Statistical Methods for Clusters of Extreme Values

Christopher A. T. Ferro, B.Sc. M.Sc.

Submitted for the degree of Doctor of Philosophy

at Lancaster University, September 2003.

I declare that the work presented in this thesis is my own, except where stated otherwise.

Christopher Ferro

21 September, 2003

Statistical Methods for Clusters of Extreme Values

Christopher A. T. Ferro, B.Sc. M.Sc.

Submitted for the degree of Doctor of Philosophy

at Lancaster University, September 2003.

Abstract

Extreme values in sequences of independent random variables tend to occur in isolation; for sequences of serially dependent variables, extremes can occur in clusters. Extreme-value theory is well developed for stationary processes and provides mathematical characterisations of the clustering of extremes. Such characterisations are useful models for the extremal behaviour of physical processes: consider storms, floods and droughts for example. Statistical applications harness extreme-value theory to make inferences about the extremes of a process based on a finite sample of data. This thesis addresses several topics in the analysis of clusters of extreme values from both univariate and multivariate processes. Key developments are the following: a theoretically justified scheme for identifying clusters in a sample; estimators for the extremal index that do not require clusters to be identified; a semi-parametric estimator for multivariate extreme-value densities; estimators for cluster summaries that exploit the asymptotic structure of clusters; and a method for modelling clusters in multivariate processes.

Acknowledgements

The work presented in this thesis was conducted under the expert supervision of Professor Jonathan Tawn. The generosity with which he gave of his time, ideas and encouragement is remarkable and it is a pleasure to record my enduring appreciation. I was fortunate also to benefit from a lengthy and very enjoyable collaboration with Dr Johan Segers, now at Tilburg University. I am grateful for the opportunities afforded me by the Engineering and Physical Sciences Research Council, which funded my period of study at Lancaster University, and for the understanding shown by the Department of Meteorology at Reading University, in particular Dr David Stephenson, who granted me time to complete this thesis. Many other friends and colleagues made their own, multifarious contributions; I hope that they will forgive me for mentioning only Omiros Papaspiliopoulos, for providing a timely courier service.

Contents

List of Figures vi

List of Tables ix

1 Introduction **1**

2 Background Theory & Methods I **4**

2.0 Introduction 4

2.1 Univariate Sequences 5

2.2 Multivariate Sequences 9

3 Background Theory & Methods II **16**

3.0 Introduction 16

3.1 Univariate Sequences 16

3.2 Multivariate Sequences 27

4 Declustering & the Extremal Index **31**

4.0 Introduction 31

4.1 Univariate Sequences 32

4.1.1 Inter-exceedance times 35

4.1.2 Moment estimators 41

4.1.3 Least-squares estimators 46

4.1.4 Maximum-likelihood estimators 48

4.1.5 Declustering and bootstrapping 51

4.1.6 Simulation study 53

4.1.7 Data example 58

4.2 Multivariate Sequences 65

4.A Proof of Theorem 4.1 70

4.B Proof of Theorem 4.2 72

4.C Proof of Theorem 4.3 76

4.D Proof of Theorem 4.4 77

5 Modelling Componentwise Maxima **84**

5.0 Introduction 84

5.1 Parametric Estimators 87

5.2 Non-parametric Estimators 89

5.3 Semi-parametric Estimators 91

5.3.1 Asymptotic motivation 92

5.3.2 Kernel estimator 94

5.3.3 Symmetric kernel estimator 97

5.3.4 Unknown component distributions 98

5.3.5 Asymptotic properties 99

5.4 Simulation Study 101

5.5 Data Example 107

5.6 Discussion 112

6 Estimating Cluster Functionals **114**

6.0 Introduction 114

6.1 Clusters & Strings 115

6.2 SWOP Estimators 118

6.3 SWAP Estimators 119

6.3.1 Modelling the string process 121

6.3.2 Modelling the expectation 124

6.4 Multiple Thresholds 128

6.5 Examples 130

6.5.1 Cluster size 130

6.5.2 Cluster excess 133

6.6 Simulation Study 136

6.7 Data Example 141

7 Inference for Multivariate Clusters **143**

7.0 Introduction 143

7.1 Estimating Failure Probabilities 144

7.2 Modelling Multivariate Clusters 147

7.2.1 Block maxima 147

7.2.2 Strings 148

7.2.3 Failure probabilities 150

7.3 Simulation Study 151

7.4 Data Example 154

7.5 Discussion 158

Bibliography 161

List of Figures

4.1 Blocks and runs estimates of the extremal index 34

4.2 Quantile-quantile plot of normalised inter-exceedance times 38

4.3 Max-autoregressive simulations for runs and intervals estimators of the extremal index 56

4.4 Markov chain simulations for runs and intervals estimators of the extremal index 57

4.5 Simulations for runs, intervals and least-squares estimators of the extremal index 59

4.6 Simulations for the indirect estimator of the cluster-size distribution 60

4.7 Simulations for the indirect, least-squares and direct estimators of the cluster-size distribution 61

4.8 Wooster temperature data 62

4.9 Extremal index estimates for the Wooster temperature data 63

4.10 Cluster-size distribution estimates for the Wooster temperature data 64

4.11 Cluster excess and run length estimates for the Wooster temperature data . . . 65

4.12 Estimates of the multivariate extremal index and cluster-size distribution . . . 70

5.1 Dependence functions of three logistic models . . . 102

5.2 Dependence functions of three Dirichlet models . . . 102

5.3 Sheerness-Lowestoft sea-level data . . . 108

5.4 Dependence function estimates for the Lowestoft-Sheerness data . . 110

5.5 Stability plots for the parameters of the kernel estimator . . . 111

6.1 Simulation results for standard, SWOP and SWAP estimators of the extremal index . . . 137

6.2 Simulation results for standard, SWOP and SWAP estimators of the cluster-size distribution . . . 138

6.3 Simulation results for standard, SWOP and SWAP estimators of the cluster excess . . . 139

6.4 Standard, SWOP and SWAP estimates for the Wooster temperature data . . . 142

7.1 Christchurch wave height and surge time series . . . 155

7.2 Christchurch wave height and surge scatter plot 156

7.3 Exploratory plots for surge 157

7.4 Exploratory plots for significant wave height 158

7.5 Diagnostic plots for surge and significant wave height 159

7.6 Block maxima for surge and significant wave height 160

List of Tables

5.1 Simulation results for estimators of A 105

5.2 Simulation results for estimators of A' 106

5.3 Simulation results for estimators of A'' 106

5.4 Simulation results for estimators of A, A' and A'' under independence 107

5.5 Parameter estimates for the Lowestoft-Sheerness data 109

7.1 Coefficients and cluster shapes for the five M_4 processes 152

7.2 Relative bias 154

7.3 Relative standard deviation 154

Chapter 1

Introduction

Extreme events can have powerful impacts on the world. Storms, floods and droughts are familiar environmental examples; consider also stock-market crashes and catastrophic engineering failures. An extreme event can often be characterised in several ways: as a single, extreme value of some quantity; as a cluster of large values occurring in quick succession; as a rare combination of two or more factors; or as a collection of large values occurring simultaneously at several locations. For example, a flood might be caused by rain from a single storm, by consecutive days of high rainfall, by a strong surge combined with high tide, or by high sea-levels at several sites along a coast. There is considerable interest in understanding and predicting the nature of extreme events such as these. With this insight, action can be taken to prevent extreme events, mitigate their effects, or plan for their impacts. Statistical analysis can help by estimating the rates at which extreme events occur, and their critical features.

But the statistical analysis of extremes is difficult. The amount of data available is usually small and estimates are often required at levels for which there is no data. For example, an estimate might be needed for the probability that a sea-wall is breached although the data record might contain no instance of this happening. A successful statistical analysis must use data efficiently and base extrapolations on justifiable theory. This is the goal of extreme-value theory.

This thesis aims to use extreme-value theory to develop statistical methods for making inferences about extreme events occurring in univariate and multivariate stationary processes. Extreme-value theory is well developed for stationary processes and a review of key results under serial independence and serial dependence is provided in Chapters 2 and 3. This work provides characterisations of extreme events under weak conditions on the underlying process. For example, clusters of large values occur according to Poisson processes in time, the maximum value in a cluster has a distribution that belongs to a certain class, and clusters are independent and identically distributed. For practical application, clusters must be identified in the data, and the distribution of the maximum and other features of clusters must be estimated. These are the principal topics discussed in this thesis.

Cluster identification, known as declustering, is currently a fairly arbitrary procedure that is not governed by any theoretical considerations. Moreover, uncertainty in the identification is not accounted for in any formal way. The declustering scheme introduced in Chapter 4 is justified by theoretical developments and supports a technique for estimating the effect of uncertainty on subsequent analyses.

The maximum value in a cluster from a univariate process has a distribution that belongs to a finite-parameter family, for which estimation is now standard. For multivariate processes, the largest value in a cluster, where 'largest' has a specific meaning, has a distribution that must satisfy certain criteria but does not belong to a finite-parameter family. Non-parametric estimators and parametric models are both in use, but neither approach is ideal: the former estimates are not differentiable; the latter are restricted. A semi-parametric estimator that is differentiable and flexible is introduced in Chapter 5.

Other features of clusters admit simple empirical estimates once clusters have been identified. These estimates rely on clusters being independent and identically distributed. Theoretical results highlight additional structure in clusters that is not exploited by these simple estimators, however, and estimators that do exploit this information are introduced in Chapter 6. An advantage of the new estimators is their ability to estimate features of a cluster at levels outwith the data. One application of this property is estimating the probability that a multivariate observation will fall in some rare 'failure region'. Such probabilities are currently estimated using the distribution of the cluster maxima, but this can lead to over-estimation. A method that avoids this bias is introduced in Chapter 7.

The theory and methods reviewed in Chapter 3 for stationary processes appear with related material and in more detail in Ferro and Segers (2003b). Most of Section 4.1 appears in Ferro and Segers (2003a).

Chapter 2

Background Theory & Methods I – Independent Sequences

2.0 Introduction

A precept of extreme-value theory is to employ as statistical models distributions that arise as limiting distributions for random variables associated with stochastic processes. These limits correspond to levels that are asymptotically extreme in some sense and, as such, justify their use as models for extreme values. The asymptotic distributions of two constructions are reviewed in this chapter for sequences of independent and identically distributed random variables: the maximum of a block of consecutive variables and the point process of consecutive variables. Univariate sequences are covered in Section 2.1; multivariate sequences are covered in Section 2.2. Statistical applications of the results are also demonstrated.

More complete introductions to extreme-value theory and methods are available in Leadbetter *et al.* (1983), Resnick (1987), Embrechts *et al.* (1997), Coles (2001) and Beirlant *et al.* (2003).

2.1 Univariate Sequences

Let $\{X_i\}_{i\geq 1}$ be a sequence of independent and identically distributed random variables with marginal distribution function F, that is each X_i is distributed according to F. Denote the maximum of n consecutive elements of the sequence by

$$M_n = \max\{X_1, \dots, X_n\}.$$

As n increases, M_n approaches the upper end-point, $\omega = \sup\{x : F(x) < 1\}$, of F and the limiting distribution of M_n is a point mass at ω. A normalisation is required to obtain a non-degenerate limit. As in the central limit theorem, a linear normalisation is traditional and the limit, as n approaches infinity, is sought for the distribution function

$$P\left(\frac{M_n - b_n}{a_n} \leq x\right) = F^n(a_n x + b_n)$$

for sequences of constants $a_n > 0$ and b_n. The following theorem, due to Fisher and Tippett (1928), characterises all of the possible limit distributions. Denote weak convergence by $\xrightarrow{w}$.

Theorem 2.1. *If there exist sequences of constants $a_n > 0$ and b_n such that*

$$P\left(\frac{M_n - b_n}{a_n} \leq x\right) \xrightarrow{w} G(x) \quad \textit{as } n \to \infty$$

for a non-degenerate distribution function G, then G is a generalised extreme-value (GEV) distribution function,

$$G(x) = \exp\left[-\left\{1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right\}^{-1/\xi}\right], \tag{2.1}$$

defined on $\{x : 1 + \xi(x-\mu)/\sigma > 0\}$ for parameters $-\infty < \mu < \infty$, $\sigma > 0$ and $-\infty < \xi < \infty$. For $\xi = 0$, G is defined by the limit as $\xi \to 0$.

The generalised extreme-value distribution function (2.1) comprises three sub-classes:

I	$\exp\left[-\exp\left\{-\left(\frac{x-\beta}{\alpha}\right)\right\}\right]$	for $x \in \mathbb{R}$,
II	$\exp\left\{-\left(\frac{x-\beta}{\alpha}\right)^{-\gamma}\right\}$	for $x > \beta$,
III	$\exp\left[-\left\{-\left(\frac{x-\beta}{\alpha}\right)\right\}^{\gamma}\right]$	for $x < \beta$,

where $\alpha > 0$ and $\gamma > 0$. These sub-classes correspond to the generalised extreme-value distribution with $\xi = 0$, $\xi > 0$ and $\xi < 0$, and are known by the names I – Gumbel, II – Fréchet and III – Weibull. 'Standard' versions of these distributions refer to the special case $\alpha = 1$, $\beta = 0$ and $\gamma = 1$.

The approximation $P(M_n \leq x) \approx G((x - b_n)/a_n)$ for large n motivates the generalised extreme-value distribution (2.1) as a model for maxima of blocks with large but finite lengths since the normalising constants can be assimilated into the location and scale parameters μ and σ. For example, the model could be fitted to a sample of annual maximum temperatures by maximising the likelihood. Although the extremes of non-stationary processes will not be treated in this thesis, any non-stationarity evident in the data can be accounted for by incorporating covariate information into the models. For example, μ could be replaced by $\alpha + \beta t$ to allow a linear trend in time t.

Block maxima contain only partial information about the extremes of a process; more detail is provided by other large observations, for which point processes yield useful limiting models. Let $\{X_i : i \in \mathcal{I}\}$ represent the locations of points, indexed by a set $\mathcal{I}$, occurring randomly in a space S. A point process Q counts the number of points in regions of S:

$$Q(A) = \sum_{i \in \mathcal{I}} \delta_{X_i}(A) \qquad A \subseteq S,$$

where $\delta_x(A) = 1$ if $x \in A$ and $\delta_x(A) = 0$ if $x \notin A$. The expected number of points in a set A is given by the intensity measure $\Lambda(A)$.

For $\{X_i\}_{i \geq 1}$ a sequence of independent and identically distributed random variables, define

$$Q_n(\cdot) = \sum_{i=1}^{n} \delta_{(i/n\,,\,(X_i - b_n)/a_n)}(\cdot).$$

This two-dimensional point process counts points on $(0,1] \times \mathbb{R}$. The following theorem, due to Pickands (1971), characterises the limit of Q_n, where the linear normalisation is used again to avoid degeneracy. Denote convergence in distribution by $\xrightarrow{d}$.

Theorem 2.2. *If there exist sequences of constants $a_n > 0$ and b_n such that*

$$P\left(\frac{M_n - b_n}{a_n} \leq x\right) \xrightarrow{w} G(x) \quad \text{as } n \to \infty$$

for a non-degenerate distribution function G with lower and upper end-points $\omega_0 = \inf\{x : G(x) > 0\}$ and $\omega_1 = \sup\{x : G(x) < 1\}$, then $Q_n \xrightarrow{d} Q$, where Q is a non-homogeneous Poisson process on $(0,1] \times (\omega_0, \omega_1)$ with intensity measure

$$\Lambda(A) = -(b-a)\log G(x) \tag{2.2}$$

on $A = (a,b) \times [x, \omega_1)$ when $0 \leq a < b \leq 1$ and $\omega_0 < x < \omega_1$.

The theorem motivates the Poisson process Q as a model for those points $(i/n, X_i)$ for which X_i exceeds a high threshold u. The model can be fitted by maximising the likelihood

$$L(\mu, \sigma, \xi) = \exp\{-\Lambda(A)\} \prod_{i:X_i>u} \lambda(i/n, X_i),$$

where $A = (0,1] \times [u, \omega_1)$ and

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

is the intensity associated with Λ at a point (t,x) in A.

Another useful model for threshold exceedances can be derived from Theorem 2.2

as follows.

$$\begin{aligned} P\left(\frac{X_1 - b_n}{a_n} > u + x \,\middle|\, \frac{X_1 - b_n}{a_n} > u\right) &\rightarrow \frac{\Lambda\{(0,1] \times (u+x, \omega_1)\}}{\Lambda\{(0,1] \times (u, \omega_1)\}} \\ &= \frac{\log G(u+x)}{\log G(u)} \\ &= \left(1 + \frac{\xi x}{\sigma_u}\right)^{-1/\xi}, \end{aligned}$$

where $\sigma_u = \sigma + \xi(u - \mu)$. This limit is called the generalised Pareto (GP) distribution (Pickands, 1975) and provides a model for the distribution of excesses over high thresholds:

$$P(X_1 - u > x \mid X_1 > u) \approx \left(1 + \frac{\xi x}{\sigma_u}\right)^{-1/\xi} \tag{2.3}$$

for large u on $\{x : x \geq 0 \text{ and } 1 + \xi x/\sigma_u > 0\}$.

2.2 Multivariate Sequences

Let $\{X_i\}_{i \geq 1}$ be a sequence of independent and identically distributed vector random variables $X_i = (X_{i1}, \ldots, X_{iD})$ with dimension D and marginal distribution function F. Many definitions are possible for the maximum of n consecutive elements of a multivariate sequence. The classical line in multivariate extremes (Sibuya, 1960) corresponds to the marginal ordering described by Barnett (1976). Another, conditional approach has been proposed recently by Heffernan and Tawn (2003). This thesis follows the classical line for which the maximum of n consecutive elements is defined to be the componentwise maximum,

$$M_n = (M_{n1}, \ldots, M_{nD}),$$

where $M_{nd} = \max\{X_{1d}, \ldots, X_{nd}\}$ for each $1 \leq d \leq D$. Non-degenerate limits are obtained for the distribution function

$$P\left(\frac{M_n - b_n}{a_n} \leq x\right) = F^n(a_n x + b_n),$$

where $x = (x_1, \ldots, x_D)$, $a_n = (a_{n1}, \ldots, a_{nD}) > 0$ and $b_n = (b_{n1}, \ldots, b_{nD})$. All operations are performed componentwise, so $a_n > 0$ is translated as $a_{n1} > 0$, ..., $a_{nD} > 0$. The following theorem, due to de Haan and Resnick (1977), characterises all of the possible limit distributions.

Theorem 2.3. *If there exist sequences of constants $a_n > 0$ and b_n such that*

$$P\left(\frac{M_n - b_n}{a_n} \leq x\right) \xrightarrow{w} G(x) \quad \text{as } n \to \infty \tag{2.4}$$

for a non-degenerate distribution function G, then each of the D one-dimensional component distributions of G is a generalised extreme-value distribution function,

$$G_d(x_d) = \exp\left[-\left\{1 + \xi_d\left(\frac{x_d - \mu_d}{\sigma_d}\right)\right\}^{-1/\xi_d}\right] \quad \text{for } 1 \leq d \leq D, \tag{2.5}$$

and $G(x) = G_*\{\zeta^{-1}(x)\}$, *where*

$$\zeta_d(x_d) = \mu_d + \frac{\sigma_d}{\xi_d}\left(x_d^{\xi_d} - 1\right) \quad \textit{for } 1 \leq d \leq D \tag{2.6}$$

and G_* *is a multivariate extreme-value distribution function with standard Fréchet components, that is*

$$G_*(x) = \exp\{-V(x)\} \tag{2.7}$$

for $x \in \mathbb{R}_+^D$, *where*

$$V(x) = \int_{S_D} \max_{1 \leq d \leq D} \left\{\frac{z_d}{x_d}\right\} \mathrm{d}H(z)$$

and H *is a non-negative, finite measure on the simplex*

$$S_D = \{z = (z_1, \ldots, z_D) \in \mathbb{R}_+^D \,:\, z_1 + \ldots + z_D = 1\}$$

for which

$$\int_{S_D} z_d \, \mathrm{d}H(z) = 1 \quad \textit{for } 1 \leq d \leq D. \tag{2.8}$$

Unlike the univariate case, the limiting distribution (2.4) is not determined by a finite number of parameters; rather, it is characterised by the $3D$ parameters of the component distributions (2.5) and the infinite-dimensional measure H. For the remainder of the thesis, it is assumed that the derivative, h, of H exists on the interior of S_D and on the interiors of all of the lower-dimensional boundaries, in which case the density of the multivariate extreme-value distribution (2.4) exists everywhere.

Estimation of multivariate extreme-value distributions, in particular when $D = 2$, will be discussed in Chapter 5. One option is to choose a parametric model for the measure H, or equivalently V, such that the conditions (2.8) are satisfied, and then fit the model to a sample of componentwise maxima by maximising the likelihood. An example is the logistic model (Gumbel, 1960), which in the bivariate case is

$$V(x_1, x_2) = \left(x_1^{-1/\alpha} + x_2^{-1/\alpha}\right)^{\alpha} \tag{2.9}$$

for dependence parameter $\alpha \in (0, 1]$. Non-parametric estimators also exist; a semi-parametric estimator is introduced in Chapter 5.

It is convenient to describe multivariate extreme-value distributions by separating the dependence structure (2.7) from the one-dimensional components (2.5). Assume, therefore, that the component distributions of F are standard Fréchet, that is $F_d(x_d) = \exp(-1/x_d)$ when $x_d > 0$ for $1 \leq d \leq D$. It follows that G_d is also standard Fréchet for $1 \leq d \leq D$; results for generalised extreme-value component distributions can be obtained by applying the change of variables (2.6). It is helpful also to introduce another coordinate system. For $x = (x_1, \ldots, x_D)$, define its radial component to be $r = x_1 + \ldots + x_D > 0$ and its angular component to be $w = (w_1, \ldots, w_D)$, where $w_d = x_d/r \in (0, 1)$.

The following theorem, due to de Haan (1985), characterises the limit of the point process

$$Q_n(\cdot) = \sum_{i=1}^{n} \delta_{(X_i - b_n)/a_n}(\cdot).$$

Theorem 2.4. *If there exist sequences of constants $a_n > 0$ and b_n such that*

$$P\left(\frac{M_n - b_n}{a_n} \leq x\right) \xrightarrow{w} G(x) \quad \text{as } n \to \infty \tag{2.10}$$

for a non-degenerate distribution function G with form (2.7), then $Q_n \xrightarrow{d} Q$, where Q is a non-homogeneous Poisson process on $\mathbb{R}_+^D \setminus \{0\}$ with intensity

$$\lambda(r, w) = r^{-2} h(w).$$

If the component distributions F_d are standard Fréchet then the normalising constants in the limit (2.10) can be set equal to $a_n = n$ and $b_n = 0$. In this case, the theorem provides a Poisson process model for points X_i/n on regions such as

$$A = \mathbb{R}_+^D \setminus \{[0, c_1) \times \ldots \times [0, c_D)\}$$

that are bounded away from the origin. If a parametric model such as the logistic (2.9) is chosen for the dependence structure then the Poisson process can be fitted by maximising the likelihood

$$\exp\{-\Lambda(A)\} \prod_{i : x_i/n \in A} \lambda(r_i, w_i),$$

where Λ is the intensity measure, $r_i = (x_{i1} + \ldots + x_{iD})/n$ and $w_{id} = x_{id}/(nr_i)$.

The Poisson process can also be fitted to points X_i on regions

$$A = \mathbb{R}_+^D \setminus \{[0, u_1) \times \ldots \times [0, u_D)\}, \tag{2.11}$$

where the u_d are high thresholds substituting for nc_d. Since $n^2\lambda(nr, w) = \lambda(r, w)$, the corresponding likelihood is proportional to

$$\exp\{-n\Lambda(A)\} \prod_{i:x_i \in A} \lambda(r_i, w_i),$$

where $r_i = x_{i1} + \ldots + x_{iD}$ and $w_{id} = x_{id}/r_i$.

This assumes that the component distributions F_d are standard Fréchet. When the F_d are unknown, the transformation $\tilde{x}_{id} = -1/\log F_d(x_{id})$ can be incorporated and the component distributions estimated simultaneously with the dependence structure. A model for each F_d is available if the generalised Pareto distribution (2.3) holds, that is

$$1 - F_d(x_d) = \{1 - F_d(u_d)\} \left\{1 + \xi_d \left(\frac{x_d - u_d}{\sigma_d}\right)\right\}^{-1/\xi_d}$$

when $x_d > u_d$. For $x_d \leq u_d$, F_d can be estimated by the empirical distribution function. Defining $\tilde{A} = \mathbb{R}_+^D \setminus \{[0, \tilde{u}_1) \times \cdots \times [0, \tilde{u}_D)\}$ and $\tilde{u}_d = -1/\log F_d(u_d)$, the likelihood becomes

$$\exp\{-n\Lambda(\tilde{A})\} \prod_{i:x_i \in A} \left\{\lambda(\tilde{r}_i, \tilde{w}_i) \prod_{d:x_{id} > u_d} \frac{\tilde{x}_{id}^2 (1 - e^{-1/\tilde{x}_{id}})^{1+\xi_d} e^{1/\tilde{x}_{id}}}{\sigma_d \{1 - F_d(u_d)\}^{\xi_d}}\right\}$$

since

$$\frac{\mathrm{d}\tilde{x}}{\mathrm{d}x} = \frac{\tilde{x}^2(1 - e^{-1/\tilde{x}})^{1+\xi} e^{1/\tilde{x}}}{\sigma\{1 - F(u)\}^{\xi}}.$$

The choice (2.11) of extreme region, A, is important here. For other choices the set of points that fall in the region can change with the parameter estimates of the component distributions. For further details on this model see Coles and Tawn (1991, 1994) and Joe *et al.* (1992).

Chapter 3

Background Theory & Methods

II – Dependent Sequences

3.0 Introduction

Serial independence was assumed throughout Chapter 2 but is rarely a valid assumption for applications. This chapter reviews key results for strongly stationary processes. The presence of serial dependence causes quantitative and qualitative changes in the behaviour of extremes.

3.1 Univariate Sequences

Let $\{X_i\}_{i\geq 1}$ be a stationary sequence of random variables with marginal distribution function F. Without any restriction on the strength of dependence, the limiting distribution of the block maximum M_n can remain obscure. For example, if $X_i = X_1$ almost surely for all $i \geq 1$ then the distribution of M_n is F. The serial dependence need be restricted at only extreme levels, however, to obtain a non-trivial limit. One such restriction is the following mixing condition introduced by Leadbetter (1983). Define $M(I) = \max\{X_i : i \in I\}$ and, for a sequence of thresholds u_n, let $I_{j,k}(u_n) = \{\{M(I) \leq u_n\} : I \subseteq \{j, \ldots, k\}\}$ be the set of all intersections of the events $\{X_i \leq u_n\}$, $j \leq i \leq k$.

Condition $D(u_n)$. For all $A \in I_{1,k}(u_n)$, $B \in I_{k+l,n}(u_n)$ and $1 \leq k \leq n-l$,

$$|P(A \cap B) - P(A)P(B)| \leq \alpha(n, l)$$

and $\alpha(n, l_n) \to 0$ as $n \to \infty$ for some $l_n = o(n)$.

Condition $D(u_n)$ ensures that any two events of the form $\{M(I_1) \leq u_n\}$ and $\{M(I_2) \leq u_n\}$ can become approximately independent as n increases when the index sets $I_i \subset \{1, \ldots, n\}$ are separated by a relatively short interval of length $l_n = o(n)$. Hence, $D(u_n)$ limits the long-range dependence between such events. It is a weak condition relative to many imposed in other areas of statistics. For example, $D(u_n)$ is satisfied for linear Gaussian processes if the correlation decays faster than the logarithm of the lag (Berman, 1964), and this is weaker than the geometric decay assumed by autoregressive models. With the restriction of $D(u_n)$, Leadbetter (1974) proves the following analogue of Theorem 2.1.

Theorem 3.1. *If there exist sequences of constants* $a_n > 0$ *and* b_n *such that*

$$P\left(\frac{M_n - b_n}{a_n} \leq x\right) \xrightarrow{w} G(x) \quad \text{as } n \to \infty$$

for a non-degenerate distribution function G*, and if* $D(u_n)$ *holds with* $u_n = a_n x + b_n$ *for each* x *such that* $G(x) > 0$*, then* G *is a generalised extreme-value distribution function* (2.1)*.*

The possible limit distributions for block maxima from stationary processes are therefore the same as in the independent case, and can be used as models in the same way. Dependence does influence the particular limit that is achieved, however, as made clear by the next theorem, due to Leadbetter (1983). Define $\{\hat{X}_i\}_{i \geq 1}$ to be the sequence of independent random variables also with marginal distribution F and let $\hat{M}_n = \max\{\hat{X}_1, \ldots, \hat{X}_n\}$.

Theorem 3.2. *If there exist sequences of constants* $a_n > 0$ *and* b_n *such that*

$$P\left(\frac{\hat{M}_n - b_n}{a_n} \leq x\right) \xrightarrow{w} G(x) \quad \text{as } n \to \infty$$

for a non-degenerate distribution function G*, if* $D(u_n)$ *holds with* $u_n = a_n x + b_n$ *for each* x *such that* $G(x) > 0$*, and if* $P\{(M_n - b_n)/a_n \leq x\}$ *converges for some* x *then*

$$P\left(\frac{M_n - b_n}{a_n} \leq x\right) \xrightarrow{w} G^{\theta}(x) \quad \text{as } n \to \infty \tag{3.1}$$

for some constant $\theta \in [0, 1]$*.*

Block maxima from stationary processes are therefore stochastically smaller than those from independent processes. The constant θ is called the extremal index and summarises the strength of dependence between extremes in a stationary sequence. It has several interpretations and will play a crucial role in Chapter 4.

The two-dimensional point process

$$Q_n(\cdot) = \sum_{i=1}^{n} \delta_{(i/n\,,\,(X_i - b_n)/a_n)}(\cdot)$$

considered in Theorem 2.2 for independent sequences has a limit for stationary processes too. The following mixing condition of Hsing (1987) is needed to ensure convergence. Let $\mathcal{F}_{j,k}(u_n\,;\,\tau_1,\ldots,\tau_s) = \sigma(\{X_i \le u_n(\tau_r)\} \,:\, j \le i \le k, 1 \le r \le s)$.

Condition $\Delta(u_n)$. For all integer $s \ge 1$, $\tau_1,\ldots,\tau_s > 0$, $A \in \mathcal{F}_{1,k}(u_n\,;\,\tau_1,\ldots,\tau_s)$, $B \in \mathcal{F}_{k+l,n}(u_n\,;\,\tau_1,\ldots,\tau_s)$ and $1 \le k \le n-l$,

$$|P(A \cap B) - P(A)P(B)| \le \alpha(n,l)$$

and $\alpha(n,l_n) \to 0$ as $n \to \infty$ for some $l_n = o(n)$.

This condition is stricter than $D(u_n)$ because the events required to become independent are more numerous and they are required to become independent at many levels simultaneously. The following result was proved by Hsing (1987).

Theorem 3.3. *If there exist sequences of constants $a_n > 0$ and b_n such that*

$$P\left(\frac{M_n - b_n}{a_n} \le x\right) \xrightarrow{w} G(x) \quad \textit{as } n \to \infty$$

for a non-degenerate distribution function G of generalised extreme-value form (2.1), if $\Delta(u_n)$ holds when $u_n(\tau) = a_n\tau + b_n$ and if $Q_n \xrightarrow{d} Q$, then Q is the point process

$$Q(\cdot) = \sum_{i=1}^{\infty}\sum_{j=1}^{K_i} \delta_{(S_i\,,\mu+\sigma\xi^{-1}\{(T_iY_{i,j})^{-\xi}-1\})}(\cdot), \tag{3.2}$$

where $\{(S_i, T_i)\}_{i\geq 1}$ are points of a homogeneous Poisson process, η, on $(0,1] \times (0,\infty)$ with Lebesgue measure, $\{Y_{i,j}\}_{j=1}^{K_i}$ are points of a point process γ_i on $[1,\infty)$ with an atom $Y_{i,1} = 1$ for each $i \geq 1$ and $\{\gamma_i\}_{i\geq 1}$ are independent, identically distributed and independent of η.

The limit process (3.2) has the following interpretation. For each i, the points $\{(S_i, T_{i,j})\}_{j=1}^{K_i}$, where

$$T_{i,j} = \mu + \sigma\xi^{-1}\{(T_iY_{i,j})^{-\xi} - 1\},$$

all occur at the same time, S_i. Also, $T_{i,1}$, which corresponds to the atom $Y_{i,1} = 1$, is the largest of the $T_{i,j}$ occurring at time S_i. This structure describes clusters of points occurring through time: K_i points occur at time S_i with cluster maximum $T_{i,1}$; the remaining points in the cluster are defined relative to $T_{i,1}$ by the random variables

$$Y_{i,j} = \frac{\left\{1+\xi\left(\frac{T_{i,j}-\mu}{\sigma}\right)\right\}^{-1/\xi}}{\left\{1+\xi\left(\frac{T_{i,1}-\mu}{\sigma}\right)\right\}^{-1/\xi}}, \quad 1 \leq j \leq K_i. \tag{3.3}$$

The collection $\{Y_{i,j}\}_{j=1}^{K_i}$ is called the string associated with the i-th cluster.

The cluster maxima, $\{(S_i, T_{i,1})\}_{i\geq 1}$, are points of a non-homogeneous Poisson process on $(0,1] \times (\omega_0, \omega_1)$ with intensity measure

$$\Lambda(A) = -(b-a)\log G(x) \tag{3.4}$$

on $A = (a,b) \times [x, \omega_1)$ when $0 \leq a < b \leq 1$ and $\omega_0 < x < \omega_1$. This is the same point process found in Theorem 2.2 for independent sequences except that the intensity measure (3.4) is reduced by a factor θ due to Theorem 3.2. For each i, the number of the remaining points in the cluster and their distribution relative to the cluster maximum is determined by the point process γ_i. The distribution of γ_i depends on the particular short-range dependence structure of the process and Mori (1977) shows that γ_i can be any point process. Nevertheless, the theorem does say that clusters are independent and the distribution of points within a cluster relative to the cluster maximum is the same for all clusters.

Statistical application of Theorem 3.3 will be discussed later in this section. Another point process is investigated first, which requires the following definitions. A marked point process Q counts for each point X_i in some space S a quantity Y_i, and has the form

$$Q(A) = \sum_{i \in \mathcal{I}} Y_i \, \delta_{X_i}(A), \quad A \subseteq S,$$

where the marks $\{Y_i : i \in \mathcal{I}\}$ are identically distributed. A compound Poisson process is a marked point process for which the points occur according to a Poisson process independently of the marks, which are themselves independent and identically distributed.

For the stationary sequence $\{X_i\}_{i\geq 1}$, the point process

$$Q'_n(\cdot) = \sum_{i\in\mathcal{I}} \delta_{i/n}(\cdot), \quad \mathcal{I} = \{i : X_i > u_n, 1 \leq i \leq n\}, \tag{3.5}$$

counts the times on $(0,1]$ at which a threshold u_n is exceeded. The limit of Q'_n is found by Hsing *et al.* (1988) under the following mixing condition. Let $\mathcal{F}'_{j,k}(u_n) = \sigma(\{X_i \leq u_n\} : j \leq i \leq k)$.

Condition $\Delta'(u_n)$. For all $A \in \mathcal{F}'_{1,k}(u_n)$, $B \in \mathcal{F}'_{k+l,n}(u_n)$ and $1 \leq k \leq n-l$,

$$|P(A\cap B) - P(A)P(B)| \leq \alpha'(n,l)$$

and $\alpha'(n,l_n) \to 0$ as $n \to \infty$ for some $l_n = o(n)$.

Define also the cluster-size distribution

$$\pi_n(j) = P\{N_{r_n}(u_n) = j \mid N_{r_n}(u_n) > 0\} \quad \text{for integer } j \geq 1, \tag{3.6}$$

where $N_r(u) = \sum_{i=1}^{r} I(X_i > u)$ is the number of exceedances in a block of length r, and denote by $\lfloor x \rfloor$ the integer part of x.

Theorem 3.4. *If there exist sequences of constants $a_n > 0$ and b_n such that, for some x,*

$$P\left(\frac{M_n - b_n}{a_n} \leq x\right) \to G(x) > 0 \quad \textit{as } n \to \infty$$

and if, for $u_n = a_n x + b_n$, $\Delta'(u_n)$ holds, there exist sequences of constants k_n and l_n such that $k_n \to \infty$, $k_n l_n = o(n)$ and $k_n \alpha'(n,l_n) \to 0$, and there exists a distribution

π such that, when $r_n = \lfloor n/k_n \rfloor$, the cluster-size distribution $\pi_n(j) \to \pi(j)$ for all positive integers j, then $Q'_n \xrightarrow{d} Q'$, where Q' is a compound Poisson process on $(0,1]$ with mark distribution π and intensity measure

$$\Lambda(A) = -(b-a)\log G(x)$$

on $A = (a,b)$ when $0 \le a < b \le 1$.

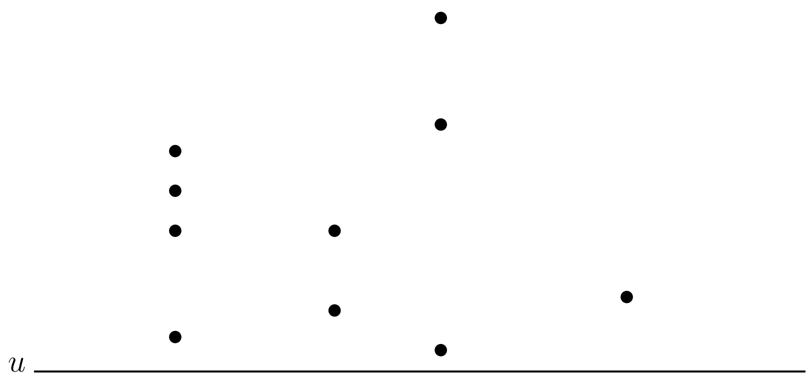
The x in the intensity measure of Q' is the particular, fixed value used to define the threshold u_n. Therefore, the result states that, in the limit, the threshold is exceeded at times that occur according to a homogeneous Poisson process, and that each event comprises multiple exceedances distributed according to the limiting cluster-size distribution, π. This agrees with the structure of the limiting two-dimensional process in Theorem 3.3.

Leadbetter (1983) notes the following connection between the cluster-size distribution and the extremal index θ:

$$\theta^{-1} = \lim_{n\to\infty} \sum_{j=1}^{\infty} j\pi_n(j), \tag{3.7}$$

that is, the extremal index is the reciprocal of the limiting mean cluster size. To see this, recall that the intensity (3.4) of clusters is smaller by a factor θ than the intensity (2.2) of exceedances in independent sequences. Since the expected number of exceedances is unaffected by the strength of the dependence, an average of $1/\theta$ exceedances must occur in each cluster.

The previous two theorems have described the asymptotic clustering behaviour of stationary processes. The following picture gives an idealised view of a stationary sequence over a high but finite threshold u.



The four clusters in the diagram might represent storms, heat waves or other extreme events. Certain features of a cluster, such as the number of exceedances, may be of particular interest. In general, cluster characteristics can be written as

$$c\{(X_i - u) : i \in \mathcal{S}\}, \quad \mathcal{S} \subset \mathbb{N},$$

for a set $\{X_i : i \in \mathcal{S}\}$ of threshold exceedances. Examples of cluster characteristics are

$$\begin{array}{ll} \text{Cluster Size} & \sum_{i \in \mathcal{S}} I\{(X_i - u) > 0\}, \\ \text{Cluster Excess} & \sum_{i \in \mathcal{S}} (X_i - u), \\ \text{Maximum Excess} & \max_{i \in \mathcal{S}} (X_i - u). \end{array} \tag{3.8}$$

The first step in making inferences about such functionals is identifying clusters in the data. Declustering has two stages: identifying those observations regarded as extreme, and grouping those observations into clusters. In the univariate case, the first stage is completed by choosing a threshold u above which the limiting behaviour described in Theorem 3.3 is believed to be a good approximation. Threshold choice involves a trade-off between bias and variance: if the threshold is too low then limiting forms may be poor models; if the threshold is too high then there will be few observations with which to make inferences. Often, if a limiting form holds at one threshold then it holds also at all higher thresholds. Therefore, one strategy is to try several thresholds before selecting u to be the lowest in the range of thresholds for which estimates appear stable and models fit the data reasonably well. Objective threshold selection procedures are described by Guillou and Hall (2001) among others; a different approach, which requires the specification of a model for the data below the threshold, has been developed by Frigessi *et al.* (2002). The second stage of declustering is not trivial: threshold exceedances do not occur at single points in time except in the limit; instead, they are slightly spread out and it may not be obvious if a group of exceedances should form one cluster or be split into two clusters. This problem is the topic of Chapter 4.

Suppose that C clusters have been identified above u, where the i-th cluster comprises exceedances $\mathcal{C}_i = \{X_j : j \in \mathcal{S}_i\}$ in some set $\mathcal{S}_i \subseteq \{1, \ldots, n\}$. Theorem 3.3 provides a model for these clusters (Leadbetter, 1991). According to the model, the cluster maxima can be thought of as points of a non-homogeneous Poisson process with intensity measure (3.4). This is the same model described in Chapter 2 and can be estimated by maximum likelihood. The theorem also justifies the assumption that clusters are independent and identically distributed. Inference for cluster functionals $c(\cdot)$ can therefore be made using the empirical values $c\{(X_j - u) : j \in \mathcal{S}_i\}$, $1 \leq i \leq C$. Estimators for cluster functionals that exploit more of the structure described by Theorem 3.3 are introduced in Chapter 6.

The statistical methods described so far have relied on point process results for quite general stochastic processes. Another approach for analysing extremes is to select a time-series model for the evolution of the process at extreme levels. For example, Smith (1992) and Smith *et al.* (1997) consider first-order Markov chains with bivariate extreme-value distributions, such as the logistic model (2.9), for consecutive elements (X_i, X_{i+1}). Such models can be fitted to threshold exceedances and then asymptotic properties of the fitted model found by simulation. Simpler models include the moving-maximum process,

$$X_i = \max_{j \geq 0}(\alpha_j Z_{i-j}) \quad \text{for } i \geq 1, \tag{3.9}$$

where the constants, $\alpha_j \geq 0$, satisfy $\sum_{j \geq 0} \alpha_j = 1$ and the Z_i are independent, standard Fréchet random variables. These processes were developed by Newell (1964) and Deheuvels (1983, 1985). A special case of the moving-maximum process, when $\alpha_j = (1-\alpha)\alpha^j$ for $0 \leq \alpha < 1$, is the max-autoregressive process (Leadbetter, 1983) defined recursively by

$$X_i = \max\{\alpha X_{i-1}, (1-\alpha)Z_i\} \quad \text{for } i \geq 1. \tag{3.10}$$

The developments in subsequent chapters will avoid making assumptions about the serial structure in the underlying process, preferring to make inferences free of any models not justified by asymptotic theory. The time-series models listed above, for which extremal properties can be obtained analytically or by simulation, will be used instead in simulation studies for assessing the performance of new estimators.

3.2 Multivariate Sequences

Let $\{X_i\}_{i\geq 1}$ be a stationary sequence of D-dimensional random variables $X_i = (X_{i1}, \ldots, X_{iD})$ with marginal distribution function F. Just as Theorem 2.1 has an analogue for stationary processes under a suitable mixing condition, so there is an extension of Theorem 2.3 for componentwise maxima. An appropriate multivariate mixing condition is introduced by Hsing (1989) and is precisely the same as the $D(u_n)$ condition except that the thresholds, u_n, are vectors. Hsing (1989) also proves the following result.

Theorem 3.5. *If there exist sequences of constants $a_n > 0$ and b_n such that*

$$P\left(\frac{M_n - b_n}{a_n} \leq x\right) \xrightarrow{w} G(x) \quad \textit{as } n \to \infty$$

for a non-degenerate distribution function G and if the multivariate version of $D(u_n)$ holds with $u_n = a_n x + b_n$ for each x such that $G(x) > 0$, then G is a multivariate extreme-value distribution function (2.4).

Inference for componentwise maxima from stationary processes can therefore be conducted with the same models used in the independence case. Dependence again influences the particular limit achieved, and Nandagopalan (1994) defines a multivariate extremal index in direct analogy with Theorem 3.2 as follows. A stationary sequence has multivariate extremal index $\theta(\cdot) \in [0,1]$ if there exist sequences of constants $a_n > 0$ and b_n such that, as $n \to \infty$,

$$P\left(\frac{\hat{M}_n - b_n}{a_n} \leq x\right) \xrightarrow{w} G(x) \quad \text{and} \quad P\left(\frac{M_n - b_n}{a_n} \leq x\right) \xrightarrow{w} G^{\theta(x)}(x).$$

Unlike its univariate counterpart, the multivariate extremal index depends on the point at which the distribution function is evaluated.

A multivariate extension of the point process (3.5) of exceedance times is

$$Q_n'(\cdot) = \sum_{i \in \mathcal{I}} (I(X_{i1} > u_{n1}), \ldots, I(X_{iD} > u_{nD}))\delta_{i/n}(\cdot), \quad \mathcal{I} = \{i : X_i \not\leq u_n\}.$$

This marked point process has marks on $\{0,1\}^D$ and counts points for which X_i exceeds u_n in at least one component. Define also a multivariate extension of the cluster-size distribution (3.6) as

$$\pi_n(j) = P\{\, N_{r_n}(u_n) = j \,|\, N_{r_n}(u_n) \neq 0\} \quad \text{for } j \in \mathbb{N}_0^D \setminus \{0\},$$

where $N_r(u) = (\sum_{i=1}^r I(X_{i1} > u_1), \ldots, \sum_{i=1}^r I(X_{iD} > u_D))$ and $\mathbb{N}_0$ is the set of non-negative integers. The limit of Q_n' is found by Nandagopalan (1994) under the multivariate version of condition $\Delta'(u_n)$, where, again, u_n is a vector.

Theorem 3.6. *If there exist sequences of constants $a_n > 0$ and b_n such that, for* *some x,*

$$P\left(\frac{M_n - b_n}{a_n} \leq x\right) \to G(x) > 0 \quad \textit{as } n \to \infty,$$

and if, for $u_n = a_n x + b_n$, $\Delta'(u_n)$ holds, there exist sequences of constants k_n and l_n such that $k_n \to \infty$, $k_n l_n = o(n)$ and $k_n \alpha'(n, l_n) \to 0$, and there exists a distribution π such that, when $r_n = \lfloor n/k_n \rfloor$, the multivariate cluster-size distribution $\pi_n(j) \to \pi(j)$ for all $j \in \mathbb{N}_0^D \setminus \{0\}$, then $Q'_n \xrightarrow{d} Q'$, where Q' is a compound Poisson process on $(0, 1]$ with mark distribution π and intensity measure

$$\Lambda(A) = -(b - a) \log G(x)$$

on $A = (a, b)$ when $0 \leq a < b \leq 1$.

The result shows how clusters (of observations extreme in at least one component) occur according to a homogeneous Poisson process in time and that each event comprises a certain number of exceedances in each component distributed according to the limiting multivariate cluster-size distribution, π.

Estimation for features of multivariate clusters has hardly been discussed in the literature. As in the univariate case, the first step is to decluster the data, but this is more complicated here because a single threshold is an insufficient demarcation. Multivariate declustering will be addressed in Chapter 4. Once clusters have been identified, however, the foregoing result justifies the assumption that they are independent. A sequence comprising single observations from each cluster is therefore an independent sequence of multivariate observations that can be modelled on regions bounded away from the origin by the Poisson process of Theorem 2.4. This is the model used by Coles and Tawn (1994), where they summarise each cluster by its componentwise maximum. A procedure for making inferences about multivariate clusters that removes the need for such summaries is introduced in Chapter 7.

Time-series models also exist for multivariate extremes. A multivariate extension of the moving-maximum process (3.9) called the M_4 process was introduced by Smith and Weissman (1996) and is defined by

$$X_{id} = \max_{l \geq 1} \max_{j \geq 0} (\alpha_{ljd} Z_{l,i-j}) \quad \text{for } i \geq 1 \text{ and } 1 \leq d \leq D, \tag{3.11}$$

where the constants, $\alpha_{ljd} \geq 0$, satisfy $\sum_{l \geq 1} \sum_{j \geq 0} \alpha_{ljd} = 1$ for $1 \leq d \leq D$ and the $Z_{l,i}$ are independent, standard Fréchet random variables. Multivariate Markov chains with extreme-value transition distributions have been studied by Perfekt (1997) but statistical methods have yet to be developed.

Chapter 4

Declustering & the Extremal Index

4.0 Introduction

In stationary sequences, extreme values can occur in clusters, for which asymptotic characterisations and methods of inference were reviewed in Chapter 3. The first step in making inferences is identifying clusters in the data, a procedure known as declustering. Declustering is discussed in this chapter, first for univariate sequences in Section 4.1. The unsatisfactory nature of declustering schemes currently in use is noted and an alternative that does not suffer from the same failings is proposed. The new scheme relies on a novel interpretation of the extremal index (3.1) as a measure of over-dispersion of times between extreme observations. This also leads to new estimators for the extremal index and the cluster-size distribution

(3.6) that, unlike other estimators, do not require data to be declustered first. A natural extension of these ideas to address declustering in multivariate sequences is presented in Section 4.2.

4.1 Univariate Sequences

Let $\{X_i\}_{i=1}^n$ be a sample from a stationary sequence of random variables with marginal distribution function F, survival function $\bar{F} = 1 - F$ and upper end-point $\omega = \sup\{x : F(x) < 1\}$. The aim is to identify within this sample approximately independent clusters of extreme observations. First a definition is required for 'extreme'. In the univariate case, this typically corresponds to an observation exceeding a chosen threshold u. All that remains is to group the exceedances into clusters.

Two popular schemes are blocks and runs declustering. Blocks declustering consists in choosing a block length, b, and partitioning the sequence $\{1, \ldots, n\}$ into $k = \lfloor n/b \rfloor$ blocks, $\{(i-1)b + 1, \ldots, ib\}$ for $1 \leq i \leq k$. The scheme stipulates that any extreme observations lying within the same block belong to the same cluster. Runs declustering consists in choosing a run length, r, and stipulates that any extreme observations separated by fewer than r non-extreme observations belong to the same cluster.

An application of these two declustering schemes is to estimation of the extremal index (3.1). The relationship (3.7) shows that the extremal index can be interpreted as the reciprocal of the mean number of exceedances in a cluster. Therefore, an estimator is the ratio of the number of clusters found by a declustering scheme to the total number of extreme observations in the sequence. Let

$$M_{i,j} = \max\{X_{i+1}, \ldots, X_j\}.$$

The estimators for the extremal index that correspond to the blocks and runs declustering schemes are

$$\bar{\theta}_n^B(u\,;\,b) = \frac{\sum_{i=1}^{k} I\{M_{(i-1)b,ib} > u\}}{\sum_{i=1}^{kb} I(X_i > u)}$$

and

$$\bar{\theta}_n^R(u\,;\,r) = \frac{\sum_{i=1}^{n} I(X_i > u, M_{i,i+r} \leq u)}{\sum_{i=1}^{n} I(X_i > u)}, \tag{4.1}$$

both of which are described by Smith and Weissman (1994). Hsing (1991) points out that a problem with these estimators, intrinsic to the nature of the declustering schemes involved, is that selection of values for the declustering parameters, b and r, is largely arbitrary. Furthermore, the choice can have a significant impact on the estimate of the extremal index, as illustrated in Figure 4.1.

Estimators for other cluster functionals (3.8) that are based on these declustering schemes are similarly sensitive to the declustering parameter. One example is the runs estimator for the cluster-size distribution, which estimates $\pi(j)$ by the ratio of the number of clusters with j exceedances to the total number of clusters.

Some attempts have been made to guide the choice of declustering parameter, for example Ledford and Tawn (2003), but the issue remains problematic. In Sec-

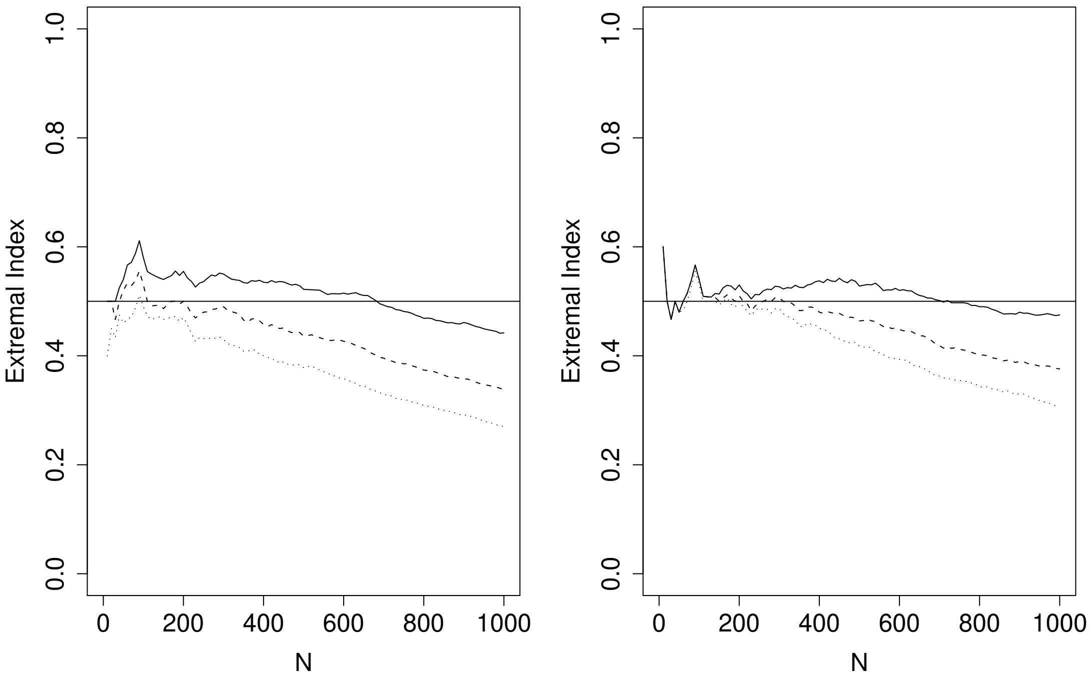


Figure 4.1: Estimates of the extremal index against number of threshold exceedances for a max-autoregressive process of length 10000 with extremal index $\theta = 0.5$. Blocks estimates (left) with $b = 10$ (——), 20 (- - -) and 30 ($\cdots$), and runs estimates (right) with $r = 1$ (——), 5 (- - -) and 9 ($\cdots$).

tion 4.1.2, new estimators for the extremal index and the cluster-size distribution are introduced that, unlike the runs, blocks and other estimators in the literature, do not depend on a declustering scheme. Two more estimation approaches are discussed in Sections 4.1.3 and 4.1.4. It is shown in Section 4.1.5 how the new estimators for the extremal index lead to a declustering scheme that is completely prescribed by the data. The performance of the estimators is evaluated with a simulation study in Section 4.1.6 and a data example in Section 4.1.7. The estimators are motivated by limit results for distributions of times between threshold exceedances, presented in the following section.

4.1.1 Inter-exceedance times

Let $T(u)$ be a random variable representing the time between consecutive exceedances of a threshold u by the process $\{X_i\}_{i\geq 1}$, that is

$$T(u) \stackrel{d}{=} \min\{n \geq 1 : \; X_{n+1} > u\} \text{ given } X_1 > u.$$

The distribution of the inter-exceedance time $T(u)$ is defined by

$$P\{T(u) > n\} = P(M_{1,n+1} \leq u \mid X_1 > u) \quad \text{for } n \geq 1.$$

The asymptotic distribution of $T(u)$ can be found easily when the sequence $\{X_i\}_{i\geq 1}$ is independent. In that case

$$P\{T(u) > n\} = F(u)^n \quad \text{for } n \geq 1,$$

so that, for $t > 0$,

$$P\{\bar{F}(u)T(u) > t\} = P\{T(u) > \lfloor t/\bar{F}(u) \rfloor\} = \exp\{\lfloor t/\bar{F}(u) \rfloor \log F(u)\}.$$

If F has no atom at its upper end-point ω then, since $\log(1+\epsilon) \sim \epsilon$ as $\epsilon \to 0$,

$$\lim_{u\to\omega} P\{\bar{F}(u)T(u) > t\} = \exp(-t) \quad \text{for } t > 0.$$

So $\bar{F}(u)T(u)$ is asymptotically standard exponentially distributed. This agrees with the result (Theorem 2.2) that the point process of exceedances has a Poisson process limit.

Now consider the general case where the sequence $\{X_i\}_{i\geq 1}$ has extremal index $\theta \in [0,1]$. The corresponding point-process limit for the exceedance times is compound Poisson (Theorem 3.4). This anticipates that the limit distribution of the inter-exceedance times will be a mixture of an exponential distribution and a point-mass on zero. This is indeed the case, as described in Theorem 4.1 below under the following mixing condition, which is stated in slightly more general terms than those introduced in Chapter 3. Let $\mathcal{F}^*_{j,k}(u) = \sigma(\{X_i \leq u\} : j \leq i \leq k)$.

Condition $\Delta^*(u_n)$. For all $A \in \mathcal{F}^*_{1,k}(u_n)$ with $P(A) > 0$, $B \in \mathcal{F}^*_{k+l,cr_n}(u_n)$ and $1 \leq k \leq cr_n - l$,

$$|P(B \mid A) - P(B)| \leq \alpha^*(cr_n, l) \quad \text{for each } c > 0$$

and there exists $l_n = o(r_n)$ such that $\alpha^*(cr_n, l_n) \to 0$ as $n \to \infty$ for all $c > 0$.

Theorem 4.1. *Let the positive integers r_n, $n \geq 1$, and the thresholds u_n, $n \geq 1$, be such that*

$$r_n \to \infty, \qquad r_n \bar{F}(u_n) \to \tau, \quad \textit{and} \quad P(M_{r_n} \leq u_n) \to \exp(-\theta\tau),$$

for some $\tau \in (0, \infty)$ and $\theta \in [0,1]$. If $\Delta^*(u_n)$ holds, then

$$P\{\bar{F}(u_n)T(u_n) > t\} \to \theta \exp(-\theta t) \quad \textit{for } t > 0$$

as $n \to \infty$.

A proof is contained in Appendix 4.A. Note that the mixing condition used is similar to that of Weissman and Novak (1998).

Theorem 4.1 says that if $\{X_i\}_{i\geq 1}$ has extremal index θ then

$$\bar{F}(u)T(u) \xrightarrow{d} T_\theta \quad \text{as } u \to \omega,$$

where T_θ is a random variable distributed according to the mixture distribution

$$(1-\theta)\epsilon_0 + \theta\mu_\theta, \tag{4.2}$$

ϵ_0 is the degenerate probability distribution at zero and μ_θ is the exponential distribution with mean $1/\theta$. Note the dual role of the extremal index: θ is both the proportion of non-zero inter-exceedance times and the reciprocal of the mean of the non-zero inter-exceedance times. These features are evident from quantile-quantile plots such as that in Figure 4.2.

In subsequent sections estimators for the extremal index are obtained by fitting the limiting model (4.2) to observed inter-exceedance times. First, however, the joint distributions of inter-exceedance times are investigated. Recall that the reciprocal of the extremal index is the limit of the mean of the cluster-size distribution (3.6). Therefore, Theorem 4.1 shows that the marginal distribution of inter-exceedance times is informative about the mean of the cluster-size distribution. In the remainder of this section, it is shown that limiting joint distributions of inter-exceedance

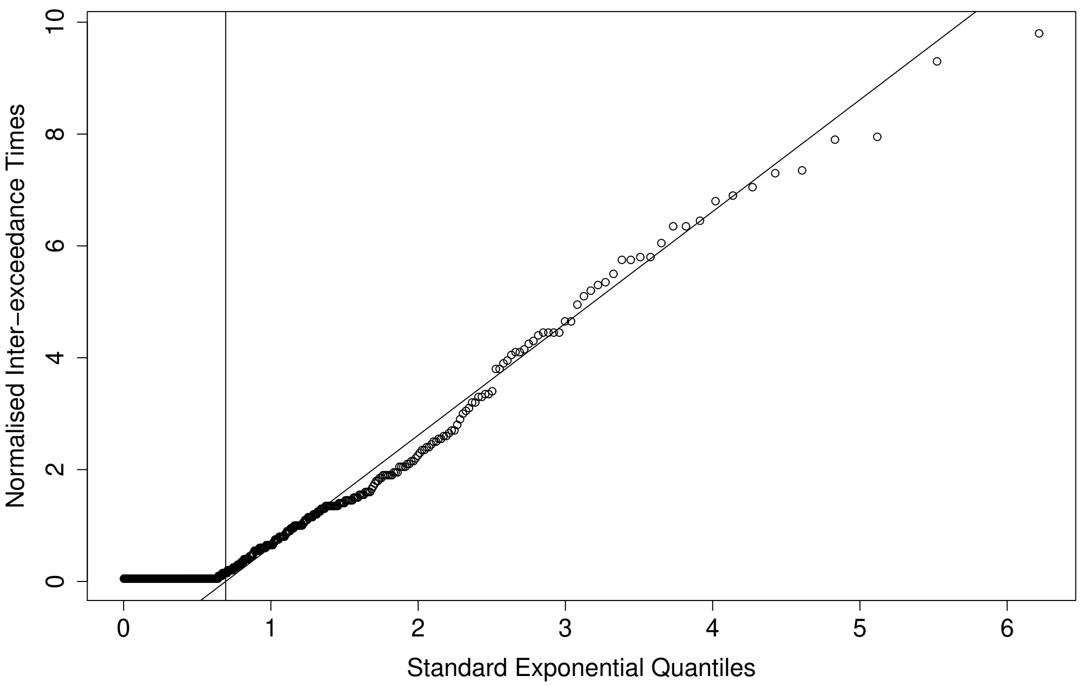


Figure 4.2: Quantile-quantile plot of 500 normalised inter-exceedance times against standard exponential quantiles for a max-autoregressive process of length 10000 with extremal index $\theta = 0.5$. The vertical line indicates the $(1-\theta)$-quantile and the sloping line has gradient $1/\theta$.

times are characterised by the limiting cluster-size distribution itself.

Let $T_1(u)$ and $T_{1+j}(u)$ be random variables representing inter-exceedance times separated by $j-1$ other inter-exceedance times. For example, the joint distribution of $T_1(u)$ and $T_2(u)$ is

$$P\{T_1(u) > s, T_2(u) > t\} = \sum_{r>s} P(M_{1,r} \leq u, X_{r+1} > u, M_{r+1,r+1+t} \leq u \mid X_1 > u)$$

for positive integers s and t.

The limit of this joint distribution will be computed under a stronger mixing condition than Theorem 4.1, namely m-dependence. For a positive integer m, the process $\{X_i\}_{i\geq 1}$ is m-dependent if, for all positive integers k, $\sigma(X_i : 1 \leq i \leq k)$ and $\sigma(X_i : i \geq k+m)$ are independent.

Theorem 4.2. *Let the thresholds u_n, $n \geq 1$, be such that*

$$n\bar{F}(u_n) \to \tau \quad \textit{and} \quad P(M_n \leq u_n) \to \exp(-\theta\tau)$$

for some $\tau \in (0, \infty)$ and $\theta \in (0, 1]$. Let also $r_n = o(n)$ be such that $\pi_n(1) \to \pi(1)$, where

$$\pi_n(1) = P\{N_{r_n}(u_n) = 1 \mid N_{r_n}(u_n) > 0\}$$

and $N_r(u) = \sum_{i=1}^{r} I(X_i > u)$. If $\{X_i\}_{i\geq 1}$ is m-dependent then

$$P\{\bar{F}(u_n)T_1(u_n) > s, \bar{F}(u_n)T_2(u_n) > t\} \to \theta\pi(1)e^{-\theta(s+t)} \quad \textit{for } s, t > 0.$$

A proof is presented in Appendix 4.B. Theorem 4.2 says that

$$\left(\bar{F}(u)T_1(u), \bar{F}(u)T_2(u)\right) \xrightarrow{d} (\tilde{T}_1, \tilde{T}_2) \quad \text{as } u \to \omega,$$

where

$$P(\tilde{T}_1 > s, \tilde{T}_2 > t) = \theta\pi(1)e^{-\theta(s+t)} \quad \text{for } s, t \geq 0.$$

This distribution has mass $1-\theta$ on each of the axes $\tilde{T}_1 = 0$ and $\tilde{T}_2 = 0$, and mass $1 - \theta\{2 - \pi(1)\}$ at $(0, 0)$.

The following result extends Theorem 4.2 to arbitrary lag.

Theorem 4.3. *Let the thresholds u_n, $n \geq 1$, be such that*

$$n\bar{F}(u_n) \to \tau \quad \textit{and} \quad P(M_n \leq u_n) \to \exp(-\theta\tau)$$

for some $\tau \in (0, \infty)$ and $\theta \in (0, 1]$. Let also $r_n = o(n)$ be such that $\pi_n(j) \to \pi(j)$ for all positive integers j. If $\{X_i\}_{i \geq 1}$ is m-dependent then

$$\left(\bar{F}(u_n)T_1(u_n), \bar{F}(u_n)T_{1+j}(u_n)\right) \xrightarrow{d} (\tilde{T}_1, \tilde{T}_{1+j}) \quad \textit{as } n \to \infty,$$

for all positive integers j, where

$$P(\tilde{T}_1 > s, \tilde{T}_{1+j} > t) = \theta e_j e^{-\theta(s+t)} \quad \textit{for } s, t \geq 0 \tag{4.3}$$

and

$$\begin{aligned} e_1 &= \pi(1), \\ e_j &= \pi(1)e_{j-1} + \ldots + \pi(j-1)e_1 + \pi(j) \quad \textit{for } j \geq 2. \end{aligned} \tag{4.4}$$

The proof of this result is discussed in Appendix 4.C.

The results of this section have shown that the limiting distributions (4.2) and (4.3) of normalised inter-exceedance times are characterised by the extremal index and the cluster-size distribution. In the following section, novel, moment-based estimators are presented for these quantities that do not require declustering, but are justified by the limiting distributions. In Sections 4.1.3 and 4.1.4, least-squares and maximum-likelihood estimators are discussed.

4.1.2 Moment estimators

Estimators for the extremal index and the cluster-size distribution can be constructed by equating theoretical moments of the limiting distributions (4.2) and (4.3) to their empirical counterparts.

Extremal index

For a high threshold u, let

$$N = N_n(u) = \sum_{i=1}^{n} I(X_i > u)$$

be the number of observations exceeding u and let $1 \leq S_1 < \ldots < S_N \leq n$ be the exceedance times. The inter-exceedance times are

$$T_i = S_{i+1} - S_i, \quad \text{for } 1 \leq i \leq N-1.$$

The second moment of T_θ with distribution (4.2) is $E(T_\theta^2) = 2/\theta$, which can be estimated to obtain a first estimator for θ. Let $\bar{F}_n(u)$ be an estimator for $\bar{F}(u)$. Then

$$\bar{\theta}_n(u) = \frac{2(N-1)}{\bar{F}_n(u)^2 \sum_{i=1}^{N-1} T_i^2}.$$

For example, if $\bar{F}_n(u) = N/n$ then

$$\bar{\theta}_n(u) = \frac{2n^2(N-1)}{N^2 \sum_{i=1}^{N-1} T_i^2}.$$

This can be improved. The first moment of T_θ is one so that θ is related to the coefficient of variation, ν, of the inter-exceedance times by

$$1 + \nu^2 = E(T_\theta^2)/E(T_\theta)^2 = 2/\theta.$$

In particular, the exceedance times are over-dispersed, with clustering in the limit, if and only if the extremal index is less than one. This relationship motivates another estimator for θ,

$$\hat{\theta}_n(u) = \frac{2\left(\sum_{i=1}^{N-1} T_i\right)^2}{(N-1)\sum_{i=1}^{N-1} T_i^2}, \tag{4.5}$$

where $\bar{F}(u)$ need not be estimated.

A further improvement is possible by considering a penultimate approximation to the limiting mixture distribution (4.2). Setting $r_n = n$ in the proof of Theorem 4.1 (see Appendix 4.A), the distribution of the inter-exceedance times can be seen to satisfy

$$P\{T(u_n) > n\} = \theta F(u_n)^{\theta n} + o(1).$$

An estimator for θ can be based on this relationship. Let T denote a random variable on the positive integers whose distribution is given by

$$P(T > n) = \theta p^{\theta n} \quad \text{for } n \geq 1, \tag{4.6}$$

where $\theta \in (0, 1]$ and $p \in (0, 1)$ can be thought of as $F(u_n)$. Note that

$$\begin{aligned}\frac{2E(T)^2}{E(T^2)} &= \frac{2\{1-(1-\theta)p^\theta\}^2}{2\theta p^\theta + \theta p^\theta(1-p^\theta) + (1-p^\theta)^2} \\ &= \theta + \theta(2 - 3\theta/2)(1-p) + O\left\{(1-p)^2\right\} \quad \text{as } p \to 1,\end{aligned}$$

which follows from

$$\begin{aligned}E(T-1) &= \sum_{n=1}^{\infty} P(T > n) = \theta p^\theta (1-p^\theta)^{-1}, \\ E\left\{\frac{T(T-1)}{2}\right\} &= \sum_{n=1}^{\infty} nP(T > n) = \theta p^\theta (1-p^\theta)^{-2}.\end{aligned}$$

Therefore, the first-order bias of $\hat{\theta}_n(u)$ is approximately $\theta(2 - 3\theta/2)\bar{F}(u)$. In contrast, the relationship

$$\frac{2\{E(T-1)\}^2}{E\{(T-1)(T-2)\}} = \theta \tag{4.7}$$

motivates the estimator

$$\hat{\theta}_n^*(u) = \frac{2\left\{\sum_{i=1}^{N-1}(T_i - 1)\right\}^2}{(N-1)\sum_{i=1}^{N-1}(T_i-1)(T_i-2)}, \tag{4.8}$$

whose first-order bias is approximately zero. Compared to estimator (4.5), the estimators for the first and second moments from which $\hat{\theta}_n^*(u)$ is constructed have been shrunk towards zero. This is intuitively appealing. The smallest observed inter-exceedance times are positive, whereas the limiting distribution (4.2) models them as zero. Estimator $\hat{\theta}_n^*(u)$ ensures that the contributions from the smallest inter-exceedance times are indeed zero; the larger inter-exceedance times are relatively unaffected.

Finally, note that $\hat{\theta}_n^*(u)$ can take values that are greater than one and is undefined if the largest inter-exceedance time is less than three. To avoid these cases, define

$$\tilde{\theta}_n(u) = \begin{cases} 1 \wedge \hat{\theta}_n(u) & \text{if } \max\{T_i : 1 \le i \le N-1\} \le 2, \\ 1 \wedge \hat{\theta}_n^*(u) & \text{if } \max\{T_i : 1 \le i \le N-1\} > 2, \end{cases} \tag{4.9}$$

and call this the intervals estimator for the extremal index.

Consistency of the intervals estimator (4.9) for m-dependent processes is stated in the following theorem, a proof of which is in Appendix 4.D. Denote convergence in probability by $\xrightarrow{p}$.

Theorem 4.4. *Let the positive integers r_n and the thresholds u_n, for $n \ge 1$, be such that*

$$r_n = o(n), \quad r_n \to \infty, \quad r_n \bar{F}(u_n) \to \tau \quad \textit{and} \quad P(M_{r_n} \le u_n) \to \exp(-\theta\tau),$$

for some $\tau \in (0, \infty)$ and $\theta \in (0, 1]$. If $\{X_i\}_{i \ge 1}$ is m-dependent, then $\tilde{\theta}_n(u_n) \xrightarrow{p} \theta$.

Cluster-size distribution

Just as the limiting distribution (4.2) has been used to derive a moment estimator for the extremal index, so the limiting distributions (4.3) can be used to derive moment estimators for the cluster-size distribution. Rewriting the recurrence relation

(4.4) yields

$$\begin{aligned} \pi(1) &= e_1, \\ \pi(j) &= e_j - \pi(1)e_{j-1} - \ldots - \pi(j-1)e_1 \quad \text{for } j \geq 2. \end{aligned} \tag{4.10}$$

Two of many possible expressions for e_j that can be derived from the joint distribution (4.3) are

$$e_j = \theta E(\tilde{T}_1 \tilde{T}_{1+j}) \quad \text{and} \quad e_j = 2E\{\min(\tilde{T}_1, \tilde{T}_{1+j})\}.$$

Recalling $E(\tilde{T}_1) = 1$ and $E(\tilde{T}_1^2) = 2/\theta$, it follows that

$$e_j = 2E(\tilde{T}_1 \tilde{T}_{1+j})/E(\tilde{T}_1^2) \quad \text{and} \quad e_j = 2E\{\min(\tilde{T}_1, \tilde{T}_{1+j})\}/E(\tilde{T}_1).$$

The advantage of these expressions is that the estimators defined by substituting empirical versions of the expectations are independent of the normalisation used for the inter-exceedance times. In particular, e_j can be estimated for each $1 \leq j \leq N-2$ by

$$\hat{e}_j = \frac{2(N-1)\sum_{i=1}^{N-1-j}(T_i - 1)(T_{i+j} - 1)}{(N-1-j)\sum_{i=1}^{N-1}(T_i - 1)(T_i - 2)}$$

or

$$\tilde{e}_j = \frac{2(N-1)\sum_{i=1}^{N-1-j}\min(T_i - 1, T_{i+j} - 1)}{(N-1-j)\sum_{i=1}^{N-1}(T_i - 1)}, \tag{4.11}$$

where a penultimate approximation is again used to shrink the expectations. Simulations indicate that estimators based on $\tilde{e}_j$ are superior to those based on $\hat{e}_j$, which are therefore discarded. The estimators, $\tilde{\pi}_n(j)$, for the cluster-size distribution obtained recursively for $j = 2,\ldots,N-2$ by substituting expression (4.11) into relation (4.10) can be constrained by

$$\tilde{\pi}_n(j) = \min\left\{\max\{0, \tilde{e}_j - \tilde{\pi}_n(1)\tilde{e}_{j-1} - \ldots - \tilde{\pi}_n(j-1)\tilde{e}_1\}, 1 - \sum_{i=1}^{j-1}\tilde{\pi}_n(i)\right\} \quad (4.12)$$

to ensure that they form a proper distribution function. These will be referred to as the direct estimators for $\pi(j)$.

4.1.3 Least-squares estimators

In this section, an approach to estimating the extremal index and the cluster-size distribution based on quantile-quantile plots like Figure 4.2 is described. The limiting mixture distribution (4.2) determines that plots of normalised inter-exceedance times against standard exponential quantiles should be piecewise-linear with a breakpoint at the $(1-\theta)$-quantile, $-\log\theta$. The segment on the left should be constant with value zero; the segment on the right should have gradient $1/\theta$. The intervals estimator (4.9) is constructed from the moments of the limiting distribution, so it ignores this structure in the distribution. Here, an attempt is made to exploit the structure evident in the quantile-quantile plot.

Extremal index

Let $\{t_i\}_{i=1}^{N-1}$ be a collection of normalised inter-exceedance times, $NT_i/(n+1)$ say, ordered so that $t_1 \leq \ldots \leq t_{N-1}$. Let $\{x_i\}_{i=1}^{N-1}$ be a collection of standard exponential quantiles, where $x_i = -\log(1 - i/N)$. Figure 4.2 is an example of t_i plotted against x_i. If the t_i were a sample from the limiting distribution (4.2) then the broken-stick model

$$t = \beta(x - \log\beta)_+ \quad \text{for } \beta \geq 1$$

would have $\beta = 1/\theta$. In fact, it is better to fit this model when the t_i are defined by $N(T_i - 1)/(n + 1)$ since then the smallest inter-exceedance times are indeed zero.

Dependence between the ordered inter-exceedance times will be ignored, but heteroskedasticity accounted for by using weighted least squares to reduce the influence of the larger inter-exceedance times. The model is fitted by minimising

$$S(\beta) = \sum_{i=1}^{N-1} w_i\{t_i - \beta(x_i - \log\beta)_+\}^2,$$

where the weights are $w_i = 1/\text{var}(t_i)$. These weights could be computed from the mixture distribution (4.2) but, since θ is unknown and $\theta < 1$ can yield infinite weights, it is assumed for the purpose of choosing the weights that the t_i form an ordered sample of size $N - 1$ from a standard exponential distribution. In this case,

$$w_i^{-1} = \sum_{j=N-i}^{N-1} j^{-2};$$

see David (1981, page 49) for example. The least-squares estimator for the extremal index is

$$\tilde{\theta}_n^{LS}(u) = \{\text{argmin } S(\beta)\}^{-1}, \tag{4.13}$$

where it is necessary to search over β in the interval $[1, N]$ only since $S(\beta)$ takes the same value for all $\beta \geq \exp(x_{N-1}) = N$. More sophisticated choices of weights, using the intervals estimator to obtain a preliminary estimate of θ for example, were found to make little difference to the performance of the least-squares estimator.

Cluster-size distribution

Least-squares estimators can be defined for the cluster-size distribution as well. The joint distribution (4.3) implies

$$P\{\min(\tilde{T}_1, \tilde{T}_{1+j}) > t\} = \theta e_j \exp(-2\theta t),$$

and e_j can be estimated by fitting a broken-stick model to the corresponding quantile-quantile plot. The extremal index can be estimated simultaneously, or a prior estimate such as $\tilde{\theta}_n^{LS}(u)$ can be plugged in. Estimates of the cluster-size distribution follow from the recurrence relation (4.10).

4.1.4 Maximum-likelihood estimators

Two models for the marginal distribution of the inter-exceedance times have been described: the limiting form (4.2) and the penultimate approximation (4.6). In this section, the possibility of using these models to define maximum-likelihood estimators for the extremal index is examined. The joint distribution of all inter-exceedance times, which can be found by applying arguments similar to those used to prove Theorem 4.3, could be used, but likelihoods will be constructed from the marginal distribution assuming independent inter-exceedance times. The joint likelihood is more complicated and suffers from the same problems that are outlined below for the marginal likelihood.

With $t_i = NT_i/n$, the log-likelihood from the limiting model (4.2) is

$$\sum_{i=1}^{N-1} \log\left\{(1-\theta)^{I(t_i=0)}\left(\theta^2 e^{-\theta t_i}\right)^{I(t_i>0)}\right\} = 2(N-1)\log\theta - \theta\sum_{i=1}^{N-1} t_i,$$

since the observed inter-exceedance times are always strictly positive. The resulting maximum-likelihood estimator for θ is $\min(1, 2/\bar{t})$, where $\bar{t}$ is the mean of the t_i. This estimator tends in distribution to one as n increases, a failure arising from the model assigning all of the inter-exceedance times to the exponential component of the mixture distribution. It is possible to circumvent this problem by grouping a certain number of the smallest inter-exceedance times and treating them as though they were equal to zero. This solution requires the choice of an auxiliary parameter however. The choice is arbitrary and can have a significant impact on estimates, which is no improvement on runs or blocks declustering.

The log-likelihood from the penultimate model (4.6) is

$$m_1 \log(1-\theta p^\theta) + \{\log\theta + \log(1-p^\theta)\}(N-1-m_1) + \theta(\log p)\sum_{i=1}^{N-1}(T_i - 1),$$

where m_1 is the number of inter-exceedance times that are equal to one and $p = F(u)$. Maximum-likelihood estimates can be found by numerical optimisation. Unfortunately, a combination of two features can cause them to perform poorly. First, the distribution (4.6), from which the likelihood is constructed, is in general not a good model for the smallest inter-exceedance times. In particular, $P(T > 1) = \theta p^\theta$ converges to θ as $p \to 1$, where θ arises as the limit of $P(M_{1,m} \leq u \mid X_1 > u)$ as $u \to \omega$ and $m \to \infty$ at an appropriate rate (O'Brien, 1987). On the other hand, $P\{T(u) > 1\} = P(X_2 \leq u \mid X_1 > u)$ is larger than $P(M_{1,m} \leq u \mid X_1 > u)$ and need not converge to θ. For example, the doubly stochastic process of Smith and Weissman (1994) with parameters $0 < \psi < 1$ and $0 < \eta < 1$ has $\lim_{u\to\omega} P\{T(u) > 1\} = 1 - \eta\psi \geq (1-\psi)/(1-\psi+\psi\eta) = \theta$. The second, troublesome feature is the sensitivity of the log-likelihood to m_1, which is clearer after noting that $\sum_{i=1}^{N-1}(T_i - 1) = S_N - S_1 - (N-1)$ is asymptotically equivalent to n as $N \to \infty$ and $N/n \to 0$. Previously made comments warn that the observed value of m_1 may well be far from the value expected under the model, in which case poor estimates result. Grouping small inter-exceedance times and modelling only the larger times is again a possible but unattractive solution.

In contrast to the problems encountered with likelihood estimation, the intervals estimator (4.9) is rather robust to deviations of the smallest inter-exceedance times from the postulated model. This is due to relationship (4.7) in which the numerator and denominator both tend to infinity as $p = F(u) \to 1$, so that the smallest inter-exceedance times play no role asymptotically and the estimator depends most heavily on the well-modelled, large inter-exceedance times.

4.1.5 Declustering and bootstrapping

Although the extremal index and the cluster-size distribution can be estimated without recourse to declustering, estimating other cluster characteristics still may require clusters to be identified. All declustering schemes proposed in the literature require an auxiliary parameter, the choice of which is largely arbitrary. This section explains how the limiting distribution (4.2) can be used to identify clusters without making an arbitrary choice. Declustering will also be shown to support a bootstrap procedure for assessing estimation uncertainty. Importantly, this procedure accounts for the uncertainty in the choice of declustering scheme.

As the limiting process of exceedance times is a compound Poisson process, inter-exceedance times can be categorised as either independent inter-cluster times (between clusters) or independent sets of intra-cluster times (within clusters). As mentioned at the end of Section 4.1.1, the extremal index is the proportion of inter-exceedance times that can be regarded as inter-cluster times. Suppose that N exceedance times, $S_1 < \ldots < S_N$, are observed, with inter-exceedance times $T_i = S_{i+1} - S_i$ for $1 \leq i \leq N-1$, and let $\hat{\theta}$ be an estimate of the extremal index. The largest $C - 1 = \lfloor \hat{\theta}(N-1) \rfloor$ inter-exceedance times can be assumed to be approximately independent inter-cluster times that divide the remainder into approximately independent sets of intra-cluster times. To be precise, if $T_{(C)}$ is the C-th largest inter-exceedance time and T_{i_j} is the j-th inter-exceedance time to exceed $T_{(C)}$, then $\{T_{i_j}\}_{j=1}^{C-1}$ is a set of approximately independent inter-cluster times. If $C = N$, set $T_{(C)} = 0$; in the case of ties, decrease C until $T_{(C-1)}$ is strictly greater than $T_{(C)}$. Let $\mathcal{T}_j = \{T_{i_{j-1}+1}, \ldots, T_{i_j - 1}\}$, where $i_0 = 0$, $i_C = N$ and $\mathcal{T}_j = \emptyset$ if $i_j = i_{j-1} + 1$. Then $\{\mathcal{T}_j\}_{j=1}^{C}$ is a collection of approximately independent sets of intra-cluster times. Furthermore, each set $\mathcal{T}_j$ has associated with it a set of threshold exceedances $\mathcal{C}_j = \{X_k : k \in \mathcal{S}_j\}$, where $\mathcal{S}_j = \{S_{i_{j-1}+1}, \ldots, S_{i_j}\}$.

This interpretation justifies a decomposition of the observed process into C clusters, where the j-th cluster comprises the exceedances $\mathcal{C}_j$. This is equivalent to runs declustering with run length $r = T_{(C)}$, which is estimated directly from the data. If $\hat{\theta}$ is the intervals estimate (4.9) then the declustering procedure is entirely automatic and justified by the limiting theory.

Suppose now that inferences are required for a cluster functional $c(\cdot)$ such as those listed in (3.8). The functional can be evaluated for each cluster to obtain values $CF_i = c\{(X_j - u) : j \in \mathcal{S}_i\}$, $1 \leq i \leq C$, that can be used to find estimates, $\widehat{CF}$, of properties of c. For example, the expectation could be estimated by $\widehat{CF} = (CF_1 + \ldots + CF_C)/C$. Assuming independence of clusters and inter-cluster times, the bootstrap can be employed as follows to obtain confidence limits.

1. Resample with replacement $C-1$ inter-cluster times from $\{T_{i_j}\}_{j=1}^{C-1}$.
2. Resample with replacement C sets of intra-cluster times (some of which may be empty) and associated exceedances from $\{(\mathcal{T}_j, \mathcal{C}_j)\}_{j=1}^{C}$.
3. Intercalate these inter-exceedance times and clusters to form a bootstrap replication of the process above the threshold.
4. Compute N for the bootstrap process, estimate θ, and decluster accordingly.

5. Compute $\widehat{CF}$ from the declustered bootstrap process.

Forming B such bootstrap processes yields collections of estimates, $\{\hat{\theta}_{(1)}, \ldots, \hat{\theta}_{(B)}\}$ and $\{\widehat{CF}_{(1)}, \ldots, \widehat{CF}_{(B)}\}$, that can be used to approximate the distributions of the original point estimates, $\hat{\theta}$ and $\widehat{CF}$. In particular, the empirical α- and $(1-\alpha)$-quantiles of each sample define $(1-2\alpha)$-confidence intervals. This procedure is demonstrated for the data-set in Section 4.1.7, and used in the simulation study of the following section. Note first that if $CF_i = I(\#\mathcal{S}_i = j)$ then $\widehat{CF}$ estimates $\pi(j)$. The estimator

$$\bar{\pi}_n(j) = C^{-1} \sum_{i=1}^{C} I(\#\mathcal{S}_i = j) \quad \text{for } j \geq 1 \tag{4.14}$$

will be referred to as the indirect estimator for the cluster-size distribution.

4.1.6 Simulation study

In this section a simulation study is conducted to investigate the performances of the various estimators introduced in this chapter. For the extremal index comparisons are made between the runs estimator (4.1), which can be written in terms of the inter-exceedance times as

$$\bar{\theta}_n^R(u\,;\,r) = N^{-1} \left\{ \sum_{i=1}^{N-1} I(T_i > r) + 1 \right\}$$

when the run length is r, the intervals estimator (4.9) and the least-squares estimator (4.13). For the cluster-size distribution only estimators for $\pi(1)$ are compared:

estimators for other elements of the distribution were found to have similar relative performance. The estimators used are the direct estimator (4.12), the indirect estimator (4.14) when both intervals declustering and runs declustering are used, and the least-squares estimator described in Section 4.1.3.

Data are simulated from two stationary processes: a max-autoregressive process (3.10) with dependence parameter α, and a first-order Markov chain with extreme-value logistic transition distribution (2.9) and dependence parameter β. The extremal index of the Markov chain for specific values of β can be found by simulation (Smith *et al.*, 1997). One-thousand sequences of length 5000 are simulated from both of these processes for each of three extremal indices, $\theta = 0.25$, 0.5 and 0.75, corresponding to $\alpha = 0.75$, 0.5 and 0.25 for the max-autoregressive process, and $\beta = 0.43$, 0.64 and 0.82 for the Markov chain. For the max-autoregressive process, $\pi(1) = \theta$; for the three Markov chains, the values of $\pi(1)$ are 0.35, 0.56 and 0.77.

Extremal index

First, the intervals estimator is compared to three runs estimators, with run lengths 1, 5 and 9, at a range of thresholds chosen so that there are N exceedances, with N ranging from 10 to 1000. Root-mean-square errors of the point estimates and the coverage probability of bootstrapped confidence intervals, computed with $B = 1000$, are reported in each case.

Figure 4.3 shows the empirical root-mean-square errors and coverage probabilities for the four estimators applied to the three max-autoregressive processes. The sensitivity of the runs estimator to run length, threshold and the true extremal index is clear; the intervals estimator is robust to these factors. This robustness is particularly evident at lower thresholds, where the coverage probability for the intervals estimator remains close to the nominal value. The cause of the poor coverage for the runs estimator can be traced to step (d) of the bootstrap algorithm: the run length is fixed so that any uncertainty in the declustering scheme is unaccounted for, resulting in confidence intervals that are too narrow. The apparently uniform decrease of root-mean-square error with threshold for the intervals estimator is only indicative of the fact that the optimal threshold for the intervals estimator is typically lower than that for the runs estimator. The reason for this is the greater sensitivity of the runs estimator to disproportionately large numbers of small inter-exceedance times, which enhances the bias at low thresholds. The intervals estimator is robust to such discrepancies. Nevertheless, the runs estimators appear to outperform the intervals estimator for high thresholds. Figure 4.4 presents the results for the Markov chains. Similar comments can be made here with regard to the robustness of the intervals estimator. Note, however, that not every choice of run length for the runs estimator leads to a performance that is superior at high thresholds to that of the intervals estimator.

In Figure 4.5, root-mean-square error is plotted for the least-squares estimator, the intervals estimator and the runs estimators of the extremal index. The least-squares estimator often performs similarly to the best runs estimator, improving on the intervals estimator at high thresholds but worsening as the threshold decreases. This behaviour is due to the break-point in the quantile-quantile plot being poorly

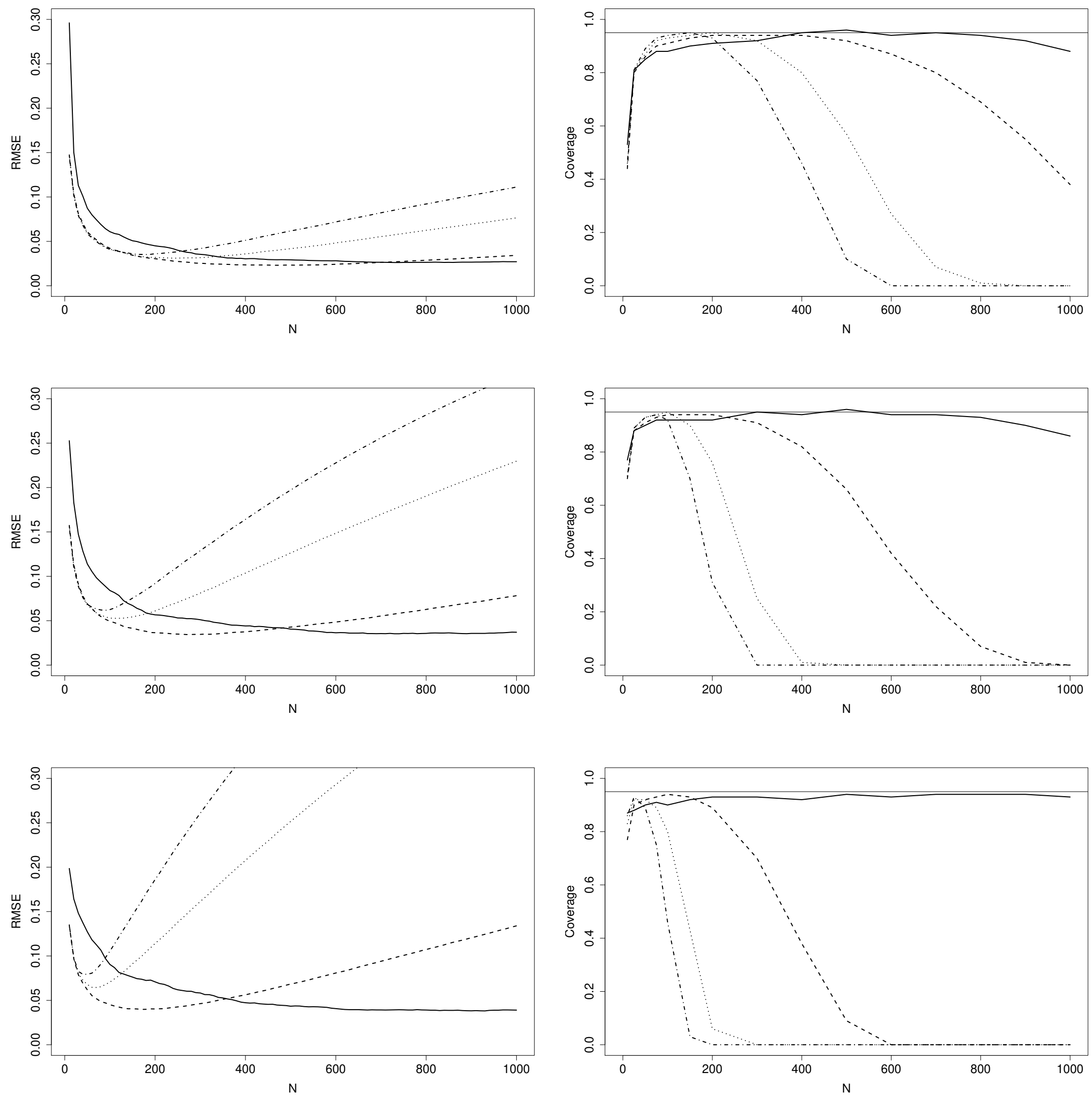


Figure 4.3: Root-mean-square errors (left) and coverage probabilities (right) for the intervals estimator (——) and runs estimators with run lengths 1 (- - -), 5 (· · ·) and 9 (-·-·-) applied to max-autoregressive sequences of length 5000 with extremal indices 0.25 (top), 0.5 (middle) and 0.75 (bottom). The nominal coverage probability (0.95) is indicated by a horizontal line.

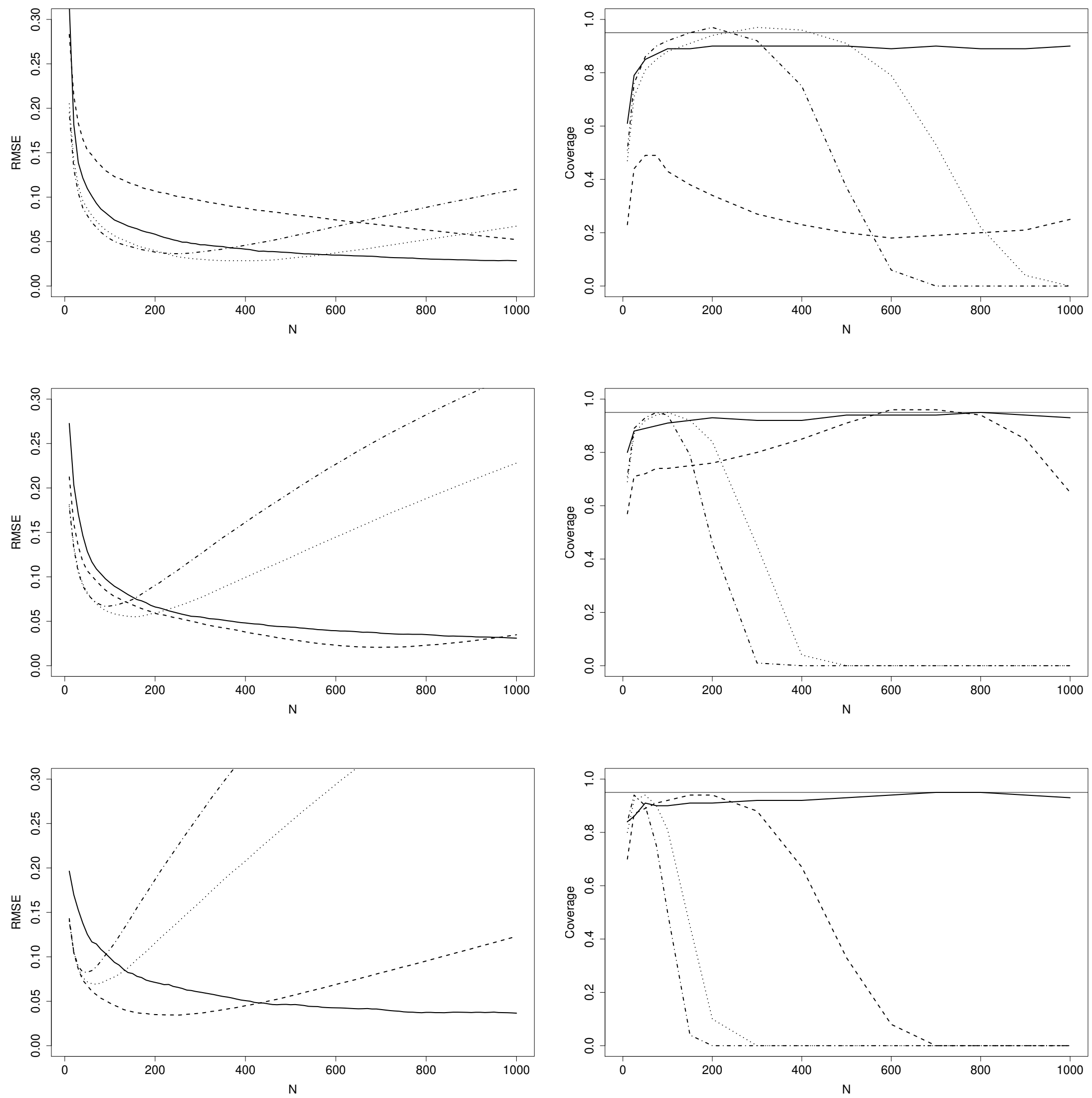


Figure 4.4: Root-mean-square errors (left) and coverage probabilities (right) for the intervals estimator (——) and runs estimators with run lengths 1 (- - -), 5 (· · ·) and 9 (-·-·-) applied to Markov chain sequences of length 5000 with extremal indices 0.25 (top), 0.5 (middle) and 0.75 (bottom). The nominal coverage probability (0.95) is indicated by a horizontal line.

defined at lower thresholds. As with the intervals estimator, the least-squares estimator does not require the choice of an auxiliary parameter, so its performance is very favourable.

Cluster-size distribution

Finally, a comparison is made of the estimators for $\pi(1)$, with N in the smaller range 10 to 500. The root-mean-square errors for the indirect estimators with both runs and intervals declustering are shown in Figure 4.6. Intervals declustering is seen to provide a robust alternative to runs declustering. The root-mean-square error for the direct estimator, the least-squares estimator and the indirect estimator with intervals declustering are compared in Figure 4.7. The least-squares estimator performs similarly to the indirect estimator, with the direct estimator performing slightly worse. The least-squares estimator for $\pi(1)$ is therefore competitive, but, as it is more expensive to compute than the indirect estimator, the cluster-size distribution and other cluster functionals are perhaps best estimated by declustering.

4.1.7 Data example

To conclude, the intervals estimator and intervals declustering are applied to a sequence of daily minimum temperatures recorded to the nearest degree Fahrenheit at Wooster, Ohio, between 1 June 1893 and 31 December 1987. The raw data exhibit strong seasonality, which is removed here by using the data from only

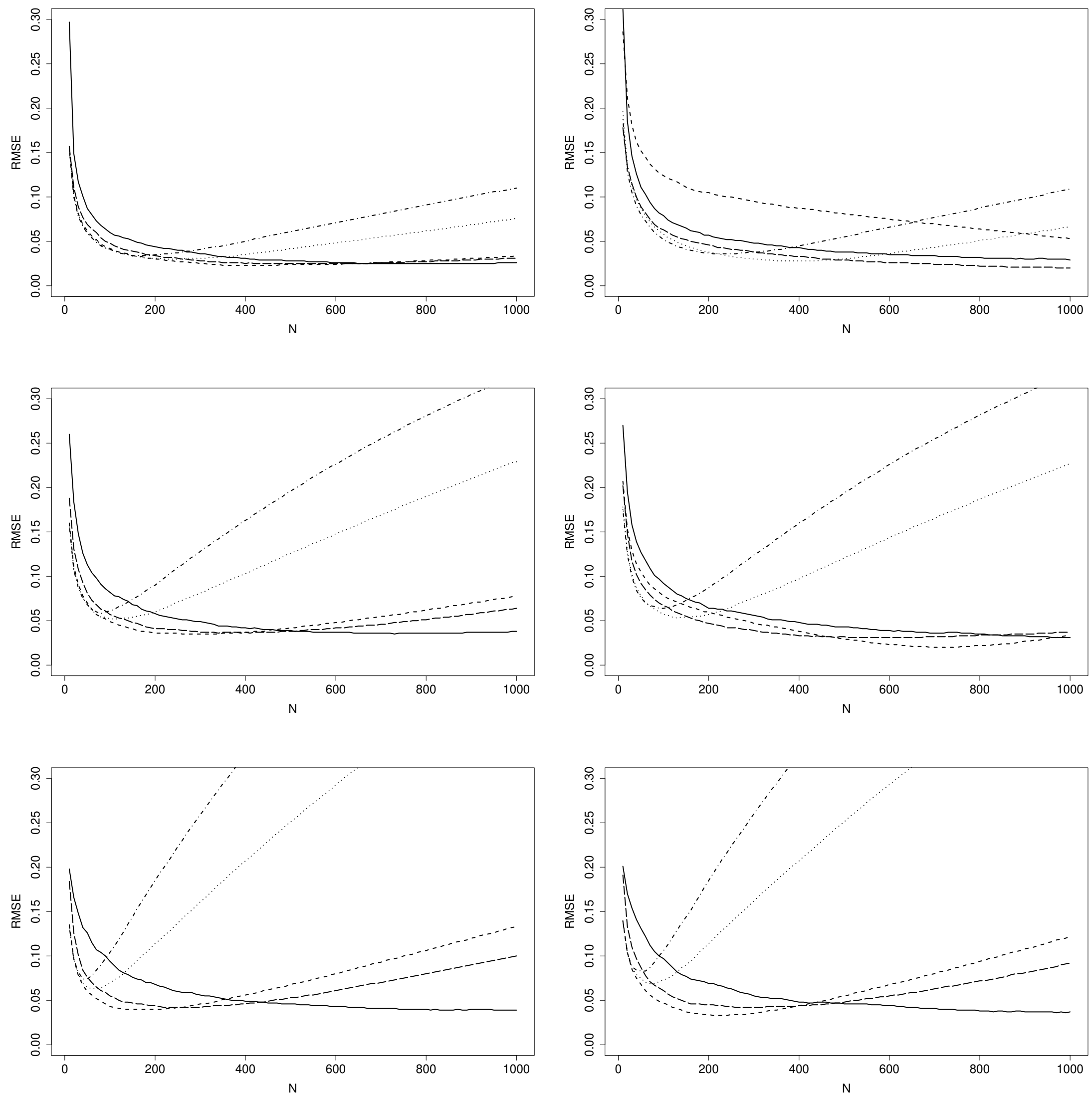


Figure 4.5: Root mean-square-error for the intervals estimator (——), the least-squares estimator (– – –) and the runs estimator with run lengths 1 (- - -), 5 (· · ·) and 9 (-·-·-) applied to max-autoregressive (left) and Markov chain (right) sequences of length 5000 with extremal indices 0.25 (top), 0.5 (middle) and 0.75 (bottom).

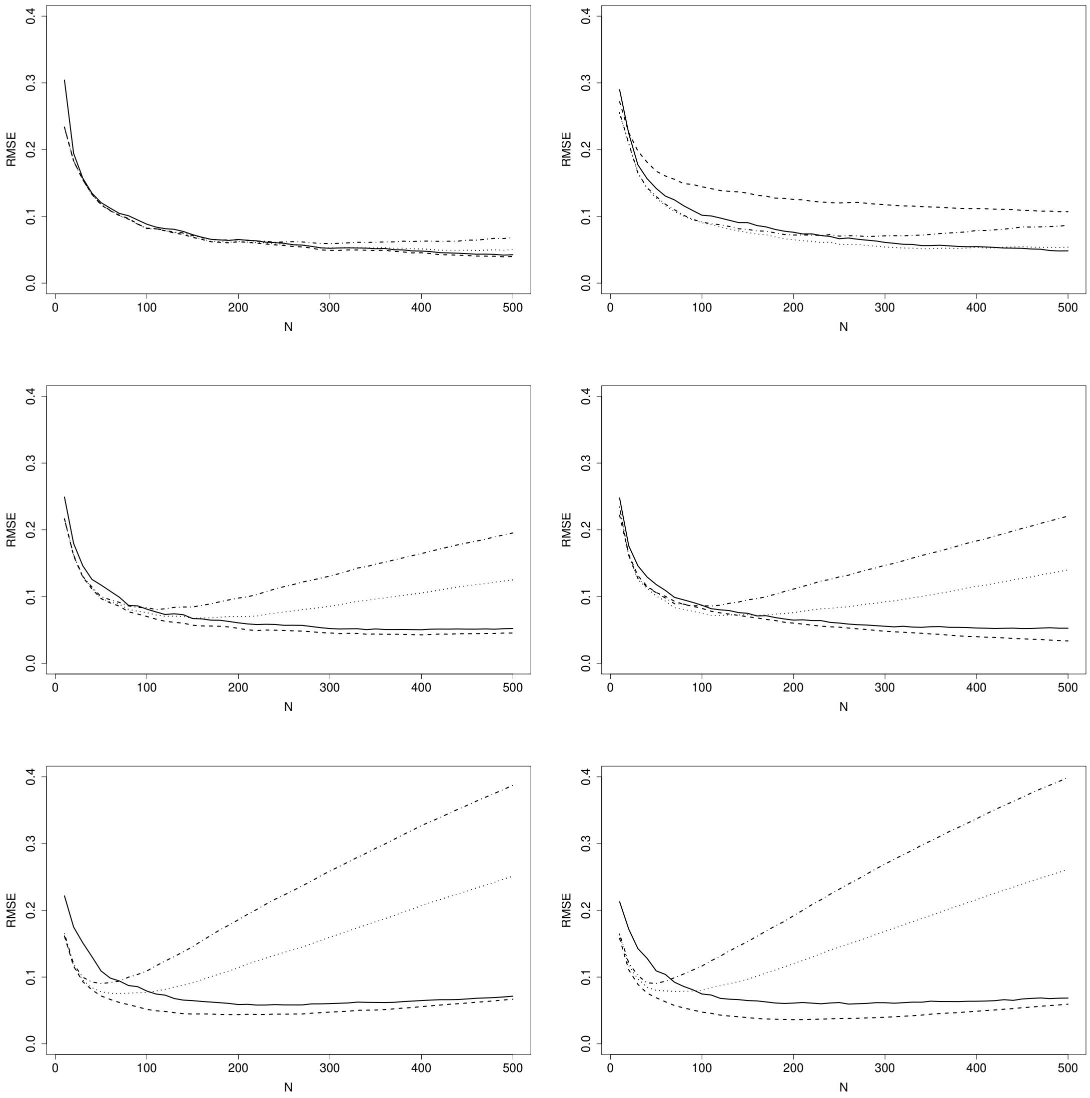


Figure 4.6: Root mean-square-error for the indirect estimator of $\pi(1)$ with intervals declustering (——) and runs declustering with run lengths 1 (- - -), 5 ($\cdots$) and 9 (-·-·-) applied to max-autoregressive (left) and Markov chain (right) sequences of length 5000 with extremal indices 0.25 (top), 0.5 (middle) and 0.75 (bottom).

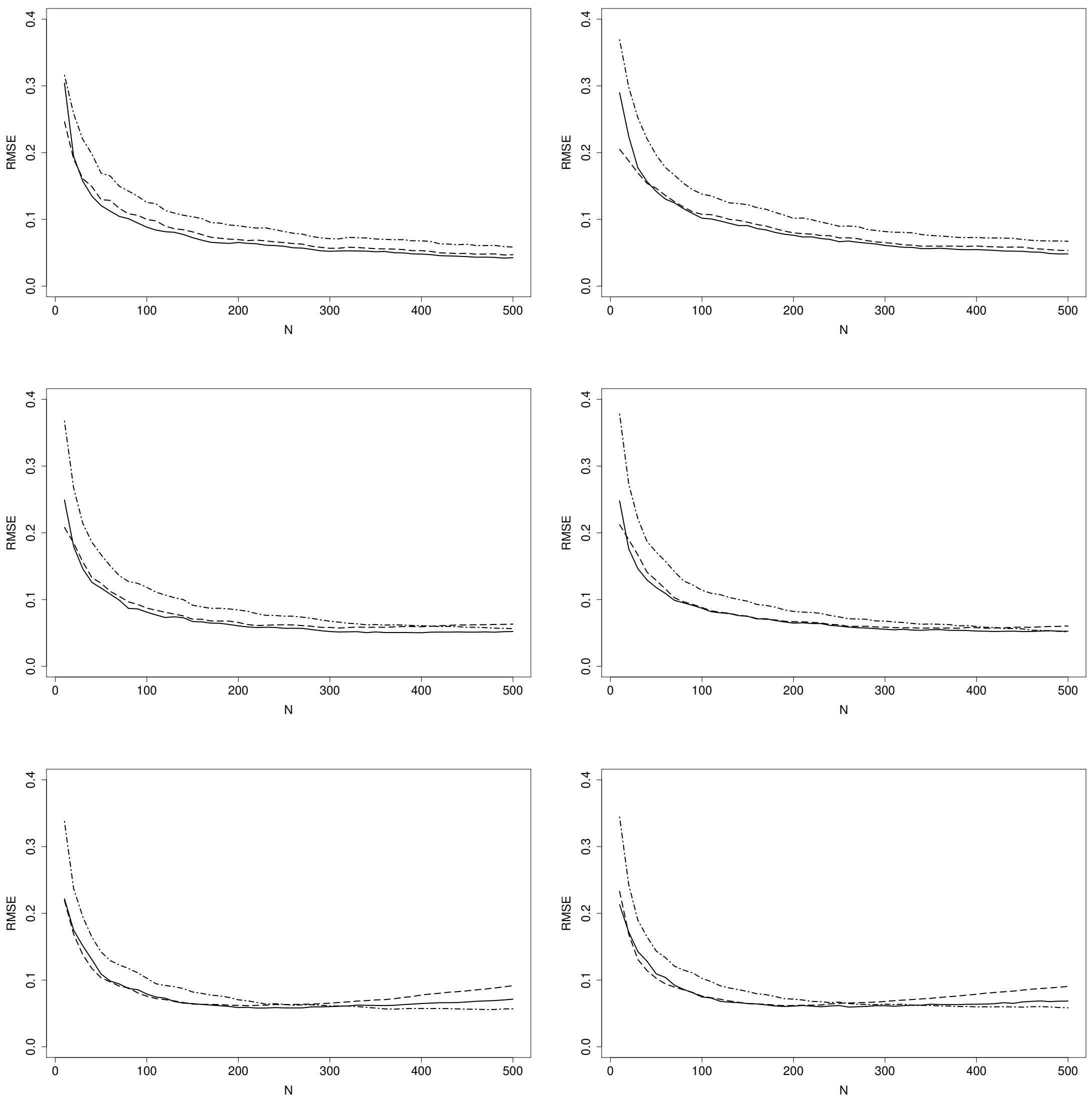


Figure 4.7: Root mean-square-error for the indirect estimator of $\pi(1)$ with intervals declustering (——), the least-squares estimator (– – –) and the direct estimator (-·-·-) applied to max-autoregressive (left) and Markov chain (right) sequences of length 5000 with extremal indices 0.25 (top), 0.5 (middle) and 0.75 (bottom).

the months December, January and February. Two missing values on 1 February 1954 and 3 January 1981 are also removed. The concatenated data are plotted in Figure 4.8. The extremal index, cluster-size distribution and mean cluster excess (3.8) of the series will be estimated at thresholds $u = 0.5, 1.5, \ldots, 14.5$. See Smith *et al.* (1997) for additional information and analysis.

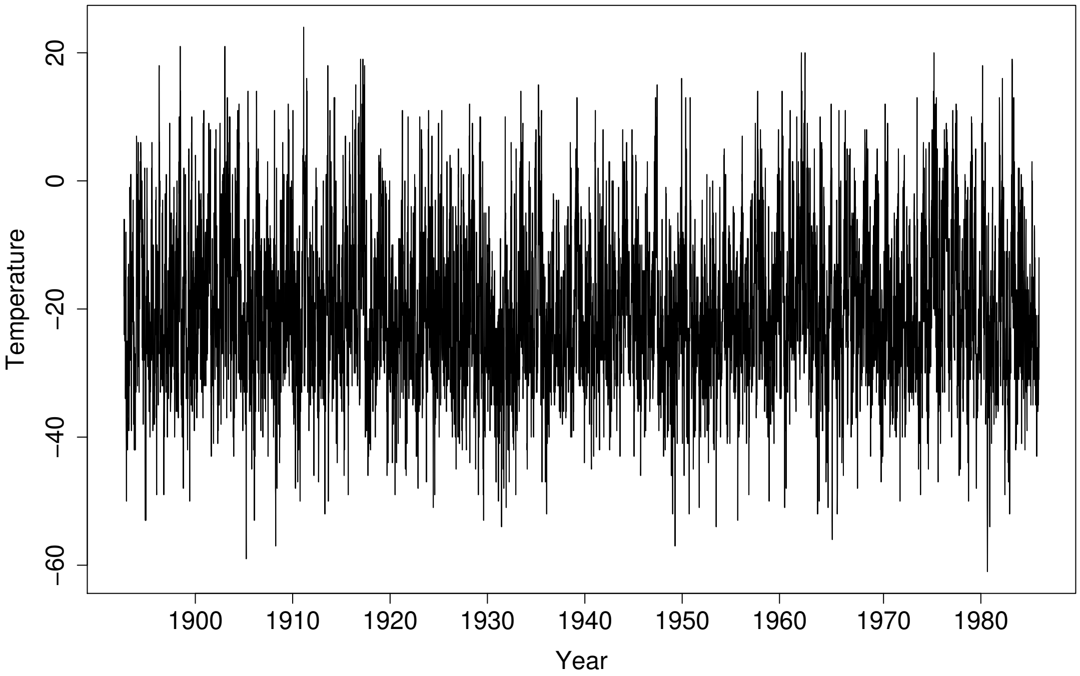


Figure 4.8: Daily minimum temperature recorded in degrees Fahrenheit below zero at Wooster. The first observation in each decade is labelled.

The intervals estimates of the extremal index are shown in Figure 4.9 with bootstrap confidence limits. For comparison, least-squares estimates and runs estimates with run length five days are also shown. The point estimates from the three methods are similar. As expected, the variance of the intervals estimator is greater than that of the runs estimator, which does not account for uncertainty

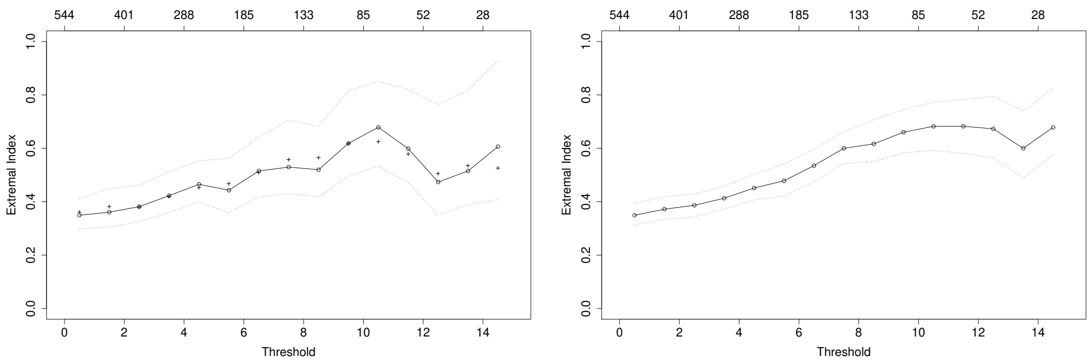


Figure 4.9: Estimates (–•–) of the extremal index for the Wooster temperature series. The intervals estimator (left) and the runs estimator with run length five (right) are shown with bootstrapped 95% confidence limits (- - -). The least-squares estimator (+) is superimposed and the number of exceedances at each threshold is indicated on the upper axis.

in the run length. Both plots support an extremal index of about 0.6, which is similar to the values found by Smith *et al.* (1997).

Estimates of the cluster-size distribution, obtained by declustering according to the intervals scheme, are in Figure 4.10. The estimates of $\pi(1)$ are similar to those of the extremal index; the estimates of $\pi(2)$ are around 0.25. Approximately 4% of clusters have more than four exceedances. Again, these values are similar to those found by Smith *et al.* (1997).

The estimates of the mean cluster excess are presented in the left-hand plot of Figure 4.11. These decrease approximately linearly with unit slope as the threshold increases before levelling out at the ten-degree threshold with mean cluster excess about six degrees. The pattern is the inverse of the extremal index, suggesting that the mean cluster excess follows the mean cluster size, with the distribution of excesses approximately stable over thresholds.

Finally, returning to the bootstrap procedure of Section 4.1.5, notice that each

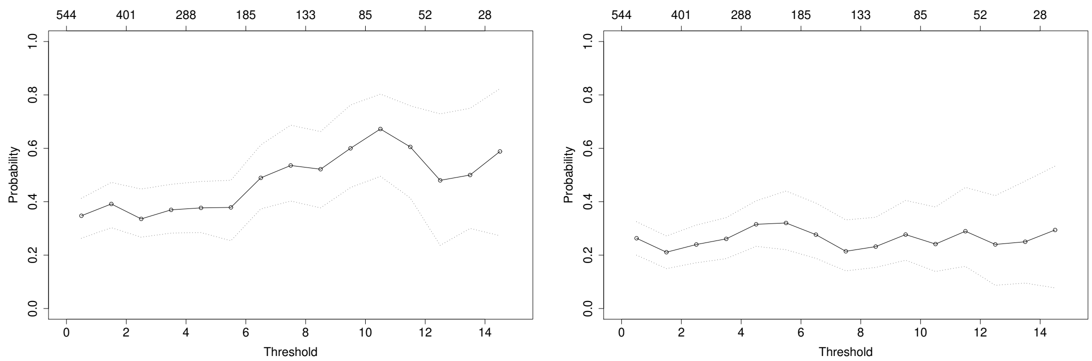


Figure 4.10: Estimates (–•–) based on intervals declustering of the cluster-size distribution, size 1 (left) and 2 (right), for the Wooster temperature series. Bootstrapped 95% confidence limits (- - -) are also shown. The number of exceedances at each threshold is indicated on the upper axis.

bootstrapped process has an associated run length, $r_{(b)} = T_{(C_{(b)})}$, where $C_{(b)} = \lceil \hat{\theta}_{(b)}(N_{(b)} - 1) \rceil$ and $N_{(b)}$ is the number of exceedances. At each threshold, the collection $\{r_{(b)}\}_{b=1}^{B}$ describes the uncertainty in the declustering scheme and indicates the extent of the uncertainty ignored by using runs declustering with an arbitrary run length. The right-hand plot in Figure 4.11 summarises the run-length distribution. A run length of about one week is suggested with the interpretation that cold periods separated by more than one week can be considered independent.

As well as stability of cluster estimates, the goodness-of-fit of the limiting model (4.2) to the observed inter-exceedance times can help to identify thresholds above which the asymptotic models are reasonable approximations. The fit can be assessed using quantile-quantile plots. For the Wooster data, all of the thresholds considered in the analysis gave good fits.

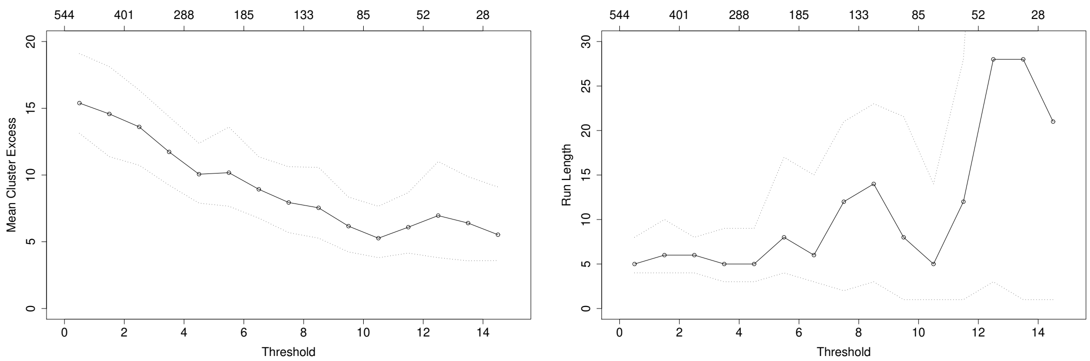


Figure 4.11: Estimates (–•–) based on intervals declustering of mean cluster excess (left) and run length (right) for the Wooster temperature series. Bootstrapped 95% confidence limits (- - -) are also shown. The number of exceedances at each threshold is indicated on the upper axis.

4.2 Multivariate Sequences

A few declustering schemes have been proposed for multivariate sequences: for example, Coles and Tawn (1991) and Nadarajah (2001). These schemes are designed to extract independent observations from a multivariate, stationary sequence: clusters are identified and then summarised by a single value, such as the componentwise maximum of the observations in the cluster. The approach of Coles and Tawn (1991) is a multivariate version of blocks declustering; that of Nadarajah (2001) is an extension of runs declustering. Both methods require the choice of one or more declustering parameters. A multivariate extension of the scheme outlined in the previous section that does not require an arbitrary choice of declustering parameter is proposed below.

Before declustering, the subset of observations considered to be extreme must be chosen. For univariate sequences this amounts to choosing a threshold; for multivariate data the decision is less straightforward. In Section 2.2, defining an observation to be extreme if and only if it falls in a set $A = \{x \in \mathbb{R}^D : x \not\leq u\}$ for some choice of high thresholds $u = (u_1, \ldots, u_D)$ allowed the marginal and dependence features of the point process model to be estimated simultaneously. Adopting this definition, the problem of multivariate declustering reduces to the problem of declustering the univariate sequence, $\{I(X_i \in A)\}_{i=1}^n$, of indicator variables. The approach of the previous section can be applied to this sequence to accomplish the task as the next result shows.

Let $A_n = \{x \in \mathbb{R}^D : x \not\leq u_n\}$ for $u_n = (u_{n1}, \ldots, u_{nD})$ and define $p(A_n) = P(X_1 \in A_n)$. Let $T(A_n)$ be the random variable representing the return time to the set A_n, with distribution

$$P\{T(A_n) > n\} = P(X_i \notin A_n \text{ for } 2 \leq i \leq n \mid X_1 \in A_n).$$

The following theorem uses a multivariate version of the $\Delta^*(u_n)$ condition of Section 4.1.1, in which the only difference is that the random variables and the thresholds have dimension D.

Theorem 4.5. *Let the thresholds u_n, $n \geq 1$, be such that*

$$nP(X_{1d} > u_{nd}) \to \tau_d \quad \textit{for } 1 \leq d \leq D$$

and $\{X_i\}_{i=1}^n$ has multivariate extremal index $\theta(\tau)$, that is

$$np(A_n) \to \nu \quad \textit{and} \quad P(M_n \leq u_n) \to \exp\{-\theta(\tau)\nu\}$$

for some $\nu \in (0, \infty)$ *and where* $\tau = (\tau_1, \ldots, \tau_D)$*. If the multivariate version of* $\Delta^*(u_n)$ *holds, then*

$$P\{p(A_n)T(A_n) > t\} \to \theta(\tau) \exp\{-\theta(\tau)t\} \quad \textit{for all } t > 0$$

as $n \to \infty$*.*

Proof. Let $\tilde{X} = \Psi(X)$ have standard Fréchet component distributions and let $\tilde{u}_n = n/\tau \sim \Psi(u_n)$. Following Proposition 2.1 of Smith and Weissman (1996), let

$$Z_i(\tau) = \max_{1 \le d \le D} \{\tau_d \tilde{X}_{id}\}.$$

Then,

$$P\{Z_1(\tau) \le n\} = P(\tilde{X}_1 \le \tilde{u}_n) \sim P(X_1 \le u_n)$$

so that the multivariate condition $\Delta^*(u_n)$ for the process $\{X_i\}_{i=1}^n$ is equivalent to the univariate analogue for the process $\{Z_i(\tau)\}_{i=1}^n$ with threshold n. Define $\bar{F}_Z(z) = P\{Z_1(\tau) > z\}$, $M_n^Z = \max_{1 \le i \le n} Z_i(\tau)$ and let $T_Z(z)$ be the random variable representing the time between exceedances of z by $\{Z_i(\tau)\}_{i=1}^n$. Now,

$$n\bar{F}_Z(n) = nP(\tilde{X}_1 \not\le \tilde{u}_n) \sim nP(X_1 \not\le u_n) \to \nu$$

and

$$P(M_n^Z \le n) = P(\tilde{M}_n \le \tilde{u}_n) \sim P(M_n \le u_n) \to \exp\{-\theta(\tau)\nu\},$$

that is $\{Z_i(\tau)\}_{i=1}^n$ has extremal index $\theta(\tau)$. Therefore, Theorem 4.1 applies and yields

$$P\{p(A_n)T(A_n) > t\} \sim P\{\bar{F}_Z(n)T_Z(n) > t\} \to \theta(\tau)\exp\{-\theta(\tau)t\} \quad \text{for } t > 0$$

as $n \to \infty$. □

The application of this result is demonstrated with a simple example due to Smith and Weissman (1996). Consider the bivariate M_4 process (3.11) with $\alpha_{101} = 1$, $\alpha_{102} = \alpha_{112} = 1/2$, that is $X_i = (X_{i1}, X_{i2})$, where

$$X_{i1} = Z_i \quad \text{and} \quad X_{i2} = \max(Z_{i-1}, Z_i)/2$$

and the Z_i are independent, standard Fréchet random variables. The multivariate extremal index satistfies $\theta(c\tau) = \theta(\tau)$ for all positive c (Nandagopalan, 1994), so the extremal index of the bivariate process has the form $\theta(\tau_1, \tau_2) = \theta(w, 1-w)$, where $w = \tau_1/(\tau_1 + \tau_2) \in [0, 1]$. It is straightforward to derive

$$\theta(w, 1-w) = \begin{cases} 1/2 & \text{if } 0 \leq w \leq 1/3, \\ 2w/(1+w) & \text{if } 1/3 \leq w \leq 1. \end{cases}$$

The limiting cluster-size distribution, where clusters contain values in the set A_n above, satisfies

$$\pi(1) = \begin{cases} 0 & \text{if } 0 \leq w \leq 1/3, \\ (3w-1)/2w & \text{if } 1/3 \leq w \leq 1 \end{cases}$$

and $\pi(2) = 1 - \pi(1)$ (Smith and Weissman, 1996).

By the theorem, an estimator for $\theta(c\tau)$ is the intervals estimator evaluated for return times to a set $A = \{x \in \mathbb{R}^2 : x \not\leq u\}$, where $u = (u_1, u_2)$ satisfies

$$n\bar{F}_1(u_1) = cw \quad \text{and} \quad n\bar{F}_2(u_2) = c(1 - w).$$

In practice, thresholds are estimated using the component empirical distribution functions, and the constant c can be chosen so that the number of points in A is the same for each w. This method, called Method I, can be adapted to any definition of the extreme set A. A simpler but less general approach, Method II, is to apply the intervals estimator to the univariate sequence $Z_i(\tau) = \max\{wX_{i1}, (1-w)X_{i2}\}$, $1 \leq i \leq n$, identified in the proof of the theorem as having extremal index $\theta(\tau)$. Other estimators are described by Smith and Weissman (1996).

The multivariate process can be declustered by applying intervals declustering to either the return times to A, or the inter-exceedance times of $\{Z_i(\tau)\}_{i=1}^n$. The cluster-size distribution and other cluster functionals can then be estimated empirically.

A single realisation of length 5000 is simulated from the M_4 process, and the extremal index and cluster-size distribution are estimated. For estimates based on return times to the extreme set A, c is chosen so that A contains 250 observations for each w. Estimates based on $\{Z_i(\tau)\}_{i=1}^n$ are evaluated at a threshold equal to the empirical 95% quantile. The estimates of the multivariate extremal index and the cluster-size distribution from the two methods are displayed in Figure 4.12.

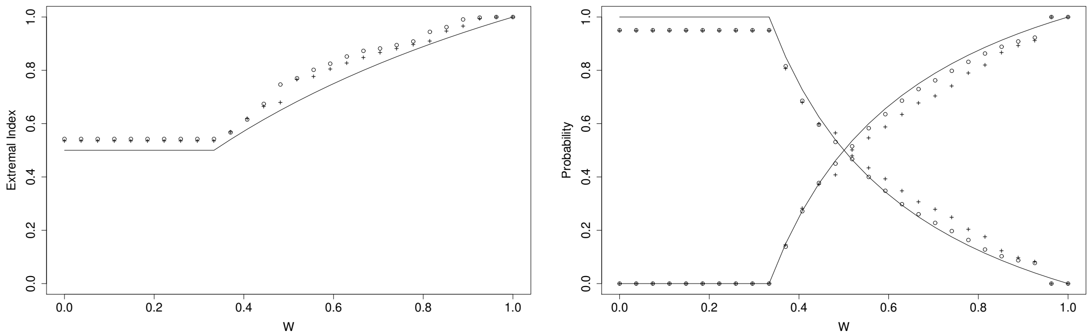


Figure 4.12: Estimates of the extremal index (left) and the first two elements of the cluster-size distribution (right) using Method I (∘) and Method II (+). In both plots, the solid lines correspond to the truth. For the cluster-size distribution, $\pi(1)$ increases and $\pi(2)$ decreases as w increases.

4.A Proof of Theorem 4.1

(The proofs in Appendices 4.A and 4.D were constructed by Johan Segers, co-author to Ferro and Segers (2003a).)

Since both $l_n = o(r_n)$ and $\alpha_n = \alpha_{r_n,l_n}(u_n) = o(1)$, there exist positive integers p_n such that

$$p_n = o(r_n), \quad r_n\alpha_n = o(p_n) \quad \text{and} \quad l_n = o(p_n).$$

(Take for instance $p_n = \lfloor\{r_n \max(l_n, r_n\alpha_n)\}^{1/2}\rfloor$.) Since $F(u_n)^{r_n} \to e^{-\tau}$, all conditions of Corollary 2.3 of O'Brien (1987) are fulfilled, and

$$P(M_{r_n} \leq u_n) - F(u_n)^{r_n P(M_{1,p_n} \leq u_n | X_1 > u_n)} \to 0.$$

But $P(M_{r_n} \leq u_n) \to e^{-\theta\tau}$ by assumption, so necessarily

$$P(M_{1,p_n} \leq u_n \mid X_1 > u_n) \to \theta.$$

Now put $t_n = \lfloor t/\bar{F}(u_n) \rfloor$. Then

$$\begin{aligned} P\{\bar{F}(u_n)T(u_n) > t\} &= P\{T(u_n) > t_n\} \\ &= P(M_{1,1+t_n} \leq u_n \mid X_1 > u_n). \end{aligned}$$

For n large enough so that $l_n < p_n$,

$$\begin{aligned} P(M_{p_n,p_n+l_n} > u_n \mid X_1 > u_n) &\leq P(M_{p_n,p_n+l_n} > u_n) + \alpha_n \\ &\leq l_n\bar{F}(u_n) + \alpha_n \to 0, \end{aligned}$$

so that

$$\begin{aligned} &P(M_{1,1+t_n} \leq u_n \mid X_1 > u_n) \\ &\quad = P(M_{1,p_n} \leq u_n,\, M_{p_n+l_n,t_n} \leq u_n \mid X_1 > u_n) + o(1). \end{aligned}$$

But $t_n \sim tr_n/\tau = O(r_n)$, so that by assumption

$$\begin{aligned} &P(M_{1,p_n} \leq u_n,\, M_{p_n+l_n,t_n} \leq u_n \mid X_1 > u_n) \\ &\quad = P(M_{p_n+l_n,t_n} \leq u_n \mid M_{1,p_n} \leq u_n,\, X_1 > u_n)P(M_{1,p_n} \leq u_n \mid X_1 > u_n) \\ &\quad = \{P(M_{p_n+l_n,t_n} \leq u_n) + o(1)\}\{\theta + o(1)\}. \end{aligned}$$

Now also $s_n = k_n - (p_n + q_n) \sim t\tau^{-1} r_n = O(r_n)$. Apply Corollary 2.3 of O'Brien (1987) again to deduce that

$$\begin{aligned} P(M_{p_n+l_n,t_n} \le u_n) &= P(M_{s_n} \le u_n) \\ &= F(u_n)^{s_n P(M_{1,p_n} \le u_n | X_1 > u_n)} + o(1) \to e^{-\theta t}. \end{aligned}$$

$\square$

4.B Proof of Theorem 4.2

Lemma 4.1. *Let $\{X_i\}_{i\ge1}$ be an m-dependent, strictly stationary sequence and let the thresholds u_n, $n \ge 1$, be such that*

$$n\bar{F}(u_n) \to \tau \quad \textit{and} \quad P(M_n \le u_n) \to \exp(-\theta\tau)$$

for some $\tau \in (0,\infty)$ and $\theta \in (0,1]$. Define also

$$\pi_n(1) = P\{N_{r_n}(u_n) = 1 \mid N_{r_n}(u_n) > 0\},$$

where $r_n = o(n)$ and $r_n \to \infty$. Then, there exists $l_n = o(r_n)$ such that $\pi_n^*(1) = \pi_n(1) + o(1)$, where

$$\pi_n^*(1) = P(M_{l_n-1} \le u_n \mid X_{l_n} > u_n, M_{l_n,2l_n-1} \le u_n).$$

Proof. By Theorem 2.1 of O'Brien (1987) there exists $l_n = o(r_n)$ such that $l_n \to \infty$ and

$$P(M_{1,l_n} \le u_n \mid X_1 > u_n) \to \theta.$$

Note that $P\{N_{r_n}(u_n) > 0\} = 1 - P(M_{r_n} \le u_n) \sim \theta\tau/k_n$, where $k_n = \lfloor n/r_n \rfloor$. Therefore,

$$\pi_n(1) = \frac{k_n}{\theta\tau} P\{N_{r_n}(u_n) = 1\}\{1 + o(1)\}.$$

Now,

$$P\{N_{r_n}(u_n) = 1\} = \sum_{i=1}^{r_n} P(X_i > u_n, X_j \le u_n \text{ for } j = 1, \ldots, r_n, j \ne i)$$

and

$$\sum_{i=1}^{l_n} P(X_i > u_n, X_j \le u_n \text{ for } j = 1, \ldots, r_n, j \ne i) \le l_n \bar{F}(u_n) \sim \tau l_n / n$$

so that

$$\pi_n(1) = \frac{k_n}{\theta\tau} \sum_{i=l_n+1}^{r_n - l_n} P(X_i > u_n, X_j \le u_n \text{ for } j = 1, \ldots, r_n, j \ne i)\{1 + o(1)\} + o(1).$$

For $1 + l_n \leq i \leq r_n - l_n$ and n large enough so that $l_n \geq m$,

$$\begin{aligned}
0 &\leq P(M_{i-l_n,i-1} \leq u_n, X_i > u_n, M_{i,i+l_n-1} \leq u_n) \\
&\quad - P(M_{i-1} \leq u_n, X_i > u_n, M_{i,r_n} \leq u_n) \\
&= P(M_{i-l_n} > u_n, M_{i-l_n,i-1} \leq u_n, X_i > u_n, M_{i,i+l_n-1} \leq u_n, M_{i+l_n-1,r_n} > u_n) \\
&\leq P(M_{i-l_n} > u_n, X_i > u_n) \\
&\leq \bar{F}(u_n)^2 = o\left(n^{-1}\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\pi_n(1) &= \frac{k_n}{\theta\tau} \sum_{i=l_n+1}^{r_n-l_n} P(M_{i-l_n,i-1} \leq u_n, X_i > u_n, M_{i,i+l_n-1} \leq u_n)\{1 + o(1)\} + o(1) \\
&= \frac{k_n}{\theta\tau}(r_n - 2l_n)P(M_{l_n-1} \leq u_n, X_{l_n} > u_n, M_{l_n,2l_n-1} \leq u_n)\{1 + o(1)\} + o(1) \\
&= \frac{k_n}{\theta\tau}(r_n - 2l_n)\pi_n^*(1)P(M_{1,l_n-1} \leq u_n \mid X_1 > u_n)\bar{F}(u_n)\{1 + o(1)\} + o(1) \\
&= \frac{k_n r_n}{\theta\tau}\pi_n^*(1)\theta\bar{F}(u_n)\{1 + o(1)\} + o(1) \\
&= \pi_n^*(1) + o(1).
\end{aligned}$$

□

Remark The definition of $\pi_n^*(1)$ is based on runs declustering rather than the usual blocks declustering. It can be read as the probability that X_{l_n} is the first exceedance in a cluster given that it is the last exceedance in a cluster, where clusters are defined by runs declustering with run length $l_n - 1$.

Turn now to the theorem itself. The conditions of Lemma 4.1 are satisfied. Let

$s_n = \lfloor s/\bar{F}(u_n) \rfloor \sim sn/\tau$, $t_n = \lfloor t/\bar{F}(u_n) \rfloor \sim tn/\tau$ and denote $\lceil wn \rceil$ by w_n. Then

$$\begin{aligned}
&P\{\bar{F}(u_n)T_1(u_n) > s, \bar{F}(u_n)T_2(u_n) > t\} \\
&= \int_{s_n}^{\infty} P(M_{1,\lceil w \rceil} \leq u_n, X_{\lceil w \rceil + 1} > u_n, M_{\lceil w \rceil + 1, \lceil w \rceil + 1 + t_n} \leq u_n \mid X_1 > u_n) \, \mathrm{d}w \\
&= \int_{s_n/n}^{\infty} nP(M_{1,w_n} \leq u_n, X_{w_n+1} > u_n, M_{w_n+1, w_n+1+t_n} \leq u_n \mid X_1 > u_n) \, \mathrm{d}w.
\end{aligned}$$

By arguments similar to those in the proof of Lemma 4.1 and for sufficiently large n, the integrand is

$$\begin{aligned}
&n\bar{F}(u_n)P(M_{1,w_n} \leq u_n, M_{w_n+1,w_n+1+t_n} \leq u_n \mid X_1 > u_n, X_{w_n+1} > u_n) \\
&= n\bar{F}(u_n)P(M_{1,l_n} \leq u_n, M_{l_n+m, w_n-l_n-m} \leq u_n, M_{w_n-l_n,w_n} \leq u_n, \\
&\qquad M_{w_n+1,w_n+l_n} \leq u_n, M_{w_n+l_n+m, w_n+1+t_n} \leq u_n \mid X_1 > u_n, X_{w_n+1} > u_n) + o(1) \\
&= n\bar{F}(u_n)P(M_{1,l_n} \leq u_n \mid X_1 > u_n)P(M_{w_n-2l_n-2m} \leq u_n)P(M_{t_n-l_n-m+1} \leq u_n) \\
&\qquad \times P(M_{w_n-l_n,w_n} \leq u_n, M_{w_n+1,w_n+l_n} \leq u_n \mid X_{w_n+1} > u_n) + o(1) \\
&= \tau\theta^2\pi(1)e^{-\theta(t+w\tau)} + o(1).
\end{aligned}$$

Finally, the integrand is dominated by $C\gamma^w$ when $n \geq n_0$ for some $C < \infty$, $\gamma \in (0,1)$ and positive integer n_0 since

$$\begin{aligned}
&nP(M_{1,w_n} \leq u_n, X_{w_n+1} > u_n, M_{w_n+1,w_n+1+t_n} \leq u_n \mid X_1 > u_n) \\
&\leq nP(M_{m,w_n+1-m} \leq u_n, X_{w_n+1} > u_n \mid X_1 > u_n) \\
&\leq n\bar{F}(u_n)P(M_{w_n-2m} \leq u_n),
\end{aligned}$$

$n\bar{F}(u_n) \to \tau \in (0, \infty)$ and $P(M_{w_n - 2m} \le u_n) \to \exp(-\theta\tau w)$. The result follows by dominated convergence. □

4.C Proof of Theorem 4.3

A proof of Theorem 4.3 follows the same lines as that of Theorem 4.2; the reasoning is outlined here. There are j exceedances between the two inter-exceedance times under consideration so that the joint tail function can be written as

$$\sum_{i=1}^{j} P\{T_1(u_n) \ge s_n, i \text{ intervening clusters}, T_{1+j}(u_n) \ge t_n\},$$

where $s_n = \lceil s/\bar{F}(u_n) \rceil$ and $t_n = \lceil t/\bar{F}(u_n) \rceil$. The summands factorise approximately into terms that behave similarly to

$$\begin{gathered} P(M_{1,l_n} \le u_n \mid X_1 > u_n) P(M_{s_n} \le u_n) P(M_{t_n} \le u_n) \\ \times P(j \text{ consecutive exceedances form } i \text{ clusters}). \end{gathered}$$

This has limit $\theta e^{-\theta(s+t)} p_{i,j}$, where

$$\begin{aligned} p_{1,j} &= \pi(j) \\ p_{i+1,j} &= \sum_{k=1}^{j-i} \pi(k) p_{i,j-k} \quad \text{for } i = 1, \dots, j-1 \end{aligned}$$

by independence of clusters. Finally, define $e_j = \sum_{i=1}^{j} p_{i,j}$ for $j \geq 1$ and observe that

$$\begin{aligned} e_j &= \sum_{i=1}^{j-1} p_{i+1,j} + \pi(j) = \sum_{i=1}^{j-1} \sum_{k=1}^{j-i} \pi(k) p_{i,j-k} + \pi(j) \\ &= \sum_{k=1}^{j-1} \pi(k) \sum_{i=1}^{j-k} p_{i,j-k} + \pi(j) = \sum_{k=1}^{j-1} \pi(k) e_{j-k} + \pi(j). \end{aligned}$$

□

4.D Proof of Theorem 4.4

Denote $\bar{F}(u_n)$ by p_n. To avoid ambiguities, let $\tilde{\theta}_n(u_n)$ be an arbitrary number, say 1, if $N \in \{0, 1\}$.

Lemma 4.2. *There exist $C > 0$, $0 < \gamma < 1$ and positive integer n_0 such that*

$$\begin{aligned} &(i) & P(M_k \leq u_n) &\leq C\gamma^{k/r_n} \\ &(ii) & P(M_{1,k} \leq u_n \mid X_1 > u_n) &\leq C\gamma^{k/r_n} \end{aligned}$$

for all integer $k \geq 1$ and $n \geq n_0$.

Proof. (i) Fix positive integer k and write $k = qm + j$ where $q = 0, 1, 2, \ldots$ and $j = 0, 1, \ldots, m-1$. Divide the sample of size k into q blocks of size m and one block of size j. By m-dependence,

$$P(M_k \leq u_n) \leq (1 - p_n)^q \leq (1 - p_n)^{-j/m} \left\{ (1 - p_n)^{r_n/m} \right\}^{k/r_n}.$$

Since $1 - p_n \to 1$ and $(1-p_n)^{r_n} \to e^{-\tau} < 1$, there exists $0 < \gamma < 1$ and a positive integer n_0 such that for all integer $n \geq n_0$,

$$\begin{aligned} (1-p_n)^{-j/m} &\leq (1-p_n)^{-1} \leq 2 \quad \text{for } j = 0, 1, \ldots, m-1, \\ (1-p_n)^{r_n/m} &\leq \gamma. \end{aligned}$$

Hence for all integer $n \geq n_0$, $P(M_k \leq u_n) \leq 2\gamma^{k/r_n}$.

(ii) For integer $k \geq m+1$, m-dependence and stationarity give

$$P(M_{1,k} \leq u_n \mid X_1 > u_n) \leq P(M_{m,k} \leq u_n) = P(M_{k-m} \leq u_n).$$

Thanks to (i), the latter is bounded by $2\gamma^{-m}\gamma^{k/r_n}$. For $k \leq m$, $2\gamma^{-m}\gamma^{k/r_n} \geq 1$. Put $C = 2\gamma^{-m}$. $\square$

Lemma 4.3. *If the positive integers s_n satisfy $s_n \sim \lambda r_n$ for some $0 < \lambda < \infty$, then*

$$P(M_{s_n} \leq u_n) \to e^{-\theta\lambda\tau} \quad \textit{and} \quad P(M_{1,s_n} \leq u_n \mid X_1 > u_n) \to \theta e^{-\theta\lambda\tau}.$$

Proof. Similar to the proof of Theorem 4.1. $\square$

By convention, set $\sum_{i=1}^{N-1}(T_i - 1) = 0$ if $N \in \{0, 1\}$.

Lemma 4.4. *$N \overset{p}{\sim} np_n$ and $\sum_{i=1}^{N-1}(T_i - 1) \overset{p}{\sim} n$.*

Proof. Clearly $E(N) = np_n$. By stationarity and m-dependence, also $\operatorname{var}(N) \leq$

$2mnp_n$. Together, $E[\{(np_n)^{-1}N - 1\}^2] \leq 2m(np_n)^{-1} \to 0$, which implies $N \overset{p}{\sim} np_n$. In particular, $N \xrightarrow{p} \infty$, so that $P(N \geq 2) \to 1$.

Next, $\sum_{i=1}^{N-1}(T_i - 1) = S_N - S_1 - (N-1) = n - (n - S_N) - S_1 - (N-1)$ if $N \geq 2$. For $c > 0$, Lemma 4.2 gives

$$P(n - S_N \geq cr_n \text{ and } N \geq 2) \leq P(M_{\lfloor cr_n \rfloor} \leq u_n) \leq C\gamma^{\lfloor cr_n \rfloor / r_n}$$

so that $n - S_N = O_p(r_n) = o_p(n)$. Similarly, $S_1 = O_p(r_n) = o_p(n)$. From the first part, it is already known that $N = o_p(n)$. $\square$

It remains to deal with the term $\sum_{i=1}^{N-1}(T_i - 1)(T_i - 2)$, which is also set equal to zero if $N \in \{0, 1\}$. Define

$$Y_n = \sum_{i=1}^{n-1} \sum_{j=1}^{n-i} (j-1) I(X_i > u_n \geq M_{i,j}).$$

Observe that if $N \geq 2$, then

$$\begin{aligned} Y_n &= \sum_{i=1}^{N-1} \sum_{j=1}^{T_i - 1} (j-1) + \sum_{j=1}^{n-S_N} (j-1) \\ &= \frac{1}{2} \sum_{i=1}^{N-1} (T_i - 1)(T_i - 2) + \frac{1}{2}(n - S_N)(n - S_N - 1). \end{aligned}$$

In particular,

$$Y_n = \frac{1}{2} \sum_{i=1}^{N-1} (T_i - 1)(T_i - 2) + O_p(r_n^2).$$

Hence, to prove the Theorem, it is sufficient to show that $n^{-1}p_n Y_n \xrightarrow{p} \theta^{-1}$. To this end, it is sufficient to show that

$$n^{-1} p_n E(Y_n) \to \theta^{-1} \quad \text{and} \quad n^{-2} p_n^2 E(Y_n^2) \to \theta^{-2}.$$

Lemma 4.5. $n^{-1} p_n E(Y_n) \to \theta^{-1}$

Proof. By stationarity of the process $\{X_n\}_{n \geq 1}$,

$$\begin{aligned} E(Y_n) &= \sum_{i=1}^{n-1} \sum_{j=1}^{n-i} (j-1) P(X_i > u_n \geq M_{i,i+j}) \\ &= p_n \sum_{j=1}^{n} (j-1)(n-j) P(M_{1,1+j} \leq u_n \mid X_1 > u_n). \end{aligned}$$

Rewrite the sum as an integral to find

$$\begin{aligned} E(Y_n) &= p_n \int_0^n (\lceil s \rceil - 1)(n - \lceil s \rceil) P(M_{1,1+\lceil s \rceil} \leq u_n \mid X_1 > u_n) \, \mathrm{d}s \\ &= p_n r_n \int_0^{n/r_n} (\lceil r_n s \rceil - 1)(n - \lceil r_n s \rceil) P(M_{1,1+\lceil r_n s \rceil} \leq u_n \mid X_1 > u_n) \, \mathrm{d}s. \end{aligned}$$

Hence

$$\begin{aligned} & n^{-1} p_n E(Y_n) \\ &\quad = p_n r_n \int_0^{n/r_n} p_n (\lceil r_n s \rceil - 1)(1 - \lceil r_n s \rceil / n) P(M_{1,1+\lceil r_n s \rceil} \leq u_n \mid X_1 > u_n) \, \mathrm{d}s. \end{aligned}$$

By Lemmas 4.2 and 4.3, the dominated convergence theorem yields

$$n^{-1} p_n E(Y_n) \to \tau^2 \int_0^\infty s \theta e^{-\theta \tau s} \, \mathrm{d}s = \theta^{-1},$$

as required. □

Lemma 4.6. $n^{-2}p_n^2 E(Y_n^2) \to \theta^{-2}$

Proof. For integers $n \geq 1$, $1 \leq i \leq n-1$ and $1 \leq j \leq n-i$, abbreviate

$$I_{i,j} = I_{i,j}^{(n)} = I(X_i > u_n \geq M_{i,i+j}).$$

Then

$$E(Y_n^2) = \sum_{i=1}^{n-1}\sum_{j=1}^{n-i}\sum_{k=1}^{n-1}\sum_{l=1}^{n-k}(j-1)(l-1)E(I_{i,j}I_{k,l}).$$

The expectations can be computed as follows:

$$\begin{aligned}
E(I_{i,j}) &= E(I_{1,j}), \\
E(I_{i,j}I_{i,l}) &= E(I_{1,j\vee l}), \\
E(I_{i,j}I_{k,l}) &= 0, \quad \text{for } i < k \leq i+j, \\
E(I_{i,j}I_{k,l}) &= E(I_{1,j})E(I_{1,l}), \quad \text{for } k \geq i+j+m.
\end{aligned}$$

Hence $E(Y_n^2) = A_n + B_n + C_n$ where

$$\begin{aligned}
A_n &= \sum_{i=1}^{n-1}\sum_{j=1}^{n-i}\sum_{l=1}^{n-i}(j-1)(l-1)E(I_{1,j\vee l}), \\
B_n &= 2\sum_{i=1}^{n-1}\sum_{j=1}^{n-i}\sum_{k=i+j+1}^{i+j+m-1}\sum_{l=1}^{n-k}(j-1)(l-1)E(I_{1,j}I_{1,l}), \\
C_n &= 2\sum_{i=1}^{n-1}\sum_{j=1}^{n-i}\sum_{k=i+j+m}^{n-1}\sum_{l=1}^{n-k}(j-1)(l-1)E(I_{1,j})E(I_{1,l}).
\end{aligned}$$

For the terms A_n and B_n,

$$
\begin{aligned}
A_n &\leq n\sum_{j=1}^{n-1}\sum_{l=1}^{n-1}(j-1)(l-1)E(I_{1,j\vee l}),\\
B_n &\leq 2mn\sum_{j=1}^{n-1}\sum_{l=1}^{n-1}(j-1)(l-1)E(I_{1,j\vee l}).
\end{aligned}
$$

Changing sums gives

$$
\sum_{j=1}^{n-1}\sum_{l=1}^{n-1}(j-1)(l-1)E(I_{1,j\vee l}) \leq 2\sum_{j=1}^{n-1}(j-1)\sum_{l=j}^{n-1}(l-1)E(I_{1,l}) = \sum_{l=1}^{n-1}(l-1)^2 lE(I_{1,l}).
$$

By Lemma 4.2, $E(I_{1,l}) = P(X_1 > u_n \geq M_{1,1+l}) \leq p_n C\gamma^{l/r_n}$. Hence both $n^{-1}A_n$ and $n^{-1}B_n$ are bounded by some positive constant times $p_n\sum_{l=1}^{\infty}(l-1)^2 l\gamma^{l/r_n}$. As $n\to\infty$, this expression is $O\left\{p_n(1-\gamma^{1/r_n})^{-4}\right\} = O(p_n^{-3})$. Therefore, both A_n and B_n are $O(np_n^{-3}) = o(n^2p_n^{-2})$.

For the term C_n, change the order of summation to obtain

$$
\begin{aligned}
C_n &= 2\sum_{j=1}^{n}\sum_{l=1}^{n}(j-1)(l-1)E(I_{1,j})E(I_{1,l})\sum_{i=1}^{n-j}\sum_{k=i+j+m}^{n-l}1\\
&= \sum_{j=1}^{n}\sum_{l=1}^{n}(j-1)(l-1)(n-l-j-m+1)(n-l-j-m)_+E(I_{1,j})E(I_{1,l}).
\end{aligned}
$$

Write the sum as an integral:

$$\begin{aligned} C_n &= \int_0^n \int_0^n (\lceil s \rceil - 1)(\lceil t \rceil - 1)(n - \lceil s \rceil - \lceil t \rceil - m + 1) \\ &\qquad \times (n - \lceil s \rceil - \lceil t \rceil - m)_+ E(I_{1,\lceil s \rceil}) E(I_{1,\lceil t \rceil})\, \mathrm{d}s\, \mathrm{d}t \\ &= r_n^2 \int_0^{n/r_n} \int_0^{n/r_n} (\lceil r_n s \rceil - 1)(\lceil r_n t \rceil - 1)(n - \lceil r_n s \rceil - \lceil r_n t \rceil - m + 1) \\ &\qquad \times (n - \lceil r_n s \rceil - \lceil r_n t \rceil - m)_+ E(I_{1,\lceil r_n s \rceil}) E(I_{1,\lceil r_n t \rceil})\, \mathrm{d}s\, \mathrm{d}t. \end{aligned}$$

Since $E(I_{1,\lceil r_n s \rceil}) = p_n P(M_{1,1+\lceil r_n s \rceil} \leq u_n \mid X_1 > u_n)$, Lemmas 4.2 and 4.3 apply, giving

$$n^{-2} p_n^2 C_n \to \tau^4 \int_0^\infty \int_0^\infty st\theta e^{-\theta\tau s} \theta e^{-\theta\tau t}\, \mathrm{d}s\, \mathrm{d}t = \theta^{-2}$$

by dominated convergence. □

Chapter 5

Modelling Componentwise Maxima

5.0 Introduction

The class of multivariate extreme-value distributions (2.7) was shown in Chapters 2 and 3 to be an asymptotically justified model for componentwise maxima from multivariate sequences. Estimation within this class is the focus of this chapter.

Let $X = (X_1, \ldots, X_D)$ be a vector random variable with distribution function G and component distribution functions G_d, $1 \leq d \leq D$. Suppose for now that the components are standard Fréchet: $G_d(x_d) = \exp(-1/x_d)$ when $x_d > 0$. Let $x = (x_1, \ldots, x_D)$ be a realisation of X with radial component $r = x_1 + \ldots + x_D$ and angular component $w = (w_1, \ldots, w_D)$, where $w_d = x_d/r$. Recall from Theorem 2.3

that G is a multivariate extreme-value distribution function if

$$G(x) = \exp\{-V(x)\} \quad \text{for all } x \in \mathbb{R}_+^D,$$

where

$$V(x) = \frac{1}{r} \int_{S_D} \max_{1 \le d \le D} \left\{ \frac{z_d}{w_d} \right\} \mathrm{d}H(z)$$

for some positive, finite measure H on the simplex

$$S_D = \{z = (z_1, \ldots, z_D) \in \mathbb{R}_+^D \,:\, z_1 + \ldots + z_D = 1\}$$

for which

$$\int_{S_D} z_d \,\mathrm{d}H(z) = 1 \quad \text{for } 1 \le d \le D.$$

In this chapter, only the bivariate case, $D = 2$, is considered. Because H is restricted to the unit simplex,

$$V(x) = \frac{1}{r} \int_0^1 \max\left\{ \frac{z}{w}, \frac{1-z}{1-w} \right\} \mathrm{d}H(z),$$

where $w = w_1$ and $z = z_1$ lie in $[0, 1]$, and H is now a positive, finite measure on $[0, 1]$ satisfying

$$\int_0^1 z \,\mathrm{d}H(z) = \int_0^1 (1 - z) \,\mathrm{d}H(z) = 1. \tag{5.1}$$

It will be convenient to write the distribution function as

$$G(x) = \exp\left\{ -\frac{A(w)}{rw(1-w)} \right\}, \tag{5.2}$$

where A is called the dependence function and satisfies the integral relationship

$$A(w) = \int_0^1 \max\{z(1-w), w(1-z)\}\, \mathrm{d}H(z). \tag{5.3}$$

Attention will also be confined to differentiable distributions, for which G has density

$$g(x) = r^{-4}\{C(w) + rA''(w)\} \exp\left\{-\frac{A(w)}{rw(1-w)}\right\}, \tag{5.4}$$

where $C(w) = \{A(w) - wA'(w)\}\{A(w) + (1-w)A'(w)\}/\{w^2(1-w)^2\}$, and the derivative of H exists on $(0,1)$. In the differentiable case, the constraints implied by (5.1) and (5.3) can be written (Tawn, 1988) explicitly in terms of A:

$$\begin{gathered} A(0) = A(1) = 1, \\ -1 \leq A'(0) \leq 0 \leq A'(1) \leq 1, \\ A''(w) \geq 0 \quad \text{for all } 0 < w < 1. \end{gathered} \tag{5.5}$$

The graphical interpretation of these constraints is that A is convex, must lie within the triangle with vertices (0,1), (1,1) and $(\frac{1}{2}, \frac{1}{2})$, and is fixed at the first two vertices. The two bounding cases, $A(w) \equiv 1$ and $A(w) = \max\{w, 1-w\}$, correspond to independence and complete dependence between the components.

Let $\{(r_i, w_i)\}_{i=1}^n$ be a sample of independent realisations of the vector random variable (R, W), where $R = X_1 + X_2$, $W = X_1/R$ and (X_1, X_2) has bivariate extreme-value distribution (5.2). Estimation for the dependence function A currently falls into two camps: parametric and non-parametric methods, which will be reviewed in Sections 5.1 and 5.2. A semi-parametric approach that attempts to overcome the deficiencies of the parametric and non-parametric estimators is described in Section 5.3. The performances of several of the estimators are compared with a simulation study in Section 5.4 and their application is illustrated with a data example in Section 5.5.

5.1 Parametric Estimators

One approach to estimating the dependence function, A, is to choose a differentiable, parametric model that satisfies the constraints (5.5) and then fit the model by maximum likelihood. This approach has a number of benefits. Foremost, the resulting estimate of G is guaranteed to be a valid extreme-value distribution function. An estimate of the density is also obtained so that the likelihood is available. Joint estimation of the dependence structure and the parameters of the component distributions is therefore possible, which can have considerable benefits, as found, for example, by Barão and Tawn (1999). This also allows uncertainty in component parameters to be incorporated into confidence statements about G; covariate information can be included in the model; and likelihood-based tests of model features (such as independence between components or equality of component shape parameters) can be performed.

The disadvantage of the parametric approach is that it restricts attention to only a subset of the possible dependence functions. Therefore, sufficiently flexible models must be used. The construction of flexible models that satisfy the constraints (5.5) is a difficult task, particularly in higher dimensions, and, while there are now quite a few parametric models available, there is still a requirement for more flexibility, as found by Joe *et al.* (1992).

One parametric model is the logistic model, introduced by Gumbel (1960) and extended by Tawn (1988). The model for the dependence function is

$$A(w) = \{(1-w)^{1/\alpha} + w^{1/\alpha}\}^{\alpha} \tag{5.6}$$

for a dependence parameter $\alpha \in (0, 1]$. The model reaches independence when $\alpha = 1$ and approaches complete dependence as $\alpha \to 0$. The logistic model is not a particularly flexible class however: the dependence function is symmetric and $A'(1) = -A'(0) = 1$.

Another model is the Dirichlet model, introduced by Coles and Tawn (1991). The model for the dependence function is

$$A(w) = (1-w)\{1 - \mathrm{Be}(\alpha_1 + 1, \alpha_2; z)\} + w\mathrm{Be}(\alpha_1, \alpha_2 + 1; z) \tag{5.7}$$

for positive dependence parameters α_1 and α_2, where $z = \alpha_1 w / \{\alpha_1 w + \alpha_2(1-w)\}$ and

$$\mathrm{Be}(a, b; z) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^z y^{a-1}(1-y)^{b-1}\,\mathrm{d}y$$

is the incomplete Beta function. The model is asymmetric unless $\alpha_1 = \alpha_2 = \alpha$, in which case the model reaches independence at $\alpha = 0$ and approaches complete dependence as $\alpha \to \infty$. Again, $A'(1) = -A'(0) = 1$.

Other parametric models include the negative logistic (Galambos, 1975), asymmetric logistic (Tawn, 1988), Gaussian (Hüsler and Reiss, 1989), negative asymmetric logistic (Joe, 1989), bilogistic (Joe *et al.*, 1992) and negative bilogistic (Coles and Tawn, 1994). All of these are listed in Kotz and Nadarajah (2000).

5.2 Non-parametric Estimators

In order to avoid the restrictions imposed by parametric models for the dependence function, non-parametric estimators have been developed. Unfortunately, this approach has a number of drawbacks. It is particularly difficult to ensure that non-parametric estimates of A satisfy the constraints (5.5), so that crude adjustments are typically required. The estimators are often not differentiable on the whole interval $[0, 1]$; consequently, G contains point masses, the likelihood has degeneracies and many of the benefits of the parametric approach are out of reach. A good non-parametric estimate of A can, nevertheless, be a valuable guide to selecting a suitable parametric model. With this motivation, two non-parametric estimators are now reviewed.

The first is introduced by Capéraà *et al.* (1997). Making the change of variables $(x_1, x_2) \to (r, w)$, the density (5.4) becomes

$$g_{R,W}(r, w) = r^{-3}\{C(w) + rA''(w)\} \exp\left\{-\frac{A(w)}{rw(1-w)}\right\}$$

and it follows that the distribution function of the angular component is

$$G_W(w) = w + w(1-w)A'(w)/A(w). \tag{5.8}$$

Capéraà *et al.* (1997) use this to derive an estimator based on the empirical distribution function, $\hat{G}_W$, of the angular components, $\{w_i\}_{i=1}^n$, specifically

$$A_c(w) = \exp\left\{(1-w)\int_0^w \frac{\hat{G}_W(z)-z}{z(1-z)}\,\mathrm{d}z - w\int_w^1 \frac{\hat{G}_W(z)-z}{z(1-z)}\,\mathrm{d}z\right\}. \tag{5.9}$$

While this estimator does satisfy some of the conditions (5.5), it is not necessarily convex, nor does it necessarily lie above $\max\{w, 1-w\}$. A valid estimator can be obtained by enforcing the constraints artificially, that is by taking the convex hull of $\max\{A_c(w), w, 1-w\}$. Note, however, that taking the convex hull pulls the estimator towards stronger dependence.

The second non-parametric estimator is a modification proposed by Hall and Tajvidi (2000) of the estimator of Pickands (1981). Pickands uses the fact that the quantity $1/\max\{(1-w)X_1, wX_2\}$ has an exponential distribution with mean $1/A(w)$ to motivate the piecewise-linear estimator

$$n\left[\sum_{i=1}^n \min\left\{\frac{1}{(1-w)x_{i1}}, \frac{1}{wx_{i2}}\right\}\right]^{-1}. \tag{5.10}$$

The asymptotic properties of this estimator, which is also the maximum-likelihood estimator for $A(w)$ when w is fixed, are investigated by Deheuvels (1991). The estimator is not guaranteed to satisfy any of the constraints (5.5) however. Hall and Tajvidi (2000) propose a normalised version,

$$A_p(w) = n\left[\sum_{i=1}^{n} \min\left\{\frac{\bar{x}_1}{(1-w)x_{i1}}, \frac{\bar{x}_2}{wx_{i2}}\right\}\right]^{-1}, \tag{5.11}$$

where $\bar{x}_d = n/\sum_{i=1}^{n} x_{id}^{-1}$. This ensures that $A(0) = A(1) = 1$, as noted by Deheuvels (1985). Again, a valid estimator is obtained by taking the convex hull of $\max\{A_p(w), w, 1-w\}$.

These estimates of the dependence function are not differentiable on the whole of $[0, 1]$. Two attempts have been made at smoothing that yield differentiable estimates. The first is proposed by Hall and Tajvidi (2000), who apply spline smoothing to both A_c and A_p. By imposing constraints on the spline functions they also ensure that the resulting estimates are valid dependence functions. The second approach is proposed by Smith *et al.* (1990), who smooth Pickands's estimator (5.10) with a kernel function. The idea of kernel smoothing will appear again in the following section.

5.3 Semi-parametric Estimators

In this section, a new, semi-parametric approach to the problem of estimating the dependence function is introduced. The approach attempts to combine the flexibility of non-parametric methods with the coherency of parametric models. The estimator yields estimates of A and its first two derivatives that are self-consistent and satisfy naturally the criteria (5.5) of a dependence function. This makes the likelihood available. The estimator is also flexible and differentiable as it is constructed using kernel density estimation. Of the current non-parametric methods, only the constrained spline-smoothing approach of Hall and Tajvidi (2000) enjoys the same benefits.

5.3.1 Asymptotic motivation

Recall that, for differentiable models, the measure H has a density, h, on $(0,1)$. Point masses, $0 \leq \phi_0 \leq 1$ and $0 \leq \phi_1 \leq 1$, can also be admitted at zero and one (Smith, 1985). In this case, the integral relationship (5.3) can be rewritten as

$$A(w) = w\phi_0 + (1-w)\phi_1 + \int_{(0,1)} \max\{z(1-w), w(1-z)\}h(z)\,\mathrm{d}z. \tag{5.12}$$

Differentiating yields

$$A'(w) = \phi_0 - \phi_1 + \int_{(0,w)} (1-z)h(z)\,\mathrm{d}z - \int_{(w,1)} zh(z)\,\mathrm{d}z \tag{5.13}$$

and

$$A''(w) = h(w). \tag{5.14}$$

Note that $A'(0) = -1 + \phi_0$ and $A'(1) = 1 - \phi_1$ so that allowing non-zero point-masses at zero and one increases the flexibility of A. The moment constraints (5.1) can also be rewritten as

$$\phi_0 + \int_{(0,1)} (1-z)h(z)\,\mathrm{d}z = \phi_1 + \int_{(0,1)} zh(z)\,\mathrm{d}z = 1. \tag{5.15}$$

The plan is to estimate ϕ_0, ϕ_1 and h under constraint (5.15). Substituting the estimates into equations (5.12) – (5.14) will yield estimates of A, A' and A'' that automatically define a valid estimate of the extreme-value density (5.4). This approach contrasts with the non-parametric estimators encountered in Section 5.2, which estimate A directly, and has the advantage that it is relatively simple to ensure that an estimator for h satisfies the appropriate conditions. This is likely to be particularly useful in higher dimensions.

To see how h might be estimated, consider the distribution of the angular component, W, conditional on a large radial component, R. By Theorem 3 of de Haan and Resnick (1977),

$$\lim_{r\to\infty} P(W \leq w \mid R > r) = \frac{H\{(0,w)\}}{H\{(0,1)\}} \quad \text{for } w \in (0,1)$$

as long as H has no mass at zero or one, that is $\phi_0 = \phi_1 = 0$. The main assumption in what follows is that the limit holds for the density,

$$q(w) = \lim_{r\to\infty} \frac{\mathrm{d}}{\mathrm{d}w} P(W \leq w \mid R > r) = \frac{h(w)}{H\{(0,1)\}} \quad \text{for } w \in (0,1), \tag{5.16}$$

and when H has point masses at zero and one. If this is the case then information about h will be found in the angular component of observations with a large radial component.

5.3.2 Kernel estimator

Rewriting the limit (5.16) yields

$$h(w) = \phi q(w) \quad \text{for } w \in (0,1),$$

where $\phi = H\{(0,1)\} = (1-\phi_0)+(1-\phi_1)$ by the moment conditions (5.15). Suppose for now that the values of ϕ_0 and ϕ_1 are known. When $\phi \neq 0$, the problem reduces to estimating the limiting angular density, q, subject to the constraints

$$\int_{(0,1)} q(z)\,\mathrm{d}z = 1 \quad \text{and} \quad \int_{(0,1)} zq(z)\,\mathrm{d}z = \psi, \tag{5.17}$$

where $\psi = (1-\phi_1)/\phi$. If $\phi = 0$ then q is arbitrary.

The shape of q influences the shapes of A and its derivatives on $(0,1)$. Parametric models force a structure on this shape that can be quite restrictive, particularly if a symmetric model is chosen. Any restriction can be avoided by estimating q non-parametrically.

First choose a radial threshold above which the limit (5.16) is believed to hold approximately. For some $1 \leq m \leq n$, the threshold used here is r_m, the m-th largest of the ordered radial components, $r_1 > \ldots > r_n$. Re-label $w_1, \ldots, w_n$ so that w_i is the angle that occured with r_i. A kernel density estimate of q is

$$q_n(w) = \frac{1}{m\lambda} \sum_{i=1}^{m} \tilde{k}\left(\frac{w - w_i}{\lambda}\right), \tag{5.18}$$

based on the angular components of the m largest observations. Define the kernel, $\tilde{k}$, to be

$$\tilde{k}\left(\frac{w-z}{\lambda}\right) = k\left(\frac{w-z}{\lambda}\right) + k\left(\frac{w+z}{\lambda}\right) + k\left(\frac{w-2+z}{\lambda}\right),$$

where k is the standard Gaussian density and λ is a smoothing parameter. This choice of kernel compensates for the mass that would spill over the boundaries of $(0,1)$ if a simple kernel such as k were used instead.

The estimator (5.18) is not constrained to have the required mean (5.17), but this can be achieved with a simple power transformation:

$$q_n^*(w) = \frac{1}{m\lambda}\sum_{i=1}^{m} \tilde{k}\left(\frac{w-w_i^{\kappa}}{\lambda}\right),$$

where $\kappa \geq 0$ solves

$$\int_{(0,1)} z q_n^*(z)\,\mathrm{d}z = \psi. \tag{5.19}$$

Note that the more common approach (Hall and Presnell, 1999) of modifying the weights $1/m$ rather than the angles w_i is not applicable here since the mean ψ could be outside the range of the w_i. An alternative would be to modify both the weights and angles, an example of which is the generalisation of Theorem 2 of Coles and Tawn (1991) described in Section 3.7.1 of Coles (1991), but this option is not considered here.

An estimator has been defined for q given ϕ_0, ϕ_1, m and λ. The four parameters could be chosen to maximise the likelihood

$$L(\phi_0, \phi_1, m, \lambda) = \prod_{i=1}^{n} g_n(x_i),$$

where g_n is obtained from (5.4) by replacing A, A' and A'' with the estimators

$$\begin{aligned} A_n''(w) &= \phi q_n^*(w), \\ A_n'(w) &= \phi_0 - \phi_1 + \phi \int_{(0,w)} (1-z) q_n^*(z)\,\mathrm{d}z - \phi \int_{(w,1)} z q_n^*(z)\,\mathrm{d}z, \\ A_n(w) &= w\phi_0 + (1-w)\phi_1 + \phi \int_{(0,1)} \max\{w(1-z), z(1-w)\} q_n^*(z)\,\mathrm{d}z. \end{aligned} \tag{5.20}$$

Unfortunately, the likelihood has a singularity when λ equals zero. To see this, consider m fixed and note that, as $\lambda \to 0$, $A_n''(w_i^\kappa) \to \infty$ for each $1 \leq i \leq m$. The likelihood contains the terms $A_n''(w_i)$, $1 \leq i \leq n$. Therefore, ϕ_0 and ϕ_1 will be chosen so that $\psi = \bar{w}$, the mean of $w_1, \ldots, w_m$, since then $\kappa \to 1$ as $\lambda \to 0$. Instead of estimating all four parameters with the likelihood, only ϕ_0, ϕ_1 and m will be estimated; λ is fixed as a function of m using the over-smoothing formula (Wand and Jones, 1995, page 61) for a standard Gaussian kernel, that is

$$\lambda = s\left(\frac{243}{70 m \pi^{3/2}}\right)^{1/5}, \tag{5.21}$$

where s is an estimate of the standard deviation of q. When an automatic choice of smoothing parameter is required, s will be set equal to the median of $s_1, \ldots, s_n$, where, for $1 \leq m \leq n$,

$$s_m^2 = m^{-1} \sum_{i=1}^{m} (w_i - \bar{w}_m)^2 \quad \text{and} \quad \bar{w}_m = m^{-1} \sum_{i=1}^{m} w_i. \tag{5.22}$$

A more subjective choice is made in the data example of Section 5.5.

By fixing λ in this way, the mean of q_n^* is constrained to lie in the interval $(\psi_\infty, \psi_0]$, where

$$\psi_\infty = \frac{1}{\lambda} \int_{(0,1)} z\tilde{k}\left(\frac{z}{\lambda}\right) \mathrm{d}z \quad \text{and} \quad \psi_0 = \frac{1}{\lambda} \int_{(0,1)} z\tilde{k}\left(\frac{z-1}{\lambda}\right) \mathrm{d}z.$$

Through (5.19) and (5.17), this imposes an additional linear constraint,

$$\frac{\psi_\infty}{1-\psi_\infty} < \frac{1-\phi_1}{1-\phi_0} \leq \frac{\psi_0}{1-\psi_0},$$

on ϕ_0 and ϕ_1.

Note finally that, because the full likelihood is used, all of the data contribute to the estimates of ϕ_0, ϕ_1, m and, through these, the dependence function and its derivatives.

5.3.3 Symmetric kernel estimator

A symmetric version of the estimator can also be defined by setting $\phi_0 = \phi_1$, $\kappa = 1$ and

$$q_n^*(w) = \frac{1}{2m\lambda} \sum_{i=1}^{m} \left\{ \tilde{k}\left(\frac{w - w_i}{\lambda}\right) + \tilde{k}\left(\frac{w - 1 + w_i}{\lambda}\right) \right\}.$$

The only constraints now are that ϕ_0 and ϕ_1 lie in $[0,1]$. The smoothing parameter becomes

$$\lambda = s\left(\frac{243}{140m\pi^{3/2}}\right)^{1/5}$$

and $s_m^2 = m^{-1}\sum_{i=1}^{m}(w_i - 1/2)^2$ is used to guide selection of s. The performance of this symmetric estimator will be compared with the unrestricted estimator in the simulation study of Section 5.4.

5.3.4 Unknown component distributions

Although the component distributions have been assumed to be standard Fréchet, this is not necessary and the maximum likelihood estimates of the component parameters can be found simultaneously. Suppose that the d-th component, here represented by $\tilde{X}_d$, has generalised extreme-value distribution with parameters (μ_d, σ_d, ξ_d). Then, transforming via (2.6) to standard Fréchet components

$$X_d = \left\{1 + \xi_d\left(\frac{\tilde{X}_d - \mu_d}{\sigma_d}\right)\right\}^{1/\xi_d}$$

yields the likelihood required for joint estimation. The terms in the likelihood are of the form

$$g(x)\frac{x_1^{1-\xi_1}x_2^{1-\xi_2}}{\sigma_1\sigma_2},$$

where the component parameters, (μ_1, σ_1, ξ_1) and (μ_2, σ_2, ξ_2), are constrained so that x_1 and x_2 are positive.

5.3.5 Asymptotic properties

Asymptotic properties for the kernel estimators described in the previous subsections have not been found, but uniform consistency can be shown for simplified estimators in a restricted setting. Define $g_{W|R}(w \mid r)$ to be the density of $P(W \leq w \mid R > r)$ and suppose that the following conditions hold:

$$\phi_0 = \phi_1 = 0, \tag{5.23}$$

$$\lim_{w \to 0} q(w) = \lim_{w \to 1} q(w) = 0, \tag{5.24}$$

$$\lim_{r \to \infty} \sup_{w \in [0,1]} |g_{W|R}(w \mid r)| < \infty, \tag{5.25}$$

$$\lim_{r \to \infty} \sup_{w \in [0,1]} |g'_{W|R}(w \mid r)| < \infty. \tag{5.26}$$

Consider the estimators (5.20) for A, A' and A'' when the parameters ϕ_0 and ϕ_1 are known to be zero and the kernel density estimator q_n^* is replaced with q_n, the simpler version (5.18) that does not incorporate the power transformation.

For the density, $g_{W|R}(\cdot \mid r_m)$, estimated by $q_n(\cdot)$ to approach $q(\cdot)$, it is necessary that r_m approaches infinity as the sample size n increases. For consistency, it is also necessary that $m \to \infty$, $\lambda \to 0$ and $m\lambda^2 \to \infty$ as $n \to \infty$ (Parzen, 1962). Recall that $\lambda \propto m^{-1/5}$, so this is achieved if $m = m_n = o(n)$ and $m_n \to \infty$ as $n \to \infty$.

Uniform consistency of A''_n follows from

$$\begin{aligned} E \sup_{w\in[0,1]} |A''_n(w) - A''(w)| &\leq 2E \sup_{w\in[0,1]} |q_n(w) - g_{W|R}(w \mid r_m)| \\ &\quad + 2 \sup_{w\in[0,1]} |g_{W|R}(w \mid r_m) - q(w)|. \end{aligned}$$

The first term approaches zero by Theorem 3A of Parzen (1962), where conditions (5.25) and (5.26) are required instead of his uniform continuity conditions because the density being estimated changes with n. The second term tends to zero under conditions (5.23) and (5.24) by Theorem 1 of Nadarajah (2000). Uniform consistency of A_n follows from

$$\begin{aligned} |A_n(w) - A(w)| &\leq w \int_0^w (1-z)|A''_n(z) - A''(z)|\,\mathrm{d}z \\ &\quad + (1-w)\int_w^1 z|A''_n(z) - A''(z)|\,\mathrm{d}z \\ &\leq \sup_{z\in[0,1]} |A''_n(z) - A''(z)|, \end{aligned}$$

and similarly for A'_n.

The conditions under which these results hold are strong: for example, condition (5.24) holds for the logistic model (5.6) only when $\alpha < 1/2$, and for the Dirichlet model (5.7) only when $\alpha_1 > 1$ and $\alpha_2 > 1$. It will be seen in the following section, however, that the estimators perform well for finite sample sizes even when conditions (5.23) – (5.26) are not satisfied.

5.4 Simulation Study

The performances of the parametric, non-parametric and semi-parametric estimators of the three previous sections are compared here for small samples with a simulation study. To simulate data from a bivariate extreme-value distribution with arbitrary dependence function, A, and standard Fréchet margins, the following algorithm (Ghoudi *et al.*, 1998) is employed. First, generate a uniform(0,1) variate, u, and then generate the angular component from distribution (5.8) by solving $G_W(w) = u$ for w. To generate the radial component, use the fact that the conditional density, $g_{R|W}(\cdot \mid W = w)$, is a mixture of two inverse Gamma densities:

$$g_{R|W}(r \mid W = w) = \pi p(r\,;\,1, \beta) + (1 - \pi)p(r\,;\,2, \beta),$$

where $\beta = A(w)/w(1-w)$, $\pi = [1 + w(1-w)C(w)/\{A(w)A''(w)\}]^{-1}$ and $p(\cdot\,;\,a, b)$ is an inverse Gamma density with parameters a and b.

Data are simulated from two bivariate extreme-value distributions: one with symmetric, logistic dependence (5.6) and another with asymmetric, Dirichlet dependence (5.7). For each model, three different sets of parameter values are considered: for the logistic model, $\alpha = 0.25, 0.5$ and 0.75; for the Dirichlet model, $(\alpha_1, \alpha_2) = (1, 9), (\frac{1}{3}, 9)$ and $(\frac{1}{9}, 1)$. These parameter values are listed in order of weakening asymptotic dependence; the shapes of the true dependence functions and their second derivatives are shown in Figures 5.1 and 5.2.

For each of the six dependence structures, samples of sizes 25, 50 and 100 are

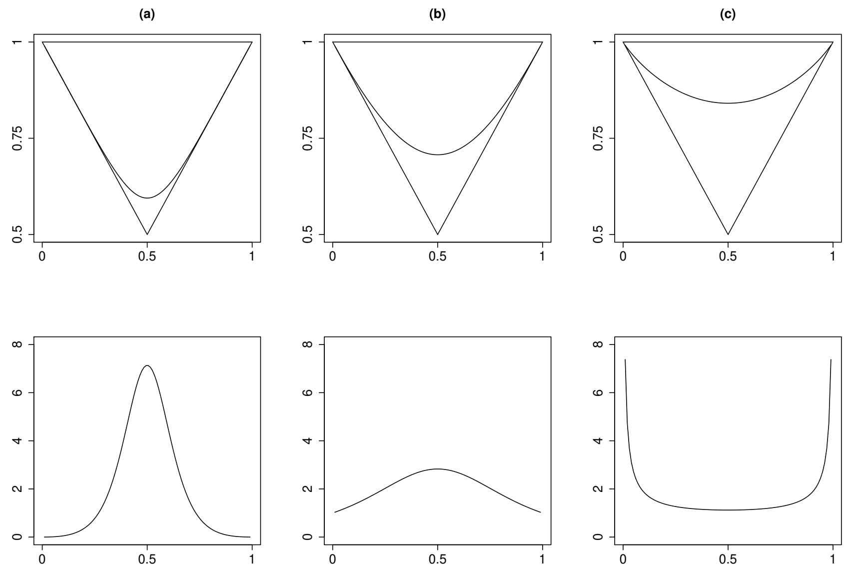


Figure 5.1: The dependence functions A (above) and A'' (below) of the three logistic models. The dependence parameters are (a) 0.25, (b) 0.5, (c) 0.75.

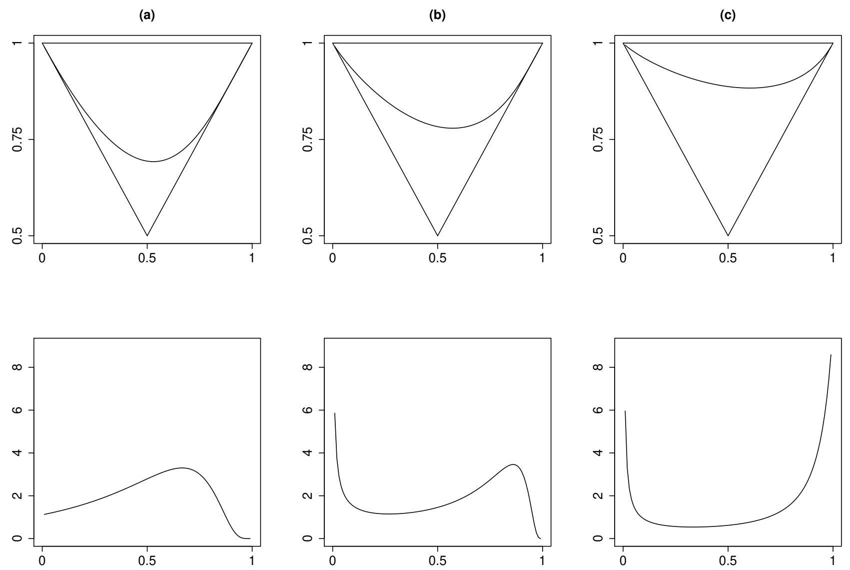


Figure 5.2: The dependence functions A (above) and A'' (below) of the three Dirichlet models. The dependence parameters are (a) $(1, 9)$, (b) $(\frac{1}{3}, 9)$, (c) $(\frac{1}{9}, 1)$.

generated. In each case the performances of four estimators for the dependence function are compared to that of the true, parametric model (labelled M) fitted by maximum likelihood. The four estimators are the kernel estimator (K) defined in (5.20), the symmetric kernel estimator (Ks), the estimator (C) of Capéraà *et al.* (1997) defined in (5.9) and the modified Pickands estimator (P) of Hall and Tajvidi (2000) defined in (5.11). Kernel estimates of A' and A'' are also compared to those of the true model: no such results for the non-differentiable, non-parametric estimators C and P are presented. The component distributions are assumed known and performance is summarised by Monte Carlo estimates of mean-integrated-square error (MISE) based on one-hundred replicated samples. The results are shown in Tables 5.1, 5.2 and 5.3.

Consider first the results in Table 5.1 for the estimates of A. Of the two non-parametric estimators, C almost always outperforms P except for the logistic model with strong dependence. Compare now the two kernel estimators. As expected, the symmetric version (Ks) is preferred only when the data are generated from the symmetric, logistic model and this superiority decreases as dependence weakens, where there is less scope for asymmetry. Impressively, the performance of Ks for the logistic data is almost equal to that of the true model fitted by maximum likelihood. The unconstrained kernel estimator (K) also performs well, typically outperforming both of the non-parametric estimators and occasionally matching the true model when the sample size is at its smallest.

The spline smoothing approach of Hall and Tajvidi (2000) has not been implemented but a tentative comparison of the results in Table 5.1 can be made with their simulation study in the case of the logistic dependence structure with $\alpha = 0.5$. Both studies produce the same ordering of the estimators M, C and P, and yield similar values of mean-integrated-square error. Hall and Tajvidi find that spline smoothing reduces the mean-integrated-square error for C by 31, 32 and 19%, and for P by 47, 30 and 21% when the sample size is 25, 50 and 100. If this improvement were replicated in the current study then estimator C would become superior to K at all sample sizes; estimator P would become superior to K at sample size 25.

Estimating the derivatives, A' and A'', is more difficult, as made clear by Tables 5.2 and 5.3. Nevertheless, the performance of the kernel estimator is encouraging, particularly for small samples of the Dirichlet data, where fitting the true model does not give vastly superior estimates. Again, the symmetric kernel estimator is significantly better than the unconstrained version for the logistic data.

The estimators are also applied to data simulated from the logistic model with $\alpha = 1$, that is independence between components. The results are in Table 5.4 and provoke comments similar to those recorded above for the other dependence structures.

It seems that the kernel estimator, K, is worthy of consideration as an estimator for the dependence function and its derivatives. Its overall performance for the small samples considered here is significantly better than that of the non-parametric estimators examined, particularly when asymmetry is present, and often compares favourably with maximum-likelihood fitting of the true model. Moreover, some

	Logistic(0.25)						Dirichlet(1,9)					
	25		50		100		25		50		100	
M	8	(1)	3	(1)	2	(0)	59	(10)	22	(4)	11	(1)
C	24	(2)	13	(2)	6	(1)	96	(11)	46	(7)	22	(2)
P	16	(2)	7	(1)	4	(1)	107	(12)	51	(8)	27	(6)
K	15	(2)	8	(2)	2	(0)	67	(9)	36	(6)	17	(2)
Ks	8	(1)	3	(1)	1	(0)	76	(10)	34	(6)	19	(2)

	Logistic(0.5)						Dirichlet($\frac{1}{3}, 9$)					
	25		50		100		25		50		100	
M	83	(12)	31	(6)	16	(2)	144	(22)	49	(7)	27	(4)
C	111	(12)	53	(8)	26	(3)	257	(37)	105	(19)	57	(7)
P	142	(14)	70	(11)	35	(7)	325	(38)	146	(24)	76	(14)
K	122	(20)	49	(8)	22	(2)	191	(27)	79	(10)	49	(6)
Ks	93	(13)	34	(6)	19	(3)	187	(29)	89	(12)	60	(5)

	Logistic(0.75)						Dirichlet($\frac{1}{9}, 1$)					
	25		50		100		25		50		100	
M	264	(37)	104	(18)	59	(8)	237	(26)	106	(14)	63	(8)
C	397	(54)	168	(28)	91	(11)	554	(88)	215	(36)	119	(14)
P	576	(69)	258	(40)	131	(21)	794	(116)	327	(50)	169	(26)
K	264	(29)	118	(16)	75	(8)	209	(21)	122	(16)	84	(10)
Ks	273	(38)	119	(20)	66	(9)	275	(29)	134	(16)	95	(11)

Table 5.1: MISE ($\times 10^5$) for five estimators for A and six models of dependence. Columns are labelled with sample size; estimated standard errors are in brackets.

features of the kernel estimator might be improved, in particular the choice of smoothing parameter which is far from optimal. A detailed comparison with the spline smoothing estimators of Hall and Tajvidi (2000) is needed to discriminate the performances of their method and the kernel estimator. Finally, notice that it is possible to apply kernel-smoothing ideas to non-parametric estimators such as C and P. These estimators for the dependence function A imply a point-mass function for A'' that could be smoothed and adjusted to ensure the appropriate moment conditions, and then integrated to obtain estimates for A and A'.

	Logistic(0.25)						Dirichlet(1,9)					
	25		50		100		25		50		100	
M	23	(3)	10	(2)	5	(1)	95	(14)	39	(6)	18	(2)
K	53	(5)	27	(3)	12	(1)	130	(12)	74	(8)	36	(3)
Ks	30	(4)	14	(2)	7	(1)	129	(12)	72	(7)	49	(3)
	Logistic(0.5)						Dirichlet($\frac{1}{3}$, 9)					
	25		50		100		25		50		100	
M	95	(13)	37	(7)	20	(3)	199	(26)	72	(10)	35	(5)
K	184	(23)	89	(10)	43	(4)	279	(30)	137	(13)	81	(7)
Ks	124	(14)	50	(7)	28	(4)	279	(28)	171	(12)	136	(6)
	Logistic(0.75)						Dirichlet($\frac{1}{9}$, 1)					
	25		50		100		25		50		100	
M	260	(37)	102	(18)	58	(8)	315	(28)	144	(17)	87	(10)
K	358	(28)	184	(17)	117	(10)	320	(23)	199	(19)	133	(11)
Ks	292	(36)	139	(20)	78	(8)	353	(29)	206	(16)	157	(11)

Table 5.2: MISE ($\times 10^4$) for three estimators for A' and six models of dependence. Columns are labelled with sample size; estimated standard errors are in brackets.

	Logistic(0.25)						Dirichlet(1,9)					
	25		50		100		25		50		100	
M	23	(3)	11	(2)	6	(1)	48	(6)	22	(3)	10	(1)
K	80	(8)	41	(3)	23	(2)	71	(6)	45	(4)	26	(2)
Ks	50	(6)	26	(3)	15	(2)	71	(5)	48	(3)	36	(2)
	Logistic(0.5)						Dirichlet($\frac{1}{3}$, 9)					
	25		50		100		25		50		100	
M	34	(4)	14	(2)	7	(1)	108	(14)	40	(6)	18	(3)
K	72	(6)	44	(3)	26	(2)	154	(9)	111	(5)	89	(3)
Ks	52	(5)	27	(3)	17	(2)	136	(6)	109	(4)	98	(2)
	Logistic(0.75)						Dirichlet($\frac{1}{9}$, 1)					
	25		50		100		25		50		100	
M	58	(9)	19	(4)	9	(1)	288	(25)	175	(16)	125	(15)
K	224	(7)	175	(6)	149	(5)	273	(14)	221	(12)	182	(7)
Ks	176	(8)	142	(6)	118	(5)	290	(10)	246	(8)	213	(6)

Table 5.3: MISE ($\times 10^2$) for three estimators for A'' and six models of dependence. Columns are labelled with sample size; estimated standard errors are in brackets.

	25		50		100	
M	184	(38)	94	(24)	42	(11)
C	1091	(146)	473	(66)	241	(27)
P	1660	(261)	636	(88)	324	(47)
K	312	(43)	163	(27)	75	(13)
Ks	225	(42)	113	(25)	53	(12)

	25		50		100	
M	181	(37)	93	(23)	41	(11)
K	384	(48)	214	(32)	97	(16)
Ks	224	(42)	115	(26)	54	(12)

	25		50		100	
M	77	(13)	53	(9)	30	(6)
K	141	(17)	100	(16)	46	(9)
Ks	36	(8)	22	(6)	11	(3)

Table 5.4: MISE ($\times 10^5$, $\times 10^4$ and $\times 10^2$) for estimators for A (top), A' (middle) and A'' (bottom) under independence. Columns are labelled with sample size; estimated standard errors are in brackets.

5.5 Data Example

In this section, the estimators of Sections 5.1, 5.2 and 5.3 are applied to a bivariate time series of seventy-four annual maximum sea-levels recorded in metres at two sites, Lowestoft (X_1) and Sheerness (X_2), on the east coast of England between 1899 and 1983. The data are plotted in Figure 5.3. First, the component distributions are estimated and fixed as the dependence structure is estimated. These results are compared with those obtained by estimating the component parameters and the dependence structure simultaneously.

Following Tawn (1988), linear trends are included in the location parameters of the component generalised extreme-value distributions by replacing (μ_1, μ_2) with $(\alpha_1 + \beta_1 t, \alpha_2 + \beta_2 t)$, where t is a time covariate for which an increment of 0.01 corresponds to one year and $t = 0$ corresponds to 1941. Fitting the components

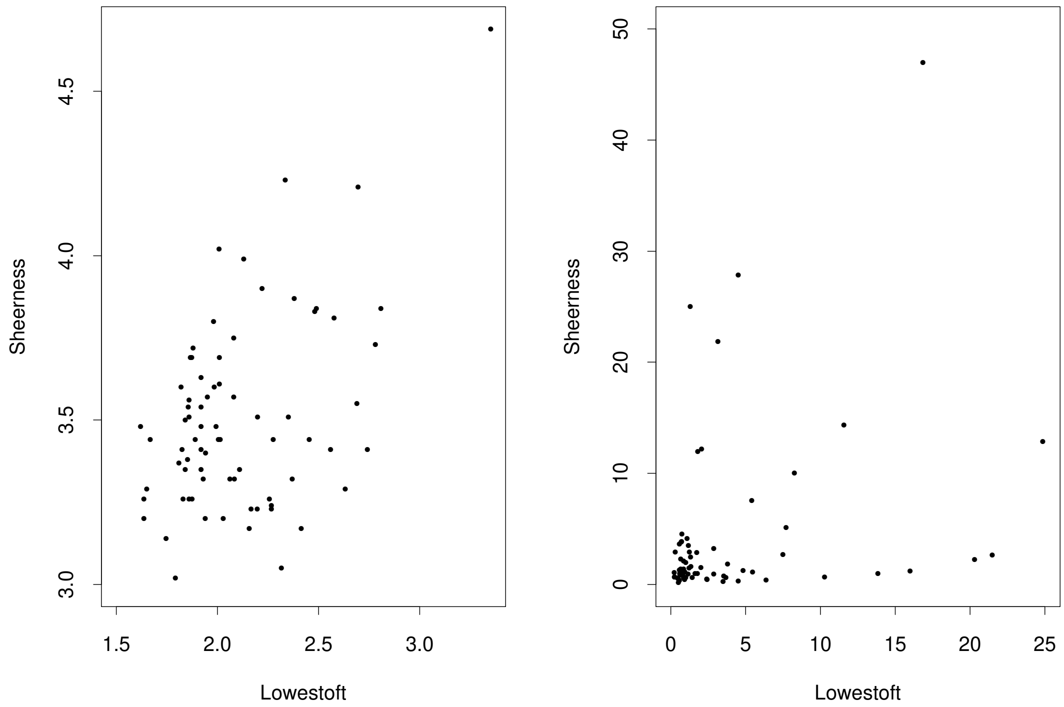


Figure 5.3: Annual maximum sea-levels (metres) recorded at Sheerness against those recorded at Lowestoft. The raw data are plotted on the left, and the data transformed to standard Fréchet components by the independence fit are plotted on the right. The outlier is not shown in the right-hand plot: it has coordinates (126, 236).

separately yields the parameter estimates displayed in the bottom half of Table 5.5. The data transformed to standard Fréchet components using these estimates are also plotted in Figure 5.3.

Having transformed to standard Fréchet components the dependence function can be estimated using the non-parametric estimators (5.9) and (5.11). These estimates are plotted in Figure 5.4; they have not been made convex for illustrative purposes. The estimates point to stronger dependence when w is less than one-half, which corresponds to relatively larger sea-levels at Sheerness than at Lowestoft.

Joint	α	β	σ	ξ
Lowestoft	1.95	0.050	0.23	0.070
Sheerness	3.39	0.379	0.19	0.019
Independence	α	β	σ	ξ
Lowestoft	1.95	0.032	0.23	0.083
Sheerness	3.39	0.369	0.19	0.072

Table 5.5: Component parameters estimated jointly (top) and under independence (bottom).

Another feature is the very strong dependence at w close to 0.94. This corresponds to the string of observations along the bottom of the right-hand plot in Figure 5.3.

For comparison, a parametric model is also fitted, the asymmetric logistic, with dependence function

$$A(w) = (1-\theta_1)(1-w) + (1-\theta_2)w + \{\theta_1^{1/\alpha}(1-w)^{1/\alpha} + \theta_2^{1/\alpha}w^{1/\alpha}\}^\alpha,$$

where $0 \le \theta_1 \le 1$, $0 \le \theta_2 \le 1$ and $0 \le \alpha \le 1$. The parameter estimates are $\theta_1 = 0.33$, $\theta_2 = 1.00$ and $\alpha = 0.51$, the log-likelihood is -310.20, and the estimated dependence function is shown in Figure 5.4. The principal difference to the non-parametric estimates is near $w = 1$: the parametric model places mass $1-\theta_1 = 0.67$ at $w = 1$, interpreting the aforementioned string of points in Figure 5.3 as events at Lowestoft that are independent of the sea-level at Sheerness.

For the kernel estimator, the smoothing parameter, λ, is set using the over-smoothing formula (5.21) with the standard deviation, s, chosen by examining the plot, shown in Figure 5.5, of s_m (5.22) for $m = 1, \ldots, n$. There is the usual trade-off between bias and variability as m decreases, but the estimates appear to stabilise around $s = 0.31$. With this choice for s, the parameter estimates are

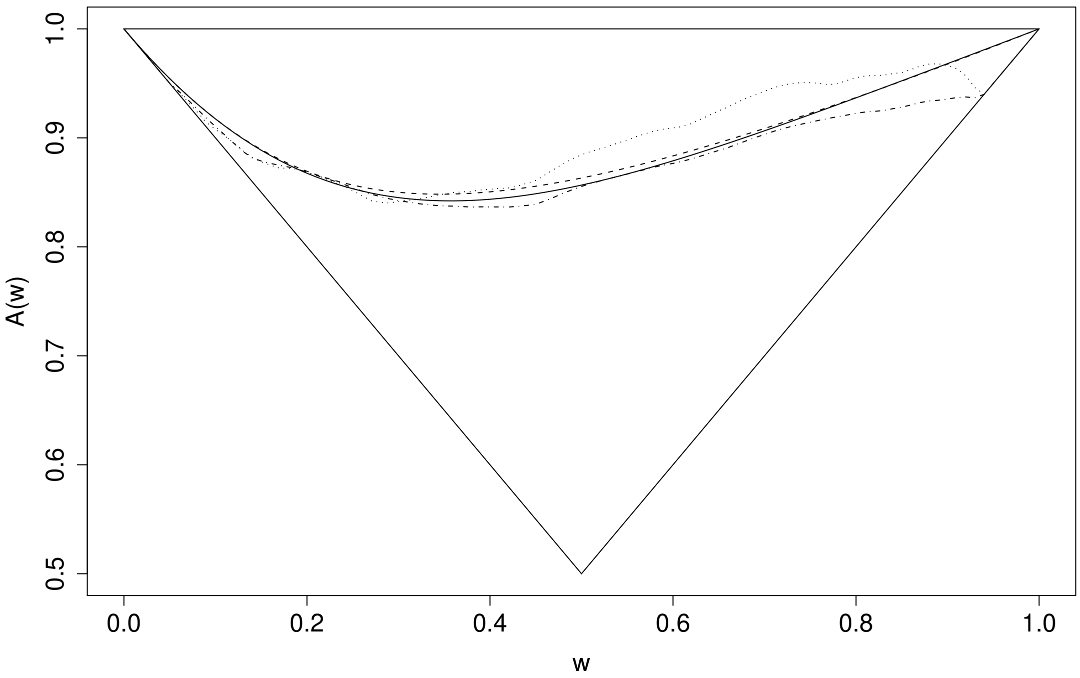


Figure 5.4: Four estimates of A for the transformed Lowestoft-Sheerness data: kernel (——), asymmetric logistic (- - -), Pickands (· · ·) and Capéraà *et al.* (-·-·-).

$m = 3$, $\phi_0 = 0.00$, $\phi_1 = 0.68$ and $\lambda = 0.22$; the log-likelihood is -310.11. The stability of parameter estimates from choosing a wider range of values for s was also investigated. Figure 5.5 indicates that the estimates for m, ϕ_0 and ϕ_1 are indeed stable when s is about 0.31. The kernel estimator places mass $\phi_1 = 0.68$ at $w = 1$, almost the same as with the asymmetric logistic model. The parametric and kernel estimates of the dependence function are also similar: see Figure 5.4.

Now the components and dependence structure are fitted simultaneously using the kernel estimator. With starting values dictated by the previous analysis, this yields the component parameter estimates in the top half of Table 5.5. Comparing these with the estimates from the separate component fits confirms the common finding

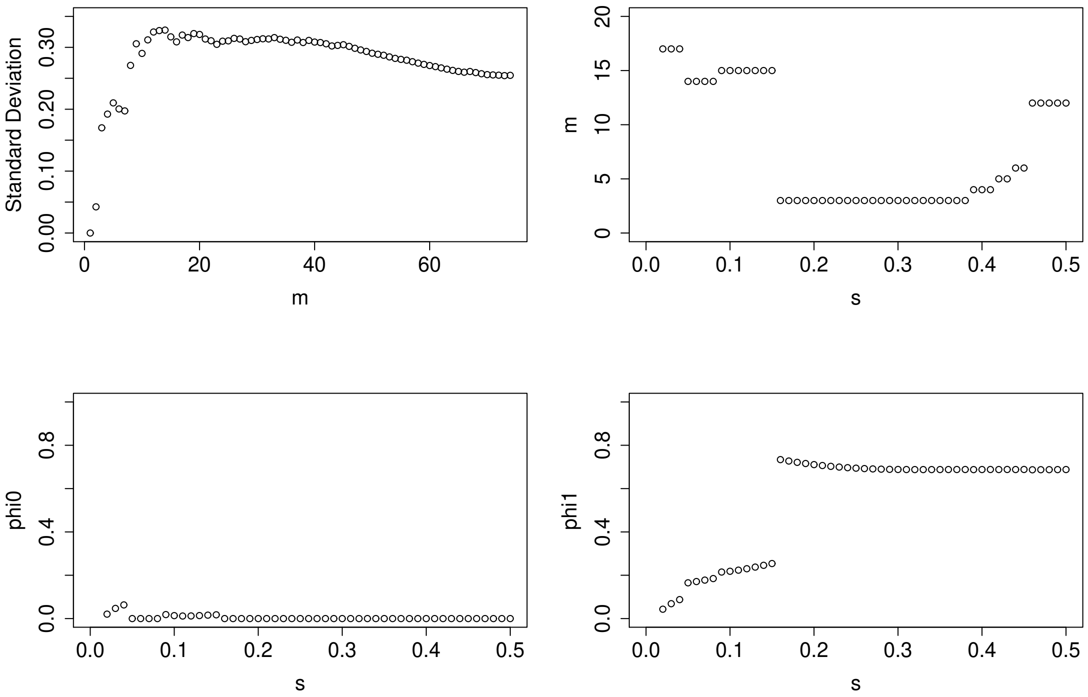


Figure 5.5: Top-left: standard deviation (s_m) of the angular component computed from the largest m observations. The remaining plots show estimates of the kernel dependence parameters with different values for s.

that joint estimation has most effect on the shape parameters. The parameter estimates of the dependence structure are $m = 14$, $\phi_0 = 0.00$, $\phi_1 = 0.72$ and $\lambda = 0.16$. Estimating the asymmetric logistic model simultaneously with the component parameters yields dependence parameters $\theta_1 = 0.28$, $\theta_2 = 1.00$ and $\alpha = 0.49$, and the two pairs of component parameter estimates are (1.96, 0.030, 0.25, 0.050) and (3.39, 0.384, 0.19, 0.023). Again, the mass placed on $w = 1$ by the two estimators is almost equal.

To conclude, the fitted models are used to estimate two failure probabilities:

$$
\begin{aligned}
p_{X_1 \cup X_2} &= P(X_1 > x_1 \text{ or } X_2 > x_2), \\
p_{X_2 | X_1} &= P(X_2 > x_2 \mid X_1 > x_1)
\end{aligned}
$$

for some threshold values x_1 and x_2. As an illustration, these are set equal to the upper p-quantiles (estimated under independence) of the two component distributions at time $t = 0.46$, corresponding to the year 1987, that is

$$
\begin{aligned}
x_1 &= \alpha_1 + \beta_1 t - \frac{\sigma_1}{\xi_1}\left[1 - \{-\log(1-p)\}^{-\xi_1}\right], \\
x_2 &= \alpha_2 + \beta_2 t - \frac{\sigma_2}{\xi_2}\left[1 - \{-\log(1-p)\}^{-\xi_2}\right].
\end{aligned}
$$

These are estimated to be $x_1 = 3.28$ and $x_2 = 4.58$ for $p = 0.01$. Of the observed data, only the outlier corresponding to the 1953 flood exceeds these thresholds: the empirical estimates of the probabilities are $\hat{p}_{X_1 \cup X_2} = 1/74 \approx 0.014$ and $\hat{p}_{X_2|X_1} = 1$. The estimates from the joint kernel fit are, with bootstrapped 90% confidence intervals, 0.012 (0.007, 0.029) and 0.20 (0.14, 0.26), while the joint asymmetric logistic fit yields 0.014 (0.004, 0.032) and 0.18 (0.04, 0.63).

5.6 Discussion

There are potentially many ways to improve the semi-parametric kernel estimator. One is the choice of smoothing parameter λ, which could be selected with a cross-validation procedure (cf. Hall and Tajvidi, 2000), or the kernel density estimator

(5.18) could be replaced with a simple histogram. Another issue is the scaling of q_n^* by ϕ to obtain the estimate (5.20) of $h = A''$. This was suggested by the limit (5.16), but when there is mass at the end points the estimate of q is likely to be contaminated by surplus observations near to zero and one: at finite levels, no angles will be observed that are precisely equal to zero or one. It may be profitable, therefore, to downweight q_n^* near the end points rather than scale uniformly over the entire interval.

Perhaps the most pressing issue concerns the likelihood. The parameter space of (ϕ_0, ϕ_1, m) is discrete, which makes optimisation expensive and invalidates standard likelihood procedures. A possible solution is to define a kernel density estimator with mass p_i on w_i, $1 \leq i \leq n$, and optimise the likelihood with respect to $(\phi_0, \phi_1, p_1, \ldots, p_n)$. The n masses could be chosen to belong to some parametric family in order to reduce the number of parameters. Another option is to formulate the estimator in a Bayesian framework and make inferences with a suitable Markov chain Monte Carlo scheme.

Chapter 6

Estimating Cluster Functionals

6.0 Introduction

In Chapter 4 techniques for identifying clusters were discussed. Chapter 5 addressed estimation of multivariate extreme-value distributions, which can be used to model cluster componentwise maxima from multivariate sequences. Simple empirical estimators were used in Chapter 4 to estimate more general cluster features (3.8). New estimators are introduced in this chapter that attempt to exploit the structure of clusters in univariate sequences identified by Theorem 3.3. In particular, it will be seen that clusters can be decomposed into a part for which there exists an asymptotically justified model, and a part for which no such model exists. Fitting the limiting model to the appropriate component of the observed clusters and combining this with an empirical description of the second component produces an estimation scheme that exploits all known cluster structure.

6.1 Clusters & Strings

Let $\{X_i\}_{i=1}^n$ be a sample from a stationary process of univariate random variables and suppose that a declustering scheme has identified C clusters, $\mathcal{C}_i = \{X_j : j \in \mathcal{S}_i\}$ for $1 \leq i \leq C$, of exceedances of a threshold u, where $\mathcal{S}_i \subseteq \{1, \ldots, n\}$. Let K_i be the number of points in $\mathcal{C}_i$ and re-label the points in the i-th cluster as $\{X_{i,j}\}_{j=1}^{K_i}$, where $X_{i,1} \geq \ldots \geq X_{i,K_i}$ for each i, and order the clusters so that $X_{1,1} \geq \ldots \geq X_{C,1}$. Define also $X_{C+1,1} = u$.

Denote the value of a cluster functional $c(\cdot)$ for the i-th cluster by

$$CF_i = c\{(X_{i,j} - u) : 1 \leq j \leq K_i\}.$$

The standard estimator for the expectation of the cluster functional is

$$\overline{CF} = C^{-1} \sum_{i=1}^{C} CF_i. \tag{6.1}$$

The distribution of the cluster functional can be estimated by choosing $CF_i = I[c\{(X_{i,j} - u) : 1 \leq j \leq K_i\} \leq x]$ for fixed x.

The estimators introduced in this chapter will require the computation of the string (3.3) associated with each cluster. Recall that a string characterises the position of cluster points relative to the cluster maximum. In order to compute strings, the marginal distribution of the exceedances must be specified. If this distribution is unknown then it can be estimated by fitting to threshold excesses the generalised Pareto distribution (2.3) with scale and shape parameters σ and ξ.

It is convenient to work with excesses transformed to have standard exponential marginal distribution. Let Ψ be this transformation and denote the transformed excesses by $\{Z_{i,j}\}_{j=1}^{K_i}$ for $1 \le i \le C$, where $Z_{i,j} = \Psi(X_{i,j} - u)$. If the marginal distribution has been estimated with the generalised Pareto model then

$$Z_{i,j} = \Psi(X_{i,j} - u) = \xi^{-1} \log \left\{ 1 + \xi \left(\frac{X_{i,j} - u}{\sigma} \right) \right\}, \tag{6.2}$$

taking the limit $Z_{i,j} = (X_{i,j} - u)/\sigma$ if $\xi = 0$.

For the standard exponential excesses, the string points for the i-th cluster are

$$Y_{i,j} = \exp(Z_{i,1} - Z_{i,j}) \quad \text{for } 1 \le j \le K_i. \tag{6.3}$$

Conversely, strings can be re-attached to cluster maxima to construct clusters:

$$Z_{i,j} = Z_{i,1} - \log Y_{i,j} \quad \text{for } 1 \le j \le K_i.$$

A consequence of the string definition (6.3) is that strings for different clusters are observed on different intervals, depending on the values of the cluster maxima, $\{Z_{i,1}\}_{i=1}^{C}$; the i-th string is observed on the interval $[1, \exp(Z_{i,1}))$.

A fundamental component of the estimators will be the value of the functional for a cluster constructed by attaching a string, $\{Y_{i,j}\}_{j=1}^{K_i}$, to a cluster maximum, Z. Denote these values by

$$CF_i(Z) = c\{\Psi^{-1}(Z - \log Y_{i,j})_+ : 1 \le j \le K_i\},$$

where the subscript i identifies the attached string. Reference will also be made to the random variable

$$CF(Z) = c\{\Psi^{-1}(Z - \log Y_j)_+ : 1 \leq j \leq K\}$$

where Z is standard exponential and $\{Y_j\}_{j=1}^{K}$ are points of the limiting string process in Theorem 3.3.

The standard estimator (6.1) is constituted from the C original clusters, which are equivalent to the clusters formed by attaching the i-th observed string to the i-th observed maximum for each i. Therefore, the standard estimator can be re-written as

$$\overline{CF} = C^{-1} \sum_{i=1}^{C} CF_i(Z_{i,1}).$$

The estimators introduced in the following sections assume that the asymptotic structure of clusters described by Theorem 3.3 holds over the threshold u, in particular, that both the strings and the maxima are independent and identically distributed and that the strings and maxima are independent of one another. Given this structure, it is reasonable to assume that each observed string could have occurred with any of the observed cluster maxima. This idea motivates the SWOP (Strings With Observed Peaks) estimator described in Section 6.2. The SWAP (Strings With Asymptotic Peaks) estimator of Section 6.3 is an extension that exploits knowledge of the distribution of the cluster maxima. The SWOP and SWAP estimators are compared with the standard estimator in a simulation study in Section 6.6 and their use is demonstrated for the Wooster temperature series in Section 6.7.

6.2 SWOP Estimators

The limiting independence of strings and maxima motivates the construction of C^2 clusters by attaching each observed string to each observed cluster maximum; C of these clusters will be the original, observed clusters. The cluster functional can be evaluated for each of the new clusters and combined to produce estimators such as

$$C^{-2}\sum_{i=1}^{C}\sum_{j=1}^{C} CF_j(Z_{i,1}) \tag{6.4}$$

for the expectation of the functional. It is hoped that constructing these additional clusters will result in better estimates. In effect, for each cluster maximum, not only the cluster shape that happened to coincide with it in the sample is observed, but also other cluster shapes that were not realised.

It was mentioned previously that the i-th string is observed only on the interval $[1, \exp(Z_{i,1}))$. Therefore, the cluster obtained from attaching the i-th string to a cluster maximum, Z, is defined on the interval $(Z - Z_{i,1}, Z]$. If the left-hand end of this interval is above zero, that is if $Z_{i,1}$ is less than Z, then the cluster will be truncated above the threshold. In effect, there will be an interval above zero in which the cluster is unobserved and the value of any cluster functional will be biased. For this reason it is inappropriate to attach the i-th string to cluster maxima greater than $Z_{i,1}$, and so estimator (6.4) is not valid. Discarding the truncated clusters leaves $C(C+1)/2$ clusters, yielding functional values $CF_j(Z_{i,1})$ for $1 \leq j \leq i \leq C$.

An estimator for the expectation of $CF(Z)$ conditional on $Z = Z_{i,1}$ is

$$i^{-1} \sum_{j=1}^{i} CF_j(Z_{i,1})$$

and so an estimator for the unconditional expectation of $CF(Z)$ is

$$\widehat{CF} = C^{-1} \sum_{i=1}^{C} i^{-1} \sum_{j=1}^{i} CF_j(Z_{i,1}),$$

which will be called the SWOP estimator. It is difficult to obtain analytically properties of the SWOP estimator because the string point process would need to be handled explicitly. Instead, its performance is evaluated in a simulation study in Section 6.6.

6.3 SWAP Estimators

The SWOP estimator of the previous section averages over only the observed (transformed) cluster maxima. In order to define strings, however, the marginal distribution must be known, or estimated with a generalised Pareto model. It seems wasteful to revert to the empirical marginal distribution once the strings have been computed. In this section, the assumption that the threshold excesses of the transformed cluster maxima are exponentially distributed is exploited. Instead of averaging over just the observed maxima, it is then possible to integrate over all possible cluster maxima above the threshold. This suggests the estimator

$$\int_0^\infty \left\{ C^{-1} \sum_{j=1}^{C} CF_j(z) \right\} e^{-z}\,\mathrm{d}z \tag{6.5}$$

for the expectation of $CF(Z)$.

Recall that attaching the i-th string to a maximum greater than $Z_{i,1}$ results in a truncated cluster and a biased value of the cluster functional. For the SWOP estimators, this was overcome by attaching to maxima only those strings that did not lead to truncated clusters; for each observed maximum there was at least one such string. The issue is more pressing here because of the integration over all maxima on $(0, \infty)$: on $(Z_{1,1}, \infty)$, all strings yield truncated clusters. A simple modification of the naive estimator (6.5) is not available.

The expectation of $CF(Z)$ can be decomposed as

$$E\{CF(Z)\} = E\{CF(Z) \mid Z \le Z_{1,1}\}(1 - e^{-Z_{1,1}}) + E\{CF(Z) \mid Z > Z_{1,1}\}e^{-Z_{1,1}}.$$

The data provide an estimator (akin to the SWOP estimator of the previous section) for the first term:

$$\tilde{E}\{CF(Z) \mid Z \le Z_{1,1}\}(1 - e^{-Z_{1,1}}) = \sum_{i=1}^{C} \int_{Z_{i+1,1}}^{Z_{i,1}} \left\{ i^{-1} \sum_{j=1}^{i} CF_j(z) \right\} e^{-z}\,\mathrm{d}z. \tag{6.6}$$

The difficulty lies in estimating the conditional expectation

$$E\{CF(Z) \mid Z > z\} \tag{6.7}$$

when $z = Z_{1,1}$. Any information on this quantity is incomplete in the sense that attaching any string to a cluster maximum $Z > Z_{1,1}$ will produce a cluster truncated above the threshold. Parametric modelling of the string point process is used to solve this problem. The conditional expectation (6.7) can then be approximated using Monte Carlo simulation or, in some cases, evaluated analytically. Another strategy that models the conditional expectation itself, as a function of z, will also be mentioned. This latter approach is less general because the model must be tailored to the functional of interest. Simulation studies have shown that the two strategies yield very similar results, so attention will be focussed on modelling the string process.

6.3.1 Modelling the string process

In applications, knowledge of how the string process behaves at unobserved levels could be available from an understanding of the mechanism generating the sequence. In this section, it is shown how such a model can be used to define a cluster-functional estimator. The idea is to choose a parametric model for the string process, fit it to the observed strings and assume that the tail of the model is a reasonable approximation to the unobserved string behaviour. Such a model enables simulation of clusters with maximum greater than $Z_{1,1}$ and so estimation of the expectation (6.7).

There is no theory to guide model choice in general, so the selection of the process with atom at $\{1\}$ and which is Poisson on $(1, \infty)$ with intensity $\lambda(\cdot)$ is arbitrary. The intensity function is chosen to decay according to a power law,

$$\lambda(y) = \beta y^{-\lambda} \quad \text{for } \lambda > 0 \text{ and } \beta \geq 0, \tag{6.8}$$

since this guards against the expectation (6.7) being infinite and yields tractable expressions for the expectation for some functionals. Denote the intensity measure on $[a, b) \subset (1, \infty)$ by

$$\Lambda[a, b) = \int_a^b \lambda(y)\,\mathrm{d}y = \begin{cases} \beta(1-\lambda)^{-1}\left(b^{1-\lambda} - a^{1-\lambda}\right) & \text{if } \lambda \neq 1, \\ \beta(\log b - \log a) & \text{if } \lambda = 1. \end{cases}$$

From a fit of the string model, an estimate of the expectation (6.7) can be obtained in the following way.

1. Generate a cluster maximum, $Z > Z_{1,1}$, by setting $Z = Z_{1,1} - \log U$, where U is a uniform random variate on $(0, 1)$.

2. Generate the number, K_0, of model string points on $[\exp(Z_{1,1}), \exp(Z))$ from a Poisson distribution with mean $\Lambda[\exp(Z_{1,1}), \exp(Z))$.

3. Generate the K_0 string points, $\{Y_{0,k}\}_{k=1}^{K_0}$, from the model by setting

$$Y_{0,k} = \begin{cases} \left\{U_k e^{(1-\lambda)Z} + (1-U_k)e^{(1-\lambda)Z_{1,1}}\right\}^{1/(1-\lambda)} & \text{if } \lambda \neq 1, \\ \exp\{U_k Z + (1-U_k)Z_{1,1}\} & \text{if } \lambda = 1, \end{cases}$$

   where $U_1, \dots, U_{K_0}$ are independent, uniform random variates on $(0,1)$.

4. Evaluate $CF(Z)$ for the string comprising the points $\{Y_{1,k}\}_{k=1}^{K_1}$ and $\{Y_{0,k}\}_{k=1}^{K_0}$.

5. Repeat these steps N times to obtain N values, $\{CF(Z_i)\}_{i=1}^{N}$, and the Monte Carlo estimator

$$\tilde{E}\{CF(Z) \mid Z > Z_{1,1}\}e^{-Z_{1,1}} = N^{-1}\sum_{i=1}^{N} CF(Z_i).$$

This computationally intensive procedure can sometimes be avoided if the cluster functional is additive. In that case the expectation of $CF(z)$ can be estimated by

$$\tilde{E}\{CF(z)\} = CF_1(z) + E\{CF_0(z)\},$$

where $CF_0(z)$ represents the value of the cluster functional obtained from attaching string points simulated from the fitted model on $[\exp(Z_{1,1}), \exp(z))$. For some functionals, the distribution of $CF_0(z)$ is tractable and it is possible to compute

$$\tilde{E}\{CF(Z) \mid Z > Z_{1,1}\}e^{-Z_{1,1}} = \int_{Z_{1,1}}^{\infty} \tilde{E}\{CF(z)\}e^{-z}\,\mathrm{d}z$$

analytically. Some examples for which this is the case are given later. Whichever method is used to estimate the conditional expectation (6.7), combining it with the estimator (6.6) yields the SWAP estimator,

$$\widetilde{CF} = \sum_{i=1}^{C} \int_{Z_{i+1,1}}^{Z_{i,1}} \left\{ i^{-1} \sum_{j=1}^{i} CF_j(z) \right\} e^{-z} \, \mathrm{d}z + \tilde{E}\{CF(Z) \mid Z > Z_{1,1}\} e^{-Z_{1,1}}, \quad (6.9)$$

for the expectation of $CF(Z)$.

It remains to fit the string model with intensity (6.8) to the C observed strings, which will be done by maximum likelihood. Discounting the atoms at $\{1\}$, the i-th string is observed on the interval $(1, \exp(Z_{i,1}))$ and the Poisson process likelihood is

$$\prod_{i=1}^{C} \exp\left\{-\Lambda\left(1, e^{Z_{i,1}}\right)\right\} \prod_{j=2}^{K_i} \lambda(Y_{i,j}).$$

Maximising the likelihood is a one-dimensional problem since the estimators satisfy the relationship

$$\beta = \begin{cases} (1-\lambda)\sum_{i=1}^{C}(K_i - 1) \Big/ \sum_{i=1}^{C}\{e^{(1-\lambda)Z_{i,1}} - 1\} & \text{if } \lambda \neq 1, \\ \sum_{i=1}^{C}(K_i - 1) \Big/ \sum_{i=1}^{C} Z_{i,1} & \text{if } \lambda = 1. \end{cases}$$

6.3.2 Modelling the expectation

In the previous section an estimator for the expectation (6.7) when $z = Z_{1,1}$ was obtained by modelling the string process. Here, an alternative approach is taken that models the expectation itself as a function of z. Note first that the expectation can be estimated when $z = Z_{k,1}$, $2 \leq k \leq C+1$, using the observed clusters:

$$\tilde{E}\{CF(Z) \mid Z > Z_{k,1}\} = (k-1)^{-1} \sum_{i=1}^{k-1} CF_i. \tag{6.10}$$

To estimate the expectation at $z = Z_{1,1}$ for a particular functional, a parametric model for $E\{CF(Z) \mid Z > z\}$ will be chosen, fitted to the C estimates (6.10) and then evaluated at $z = Z_{1,1}$. This approach will be illustrated for when the functional is cluster size, that is

$$CF_i(Z) = \sum_{j=1}^{K_i} I(Z - \log Y_{i,j} > 0),$$

and a suitable parametric model will be identified by examining the behaviour of the expectation for some simple processes.

Let $\{Z_i\}_{i\geq 1}$ be a process with standard exponential marginal distribution and define the conditional cluster-size distribution to be

$$\pi_n(j\,;\,z) = P\left\{ \sum_{i=1}^{r_n} I(Z_i > u_n) = j \;\middle|\; \max_{1\leq i\leq r_n} (Z_i) > u_n + z \right\}$$

for a threshold u_n, block size $r_n = o(n)$ and a constant $z \geq 0$. This is the cluster-size distribution conditional on the cluster maximum being bounded below by $u_n + z$. If

$$\pi(j\,;\,z) = \lim_{n\to\infty} \pi_n(j\,;\,z)$$

then $E\{CF(Z) \mid Z > z\} = \sum_{j\geq 1} j\pi(j\,;\,z)$, which will now be computed for a moving-maximum process (3.9).

Let $X_i = \max_{j\geq 1}(\alpha_j W_{i-j})$, where each $\alpha_j \geq 0$, $\sum_{i\geq 1}\alpha_i = 1$ and the W_i are independent, standard Fréchet innovations. Denote by $\alpha_{(j)}$ the j-th largest of the α_j, let $Z_i = -\log\{1-\exp(-1/X_i)\}$ be the transformed process with standard exponential margin and set $u_n = -\log\{1-\exp(-1/n)\}$. For large n, clusters tend to be formed from a single innovation, say W_0, giving

$$\begin{aligned}
\pi_n(j\,;\,z) &\sim P\left\{\sum_{i=1}^{r_n} I(\alpha_i W_0 > n) = j \,\middle|\, \max_{1\leq i\leq r_n}(\alpha_i W_0) > ne^z\right\} \\
&\sim P\left\{\frac{n}{\alpha_{(j)}} < W_0 \leq \frac{n}{\alpha_{(j+1)}} \,\middle|\, W_0 > \frac{ne^z}{\alpha_{(1)}}\right\} \\
&= \frac{\exp\{-\alpha_{(j+1)}/n\} - \exp[-1/\max\{n/\alpha_{(j)}, ne^z/\alpha_{(1)}\}]}{1-\exp\{-\alpha_{(1)}/(ne^z)\}}\, I\left\{e^z < \frac{\alpha_{(1)}}{\alpha_{(j+1)}}\right\}.
\end{aligned}$$

The limit is

$$\pi(j\,;\,z) = \frac{\min\{\alpha_{(j)}e^z, \alpha_{(1)}\} - \alpha_{(j+1)}e^z}{\alpha_{(1)}}\, I\left\{e^z < \frac{\alpha_{(1)}}{\alpha_{(j+1)}}\right\}.$$

Note that $\pi(\cdot\,;\,z) \to \pi(\cdot)$, the unconditional cluster-size distribution, as z tends to zero.

Now let $k = k(z)$ satisfy

$$\alpha_{(1)}/\alpha_{(k)} < e^z \leq \alpha_{(1)}/\alpha_{(k+1)}. \tag{6.11}$$

Then, the conditional expected cluster size is

$$E\{CF(Z) \mid Z > z\} = \sum_{j \geq 1} j\pi(j\,;\,z) = k + \frac{e^z}{\alpha_{(1)}}\left\{1 - \sum_{j=1}^{k} \alpha_{(j)}\right\}.$$

This shows how the conditional expectation depends on z. Unfortunately, the moving-maximum process is too general to suggest a useful parametric model. Consider instead the special case of the max-autoregressive process, for which $\alpha_{(j)} = (1-\alpha)\alpha^{j-1}$. In this case, condition (6.11) becomes

$$\alpha^{1-k} < e^z \leq \alpha^{-k} \iff -z/\log\alpha \leq k < 1 - z/\log\alpha,$$

and if $\exp(z) = \alpha^{-k}$ then

$$E\{CF(Z) \mid Z > z\} = k + (1-\alpha)^{-1}\alpha^k e^z = (1-\alpha)^{-1} - z(\log\alpha)^{-1}.$$

This suggests the linear model

$$E\{CF(Z) \mid Z > z\} = \gamma + \beta z, \qquad \beta \geq 0 \text{ and } \gamma \geq 1. \tag{6.12}$$

Given the sample values (6.10), least-squares estimates of β and γ can be obtained by minimising

$$\sum_{k=2}^{C+1}\left[\tilde{E}\{CF(Z) \mid Z > Z_{k,1}\} - \gamma - \beta Z_{k,1}\right]^2.$$

A model has been proposed for the conditional expectation in light of results for the max-autoregressive process. To assess its wider applicability, one-thousand clusters with maximum over z, for each $z = 0, 1, \ldots, 10$, were generated from a Markov chain with logistic dependence structure. Plotting the mean cluster size against z produced a near-linear curve (not shown): surprisingly, the model (6.12) is an excellent fit.

Let $\hat{\gamma} + \hat{\beta} Z_{1,1}$ be the estimator for the conditional expectation (6.7). Combining this with the estimator (6.6) produces an estimator for the expected cluster size:

$$\widetilde{CF} = \sum_{i=1}^{C} \int_{Z_{i+1,1}}^{Z_{i,1}} \left\{ i^{-1} \sum_{j=1}^{i} CF_j(z) \right\} e^{-z}\,\mathrm{d}z + (\hat{\gamma} + \hat{\beta} Z_{1,1}) e^{-Z_{1,1}}. \tag{6.13}$$

The linear model (6.12) seems to be a reasonable choice for expected cluster size. For other cluster functionals, different models may be required. The simulation study described later in this chapter was conducted for estimator (6.13) and its performance was found to be very similar to the SWAP estimator (6.9). Given the greater generality of the SWAP estimator, the method described in this section is not pursued any further.

6.4 Multiple Thresholds

Inferences about cluster functionals may be required at a range of thresholds u. For the SWOP and SWAP estimators, this means that estimates of the marginal distribution and, for SWAP, the string model must be obtained at each threshold. This could be achieved simply by re-estimating the generalised Pareto and string models at each threshold. More stable estimates are likely to be obtained, however, if the marginal transformation (6.2) is made at the lowest threshold over which the generalised Pareto model is believed to hold. If this threshold is u_0 then the parameters (σ_0, ξ_0) can be estimated from the excesses over u_0. For thresholds $u_1 > u_0$, the generalised Pareto model still holds but with parameters $\xi_1 = \xi_0$ and $\sigma_1 = \sigma_0 + \xi_0(u_1 - u_0)$. Therefore, the parameters need to be estimated at only u_0, after which the parameters at all higher thresholds are determined.

A similar property holds for the string model. For clusters identified over the threshold u_1, the j-th string point of the i-th cluster is

$$\begin{aligned} Y^1_{i,j} &= \left\{ \frac{1 + \xi_1(X_{i,1} - u_1)/\sigma_1}{1 + \xi_1(X_{i,j} - u_1)/\sigma_1} \right\}^{1/\xi_1} \\ &= \left\{ \frac{\sigma_1 + \xi_1(X_{i,1} - u_1)}{\sigma_1 + \xi_1(X_{i,j} - u_1)} \right\}^{1/\xi_1} \\ &= \left\{ \frac{\sigma_0 + \xi_0(X_{i,1} - u_0)}{\sigma_0 + \xi_0(X_{i,j} - u_0)} \right\}^{1/\xi_0} \\ &= Y^0_{i,j}. \end{aligned}$$

So, the string elements are unaffected by changing threshold; only the interval over which strings are observed is changed. This means that, at sufficiently high levels, the parameters (β, λ) of the string model are expected to be the same at all thresholds. As for the parameters of the generalised Pareto distribution, therefore, the parameters of the string model can be estimated at one threshold and fixed at these values for all higher thresholds.

In summary, the cluster-functional estimators are implemented in the following way. First, set u_0 to be the lowest threshold above which the extremal behaviour of the process $\{X_i\}_{i\geq 1}$ is believed to be stable. Apply a declustering scheme at this threshold and identify the clusters $\{X_{i,j}\}_{j=1}^{K_i}$ for $1 \leq i \leq C$. Standard estimators can now be evaluated. Next, fit the generalised Pareto distribution to the excesses of u_0, obtaining parameter estimates (σ_0, ξ_0). Transform the clusters to have standard exponential margins and determine the strings $\{Y_{i,j}\}_{j=1}^{K_i}$ for $1 \leq i \leq C$. SWOP estimators can now be evaluated. Fit the string model to these strings, obtaining parameter estimates (β, λ). SWAP estimators can now be evaluated. For estimates at thresholds u_1 greater than u_0, re-apply the declustering scheme, use the parameters of the generalised Pareto distribution implied by the estimates at u_0 and, for the string model, use the parameter estimates at u_0. In the simulation study of Section 6.6, estimates obtained by re-estimating the generalised Pareto distribution at each threshold will also be considered.

6.5 Examples

In this section the forms of the standard, SWOP and SWAP estimators are made explicit for two cluster functionals: the cluster size and cluster excess (3.8).

6.5.1 Cluster size

The functional for cluster size has value $CF_i = K_i$ for the i-th cluster. Estimators will be reported for the expected cluster size, $1/\theta$, and the mass function, π, of the cluster-size distribution.

The standard estimators are

$$\begin{aligned}\bar{\theta}^{-1} &= C^{-1}\sum_{i=1}^{C} K_i,\\ \bar{\pi}(k) &= C^{-1}\sum_{i=1}^{C} I(K_i = k).\end{aligned}$$

Since cluster size is invariant to the scale on which it is computed, the SWOP estimators are

$$\begin{aligned}\hat{\theta}^{-1} &= C^{-1}\sum_{i=1}^{C} i^{-1}\sum_{j=1}^{i} K_{i,j},\\ \hat{\pi}(k) &= C^{-1}\sum_{i=1}^{C} i^{-1}\sum_{j=1}^{i} I(K_{i,j} = k),\end{aligned}$$

where $K_{i,j} = \sum_{k=1}^{K_j} I(Z_{i,1} - \log Y_{j,k} > 0)$.

Cluster size is an additive functional and the SWAP estimators can be written in closed form. First note that, for $z > Z_{1,1}$, the value, $CF_0(z)$, of the cluster functional obtained from attaching string points simulated from the string model on $[\exp(Z_{1,1}), \exp(z))$ is a Poisson random variable with mean

$$E\{CF_0(z)\} = \Lambda[e^{Z_{1,1}}, e^z) = \beta(1-\lambda)^{-1}\left\{e^{(1-\lambda)z} - e^{(1-\lambda)Z_{1,1}}\right\},$$

taking the limit, $\beta(z - Z_{1,1})$, when $\lambda = 1$. Now, $CF_1(z) = K_1$ for $z \geq Z_{1,1}$ and so

$$\int_{Z_{1,1}}^{\infty} \tilde{E}\{CF(z)\}e^{-z}\,\mathrm{d}z = K_1 e^{-Z_{1,1}} + \beta\lambda^{-1}e^{-\lambda Z_{1,1}}.$$

Furthermore,

$$
\begin{aligned}
&\sum_{i=1}^{C}\int_{Z_{i+1,1}}^{Z_{i,1}} i^{-1}\sum_{j=1}^{i} CF_j(z)e^{-z}\,\mathrm{d}z \\
&\quad= \sum_{i=1}^{C} i^{-1}\sum_{j=1}^{i}\sum_{k=1}^{K_j}\int_{Z_{i+1,1}}^{Z_{i,1}} I(z-\log Y_{j,k}>0)e^{-z}\,\mathrm{d}z \\
&\quad= \sum_{i=1}^{C} i^{-1}\sum_{j=1}^{i}\sum_{k=1}^{K_j}\int_{m_{i+1,j,k}}^{m_{i,j,k}} e^{-z}\,\mathrm{d}z \\
&\quad= \sum_{i=1}^{C} i^{-1}\sum_{j=1}^{i}\sum_{k=1}^{K_j}(e^{-m_{i+1,j,k}}-e^{-m_{i,j,k}}),
\end{aligned}
$$

where $m_{i,j,k} = \max\{Z_{i,1}, \log Y_{j,k}\}$, and the SWAP estimator for the expected cluster size is

$$
\tilde{\theta}^{-1} = \sum_{i=1}^{C} i^{-1}\sum_{j=1}^{i}\sum_{k=1}^{K_j}(e^{-m_{i+1,j,k}}-e^{-m_{i,j,k}}) + K_1e^{-Z_{1,1}} + \beta\lambda^{-1}e^{-\lambda Z_{1,1}}.
$$

For the cluster-size distribution, define $Y_{j,k} = \infty$ for $k > K_j$ and $1 \le j \le C$. Then, for positive integer k,

$$
\begin{aligned}
&\sum_{i=1}^{C}\int_{Z_{i+1,1}}^{Z_{i,1}} i^{-1}\sum_{j=1}^{i} I\{CF_j(z)\ge k\}e^{-z}\,\mathrm{d}z \\
&\quad= \sum_{i=1}^{C} i^{-1}\sum_{j=1}^{i}\int_{Z_{i+1,1}}^{Z_{i,1}} I(z-\log Y_{j,k}>0)e^{-z}\,\mathrm{d}z \\
&\quad= \sum_{i=1}^{C} i^{-1}\sum_{j=1}^{i}\int_{m_{i+1,j,k}}^{m_{i,j,k}} e^{-z}\,\mathrm{d}z \\
&\quad= \sum_{i=1}^{C} i^{-1}\sum_{j=1}^{i}(e^{-m_{i+1,j,k}}-e^{-m_{i,j,k}}).
\end{aligned}
$$

For $z > Z_{1,1}$,

$$P\{CF_1(z) + CF_0(z) = k\} = \frac{I(k \geq K_1)}{(k - K_1)!} E\{CF_0(z)\}^{k-K_1} \exp[-E\{CF_0(z)\}]$$

and so the contribution to the SWAP estimator from the string model requires numerical approximation. If cluster sizes $k < K_1$ only are of interest, however, the contribution from the string model is zero and the SWAP estimator is

$$\tilde{\pi}(k) = \sum_{i=1}^{C} i^{-1} \sum_{j=1}^{i} \left\{ \left(e^{-m_{i+1,j,k}} - e^{-m_{i,j,k}} \right) - \left(e^{-m_{i+1,j,k+1}} - e^{-m_{i,j,k+1}} \right) \right\}.$$

6.5.2 Cluster excess

The functional for cluster excess has value $CF_i = \sum_{j=1}^{K_i} (X_{i,j} - u)$ for the i-th cluster. Estimators for the expected cluster excess, A, and the cluster-excess distribution, F_A, will be reported.

The standard estimators are

$$\begin{aligned} \bar{A} &= C^{-1} \sum_{i=1}^{C} \sum_{j=1}^{K_i} (X_{i,j} - u), \\ \bar{F}_A(a) &= C^{-1} \sum_{i=1}^{C} I \left\{ \sum_{j=1}^{K_i} (X_{i,j} - u) \leq a \right\}. \end{aligned}$$

The SWOP estimators are

$$\hat{A} = C^{-1}\sum_{i=1}^{C} i^{-1}\sum_{j=1}^{i}\sum_{k=1}^{K_j}\Psi^{-1}(Z_{i,1}-\log Y_{j,k})_+,$$
$$\hat{F}_A(a) = C^{-1}\sum_{i=1}^{C} i^{-1}\sum_{j=1}^{i} I\left\{\sum_{k=1}^{K_j}\Psi^{-1}(Z_{i,1}-\log Y_{j,k})_+ \leq a\right\}.$$

The SWAP estimators require some working. The expectation $E\{CF_0(z)\}$ is computed first for $z > Z_{1,1}$. Let $K_0(z)$ represent the number of points simulated from the string model in $[\exp(Z_{1,1}), \exp(z))$. Conditional on $K_0(z)$, the simulated points are independent with density $\beta y^{-\lambda}/\Lambda[\exp(Z_{1,1}), \exp(z))$ on $[\exp(Z_{1,1}), \exp(z))$. Let Y represent an arbitrary point simulated from this density and note that $z - \log Y > 0$ due to the support of Y. Now, $E\{K_0(z)\} = \Lambda[\exp(Z_{1,1}), \exp(z))$ and so

$$E\{CF_0(z)\} = \Lambda[e^{Z_{1,1}}, e^z)E\left\{\Psi^{-1}(z-\log Y) - u \mid K_0(z)\right\}.$$

Some algebra shows that this is equal to

$$\frac{\beta\sigma}{\xi(1-\lambda)(1-\xi-\lambda)}\left[\xi e^{(1-\lambda)z} + \left\{(1-\xi-\lambda)-(1-\lambda)e^{\xi(z-Z_{1,1})}\right\}e^{(1-\lambda)Z_{1,1}}\right],$$

taking the appropriate limit when $\xi = 0, \lambda = 1$ or $\xi + \lambda = 1$. Integrating yields

$$\int_{Z_{1,1}}^{\infty} E\{CF_0(z)\}e^{-z}\,\mathrm{d}z = \frac{\beta\sigma e^{-\lambda Z_{1,1}}}{\lambda(1-\xi)}$$

for $\lambda > 0$ and $\xi < 1$. The integral diverges when $\xi \geq 1$, which corresponds to the generalised Pareto distribution for the cluster maximum excess having infinite mean. Also,

$$\int_{Z_{1,1}}^{\infty} CF_1(z)e^{-z}\,\mathrm{d}z = \frac{\sigma}{\xi}\sum_{k=1}^{K_1}\left\{\frac{Y_{1,k}^{-\xi}}{1-\xi}e^{-(1-\xi)Z_{1,1}} - e^{-Z_{1,1}}\right\}$$

for $z > Z_{1,1}$ and $\xi < 1$, taking the limit when $\xi = 0$.

The contribution to the estimator for $z < Z_{1,1}$ is

$$\sum_{i=1}^{C} i^{-1}\sum_{j=1}^{i}\sum_{k=1}^{K_j} I_{i,j,k},$$

where

$$\begin{aligned} I_{i,j,k} &= \int_{Z_{i+1,1}}^{Z_{i,1}} \Psi^{-1}(z-\log Y_{j,k})_+ e^{-z}\,\mathrm{d}z \\ &= \frac{\sigma}{\xi}\left[\frac{Y_{j,k}^{-\xi}}{1-\xi}\left\{e^{-(1-\xi)m_{i+1,j,k}} - e^{-(1-\xi)m_{i,j,k}}\right\} - \left(e^{-m_{i+1,j,k}} - e^{-m_{i,j,k}}\right)\right], \end{aligned}$$

taking appropriate limits when $\xi = 0$ or $\xi = 1$. Therefore, the SWAP estimator for the expected cluster excess is infinity if $\xi \geq 1$ and

$$\tilde{A} = \sum_{i=1}^{C} i^{-1}\sum_{j=1}^{i}\sum_{k=1}^{K_j} I_{i,j,k} + \frac{\sigma}{\xi}\sum_{k=1}^{K_1}\left\{\frac{Y_{1,k}^{-\xi}}{1-\xi}e^{-(1-\xi)Z_{1,1}} - e^{-Z_{1,1}}\right\} + \frac{\beta\sigma e^{-\lambda Z_{1,1}}}{\lambda(1-\xi)}$$

if $\xi < 1$ with the appropriate limit at $\xi = 0$.

For the distribution function, the contribution from $z < Z_{1,1}$ is

$$\sum_{i=1}^{C} i^{-1}\sum_{j=1}^{i} I_{i,j},$$

where

$$I_{i,j} = \int_{Z_{i+1,1}}^{Z_{i,1}} I\left\{\sum_{k=1}^{K_j} \Psi^{-1}(z - \log Y_{j,k})_+ \leq a\right\} e^{-z}\,\mathrm{d}z = e^{-m^*_{i+1,j}} - e^{-m^*_{i,j}},$$

$m^*_{i,j} = \min\{Z_{i,1}, z^*_j\}$ and z^*_j satisfies

$$\sum_{k=1}^{K_j} \Psi^{-1}(z^*_j - \log Y_{j,k})_+ = a.$$

If only values $a < CF_1$ are of interest then the contribution from $z > Z_{1,1}$ is zero. Otherwise, $P\{CF_0(z) \leq a\}$ requires numerical approximation.

6.6 Simulation Study

The performances of the standard, SWOP and SWAP estimators are compared through a simulation study. Three quantities are estimated: the extremal index, as the reciprocal of mean cluster size, the probability $\pi(1)$ of a cluster having size 1, and the mean cluster excess. The estimators are applied to 100 realisations of length 5000 from three max-autoregressive processes, with extremal indices 0.25, 0.5 and 0.75. The processes are generated with standard exponential margin, in which case the mean cluster excess has a finite limit as the threshold increases. In fact, the mean cluster excess is equal to the mean cluster size since the expected excess for a single exceedance is 1. The three quantities are estimated at ten thresholds, corresponding to the empirical p-quantiles for $p = 0.90, 0.91, \ldots, 0.99$, and the estimators are applied in three situations: when the margin is assumed known, when the margin is estimated at the lowest threshold, and when the margin is re-estimated at each threshold. The status of knowledge about the margin does not affect the standard estimator (6.1). The parameters of the string process model in the SWAP estimator are always estimated at only the lowest threshold. Intervals declustering is implemented at each threshold to identify clusters. The performance of the estimators is summarised by the Monte Carlo estimates of root mean square error (RMSE) reproduced in Figures 6.1 – 6.3.

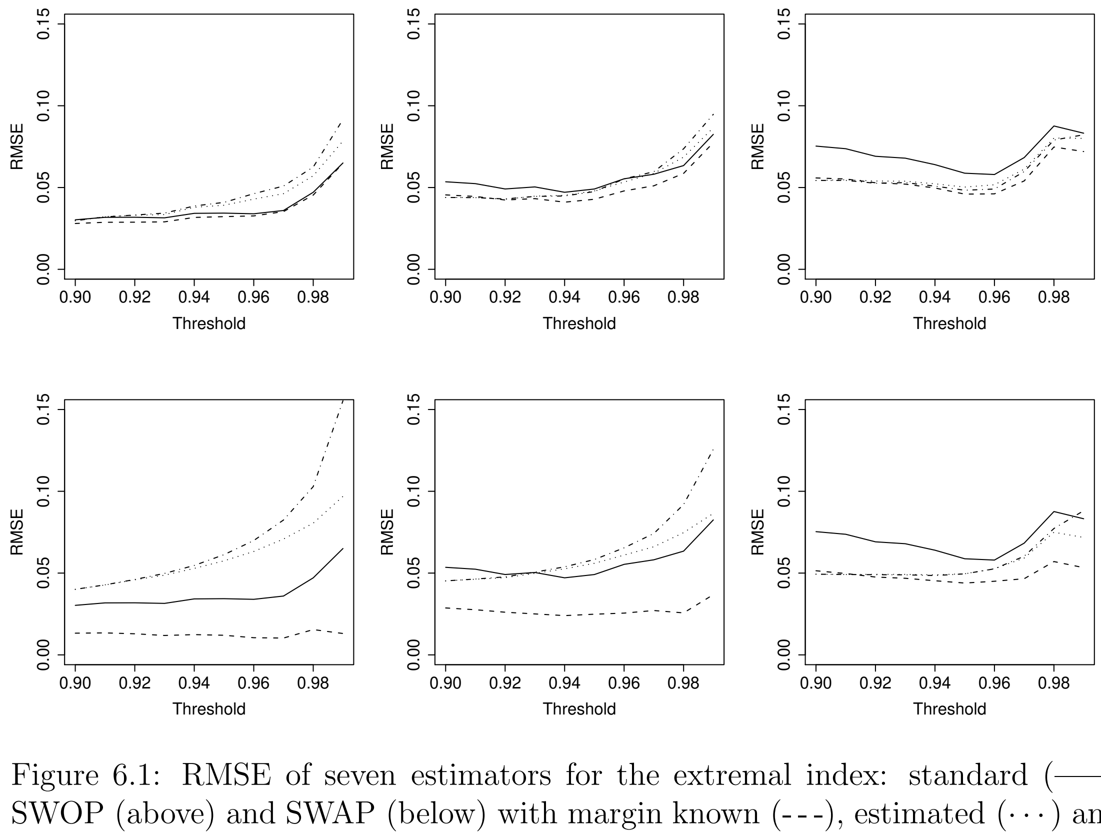


Figure 6.1: RMSE of seven estimators for the extremal index: standard (——), SWOP (above) and SWAP (below) with margin known (- - -), estimated (· · ·) and re-estimated (-·-·-). The true extremal index is 0.25 (left), 0.5 (middle) and 0.75 (right).

Consider first the relative performance of the standard and SWOP estimators, seen in the first rows of Figures 6.1 – 6.3. When the margin is known, the SWOP

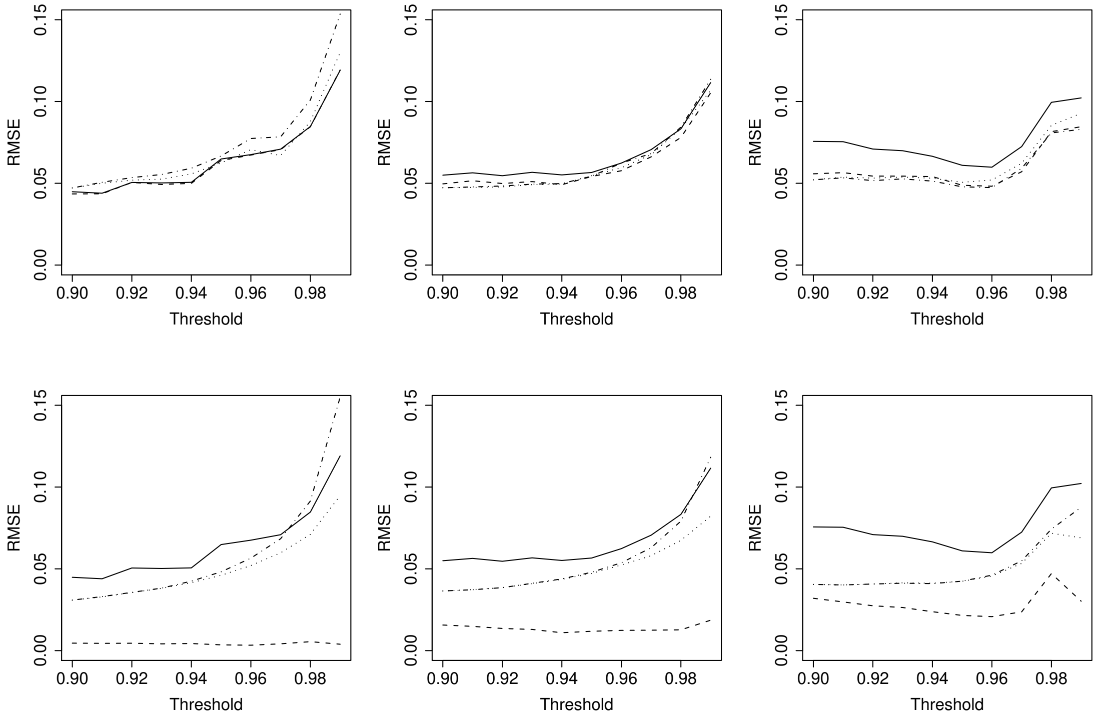


Figure 6.2: RMSE of seven estimators for $\pi(1)$: standard (——), SWOP (above) and SWAP (below) with margin known (- - -), estimated ($\cdots$) and re-estimated (-·-·-). The true extremal index is 0.25 (left), 0.5 (middle) and 0.75 (right).

estimator outperforms the standard estimator to a small degree in every case, and this superiority is greater when the extremal index is greater. The performance of the SWOP estimator deteriorates when the marginal distribution is estimated, but remains slightly superior to the standard estimator, at least when the extremal index is large. Bias and variance estimates (not shown) indicate that the SWOP estimators produce smaller clusters (larger extremal index and $\pi(1)$, smaller mean cluster excess) and, except when the margin is known, have greater variability than the standard estimator. Standard estimators tend to over-estimate clusters. When the margin is known, therefore, the superiority of the SWOP estimator is due to a reduction in variance and bias; when the margin is unknown, there is a

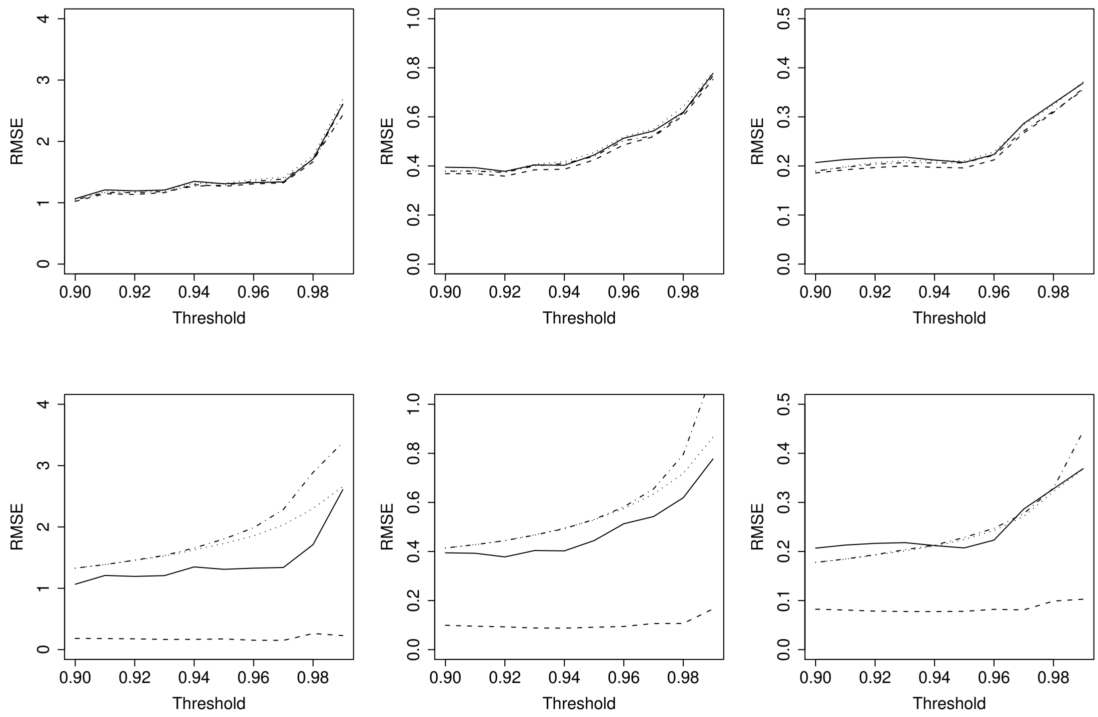


Figure 6.3: RMSE of seven estimators for the mean cluster excess: standard (——), SWOP (above) and SWAP (below) with margin known (- - -), estimated (· · ·) and re-estimated (-·-·-). The true extremal index is 0.25 (left), 0.5 (middle) and 0.75 (right).

trade-off between a reduction in bias and an increase in variance.

Consider now the performance of the SWAP estimators, seen in the second rows of Figures 6.1 – 6.3. When the margin is known, the SWAP estimator significantly outperforms the standard and SWOP estimators in every case. The performance is severely diminished when the margin is estimated, however, perhaps remaining superior only when the extremal index is large. Again, SWAP estimators produce smaller clusters than the standard estimator and, except when the margin is known, have greater variability.

The simulation study was repeated, replacing the max-autoregressive processes by

Markov chains with bivariate extreme-value logistic dependence structure (5.6). Although the extremal behaviours of the two processes are different, with the max-autoregressive clusters following a deterministic pattern and those of the Markov chains being stochastic, the results were qualitatively similar. The benefit of knowing the marginal distribution on the performance of the SWAP estimator was markedly reduced however.

When the marginal distribution is known, creating extra clusters by interchanging strings between cluster maxima has been found to reduce the bias and variance of estimators for some cluster functionals. When the margin is estimated, however, this procedure inflates the variance of the estimators, particularly when the clustering in the data is strong. Any error in the marginal estimation causes a mis-specification of the strings and, consequently, of the extra clusters. The effect is magnified for the SWAP estimator because strings are attached to all possible maxima over the threshold, not just the observed maxima, and because the strings are used to estimate the string model. The SWOP and SWAP estimators perform better when the clustering is weak because declustering becomes easier: with strong dependence, the scope for nominally independent clusters to be identified as a single cluster is greater, and all of the clusters formed from the corresponding string will be erroneous. The need for precise declustering is emphasised by the results for the Markov chains, for which declustering is more difficult compared to max-autoregressive processes.

The issues of marginal estimation and declustering appear to be of primary importance for improving the SWOP and SWAP estimators. No attempt was made to optimise these aspects in the simulation study, by careful selection of threshold for example, so this might be considered in more detail. Several methods were tried of reducing the influence of strings from clusters with large maxima, since these clusters tend to be larger and so are less likely to be declustered successfully. For example, the string from only the i-th cluster was used to define string points on $[\exp(Z_{i+1,1}), \exp(Z_{i,1}))$, and strings were created by drawing at random from the empirical distribution of all observed string points. Recall also that the string model for the SWAP estimator was fitted to all observed string points, but that the model is used to generate string points only above $\exp(Z_{1,1})$. The aim is to estimate just the tail of the string process, therefore, and so attempts were made to fit the model to only those string points over some threshold. None of these modifications to the creation of strings had a significant impact on the estimators.

6.7 Data Example

To conclude, the estimators are applied to the Wooster temperature series of Section 4.1.7. There, standard estimates of the extremal index, cluster-size distribution and the mean cluster excess were computed at several different thresholds. Here, the standard, SWOP and SWAP estimators are applied, re-estimating the parameters of the generalised Pareto and string models at each threshold. The estimates, with bootstrapped confidence intervals are displayed in Figure 6.4. The relative bias noted in the simulation is evident, with the standard estimator producing larger clusters. The variances of the standard and SWOP estimators are similar; the variance of the SWAP estimator is typically lowest. For thresholds above six degrees, the estimators are in broad agreement, with both the extremal index and $\pi(1)$ between about 0.5 and 0.7. These are similar to the values found with the Markov chain model of Smith *et al.* (1997).

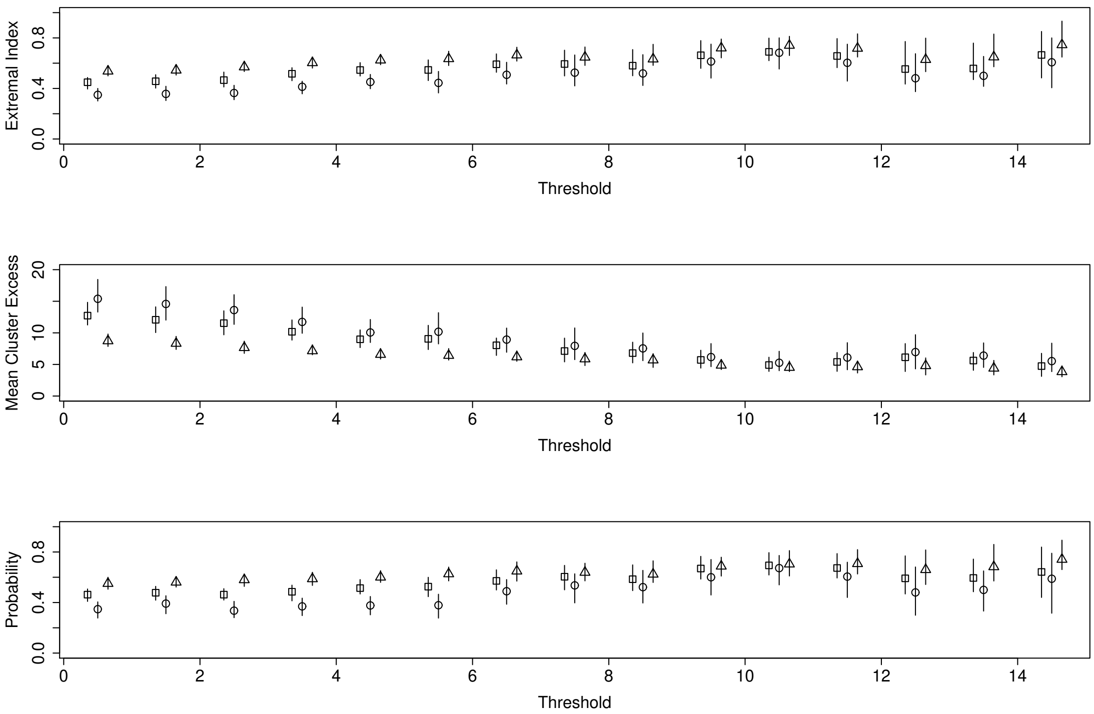


Figure 6.4: Standard (◯), SWOP (□) and SWAP (△) estimates of the extremal index (top), mean cluster excess (middle) and $\pi(1)$ (bottom) against threshold, with bootstrapped 90% confidence intervals (|).

Chapter 7

Inference for Multivariate Clusters

7.0 Introduction

In the previous chapter, new estimators for features of clusters in univariate processes were defined. These estimators decomposed clusters into maxima and strings, where a string defines the points in a cluster relative to the cluster maximum. In this chapter, the decomposition is extended to the multivariate setting. Combining the empirical description of strings with an extreme-value distribution for the maxima provides a model for the multivariate process, including its clustering behaviour, at levels outwith the data. Previous attempts at inference for multivariate processes have modelled only summaries of clusters, with estimates depending on the particular summary chosen.

One important problem in multivariate extremes is the estimation of failure probabilities, that is the probability that an observation falls in some rare region of the sample space. Such a failure region might, for example, characterise the conditions under which a sea-wall is breached, with respect to wave height, wave period, surge and tide. See Coles and Tawn (1994) for several other examples. Estimation of failure probabilities is reviewed in Section 7.1, where shortcomings of current methods are noted. The decomposition of multivariate clusters into maxima and strings is presented in Section 7.2 together with methods for fitting the model and estimating failure probabilities. A simulation study illustrating the benefits of the new approach is documented in Section 7.3 and a data example is provided in Section 7.4.

7.1 Estimating Failure Probabilities

Let $\{X_i\}_{i=1}^n$ be a stationary sequence of D-dimensional random variables $X_i = (X_{i1}, \ldots, X_{iD})$ with unknown marginal distribution function F. Suppose that it is of interest to estimate the probability $P(X \in B)$ that an arbitrary observation X falls in some failure region $B \subset \mathbb{R}^D$.

Theorem 2.4 motivates a Poisson process model for points X_i on regions bounded away from the origin when those points are from an independent sequence. For stationary sequences, such a model is not applicable because extremes will form clusters, as described by Theorem 3.6. This problem can be avoided by filtering the sequence to obtain approximately independent observations. For example, Coles and Tawn (1994) partition the original data into blocks and retain only the componentwise maximum from each block, using arguments concerning the physical process generating the data to justify independence. Similarly, de Haan and Sinha (1999) retain only the componentwise maxima from subjectively identified storms.

Parametric and semi-parametric methods exist for estimating failure probabilities. Parametric methods rely on fitting the Poisson process model to the filtered data on some region bounded away from the origin, as described in Section 2.2. A model for the dependence structure, such as the logistic model (5.6) or the kernel estimator of Chapter 5, must be selected. The failure probability can be estimated by computing the probability that a point of the fitted Poisson process falls in B. This approach is developed by Coles and Tawn (1991, 1994) and Joe *et al.* (1992). A semi-parametric approach has been developed by de Haan and de Ronde (1998) and de Haan and Sinha (1999). This relies on a homogeneity property of the Poisson process measure to relate the failure probability to the probability that a point falls in a set $B_c = \{x : cx \in B\}$, where c shrinks B until it contains some of the observations. The latter probability can then be estimated by the proportion of observations in B_c.

Whether the parametric or semi-parametric approach is followed, the resulting failure probability estimate depends crucially on the filtering used to obtain independent observations. For example, consider componentwise maxima. Suppose that the original data have been partitioned into $n_b = \lfloor n/b \rfloor$ blocks of length b:

$$\{X_j : (i-1)b + 1 \leq j \leq ib\} \quad \text{for } 1 \leq i \leq n_b. \tag{7.1}$$

The block componentwise maxima, hereafter called the block maxima, are

$$X_i^* = (X_{i1}^*, \ldots, X_{iD}^*) \quad \text{for } 1 \leq i \leq n_b,$$

where $X_{id}^* = \max\{X_{jd} : (i-1)b + 1 \leq j \leq ib\}$. Note that a block maximum does not necessarily coincide with any actual observation in the block. Let the distribution of a block maximum be F^* with component distributions F_d^*. For failure regions satisfying

$$X_1 \in B \text{ or } X_2 \in B \implies (\max\{X_{11}, X_{21}\}, \ldots, \max\{X_{1D}, X_{2D}\}) \in B, \tag{7.2}$$

the probability of a block maximum X^* falling in B is at least as large as the probability for an arbitrary observation X in the block. Basing inferences solely on the block maxima therefore leads to potential over-estimation of the failure probability.

The failure probability $P(X \in B)$ factorises as

$$P(X \in B \mid X^* \in B)P(X^* \in B) + P(X \in B \mid X^* \notin B)P(X^* \notin B), \tag{7.3}$$

which is approximated by $P(X^* \in B)$ if only block maxima are used. The method described in the following section also begins by estimating $P(X^* \in B)$, but then estimates the correction factor $P(X \in B \mid X^* \in B)$. If B satisfies condition (7.2) then $P(X \in B \mid X^* \notin B)$ equals zero and nothing more is required. This is likely to be the case for most failure regions of interest. If condition (7.2) does not hold then estimates of the second term in factorisation (7.3) can also be obtained.

7.2 Modelling Multivariate Clusters

7.2.1 Block maxima

Let n_c of the n_b blocks (7.1) contain points in a region $A = \{x \in \mathbb{R}^D : x \not\leq u\}$ for some choice of high thresholds $u = (u_1, \ldots, u_D)$. For suitably large block length, these n_c collections of points can be considered independent clusters because of the limiting behaviour described by Theorem 3.6. This motivates the Poisson process of Theorem 2.4 as a model for block maxima in A if the component distributions F_d^*, $1 \leq d \leq D$, are standard Fréchet. If the component distributions are unknown then they can be estimated using generalised Pareto forms above thresholds and empirical distribution functions below thresholds, as detailed in Section 2.2. The transformation of the d-th component to standard Fréchet scale is $-1/\log F_d^*(\cdot)$. Define also the transformation

$$\Psi_d^*(\cdot) = -\log\{1 - F_d^*(\cdot)\} \tag{7.4}$$

of the d-th component to standard exponential scale.

7.2.2 Strings

As discussed in Section 7.1, modelling only block maxima can cause failure probabilities to be over-estimated. To obtain unbiased estimates it is necessary to retain information about the distribution of actual points within a block. For extreme failure regions, only the most extreme points in a block, such as those that fall in the region A, are relevant. Suppose again that n_c of the n_b blocks (7.1) contain points in A. Denote these n_c clusters of points by $\{X_j : j \in \mathcal{S}_i\}$, $1 \leq i \leq n_c$, where $\mathcal{S}_i$ contains the indices of the points in A for the i-th cluster. It was seen in Chapter 6 that for univariate clusters the distribution of cluster points relative to their cluster maximum is described by a string. An analogous decomposition for multivariate clusters is given below. This characterises the position of extreme points in a block relative to the block maximum and so provides the necessary information for estimating failure probabilities accurately.

Recall from Section 6.1 that the marginal distribution, F, must be specified in order to define strings associated with clusters. Let Ψ_d transform the d-th component to have standard exponential distribution: write $Z_{jd} = \Psi_d(X_{jd})$, $j \in \mathcal{S}_i$, for the transformed points of the i-th cluster and $Z^*_{id} = \Psi^*_d(X^*_{id})$ for the transformation of the block maximum X^*_i. The D-dimensional string point corresponding to observation X_j in the i-th cluster is defined to be $Y_{ij} = (Y_{ij1}, \ldots, Y_{ijD})$, where

$$Y_{ijd} = \exp(Z^*_{id} - Z_{jd});$$

the string associated with the i-th cluster is the collection $Y_i = \{Y_{ij} : j \in \mathcal{S}_i\}$.

This string definition is analogous to the univariate definition (6.3) for each component. The treatment of multivariate clusters in this way pre-empts theoretical developments: no multivariate extension of Theorem 3.3 has yet been published. The definition of strings seems natural nevertheless, and at least encompasses the structure of clusters in the M_4 process (3.11).

The choice of declustering region A means that clusters can include points below the threshold in up to $D-1$ components. In contrast to the univariate case, therefore, the transformation Ψ_d must be applicable above and below u_d. The method of Section 2.2 can be used, that is $\Psi_d(\cdot) = -\log\{1 - F_d(\cdot)\}$, where F_d is modelled with a generalised Pareto distribution above u_d and the empirical distribution function below u_d. Note that this transformation is based on all observations $\{X_i\}_{i=1}^n$, not just the block maxima that were used to estimate the transformation (7.4).

Identifying clusters by grouping observations into blocks does not accord with the theoretically motivated intervals declustering scheme proposed in Section 4.2. Intervals declustering considers points only in the extreme region A, however, and so yields only the n_c cluster maxima, not the n_b block maxima. If only the cluster maxima are available then estimation of the component distributions required for the Poisson process model is impossible without making parametric assumptions. The problem is the estimation below thresholds: empirical distributions are inappropriate because any points with one component below threshold have at least one other component above threshold; furthermore, points with components below thresholds could be sparse. For this reason, blocks declustering is used in this chapter. As a compromise, the block length b is chosen to yield the same number of clusters in A as would be identified by intervals declustering.

7.2.3 Failure probabilities

Block maxima have been modelled with a Poisson process and clusters of extreme points within blocks have been identified with strings. If a block maximum X^* is simulated from the fitted Poisson process and transformed to a block maximum $Z^* = \Psi^*(X^*)$ with standard exponential components then an observed string $Y_i = \{Y_{ij} : j \in \mathcal{S}_i\}$ can be attached to form cluster points Z_j with $Z_{jd} = Z_d^* - \log Y_{ijd}$. These points can then be transformed back to the original space by inverting Ψ^*. This enables the failure probability $P(X \in B)$ to be estimated in the following way.

First simulate a large number, n_s, of block maxima from the Poisson process model in a region B_0 containing the failure region B. This is perhaps most easily done in standard Fréchet space. Let $\tilde{B}$ be the failure region B transformed to standard Fréchet space using the transformations $-1/\log F_d^*(\cdot)$. Block maxima can then be simulated in a region $\tilde{B}_0 = \{x \in \mathbb{R}^D : x_1 + \ldots + x_D > r_0\}$ with r_0 chosen so that $\tilde{B} \subset \tilde{B}_0$ (Nadarajah, 1999). Clusters can be constructed by transforming to standard exponential space, attaching the n_c observed strings to each of the simulated block maxima and transforming back to the original space. The probability that a block maximum falls in $\tilde{B}_0$ is $1 - \exp\{-\Lambda(\tilde{B}_0)\} = 1 - \exp(-D/r_0)$, where Λ is the intensity measure of the Poisson process; see Coles and Tawn (1994). If n_m of the simulated block maxima fall inside B then a Monte Carlo estimate of $P(X^* \in B)$ is $(n_m/n_s)\{1-\exp(-D/r_0)\}$. To each maximum, n_c strings are attached, creating $n_m n_c$ blocks with block maximum in B. Nominally, there are b points in each block, but the strings only describe the extreme cluster points; the remaining points in a block are assumed to be negligible. If n_x is the total number of points in B summed over all $n_m n_c$ blocks with block maximum in B then an estimate of the correction factor $P(X \in B \mid X^* \in B)$ is $n_x/(b n_m n_c)$. If $P(X \in B \mid X^* \notin B)$ is non-zero then a similar Monte Carlo approximation can be found for the second term in the factorisation (7.3).

In Chapter 6, it was noted that attaching strings to maxima outside the range of the data leads to a truncation of clusters above the threshold. The same issue arises in the multivariate case and can cause under-estimation of the correction factor. One remedy is to model strings in censored regions, as was done for the SWAP estimator in Section 6.3, but this will not be attempted here. Instead, the performance of the method will be investigated with a simulation study, ignoring the issue of censoring.

7.3 Simulation Study

To illustrate the practical effect of estimating the correction factor, the model of the previous section is applied to data simulated from five M_4 processes (3.11).

Bivariate processes are used, with form

$$X_{id} = \max_{1 \le l \le L} \max_{1 \le j \le J} (\alpha_{ljd} Z_{l,i-j}) \quad \text{for } i \ge 1 \text{ and } d = 1, 2,$$

where the $Z_{l,i}$ are independent, standard Fréchet random variables. The coefficients are constrained to ensure that the component distributions of the X_i are also standard Fréchet, and $L = 1$ so that limiting clusters have only a single deterministic shape. The coefficients of the five processes are listed in Table 7.1 together with a sketch of the limiting cluster shape. The coefficients are such that cluster maxima tend to lie on the line $x_1 = x_2$ and are not members of clusters, except for process 1 where clusters comprise just single points. Processes 2 and 3 have two points per cluster, each of which contributes a coordinate to the cluster maximum. The distance between the cluster points and the cluster maximum is greater in process 3. Processes 4 and 5 are the same as processes 2 and 3 except that clusters have an additional point that does not contribute to the maximum.

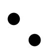
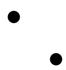
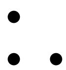
Process	J	$\{\alpha_{1j1}\}_{j=1}^{J}$	$\{\alpha_{1j2}\}_{j=1}^{J}$	Shape
1	1	$\{1\}$	$\{1\}$	
2	2	$\{1, 2\}/3$	$\{2, 1\}/3$	
3	2	$\{1, 4\}/5$	$\{4, 1\}/5$	
4	3	$\{1, 1, 2\}/4$	$\{2, 1, 1\}/4$	
5	3	$\{1, 1, 4\}/6$	$\{4, 1, 1\}/6$	

Table 7.1: Coefficients and cluster shapes for the five M_4 processes.

For each of the five processes, $N = 100$ sequences of length 1000 are simulated and a failure probability $P(X \in B)$ is estimated for each data-set. The failure region is $B = \{x \in \mathbb{R}^2 : x_1 + x_2 > v\}$ with v chosen so that the true failure probability is $p_0 = 1 - \exp(-2/v) = 0.001$. To estimate p_0, intervals declustering is applied with the region $A = \{x \in \mathbb{R}^2 : x \not\leq u\}$, where $u = (u_1, u_2)$ is defined by the empirical 95% quantiles of the two components. The block length is then chosen to produce the same number of clusters in A. Instead of fitting the full Poisson process model to the block maxima, the component distributions are estimated separately and the limiting dependence structure, for which block maxima fall on the line $x_1 = x_2$, is assumed known.

For each data-set, Monte Carlo estimates of $P(X^* \in B)$ and $P(X \in B)$ are computed by simulating one-thousand block maxima from the fitted Poisson process. Their performances as estimates of p_0 are summarised with two measures: relative bias and relative standard deviation, defined for estimates $\{\hat{p}_i\}_{i=1}^N$ as

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\hat{p}_i - p_0}{p_0}\right) \quad \text{and} \quad \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\hat{p}_i - \bar{p}}{p_0}\right)^2},$$

where $\bar{p} = \sum_{i=1}^{N} \hat{p}_i / N$. The simulation results are presented in Tables 7.2 and 7.3. They show that $P(X^* \in B)$ vastly over-estimates p_0 and that this is greater when the distance from actual points to the block maximum is larger. Estimating the correction factor removes the bias almost completely and reduces the variance. The correction is equally effective for all five processes.

Process	1	2	3	4	5
$P(X^* \in B)$	1.32	6.09	7.11	6.57	9.14
$P(X \in B)$	0.02	0.09	0.11	−0.01	0.04

Table 7.2: Relative bias

Process	1	2	3	4	5
$P(X^* \in B)$	2.98	6.25	6.88	7.93	8.91
$P(X \in B)$	0.70	0.90	0.88	0.99	0.96

Table 7.3: Relative standard deviation

7.4 Data Example

In this section, the model for multivariate clusters is applied to an oceanographic, bivariate time series. The variables, both measured in metres, are surge (X_1) and significant wave height (X_2). Surge is the measured sea-level with the tidal component removed; significant wave height is approximately equal to the mean height of the highest one-third of the waves, and is hindcast. The data were recorded hourly at Christchurch on the south coast of England between 1 April 1978 and 31 March 1990. See Hawkes *et al.* (1998) for additional information and analysis. The raw data exhibit strong seasonality, removal of which is achieved here by using data from only the months December and January. Missing values are avoided by restricting attention to the period December 1980 to January 1987. The two variables are plotted separately in Figure 7.1, and jointly in Figure 7.2.

The first step is to select thresholds over which the two series exhibit stable extremal behaviour. To aid this selection, estimates of the extremal index, the mean excess and the parameters (σ, ξ) of the generalised Pareto distribution are plotted over a range of thresholds in Figures 7.3 and 7.4. The scale parameters of the

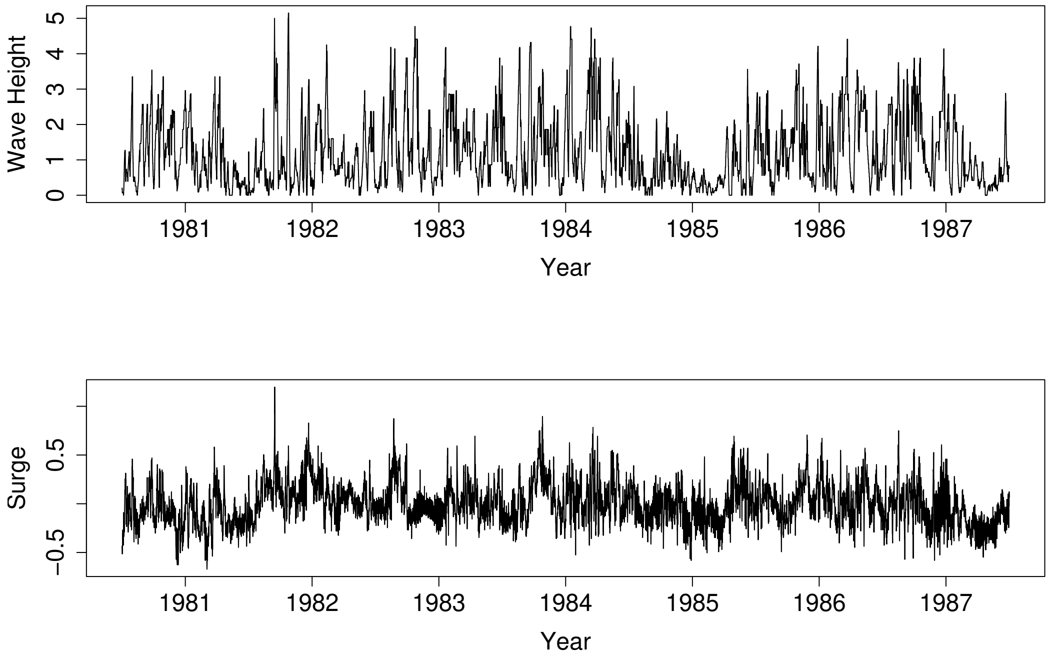


Figure 7.1: Significant wave height (above) and surge (below) recorded hourly in metres at Christchurch. The first observation in each year is labelled.

generalised Pareto distributions have been modified to $\sigma - \xi u$, since this quantity should be approximately constant across thresholds: see Section 6.4. For surge, stability appears to be achieved at $u_1 = 0.4$ if the rogue estimates of the generalised Pareto parameters at threshold 0.6 are ignored. For wave height, the estimates of the scale parameter are confusing. The threshold $u_2 = 3.0$ is tentatively selected. Diagnostic plots for the fit of the generalised Pareto distribution to the 402 surge exceedances and to the 625 wave exceedances are shown in Figure 7.5. Allowing for discreteness, the fit is acceptable for both variables.

Multivariate intervals declustering with extreme region $A = \{x \in \mathbb{R}^2 : x \not\leq u\}$ yields an extremal index estimate of 0.07 and 61 clusters. Blocks declustering also yields 61 clusters if the block length is $b = 82$. This partitions the data into 128

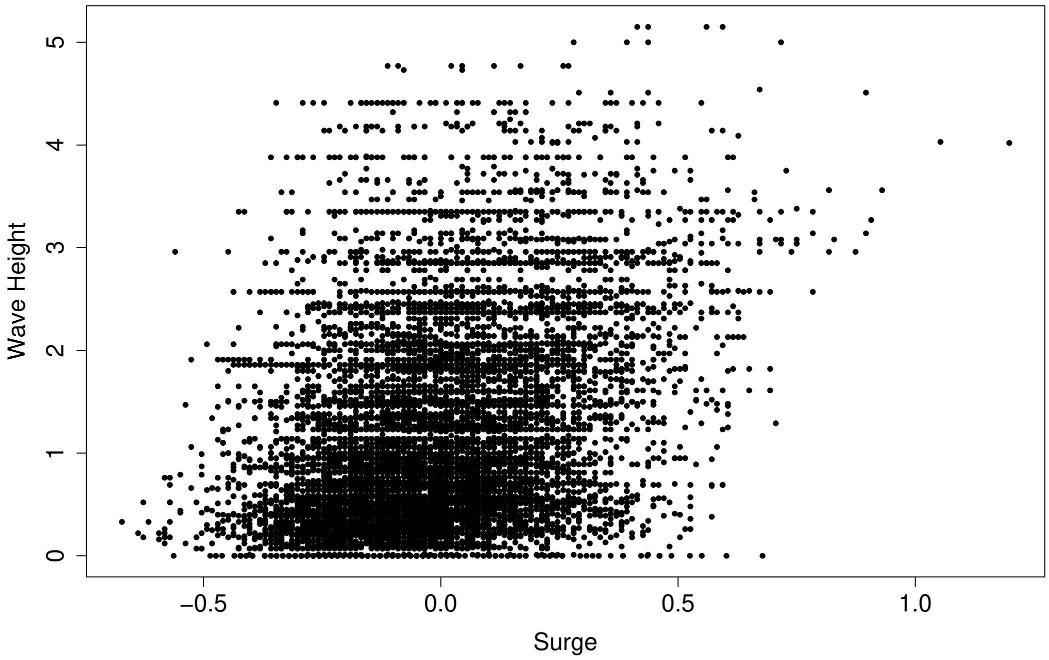


Figure 7.2: Significant wave height against surge, both recorded hourly in metres at Christchurch.

blocks, of which the maxima are plotted in Figure 7.6.

The Poisson process model is fitted to the block maxima with logistic dependence structure (5.6). The parameter estimates for the generalised Pareto components are $(0.20, 0.02)$ for surge and $(1.55, -0.67)$ for wave height; the dependence parameter estimate is 0.45.

The failure region $B = \{x \in \mathbb{R}^2 : x_1 + x_2/8 > v\}$ with $v = 2$ is used to illustrate the estimation of failure probabilities. This form of failure region approximates the conditions under which certain types of sea-wall are breached: see Coles and Tawn (1994). One-thousand cluster maxima are simulated from the fitted model in a region containing B, and superimposed on Figure 7.6. Note the upper limit of $u_2 - \sigma_2/\xi_2 = 5.32$ metres for the estimated significant wave height distribution. The

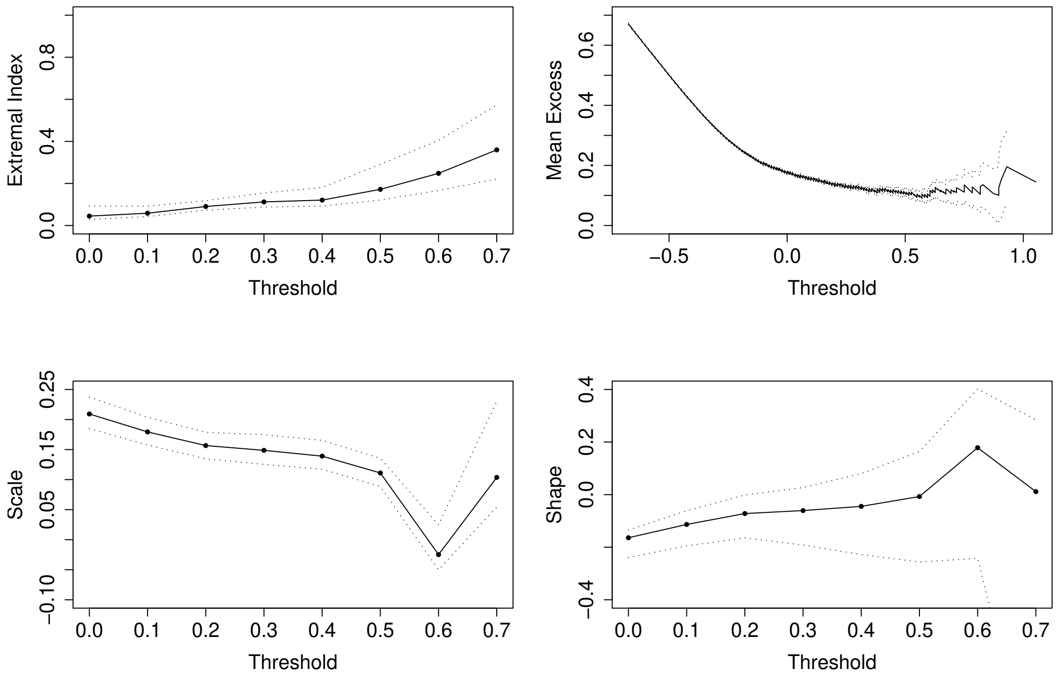


Figure 7.3: Estimates (–●–) for surge of the extremal index, mean excess, and parameters of the generalised Pareto distribution with bootstrapped 90% confidence intervals $(\cdots)$.

estimate of $P(X^* \in B)$ is 0.0039. For $n_m n_c = 38\,918$ clusters with a maximum in B, 163 955 out of the nominal $bn_m n_c = 3\,191\,276$ points are in B. So the estimate of the correction factor is $163\,955/3\,191\,276 = 0.051$ and the estimate of $P(X \in B)$ is 0.00020.

The estimate of $P(X^* \in B)$ is in accordance with the data: there are 128 blocks in the sample, none of which has maximum in B, suggesting that the probability should be less than $1/128 = 0.0078$. There are 10 416 observations in the sample, and again none of these is in B, which suggests that $P(X \in B)$ should be less than $1/10\,416 = 0.0001$. Although the estimate is double this value, it is not at odds with the data because points will tend to occur in B in clusters. Indeed, the

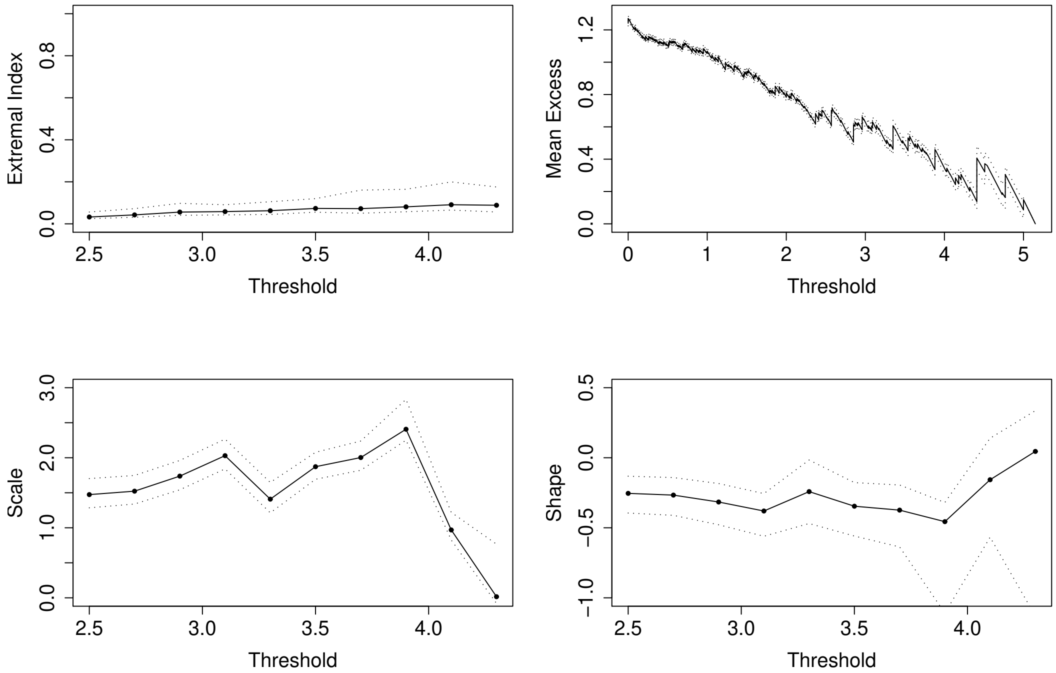


Figure 7.4: Estimates (–•–) for significant wave height of the extremal index, mean excess, and parameters of the generalised Pareto distribution with bootstrapped 90% confidence intervals ($\cdots$).

average number of points in B among the simulated clusters that have a point in B is about 4.

7.5 Discussion

The technique described in this chapter allows failure probabilities to be estimated without prior filtering of the data, which can lead to over-estimation. Only simple failures, caused by a single observation falling in a failure region, have been considered, but failures might result from more complex events. For example, failure may occur only if two consecutive observations fall in the failure region. The probabilities of such failures cannot be estimated using filtered data, but are handled

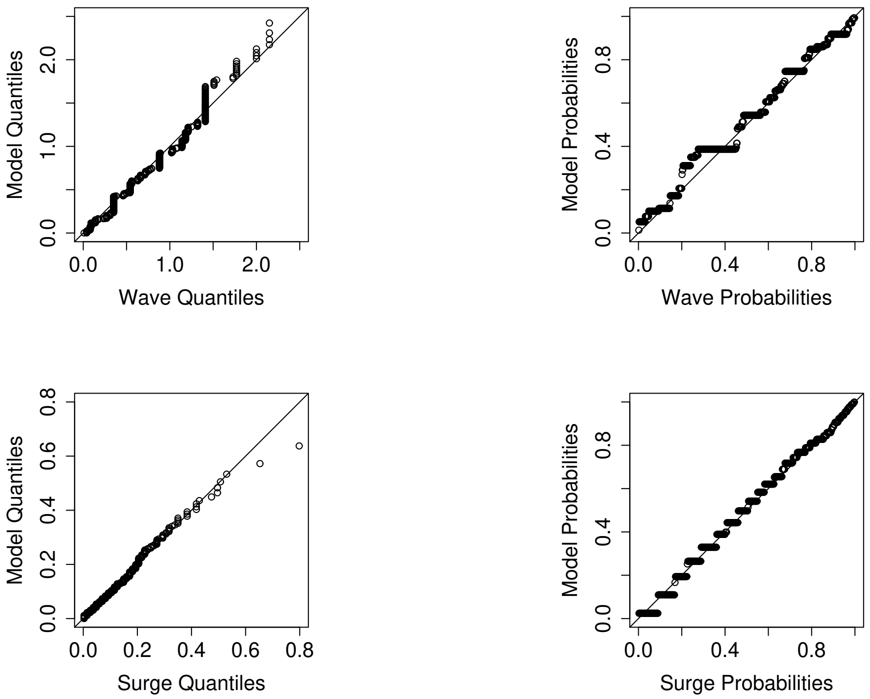


Figure 7.5: Quantile plots (left) and probability plots (right) for the generalised Pareto fit at the chosen thresholds for significant wave height (top) and surge (bottom).

naturally when clusters are modelled.

Obtaining confidence intervals for failure probabilities using the method of this chapter is complicated. In order to account for uncertainty in the block length, the strings and the Poisson process model, it is perhaps simplest to apply a block bootstrap to the original multivariate sequence. Using the block length identified by the declustering procedure would appear to be a natural choice for bootstrapping, but this issue has not been explored here.

Finally, estimates of the failure probability $P(X \in B)$ can be obtained from $P(X^* \in B)$ in another way if B has a particular form. Suppose that the se-

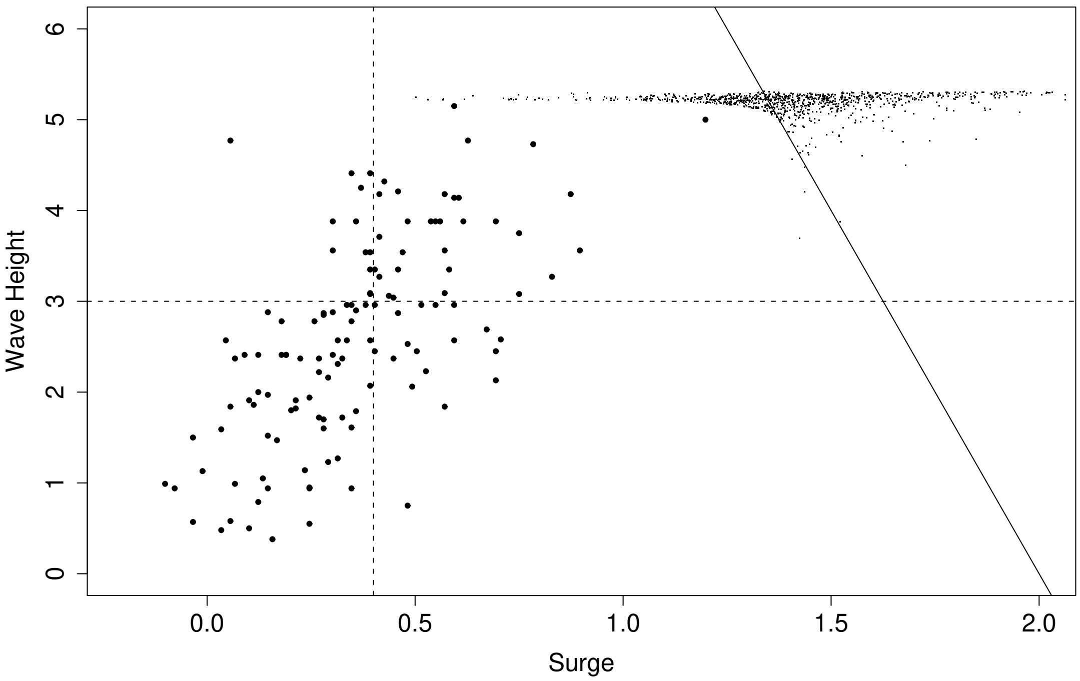


Figure 7.6: Block maxima (●), component thresholds (- - -), boundary of the failure region (——) and simulated cluster maxima (·).

quence has multivariate extremal index $\theta(\cdot)$ satisfying limit (3.2). Then, for $B = \{x \in \mathbb{R}^D : x \not\leq u\}$, and a suitably large block length b,

$$P(X^* \notin B) \approx P(X \notin B)^{\theta(u)b}.$$

Combining an estimate of $\theta(u)$, perhaps obtained using one of the methods described in Section 4.2, with an estimate of $P(X^* \notin B)$ yields an estimate of $P(X \in B)$. Whether or not such approximations can be exploited for general failure regions B remains an open question.

Bibliography

Barão, M. I. and Tawn, J. A. (1999) Extremal analysis of short series with outliers: sea-levels and athletics records. *Appl. Statist.*, **48**, 469–487.

Barnett, V. (1976) The ordering of multivariate data (with discussion). *J. Roy. Statist. Soc. A*, **139**, 318–355.

Beirlant, J., Teugels, J. and Goegebeur, Y. (2003) *Statistics of Extremes: Theory and Applications*. London: Wiley. In preparation.

Berman, S. M. (1964) Limit theorems for the maximum term in stationary sequences. *Ann. Math. Statist.*, **35**, 502–516.

Capéraà, P., Fougères, A.-L. and Genest, C. (1997) A nonparametric estimation procedure for bivariate extreme value copulas. *Biometrika*, **84**, 567–577.

Coles, S. G. (1991) *Statistical Methodology for the Multivariate Analysis of Environmental Extremes*. PhD thesis, Sheffield University.

Coles, S. G. (2001) *An Introduction to Statistical Modeling of Extreme Values*. London: Springer.

Coles, S. G. and Tawn, J. A. (1991) Modelling extreme multivariate events. *J. Roy. Statist. Soc. B*, **53**, 377–392.

Coles, S. G. and Tawn, J. A. (1994) Statistical methods for multivariate extremes: an application to structural design (with discussion). *Appl. Statist.*, **43**, 1–48.

David, H. A. (1981) *Order Statistics.* Second edition. London: Wiley.

Deheuvels, P. (1983) Point processes and multivariate extreme values. *J. Mult. Anal.*, **13**, 257–273.

Deheuvels, P. (1985) Point processes and multivariate extreme values II. In *Multivariate Analysis VI* (P. R. Krishnaiah, Ed.), 145–164. Amsterdam: North Holland.

Deheuvels, P. (1991) On the limiting behaviour of the Pickands estimator for bivariate extreme-value distributions. *Statist. Probab. Letters*, **12**, 429–439.

Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997) *Modelling Extremal Events.* Springer.

Ferro, C. A. T. and Segers, J. (2003a) Inference for clusters of extreme values. *J. Roy. Statist. Soc. B*, **65**, 545–556.

Ferro, C. A. T. and Segers, J. (2003b) Extreme value methods in time series analysis. To appear in *Statistics of Extremes: Theory and Applications* (J. Beirlant, J. Teugels and Y. Goegebeur). London: Wiley.

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

Frigessi A., Haug, O. and Rue, H. (2002) A dynamic mixture model for unsupervised tail estimation without threshold selection. *Extremes*, **5**, 219–235.

Galambos, J. (1975) Order statistics of samples from multivariate distributions. *J. Amer. Statist. Assoc.*, **70**, 674–680.

Ghoudi, K., Khoudraji, A. and Rivest, L.-P. (1998) Propriétés statistiques des copules de valeurs extrêmes bidimensionnelles. *Canad. J. Statist.*, **26**, 187–197.

Guillou, A. and Hall, P. (2001) A diagnostic for selecting the threshold in extreme value analysis. *J. Roy. Statist. Soc. B*, **63**, 293–305.

Gumbel, E. J. (1960) Bivariate exponential distributions. *J. Amer. Statist. Assoc.*, **55**, 698–707.

de Haan, L. (1985) Extremes in higher dimensions: the model and some statistics. *Proc. 45th Session, Int. Stat. Inst.*, paper 26.3. Amsterdam: International Statistical Institute.

de Haan, L. and Resnick, S. I. (1977) Limit theory for multivariate sample extremes. *Z. Wahr. Ver. Geb.*, **40**, 317–337.

de Haan, L. and de Ronde, J. (1998) Sea and wind: multivariate extremes at work. *Extremes*, **1**, 7–45.

de Haan, L. and Sinha, A. K. (1999) Estimating the probability of a rare event. *Ann. Statist.*, **27**, 732–759.

Hall, P. and Presnell, B. (1999) Density estimation under constraints. *J. Comp. Graph. Stat.*, **8**, 259–277.

Hall, P. and Tajvidi, N. (2000) Distribution and dependence-function estimation for bivariate extreme-value distributions. *Bernoulli*, **6**, 835–844.

Hawkes, P. J., Gouldby, B., Yarde, A., Tawn, J. A. and Bortot, P. (1998) The joint probability of waves and water levels: JOIN-SEA. HR Wallingford, report SR 537.

Heffernan, J. and Tawn, J. A. (2003) A conditional approach for multivariate extreme values. To appear in *J. Roy. Statist. Soc. B*.

Hsing, T. (1987) On the characterization of certain point processes. *Stochastic Process. App.*, **26**, 297–316.

Hsing, T. (1989) Extreme value theory for multivariate stationary sequences. *J. Mult. Anal.*, **29**, 274–291.

Hsing, T. (1991) On tail index estimation using dependent data. *Ann. Statist.*, **19**, 1547–1569.

Hsing, T., Hüsler, J. and Leadbetter, M. R. (1988) On the exceedance point process for a stationary sequence. *Probab. Theory Related Fields*, **78**, 97–112.

Hüsler, J. and Reiss, R.-D. (1989) Maxima of normal random vectors: between independence and complete dependence. *Statist. Probab. Letters*, **7**, 283–286.

Joe, H. (1989) Families of min–stable multivariate exponential and multivariate extreme value distributions. *Statist. Probab. Letters*, **9**, 75–82.

Joe, H., Smith, R. L. and Weissman, I. (1992) Bivariate threshold methods for extremes. *J. Roy. Statist. Soc. B*, **54**, 171–183.

Kotz, S. and Nadarajah, S. (2000) *Extreme Value Distributions: Theory and Applications*. Imperial College Press.

Leadbetter, M. R. (1974) On extreme values in stationary sequences. *Z. Wahr. Ver. Geb.*, **28**, 289–303.

Leadbetter, M. R. (1983) Extremes and local dependence in stationary sequences. *Z. Wahr. Ver. Geb.*, **65**, 291–306.

Leadbetter, M. R. (1991) On a basis for 'Peaks over Threshold' modeling. *Statist. Probab. Letters*, **12**, 357–362.

Leadbetter, M. R., Lindgren, G. and Rootzén, H. (1983) *Extremes and Related Properties of Random Sequences and Processes*. New York: Springer.

Ledford, A. W. and Tawn, J. A. (2003) Diagnostics for dependence within time series extremes. *J. Roy. Statist. Soc. B*, **65**, 521–543.

Mori, T. (1977) Limit distributions of two-dimensional point processes generated by strong-mixing sequences. *Yokohama Math. J.*, **25**, 155–168.

Nadarajah, S. (1999) Simulation of multivariate extreme values. *J. Statist. Comput. Simul.*, **62**, 395–410.

Nadarajah, S. (2000) Approximations for bivariate extreme values. *Extremes*, **3**, 87–98.

Nadarajah, S. (2001) Multivariate declustering techniques. *Environmetrics*, **12**, 357–365.

Nandagopalan, S. (1994) On the multivariate extremal index. *J. Res. Natl. Inst. Stand. Technol.*, **99**, 543–550.

Newell, G. F. (1964) Asymptotic extremes for m-dependent random variables. *Ann. Math. Statist.*, **35**, 1322–1325.

O'Brien, G. L. (1987) Extreme values for stationary and Markov sequences. *Ann. Probab.*, **15**, 281–291.

Parzen, E. (1962) On estimation of a probability density function and mode. *Ann. Math. Statist.*, **33**, 1065–1076.

Perfekt, R. (1997) Extreme value theory for a class of Markov chains with values in $\mathbb{R}^d$. *Adv. Appl. Probab.*, **29**, 138–164.

Pickands, J. III (1971) The two-dimensional Poisson process and extremal processes. *J. Appl. Probab.*, **8**, 745–756.

Pickands, J. III (1975) Statistical inference using extreme order statistics. *Ann. Statist.*, **3**, 119–131.

Pickands, J. III (1981) Multivariate extreme value distributions. *Proc. 43rd Session, Int. Stat. Inst.*, 859–878.

Resnick, S. I. (1987) *Extreme Values, Regular Variation, and Point Processes.* New York: Springer–Verlag.

Sibuya, M. (1960) Bivariate extreme statistics. *Ann. Inst. Statist. Math.*, **11**, 195–210.

Smith, R. L. (1985) Statistics of extreme values. *Proc. 45th Session, Int. Stat. Inst.*, paper 26.1. Amsterdam: International Statistical Institute.

Smith, R. L. (1992) The extremal index for a Markov chain. *J. Appl. Probab.*, **29**, 37–45.

Smith, R. L., Tawn, J. A. and Coles, S. G. (1997) Markov chain models for threshold exceedances. *Biometrika*, **84**, 249–268.

Smith, R. L., Tawn, J. A. and Yuen, H. K. (1990) Statistics of multivariate extremes. *Int. Statist. Rev.*, **58**, 47–58.

Smith, R. L. and Weissman, I. (1994) Estimating the extremal index. *J. Roy. Statist. Soc. B*, **56**, 515–528.

Smith, R. L. and Weissman, I. (1996) Characterization and estimation of the multivariate extremal index. Unpublished.

Tawn, J. A. (1988) Bivariate extreme value theory: models and estimation. *Biometrika*, **75**, 397–415.

Wand, M. P. and Jones, M. C. (1995) *Kernel Smoothing*. New York: Chapman & Hall.

Weissman, I. and Novak, S. Y. (1998) On blocks and runs estimators of the extremal index. *J. Statist. Plan. Inf.*, **66**, 281–288.