest to us here is not the correlation in the middle of the distribution, but instead the dependence that exists in the lower tail (lower tail rather than upper tail, because our focus is drought). Therefore, the variables are transformed empirically to the unit Fréchet distribution (small values of precipitation corresponding to large values on Fréchet scale), with the results shown in the right-hand panel of Figure 4.4.

The effect of the Fréchet transformation is to highlight the most extreme observations in each variable. However, the most interesting observations are those that

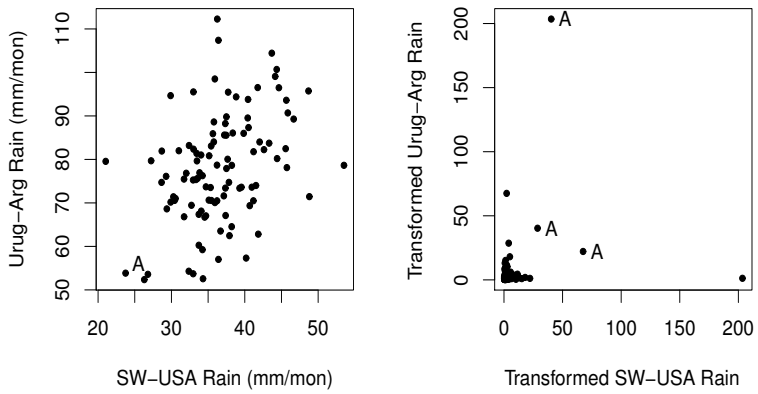

Figure 4.4 *Left: Plot of U.S. annual precipitation means over latitudes 25–35°N, longitudes 95–120°W, against Argentina annual precipitation means over latitudes 30–40°S, longitudes 50–65°W, 1901–2002. Right: Same data with empirical transformation to unit Fréchet distribution. Observations near the letter A in the left-hand plot and marked by A in the right-hand plot refer to simultaneous occurrences of extremely low precipitation in both locations. Data from gridded monthly precipitation means archived by the Climate Research Unit of the University of East Anglia (http://www.cru.uea.ac.uk/cru/data/hrg/timm/grid/CRU_TS_2_1.html, accessed November 15, 2012).*

are not close to either of the axes, because these correspond to observations that are extreme in both variables. In particular, the triangle of observations near the letter A in the left-hand plot are transformed into the observations marked A in the right-hand plot, which are all far from either axis. This is empirical evidence that there is indeed dependence between the most extreme values in this example. To go further, we have fitted one of the standard extremal dependence models—the logistic model, for which a detailed methodology based on events exceeding a threshold was developed by Coles and Tawn [32]. We have used rather a low threshold (2.5 on the unit Fréchet scale) in order to illustrate the applicability of the method; ideally, we would like to use a longer series and a higher threshold. An intuitive way to understand the effect of this model is to show how the probability of a jointly extreme event in both variables is inflated compared with what it would be if the variables were independent. For example, if we consider the 10-year return level (the value of each variable that would be exceeded with a probability of 0.1 in a single year), if the variables were independent, the probability that both 10-year return values would be exceeded is 0.01. Under the logistic model fitted to this dataset, the joint probability is 0.027 — an increase of 2.7 over the independent case. For more extreme events, the relative

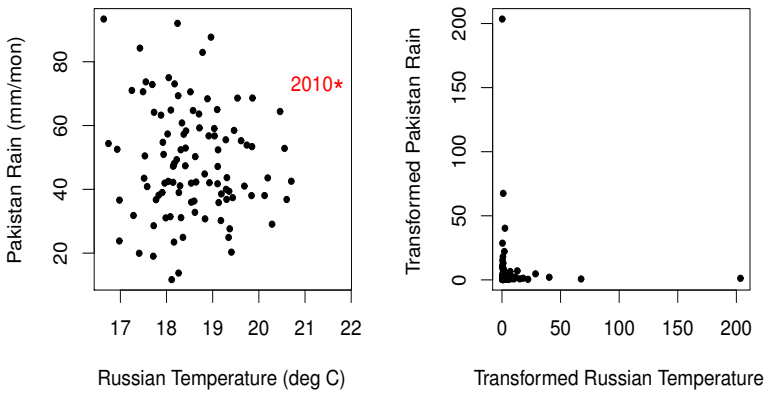

Figure 4.5 *Left: Plot of June, July, and August (JJA) Russian temperature means against Pakistan JJA precipitation means, 1901–2002. Right: Same data with empirical transformation to unit Fréchet distribution. Data from Climatic Research Unit, as in Figure 4.4. The Russian data were averaged over 45–65°N, 30–60°E, while the Pakistan data were averaged over 32–35°N, 70–73°E, same as in Lau and Kim [138].*

increase in joint probability compared with the independent case is larger — 4.7 for the 20-year return level, and 10.8 for the 50-year return level. However, confidence intervals for these relative increases in joint probability are quite wide. For example, for the 50-year return level, a 90% confidence interval is (2.1, 18.8), obtained by bootstrapping.

The logistic model, although very widely used in bivariate extreme value modeling, has a couple of well-documented disadvantages: It assumes symmetry between the two variables, and it has also a property known as asymptotic dependence, which might not be satisfied in practice. Recent work by Ramos and Ledford [188, 189] has suggested an alternative, more complicated, model that does not make those assumptions. They called this model the η-asymmetric logistic model, but for the present discussion we shall call it the Ramos–Ledford model. The estimation procedure used here follows Section 4.1 of [188]. Under this model, the estimated probability ratios are very similar to those of the logistic model, although the confidence intervals are somewhat wider. A summary of all the estimates and confidence intervals is in Table 4.1.

Example 2. Lau and Kim [138] have provided evidence that the 2010 Russian heat wave and the 2010 Pakistan floods were derived from a common set of meteorological conditions, implying a physical dependence between these two very extreme events. Using the same data source as for Example 1, we have constructed summer

	Logistic Model		Ramos-Ledford Model	
Period	Estimate	90% CI	Estimate	90% CI
10-year	2.7	(1.2, 4.2)	2.9	(1.2, 5.0)
20-year	4.7	(1.4, 7.8)	4.9	(1.2, 9.6)
50-year	10.8	(2.1, 18.8)	9.9	(1.4, 23.4)

Table 4.1 *Estimates of the Increase in Probability of a Joint Extreme Event in Both Variables, Relative to the Probability Under Independence, for the United States/Uruguay–Argentina Precipitation Data*

temperature means over Russia and precipitation means over Pakistan corresponding to the spatial areas used by Lau and Kim. Figure 4.5 shows a scatterplot; the left-hand plot is of the raw data, and the right-hand plot is of the data after transformation to the unit Fréchet distribution (with the largest values on the original plot corresponding to the largest value on Fréchet scale, because the right-hand tail is of interest here). Because the data source goes up only to 2002, we have approximated the 2010 values using a different data source (the National Centers for Environmental Prediction); this data point is shown in the left-hand panel of Figure 4.5 but is not included in the subsequent analysis. The 2010 value is clearly an outlier for temperature but not for precipitation. It should be noted that, while the 2010 Pakistan flooding was severe, the overall rainfall over northern Pakistan was not unprecedented. This is because the heavy rain was concentrated in a very small area over the upper Indus river basin, over a few days (Dr. W.K. Lau, Chief of Atmospheres, National Aeronautics and Space Administration, 2012, personal communication).

In contrast with Figure 4.4, the right hand plot of Figure 4.5 shows virtually no data point away from the axes, indicating that there is no evidence of dependence in the upper tail of the distribution. This is confirmed by repeating the same analyses as for Example 1, with results shown in Table 4.2. For the logistic model, which is constrained to positive dependence between the two variables, the point estimates and confidence intervals (for the ratio of joint probability to the independent case) are all very close to 1. Under the Ramos-Ledford model, which does not have that constraint, the estimated probability ratios are < 1 (indicating negative dependence), but the confidence intervals include 1. With either set of results, the net conclusion is that there is no evidence against the hypothesis of independence in the right hand tail of the distribution.

	Logistic Model		Ramos-Ledford Model	
Period	Estimate	90% CI	Estimate	90% CI
10-year	1.01	(1.00, 1.01)	0.33	(0.04, 1.4)
20-year	1.02	(1.00, 1.03)	0.21	(0.008, 1.8)
50-year	1.05	(1.01, 1.07)	0.17	(0.001, 2.9)

Table 4.2 *Similar to Table 4.1, but for the Russia-Pakistan Dataset*

Conclusions. Example 1 confirms and extends the results of Herweijer and Sea-

ger [114] by showing that the interdependence of drought conditions in the two given regions of the United States and South America extends to the tail of the distribution, although the confidence intervals for the probability ratios are still fairly wide as a result of the relatively small number of data points (102). However, Example 2 shows no evidence at all that there is any tendency for extreme high temperatures in Russia to be associated with extreme high precipitation in Pakistan; in other words, the 2010 event may have been truly an outlier without precedent in history. This should however be qualified by noting that the dataset used, consisting of monthly averages over half-degree grid cells, cannot be expected to reproduce extreme precipitation events over very short time and spatial scales, and it remains possible that an alternative data source, using finer-scale data, would produce a different conclusion.

4.7 Other Approaches to Multivariate Extremes

4.7.1 The conditional approach of Heffernan and Tawn

The asymptotically dependent models of classical multivariate extreme value theory and the asymptotically dependent approach started by Ledford and Tawn have dominated most of the recent literature on multivariate extremes, but there is also a third approach, introduced by Heffernan and Tawn [113]. This is the *conditional* approach to multivariate extreme value theory. The idea can be summarized as follows: suppose $\mathbf{Y}$ is a p-dimensional random vector but only one component Y_i (where $1 \leq i \leq p$) is extreme. What is the conditional distribution of $\mathbf{Y}_{(i)}$ — the vector $\mathbf{Y}$ omitting the ith component — given that Y_i exceeds some high threshold? They show that under suitable conditions, one can normalize the components of $\mathbf{Y}_{(i)}$ to get a non-degenrate limiting distribution as Y_i approaches its upper endpoint. As an example, the authors consider the distribution of a suite if air pollutants given that one of them becomes extreme. This is relevant to air pollution regulation questions since regulations are often focused on a single pollutant (e.g. particulate matter or ozone) but regulators are concerned that other pollutants may take very high values at the same time. Heffernan and Tawn provided some general theory on this approach and proposed some possible parametric models to take account of it.

The approach has found subseqent appications including applications to spatial extremes (references to be added).

4.7.2 Combining AD and AI models: The approach of Wadsworth et al.

References: [250, 251]

From the abstract of [251]:

Different dependence scenarios can arise in multivariate extremes, entailing careful selection of an appropriate class of models. In bivariate extremes, the variables are either asymptotically dependent or are asymptotically independent. Most available statistical models suit one or other of these cases, but not both, resulting in a stage in the inference that is unaccounted for, but can substantially impact subsequent extrapolation. Existing modelling solutions to this problem are either applicable only on sub-domains, or appeal to multiple limit theories. We introduce a unifed

representation for bivariate extremes that encompasses a wide variety of dependence scenarios, and applies when at least one variable is large. Our representation motivates a parametric model that encompasses both dependence classes. We implement a simple version of this model, and show that it performs well in a range of settings.

One concern that these authors express about the Ramos-Ledford approach [188] is that it is only applicable in the case where both variables are above a high threshold. In many cases, we are interested in data where only one component is extreme, which is reflective of the conditional approach of Heffernan and Tawn [113]. This paper proposes an alternative family of normalizations that encompasses both approaches.

4.7.3 De Haan and de Ronde

This was an applied paper [104] summarizing the result of a multi-year project to assess flooding risk on the Dutch coast. The authors collected 13 years of data measuring the joint distribution of sea level (SWL) and wave height (HmO) during storms along the North Sea coast. If the combination of SWL and HmO crosses a certain line, flooding occurs. The Dutch government has set a target that such events should be "10,000-year events", i.e. a 0.0001 failure probability in a given year. None of the events in the 13-year dataset comes anywhere close to this boundary, but there is obvious concern with simply extrapolating standard distributions to the boundary point. In proposing a solution to this problem, de Haan and de Ronde go through the theory of bivariate extreme value distributions, focusing on the classical formulation of bivariate exteme value theory but also acknowledging the (at the time, recent) theories of asymptotic independence.

One aspect of this problem is the following: suppose we have bivariate random variables (X,Y) but failure is define by some univariate function sch as $h(X,Y) > t$ where h is a scalar and t is a failure threshold. Do we apply bivariate extreme value theory to the $\{(X_i,Y_i), i = 1,\ldots,n\}$ pairs and then apply the resulting distributions to estimate the probability $\Pr\{h(X,Y) > t\}$, or do we simply calculate $H_i = H(X_i,Y_i)$ for each pair of observations and then apply the standard univariate techniques, such as fitting the generalized Pareto distributions to the exceedances of a high threshold, to the scalar observations $\{H_i\}$? The latter are sometimes called "structure variables". A very similar problem was discussed in an earlier paper by Coles and Tawn [33], who gave general arguments why one might in practice prefer the bivariate distribution approach over the structure variable approach/

4.7.4 General max-stable approach

Segers [213] wrote an elegant review of the classical "max-stable" approach to multivariate extremes, focusing on the role of copulas and max-stable models for copulas. The main contribution of the paper was an approach to generating families of multivariate extreme models for which there are still only a limited class of widely used models.

4.7.5 Multivariate generalized Pareto distributions

Rootzén and Tajvidi [202] proposed a way of defining "multivariate generalized Pareto distributions" that fulfil the twin properties that they arise as limits of exceedances over thresholds and that their form is preserved under change of the threshold level. They acknowledged the connection with earlier approaches such as [32, 131] but their paper was the first to propose a specific family of models with these properties.

4.7.6 High-dimensional multivariate extremes

From a letter I wrote in support of Dan Cooley's nomination as a Fellow of the American Statistical Association:

One of his more distinctive contributions is an entirely new method for high-dimensional multivariate extremes. Multivariate extreme value theory is concerned with joint distributions for the extremes of two or more variables; it originated in the 1950s with studies of the bivariate case, and was extended to arbitrary dimensions in the 1970s, but the statistical theory was always a challenge in high dimensions because of the curse of dimensionality issues that always arise in such problems. Dan and his collaborators have resolved that issue by proposing a version of principal components analysis (PCA) for high-dimensional extremes. Conventional PCA is based on either the covariance or the correlation matrix and pays no particular attention to extremes. The new method first defines a pairwise dependence measure based on exceedances over high thresholds, and then uses an ingenious transformation to define a tail pairwise dependence matrix (TPDM) with properties similar to a covariance matrix. In particular, in this transformed space it is possible to perform an eigenvector decomposition and select large-eigenvalue components similar to a conventional PCA. The mathematical theory of this approach was worked out by Cooley and Thibaud [38] and a striking application to US precipitation data was given by Jiang, Cooley and Wehner [129]. In that paper, the resulting components show strong spatial patterns in the data and also allow the authors to identify which of the components are correlated with the El Niño (ENSO) signal, which is an important technique for distinguishing short-term fluctuations in meteorological data from long-term trends due to climate change. Thus, the method has shown practical applications as well as theoretical innovation.

The TPDM idea has been used in other papers as well, for example Fix, Cooley and Thibaud (2020) where it was applied to a spatial process using an analog of the simultaneous autoregressive (SAR) construction that is familiar in spatial statistics. This leads to a spatial model indexed by a single parameter ρ that can be estimated quickly from a large number of spatial locations. They show, for example, that this method leads to quick estimates of spatial extremes that perform well in comparisons with the Brown-Resnick process, a well-established method that is computationally intensive.

Another innovative paper was Cooley, Thibaud, Castillo and Wehner (2019). This is more limited in that is only concerns bivariate data, but it is a nonparametric approach that gets away from the idea that "extreme events" occur only when both

the considered variables are extreme. For example, in considering the joint effects of temperature and humidity on human health, it is rarely the case that temperature and humidity are both extreme on the same day. However, it is still true that a combination of high temperature and high humidity are the most significant events for adverse health outcomes. Michael Wehner is a well-known climate scientist who has been active on both IPCC (international) and US climate assessments; he has several times remarked to me how important he considers Dan's work for assessing the impacts of extreme climate events. Another contribution of a more technical nature in this paper is its dual treatment of the "asymptotically dependent" and "asymptotically independent" cases of bivariate extremes; this is another area to which Dan has contributed extensively.

Chapter 5

Spatial Extremes

Many modern data sources are spatial in nature. This is especially true in environmental fields such as climate and air pollution. As an example the Global Historical Climatological Network[1] includes daily data on numerous climate variables measured at weather stations across the world. In the United States, the Environmental Protection Agency's Air Quality System[2] provides data on the major components of air pollution. New sources of data are available from satellite observations such as the Orbiting Carbon Observatory-2 (OCO-2) of the National Aeronautics and Space Administration[3]. In addition, climate data from computer models, both historical and future projections, are available though the sequence of experiments of the Coupled Model Intercomparison Project [240, 68]. Similar sources exists for hydrological, oceanographic and a whole host of other data types. Much of modern environmental statistics is concerned with new statistical methods for analyzing these vast data sources, and extreme value theory is no exception.

Broadly speaking, problems of spatial extremes are of two types. One the one hand are the same problems that arise in univariate extreme value theory, such as estimating probabilities of high-level exceedances, or long-period return values, but at many sites simultaneously rather than just one site at a time. In this context, ideas of "borrowing strength" and exploiting spatial smoothness (in other words, extreme value parameters at neighboring sites may be expected to be similar) are critical towards making the most computationally and statistically efficient use of data, but they don't directly address issues of spatial dependence, e.g. whether a particularly severe rainstorm leads to extreme precipitation at several sites simultaneously (or how large an area is affected). Problems of that nature require attention to the extreme dependence properties of observations at multiple sites, which we have already discussed extensively in Chapter 4 on multivariate extremes. From that point of view, spatial extreme value theory is the extension of multivariate extreme value theory to infinitely many components, but where the spatial structure of the system suggests specific classes of statistical models. In this context, the family of *max-stable processes* is of particular importance.

[1] https://www.ncdc.noaa.gov/data-access/land-based-station-data/land-based-datasets/global-historical-climatology-network-ghcn

[2] https://www.epa.gov/aqs

[3] https://oco.jpl.nasa.gov/

Davison et al. [49] distinguished three types of spatial extreme value analysis, based on *latent variable models*, *copula models* and *max-stable models*. Broadly speaking, latent variable models are used in answering the first of the questions just described, where we are trying to describe univeriate extremal properties simultaneously at multiple sites, while copula and max-stable models are trying to deal directly with spatial dependence. Here, we focus primarily on the latent variable and max-stable approaches, introduced in Sections 5.1 and 5.2 respectively, while copulas and some more recent approaches are covered in 5.5.

There have already been a number of review papers on spatial extremes — apart from [49], there was also a review by Cooley and co-authors [39], and we shall also quote extensively from a more recent review by Davison and co-authors [48].

5.1 The Latent Process Approach

To introduce this topic, we discuss in detail the model from a recent paper by Russell and co-authors [204]. After that, we shall show something of the history of these approaches, which have used both Bayesian and non-Bayesian methods, though in some respects the non-Bayesian approaches have shown more flexibility in handling the random noise component of the model.

In the block maxima approach to extremes, the distribution of a block maximum at a particular site is modeled through the Generalized Extreme Value (GEV) distribution,

$$\Pr\{Z \le z\} \quad = \quad \exp\left\{-\left(1+\xi\frac{z-\mu}{\psi}\right)_+^{-1/\xi}\right\}, \tag{5.1}$$

where μ, ψ and ξ are the usual GEV location, shape and scale parameters. We will often write this in the form $Z \sim \mathrm{GEV}(\mu,\psi,\xi)$.

This leads to the following formula for the p-quantile of the distribution:

$$Z_p(\mu,\psi,\xi) = \begin{cases} \mu - \frac{\psi}{\xi}\{1-(-\log p)^{-\xi}\} & \text{if } \xi \neq 0, \\ \mu - \psi\log(-\log p) & \text{if } \xi = 0. \end{cases} \tag{5.2}$$

When blocks correspond to years (this is the most common assumption, though it is also possible to compare blocks of different lengths to improve the bias-variance tradeoff), we often identify the $1-1/r$ quantile with the *r-year return level*, thus

$$\mathrm{RL}_r \quad = \quad Z_{1-1/r}(\mu,\psi,\xi). \tag{5.3}$$

For the rest of this section, we shall indeed assume blocks correspond to years. However, the parameters μ, ψ and ξ may be dependent on both space and time, and this in essence is at the core of our proposed modeling procedure.

Suppose $Y_t(\mathbf{s})$ is the annual maximum in year t and location $\mathbf{s}$, then consider a model of the form

$$Y_t(\mathbf{s}) \quad \sim \quad \mathrm{GEV}(\mu_t(\mathbf{s}),\psi_t(\mathbf{s}),\xi_t(\mathbf{s})) \tag{5.4}$$

where $\mu_t(\mathbf{s}), \psi_t(\mathbf{s}), \xi_t(\mathbf{s})$ are the location, scale and shape parameters at location $(\mathbf{s})$ in year t.

The paper [204] was especially concerned with the relationship between hurricane-season annual maximum precipitation and spring-season mean sea surface temperatures in the Gulf of Mexico, denoted SST_t, so they considered a model of the form

$$\begin{aligned} \mu_t(\mathbf{s}) &= \theta_1(\mathbf{s}) + \mathrm{SST}_t\,\theta_2(\mathbf{s}), \\ \log \psi_t(\mathbf{s}) &= \theta_3(\mathbf{s}) + \mathrm{SST}_t\,\theta_4(\mathbf{s}), \\ \xi_t(\mathbf{s}) &= \theta_5(\mathbf{s}) \end{aligned} \tag{5.5}$$

Here, the model for $\psi_t(\mathbf{s})$ was expressed on a logarithmic scale which is more natural given that $\psi_t(\mathbf{s}) \geq 0$, while in common with many other studies, they treated the shape parameter $\xi_t(\mathbf{s})$ as constant in time (but not space).

Define also

$$\boldsymbol{\theta}(\mathbf{s}) = \begin{bmatrix} \theta_1(\mathbf{s}) \\ \vdots \\ \theta_5(\mathbf{s}) \end{bmatrix}. \tag{5.6}$$

The objective is to come up with a spatial model for the five-dimensional spatial process $\boldsymbol{\theta}(\mathbf{s})$ as a function of $\mathbf{s}$ in some domain $\mathscr{D}$, typically represented as a subset of $\mathbb{R}^2$.

In general, of course, time-dependent covariates should be chosen to be relevant to the specific application of interest; the model (5.5) is intended to be representative of a general class of models for which the GEV parameters are expressed as functions of known covariates and a vector spatial process $\boldsymbol{\theta}(\mathbf{s}),\ \mathbf{s} \in \mathscr{D}$, and the objective is to come up with a suitable model for the process $\boldsymbol{\theta}$.

5.1.1 Background on Spatial Statistics

At this point, it is useful to give a little background on models that are used for spatial processes, with specific reference to Gaussian spatial processes, which will be our main emphasis for the following development. Some standard references for this material include books by Cressie [41] and by Schabenberger and Gotway [210].

A standard model for a univariate spatial process is a Gaussian process model of the form

$$\mathbf{Z} \sim \mathscr{N}_n(X\beta, V), \tag{5.7}$$

where $\mathbf{Z} = \begin{bmatrix} Z(\mathbf{s}_1) \\ \vdots \\ Z(\mathbf{s}_n) \end{bmatrix}$ is a vector of observations at n spatial locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$ in a domain $\mathscr{D} \subset \mathbb{R}^k$ of dimension $k \geq 1$, $\mathscr{N}_n$ denotes the n-dimensional normal

distribution, X is an $n \times p$ matrix of known covariates, β is a p-dimensional vector of regression parameters, and V is a $n \times n$ covariance matrix with entries $\{v_{ij}\}$ where

$$v_{ij} = \text{Cov}\{Z(\mathbf{s}_i), Z(\mathbf{s}_j)\}$$

is the covariance between two observations measured at sites $\mathbf{s}_i$ and $\mathbf{s}_j$.

In general, V can be arbitrary subject to the constraint that it is a non-negative definite matrix. In practice, it is usual to impose further assumptions on V:

(a) If v_{ij} depends on the sites $\mathbf{s}_i,\ \mathbf{s}_j$ only through the vector difference $\mathbf{s}_i - \mathbf{s}_j$, the process is said to be *stationary*;

(b) If v_{ij} in invariant under rotations, so that $v_{ij} = \text{Cov}\{Z(Q\mathbf{s}_i), Z(Q\mathbf{s}_j)\}$ for any rotation matrix Q, the process is said to be *isotropic*;

(c) The most common assumption is that the process is both stationary and isotropic, in which case we may write

$$v_{ij} = \sigma^2 C(d_{ij}),\ d_{ij} = ||\mathbf{s}_i - \mathbf{s}_j||, \tag{5.8}$$

where σ^2 is the (assumed constant) variance of the process and $C(d)$ is the correlation between two sites separated by a scalar distiance d; in this case, the process is called *homogeneous*.

There are a number of standard parametric models for homogeneous spatial processes, for example:

$$C(d) = \exp(-d/d_0),\ \text{(Exponential correlation)} \tag{5.9}$$

$$C(d) = \exp\{-(d/d_0)^2\},\ \text{(Gaussian correlation)} \tag{5.10}$$

$$C(d) = \exp\{-(d/d_0)^p\},\ 0 < p \leq 2,\ \text{(Exponential-power correlation)} \tag{5.11}$$

$$C(d) = \begin{cases} 1 - \frac{3}{2}\frac{d}{d_0} + \frac{1}{2}\left(\frac{d}{d_0}\right)^3, & d \leq d_0, \\ 0, & d \geq d_0, \end{cases} \text{(Spherical correlation)} \tag{5.12}$$

$$C(d) = \frac{1}{2^{\phi_2 - 1}\Gamma(\phi_2)} \cdot \left(\frac{2\sqrt{\phi_2}d}{\phi_1}\right)^{\phi_2} \cdot K_{\phi_2}\left(\frac{2\sqrt{\phi_2}d}{\phi_1}\right).\ \text{(Matérn correlation)} \tag{5.13}$$

All of these are valid in any dimension k except for the spherical model (5.12), which is valid only for $k = 1, 2, 3$. In each of (5.9)–(5.12), the parameter d_0 is called the *range* and loosely represents the range of distances for which the spatial correlation is effective, though only for the spherical model (5.12) is it exactly true that $C(d) = 0$ when $d \geq d_0$. In (5.13), $\Gamma(\cdot)$ represents the standard gamma function and $K_{\phi_2}(\cdot)$ is a modified Bessel function of the third kind of order ϕ_2. The parameters are $\phi_1 > 0,\ \phi_2 > 0$ where ϕ_1 plays essentially the same role as d_0 in (5.9)–(5.12) and ϕ_2 represents a shape parameter; the case $\phi_2 = \frac{1}{2}$ is equivalent to exponential (5.9) and the limit $\phi_2 \to \infty$ is equivalent to Gaussian (5.10). The process is generally named after Matérn [155] though it had earlier been derived by Whittle [261].

The spherical model often appeals to beginners because of its simple algebraic

structure and the fact that $C(d) = 0$ when $d \geq d_0$; however, the lack of smoothness near $d = 0$ and $d = d_0$ makes it less appealing as a statistical model. Wendland [260] sought to remedy this by proposing a general scheme to construct covariance functions, expressible through simple polynomials with finite range, that are non-negative definite in any dimension and have additional smoothness properties. An example is the Wendland 2.2 model

$$C(d) = \begin{cases} \left(1-\frac{d}{d_0}\right)^6 \left\{ \frac{35}{3}\left(\frac{d}{d_0}\right)^2 + 6\left(\frac{d}{d_0}\right) + 1 \right\}, & d \leq d_0, \\ 0, & d \geq d_0, \end{cases} \tag{5.14}$$

valid in dimension $k = 2$ or 3. These functions have been implemented in the R functions `Wendland` or `Wendland2.2` in the package *fields* [168].

5.1.1.1 Intrinsic Stationarity and the Semivariogram

An alternative (and, in fact, slightly more general) way of characterizing stationary spatial processes is through the *semivariogram* rather than the spatial variance or correlation function.

In general, a semivariogram is a function of a pair of spatial coordinate vectors,

$$\gamma(\mathbf{s}_1, \mathbf{s}_2) = \frac{1}{2}\mathrm{E}\{(Z(\mathbf{s}_1) - Z(\mathbf{s}_2))^2\}. \tag{5.15}$$

Without the multiplier $\frac{1}{2}$, this would be called the variogram rather than semivariogram. In practice, the most common examples assume constant mean, stationary and isotropy, in which case the semivariogram reduces to a function of scalar distance,

$$\gamma(\mathbf{s}_1, \mathbf{s}_2) = \gamma_0(||\mathbf{s}_1 - \mathbf{s}_2||). \tag{5.16}$$

Such a process is called *intrinsically stationary* to distinguish it from (5.8); there are processes which are intrinsically stationary without being stationary; an example is (5.17) below.

If $Z(\mathbf{s})$, $\mathbf{s} \in \mathscr{D}$ is a homogeneous process with variance σ^2 and correlation function $C(\cdot)$ defined by (5.8), then it is readily varified that

$$\gamma_0(d) = \sigma^2(1 - C(d))$$

and hence, any of (5.9)–(5.14) may be expressed in terms of its semivariogram.

However, there are also semivariograms that do not correspond to homogenous covariance functions, of which one of the best known is

$$\gamma_0(d) = \phi_1 + \phi_2 d^\lambda \tag{5.17}$$

where $0 \leq \lambda < 2$; this will be particularly useful in Section 5.2.

5.1.1.2 Lattice Models

The spatial models developed up to this point rely heavily on the geographic coordinates of the locations $\mathbf{s}_1, ..., \mathbf{s}_n$ of the n data points. An alternative approach, originally introduced by Besag [19], is often considered more appropriate when the observations are arranged on na regular lattice. The simplest form of such model is the *conditional autoregressive* process, usually written CAR. The simplest form of such a model assumes conditional distributions

$$Z_i \mid Z_j,\ j \neq i \quad \sim \quad \mathscr{N}\left(\frac{\sum_j w_{i,j}}{w_{i+}}, \frac{\lambda}{w_{i+}}\right) \tag{5.18}$$

where Z_i is the observation at the ith lattice point, $\lambda > 0$ is a fixed conditional variance parameter and $\{w_{i,j}\}$ are weights: the most common specification sets $w_{i,j} = 1$ if the lattice points i and j are neighbors, 0 otherwise. Although the basic model (5.18) is very simple, it is often used as one component of a Bayesian hierarchical model for spatial data; an example will be seen in Section 5.1.4.

5.1.1.3 Estimation of Gaussian Spatial Processes

We return to the model (5.7), where we assume $\mathbf{Z}$ is $n \times 1$, X is $n \times p$, β is $p \times 1$ and the $n \times n$ matrix V is of the form $\sigma^2 C(\boldsymbol{\phi})$ where σ^2 is an assumed constant variance and $C(\boldsymbol{\phi})$ a spatial correlation function depending on some finite-dimensional parameter vector $\boldsymbol{\phi}$; any of the models (5.9)–(5.14) would be suitable candidates for this. In this section, we assume finite-dimensional $\boldsymbol{\phi}$ because that facilitates estimation by the method of maximum likelihood or its close relative, the restricted maximum likelihood (REML) method. Another reason for restricting to parametrically specified correlation functions is that each of the functions (5.9)–(5.14) has been proved to be positive definite, so in this way we avoid the potential awkwardness of a spatial correlation estimate that does not satisfy that constraint.

To compute the maximum likelihood estimator (MLE), the parameters $(\beta, \sigma^2, \boldsymbol{\phi})$ are chosen to maximize the likelihood function

$$f(\mathbf{Z} \mid \beta, \sigma^2, \boldsymbol{\phi}) \quad = \quad (2\pi\sigma^2)^{-n/2} |C(\boldsymbol{\phi})|^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{Z} - X\beta)^T C(\boldsymbol{\phi})^{-1}(\mathbf{Z} - X\beta)\right\}. \tag{5.19}$$

Alternatively, the restricted maximum likelihood (REML) estimator replaces the right side of (5.19) by

$$(2\pi\sigma^2)^{-(n-p)/2} |C(\boldsymbol{\phi})|^{-1/2} |X^T C(\boldsymbol{\phi})^{-1} X|^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{Z} - X\beta)^T C(\boldsymbol{\phi})^{-1}(\mathbf{Z} - X\beta)\right\}. \tag{5.20}$$

A third alternative is Bayesian estimation: define a prior density $\pi(\beta, \sigma^2, \boldsymbol{\phi})$ and integrate the prior $\times$ likelihood function $\pi(\beta, \sigma^2, \boldsymbol{\phi}) f(\mathbf{Z} \mid \beta, \sigma^2, \boldsymbol{\phi})$. This is usually achieved by some version of a Markov chain Monte Carlo (MCMC) procedure.

5.1.1.4 Spatial Models with Measurement Error

An important subclass of spatial models arises when each spatial observation is measured with some quantifiable measurement error. Sometimes the "observation" is itself a parameter estimate from some regression model or generalized linear model (GLM), such as a secondary analysis at each spatial location resulting in a parameter estimate and standard error, which is then combined across multiple locations to produce a regional or national estimate for the parameter of interest. Models of this structure have been extensively studied in epidemiology [60, 14, 59, 236] and also in the present context of extreme value estimation (references to follow), but here we follow a paper by Holland et al. [116], which laid out many of the issues that arise in this kind of analysis. (The notation of [116] has been changed to make it more consistent with what we are using for the precipitation extremes analysis.)

The objective of the paper [116] was to study time trends in sulfur dioxide (SO_2) across a network of rural monitoring sites in the eastern U.S. At each site, a generalized additive model (GAM) was fitted to estimate the linear trend in SO_2 as a function of weather and a number of other covariates. We omit the details of this part of the analysis as they are not relevant for the following discussion. The outcome of this initial GAM analysis was an estimate of the trend at each location, together with its estimated standard error (SE). The objective of the spatial part of the analysis was to combine these estimates across sites, possibly including additional covariates (e.g. latitude and longitude), to calculate regional estimates for different regions defined within the overall study area.

In this formulation, we may assume an unobserved "true trend" process $\theta(\mathbf{s})$ measured at each of a number of sites $\mathbf{s} = \mathbf{s}_1, ..., \mathbf{s}_n$. Represented in vector notation as $\boldsymbol{\theta} = \begin{bmatrix} \theta(\mathbf{s}_1) \\ \vdots \\ \theta(\mathbf{s}_n) \end{bmatrix}$, we assume the model

$$\boldsymbol{\theta} \sim \mathcal{N}_n(X\beta, \sigma^2 C(\boldsymbol{\phi})), \tag{5.21}$$

similar to (5.7), with some correlation matrix C depending on parameters $\boldsymbol{\phi}$.

However, in addition to (5.21), we assume an "estimated trend" of the form

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta} + \mathbf{e},\ \mathbf{e} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}, \tag{5.22}$$

in other words, the estimated trend $\hat{\theta}(\mathbf{s}_i)$ at site $\mathbf{s}_i$ ($1 \le i \le n$) consists of the unobserved true trend $\theta(\mathbf{s}_i)$ plus a measurement error e_i.

To complete the model, we assume

$$\mathbf{e} \sim \mathcal{N}_n(0, W) \tag{5.23}$$

with some covariance matrix W which, for the moment, we assume *known*. We also assume the measurement error $\mathbf{e}$ is independent of the true process $\boldsymbol{\theta}$.

Combining (5.21), (5.22), (5.23), we have

$$\hat{\boldsymbol{\theta}} \quad \sim \quad \mathcal{N}_n(X\beta, \sigma^2 C(\boldsymbol{\phi}) + W) \tag{5.24}$$

from which we can estimate the parameters β, σ^2, $\boldsymbol{\phi}$ by maximum likelihood or REML. It is also possible to calculate the conditional distribution of $\boldsymbol{\theta}$ given $\hat{\boldsymbol{\theta}}$ in the form

$$\mathcal{N}_n\left[X\beta + V(V+W)^{-1}(\hat{\theta} - X\beta), V(V+W)^{-1}W\right] \tag{5.25}$$

where $V = \sigma^2 C(\boldsymbol{\phi})$. (5.25) is, in effect, a kriging formula that may be derived by first writing down the joint distribution of $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$ and then inverting. Estimates and standard deviations of integrated quantities, such as regional averages, may be computed from (5.25).

We still have to discuss how the error covariance matrix W is estimated in practice. Given that we have an estimated standard error of each $\hat{\theta}(\mathbf{s}_i)$ from the initial regression analysis, the simplest estimator of W is a diagonal matrix where the diagonal entries are the squares of the standard errors. This in effect assumes that the errors e_i are independent at each site, an assumption that may not be correct.

As an alternative to the independence assumption, [116] used a bootstrap method to estimate, not only the variances of $\hat{\theta}(\mathbf{s}_i)$ at each site $\mathbf{s}_i$, but also the covariances across sites; these bootstrap estimate were used to estimated the full covariance matrix W. Comparisons showed that the full covariance matrix improved on the diagonal approximation as assessed by the maximized likelihood of the model (5.24), and that estimates and standard errors of the desired regional averages were meaningfully different when computed with diagonal or non-diagonal W (Table 1 and Figure 8 of [116].

Summary:

(a) The model defined by (5.21)–(5.23) is a viable approach for taking account of measurement error in the estimates $\hat{\theta}(\mathbf{s}_i)$;

(b) The spatial model parameters β, σ^2, $\boldsymbol{\phi}$ may be estimated by using maximum likelihood, REML or Bayesian estimation in the combined model (5.24), and predictions obtained using (5.25) (although we have not attempted to write downt he formula here, this is easily extended to obtain predictions and prediction variances of $\theta(\mathbf{s})$ at unmeasured sites as well);

(c) Although the simplest version of the model assumes diagonal W, it may also be beneficial to consider non-diagonal W.

5.1.2 Application to Precipitation Extremes Example

As recalled from (5.4)–(5.6), the model we are considering allows for a five-dimensional vector of extreme value parameters, $\boldsymbol{\theta}(\mathbf{s})$, for each site $\mathbf{s}$ at which we have measurements. At each weather station $\mathbf{s}_i$, $i = 1, ..., n$, we can compute estimates $\hat{\boldsymbol{\theta}}(\mathbf{s}_i)$ by maximum likelihood, including the estimated covariance matrix of the estimates at each site. As in Section 5.1.1.4, we can represent this as the sum of

some smooth latent process $\boldsymbol{\theta}(\mathbf{s})$, $\mathbf{s} \in \mathscr{D}$, and a random error process. However, there is an additional twist, that this is now a 5-variate spatial process, whereas the model described in Section 5.1.1.4 was univariate.

Following [204], the model for $\boldsymbol{\theta}$ does not involve any spatial covariates and may be written

$$\boldsymbol{\theta}(\mathbf{s}) \quad = \quad \beta + \eta(\mathbf{s}) \tag{5.26}$$

where $\beta \in \mathbb{R}^5$ and η is a 5-dimensional spatial process with mean 0. The model (5.26) is more complicated than those of Section 5.1.1 because the process η is multivariate; however, a common way to deal with that is to use *co-regionalization* [249, 76, 11]. We write

$$\eta(\mathbf{s}) \quad = \quad A\delta(\mathbf{s}) \tag{5.27}$$

where $\delta(\mathbf{s}) = \begin{pmatrix} \delta_1(\mathbf{s}) \\ \vdots \\ \delta_5(\mathbf{s}) \end{pmatrix}$ with $\delta_1, \ldots, \delta_5$ independent univariate zero-mean spatial processes and A is a 5×5 lower triangular matrix.

For the individual δ_ℓ processes, $\ell = 1, \ldots, 5$, [204] assumed exponential covariances,

$$\mathrm{E}(\delta_\ell(\mathbf{s}_i)\delta_\ell(\mathbf{s}_j) \quad = \quad \exp\left(-\frac{||\mathbf{s}_i - \mathbf{s}_j||}{\rho_\ell}\right) \tag{5.28}$$

with range parameters $\rho_1, \ldots, \rho_5$.

To develop a two-stage inference in this model, we first assume estimates

$$\hat{\boldsymbol{\theta}}(\mathbf{s}) \quad = \quad \begin{pmatrix} \hat{\theta}_1(\mathbf{s}) \\ \vdots \\ \hat{\theta}_5(\mathbf{s}) \end{pmatrix}$$

by maximum likelihood at each observed site $\mathbf{s}$.

Next, we assume

$$\hat{\boldsymbol{\theta}}(\mathbf{s}) \quad = \quad \boldsymbol{\theta}(\mathbf{s}) + \mathbf{e}(\mathbf{s})$$

where $\mathbf{e}(\mathbf{s}) = \begin{pmatrix} e_1(\mathbf{s}) & \dots & e_5(\mathbf{s}) \end{pmatrix}^T$ and we assume

$$\begin{pmatrix} e_1(\mathbf{s}_1) \\ \vdots \\ e_1(\mathbf{s}_n) \\ e_2(\mathbf{s}_1) \\ \vdots \\ e_2(\mathbf{s}_n) \\ \vdots \\ e_5(\mathbf{s}_1) \\ \vdots \\ e_5(\mathbf{s}_n) \end{pmatrix} \sim \mathcal{N}_{5n}[0, W]$$

with W some $5n \times 5n$ matrix that we still have to define.

If we define vectors $\Theta = \begin{pmatrix} \theta_1(\mathbf{s}_1) & \dots & \theta_1(\mathbf{s}_n) & \dots & \theta_5(\mathbf{s}_1) & \dots & \theta_5(\mathbf{s}_n) \end{pmatrix}^T$ and $\hat{\Theta}$ similarly, the model becomes

$$\Theta \sim \mathcal{N}_{5n}\left[\beta \otimes \mathbf{1}_n, \Sigma_{A,\rho} + W\right]$$

where $\otimes$ denotes Kronecker product, $\mathbf{1}_n$ is an n-dimensional vector of ones, and $\Sigma_{A,\rho}$ is the assumed covariance matrix of Θ, which depends on matrix A and exponential range parameters $\rho = \begin{pmatrix} \rho_1 & \dots & \rho_5 \end{pmatrix}$ through (5.27) and (5.28).

The model will therefore be fully defined as soon as we specify W, and there are three possibilities for that:

(a) The simplest model is to assume independence of the GEV estimation errors between sites. This does not lead to a diagonal matrix W, because the 5×5 covariance matrix at each site will be derived from the observed information matrix at each site, which is not diagonal, but this is the multidimensional equivalent of assuming diagonal W in Section 5.1.1.4.

(b) A second possibility is to estimate W through a block bootstrap approach, applied simultaneously across all sites, which will allow estimation of covariances between sites as well as within sites. This would be the direct generalization of the method proposed in Section 5.1.1.4, but the disadvantage of this approach is that the dimension of the covariance matrix is much larger, and it is by now well known that the standard sample covariance matrix does not perform well in high dimensions, some form of regularization being needed [45, 20].

(c) The proposed form of regularization is covariance tapering [80]. Exploiting the fact that the term-by-term product of two positive definite covariance matrices is also positive definite, the proposed estimator takes the sample covariance matrix from the block bootstrap and multiplies it term-by-term by a tapering matrix, which in this example is taken to be the Wendland 2.2 covariance function (5.14), with range d_0 replaced by a user-defined maximum radius λ. This produces a covariance matrix that is sparse in the sense that any sites more than a distance

λ apart have covariance zero, but it is also faithful to the sample covariances at small distances. The authors [204] took $\lambda = 75$ km. based on the intuition that this is about the maximum possible radius of a single rainstorm, though they also performed sensitivity analyses to show that the resulting estimates are not highly dependent on the choice of λ.

With W specified in this way, the remaining parameters β, A, ρ were estimated by the method of maximum likelihood applied to $\hat{\Theta}$, and kriging was used to construct estimates $\tilde{\boldsymbol{\theta}}(\mathbf{s})$ and their mean squared prediction errors (MSPEs) at all sites $\mathbf{s} \in \mathscr{D}$ (not restricted to sampling sites).

5.1.3 Results

In this section, we briefly summarize how the results of this analysis when they were applied tpo the Gulf of Mexico data. For further information, the reader is referred to the original paper [204].

The source of precipitation data was the Global Historical Climatological Network (GHCN) mentioned at the start of this chapter. From the stations in this network, 326 weather stations with nearly complete records were identified from six U.S. states bordering the Gulf of Mexico (Florida, Georgia, Alabama, Mississippi, Louisiana and the eastern half of Texas). The authors somewhat arbitrarily chose 1949 as the start year of the analysis, and 2016 as the finish year (i.e. the year before Hurricane Harvey, so as not to be biased by including data from Harvey itself). For each station and each year, the highest 7-day total precipitation from June to November was calculated; this portion of the year is generally defined to be the Atlantic hurricane season. Sea surface temperature (SST) data were taken from the "HadISST"[4] dataset and used to calculate March–June average SST over an area roughly corresponding to the Gulf of Mexico (21°–29° N, 83°–97° W). The March-June period was chosen because it corresponds to the Spring period when high SSTs are considered to be driving the strength of the following hurricane season. After rescaling to a mean of 0 and a standard deviation of 1, the authors defined rescaled SST equal to –1 to be a "low SST" season rescaled SST equal to +1 to be a "high SST" season; on this scale, the 2017 SST corresponded to a rescaled value of about 1.7. They then used the method previously described to estimate, on a pixel by pixel basis, the probability of observing, in a single season, a seven-day precipitation in excess of 70 cm., which was the value actually observed in Houston during 2017. The results are plotted in Figure 5.1.

The results show, first, that it is indeed the eastern Texas region (including Houston itself) where the probability of such high precipitation is greatest, but that the probability itself varies substantially with SST, about 0.002 (i.e. a one-in-500-year event) at low SST, rising to about 0.004 at high SST and about 0.007 (about a one-in-140-year event) at 2017 SST. These estimated return probabilities are substantially higher than those found by other researchers, with the exception of [197] who used very similar methods but without the spatial component of this analysis. For the fu-

[4] https://www.metoffice.gov.uk/hadobs/hadisst/

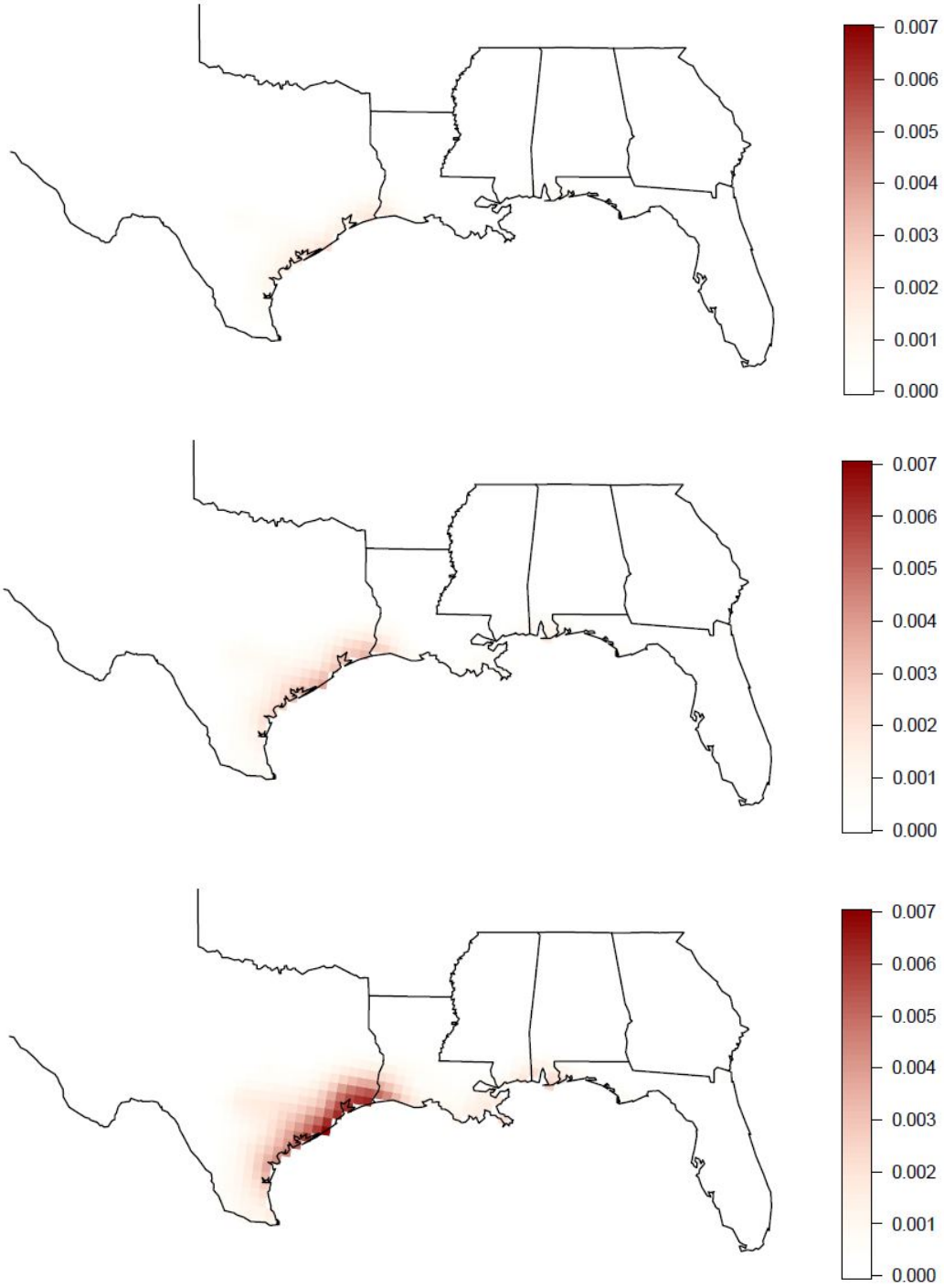

Figure 5.1 *Estimated probability that the annual maximum seven-day rainfall event exceeds 70 cm. under three scenarios: low SST (top); high SST (middle); 2017 SST (bottom). From [204].*

ture, if current trends continue, SSTs in the Gulf of Mexico in a few decades will reach levels well about the 2017 value, so it is to be antipicated that the probability of such extreme events in the future will substantially rise.

5.1.4 Literature Review

There are many precedents to the ideas outlined in this section; previous references include [27, 229, 264, 69, 36, 81, 207, 208, 37, 212, 245, 112, 63, 246, 205].

To the best of our knowledge, the earliest published model of this form was the paper by Casson and Coles [27]. These authors developed a model for extreme windspeeds as a function of location $\mathbf{s}$ in a space $\mathscr{S}$, where for each location, the model was of point process form (Chapter 1), with extreme value parameters $\mu(\mathbf{s}),\ \psi(\mathbf{s}),\ \xi(\mathbf{s})$. For the parameters $\mu(\mathbf{s}),\ s \in \mathscr{S}$, they assumed a model equation of the form

$$h_\mu(\mu(\mathbf{s})) \quad = \quad f_\mu(\mathbf{s};\beta_\mu) + Z_\mu)\mathbf{s};\alpha_\mu),$$

with models of similar structure for the other extreme value parameters $\psi(\mathbf{s}),\ \xi(\mathbf{s}),\ \mathbf{s} \in \mathscr{S}$. In this equation, h_μ is a known link function (usually assumed either the identity or a log link function), $f_\mu(\mathbf{s};\beta_\mu)$ is a regression model depending on finite-dimensional parameters β_μ, and $Z_\mu)\mathbf{s};\alpha_\mu)$ is a continuous-parameter spatial process (e.g. a Gaussian process with one of the covariance functions discussed in Section 5.1.1 depending on a finite-dimensional parameter α_μ. At the j'th site $\mathbf{s}_j$, the observed data $\mathscr{E}_j$ are assumed to consist of a set of high-threshold exceedances whose distribution are defined by a point process extreme value model with parameters $\mu(\mathbf{s}_j, \psi(\mathbf{s}_j, \xi(\mathbf{s}_j$. [27] assumed the three spatial processes $Z_\mu\ , Z_\psi,\ Z_\xi$ were independent though they remarked that this assumption is easily generalized (although they did not state an explicit alternative model, the formula (5.27) adopted in this chapter would be one example of how to do that). To complete the specification of the model, it is necessary to define prior distributions for the parameters $\alpha_\mu, \beta_\mu, \alpha_\psi, \beta_\psi, \alpha_\xi, \beta_\xi$; they assumed uniform priors, though again, they noted that more general prior distributions could have been assumed with no change in the basic methodology.

In the rest of their paper, they described a detailed Metropolis-Hastings algorithm for constructing posterior distributions in this model, demonstrated its efficiency on simulated data, and then discussed a real-world example based on hurricane data at 55 locations on the U.S. east coast. They noted, in particular, that this approach may be used to calculate predictive distributions for extreme winds at locations both on and off the measurement grid, which is something that cannot be done with more conventional method.

At the end of their paper, they noted,

"The biggest limitation of our model is the assumption of conditional independence of data given the latent extreme value parameters. Most real-life examples would require a more detailed consideration of spatial dependence, and the development of a spatial regression model which can handle data that are spatially dependent after allowance for parameter variation remains an important research objective."

The assumption they are referring to is equivalent to the assumption that the covariance matrix W in (5.23) is diagonal, and as we have noted, this assumption would indeed be restrictive. This point remains a limitation of Bayesian methods for this kind of problem, though as we shall see in the remainder of this chapter, by now a number of alternative models are available for spatial extremes, and these could be used to extend the Casson-Coles model.

We shall not attempt a complete review of all the other methods using latent processes, but summarize a few of them here:

(a) Cooley et al. [36] constructed a variant of the Casson-Coles model based on the generalized Pareto distributon (GPD) rather than the point process representation used by Casson and Coles. In this formulation, $\sigma(\mathbf{s})$ and $\xi(\mathbf{s})$ represent the scale and shape parameters of the GPD at location $\mathbf{s}$, and there is also a third process, that they wrote $\zeta(\mathbf{s})$, corresponding to the marginal probability of exceeding the threshold at site $\mathbf{s}$. Here, the threshold itself was treated as fixed (same at every site), after considering several alternative values. An alternative approach might be to define the threshold at each site as a fixed percentile (e.g. 95th or 99th) of the observations at each site, but that would create complications when it came to interpolating the data between stations. They assumed that $\log \sigma(\mathbf{s})$, $\xi(\mathbf{s})$ and $\text{logit}\ \zeta(\mathbf{s})$ are independent Gaussian processes, and since the assumptions of the model imply that process of threshold exceedances is conditionally independent of the process of exceedance times, they essentially reduced the Hastings-Metropolis sampler to two separate simulations, one to estimate $\sigma(\mathbf{s})$ and $\xi(\mathbf{s})$ based on the exceedances, the other to estimate $\zeta(\mathbf{s})$ based on the exceedance times. For the means and covariance functions of the three spatial processes they considered linear regression functions and exponential covariance functions, similar to Casson-Coles, though the priors for the spatial parameters were taken uniform over finite intervals following the warning of [15] (see also [11]) against using improper priors for these models. Like Casson-Coles, they noted that the model was implictly assuming the exceedances at different space-time locations were indepndent given the latent processes; in this case, they argued that the temporal and spatial separation between observations was sufficient to make that a reasonable assumption. For the specific application considered in [36], the model for $\xi(\mathbf{s})$ was reduced to two values (one for mountains, the other for plains) after noting that a fully spatial $\xi(\mathbf{s})$ did not improve on this.

The application discussed in [36] was to extreme precipitations in the Front Range region of Colorado. This refers to the foothills of the Rocky Mountains, where both elevation and weather conditions change very rapidly over short distances. They argued that a traditional stationary spatial statistics model based on latitude and longitude may not adequately reflect the topograpy of the region, and therefore defined the spatial model instead in terms of *climate coordinates*. Specifically, spatial locations were defined by their elevation and their mean summer precipitation (MSP), defined by averaging over the months April through October. When the resulting spatial fields are translated back into latitude-longitude coordinates for mapping purposes, they produce very sharply defined images on which the influence of the mountains is clearly seen.

(b) The first paper by Sang and Gelfand [207] presented itself as a spatio-temporal extension of the model of [36]. The paper was motivated by the study of precipitation extremes in the Cape Floristic Region of South Africa (CFR). The region was divided into grid cells of size 10 km.^{2} and annual maxima of daily precipitation calculated for each of the 50 years 1950–1999. Specifically, $Y_{i,t}$ denotes the annual maximum for year t at the ith location. It is assumed that the distribution of $Y_{i,t}$ is of GEV form (5.1) with parameters $\mu_{i,t}, \psi_{i,t}, \xi_{i,t}$; however, noting the difficulties in estimating ξ that had been mentioned by earlier authors including [36], they treated $\xi_{i,t}$ as constant (not varying witih i or t). Exploratory analysis suggested that both $\mu_{i,t}$ and $\psi_{i,t}$ were highly correlated in space, and in addition, that $\mu_{i,t}$ showed a linear time trend, so they sought a statistical model that would encompass those features.
The authors therefore sought a hierarchical model for the parameters $\mu_{i,t}, \psi_{i,t}, \xi_{i,t}$ where they initially set $\psi_{i,t} = \psi_i$ and $\xi_{i,t} = \xi$. For $\mu_{i,t}$, they wrote

$$\mu_{i,t} \mid \beta, W_{i,t}, \tau^2 \quad \sim \quad \mathcal{N}(\mathbf{X}_i^T \beta + W_{i,t}, \tau^2),$$

where $\mathbf{X}_i$ is a set of fixed spatial covariates, β and τ^2 are respectively a vector of spatial regression parameters and a nugget variance, and for $W_{i,t}$, they proposed four models:

$$\begin{aligned}
\text{Model A} \quad &: \quad W_{i,t} = \psi_i + \delta_t,\ \delta_t = \phi\,\delta_{t-1} + w_t,\ w_t \sim \mathcal{N}(0, W_0^2),\ \text{(IID)} \\
\text{Model B} \quad &: \quad W_{i,t} = \psi_i + \rho(t - t_0), \\
\text{Model C} \quad &: \quad W_{i,t} = \psi_i + (\rho + \rho_i)(t - t_0), \\
\text{Model D} \quad &: \quad W_{i,t} = \psi_i \delta_t,\ \delta_t = \phi\,\delta_{t-1} + w_t,\ w_t \sim \mathcal{N}(0, W_0^2)\ \text{(IID)}.
\end{aligned}$$

In Model A or D, $\{\delta_t\}$ is a time series of AR(1) structure, independent of all the spatial processes. To complete the specification of the model, it is necessary to specify joint distributions of the processes $\psi_i, \log \psi_i$ and, in the case of Model C, ρ_i. These are all continuous random variables with range $(-\infty, \infty)$, so a joint Gaussian process is appropriate. Sang and Gelfand used the same concept of coregionalization as was used in Section 5.1.2 (recall (5.27)) to reduce the model to two or three independent Gaussian processes, and for those, noting the lattice structure of the data, they assumed independent CAR models of the same stucture as (5.18).
The analysis of this model consists of sequentially updating all the unknown parameters by a Markov chain Monte Carlo (MCMC) procedure; the full algorithm is described in detail in an appendix to the paper [207]. In the paper, they fitted the model to the data from 1950-1998, holding out the last year of data for validation purposes. They fitted all of models A–D, noting that Model A fitted best when assessed by DIC (a common model selection criterion for Bayesian hierarchical models), but Model C performed better on the validation exercise and was preferred overall. They then computed the return values over the region for any year t using the formula (5.2), and noted how these had changed over time. In particular, they noted that extreme rainfalls had become more frequent in some

parts of the region but less frequent in others. An advantage of the fully Bayesian formulation of the problem is that since the MCMC procedure leads to full posterior distributions for all the unknown, it is possible to place uncertainty bounds on each of the estimates of return values and their trends.

(c) The second paper by Sang and Gelfand [208] was the first example using the latent process approach that did not assume that the extreme values at each site were conditionally independent given the latent process. In the introduction to their paper they noted that the hierarchical model of [207] (and, by extension, all earlier examples of the latent process approach) would produce discontinous predicted surfaces for the main variables of interest, i.e. the annual maxima at each location. That would not necessarily be of concern if the main focus was on return values rather than prediction itself, but we have also noted the broader issue that estimation of the spatial dependence model may be biased if the conditional covariance matrix (W in (5.23) is misspecified, so from that point of view as well, it is desirable to have a model than incorporates such dependence. Similarly to [207], they assume $Y(\mathbf{s})$ is the annual maximum at site $\mathbf{s}$, and that its conditional distribution given spatial processes $\mu(\mathbf{s}), \psi(\mathbf{s}), \xi(\mathbf{s})$ is of GEV structure (5.1), but unlike [207], they do not assume that these distributions are conditional independent for each $\mathbf{s}$. Instead, they define

$$Z(\mathbf{s}) \quad = \quad \left(1+\xi(\mathbf{s})\frac{Y(\mathbf{s})-\mu(\mathbf{s})}{\psi(\mathbf{s})}\right)^{1/\xi(\mathbf{s})}$$

which has unit Fréchet conditional margins, i.e. $G(z) = \Pr\{Z(\mathbf{s}) \le z \mid \mu(\mathbf{s}), \psi(\mathbf{s}), \xi(\mathbf{s})\} = e^{-1/z},\ z \ge 0$. In addition to defining the processes $\mu(\mathbf{s}), \psi(\mathbf{s}), \xi(\mathbf{s})$, the specification of the model needs to consider the joint distributions of $Z(\mathbf{s})$, as $\mathbf{s}$ ranges over the domain $\mathscr{S}$.

To define such a process, they follow the *copula approach*, which means that the process $Z(\mathbf{s})$ is defined through marginal transformations of some random process with prescribed marginal distributions, which these authors took to be standard normal. Specifically, they defined

$$Z(\mathbf{s}) \quad = \quad G^{-1}\Phi(Z^*(\mathbf{s}))$$

where G^{-1} is the inverse of the Fréchet distribution function G, Φ is the standard normal distribution function, and $Z^*(\mathbf{s}),\ \mathbf{s} \in \mathscr{S}$ is some random process on $\mathscr{S}$ with standard normal marginal distributions. It is not actually necessary, for such a representation to be valid, that the joint distributions of Z^* be multivariate normal, but of course that is the most natural assumption to adopt, and followed by [208]. From the point of view of multivariate extreme value theory, with its extensive considerations of distributions that are either asymptotically dependent or asymptotically independent, this assumption falls in the asymptotically independent class; in particular, the process is not max-stable, a point of view that will be fully developed in Section 5.2. The authors [208] noted this distinction but commented that it was beyond the scope of the current work to develop this point further; indeed, noting subsequent developments that will be further explored in

Sections 5.2 and 5.4, a max-stable model is plausibly a viable alternative to the Sang–Gelfand model. For the present discussion, we continue to follow the paper [208].

Continuing this line of thinking, Sang and Gelfand assumed the process $Z(\mathbf{s}),\ \mathbf{s} \in \mathscr{S}$ to be a Gaussian process with standard normal margins and a correlation function of the form $\rho(\mathbf{s},\mathbf{s}';\boldsymbol{\theta})$ depending on unknown parameters $\boldsymbol{\theta}$. Although any non-negative definite correlation function ρ could be assumed, they advocated using a stationary Matérn process as defined by (5.13).

The rest of the model specification requires again specifying a joint spatial distribution for the process $\mu(\mathbf{s}), \psi(\mathbf{s}), \xi(\mathbf{s})$. Although they noted and briefly discussed that a fully spatio-temporal representation is possible, they also noted that estimating such a model through a standard MCMC approach would be highly computationally intensive. Instead, they discussed fitting the model to one year's data at a time, for example through the representation

$$Y(\mathbf{s}) \quad = \quad \mathbf{X}(\mathbf{s})^T\beta + W(\mathbf{s}) + \frac{\psi}{\xi}\left(Z(\mathbf{s})^{\xi} - 1\right)$$

where $\mathbf{X}(\mathbf{s})$ is a vector of spatial covariates at $\mathbf{s}$, β are fixed regression coefficients, ψ and ξ are treated as constants (independent of $\mathbf{s}$), and $W(\mathbf{s})$ is a second (independent of Z^*) Gaussian spatial process with Matérn or some other standard spatial covariance function. Once again, the proposed method of estimation uses an MCMC algorithm.

(d) Another development about the same time was [85] that built on the work of [124]. Again motivated by the problem of rainfall extremes, Ghosh and Mallick [85] considered a model of the form

$$Y_{i,t} \quad \sim \quad \mathrm{GEV}(\mathbf{X}_i^T\beta_t, \psi, \xi)$$

where $Y_{i,t}$ is the annual maximum in year t and location i, $\mathbf{X}_i$ is a vector of spatial covariates for location i, and β_t, ψ, ξ are parameters. In this case, spatial dependence is induced, not through a spatial process for the GEV parameters, but directly in the conditional distributions: specifically, it is assumed that the joint density of $\{Y_{i,t},\ i = 1,...,n\}$ given β, ψ, ξ, for n locations in year t, is of "t-copula" form:

$$t_{\Sigma,k}\left(T_k^{-1}(F^{\mathrm{GEV}}(Y_{1,t};\mathbf{X}_1^T\beta_t,\psi,\xi)),\ldots,T_k^{-1}(F^{\mathrm{GEV}}(Y_{n,t};\mathbf{X}_n^T\beta_t,\psi,\xi))\right)\cdot\prod_{i=1}^{n}\frac{f^{\mathrm{GEV}}(Y_{i,t};\mathbf{X}_i^T\beta_t,\psi,\xi)}{t_k\left(T_k^{-1}(F^{\mathrm{GEV}}(Y_{i,t};\mathbf{X}_i^T\beta_t,\psi,\xi))\right)}.$$

Here F^{GEV} and f^{GEV} are respectively the distribution function and density of the three-parameter GEV distribution; T_k and t_k are respectively the distribution function and density of a univariate t distribution with k degrees of freedom; and $t_{\Sigma,k}$ is the density function of a multivariate t distribution with k degrees of freedom generated by the covariance matrix Σ, which in this case may be the covariance matrix of an arbitrary spatial process over the n sampling locations. In words, this is a copula model based on an n-dimensional t distribution rather than multivariate normal; in other respects, it is similar to the model of [208]; it differs from

[208] by not having an additional spatial process for the dependence of the GEV parameters; however, their model did allow for the p-dimensional time-dependent processes $\beta_t = \begin{pmatrix} \beta_{1,t} & \dots \beta_{p,t} \end{pmatrix}^T$ to be dependent through models of the form

$$\beta_{j,t} \quad = \quad \alpha_j \beta_{j,t-1} + \varepsilon_t,\ \varepsilon_t \sim \mathcal{N}(0, \sigma_j^2),\ j = 1, ..., p$$

which extends the time-dependent model of [124]. As in the other Bayesian models reviewed in this section, the model is completed by specifying priors for the additional unknown parameters, and fitted by MCMC. In addition to an annual maximum analysis fitted directly by the GEV distribution, they also proposed a threshold exceedance version where the generalized Pareto distribution (GPD) is fitted to the exceedances over a threshold at each site; in this case, the conditional distributions of the annual maxima are estimated by fitting the GPD but the copula model is applied to the annual maxima.

Comparing the three Bayesian papers of [207, 208, 85], it seems obvious that different components of these models could be mixed in different ways. For example, the multivariate t_k copula of [85] could also have been applied to the model of [208], and in the sense that the limit of t_k as $k \to \infty$ is multivariate normal, is a strict generalization. Similarly, the main example of [208] used spatial model fitting without a temporal component, but either of the approaches of [207] or [85] could have been used to add a temporal component, or one could exploit recent developments in the general theory of spatial or spatio-temporal statistics [42, 134] to construct still more general spatio-temporal models for the GEV model parameters. The principal limitation of such constructions is computational: the required MCMC computations to fit a full spatio-temporal model to the latent process, combined with a copula model for the annual maxima or peaks over a threshold, are considerable; but it is to be hoped that with more advanced computer hardware and with new algorithms for Bayesian statistics, such as those exploting Langevin or Hamiltonian dynamics [86], this will not be viewed as such an obstacle in the future.

(e) Tye and Cooley (2015) [246] was another paper about the Colorado Front Range, this one motivated by the 2013 extreme flood in the region around Boulder. They used data from 71 weather stations in the Front Range, as well as gridded monthly precipitation values from the Performance Reporting Information System (PRISM). From this dataset, they calculated summer (April–October) mean rainfalls over a 4 km. grid, which together with elevation, allowed them again to transform to "climate space", similar to [36]. In this case, they used annual maximum daily precipitation at each station and fitted generalized extreme value (GEV) values $\mu(\mathbf{s}_i), \psi(\mathbf{s}_i), \xi(\mathbf{s}_i)$ at each station $\mathbf{s}_i,\ i = 1, ..., n$, subsequently fitting the process $\mathbf{Z} = \begin{pmatrix} \mu & \log \psi & \xi \end{pmatrix}^T$ as a three-dimensional Gaussian process with dependence structure proposed by [249, 76], the same as in [204]. The fitting process in this case was not Bayesian, but essentially a two-stage hierarchical model as previously described in Section 5.1.1.4. Like the earlier paper by [27] and [36], they assumed conditional independence among sites given the latent processes, though they did check this assumption empirically. For the specific question of

estimating the probability of a 2013-like extreme, they showed substantial differences between the spatial approach and the most obvious alternative, which would be to fit the GEV at each site without any spatial smoothing; such "borrowing strength" of data from one site to another is an important selling point of this whole approach, especially when trying to model the probability of an extreme event for which direct data are necessarily limited. They also demonstrated the difference between model fits when the 2013 data were included or not, though they argued that the resulting changes in estimating extreme event probabilities are not particularly large when taking into account the uncertainties of those extreme event probabilities as characterized by their standard errors. This point is important because it illustrates the robustness of extreme probability estimations when outlying data are added.

(f) Russell et al. (2016) [205] proposed an extension to the spatial latent process technique that combined this method with bivariate extreme value theory, based on the kinds of models introduced in Chapter 4. The bivariate extreme value technique that they use was developed in an earlier paper by Russell and co-authors [206], but since its purpose is incidental to the present chapter, we content ourselves with a brief summary here.

A common formulation of bivariate extreme value theory assumes bivariate random vectors $\mathbf{Z} = (Z_1,\ Z_2) \in [0,\infty)^2$ which is transformed to $R = ||\mathbf{Z}||,\ \mathbf{W} = \mathbf{Z}/||\mathbf{Z}||$ where $||\cdot||$ denotes any standard norm (subsequently taken to be the L_1 norm, $||\mathbf{Z}|| = Z_1 + Z_2$). The distribution is said to be (bivariate) regularly varying if

$$n\Pr\left\{\frac{R}{b(n)} > r,\ \mathbf{W} \in B\right\} \quad \stackrel{v}{\to} \quad r^{-\alpha}H(B) \text{ as } n \to \infty,$$

where $\alpha > 0,\ b(n) > 0,\ H$ is a finite measure on the Borel sets of the unit sphere $S = \{\mathbf{z} :\ ||\mathbf{z}|| = 1\}$ and convergence is vague over the set $[0,\infty)^2 \setminus \{\mathbf{0}\}$. For the specific models assumed in [205], they assumed the marginal distributions of Z_1 and Z_2 have been previously transformed to unit Fréchet, so there is no loss of generality by writing $b(n) =$, and $\alpha = 1$.

As is well known, the measure H represents the strength of extremal dependence between the two components, the extreme cases being complete dependence, when H is concentrated on $\left(\frac{1}{2},\frac{1}{2}\right)$, and independence, when H puts measure $\frac{1}{2}$ on each of the extreme points $(0,1)$ and $(1,0)$. [206, 205] reduced that to a metric

$$\gamma \quad = \quad \int_{[0,1]} |2w-1|H(dw)$$

where $\gamma \in [0,1]$, the extreme cases $\gamma = 1$ and $\gamma = 0$ corresponding to complete dependence and independence. Given sample values $(Z_{i,1}, Z_{i,2}),\ i = 1,...,n$ transformed to $R_i = ||\mathbf{Z}_i||,\ W_i = Z_{i,1}/||\mathbf{Z}_i||$, they proposed the estimator

$$\hat{\gamma} \quad = \quad \frac{\sum_{i=1}^n \delta(R_i)|2W_i - 1|}{\sum_{i=1}^n \delta(R_i)} \tag{5.29}$$

where $\delta(R)$ is some non-decreasing function of R (designed to give greatest weight to the most extreme values). In particular, a common threshold-based choice would be $\delta(r) = 1$ if $r > r_0$, 0 otherwise, which would correspond to including only those values for which R exceeds the threshold r_0. The paper [205] discusses this choice, but ultimately favors a continuously increasing δ (based on the standard normal distribution function) to simplify the optimization algorithm which is described next.

The concept of the paper was to study how different meteorological covariates affect ground-level ozone. If the meteorological covariates at site $\mathbf{s}_i$ and time t are written $\mathbf{X}_{it} = \begin{pmatrix} x_{it1} & \dots & x_{itp} \end{pmatrix}$ and the ozone level at time t is Y, the objective, as in a standard regression problem, was to find, for each i, linear combinations $\sum_{j=1}^{p} x_{itj}\beta_{ij}$ that were highly correlated with $Y_{it},\ t = 1,...,T$. However, the adverse health effects of ozone are largest when the ozone level is high; therefore, instead of a conventional correlation coefficient, they sought parameter vector $\beta_i = \begin{pmatrix} \beta_{i1} & \dots & \beta_{ip} \end{pmatrix}$ to maximize γ. We omit the full details of their algorithm but the basic concept was:

(i) Transform the marginal distribution of $Y_{it},\ t = 1,...,T$ to unit Fréchet for each station $\mathbf{s}_i$ (by fitting a gamma distribution to values below a threshold, GPD above, and applying a probability integral transformation);

(ii) For any candidate set of regression coefficients β_i, transform the distribution of $\mathbf{X}_{it}^T\beta_i,\ t = 1,...,T$ to unit Fr'echet;

(iii) Estimate $\hat{\gamma}(\beta_i)$ by (5.29) applied to the transformed values of $\mathbf{X}_{it}^T\beta_i$ and Y_{it} as $t = 1,\dots,T$ for each i.

The optimal parameters β_i are found for each $\mathbf{s}_i$, along with a $p \times p$ estimated covariance matrix estimated by a pairwise bootstrap procedure.

Up to this point, this description of the method of [205] has focused on the definition of the parameters $\beta_i\beta(\mathbf{s}_i)$ and their estimation (including a covarance matrix for the errors) at each site $\mathbf{s}_i$. The rest of the paper proposed a method for spatial interpolation similar to the method described earlier in this section. The authors used six meteorological covariates and therefore needed to define a 6-dimensional Gaussian process $\beta(\mathbf{s})$. The structure of the model is similar to that in (5.26) and (5.27) where they again assumed exponential spatial covariance functions for the components of the process $\delta(\mathbf{s})$. The analysis was also similar to that of Section 5.1.2 but with a simpler specification for the error covariance matrix W: in the analysis of [205], the errors $e_j(\mathbf{s})$ are assumed to be independent at different sites. The analysis was applied to meteorological and ozone data from the United States Environmental Protection Agence (EPA) covering EPA regions 3 and 4, which covers a regions of the south-east US stretching roughly from Mississippi to Pennsylvania.

The outcome of this process was a field of interpolated values for the regression parameters $\beta(\mathbf{s})$ that, in turn, allowed for some qualitative statements about which meteorological variables most influence extreme ozone. For example, the authors concluded that temperature was more important in the northern part of the region, while relative humidity and a variable they called turbulent kinetic energy are

more important in the southern part. These kinds of conclusions are important for designing an emissions control strategy to minimize the frequency of high-ozone events.

5.1.5 Summary

The latent process model has been applied in many problems for spatial distribution of extremes since 1999. It can be fitted using either Bayesian or non-Bayesian (maximum likelihood or REML) approaches. Compared with the naïve approach of fitting independent GEV distributions to each site, the model generates smooth surfaces for the GEV parameters or for quantities derived from them (in particular, return values). Thus, the model provides interpolated values between the observation sites, and because of the "borrowing strength" property of hierarchical Bayesian methods, it may be expected also to lead to improved estimates of the GEV parameters at the observation sites. The principal limitation of these methods as they have been currently applied is that, in many cases, they have assumed conditional independence of the site-specific extremes given the latent processes for the GEV parameters. However, there are ways around that restriction as has been shown both in Bayesian [208, 85] and non-Bayesian [204] analyses. It is to be expected that, in future work, the connection with max-stable processes, to which we turn next, will be developed more strongly; however, as these examples show, the fundamental ideas are not restricted to any specific class of spatial processes and could in principle be developed for other classes as well.

5.2 Max-Stable Processes

The discussion of Section 5.1 has shown the potential to contruct very rich hierarchical models combining spatio-temporal processes for the GEV or GPD model parameters with different approaches for modeling the conditional distributions of extremes given those model parameters, but there is a limitation: the only approaches up to this point that have allowed for conditional dependence have used copula models, either multivariate normal in [208] or multivariate t in [85]; however, we know from Chapter 4 that much more general models for multivariate extremes are available. In particular, there is a family of processes known as *max-stable processes*, directly generalizing the asymptotically dependent classes of multivariate extreme value distributions, and these have attracted much attention in recent years. In this section, we review the basic theory of max-stable processes; more recent developments are in Section 5.4. Our development follows to some extent a recent review by Davison, Huser and Thibaud [48] and also draws on earier review papers such as [49, 39].

5.2.1 Background on Poisson processes

The basic theory of max-stable processes relies rather heavily on some basic results about Poisson processes, so we begin by reviewing those. Although there are many books covering the basic theory, we shall draw particularly on [196].

Following the notation and terminology of [196], we consider point processes on a set $\mathbb{E}$ where $\mathbb{E}$ is a *nice space*, which technically means a locally compact topological space with a countable base; in practice, $\mathbb{E}$ is usually $\mathbb{R}^d$ or $\mathbb{R}^d_+$ or perhaps $\mathbb{R}^d_+ \setminus \{\mathbf{0}\}$ when we wish to exclude specifically the point at $\mathbf{0}$; here d is a finite dimension.

Let N be a point process on the space $\mathbb{E}$ whose BOrel sets are denoted $\mathscr{E}$. Technically, N is a mapping from a probability space to $M_p(\mathbb{E})$, the space of Radon point measures on $\mathbb{E}$. A point measure is a measure N on $\mathbb{E}$ that puts all its mass on a discrete set of points in $\mathbb{E}$; colloquially, for any set $A \in \mathscr{E}$, $N(A)$ counts the number of points in A. Radon means that $N(K) < \infty$ for any compact set K; in addition, all the point measures we shall consider will be *simple* in the sense that they do not have multiple points; thus $N(\{x\})$ has to be either 0 or 1 for any subset $\{x\}$ consisting of a single point x.

Definition 1. N is a *Poisson process with mean measure* μ, also known as a *Poisson random measure with mean measure* μ (PRM(μ)) if the following properties hold:

(a) For $A \in \mathscr{E}$ and any $k = 0, 1, 2, \ldots$,

$$\Pr\{N(A) = k\} = \begin{cases} \frac{e^{-\mu(A)}(\mu(A))^k}{k!} & \text{if } \mu(A) < \infty, \\ 0 & \text{if } \mu(A) = \infty. \end{cases}$$

(b) If $A_1, \ldots, A_k$ are disjoint subsets of $\mathbb{E}$ in $\mathscr{E}$, then $N(A_1), \ldots, N(A_k)$ are independent random variables.

A particularly important class of Poisson processes is the *homogeneous Poisson process with rate 1 on* $[0,\infty)$: this corresponds to the case where $\mathbb{E} = [0,\infty)$ and μ is Lebesgue measure; in particular $\mu((a,b)) = b - a$ for any $0 \le a \le b < \infty$. There is a standard representation of this process in terms of exponential random variables (Proposition 5.1 of [196]):

Proposition 1. Let $\{E_j,\ j \ge 1\}$ be independent random variables with a standard exponential distribution ($\Pr\{E_j \le x = 1 - e^{-x}$ for any $x \in (0,\infty)$). Let N be the point process whose nth point is at $\sum_{j=1}^n E_j,\ n = 1, 2, \ldots$. Then N is a homogeneous Poisson process with rate 1 on $[0,\infty)$,

In the context of max-stable processes, a particularly important role is played by the process whose points are inverses (reciprocals) of this process, in other words, the process whose nth largest point is at $\frac{1}{\sum_{j=1}^n E_j}$ in the notation of Proposition 1. With slight abuse of notation, we shall call this the *inverse Poisson process*:

Definition 2. N is an inverse Poisson process (IPP) on $(0,\infty)$ if it is a Poisson process with measure ν defined by

$$\nu((x,\infty)) \quad = \quad \frac{1}{x} \text{ for any } x > 0.$$

The next topic to discuss is *mappings of Poisson processes*. Suppose there is a function $T : \mathbb{E} \to \mathbb{E}'$ where $\mathbb{E}'$ is some other nice space. Define the inverse map by the property

$$T^{-1}(A') \quad = \quad \left\{e \in \mathbb{E} :\ T(e) \in A'\right\} \text{ for any } A' \in \mathbb{E}'.$$

The following result is Proposition 5.2 of [196]:

Proposition 2. Suppose $T : \mathbb{E} \to \mathbb{E}'$ is a measurable mapping of $\mathbb{E}$ to $\mathbb{E}'$ such that if K' is a compact set in $\mathbb{E}'$ then $T^{-1}K'$ is a compact set in $\mathbb{E}$. If N is a PRM(μ) on $\mathbb{E}$, $N' = N \circ T^{-1}$ is a PRM(μ') on $\mathbb{E}'$, where $\mu' = \mu \circ T^{-1}$.

In words: if each point x in the process N is transformed into a point $T(x)$ in the process N', then N' is also a Poisson process and $\mu'(A') = \mu(T^{-1}(A'))$ for each $A' \subset \mathbb{E}'$.

As a concrete example, we prove the following:

Proposition 3. Suppose N is an IPP on $(0,\infty)$ with points $\{R_i,\ i = 1,2,\ldots\}$. Also let $\{W_i,\ i = 1,2,\ldots\}$ be a sequence of IID random variables on $(0,\infty)$ with mean 1, also assumed to be independent of the process $\{R_i,\ i = 1,2,\ldots\}$. Then the process

$$N' \quad = \quad \{R_i W_i,\ i = 1,2,\ldots\}$$

is also an IPP.

Proof. First expand the definition of N to be a two-dimensional Poisson process on $(0,\infty)^2$ with points $\{(R_i,W_i)\}$ and with

$$\mu((x,\infty) \times A) \quad = \quad \frac{F(A)}{x}$$

for any measurable set $A \subset (0,\infty)$, where F is the probability measure of the random variables W_i. Since the map

$$T((R_i,W_i)) \quad = \quad R_i W_i$$

transforms $(0,\infty)^2$ into $(0,\infty)$ and is measurable, Proposition 2 applies that for any $z > 0$,

$$\begin{aligned}
\mu'((z,\infty) &= \mu(\{(x,w) :\ xw > z\}) \\
&= \int_{(0,\infty)} F(dw) \cdot \int_{(0,\infty)} I\left(x > \frac{z}{w}\right) \nu(dx) \\
&= \int_{(0,\infty)} F(dw) \cdot \frac{w}{z} \\
&= \frac{1}{z}
\end{aligned}$$

since $\int_{(0,\infty)} wF(dw) = 1$ by the assumption that each W_i has mean 1. Thus, the measure μ' is the same as ν, the measure of an IPP, so N' is also an IPP.

Our final (elementary) result about IPPs is the following:

Proposition 4. Suppose N is an IPP with points $\{R_i,\ i = 1,2,\ldots\}$. Also let $M = \max_{i=1,2,\ldots} R_i$. Then M has a unit Fréchet distribution.

Proof. The statement $M < z$ is equivalent to $N((z,\infty)) = 0$; this has probability $e^{-\mu((z,\infty))} = e^{-1/z}$ which is the distribution function of a unit Fréchet random variable.

Although the full point process result is of interest, the most important conclusion to come out of this is: if $\{R_i\}$ is an IPP and $\{W_i\}$ is an independent sequence of IID random variables with mean 1, then $\max_i R_i W_i$ has a unit Fréchet distribution. Next, we shall see how this result may be generalized in a stochastic process context.

5.2.2 Constructing a max-stable process

In this section, we describe an extension to the model of Section 5.2.1 that defines a stochastic process $Z(x)$, where x ranges over some space $\mathscr{X}$, whose marginal distributions are unit Fréchet, in other words, $\Pr\{Z(x) \leq z\} = e^{-1/z}$ for any $z > 0$, at each point $x \in \mathscr{X}$. It also has the property of being *max-stable* in a sense we shall make precise. If $\mathscr{X}$ is a finite set with k elements, the result is a k-dimensional multivariate extreme value distribution in the same sense as in Chapter 4. However the main cases in which we are interested are for $\mathscr{X}$ a subset of $\mathbb{R}^d$ for some $d \geq 1$, which would allow for stochastic processes in time or space or both.

For this construction, we assume $\{W_i(x),\ x \in \mathscr{X}\}$ to be IID stochastic processes on the space $\mathscr{X}$, with $W_i \geq 0$ for all x, and common mean 1: $\mathrm{E}\{W_i(x)\} = 1$ for all $x \in \mathscr{X}$. Technically, we assume that sample paths $W_i(x),\ x \in \mathscr{X}$ lie in some "nice space" $\mathbb{E}$, where $\mathbb{E}$ could be, for example, the space of continuous functions over $\mathscr{X}$. However, specific properties, such as continuity, are not a requirement of the construction. We may assume the existence of a probability space $(\mathbb{E}, \mathscr{E}, \mathscr{P})$, where $\mathscr{E}$ is a space of Borel sets on $\mathbb{E}$ and $\mathscr{P}$ is a probability measure. In this space, consider a measure μ on $(0,\infty] \times \mathbb{E}$ with the property that

$$\mu((a,b) \times A) \quad = \quad \left(\frac{1}{a} - \frac{1}{b}\right) \cdot \mathscr{P}(A) \tag{5.30}$$

whenever $0 < a < b \leq \infty$ and $A \in \mathscr{E}$. Such a measure is Radon and therefore, by the definitions of Section 5.2.1, defines a Poisson process over $(0,\infty] \times \mathbb{E}$. We write the points of this Poisson process as (R_i, W_i) where $R_i \in (0,\infty)$, $W_i \in \mathbb{E}$. Alternatively, we may write explicitly $W_i(x),\ x \in \mathscr{X}$ to remind ourselves that W_i is a stochastic process on the space $\mathscr{X}$.

Now define

$$Z(x) \quad = \quad \max_i R_i W_i(x),\ x \in \mathscr{X}.$$

Proposition 4 guarantees that $Z(x)$ is finite for all x, and indeed, has a unit Fréchet marginal distribution. One consequence of this is that if $Z_1(\cdot), \ldots, Z_n(\cdot)$ are IID copies of the process $Z(\cdot)$, then for any fixed x, $\frac{1}{n}\max(Z_1(x), \ldots, Z_n(x))$ has the same distribution as $Z(x)$. In other words, $Z(x)$ is *max-stable* for each fixed $x \in \mathscr{X}$. The purpose of this section is to show that something much stronger: this property of max-stability, in a sense that we shall define, holds for the whole process $\{Z(x),\ x \in \mathscr{X}\}$ and not just for a fixed value of x.

Consider the event

$$\{Z(x) \leq z(x) \text{ for all } x \in \mathscr{D}\}, \tag{5.31}$$

where $\mathscr{D} \subset \mathscr{X}$ and $z(x),\ x \in \mathscr{X}$ is a function from $\mathscr{X}$ to $(0,\infty]$. We allow infinite values for the following reason: if we define $z(x) = +\infty$ on $\mathscr{X} \setminus \mathscr{D}$, and replace $\mathscr{D}$ by $\mathscr{X}$ in (5.31), the truth or falsity of (5.31) is unchanged. It simplifies the notation to allow this case and always take $\mathscr{D} = \mathscr{X}$ in (5.31). The conditions on $z(\cdot)$ are very mild — essentially, we want to ensure that $\inf_{x \in \mathscr{X}} \frac{w(x)}{z(x)} \in \mathscr{E}$ whenever $w \in \mathbb{E}$ —

but for most purposes it would suffice simply to require that $z(x)$ be a measurable function of $x \in \mathscr{X}$.

With these definitions, the event (5.31) occurs if and only if $R_i W_i(X) \leq z(x)$ for all x, and in terms of the point process $\{R_i, W_i(\cdot)\}$, that is equivalent to saying that the set

$$B = \left\{ (r,w) : \ r > \inf_x \frac{z(x)}{w(x)} \right\} \ \in (0,\infty) \times \mathbb{E}$$

is empty. That probability is $e^{-\mu(B)}$, where μ is the measure defined by (5.30). Noting that the first component of μ has density $\frac{1}{r^2}$ on $0 < r \leq \infty$, we calculate

$$\begin{aligned} \mu(B) &= \int_{(0,\infty]} \frac{dr}{r^2} \int_{\mathbb{E}} d\mathscr{P} \cdot I\left\{ r > \inf_x \frac{z(x)}{w(x)} \right\} \\ &= \int_{\mathbb{E}} d\mathscr{P} \sup_x \frac{w(x)}{z(x)} \\ &= \mathrm{E}\left\{ \sup_{x \in \mathscr{X}} \frac{W(x)}{z(x)} \right\} \end{aligned}$$

where $W(x)$ is any of the IID processes $\{W_i(x)\}$ and the symbol E denotes expectation.

We therefore derive the formula

$$\Pr\{Z(x) \leq z(x) \text{ for all } x \in \mathscr{D}\} \quad = \quad \exp\{-V(z(x),\ x \in \mathscr{X})\} \tag{5.32}$$

where

$$V(z(x),\ x \in \mathscr{X}) \quad = \quad \mathrm{E}\left\{ \sup_{x \in \mathscr{X}} \frac{W(x)}{z(x)} \right\} \tag{5.33}$$

together with the convention, noted previously, that we define $z(x) = +\infty$ when $x \notin \mathscr{D}$.

Formulas (5.32) and (5.33) define our basic computational formula for max-stable processes. We note several consequences:

(a) If $\mathscr{D}$ is a one-point set, say $\mathscr{D} = \{x\}$, then

$$V(z(x),\ x \in \mathscr{X}) \quad = \quad \mathrm{E}\left\{ \frac{W(x)}{z(x)} \right\} \ = \ \frac{1}{z(x)}$$

confirming that all the marginal distributions of $Z(x)$ are unit Fréchet.

(b) If $a > 0$, then

$$\begin{aligned} V(az(x),\ x \in \mathscr{X}) &= \mathrm{E}\left\{ \sup_{x \in \mathscr{X}} \frac{W(x)}{az(x)} \right\} \\ &= \frac{1}{a} V(z(x),\ x \in \mathscr{X}) \end{aligned}$$

so the function V is *homogeneous of order* –1. This mimics the corresponding property of the finite-dimensional V used in Chapter 4.

(c) Suppose $Z_1(\cdot), \ldots, Z_n(\cdot)$ are IID copies of the process $Z(\cdot)$. Then

$$\begin{aligned} \Pr\left\{Z_j(x) \le nz(x) \text{ for all } x \in \mathscr{D},\ j = 1, \ldots, n\right\} &= \Pr^n \{Z(x) \le nz(x) \text{ for all } x \in \mathscr{D}\} \\ &= \exp\{-nV(nz(x),\ x \in \mathscr{X})\} \\ &= \exp\{-V(z(x),\ x \in \mathscr{X})\} \quad (5.34) \end{aligned}$$

which confirms the fundamental max-stability property, that $\frac{1}{n}\max\{Z_1(x), \ldots, Z_n(x)\}$, $x \in \mathscr{D}$ is identical in distribution as $Z(x),\ x \in \mathscr{D}$ (in particular, all finite-dimensional distributions are the same).

(d) Suppose the function $z(x)$ is a constant z on $x \in \mathscr{D} \subset \mathscr{X}$. Then

$$\Pr\{Z(x) \le z \text{ for all } x \in \mathscr{D}\} = \exp\left(-\frac{\theta_{\mathscr{D}}}{z}\right) \quad (5.35)$$

where $\theta_{\mathscr{D}} = \mathrm{E}\left\{\max_{x \in \mathscr{D}} W(x)\right\} > 0$. This is known as the *extremal coefficient*, which is analogous to the extremal index in the theory of extremes in stochastic sequences. For a finite set, say $|\mathscr{D}| = k$, we have $1 \le \theta_{\mathscr{D}} \le k$. [Proof: If $\mathscr{D} = \{x_1, \ldots, x_k\}$, then $\mathrm{E}\{W(x_1)\} \le \mathrm{E}\left\{\max_{1 \le j \le k} W(x_j)\right\} \le \sum_{j=1}^k \mathrm{E}W(x_j)$.]

5.3 Probability Calculations for Max-Stable Processes

5.3.1 Brown-Resnick Process

This name is given to a process where $W(x) = \exp\{\varepsilon(x) - \sigma^2(x)/2\}$ where $\{\varepsilon(x),\ x \in \mathscr{X}\}$ is a zero-mean Gaussian process with variance function $\mathrm{Var}\{\varepsilon(x)\} = \sigma^2(x)$. The most obvious example is when ε is a stationary isotropic process, for which $\sigma^2(x)$ is a constant and $\mathrm{Cor}(\varepsilon(x), \varepsilon(x')) = \rho(||x - x'||)$. However, for a reason to be explained below, this is very often not a good choice and it is better to make ε an *intrinsically stationary* process for which $\mathrm{Var}\{\varepsilon(x) - \varepsilon(x')\} = \gamma(||x - x'||)$ with some scalar function $\gamma(\cdot)$ for which $\gamma(t) \to \infty$ as $t \to \infty$. A common choice (valid in any dimension) is $\gamma(t) = c_0 + c_1 t^\lambda$ where $c_0 \ge 0,\ c_1 \ge 0,\ 0 \le \lambda < 2$; the special case of dimension 1 where $c_0 = 0,\ \lambda = 1$ is known as Brownian motion (a.k.a. the Wiener process).

We define a max-stable process $Z(x) = \max_{i \ge 1} R_i W_i(x)$ where $\{R_i,\ i \ge 1\}$ is an inverse Poisson process and $\{W_i(\cdot),\ i \ge 1\}$ are IID copies of the process $W(\cdot)$; then at two locations $x = x_1, x_2$ and associated random variables $W^{(j)} = W(x_j),\ j = 1, 2$ we have

$$\Pr\{Z(x_1) \le z_1, Z(x_2) \le z_2\} = \exp\{-V(z_1, z_2)\} \quad (5.36)$$

for any $z_1 > 0,\ z_2 > 0$, where $V(z_1, z_2)$ is defined by (5.37) below. In this section, we establish the explicit formula (5.39).

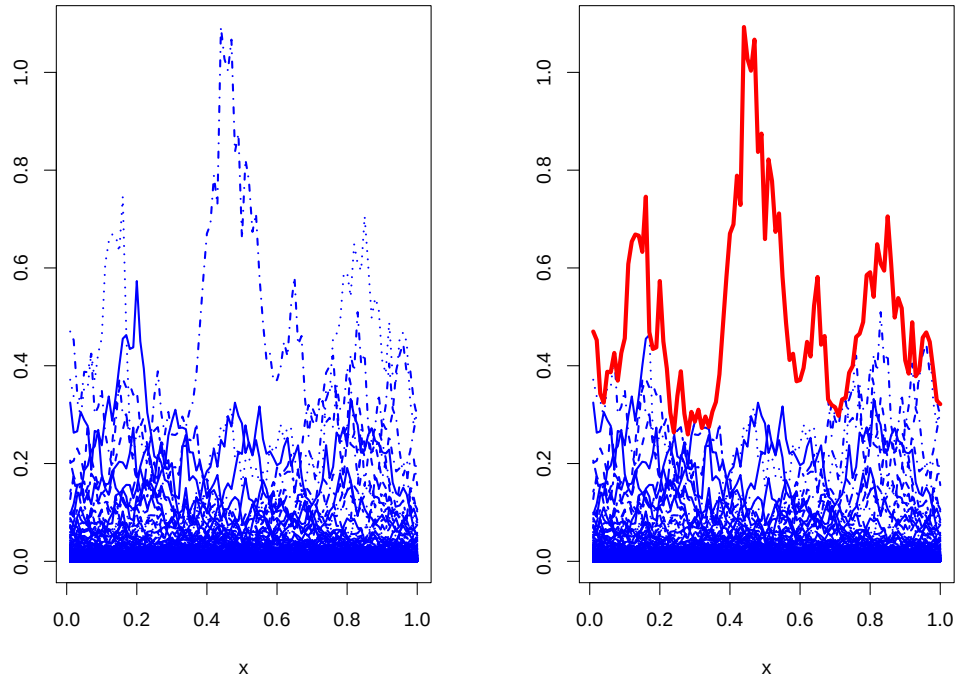

Figure 5.2 *Left hand plot: superimposed processes $R_iW_i(x)$, where $\{R_i\}$ is an IPP and $W_i(x) = \exp\{\varepsilon_i(x) - 1/2\}$ for independent Gaussian process $\{\varepsilon(x)\}$. Different indices i are indicated by different line types. Right hand plot: same, with the pointwise maximum process $Z(x) = \max_i R_iW_i(x)$ superimposed in red. Adapted from a figure in [48].*

Given random variables $W^{(j)} = \exp(\varepsilon_j - \sigma_j^2/2)$ where the joint distribution of $(\varepsilon_1, \varepsilon_2)$ is bivariate normal with means 0, variances σ_1^2 and σ_2^2, and correlation ρ, then for any $z_1 > 0,\ z_2 > 0$, we want to calculate

$$V(z_1, z_2) \quad = \quad \mathrm{E}\left\{\max\left(\frac{W^{(1)}}{z_1}, \frac{W^{(2)}}{z_2}\right)\right\}. \tag{5.37}$$

Define $a^2 = \mathrm{Var}(\varepsilon_1 - \varepsilon_2) = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2$, and write

$$\begin{aligned} \varepsilon_1 &= b\varepsilon_3 + d\varepsilon_4, \\ \varepsilon_2 &= -c\varepsilon_3 + d\varepsilon_4, \end{aligned}$$

where $b = \frac{\sigma_1(\sigma_1 - \rho\sigma_2)}{a}$, $c = \frac{\sigma_2(\sigma_2 - \rho\sigma_1)}{a}$, $d = \frac{\sigma_1\sigma_2\sqrt{1-\rho^2}}{a}$, and $\varepsilon_3,\ \varepsilon_4$ are independent $N(0,1)$ variables. Also note that $\varepsilon_1 - \varepsilon_2 = a\varepsilon_3$.

We also note that

$$\frac{W^{(1)}}{z_1} > \frac{W^{(2)}}{z_2} \quad \Longleftrightarrow \quad \varepsilon_3 > \frac{1}{a}\log\frac{z_1}{z_2} + \frac{\sigma_1^2 - \sigma_2^2}{2}.$$

Therefore,

$$\begin{aligned} V(z_1,z_2) &= \frac{1}{z_1}\mathrm{E}\left\{\exp\left(b\varepsilon_3 + d\varepsilon_4 - \frac{\sigma_1^2}{2}\right) I\left(\varepsilon_3 > \frac{1}{a}\log\frac{z_1}{z_2} + \frac{\sigma_1^2-\sigma_2^2}{2a}\right)\right\} \\ &\quad + \frac{1}{z_2}\mathrm{E}\left\{\exp\left(-c\varepsilon_3 + d\varepsilon_4 - \frac{\sigma_2^2}{2}\right) I\left(\varepsilon_3 < \frac{1}{a}\log\frac{z_1}{z_2} + \frac{\sigma_1^2-\sigma_2^2}{2a}\right)\right\} \end{aligned} \tag{5.38}$$

We also note that if $\varepsilon \sim N[0,1]$, for any real t and c,

$$\mathrm{E}\left\{e^{t\varepsilon} I(\varepsilon < c)\right\} = e^{t^2/2}\Phi(c-t)$$

with $\Phi(\cdot)$ the cumulative distribution function of $N[0,1]$.

Each of the terms in (5.38) factorizes into separate terms that depend on the (independent) random variables ε_3 and ε_4, and we trivially have that

$$\mathrm{E}\left\{\exp\left(\sigma\sqrt{\frac{1+\rho}{2}}\varepsilon_4\right)\right\} = e^{d^2/2}.$$

For the ε_3 terms in (5.38), we have

$$\begin{aligned} \mathrm{E}\left\{e^{b\varepsilon_3} I\left(\varepsilon_3 > \frac{1}{a}\log\frac{z_1}{z_2} + \frac{\sigma_1^2-\sigma_2^2}{2a}\right)\right\} &= e^{b^2/2}\Phi\left(-\frac{1}{a}\log\frac{z_1}{z_2} - \frac{\sigma_1^2-\sigma_2^2}{2a} + b\right) = e^{b^2/2}\Phi\left(-\frac{1}{a}\log\frac{z_1}{z_2} + \frac{a}{2}\right), \\ \mathrm{E}\left\{e^{-c\varepsilon_3} I\left(\varepsilon_3 < \frac{1}{a}\log\frac{z_1}{z_2} + \frac{\sigma_1^2-\sigma_2^2}{2a}\right)\right\} &= e^{c^2/2}\Phi\left(\frac{1}{a}\log\frac{z_1}{z_2} + \frac{\sigma_1^2-\sigma_2^2}{2a} + c\right) = e^{c^2/2}\Phi\left(\frac{1}{a}\log\frac{z_1}{z_2} + \frac{a}{2}\right), \end{aligned}$$

since we can check that $b - \frac{\sigma_1^2-\sigma_2^2}{2a} = c + \frac{\sigma_1^2-\sigma_2^2}{2a} = \frac{a}{2}$. Hence from (5.38),

$$\begin{aligned} V(z_1,z_2) &= \frac{1}{z_1}\exp\left\{-\frac{\sigma_1^2}{2} + \frac{d^2}{2} + \frac{b^2}{2}\right\}\Phi\left(-\frac{1}{a}\log\frac{z_1}{z_2} + \frac{a}{2}\right) + \frac{1}{z_2}\exp\left\{-\frac{\sigma_2^2}{2} + \frac{d^2}{2} + \frac{c^2}{2}\right\}\Phi\left(\frac{1}{a}\log\frac{z_1}{z_2} + \frac{a}{2}\right) \\ &= \frac{1}{z_1}\Phi\left(\frac{1}{a}\log\frac{z_2}{z_1} + \frac{a}{2}\right) + \frac{1}{z_2}\Phi\left(\frac{1}{a}\log\frac{z_1}{z_2} + \frac{a}{2}\right) \end{aligned} \tag{5.39}$$

since $b^2 + d^2 = \sigma_1^2$ and $c^2 + d^2 = \sigma_2^2$.

Note that as $a \to \infty$, $V(z_1,z_2) \to \frac{1}{z_1} + \frac{1}{z_2}$, and as $a \to 0$, $V(z_1,z_2) \to \frac{1}{\min(z_1,z_2)}$, corresponding to the cases of independence and perfect dependence ($W^{(1)} = W^{(2)}$ with probability one) respectively. However, this shows the disadvantage of taking ε to be a stationary (rather than intrinsically stationary) process: in order for $Z(x_1)$ and $Z(x_2)$ to be asymptotically independent as $||x_1 - x_2|| \to \infty$, we need $\gamma(||x_1 - x_2||) \to \infty$, and this is not true in the strictly stationary case.

The formula (5.39) was given by [48]. Curiously, exactly the same formula arises in the so-called Smith process [226], though the method of calculation is completely different. However, unknown to the author at the time, the formula was in fact discovered in yet another context by Hüsler and Reiss [125].

5.3.2 Extremal t Process

This is given by [48] but attributed to the papers [166, 173, 241]. It corresponds to the process $W(x) = m_\alpha \varepsilon(x)_+^\alpha$ where $\alpha > 0$ and $\{\varepsilon(x)\}$ is a stationary Gaussian process with mean 0, common variance 1 and covariance function $\mathrm{Cov}\{\varepsilon(x_1), \varepsilon(x_2)\} = c_\alpha(x_1, x_2)$. Here $m_\alpha = \frac{\pi^{1/2} 2^{1-\alpha/2}}{\Gamma((\alpha+1)/2)}$ is the normalizing constant required to make $\mathrm{E}\{W(x)\} = 1$. According to [48], the bivariate joint distributions are derived from

$$V(z_1, z_2) = \frac{1}{z_1} T_{\alpha+1}\left(-\frac{c}{b} + \frac{1}{b}\left(\frac{z_2}{z_1}\right)^{1/\alpha}\right) + \frac{1}{z_2} T_{\alpha+1}\left(-\frac{c}{b} + \frac{1}{b}\left(\frac{z_1}{z_2}\right)^{1/\alpha}\right) \tag{5.40}$$

where $c = c_\alpha(x_1, x_2)$, $b^2 = \frac{1-c^2}{\alpha+1}$ and $T_{\alpha+1}(\cdot)$ is the cumulative distribution function of the $t_{\alpha+1}$ distribution. Under certain circumstances the limit $\alpha \to \infty$ is the Brown-Resnick process, so in that sense, the extremal t process is a generalization of the Brown-Resnick process. For finite α, the $T_{\alpha+1}$ expressions in (5.40) are bounded away from 1, so the independent case $V(z_1, z_2) = \frac{1}{z_1} + \frac{1}{z_2}$ cannot arise even as $||x_1 - x_2|| \to \infty$.

5.3.3 Smith Process

This name has been widely given to a process first explored in an unpublished paper [226]. It corresponds to $W(x) = \frac{f(x;Y)}{f_Y(Y)}$ where Y is a random variable on some space $\mathscr{Y}$ with density $f_Y(\cdot)$, and $f(x;y)$ is a family of densities with $\int_{\mathscr{Y}} f(x;y)dy = 1$ for each x. When $f(x;y)$ is a multivariate Gaussian density centered at y the bivariate joint distributions are again of functional form (5.39). Typical sample paths from this process consist of segments from the family of densities $f(x;y)$ and these are generally regarded as too smooth to represent real environmental extremes. Nevertheless the process has some appealing mathematical properties.

5.3.4 Schlather Process

This was the second specific example of a max-stable process to be proposed [211]. As noted by [48], it corresponds to the $\alpha = 1$ special case of the extremal t process. It has the disadvantage that the dependence between two spatial locations remains positive even as distance tends to infinity, and though some modifications of the process have been propsoed to deal with that, the Brown-Resinick and extremal t processes are probably better suited to practical applications.

5.3.5 The Reich-Shaby Model

This was proposed by Reich and Shaby [190].

Define $X(s) = U(s)\theta(s)$ for all $s \in S$ for some spatial region S, where

1. $\Pr\{U(s) \le u\} = \exp\left(-u^{-1/\alpha}\right)$ independently for each s ($u \in (0,\infty)$, $\alpha \in (0,1)$),

2. $\theta(s) = \left(\sum_\ell A_\ell w_\ell(s)^{1/\alpha}\right)^\alpha$ where $\sum_\ell w_\ell(s) = 1$ for all $s \in S$ and $E\left(e^{-tA_\ell}\right) = \exp\left(-t^\alpha\right)$ independently for each ℓ.

Let Θ denote the process defined by all $\theta(s)$, $s \in S$. If we are given a finite set of $s \in S$ and values $x(s)$, we calculate

$$\begin{aligned}
\Pr\{X(s) \le x(s) \text{ for all } s\} &= E\left[\Pr\{U(s)\theta(s) \le x(s) \text{ for all } s \mid \Theta\}\right] \\
&= E\left[\prod_s \exp\left\{-x(s)^{-1/\alpha}\theta(s)^{1/\alpha}\right\}\right] \\
&= E\left[\exp\left\{-\sum_s x(s)^{-1/\alpha}\sum_\ell A_\ell w_\ell(s)^{1/\alpha}\right\}\right] \\
&= E\left[\exp\left\{-\sum_\ell A_\ell\left(\sum_s x(s)^{-1/\alpha}w_\ell(s)^{1/\alpha}\right)\right\}\right] \\
&= \exp\left\{-\sum_\ell\left(\sum_s x(s)^{-1/\alpha}w_\ell(s)^{1/\alpha}\right)^\alpha\right\}. \qquad (5.41)
\end{aligned}$$

Proof that (5.41) has the correct marginal distributions:

Suppose $x(s_0) = x \in (0,\infty)$ for some s_0, $x(s) = +\infty$ for all $s \ne s_0$. Then

$$\begin{aligned}
\Pr\{X(s_0) \le x\} &= \Pr\{X(s) \le x(s) \text{ for all } s\} \\
&= \exp\left\{-\sum_\ell\left(\sum_s x(s)^{-1/\alpha}w_\ell(s)^{1/\alpha}\right)^\alpha\right\} \\
&= \exp\left\{-\sum_\ell\left(x^{-1/\alpha}w_\ell(s_0)^{1/\alpha}\right)^\alpha\right\} \\
&= \exp\left(-x^{-1}\sum_\ell w_\ell(s_0)\right) \\
&= \exp\left(-x^{-1}\right)
\end{aligned}$$

because we fixed $\sum_\ell w_\ell(s) = 1$ for all $s \in S$.

Proof that (5.41) is max-stable:

The key property is to show that

$$\Pr\{X(s) \le nx(s) \text{ for all } s\}^n = \Pr\{X(s) \le x(s) \text{ for all } s\} \text{ for all } n \ge 1. \qquad (5.42)$$

According to (5.41),

$$\begin{aligned}\Pr\{X(s) \le nx(s) \text{ for all } s\}^n &= \exp\left\{-n\sum_\ell \left(\sum_s (nx(s))^{-1/\alpha} w_\ell(s)^{1/\alpha}\right)^\alpha\right\} \\ &= \exp\left\{-\sum_\ell \left(\sum_s (x(s))^{-1/\alpha} w_\ell(s)^{1/\alpha}\right)^\alpha\right\} \\ &= \Pr\{X(s) \le x(s) \text{ for all } s\}\end{aligned}$$

and this establishes (5.42).

5.4 Inference for Max-Stable Processes

We now consider the problem of first defining and then estimating the parameters of a max-stable process.

A typical scenario is that we observe $Y_i(\mathbf{s}_j)$, the annual maximum of a process in year i at site $\mathbf{s}_j$, where $i = 1,\ldots,n$ and there is a finite set of sampling locations $\mathscr{D} = \{\mathbf{s}_1,\ldots,\mathbf{s}_d\}$.

As a first step, we assume that the GEV parameters at site $\mathbf{s}$ are written $\mu(\mathbf{s};\boldsymbol{\phi}), \psi(\mathbf{s};\boldsymbol{\phi}), \xi(\mathbf{s};\boldsymbol{\phi})$ where we use $\boldsymbol{\phi}$ as a generic notation for the full set of parameters of the GEV distributions. For example, it's possible that μ, ψ, ξ are evaluated independently at each site, or are common to all sites, or are smooth functions of site characteristics such as latitude, longitude, elevation, mean surface pressure, etc. They may also be dependent on time-varying covariates (e.g. in the example of Section 5.1, the time-varying covariate was sea-surface temperature) and in that case we may write the GEV parameters for year i as $\mu_i(\mathbf{s};\boldsymbol{\phi}), \psi_i(\mathbf{s};\boldsymbol{\phi}), \xi_i(\mathbf{s};\boldsymbol{\phi})$, but we won't consider that explicitly as a separate case.

Anpther feature is that the *form* of dependence on site characteristics could be parametric (e.g. linear, quadratic) or nonparameteric (e.g. represented by thin-plate splines) but we treat those as equivalent here, e.g. spline representations become parametric models once the number and shape of the spline basis functions and their centers are specified.

The net result is a transformation

$$Z_i(\mathbf{s}_j;\boldsymbol{\phi}) = \left\{1 + \xi_i(\mathbf{s}_j;\boldsymbol{\phi})\frac{Y_i(\mathbf{s}_j) - \mu_i(\mathbf{s}_j;\boldsymbol{\phi})}{\psi_i(\mathbf{s}_j;\boldsymbol{\phi})}\right\}_+^{1/\xi_i(\mathbf{s}_i;\boldsymbol{\phi})} \tag{5.43}$$

which transforms all the marginal distributions to unit Fréchet ($F(z) = e^{-1/z}$, $z \ge 0$).

Then we assume that the joint distributions of $\{Z_i(\mathbf{s};\boldsymbol{\phi}),\ s \in \mathscr{D}\}$ follow some max-stable process indexed by parameters $\boldsymbol{\theta}$ (e.g. one possibility is a Brown-Resnick process generated by the power-law variogram model, in which case $\boldsymbol{\theta} = \begin{pmatrix} c_0 & c_1 & \lambda \end{pmatrix}$).

There are then two strategies:

(i) Estimate $\boldsymbol{\phi}$ first, in effect assuming independence from site to site, transform each $Y_i(\mathbf{s}_j)$ to $Z_i(\mathbf{s}_j;\boldsymbol{\phi})$, then proceed as if $\{Z_i(\mathbf{s}_j,\boldsymbol{\phi})\}$ were exactly a max-stable process with unit Fréchet margins;

(ii) Estimate $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ together.

Theoretical results (e.g. [216, 215]) nearly always show that (ii) is better, but in practice, researchers often opt for (i) because it is so much easier computationally. Nevertheless, recent Bayesian approaches [266, 267] have explored the possibilities of estimating both sets of parameters simultaneously.

For now, we focus on the problem of estimating $\boldsymbol{\theta}$, assuming that $\boldsymbol{\phi}$ is known. There are various plausible method, but we focus here on the *method of composite likelihood* before going on to consider some of the issues associated with a full likelihood calculation.

5.4.1 Method of composite likelihood

The difficulty here is that, while the formulas (5.32)–(5.33) define the d-dimensional joint distributions of $Z_i(\mathbf{s}_j)$ for any set of $\mathbf{s}_1,\ldots,\mathbf{s}_d$, there is no direct method for calculating the joint density and hence the likelihood function, which is needed for exact maximum likelihood or Bayesian estimation. Nevertheless, for many of the standard models for max-stable processes, including the Brown-Resnick, Extremal t, Smith and Schlather models, it is possible to write a closed-form expression for the joint density of any $d = 2$ sampling points. This fact prompted Padoan *et al.* [175] to propose the composite likelihood method, which we onw describe.

Suppose we observed $Z_i(\mathbf{s}),\ i = 1,\ldots,n$, for d sampling points $\mathbf{s}_j,\ j = 1,\ldots,d$. We assume these are dependent observations for each i, but the processes for different values of $i \in \{1,\ldots,n\}$ are independent. We consider a composite log likelihood of the form

$$\mathrm{CL}(\boldsymbol{\theta}) \quad = \quad \sum_{i=1}^{n}\sum_{j=2}^{d}\sum_{j'=1}^{j} w_{jj'}\ell_{i,j,j'}(\boldsymbol{\theta}) \tag{5.44}$$

where $\ell_{i,j,j'}(\boldsymbol{\theta}) = \log f(z_{ij}, z_{ij'}; \boldsymbol{\theta})$, f being the bivariate joint density of $z_{ij}, z_{ij'}$ when the parameter vector is $\boldsymbol{\theta}$. Note that we can restrict the inner sum to indices $j' < j$ because the joint density is symmetric in j and j' and the case $j = j'$ reduces to the univeriate density which is known to be unit Fréchet and therefore independent of $\boldsymbol{\theta}$. The function $w_{jj'}$ is some fixed set of weights, which could be identically 1 or could be some deistance-weighted function, for example $w_{jj'} = 1$ if the distance between sampling points $\mathbf{s}_j$ and $\mathbf{s}'_j$ is less than some threshold distance D, and 0 otherwise.

The composite maximum likelihood estimator $\hat{\boldsymbol{\theta}}_C$ is the value of $\boldsymbol{\theta}$ that maximizes $\mathrm{CL}(\boldsymbol{\theta})$.

To examine the properties of this estomator, consider the expressions

$$\begin{aligned}
\hat{K} &= \sum_i \sum_{j'<j} \frac{\partial \ell_{i,j,j'}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \ell_{i,j,j'}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}, \\
\hat{J} &= -\sum_i \sum_{j'<j} \frac{\partial^2 \ell_{i,j,j'}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T},
\end{aligned}$$

both evaluated at $\hat{\boldsymbol{\theta}}_C$.

We use the "sandwich estimator" $\hat{J}^{-1}\hat{K}\hat{J}^{-1}$ as an estimator of the covariance matrix of $\hat{\boldsymbol{\theta}}_C$. Padoan *et al.* showed that this is asymptotically a consistent estimator of the covariance matrix of $\boldsymbol{\theta}_C$ and proved a central limit theorem of the form

$$(\hat{J}^{-1}\hat{K}\hat{J}^{-1})^{-1/2}\left(\hat{\boldsymbol{\theta}}_C - \boldsymbol{\theta}\right) \xrightarrow{d} \mathcal{N}_p(0, I_p)$$

where p is the dimension of $\boldsymbol{\theta}$ and I_p is the $p \times p$ identity matrix.

They also defined a "composite likelihood information criterion"

$$\text{CLIC} \quad = \quad 2\left\{\text{tr}(\hat{J}^{-1}\hat{K}) - \text{CL}(\hat{\boldsymbol{\theta}}_C)\right\}$$

to discriminate among models, the best model among some finite family of models being judged to be the one that minimizes the CLIC. Precise statements and proofs of these results are in [175].

5.4.2 Progress towards exact maximum likelihood

5.5 Other Approaches to Spatial Extremes

See papers by [190, 191]

Others refs are [23, 22, 226, 166, 241, 173, 211, 125, 49, 50, 100, 78, 13, 209, 213, 52, 28, 39]

Bibliography

[1] R.L. Adler. Weak convergence results for extremal processes generated by dependent random variables. *Annals of Probability*, 6:660–667, 1978.

[2] M.A. Ben Alaya, F. Zwiers, and X. Zhang. An evaluation of block-maximum-based estimation of very long return period precipitation extremes with a large ensemble climate simulation. *Journal of Climate*, 33:6957–6970, 2020.

[3] C.W. Anderson. Extreme value theory for a class of discrete distributions with applications to some stochastic processes. *Journal of Applied Probability*, 7:99–113, 1970.

[4] C.W. Anderson. Local limit theorems for the maxima of discrete random variables. *Math. Proc. Camb. Phil. Soc.*, 88:161–165, 1980.

[5] R. Arratia, L. Goldstein, and L. Gordon. Two moments suffice for Poisson approximations: The Chen-Stein method. *Annals of Probability*, 17 (1):9–25, 1989.

[6] R. Arratia, L. Goldstein, and L. Gordon. Poisson approximation and the Chen-Stein method. *Statistical Science*, 5 (4):403–424, 1990.

[7] R. Arratia, L. Gordon, and M. Waterman. An extreme value theory for sequence matching. *Annals of Statistics*, 14:971–993, 1986.

[8] R. Arratia, L. Gordon, and M.S. Waterman. The Erdos-Renyi law in distribution, for coin tossing and sequence matching. *Annals of Statistics*, 18 (2):539–570, 1990.

[9] S. Asmussen. *Applied Probability and Queues*. Wiley, Chichester, 1987.

[10] A. A. Balkema and L. de Haan. Residual life time at great age. *Annals of Probability*, 2 (5):792–804, 1974.

[11] S. Banerjee, B. Carlin, and A. Gelfand. *Hierarchical Modeling and Analysis for Spatial Data (Second Edition)*. Boca Raton, FL: Chapman & Hall/CRC Press, 2014.

[12] A.D. Barbour, L. Holst, and S. Janson. *Poisson Approximation*. Oxford University Press, 1992.

[13] Jan Beirlant, Frederico Caeiro, and M. Ivette Gomes. An overview and open research topics in statistics of univariate extremes. *REVSTAT*, 10(1):1–31, 2012.

[14] M.L. Bell, A. McDermott, S.L. Zeger, J.M. Samet, and F. Dominici. Ozone and short-term mortality in 95 US urban communities, 1987–2000. *Journal of*

the American Medical Association, 292 (19):2372–2378, 2004.

[15] J. Berger, V. De Oliveira, and B. Sanso. Objective Bayesian analysis of spatially correlated data. *Journal of the American Statistical Association*, 96:1361–1374, 2001.

[16] S.M. Berman. Limiting distribution of the maximum term in a sequence of dependent random variables. *Annals of Mathematical Statistics*, 33:894–908, 1962.

[17] S.M. Berman. Limit theorems for the maximum term in stationary sequences. *Annals of Mathematical Statistics*, 35:502–516, 1964.

[18] J.-M. Bernardo. Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. Ser. B*, 41(2):113–147, 1979.

[19] J. Besag. Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B*, 36:192–236, 2016.

[20] Peter J. Bickel and Elizaveta Levina. Covariance regularization by thresholding. *Annals of Statistics*, 36 (6):2577–2604, 2008.

[21] T. Bollerslev. Generalized autoregressive conditional heteroscedasticy. *Econometrics*, 31:307–327, 1982.

[22] G. Bopp, B. Shaby, and R. Huser. A hierarchical max-infinitely divisible process for extreme areal precipitation over watersheds. *Arxiv 1804.04588*, 2018.

[23] B.M. Brown and S. Resnick. Extreme values of independent stochastic processes. *Journal of Applied Probability*, 14:732–739, 1977.

[24] A. Bücher and J. Segers. On the maximum likelihood estimator for the generalized extreme-value distribution. *Extremes*, 20:839–872, 2017.

[25] P. Capéraà, A.-L. Fougères, and C. Genest. A nonparametric estimation procedure for bivariate extreme value copulas. *Biometrika*, 84(3):567–577, 1997.

[26] D.E. Cartwright and M.S. Longuet-Higgins. The statistical distribution of the maxima of a random function. *Proc. R. Soc. Lond. A*, 237:212–232, 1956.

[27] E. Casson and S.G. Coles. Spatial regression models for extremes. *Extremes*, 1:449–468, 1999.

[28] V. Chavez-Demoulin and A.C. Davison. Modelling time series extremes. *REVSTAT*, 10(1):109–133, 2012.

[29] Daniel Clarkson, Emma Eastoe, and Amber Leeson. The importance of context in extreme value analysis with application to extreme temperatures in the USA and Greenland. *Applied Statistics, in press*, 2023.

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

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

[32] S.G. Coles and J.A. Tawn. Modelling extreme multivariate events. *Journal of the Royal Statistical Society, Series B*, 53(2):377–392, 1991.

[33] S.G. Coles and J.A. Tawn. Statistical methods for multivariate extremes: an application to structural design (with discussion). *Applied Statistics*, 43 (1):1–48, 1994.

[34] Stuart Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer-Verlag, London, https://doi.org/10.1007/978-1-4471-3675-0, 2001.

[35] Stuart Coles, Janet Heffernan, and Jonathan Tawn. Dependence measures for extreme value analyses. *Extremes*, 2 (4):339–365, 1999.

[36] D. Cooley, D. Nychka, and P. Naveau. Bayesian spatial modeling of extreme precipitation return levels. *Journal of the American Statistical Association*, 102 (479):824–840, 2007.

[37] D. Cooley and S. R. Sain. Spatial hierarchical modeling of precipitation extremes from a regional climate model. *Journal of Agricultural, Biological, and Environmental Statistics*, 15 (3):381–402, 2010.

[38] D. Cooley and E. Thibaud. Decompositions of dependence for high-dimensional extremes. *Biometrika*, 106:587–604, 2019.

[39] Daniel Cooley, Jessi Cisewski, Robert J. Erhardt, Soyoung Jeon, Elizabeth Mannshardt, Bernard Oguna Omolo, and Ying Sun. A survey of spatial extremes: Measuring spatial dependence and modeling spatial effects. *REVSTAT*, 10(1):135–165, 2012.

[40] National Research Council. *Climate and Social Stress: Implications for Security Analysis*. Washington, DC: The National Academies Press, 2013.

[41] N.A.C. Cressie. *Statistics for Spatial Data, Second Edition*. Wiley, New York, 1993.

[42] Noel Cressie and Christopher K. Wikle. *Statistics for Spatio-Temporal Data*. Wiley ISBN: 978-0-471-69274-4, 2011.

[43] M.J. Crowder. A multivariable distribution with Weibull connections. *Journal of the Royal Statistical Society, Series B*, 51:93–108, 1989.

[44] S. Csörgő, P. Deheuvels, and D.M. Mason. Kernel estimates of the tail index of a distribution. *Annals of Statistics*, 13:1050–1077, 1985.

[45] M.J. Daniels and R.E. Kass. Shrinkage estimators for covariance matrices. *Biometrics*, 57:1173–1184, 2001.

[46] R.W.R. Darling and M.S. Waterman. Extreme value distribution for the largest cube on a random lattice. *SIAM J. Appl. Math.*, 46:118–132, 1986.

[47] R.A. Davis. Limit laws for the maximum and minimum of stationary sequences. *Z. Wahrsch. v. geb.*, 61:31–42, 1982.

[48] A.C. Davison, R. Huser, and E. Thibaud. Spatial Extremes. *Handbook of Environmental and Ecological Statistics (A. Gelfand, M. Fuentes, J, Hoeting, R. Smith, eds.)*, pages 711–744, 2019.

[49] A.C. Davison, S.A. Padoan, and M. Ribatet. Statistical modeling of spatial

extremes (with discussion). *Statistical Science*, 27:161–186, 2012.

[50] A.C. Davison and N.I. Ramesh. Local likelihood smoothing of sample extremes. *Journal of the Royal Statistical Society, Series B*, 62:191–208, 2000.

[51] A.C. Davison and R.L. Smith. Models for exceedances over high thresholds (with discussion). *Journal of the Royal Statistical Society, Series B*, 52(3):393–442, 1990.

[52] Miguel de Carvalho and Alexandra Ramos. Bivariate extreme statistics, II. *REVSTAT*, 10(1):83–107, 2012.

[53] P. Deheuvels. On the limiting behavior of the pickands estimator for bivariate extreme-value distributions. *Statistics and Probability Letters*, 12:429–439, 1991.

[54] Paul Deheuvels. Caracterisation complète des lois extrêmes multivariées et de la convergence des types extremes. *Publ. Inst. Statist. Univ. Paris*, 23:1–36, 1978.

[55] G.E. Denzel and G.L. O'Brien. Limit theorems for extreme values of chain-dependent processes. *Annals of Probability*, 3:773–779, 1975.

[56] Jean Diebolt, Armelle Guillou, Philippe Naveau, and Pierre Ribereau. Improving probability-weighted moment methods for the generalized extreme value distribution. *REVSTAT – Statistical Journal Volume*, 6:33–50, 04 2008.

[57] C. Dombry. Existence and consistency of the maximum likelihood estimators for the extreme value index within the block maxima framework. *Bernoulli*, 21(1):420–436, 2015.

[58] C. Dombry and A. Ferreira. Maximum likelihood estimators based on the block maxima method. *Bernoulli*, 25:1690–1723, 2019.

[59] F. Dominici, A. McDermott, M. Daniels, S.L Zeger, and J.M. Samet. Revised analyses of the National Morbidity, Mortality, and Air Pollution Study: mortality among residents of 90 cities. *Journal of Toxicology and Environmental Health Part A*, 68:1071–1092, 2005.

[60] F. Dominici, J.M. Samet, and S.L. Zeger. Combining evidence on air pollution and daily mortality from the 20 largest US cities: a hierarchical modelling strategy. *Journal of the Royal Statistical Society, Series A*, 163(3):263–302, 2000.

[61] Holger Drees, Ana Ferreira, and Laurens de Haan. On maximum likelihood estimation of the extreme value index. *Annals of Applied Probability*, 14 (3):1179–1201, 2004.

[62] Meyer Dwass. Extremal processes. *Annals of Mathematical Statistics*, 35 (4):1718–1725, 1964.

[63] A. V. Dyrrdal, A. Lenkoski, T. L. Thorarinsdottir, and F. Stordal. Bayesian hierarchical modeling of extreme hourly precipitation in Norway. *Environmetrics*, 26 (2):89–106, 2015.

[64] W. Dziubdziela. Limit laws for kth order statistics from strong mixing pro-

cesses. *Journal of Applied Probability*, 21:720–729, 1984.

[65] W. Dziubdziela. Limit laws for the kth order statistics from conditionally mixing arrays of random variables. *Journal of Applied Probability*, 23:679–687, 1986.

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

[67] R.F. Engle. Autoregressive conditional heteroscedastic models with estimates of the variance of the united kingdom inflation. *Econometrica*, 50:987–1007, 1982.

[68] Veronika Eyring, Sandrine Bony, Gerald A. Meehl, Catherine A. Senior, Bjorn Stevens, Ronald J. Stouffer, and Karl E. Taylor. Overview of the Coupled Model Intercomparison Project Phase 6(CMIP6) experimental design and organization. *Geosci. Model Dev.*, 9:1937–1958, 2016.

[69] L. Fawcett and D. Walshaw. A hierarchical model for extreme wind speeds. *Applied Statistics*, 55:631–646, 2006.

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

[71] W. Feller. *An Introduction to Probability Theory and Its Applications, Volume II (Second edition)*. John Wiley & Sons, New York, 1971.

[72] Ana Ferreira and Laurens de Haan. On the block maxima method in extreme value theory: PWM estimators. *Annals of Statistics*, 43:276–298, 2015.

[73] H. Ferreira and M. Ferreira. Estimating the extremal index through local dependence. *Annales de l'Institut Henri Poincaré - Probabilités et Statistiques*, 54 (2):587–605, 2018.

[74] S. Ferreira and L. de Haan. On the block maxima method in extreme value theory: PWM estimators. *Annals of Statistics*, 43 (1):276–298, 2015.

[75] C.A.T. Ferro and J. Segers. Inference for clusters of extreme values. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 65 (2):545–556, 2003.

[76] A. O. Finley, S. Banerjee, A. R. Ek, and R. E. McRoberts. Bayesian multivariate process modeling for prediction of forest attributes. *Journal of Agricultural, Biological, and Environmental Statistics*, 13 (1):60–83, 2008.

[77] 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.

[78] R. de Fondeville and A.C. Davison. High-dimensional peaks over threshold inference. *Biometrika*, 105:575–592, 2018.

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

[80] R. Furrer, M. G. Genton, and D. Nychka. Covariance tapering for interpolation

of large spatial datasets. *Journal of Computational and Graphical Statistics*, 15 (3):502–523, 2006.

[81] C. Gaetan and M. Grigoletto. A hierarchical model for the analysis of spatial rainfall extremes. *Journal of Agricultural, Biological, and Environmental Statistics*, 12 (4):434–449, 2007.

[82] J. Galambos. *The Asymptotic Theory of Extreme Order Statistics, Second Edition*. Krieger, Florida. First edition published by Wiley, 1978, 1987.

[83] J. Geffroy. Contributions à la theorie des valeurs extremes. *Publ Inst Statist Univ Paris*, 7/8:37–185, 1958/59.

[84] Andrew Gelman, John Carlin, Hal Stern, David Dunson, Aki Vehtari, and Donald Rubin. *Bayesian Data Analysis (third edition)*. Chapman and Hall/CRC Press, New York, 2020.

[85] Souparno Ghosha and Bani K. Mallick. A hierarchical Bayesian spatio-temporal model for extreme precipitation events. *Environmetrics*, 22:192–204, 2011.

[86] Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods (with discussion). *Journal of the Royal Statistical Society, Series B*, 73 (2):123–214, 2011.

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

[88] C.M. Goldie. Implicit renewal theory and tails of solutions of random equations. *Annals of Applied Probability*, 1 (1):126–166, 1991.

[89] C.M Goldie and R.L. Smith. Slow variation with remainder: A survey of the theory and its applications. *Quarterly Journal of Mathematics Oxford Series (2)*, 38:45–71, 1987.

[90] M.I. Gomes. An i-dimensional limiting distribution function of largest values and its relevance to the statistical theory of extremes. *Statistical Distributions in Scientific Work (C. Taillie, G. P. Patil and B. A. Baldessari, eds.), Reidel, Dordrecht*, 6:389–410, 1981.

[91] L. Gordon, M.F. Schilling, and M.S. Waterman. An extreme value theory for long head runs. *Prob. Th. Rel. Fields*, 72:279–287, 1986.

[92] R. Grübel. Algorithm as 265: G/g/1 via fast fourier transform. *Applied Statistics*, 40:355–365, 1991.

[93] E.J. Gumbel. *Statistics of Extremes*. Columbia University Press, New York, 1958.

[94] E.J. Gumbel. Bivariate exponential distributions. *Journal of the American Statistical Association*, 55:698–707, 1960.

[95] E.J. Gumbel. Distributions des valeurs extrêmes en plusiers dimensions. *Publications de l'Institut de Statistique de l'Université de Paris*, 9:171–173, 1960.

[96] E.J. Gumbel and N. Goldstein. Analysis of empirical bivariate extremal distributions. *Journal of the American Statistical Association*, 59:794–816, 1964.

[97] E.J. Gumbel and C.K. Mustafi. Some analytical properties of bivariate extreme value distributions. *Journal of the American Statistical Association*, 62:569–588, 1967.

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

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

[100] L. De Haan. A spectral representation for max-stable processes. *Annals of Probability*, 12:1194–1204, 1984.

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

[102] L. de Haan and S. Resnick. Limit theory for multivariate sample extremes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 40:317–337, 1977.

[103] L. de Haan, S.I. Resnick, H. Rootzén, and C.G. de Vries. Extremal behaviour of solutions to a stochastic difference equations with applications to ARCH processes. *Stochastic Processes and their Applications*, 32:213–224, 1989.

[104] L. de Haan and J. de Ronde. Sea and wind: Multivariate extremes at work. *Extremes*, 1 (1):7–45, 1998.

[105] L. de Haan and U. Stadtmüller. Generalized regular variation of second order. *Journal of the Australian Mathematical Society, Series A*, 61:381–395, 1996.

[106] Heikki Haario, Eero Saksman, and Johanna Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7 (2):223–242, 2001.

[107] E. Haeusler and J.L. Teugels. On asymptotic normality of Hill's estimator for the exponent of regular variation. *Annals of Statistics*, 13:743–756, 1985.

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

[109] P. Hall and A.H. Welsh. Best attainable rates of convergence for estimates of parameters of regular variation. *Annals of Statistics*, 12:1079–1084, 1984.

[110] P. Hall and A.H. Welsh. Adaptive estimates of parameters of regular variation. *Annals of Statistics*, 13:331–341, 1985.

[111] Peter Hall. On some simple estimates of an exponent of regular variation. *Journal of the Royal Statistical Society, Series B*, 44:37–42, 1982.

[112] M.J. Heaton, M. Katzfuss, S. Ramachandar, K. Pedings, E. Gilleland, E. Mannshardt, and R.L. Smith. Spatio-temporal models for large-scale indicators of extreme weather. *Environmetrics*, 22 (3):294–303, 2010.

[113] J.E. Heffernan and J.A. Tawn. A conditional approach for multivariate extreme values (with discussion). *Journal of the Royal Statistical Society, Series B*, 66(3):497–546, 2004.

[114] C. Herweijer and R. Seager. The global footprint of persistent extra-tropical drought in the instrumental era. *International Journal of Climatology*, 28

(13):1761–1774, 2008.

[115] Bruce M. Hill. A simple general approach to inference about the tail of a distribution. *Annals of Statistics*, 3 (5):1163–1174, 1975.

[116] D.M. Holland, V. De Oliveira, L.H. Cox, and R.L. Smith. Estimation of regional trends in sulfur dioxide over the eastern united states. *Environmetrics*, 11:373–393, 2000.

[117] Gerard Hooghiemstra and Ludolf E. Meester. The extremal index in 10 seconds. *Journal of Applied Probability*, 34 (3):818–822, 1997.

[118] J.R.M. Hosking, J.R. Wallis, and E.F. Wood. Estimation of the generalized extreme value distribution by the method of probability-weighted moments. *Technometrics*, 27:251–261, 1985.

[119] P. Hougaard. A class of multivariate failure time distributions. *Biometrika*, 73:671–678, 1986.

[120] T. Hsing. Extreme value theory for suprema of random variables with regularly varying tail probabilities. *Stochastic Processes and their Applications*, 22:51–57, 1987.

[121] T. Hsing. On the characterization of certain point processes. *Stochastic Processes and their Applications*, 26:297–316, 1987.

[122] T. Hsing. On the extreme order statistics for a stationary sequence. *Stochastic Processes and their Applications*, 29:155–169, 1988.

[123] T. Hsing, J. Hüsler, and M.R. Leadbetter. On the exceedance point process for a stationary sequence. *Probability Theory and Related Fields*, 78:97–112, 1988.

[124] Gabriel Huerta and Bruno Sansó. Time-varying models for extreme values. *Environmental and Ecological Statistics*, 14:285–299, 2007.

[125] J. Hüsler and R.-D. Reiss. Maxima of normal random variables: between independence and complete dependence. *Statistics and Probability Letters*, 7:283–286, 1989.

[126] Harold Jeffreys. *Theory of Probability, First Edition*. Oxford University Press (second edition available from https://archive.org/details/in.ernet.dli.2015.2608/page/n5/mode/2up), 1939.

[127] A.F. Jenkinson. The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quarterly Journal of the Royal Meteorological Society*, 81 (348):158–171, 1955.

[128] A.F. Jenkinson. Statistics of extremes. *Estimation of Maximum Floods: WMO Techical Note 98, https://library.wmo.int/doc_num.php?explnum_id=3444*, pages 183–228, 1969.

[129] Y. Jiang, D. Cooley, and M.P. Wehner. Principal component analysis for extremes and application to US precipitation. *Journal of Climate*, 33 (15):6441–6451, 2020.

[130] H. Joe. Families of min-stable multivariate exponential and multivariate ex-

treme value distributions. *Statistics and Probability Letters*, 9:75–82, 1989.

[131] H. Joe, R.L. Smith, and I. Weissman. Bivariate threshold methods for extremes. *Journal of the Royal Statistical Society, Series B*, 54(1):171–183, 1992.

[132] Harry Joe. *Multivariate Models and Multivariate Dependence Concepts*. Chapman & Hall/CRC Monographs on Statistics and Applied Probability, 1997.

[133] Olav Kallenberg. *Random Measures, Theory and Applications*. Springer International Publishing Switzerland 2017, 2017.

[134] Matthias Katzfuss. A multi-resolution approximation for massive spatial datasets. *Journal of the American Statistical Association*, 112 (517):201–214, 2017.

[135] H. Kesten. Random difference equations and renewal theory for products of random matrices. *Acta Mathematica*, 131:207–248, 1973.

[136] S. Kotz and N.L. Johnson. *Breakthroughs In Statistics, Volume 1*. Springer, New York, 1992.

[137] John Lamperti. On extreme order statistics. *Annals of Mathematical Statistics*, 35 (4):1726–1737, 1964.

[138] W.K.M. Lau and K.-M. Kim. The 2010 Pakistan flood and Russian heat wave: Teleconnection of hydrometeorological extremes. *Journal of Hydrometeorology*, 13:392–403, 2012.

[139] F. Laurini and J Tawn. The extremal index for a GARCH(1,1) process. *Extremes*, 15:511–529, 2012.

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

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

[142] M.R. Leadbetter. Extremes and local dependence in stationary sequences. *Z. Wahrsch. v. geb.*, 65:291–306, 1983.

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

[144] A.W. Ledford and J.A. Tawn. Statistics for near independence in multivariate extreme values. *Biometrika*, 83(1):169–187, 1996.

[145] A.W. Ledford and J.A. Tawn. Statistics for near independence in multivariate extreme values. *Biometrika*, 83(1):169–187, 1996.

[146] A.W. Ledford and J.A. Tawn. Modeling dependence within joint tail regions. *Journal of the Royal Statistical Society, Series B*, 59(2):475–499, 1997.

[147] A.W. Ledford and J.A. Tawn. Modelling dependence within joint tail regions. *Journal of the Royal Statistical Society, Series B*, 59(2):475–499, 1997.

[148] A.W. Ledford and J.A. Tawn. Diagnostics for dependence within time series extremes. *Journal of the Royal Statistical Society, Series B*, 65(2):521–543, 2003.

[149] R.M. Loynes. Extreme values in uniformly mixing stationary stochastic processes. *Annals of Mathematical Statistics*, 36:993–999, 1965.

[150] M.B. Marcus and M. Pinsky. On the domain of attraction of $\exp(-e^{-x})$. *J. Math. Anal. Appl.*, 28:440–449, 1969.

[151] A.W. Marshall and I. Olkin. A multivariate exponential distribution. *Journal of the American Statistical Association*, 62:30–44, 1967.

[152] A.W. Marshall and I. Olkin. Domains of attraction of multivariate extreme value distributions. *Annals of Probability*, 11:168–177, 1983.

[153] A.P. Martins and H. Ferreira. The multivariate extremai index and the dependence structure of a multivariate extreme value distribution. *Sociedad de Estadzıistica e Investigación Operativa*, 14 (2):433–448, 2005.

[154] Eduardo S. Martins and Jery R. Stedinger. Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data. *Water Resources Research*, 36(3):737–744, 2000.

[155] B. Matérn. *Spatial Variation*. Springer, New York, 1960 (republished 1986 as Lecture Notes in Statistics, Vol. 36).

[156] K. Maulik and S.I. Resnick. Characterizations and examples of hidden regular variation. *Extremes*, 7 (1):31–67, 2004.

[157] D. McFadden. Modelling the choice of residential location. *Spatial Interaction Theory and Planning Models, edited by A. Karlqvist, L. Lundquist, F. Snickers and J. Weibull, North Holland, Amsterdam*, pages 75–96, 1978.

[158] D.G. Mejzler. On a theorem of b.v. gnedenko. *Sb. Trudov Inst. Mat. Akad. Nauk. Ukrain. SSR*, 12:31–35, 1949.

[159] Sean P. Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Springer, London, 1993.

[160] T. Mikosch and C. Stărică. Limit theory for the sample autocorrelations and extremes of a GARCH(1,1) process. *Annals of Statistics*, 28 (5):1427–1451, 2000.

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

[162] T. Mori. Limit distributions of two-dimensional point processes generated by strong mixing sequences. *Yokohama Math. J.*, 25:155–168, 1977.

[163] S. Nandagopalan. *Multivariate Extremes and Estimation of the Extremal Index*. PhD Dissertation, University of North Carolina, Chapel Hill, 1990.

[164] S. Nandagopalan. Inference for the limiting cluster size distribution of extreme values. *Journal of Research of the National Institute of Standards and Technology*, 99 (4):543–550, 1994.

[165] G.F. Newell. Asymptotic extremes for m-dependent random variables. *Annals*

of Mathematical Statistics, 35:1322–1325, 1964.

[166] A.K. Nikoloulopoulos, H. Joe, and H. Li. Extreme value properties of multivariate t copulas. *Extremes*, 12:129–148, 2009.

[167] E. Nummelin. *General Irreducible Markov Chains and Non-Negative Operators*. Cambridge University Press, 1984.

[168] Douglas Nychka, Reinhard Furrer, John Paige, and Stephan Sain. fields, 2017. R package version 10.3.

[169] G.L. O'Brien. Limit theorems for the maximum term of a stationary process. *Annals of Probability*, 2:540–545, 1974.

[170] G.L. O'Brien. The limiting distribution of maxima of random variables defined on a denumerable Markov chain. *Annals of Probability*, 2:103–111, 1974.

[171] G.L. O'Brien. The maximum term of uniformly mixing stationary processes. *Z. Wahrsch. v. geb.*, 30:57–63, 1974.

[172] G.L. O'Brien. Extreme values for stationary and Markov sequences. *Annals of Probability*, 15:281–291, 1987.

[173] T. Opitz. Extremal t processes: elliptical domain of attraction and a spectral representation. *Journal of Multivariate Analysis*, 122:409–413, 2013.

[174] S. Padoan and S. Rizzelli. Empirical bayes inference for the block maxima method. *Preprint*, 2023.

[175] S. A. Padoan, M. Ribatet, and S. A. Sisson. Likelihood-based inference for max-stable processes. *Journal of the American Statistical Association*, 105:489:263–277, 2010.

[176] R. Perfekt. Extremal behaviour of stationary markov chains with applications. *Annals of Applied Probability*, 4:529–548, 1994.

[177] R. Perfekt. Extreme value theory for a class of markov chains with values in $\mathbb{R}^d$. *Advances In Applied Probability*, 29:138–164, 1997.

[178] Sjoukje Y. Philip, Sarah F. Kew, Geert Jan van Oldenborgh, Wenchang Yang, Gabriel A. Vecchi, Faron S. Anslow, Sihan Li, Sonia I. Seneviratne, Linh N. Luu, Julie Arrighi, Roop Singh, Maarten van Aalst, Mathias Hauser, Dominik L. Schumacher, Carolina Pereira, Marghidan, Kristie L Ebi, Rémy Bonnet, Robert Vautard, Jordis Tradowsky, Dim Coumou, Flavio Lehner, Michael Wehner, Chris Rodell, Roland Stull, Rosie Howard, Nathan Gillett, and Friederike E L Otto. Rapid attribution analysis of the extraordinary heatwave on the Pacific Coast of the US and Canada June 2021. *Earth System Dynamics*, 13 (4):1689–1713, 2022, doi: https://doi.org/10.5194/esd-13-1689-2022.

[179] J. Pickands. Multivariate extreme value distributions. *Bulletin of the International Statistical Institute*, 49:859–878, 1981.

[180] J. Pickands III. Multivariate negative exponential and extreme value distributions. 1969.

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

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

[183] J. Pickands III. The continuous and differentiable domains of attraction of the extreme value distributions. *Annals of Probability*, 14 (3):996–1004, 1986.

[184] P. Pirazzoli. Maree estreme a Venezia (periodo 1872–1981). *Acqua Aria*, 10:1023–1039, 1982.

[185] P. Prescott and A.T. Walden. Maximum likelihood estimation of the parameters of the generalized extreme value distribution. *Biometrika*, 67:723–724, 1980.

[186] P. Prescott and A.T. Walden. Maximum likelihood estimation of the parameters of the generalized extreme value distribution from censored samples. *Journal of Statistical Computation and Simulation*, 16:241–250, 1983.

[187] X. Qin, R.L. Smith, and R.E. Ren. Modelling multivariate extreme dependence. *Proceedings of the 2008 Joint Stastistical Meetings, Risk Analysis Section, American Statistical Association*, pages 3089–3096, 2009.

[188] A. Ramos and A.W. Ledford. A new class of models for bivariate joint tails. *Journal of the Royal Statistical Society, Series B*, 71(1):219–241, 2009.

[189] Alexandra Ramos and Anthony Ledford. An alternative point process framework for modeling multivariate extreme values. *Communications in Statistics - Theory and Methods*, 40:12:2205–2224, 2011.

[190] B.J. Reich and B.A. Shaby. A hierarchical max-stable spatial model for extreme precipitation. *Annals of Applied Statistics*, 6:1430–1451, 2012.

[191] B.J. Reich and B.A. Shaby. A spatial Markov model for climate extremes. *Journal of Computational and Graphical Statistics*, 28(1):117–126, 2019.

[192] S. Resnick. Weak convergence to extremal processes. *Annals of Probability*, 3:951–960, 1975.

[193] S. Resnick and M. Neuts. Limit laws for maxima of a sequence of random variables defined on a markov chain. *Advances In Applied Probability*, 2:323–343, 1970.

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

[195] S.I. Resnick. Hidden regular variation, second order regular variation and asymptotic independence. *Extremes*, 5 (4):303–336, 2002.

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

[197] M. D. Risser and M. F. Wehner. Attributable human-induced changes in the likelihood and magnitude of the observed extreme precipitation during Hurricane Harvey. *Geophysical Research Letters*, 44 (24):12457–12464, 2017.

[198] C.Y. Robert. Estimating the multivariate extremal index function. *Bernoulli*,

14 (4):1027–1064, 2008.

[199] C.Y. Robert. Inference for the limiting cluster size distribution of extreme values. *Annals of Statistics*, 37 (1):271–310, 2009.

[200] Michael E. Robinson and Jonathan A. Tawn. Statistics for exceptional athletics records. *Applied Statistics*, 44 (4):499–511, 1995.

[201] H. Rootzén. Maxima and exceedances of stationary Markov chains. *Advances in Applied Probability*, 20:371–390, 1988.

[202] H. Rootzén and N. Tajvidi. Multivariate generalized pareto distributions. *Bernoulli*, 12 (5):917–930, 2006.

[203] N. Ross. Fundamentals of Stein's method. *Probability Surveys*, 8:210–293, 2011.

[204] B Russell, M. Risser, R.L. Smith, and K.E. Kunkel. Investigating the association between late spring Gulf of Mexico sea surface temperatures and US Gulf Coast precipitation extremes with focus on Hurricane Harvey. *Environmetrics*, 31(2):e2595, 2020.

[205] B. T. Russell, D. S. Cooley, W. C. Porter, and C. L. Heald. Attributable human-induced changes in the likelihood and magnitude of the observed extreme precipitation during Hurricane Harvey. *Environmetrics*, 27 (6):334–344, 2016.

[206] B.T. Russell, D.S. Cooley, W.C. Porter, B.J. Reich, and C.L. Heald. Data mining to investigate the meteorological drivers for extreme ground level ozone events. *Annals of Applied Statistics*, 10 (3):1673–1698, 2016.

[207] H. Sang and A. E. Gelfand. Hierarchical modeling for extreme values observed over space and time. *Environmental and Ecological Statistics*, 16 (3):407–426, 2009.

[208] H. Sang and A. E. Gelfand. Continuous spatial process models for spatial extreme values. *Journal of Agricultural, Biological and Environmental Statistics*, 15:49–65, 2010.

[209] Carl Scarrott and Anna MacDonald. A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT*, 10(1):33–60, 2012.

[210] Oliver Schabenberger and Carol A. Gotway. *Statistical Methods for Spatial Data Analysis*. Chapman & Hall/CRC Texts in Statistical Science, 2004.

[211] M. Schlather. Models for stationary max-stable random fields. *Extremes*, 5:33–44, 2002.

[212] E. M. Schliep, D. Cooley, S. R. Sain, and J. A. Hoeting. A comparison study of extreme precipitation from six different regional climate models via spatial hierarchical modeling. *Extremes*, 13 (2):219–239, 2010.

[213] Johan Segers. Max-stable models for multivariate extremes. *REVSTAT*, 10(1):61–82, 2012.

[214] Francesco Serinaldi. Analysis of inter-gauge dependence by kendall's τ, upper tail dependence coefficient, and 2-copulas with application to rainfall fields. *Stoch Environ Res Risk Assess*, 22:671–688, 2008.

[215] Daoji Shi. Fisher information for a multivariate extreme value distribution. *Biometrika*, 82 (3):644–649, 1995.

[216] Daoji Shi, R.L. Smith, and S.G. Coles. Joint versus marginal estimation for bivariate extremes. *Unpublished, available at http://rls.sites.oasis.unc.edu/postscript/rs/maest.pdf*, 1992.

[217] M. Sibuya. Bivariate extreme statistics. *Ann Inst Statist math*, 11:195–210, 1960.

[218] A. Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8:229–231, 1959.

[219] Richard L. Smith. Extreme value theory based on the r largest annual events. *Journal of Hydrology*, 86:27–43, 1986.

[220] Richard L. Smith. Letter to the editors: Statistics for exceptional athletics records. *Applied Statistics*, 46 (1):123–128, 1997.

[221] R.L. Smith. Uniform rates of convergence in extreme-value theory. *Advances in Applied Probability*, 14 (3):600–622, 1982.

[222] R.L. Smith. Maximum likelihood estimation in a class of non-regular cases. *Biometrika*, 72:67–92, 1985.

[223] R.L. Smith. Statistics of extreme values. *Bulletin of the International Statistical Institute, Paper 26.1*, 51:1–17, 1985.

[224] R.L. Smith. Estimating tails of probability distributions. *Annals of Statistics*, 15 (3):1174–1207, 1987.

[225] R.L. Smith. Extreme value analysis of environmental time series: An example based on ozone data (with discussion). *Statistical Science*, 4:367–393, 1989.

[226] R.L. Smith. Max-stable processes and spatial extremes. Technical Report, University of North Carolina at Chapel Hill. 1990.

[227] R.L. Smith. The extremal index for a Markov chain. *Journal of Applied Probability*, 29 (1):37–45, 1992.

[228] R.L. Smith. Multivariate threshold methods. *In Extreme Value Theory and Applications, Eds. J. Galambos, J. Lechner and E. Simiu, Dordrecht: Kluwer*, pages 225–248, 1994.

[229] R.L. Smith. Statistics of extremes, with applications in environment, insurance and finance. *Extreme Values in Finance, Telecommunications and the Environment, edited by B. Finkenstadt and H. Rootzen, Chapman and Hall/CRC Press, London*, pages 1–78, 2003.

[230] R.L. Smith and D.J. Goodman. Bayesian risk analysis. *Extremes and Integrated Risk Management, Chapter 17 (Risk Books, London, edited by P. Embrechts)*, pages 235–251, 2000.

[231] R.L. Smith, J.A. Tawn, and S.G. Coles. Markov chain models for threshold exceedances. *Biometrika*, 84(2):249–268, 1997.

[232] R.L. Smith, J.A. Tawn, and H.-K. Yuen. Statistics of multivariate extremes. *International Statistical Review*, 58:47–58, 1990.

[233] R.L. Smith and I. Weissman. Maximum likelihood estimation of the lower tail of a probability distribution. *Journal of the Royal Statistical Society, Series B*, 47 (2):285–298, 1985.

[234] R.L. Smith and I. Weissman. Estimating the extremal index. *Journal of the Royal Statistical Society, Series B*, 56:515–528, 1994.

[235] R.L. Smith and I. Weissman. Characterization and estimation of the multivariate extremal index. *Technical Report, University of North Carolina at Chapel Hill*, 1996.

[236] R.L. Smith, B. Xu, and P. Switzer. Reassessing the relationship between ozone and short-term mortality in U.S. urban communities. *Inhalation Toxicology*, 21 (S2):37–61, 2009.

[237] J.A. Tawn. Bivariate extreme value theory: Models and estimation. *Biometrika*, 75(3):397–415, 1988.

[238] J.A. Tawn. Modelling multivariate extreme value distributions. *Biometrika*, 77(2):245–253, 1990.

[239] Jonathan A. Tawn. An extreme-value theory model for dependent observations. *Journal of Hydrology*, 101:227–250, 1988.

[240] Karl E. Taylor, Ronald J. Stouffer, and Gerald A. Meehl. An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, 93:485–498, 2012.

[241] E. Thibaud and T. Opitz. Efficient inference and simulation for elliptical Pareto processes. *Biometrika*, 102:855–870, 2015.

[242] J. Tiago de Oliveira. Extremal distributions. *Rev. Fac. Ciencias Lisboa, 2 ser., A, Mat.*, VII, 1958.

[243] J. Tiago de Oliveira. Bivariate models for extremes: Statistical decision. *Statistical Extremes and Applications (Tiago de Oliveira, ed.), Reidel, Dordrecht*, pages 131–153, 1984.

[244] J. Tiago de Oliveira. Intrinsic estimation of the dependence structure for bivariate extremes. *Statistics and Probability Letters*, 8:213–218, 1989.

[245] K. F. Turkman, M. A. A. Turkman, and J. M. Pereira. Asymptotic models and inference for extremes of spatio-temporal data. *Extremes*, 13:375–397, 2010.

[246] M. R. Tye and D. Cooley. A spatial model to examine rainfall extremes in Colorado's Front Range. *Journal of Hydrology*, 530 (Supplement C):15–23, 2015.

[247] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

[248] W. Vervaat. On a stochastic difference equation and a representation of non-negative infinitely divisible random variables. *Advances in Applied Probability*, 11 (4):750–783, 1979.

[249] H. Wackernagel. *Multivariate Geostatistics*. Springer Science & Business Media, 2003.

[250] J.L. Wadsworth and J.A. Tawn. A new representation for multivariate tail

probabilities. *Bernoulli*, 19 (5B):2689–2714, 2013.

[251] J.L. Wadsworth, J.A. Tawn, A.C. Davison, and D.M. Elton. Modelling across extremal dependence classes. *Journal of the Royal Statistical Society, Series B*, 79(1):149–175, 2017.

[252] G.S. Watson. Extreme values in samples from m-dependent stationary stochastic processes. *Annals of Mathematical Statistics*, 25:798–800, 1954.

[253] W. Weibull. A statistical distribution function of wide applicability. *Journal of Applied Mechanics*, 18:293–297, 1951.

[254] I. Weissman. Extremal processes generated by independent nonidentically distributed random variables. *Annals of Probability*, 3:172–177, 1975.

[255] I. Weissman. Multivariate extremal processes generated by independent nonidentically distributed random variables. *Journal of Applied Probability*, 12:477–487, 1975.

[256] I. Weissman. On location and scale functions for a class of limiting processes with applications to extreme value theory. *Annals of Probability*, 3:178–181, 1975.

[257] 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.

[258] R.E. Welsch. A weak convergence theorem for order statistics from strong mixing processes. *Annals of Mathematical Statistics*, 42:1637–1646, 1971.

[259] R.E. Welsch. Limit laws for extreme order statistics from strong mixing processes. *Annals of Mathematical Statistics*, 43:439–446, 1972.

[260] H. Wendland. Piecewise polynomial, positive de
nite and compactly supported radial functions of minimal degree. *Advances in Computational Mathematics*, 4 (1):389–396, 1995.

[261] P. Whittle. On stationary processes in the plane. *Biometrika*, 41:434–449, 1954.

[262] H.K. Yuen. *Estimation of multivariate extreme value distributions by a kernel method with an application to non-Gaussian time series*. PhD Thesis, University of Surrey, 1988.

[263] S. Yun. The extremal index of a higher-order stationary Markov chain. *Annals of Applied Probability*, 8 (2):408–437, 1998.

[264] S. Yun and R.L. Smith. Spatial trends and spatial extremes in south korean ozone. *Journal of the Korean Statistical Society*, 32 (4):313–335, 2003.

[265] L. Zhang and B.A. Shaby. Uniqueness and global optimality of the maximum likelihood estimator for the generalized extreme value distribution. *Biometrika*, 109 (3):853–864, 2021.

[266] L. Zhang, B.A. Shaby, and J.L. Wadsworth. Hierarchical transformed scale mixtures for flexible modeling of spatial extremes on datasets with many locations. *Journal of the American Statistical Association*, 117:539:1357–1369,

2022.

[267] Likun Zhang, Mark Risser, Michael Wehner, and Travis A. O'Brien. Explaining the unexplainable: leveraging extremal dependence to characterize the 2021 pacific northwest heatwave. *https://arxiv.org/pdf/2307.03688.pdf*, 2023.

[268] Likun Zhang and Benjamin A. Shaby. Reference priors for the generalized extreme value distribution. *Statistica Sinica*, to appear, 2024.

[269] Z. Zhang and R.L. Smith. The behavior of multivariate maxima of moving maxima processes. *Journal of Applied Probability*, 41(4):1113–1123, 2004.