Statistical models for dependent and non-stationary extreme events.

Emma Eastoe, B.Sc.

Submitted for the degree of Doctor of Philosophy

at Lancaster University,

November 2007.

Statistical models for dependent and non-stationary extreme events.

Emma Eastoe, B.Sc.

Submitted for the degree of Doctor of Philosophy at Lancaster University, November 2007.

Abstract

Extreme value methods are used in a wide range of applications, for example they may be used for modelling wave heights and river levels in hydrology, wind speeds in structural engineering and share price return levels in economics. Many statistical models and methods of inference exist for the extreme values of univariate sequences of independent and identically distributed (IID) random variables. However, in most applications, the data sets are not IID and are often multivariate, and yet methods for modelling the extremes of sequences which fail to fulfil one, or both, of the IID assumptions and (or) are multivariate remain the subject of ongoing research. The work contained in this thesis is a contribution to this area.

Most of our work has been motivated by a multivariate air pollution data set, which shows complex seasonal trends and covariate relationships. We begin with a model for the extremes of a univariate sequence which displays short-range dependence within the sample extremes. Next we propose a method for modelling the extremes of a non-stationary univariate process; we then extend this methodology to model a multivariate process with non-stationary marginal and dependence structures. Finally we consider a new estimator for the dependence structure of a sequence of multivariate extremes which are pairwise dependent.

Acknowledgements

I would like to acknowledge the input into this thesis by my three supervisors. Thanks to Jan Heffernan, who initiated the project and introduced me to the world of extreme values, whilst guiding my first attempts at research. Also to Jon Tawn, who took over my supervision during the second half of the project, and has taught me some more about the statistical application of extreme value theory as well as providing much input on how to write up my research as publishable papers. Finally to Crispin Halsall, of the Department of Environmental Sciences also at Lancaster University, who provided some interesting information on atmospheric chemistry and air pollution. Acknowledgement is also due to the UK Meteorological Office and the British Atmospheric Data Centre for provision of the meteorological data used in Chapters 3 and 4.

I would also like to thank the people in the Department of Mathematics and Statistics at Lancaster University for an interesting and enjoyable three years, and in particular thanks to my fellow PhD students, especially for making my time as student rep completely without incident. Special mention to Gerwyn Green, John Minty and John Speakman, with whom I have shared an office for the last few years, also without incident! Thanks to Chris Sherlock, for putting up with all my doubts that this thesis would ever come into being and for providing much support and encouragement. Lastly thanks to my parents, Gay and Richard, and my brothers, Pete, Phil and Duncan, for their understanding and for occasionally reminding me that there might be more important things in life than statistics.

Declaration

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

Chapters 2 and 3 are joint work with Jonathan Tawn and have both been submitted separately for publication. Chapter 5 is joint work with both Jan Heffernan and Jonathan Tawn and has also been submitted for publication. I would like to acknowledge comments from various anonymous referees which have aided improvement of the work in all three chapters.

The computing code used has mostly been my own, although I have made use of the R libraries ismev and evd written by Alec Stephenson, and some of the code required for computations in Chapter 5 was written by Jan Heffernan.

Emma Eastoe

Contents

List of Figures

Chapter 1

Introduction

This chapter serves to introduce both the basic concepts of extreme value theory and statistical inference on extremes and also to discuss some of the particular problems associated with modelling extreme air pollution events. It may also be seen as a literature review covering some of the main advances in methods for the analysis of extremes over the last couple of decades.

1.1 Thesis outline

Besides this introduction, the thesis has four chapters. Of these, Chapters 2, 3 and 5 have been submitted individually for publication. Consequently references and appendices are given at the end of each chapter, rather than at the end of the thesis. Further, because each chapter is self-contained, the notation is not necessarily consistent between them; this should not however pose a problem since the necessary notation is defined in the introduction to each chapter. Chapter 4 is an extension of Chapter 3 and contains ongoing research.

The motivation for the work in Chapters 2, 3 and 4 comes from an attempt to model the extreme values of a series of surface level air pollution data, consisting of maximum daily concentrations of nitric oxide (NO), nitrogen dioxide (NO_2) and ozone (O_3). These data are discussed in greater detail in Section 1.4, along with a brief review of other published attempts to model extreme air pollution events. In the course of analysing the data several results arose which caused us to question existing modelling methods and to search for possible alternative approaches.

The question of how to analyse the local extremes of a sequence of dependent events is tackled in Chapter 2. We propose an approach that accounts for the fact that, in practice, only sub-asymptotic, rather than asymptotic, levels of a process are observed. This is in contrast to existing methods which are motivated by an underlying theory which relies on observing asymptotic levels. The method is illustrated using the ozone data.

In Chapter 3 we develop new methodology for modelling the extremes of a non-stationary process and compare this to existing methodology. Our proposal is to pre-process the data using covariates, thus removing non-stationarity from the whole data-set, and then use existing methodology to model the extremes of the pre-processed data. Certainly for the data sets we have looked at, this method has computational advantages. We suggest that it also has a theoretical advantage when compared to the existing method and that it also simplifies model interpretation. Ultimately we use our method to estimate extreme levels of ozone using NO, NO_2 and a range of meteorological variables as covariates.

We attempt to extend this work to analyse multivariate extremes in Chapter 4. We consider a hierarchical approach for estimating extreme levels of ozone given values of NO, NO_2 and the meteorological covariates, by first modelling NO conditional on the covariates, then NO_2 conditional on both the covariates and NO and finally ozone conditional on the covariates, NO and NO_2. This has the advantages that we can account for uncertainty in the NO and NO_2 measurements, which we cannot if we simply treat them as covariates, and that we can extrapolate into the tails of their distributions.

Finally, in Chapter 5 we present a new non-parametric estimator to measure the degree of association between dependent extreme random variables. This extends the work of Heffernan and Tawn (2004), see Section 1.3, in the special case of asymptotic dependence. We show that, in this special case, their model is at least comparable with existing competitors, but has the advantage of extending to cover a much broader range of dependence structures.

1.2 Extreme value theory

The aim of extreme value theory is to provide probabilistic results which allow the characterisation of the tail behaviour of any probability distribution without requiring knowledge of the form of this underlying distribution. This allows us to develop methods for statistical inference on extreme values which need not use information from data observed in the body of the distribution. In this section we state some of the main results from extreme value theory which will be useful in later chapters. We omit proofs, since these are well documented elsewhere. For further details on the univariate case see Leadbetter *et al.* (1983) and Resnick (1987); references for the multivariate case are given in Section 1.2.2.

1.2.1 Univariate case

Let us suppose that we have an independent and identically distributed (IID) sequence of random variables $\{Y_i\}$ with distribution function F. We say that F is the marginal distribution of the sequence. Let $M_{r,n}$ denote the rth largest of the first n of these, so that, for example, the *block maxima* is given by $M_{1,n} = \max\{Y_1, \ldots, Y_n\}$. We begin with a result concerning the asymptotic distribution of the block maxima which is formally stated as follows.

Theorem 1.2.1 *If $Y_1, \ldots, Y_n$ are IID random variables with distribution function F and $\{a_n > 0\}$, $\{b_n\}$ are sequences of normalising constants such that, as $n \to \infty$,*

$$\Pr\left(\frac{M_{1,n} - b_n}{a_n} \leq z\right) \to G(z)$$

where G is a non-degenerate distribution function, then G takes the form of the *generalised extreme value distribution (GEVD) which is defined by*

$$G(z) = \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]_{+}^{-1/\xi}\right\}, \qquad -\infty < z < \infty,\ \sigma > 0 \qquad (1.2.1)$$

where $a_{+} = \max\{a, 0\}$.

The GEVD parameters (μ, σ, ξ) are referred to as the location, scale and shape parameters respectively. As $\xi \to 0$ we take limits to obtain the Gumbel distribution which has the form

$$G(z) = \exp\{-\exp(-(z-\mu)/\sigma)\}, \qquad -\infty < z < \infty.$$

We say that F is in the *domain of attraction* of G and the sign of ξ is determined by the rate of decay of F. If $\xi < 0$, the (negative) Weibull case, then F has a finite upper end point *i.e.* F is light tailed. If $\xi > 0$, the Fréchet case, then F is heavy tailed with infinite upper end point. In the limit as $\xi \to 0$, the Gumbel case, the tail of F decays exponentially. Examples of distributions lying in each of the domains of attraction are; for the (negative) Weibull case, the uniform and beta distributions, for the Fréchet case, the inverse and log gamma distributions and for the Gumbel case, the normal, gamma, logistic, log normal and exponential distributions.

The second important result in this section refers to all the extremes of the sequence $\{Y_i\}$ rather than just the block maxima. It concerns the asymptotic distribution of the point process P_n, which is defined by

$$P_n = \left\{\left(\frac{i}{n+1}, \frac{Y_i - b_n}{a_n}\right) : i = 1, \ldots, n\right\} \qquad (1.2.2)$$

and is stated as follows.

Theorem 1.2.2 *If $Y_1, \ldots, Y_n$ are IID random variables with distribution F and $\{a_n > 0\}$, $\{b_n\}$ are sequences of normalising constants such that Theorem 1.2.1 holds then, as $n \to \infty$, the point process $P_n \to P$, where P is a non-homogeneous*

Poisson process on $A = [0,1] \times [v, \infty)$ *with intensity measure*

$$\lambda(t,x) = \sigma^{-1}\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-1/\xi-1}. \tag{1.2.3}$$

Here $v > \inf\{z : G(z) > 0\}$*.*

Note that the parameters in the limiting non-homogeneous Poisson process are exactly the GEVD parameters and so are independent of the level v.

The final result given in this section for univariate IID extremes is a consequence of both the block maxima and Poisson process results; we defer a discussion of the link between the three results until Chapter 2 where it is stated in the slightly more general case of stationarity. The result concerns the exceedances made by the normalised sequence $\{(Y_i - b_n)/a_n\}$ of some high level u and is stated formally as follows.

Theorem 1.2.3 *If* $Y_1, \ldots, Y_n$ *are IID random variables and* $\{a_n > 0\}$*,* $\{b_n\}$ *are sequences of normalising constants such that Theorems 1.2.1 and 1.2.2 hold, then, as* $n \to \infty$*,*

$$\Pr\left(\frac{Y_i - b_n}{a_n} > u + v \,\middle|\, \frac{Y_i - b_n}{a_n} > u\right) \to \left[1 + \frac{\xi v}{\psi_u}\right]_+^{-1/\xi}, \qquad v > 0 \text{ and } \psi_u > 0. \tag{1.2.4}$$

This limiting distribution is known as the generalised Pareto distribution (GPD).

The GPD shape parameter ξ is the same as the GEVD shape parameter, and as such is invariant to selection of the threshold u. However the GPD scale parameter ψ_u does depend on the threshold and is related to this and the GEVD parameters by the expression

$$\psi_u = \sigma + \xi(u - \mu).$$

Note that evaluation of the normalising constants $\{a_n > 0\}$ and $\{b_n\}$ used in Theorems 1.2.1 - 1.2.3 requires knowledge of the exact distributional form of F.

For example, if $\{Y_i\}$ has unit exponential margins, so that $F(y) = 1 - \exp\{-y\}$ for $y > 0$, then we have $a_n = 1$ and $b_n = \log n$, since, as $n \to \infty$,

$$\begin{aligned}
\Pr\left(\frac{M_{1,n} - b_n}{a_n} \leq z\right) &= F^n(a_n z + b_n) \\
&= (1 - \exp\{-(a_n z + b_n)\})^n \\
&\sim 1 - n\exp\{-(a_n z + b_n)\} + \frac{n(n-1)}{2}\exp\{-2\,(a_n z + b_n)\} - \ldots \\
&\sim \exp\{-\exp(-z)\} \quad \text{taking } a_n = 1 \text{ and } b_n = \log n.
\end{aligned}$$

However, since the whole point of extreme value theory is to develop a method of inference for the tails that is independent of the underlying distribution, for purposes of inference, the normalising constants are usually absorbed into the GEVD location and scale parameters.

Now suppose that the sequence of random variables $\{Y_i\}$ are not IID but are stationary. We are still concerned with making inference on the tails of the marginal distribution F of the sequence rather than the joint distribution. Depending on the nature of the dependence in the sequence, it is possible that the extremes of the sequence will occur in *clusters*, so that seeing one extreme event makes it more likely that the following event will also be extreme. If such clustering occurs we say that the sequence is *asymptotically dependent*; the formal definition of this states that two random variables X_1 and X_2 are asymptotically dependent if, as $x \to \infty$,

$$\Pr(X_2 > x | X_1 > x) \to \tau, \quad \text{where } \tau > 0.$$

If $\tau = 0$ we say that X_1 and X_2 are asymptotically independent. This is a concept that will also be useful when we discuss multivariate extremes. If a sequence is asymptotically independent it has clusters of size 1 in the limit.

We now relate the behaviour of the extremes of the stationary sequence $\{Y_i\}$ to that of the associated IID sequence $\{\tilde{Y_i}\}$, which has the same margins as $\{Y_i\}$, but is independent. To do this we must first impose a mixing condition on the sequence $\{Y_i\}$ to restrict long-term dependence. There are many possible such conditions, one of which is the so-called $D(u_n)$ condition. This states that the events that two block maxima exceed some high level u_n are independent, so long as the blocks are sufficiently far apart. The formal result for the distribution of $M_{1,n} = \max\{Y_1, \ldots, Y_n\}$ is then given as follows.

Theorem 1.2.4 *Let $Y_1, \ldots, Y_n$ be a stationary sequence of random variables with marginal distribution F and let $\tilde{Y}_1, \ldots, \tilde{Y}_n$ be the associated IID sequence. Suppose that there exist sequences of normalising constants $\{a_n > 0\}$, $\{b_n\}$ such that Theorem 1.2.1 is satisfied for $\{\tilde{Y}_i\}$ and the condition $D(u_n)$ holds for $u_n = a_n z + b_n$ where z is such that the GEV distribution function in Theorem 1.2.1 is strictly greater than zero. Then, as $n \to \infty$,*

$$\Pr\left(\frac{M_{1,n} - b_n}{a_n} \le z\right) \to G^{\theta}(z) \tag{1.2.5}$$

where $0 \le \theta \le 1$ and G is the GEV distribution function defined in Theorem 1.2.1.

The constant θ is referred to as the *extremal index.* If the sequence is asymptotically independent then $\theta = 1$ and if $\theta < 1$ it is asymptotically dependent; the level of asymptotic dependence strengthens as $\theta \to 0$.

Similarly we can obtain analogues to the point process and threshold exceedance results given in Theorems 1.2.2 and 1.2.3. Rather than reproduce these here we refer the reader to Chapter 2 where these results are stated as necessary for our purposes.

1.2.2 Multivariate case

The theory for the multivariate case mirrors that of the univariate case in that the first result described is an analogue of the block maxima result given in Theorem 1.2.1, and the second an analogue of the point process approach described in Theorem 1.2.2. Note that it is standard practise in much of the extremes literature to assume fixed marginal distributions, thus separating out marginal and dependence effects; we follow this procedure. Further, we consider the IID case only.

For the first result, define a sequence of IID random variables $\{\mathbf{Y}_i\}$ where each random variable is a p-dimensional vector $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{ip})$ with joint distribution function F. We assume that the marginal distributions are all unit Fréchet, *i.e.* have distribution function $F_j(y) = \exp\{-1/y\}$ for $y > 0$ and $j = 1, \ldots, p$. We define the componentwise maxima to be the vector $\mathbf{M}_n$ with jth component $\mathbf{M_{n,j}} = \max_{1\leq i\leq n}\{Y_{ij}\}$ for $j = 1, \ldots, p$. Then we have the following result for the asymptotic distribution of the normalised componentwise maxima

Theorem 1.2.5 *If $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ is a sequence of IID p-dimensional random variables with joint distribution F, unit Fréchet margins and componentwise maxima $\mathbf{M}_n$ defined above then, as $n \to \infty$,*

$$\Pr\left(\frac{\mathbf{M}_n}{n} \leq \mathbf{z}\right) \to G(\mathbf{z})$$

where G is a non-degenerate distribution with distribution function

$$G(\mathbf{z}) = \exp\left\{-\int_{S_p} \max_{1\leq j\leq p}\left(\frac{w_j}{z_j}p \;\mathrm{d}H(\mathbf{w})\right)\right\} \qquad (1.2.6)$$

where $S_p = \left\{\mathbf{w} : \sum_{j=1}^{p} w_j = 1\right\}$ is the unit simplex and $H(\cdot)$ is a distribution function referred to as the spectral measure *on S_p satisfying*

$$\int_{S_p} w_j \;\mathrm{d}H(\mathbf{w}) = \frac{1}{p}, \qquad j = 1, \ldots, p. \qquad (1.2.7)$$

The normalising constant n^{-1} follows from the fact that the margins are unit Fréchet, since in this case

$$\Pr(M_n/n \leq z) = F^n(nz) = \exp\{-n/(nz)\} = \exp\{-1/z\}$$

which is a GEVD with parameters (1,1,1); thus the unit Fréchet distribution is in its own domain of attraction. The distribution G defined in equation (1.2.6) is known as the *multivariate extreme value distribution* (MVEVD).

Now we consider the point process approach. Consider the point process

$$P_n^* = \left\{ n^{-1}\mathbf{Y}_i : i = 1, \ldots, n \right\}.$$

Now define the pseudo-radial and -angular co-ordinates

$$R = \sum_{j=1}^{p} Y_j, \qquad W_j = Y_j/R, \qquad j = 1, \ldots, p. \tag{1.2.8}$$

Then the asymptotic behaviour of P_n is characterised as follows.

Theorem 1.2.6 *If $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ is a sequence of IID p-dimensional random variables with joint distribution F, unit Fréchet margins and which satisfies Theorem 1.2.5 then, as $n \to \infty$, the point process $P_n^* \to P^*$, where P^* is a non-homogeneous Poisson process on $\mathbb{R}_+^p \setminus \{\mathbf{0}\}$. The intensity measure for the limiting Poisson process P^* is given in terms of the pseudo-radial and -angular co-ordinates as*

$$\lambda(\mathrm{d}r \times \mathrm{d}\mathbf{w}) = \frac{\mathrm{d}r}{r^2}\; p \; \mathrm{d}H(\mathbf{w}). \tag{1.2.9}$$

Both this and the componentwise maxima result are discussed in greater detail in Chapter 5.

Both Theorems 1.2.5 and 1.2.6 suggest that we can characterise the distribution of asymptotically dependent multivariate extremes using the spectral measure $H(\cdot)$. There is an alternative and equivalent characterisation of this, known as the *Pickands dependence function*, which, for $\mathbf{t} \in S_p$, is defined as

$$A(\mathbf{t}) = \int_{S_p} \max\{w_j t_j\} p \; \mathrm{d}H(\mathbf{w}).$$

In order for the moment constraint (1.2.7) on H to be satisfied, the Pickands dependence function satisfies the constraint that $\max_{1\leq j\leq p}\{t_j\} \leq A(\mathbf{t}) \leq 1$.

If all pairs of the vector $\mathbf{t}$ are asymptotically independent random variables, the measure H degenerates to point mass on the end points of its support. In such cases neither H nor A is of any use in characterising the degree of extremal association. What to do in this case is still very much an active area of research, with recent work on conditional modelling by Heffernan and Tawn (2004) suggesting a plausible way forward, further details can be found on this in Section 1.3 below and in Chapter 5.

1.3 Statistical inference

The results in Section 1.2 have led to a variety of methods for making statistical inference on extremes. Beirlant *et al.* (2004) and Coles (2001) both provide details beyond the outline given here. First, some general points on inference for extremes; generally we focus on estimation of events even further into the tails of a distribution than we have already observed; for example from a 50-year series of data we might wish to estimate the level exceeded only once in the next 100 years. However, inference on extremes also suffers from a scarcity of data; by definition extremes are rare events, a fact which we must be aware of if we are to attempt extrapolation.

In this thesis, we focus on parametric inference, and in particular we shall mostly use the maximum likelihood approach. Following standard notation we use $\hat{\theta}$ to denote the maximum likelihood estimate (MLE) of the parameter θ. However we note that, since the rise in popularity of Markov Chain Monte Carlo (MCMC) methodology as a tool for sampling from (posterior) distributions, increasing amounts of research has been conducted into Bayesian inference on extremes, for example Coles and Powell (1996) and Coles and Tawn (1996). More recent publications have expanded the use of hierarchical modelling to model spatial extremes, see Cooley *et al.* (2006) and Craigmile *et al.* (2005). We do use some Bayesian methodology in Chapter 4. Aside from this, other recent research has looked away from parametric inference to non-parametric (Hall and Tajvidi, 2000) and semi-parametric inference (Davison and Ramesh, 2000, and Pauli and Coles, 2001) for extremes.

1.3.1 Univariate case

There are three main approaches to modelling the extremes of a data set, if the data are assumed to be IID. In the first instance, if only block maxima (for example annual maxima) are available, then following Theorem 1.2.1, it is usual to fit the GEVD defined in equation (1.2.1). If the entire data set is available this block maxima approach is seen to be wasteful of the data, and since extremes are rare we wish to include as much data as we can in the analysis. Instead one could either fit a point process model, following Theorem 1.2.2, or take a threshold exceedances approach, following Theorem 1.2.3. It is the latter that we focus on here, as it provides the foundation for both Chapters 2 and 3.

The threshold exceedances approach was popularised by Davison and Smith (1990). Suppose we have observed data $\mathbf{y} = (y_1, \ldots, y_n)$. A threshold u is first selected; various diagnostics, such as mean residual life plots, are available for doing this in the IID case, see Coles (2001) for details. Threshold selection amounts to a bias-variance tradeoff; if u is chosen too low, the parameter estimates are biased since the asymptotic result of Theorem 1.2.3 fails to hold, but if u is chosen too high, there are too little data and so there is huge uncertainty in the inference. The rate and size of the exceedances $E_u = \{y_i : y_i > u\}$ of this threshold are then modelled, using the GPD as a model for the sizes. The rate parameter ϕ_u, interpreted as the probability of an arbitrary observation exceeding the threshold, *i.e.* $\phi_u = \Pr[Y > u]$, has MLE given by the observed proportion of exceedances,

$$\hat{\phi}_u = \frac{|E_u|}{n},$$

whereas the MLE's of the GPD parameters have no closed form and numerical optimisation is required to maximise the likelihood

$$L(\sigma_u, \xi; \mathbf{y}) = \prod_{j: y_j \in E_u} \sigma_u^{-1} \left[1 + \xi \left(\frac{y_j - u}{\sigma_u} \right) \right]_+^{-1/\xi - 1}.$$

Inference in the IID case is straightforward; what is of more interest is how to carry out inference when either, or both, of the IID assumptions cannot be satisfied. A brief review is given here and further details are given in Chapter 2, in the case of dependent data, and in Chapter 3 in the case of non-identically distributed data. The threshold exceedances method can be extended to incorporate failures in either assumption.

Under the assumption that the data are stationary, standard practise is to model only the local maxima of the threshold exceedances. The initial step is to select the threshold u; we try to select the lowest possible threshold above which asymptotic marginal and dependence properties of the data appear stable. To assess the stability of the asymptotic marginal properties, we consider mean residual life plots or plots of the GPD parameters across a range of thresholds, for details see Coles (2001). To assess the stability of asymptotic dependence properties, we estimate the extremal index θ introduced in Theorem 1.2.4 for a range of thresholds and search for the lowest threshold above which the estimates for θ are constant, under the assumption that at this level the estimated extremal index has attained its asymptotic value. The modelling procedure then has two further steps; first *decluster* the exceedances of the threshold u to extract the independent local (cluster) maxima and then estimate the rate and GPD parameters for these cluster maxima using the methods described above for IID data.

There are many declustering schemes; for an overview see Chapter 10 in Beirlant *et al.* (2004) and recent proposals can be found in Ferro and Segers (2003) and Laurini and Tawn (2003). The simplest declustering schemes are the runs and blocks methods. In the former clusters are defined as being separated by $m - 1$ consecutive non-exceedances for some pre-determined run length m; Smith (1989) uses runs declustering to analyse extreme levels of ground-level ozone. In the blocks method, the data are split into consecutive blocks of pre-determined length r and any exceedances in the same block are said to belong to the same cluster.

Under the assumption that the data are non-stationary most work follows Davison and Smith (1990), who propose combining the approach for stationary data with regression modelling so that functions of the parameters of the extremes model are modelled as linear functions of covariates $\mathbf{x}$. This approach essentially comes down to maximising two likelihoods. The first, to estimate the rate parameter $\phi_u(\mathbf{x})$, takes the straightforward form of the likelihood for Bernoulli random variables. Let C_u denote the set of cluster maxima associated with the threshold u, then the second, to estimate the GPD parameters, is as follows,

$$L(\boldsymbol{\sigma}_u, \boldsymbol{\xi}; \mathbf{y}) = \prod_{i:y_i \in C_u} \sigma_u(\mathbf{x}_i)^{-1} \left[1 + \xi(\mathbf{x}_i) \left(\frac{y_i - u}{\sigma_u(\mathbf{x}_i)} \right) \right]_+^{-1/\xi(\mathbf{x}_i)-1}$$

where the parameter coefficients $\boldsymbol{\phi}_u$, $\boldsymbol{\sigma}_u$, $\boldsymbol{\xi}$ are such that

$$\log \frac{\phi_u(\mathbf{x})}{1-\phi_u(\mathbf{x})} = \boldsymbol{\phi}_u' \mathbf{x}, \quad \log \sigma_u(\mathbf{x}) = \boldsymbol{\sigma}_u' \mathbf{x}, \quad \xi(\mathbf{x}) = \boldsymbol{\xi}' \mathbf{x}.$$

The logit and log link functions are required to ensure that the rate and scale parameters lie in the correct parameters spaces; that is $0 < \phi_u(\mathbf{x}) < 1$ and $\sigma_u(\mathbf{x}) > 0$. Recent work by Chavez-Demoulin and Davison (2005) looks at using generalised additive models rather than generalised linear ones.

1.3.2 Multivariate case

How best to make statistical inference on multivariate extremes remains a subject very much open to debate, in the IID case as much as when one or both of these assumptions are violated. Initial work focused on inference for componentwise maxima; following the asymptotic result from Theorem 1.2.5, one tries to fit the MVEVD defined in equation (1.2.6) to an observed sequence of componentwise maxima. However, this requires us to estimate the spectral measure H. Many suggestions have been made about how to do this, both non-parametric (see the review by Abdous and Ghoudi, 2005) and parametric (for example, Coles and Tawn, 1991); we concentrate on the latter here. Since no finite parametric family exists for H, we must specify some flexible model for H. The simplest parametric model is the logistic, which is symmetric, and is defined in the bivariate case as

$$h(w) = \frac{1-\alpha}{\alpha}\left[w(1-w)\right]^{1/\alpha-2}\left[(1-w)^{1/\alpha}+w^{1/\alpha}\right]^{\alpha-2}, \quad 0 \leq w \leq 1, \quad 0 < \alpha \leq 1,$$

where $h(w) = \mathrm{d}H/\mathrm{d}w$ is the density associated with $H(w)$. Asymptotic independence occurs when $\alpha = 1$ and asymptotic dependence increases as $\alpha \to 0$. Further examples of the spectral measure are given in Chapter 5.

However, as in the univariate case, use of the componentwise maxima only is wasteful if a full data set is available. Further, the analysis of componentwise maxima does not always make practical sense, since there is no reason that the maxima of the different components should occur in the same observation. Coles and Tawn (1994) suggest a method based on the limiting Poisson process characterisation for multivariate extremes given in Theorem 1.2.6. Given observed data $(\mathbf{y}_1, \ldots, \mathbf{y}_n)$, where $\mathbf{y}_i$ is a p-dimensional vector, first calculate the pseudo-radial and -angular co-ordinates as defined in equation (1.2.8). Define the extreme points by all those with large radial co-ordinates *i.e.* for large r our extremes are the set $E_r = \{\mathbf{y}_i : r_i > r,\ i = 1, \ldots, n\}$ and then estimate the spectral measure H using the angular co-ordinates associated with the observations in E_r.

Both of the above methods of inference work only for asymptotically dependent data, since, as discussed, the spectral measure degenerates when the data are asymptotically independent. This raises two questions; first, can we establish whether the data are asymptotically dependent before going to the effort of carrying out inference and, if they are not, how can we make inference in the asymptotically independent case?

The first of these questions, in the case of bivariate random variables, was tackled by Ledford and Tawn (1997) who propose the following model to characterise the level of asymptotic (in)dependence. Suppose that the random variables

(Y_1, Y_2) have unit Fréchet margins, then under the weak assumption that their joint survivor function is regularly varying, the model states that

$$\Pr(Y_1 > y, Y_2 > y) \sim \mathcal{L}(y) y^{-1/\eta}, \quad \eta \in (0, 1]. \tag{1.3.1}$$

Here the constant η is referred to as the *coefficient of tail dependence* and has the interpretation that if (Y_1, Y_2) are asymptotically dependent then $\eta = 1$, otherwise if (Y_1, Y_2) are asymptotically independent then, if they are positively associated $\eta > 0.5$, if they are near independent $\eta = 0.5$ and if they are negatively associated $\eta < 0.5$. Estimation of η is straightforward. Let $T = \min\{Y_1, Y_2\}$ then $\Pr(Y_1 > y, Y_2 > y) = \Pr(T > y)$. Now for large u, using the model in equation (1.3.1),

$$\begin{aligned} \Pr(T > u + y | T > u) &\sim \frac{\mathcal{L}(u+y)(u+y)^{-1/\eta}}{\mathcal{L}(u) u^{-1/\eta}} \\ &\sim \left(1 + \frac{y}{u}\right)^{-1/\eta} \end{aligned}$$

where the second approximation holds by the definition of the slowly varying function $\mathcal{L}(\cdot)$. Thus η is estimated as the shape parameter in the GPD model for the threshold exceedances of the series $T_i = \{\min \mathbf{Y}_i\}$. This gives a gauge of whether the data are asymptotically (in)dependent, and further the strength of any asymptotic independence; we shall use this in Chapter 2.

As mentioned at the end of Section 1.2, Heffernan and Tawn (2004) propose a method for modelling any multivariate extremes, whether asymptotically dependent or independent. They also first assume that the margins are fixed, in this case to a Gumbel distribution. Their idea is then to systematically condition on each component being extreme and model the distribution of the remaining components, *i.e.* for $i = 1, \ldots, p$, they model

$$\Pr(\mathbf{Z}_{|i} \leq \mathbf{z}_{|i} | Y_i = y_i) = G_{|i}(\mathbf{z}_{|i}), \quad y_i > u_{Y_i} \tag{1.3.2}$$

where, if $\mathbf{Y}_{-i}$ denotes the vector $\mathbf{Y}$ with the ith component removed, we define the residuals $\mathbf{Z}_{|i}$ as

$$\mathbf{Z}_{|i} = \frac{\mathbf{Y}_{-i} - a_{|i}(y_i)}{b_{|i}(y_i)} \tag{1.3.3}$$

and $a_{|i}(\cdot)$ and $b_{|i}(\cdot)$ are normalising functions defined by

$$\begin{aligned} a_{|i}(y) &= a_{|i}y + I[a_{|i} = 0, b_{|i} < 0]\{c_{|i} - d_{|i}\log(y)\} \\ b_{|i}(y) &= y^{b_{|i}} \end{aligned}$$

where the constants satisfy $0 \leq a_{j|i} \leq 1$, $-\infty < b_{j|i} < 1$, $-\infty < c_{j|i} < \infty$ and $0 \leq d_{j|i} \leq 1$.

For estimation purposes, Heffernan and Tawn (2004) suggest treating the distribution $G_{|i}(\cdot)$ as having mutually independent and Gaussian components. Estimation then becomes a regression problem. Suppose that the residuals $\mathbf{Z}_{|i}$ have two finite moments, $\boldsymbol{\mu}_{|i}$ and $\boldsymbol{\sigma}_{|i}$, then the mean $\mu_{|i}(y_i)$ and standard deviation $\sigma_{|i}(y_i)$ of $\mathbf{Y}_{-i}|Y_i = y_i$ can be found using equation (1.3.3). The parameters are estimated by maximising the following objective function, for $i = 1, \ldots, p$,

$$Q_{|i}(\mathbf{a}_{|i}, \mathbf{b}_{|i}, \mathbf{c}_{|i}, \mathbf{d}_{|i}, \boldsymbol{\mu}_{|i}, \boldsymbol{\sigma}_{|i}) = -\sum_{j \neq i} \sum_{k=1}^{n_{u_{Y_i}}} \left[\log\{\sigma_{j|i}(y_{i|i,k})\} + \frac{1}{2} \left\{ \frac{y_{j|i,k} - \mu_{j|i}(y_{i|i,k})}{\sigma_{j|i}(y_{i|i,k})} \right\}^2 \right]$$

where $n_{u_{Y_i}}$ is the number of exceedances of the threshold u_i by the ith component and $y_{j|i,k}$ denotes the kth observation associated with an exceedance of the threshold u_{Y_i} by the ith component y_i of the jth component of the vector $\mathbf{y}$. Making the further assumption of independence between the conditional distributions allows simultaneous estimation of all the conditionals.

This approach has the advantage that it can be used regardless of the asymptotic dependence structure of the components. Further it does not become difficult to fit in higher dimensions as many multivariate extremes methods do. We explore further the performance of this approach under the assumption of asymptotic dependence in Chapter 5.

1.4 Air pollution data

The statistical analysis of extreme values is used in numerous applications; examples include hydrology, meteorology, engineering, finance, economics, reinsurance and telecommunications. The environmental sciences offer a rich variety of extreme value problems; for example predicting unusually strong winds, high waves and excessive river levels. In this thesis we focus on the estimation of extremely high concentrations of certain air pollutants.

Chapters 2, 3 and 4 involve the analysis of the extreme values of a sequence of surface-level ozone data. There are two data sets involved, both were observed in urban areas in the UK; the data in Chapter 2 comes from Swansea and that in Chapter 3 from Reading. The data were produced as part of the UK governments Air Quality Monitoring Network and can be freely downloaded from the internet, see Chapter 2. We also have NO (nitric oxide) and NO_2 (nitrogen dioxide) data available at the two sites; collectively these two chemicals are referred to as NO_X. All the data are in the form of daily maxima of hourly readings. Plots of the data sets are shown in the relevant chapters.

The analysis of extreme air pollution is important for several reasons; primarily because large concentrations of a contaminant generally have worse effects than smaller concentrations. In the case of ozone these detrimental effects include causing damage to human health (Huang *et al.*, 2005) and loss of crops and forests. Since increased concentrations lead to dire consequences, statistical analysis can also be used to look for trends and patterns in concentration levels; for example, are concentration levels increasing and are they particularly high under certain meteorological conditions, or in the presence of high concentrations of other contaminants?

Much work has looked at the statistical analysis of surface-level ozone. An excellent review paper is by Thompson *et al.* (2001). In the extremes literature, Küchenhoff and Thamerus (1996) use GEVD and GPD models to model extreme ozone and NO_2 levels using IID models; they also consider using a logistic regression model for the rate of threshold exceedances, whereas Smith (1989) uses a point process model to look for trends. Coles and Pan (1996) use the threshold exceedances method to analyse extreme values of NO_2, incorporating as much structural information through covariates as possible. Heffernan and Tawn (2004) apply their conditional multivariate model to a data set of five pollutants, including ozone and NO_X. We offer what we believe to be an improvement over these methods in Chapters 3 and 4 by using our proposed methodology to better take into account the structure of the ozone data and its relationships with both NO_X and various meteorological variables.

References

Abdous, B. and Ghoudi, K. (2005) Nonparametric estimators of multivariate extreme dependence functions, *J. Nonparametric Statist.*, **17:8**, 915-935.

Beirlant, J., Goegebeur, Y., Segers, J. and Teugals, J. (2004) *Statistics of extremes: Theory and Applications*, John Wiley and Sons Ltd, Chichester, UK.

Chavez-Demoulin, V. and Davison, A.C. (2005) Generalized additive modelling of sample extremes. *Applied Statistics*, **54**:1, 207-222.

Coles, S. (2001) *An introduction to statistical modelling of extreme values*, Springer, London.

Coles, S. and Powell, E.A. (1996) Bayesian Methods in Extreme Value Modelling: A Review and New Developments.*International Statistical Review*, **64**:1, 1 19-136.

Coles, S. and Tawn, J.A. (1991) Modelling Extreme Multivariate Events. *Journal of the Royal Statistical Society B*, **53**:2, 377-392.

Coles, S. and Tawn, J.A. (1994) Statistical Methods for Multivariate Extremes: an Application to Structural Design. *Applied Statistics*, **43**:1, 1-48.

Cooley, D., Naveau, P. and Jomelli, V. (2006) A Bayesian Hierarchical Extreme Value Model for Lichenometry. *Environmetrics*, **17**, 555-574.

Craigmile, P.F., Cressie, N., Santner, T.J. and Rao, Y. (2005) A Loss function approach to identifying environmental extremes. *Extremes* **8**:3, 143-159.

Davison, A.C. and Ramesh. N.I. (2000) Local likelihood smoothing of sample extremes. *Journal of the Royal Statistical Society B*, **62**:1, 191-208.

Davison, A.C. and Smith, R.L. (1990) Models for Exceedances over High Thresholds. *J. Roy. Statist. Soc. B.* **62**, 191-208.

Ferro, C.A.T. and Segeres, J. (2003) Inference for clusters of extreme values. *Journal of the Royal Statistical Society B* **65**:2, 545-556.

Hall, P. and Tajvidi, N. Nonparametric Analysis of Temporal Trends When Fitting Parametric Models to Extreme-Value Data. *Statistical Science*, **15**:2, 153-167.

Heffernan, J.E. and Tawn, J.A. (2004) A conditional approach for multivariate extremes. *Journal of the Royal Statistical Society B*, **66**-3, 497-546.

Huang, Y., Dominici, F. and Bell, B.L. (2005) Bayesian hierarchical distributed lag models for summer ozone exposure and cardio-respiratory mortality. *Environmetrics*, **16**: 547-562.

Küchenhoff, H. and Thamerus, M. Extreme value analysis of Munich air pollution data. *Environmental and Ecological Statistics*, **3**, 127-141.

Laurini, F. and Tawn, J.A. (2003) New estimators for the Extremal Index and Other Cluster Characteristics. *Extremes*, **6**, 189-211.

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

Ledford, A.T. and Tawn, J.A. (1997) Modelling dependence within joint tail regions. *Journal of the Royal Statistical Society B*, **59**, 475-499.

Pauli, F. and Coles, S. (2001) Penalized likelihood inference in extreme value analyses. *Journal of Applied Statistics* **28**:5, 547-560

Resnick, S. I. (1987) *Extreme values, regular variation, and point processes*, Springer-Verlag, New York.

Smith, R.L. (1989) Extreme Value Analysis of Environmental Time Series: An Application to Trend Detection in Ground-Level Ozone. *Statistical Science* **4**:4, 367-393.

Thompson, M.L., Reynolds, J., Cov, L.H., Guttorp, P. and Sampson, P.D. (2001) A review of statistical methods for the meteorological adjustment of tropospheric ozone. *Atmospheric Environment*, **35**, 617-630.

Chapter 2

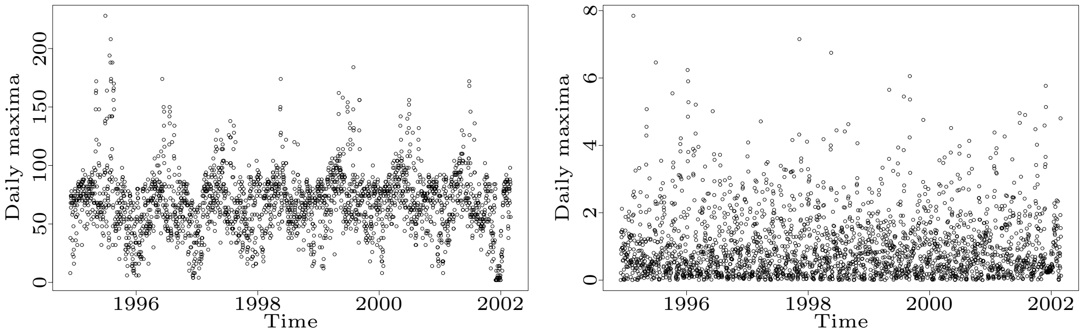
The distribution for the cluster maxima of exceedances of sub-asymptotic thresholds.

2.1 Introduction

This paper considers the analysis of the largest daily maxima values of surface-level ozone (O_3) that exceed a high threshold. Due to short-range temporal dependence between the daily maxima we are actually interested in modelling the local maxima of these threshold exceedances, since these can be considered as a sample of independent random variables with identical distributions. The data are from an urban station (Swansea, South Wales, UK), part of an air quality automatic monitoring network, run on behalf of the UK Government's Department for Environment, Food and Rural Affairs (DEFRA). The data can be downloaded from the website

`http://www.airquality.co.uk/archive/data_and_statistics.php`

The motivation for modelling large values of air pollution data sets is that these levels cause most concern when considering the various impacts of anthropogenic air pollution on human and animal health, materials, crops and forests and biological diversity. Further, air pollution control standards are mostly specified in terms of exceedances of high thresholds (Colls, 2002).



(a) (b)

Figure 2.1: Daily maxima data of O_3 shown on (a) the original non-stationary scale and (b) after local, nonparametric transformation to a stationary scale. The marginal distribution for the transformed data is standard exponential.

The study of extreme values in air pollution time series raises interesting modelling issues, since both the original series and the extremes show short-range temporal dependence and non-stationarity. This paper focuses on dealing with the extremal short-range dependence, which results in *clustering* of the extreme values. However, we must first deal with the non-stationarity in the extremes. Several ways to do this have been considered, for example by Smith (1989), Küchenhoff and Thamerus (1996), Hall and Tajvidi (2000), Ramesh and Davison (2002) and in Chapter 3 of this thesis. We remove the nonstationarity, using a local nonparametric transformation (see Section 2.6), the result of which is shown in Figure 2.1 for the ozone data which is presented on the original scale and again following transformation to a stationary series with standard exponential margins. The reason for this marginal choice is explained later in this section.

We assume that $\{X_i\}$ is a stationary series with marginal distribution function F, having upper endpoint x_+ (so that $F(x) \to 1$ as $x \to x_+$). Under stationarity, a standard approach to modelling the extremes of this series is the *peaks over threshold* (POT) method (Davison and Smith, 1990), an approach previously used on air pollution data by, for example, Smith (1989) and Küchenhoff and Thamerus (1996). This method defines the extremes to be all exceedances of a high threshold. The method then has three steps. First an appropriate high threshold u is selected. The threshold exceedances are then declustered to identify independent clusters of exceedances. Finally, a *generalised Pareto* (GP) distribution is fitted to the cluster maxima data, where if Y is a $\mathrm{GP}(\sigma_u, \xi)$ random variable then, for $v > 0$, it has the conditional survivor function

$$\bar{W}_u(v) = \Pr\left[Y > u + v | Y > u\right] = \left[1 + \frac{\xi v}{\sigma_u}\right]_+^{-1/\xi} \tag{2.1.1}$$

where $z_+ = \max(0, z)$, σ_u ($\sigma_u > 0$) is a scale parameter and ξ is a shape parameter. An alternative, but more complicated modelling approach, is to model both the distribution of all the exceedances and the dependence structure of the clusters; however we do not discuss this approach further here.

The justification for the GP distribution to model cluster maxima relies on asymptotic approximations concerning both the marginal tail behaviour and the dependence structure of exceedances of u by the series $\{X_i\}$ as $u \to x_+$. For independent and identically distributed (IID) random variables the cluster maxima are simply arbitrary exceedances of u so the POT method is strongly supported by the asymptotic theory of Pickands (1971, 1975) concerning the marginal tail behaviour. Here we assess only the validity of the POT method asymptotic approximations for the dependence structure and so remove issues about the marginal convergence by selecting the marginal distribution of $\{X_i\}$ so that the arbitrary exceedances of any threshold follow a GP distribution exactly. Specifically, we take $\{X_i\}$ to have a $\mathrm{GP}(\sigma_{u_1}, \xi)$ distribution above threshold u_1, so from the GP distribution threshold stability property, the arbitrary exceedances of $u_2 > u_1$ have $\mathrm{GP}(\sigma_{u_2}, \xi)$ distribution, where $\sigma_{u_2} = \sigma_{u_1} + \xi(u_2 - u_1)$, see Davison and Smith (1990). This strategy explains our choice to transform the data example to have a standard exponential marginal distribution, since this is equivalent to a $\mathrm{GP}(1, \xi)$ distribution, with $\xi \to 0$.

The second step in the POT method is to decluster the threshold exceedances into independent clusters. One way to define these clusters is the *runs method* in which a run length m is defined so that any exceedances of the threshold u separated by at least $m-1$ consecutive non-exceedances are considered independent (Smith and Weissman, 1994). Thus clusters are groups of extreme values in which any two consecutive cluster members are separated by, at most, $m-2$ non-exceedances and separate clusters are independent. The *extremal index*, θ, (Leadbetter *et al.*, 1983) is an asymptotic parameter measuring the strength of clustering of extreme values in the series. It is interpreted as the reciprocal of the mean limiting cluster size, as $u \to x_+$ and $m \to \infty$, hence $0 \leq \theta \leq 1$, where $\theta = 1$ if the extremes are independent. From the definitions of an extreme value and a cluster we estimate the extremal index as a function of both threshold u and run length m, thus $\hat{\theta} = \hat{\theta}(u, m)$. If $M_{k,j} = \max_{k \leq i \leq j}\{X_i\}$ then, following O'Brien (1987), we define

$$\theta(u, m) = \Pr(M_{2,m} < u | X_1 > u), \tag{2.1.2}$$

which we call the threshold-based (or *sub-asymptotic*) extremal index (Bortot and Tawn, 1998 and Ledford and Tawn, 2003).

Asymptotic theory also justifies the choice of the GP distribution for *cluster maxima*, see, for example, Smith (1989), Leadbetter (1991) and Smith, Tawn and Coles (1997). A critical feature of this derivation is that the asymptotic parameter θ is independent of the level at which the extremes are defined. We will show that the GP distribution model is the appropriate model choice for cluster maxima above a sub-asymptotic threshold u if and only if $\theta(x, m)$ exhibits stability over u, *i.e.* $\theta(x, m) = \theta(u, m)$ for all $x > u$. As Ledford and Tawn (2003) have identified broad classes of processes which have unstable $\theta(x, m)$ for any $x < x_+$ our findings suggest that there may be a better distribution than the GP for modelling cluster maxima. Specifically, we consider the case of a series with independent extremes, *i.e.* $\theta = 1$, but, for which, the sub-asymptotic extremal index is significantly less than one, *i.e.* $\theta(x, m) < 1$ for $x < x_+$ and is increasing with x. Any Gaussian process with correlation strictly less than 1 is an example of such a process.

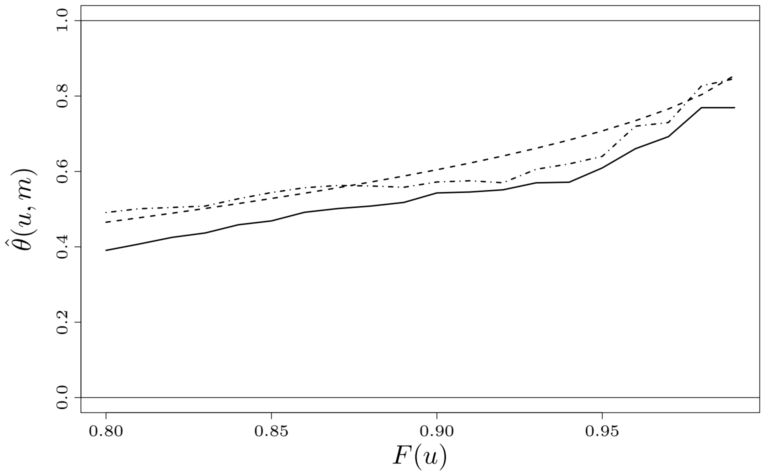


Figure 2.2: Estimate of the sub-asymptotic extremal index $\theta(u, m)$ for the ozone data (shown in Figure 2.1) using the runs estimator (full line), over a range of thresholds (80- to 99% quantiles) and with run length $m = 3$. Also shown are empirical (dash-dot line) and model-based (dashed line) estimates of $\theta(u, m)$, using the approximation of Ledford and Tawn (2003) given in equation (2.2.11). A 99% threshold was used to estimate the parameters in the model-based approach.

For the ozone data in Figure 2.1 we assess the need for an alternative distribution to the GP distribution for modelling cluster maxima by looking for stability in $\theta(x, m)$. The estimated sub-asymptotic extremal index for the ozone data is displayed in Figure 2.2 for a range of thresholds and a run length $m = 3$. The runs estimator for the extremal index (Smith and Weissman, 1994) used in this plot requires choice of both threshold and run length. More sophisticated techniques such as the intervals estimator (Ferro and Segers, 2003) incorporate automatic choice of one of these parameters into the estimation procedure. However all the available estimators give similar results. For our data, a run length of $m = 3$ gave most consistency between the runs estimator and the intervals estimator (not shown).

Figure 2.2 shows that as the threshold increases the estimate of the sub-asymptotic extremal index gets closer to 1. This suggests that although the threshold exceedances are clustered at sub-asymptotic levels, they are independent in the limit with clusters of size 1 (Ledford and Tawn, 2003). This lack of stability in the estimates of $\theta(u, m)$ across u within the range of the data, suggests that the GP distribution may not be appropriate to model cluster maxima. In this paper we will propose an alternative distribution for the cluster maxima of sub-asymptotic thresholds and consider when the GP distribution is a good approximate distribution for the cluster maxima of such thresholds.

In Section 2.2 we review extreme value theory for univariate stationary processes. The asymptotic theory discussed in Section 2.2 motivates our derivation, in Section 2.3, of the distribution for the cluster maxima of threshold exceedances when the sub-asymptotic extremal index does not stabilise and the process is asymptotically independent. In Section 2.4 we discuss how to make inference using the distribution derived in Section 2.3. We present a simulation study to compare our distribution and the usual asymptotically motivated GP distribution in Section 2.5. Section 2.6 compares the two models for the case of the Swansea air pollution data.

2.2 Background Results

In this section we introduce the block maxima and point process methods as alternative approaches for modelling extremes of a stationary series. We show how these methods are related, to each other and to the POT method introduced in Section 2.1. This is key to the model derived in Section 2.3.

Before considering the distribution of the maxima of the *stationary* series $\{X_i\}$, it is helpful to first look at the case of the associated independent series $\{\tilde{X}_i\}$. The independent series $\{\tilde{X}_i\}$ is chosen to have the same univariate marginal distribution F as the original series $\{X_i\}$. Let $\{a_n > 0\}$ and $\{b_n\}$ be sequences of constants and denote $\tilde{M}_{k,j} = \max_{k \leq i \leq j}\{\tilde{X}_i\}$ to be the analogue of $M_{j,k}$ for the associated independent series. Then there is well established asymptotic theory (Leadbetter *et al.*, 1983) for the limiting distributions of the normalised maxima, $a_n^{-1}(M_{1,n} - b_n)$ and $a_n^{-1}(\tilde{M}_{1,n} - b_n)$, of both series, as $n \to \infty$.

In the IID case the asymptotic theory states that, if the limiting distribution of the (normalised) maxima is non-degenerate, then it belongs to the *generalised extreme value* (GEV) class of distributions. That is, as $n \to \infty$,

$$\Pr\left(\frac{\tilde{M}_{1,n} - b_n}{a_n} \leq z\right) \to G(z) \tag{2.2.1}$$

with the limiting distribution function G being

$$G(z) = \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\psi}\right)\right]_+^{-1/\xi}\right\}, \tag{2.2.2}$$

where $[y]_+ = \max(y, 0)$ and the three parameters are location (μ), scale ($\psi > 0$) and shape (ξ). The shape parameter is negative if the underlying distribution has finite upper endpoint ($x_+ < \infty$), is zero if the tail of the underlying distribution decreases exponentially and is positive if the tail decreases polynomially. The result holds under a necessary and sufficient condition on the distribution function F (*eg.* Leadbetter *et al.*, 1983), which is satisfied for most distributions of interest.

We return to the original stationary series $\{X_i\}$, with maxima $M_{1,n}$. To obtain a result for the distribution of the $M_{1,n}$, in addition to the limit (2.2.1) holding, a weak mixing condition that limits the amount of long-range dependence in the extremes, denoted $D(u_n)$ by Leadbetter *et al* (1983), is assumed to hold. Under these assumptions then, as $n \to \infty$,

$$\Pr\left(\frac{M_{1,n} - b_n}{a_n} \leq z\right) \to G^{\theta}(z) \tag{2.2.3}$$

where θ is the extremal index of the stationary series $\{X_i\}$, see Leadbetter *et al.* (1983). When the series is independent $\theta = 1$, however it is also the case that $\theta = 1$ for a broad class of dependent processes, see Leadbetter *et al.* (1983) and Ledford and Tawn (2003).

Before giving the point process characterisation for extremes we need to identify clusters. We additionally assume that the short-range dependence structure of the series $\{X_i\}$ satisfies the $D^{(m)}(u_n)$ condition of Chernick *et al.* (1991), *i.e.* for the sequence $u_n = a_n u + b_n$ and fixed m, as $n \to \infty$,

$$\Pr\left(M_{2,m} < u_n, M_{m+1,s_n} > u_n | X_1 > u_n\right) \to 0$$

where $s_n = o(n)$. Under this condition the cluster ends once there have been $m-1$ consecutive non-exceedances. It does not however place any limit on cluster size. Let $\{t_k : X_{t_k} \geq u_n\}$ be the exceedance times of the threshold u_n. Given n, u_n and m, suppose that the series has r_n clusters identified under the $D^{(m)}(u_n)$ condition, then we define the sequence $\{q_j : t_{q_j+1} - t_{q_j} \geq m, j = 1, ..., r_n\}$ to be the cluster end point times so that we can partition the exceedance times $\{t_k\}$ into r_n clusters where the j^{th} cluster consists of exceedance times $C_j = \{t_{q_{j-1}+1}, t_{q_{j-1}+2}, ..., t_{q_j}\}$. Using this definition of the clusters, the 2-dimensional point process P_n consisting of the times (T_j) and sizes (Y_j) of the cluster maxima is defined to be

$$P_n = \left\{\left(\frac{T_j}{n+1}, \frac{Y_j - b_n}{a_n}\right) : Y_j = \max(X_i : i \in C_j),\ T_j = (i \in C_j : X_i = Y_j),\ 1 \leq j \leq r_n\right\} \tag{2.2.4}$$

Note that the time index (T_j) has been transformed to the unit interval and the threshold exceedances (Y_j) have been normalised using the same constants $\{a_n\}$ and $\{b_n\}$ as in equation (2.2.1).

Assuming that the $D(u_n)$ and $D^{(m)}(u_n)$ conditions hold and that the distribution of the maxima of the associated independent series $\{\tilde{X}_i\}$ converges to the non-degenerate distribution G of equation (2.2.1) then the point process P_n converges in distribution to a non-homogeneous Poisson process P on the region $A = (0, 1] \times [v, \infty]$, as $n \to \infty$, see Smith (1989), where $v > u$. The intensity density for P is then

$$\lambda(t, x) = \theta\psi^{-1}\left[1 + \xi\left(\frac{x - \mu}{\psi}\right)\right]_+^{-1/\xi - 1} \tag{2.2.5}$$

where (μ, ψ, ξ) are the parameters of the limiting distribution of $a_n^{-1}(\tilde{M}_{1,n} - b_n)$

and θ is the extremal index. The integrated intensity for the limiting process, on the region A, is

$$\begin{aligned}\Lambda(A) &= \int_v^\infty \int_0^1 \lambda(t,x)\ dt\ dx \\ &= \theta\left[1+\xi\left(\frac{v-\mu}{\psi}\right)\right]_+^{-1/\xi}. \qquad (2.2.6)\end{aligned}$$

The maxima result (2.2.3) may be derived from the point process convergence result. Let $N_n(A)$ be the number of points of the process P_n on the region A and let $N(A)$ be the equivalent number for the limiting Poisson process P. Then, taking limits as $n \to \infty$,

$$\begin{aligned}\Pr\left(\frac{M_{1,n}-b_n}{a_n} \leq v\right) &= \Pr(N_n(A)=0) \\ &\to \Pr(N(A)=0) = \exp\{-\Lambda(A)\} \\ &= \exp\left\{-\theta\left[1+\xi\left(\frac{v-\mu}{\psi}\right)\right]_+^{-1/\xi}\right\}, \qquad (2.2.7)\end{aligned}$$

which recovers the block maxima results of equation (2.2.3).

We can also use the point process result to derive the limiting conditional distribution of the cluster maxima of the threshold exceedances. We shall drop the index notation and denote a generic cluster maxima by Y. Observe that as $n \to \infty$, for $v > 0$,

$$\begin{aligned}\Pr\left(\frac{Y-b_n}{a_n} > u+v \middle| \frac{Y-b_n}{a_n} > u\right) &\to \frac{\int_{u+v}^\infty \int_0^1 \lambda(t,x) dt\ dx}{\int_u^\infty \int_0^1 \lambda(x) dt\ dx} \\ &= \frac{\theta\left[1+\xi\left(\frac{u+v-\mu}{\psi}\right)\right]_+^{-1/\xi}}{\theta\left[1+\xi\left(\frac{u-\mu}{\psi}\right)\right]_+^{-1/\xi}} \\ &= \left[1+\frac{\xi v}{\sigma_u}\right]_+^{-1/\xi} \qquad (2.2.8)\end{aligned}$$

where $\sigma_u = \psi + \xi(u-\mu)$. Note that the extremal index cancels. Result (2.2.8) shows that the limiting distribution for the cluster maxima is the GP distribution of equation (2.1.1). This derivation also relates the GEV and GP parameters. If the block maxima converge to a GEV(μ, ψ, ξ) distribution and the cluster maxima of the threshold exceedances converge to a GP(σ_u, ξ) distribution, the shape parameters (ξ) are identical and the scale parameters are linked by the above expression. Thus, the GP scale parameter is threshold dependent, but the shape is threshold invariant. The asymptotic result of equation (2.2.8) gives us the limiting conditional distribution of the cluster maxima of threshold exceedances, which motivates the POT method introduced in Section 2.1. The justification is that for a high enough threshold the distribution of the cluster maxima above this threshold is approximately the limiting distribution of equation (2.2.8).

Finally we define asymptotic (in)dependence and introduce the characterisation of the sub-asymptotic extremal index given by Ledford and Tawn (2003). The extremal dependence structure of the random variables (X_0, X_ω), at lag ω, assumed to have identical margins, is defined as follows. Taking $x \to x_+$ the probability of one variable being extreme conditional on the other also being extreme is

$$\Pr(X_\omega > x | X_0 > x) \to \begin{cases} 0 & \text{if asymptotically independent} \\ t_\omega > 0 & \text{if asymptotically dependent.} \end{cases} \tag{2.2.9}$$

If the series is independent for lags of at least m and the series is asymptotically independent for all lags ω up to $m-1$, then the extremal index is 1. If the series is asymptotically dependent for any lag up to $m-1$ then the extremal index is less than 1, see Ledford and Tawn (2003).

The model for the sub-asymptotic threshold-based extremal index given by Ledford and Tawn (2003) follows from work by Ledford and Tawn (1996) and Coles *et al.* (1999). Under this model the joint survivor function at lag ω is

$$\Pr(X_0 > x, X_\omega > x) = \mathcal{L}_\omega\left(\frac{1}{\bar{F}(x)}\right)\bar{F}(x)^{2/(1+\bar{\chi}_\omega)} \tag{2.2.10}$$

where $\bar{F}$ is the marginal survivor function, $\mathcal{L}_\omega(\cdot)$ is a slowly varying function at infinity and $\bar{\chi}_\omega$ ($-1 < \bar{\chi}_\omega \leq 1$) is a measure of tail dependence (Coles *et al.*, 1999). If the variables are asymptotically dependent then $\bar{\chi}_\omega = 1$ else they are asymptotically independent. If the variables are positively associated in the extremes then $0 < \bar{\chi}_\omega < 1$, if they are independent $\bar{\chi}_\omega = 0$ and if they are negatively associated in the extremes $-1 < \bar{\chi}_\omega < 0$. Note that this is a minor adaptation from that given by Ledford and Tawn (2003) to allow for a general marginal distribution.

Ledford and Tawn (2003) use the tail dependence structure of equation (2.2.10) to examine short-range sub-asymptotic tail dependence (clustering) in asymptotically independent time series, giving a model for the sub-asymptotic extremal index, $\theta(u, m)$. They model the tail dependence structure of the time series at a range of lags using the distribution given in equation (2.2.10). Since the aim is to model clustering the only informative lags are $\omega = 1, \ldots, (m-1)$, as consecutive threshold exceedances separated by at least lag m are assumed to belong to separate clusters, and therefore are independent. Let $\bar{\chi}^{(m)} = \max\{\bar{\chi}_\omega : \omega = 1, ..., m-1\}$, then there are two cases of interest, $\bar{\chi}^{(m)} < 1$ and $\bar{\chi}^{(m)} = 1$, which we will consider in turn.

When $\bar{\chi}^{(m)} < 1$ the process is asymptotically independent at all lags so $\theta = 1$ and Ledford and Tawn (2003) use the definition of the extremal index given in equation (2.1.2), to give the asymptotic form of the sub-asymptotic extremal index as $u \longrightarrow x_+$,

$$\begin{aligned} 1 - \theta(u, m) &= 1 - \Pr[M_{1,m-1} < u | X_0 > u] \\ &\sim \mathcal{L}^{(m)}\left(\frac{1}{\bar{F}(u)}\right) \bar{F}(u)^{\zeta^{(m)}} \end{aligned} \qquad (2.2.11)$$

where $\zeta^{(m)} = (1 - \bar{\chi}^{(m)})/(1 + \bar{\chi}^{(m)})$, $\mathcal{L}^{(m)}(x)$ is a slowly varying function defined by $\mathcal{L}^{(m)}(x) = \sum_{\omega \in \omega^{(m)}} \mathcal{L}_\omega(x)$ and $\omega^{(m)} = \{\omega \in (1, \ldots, m-1) : \bar{\chi}_\omega = \bar{\chi}^{(m)}\}$. Note that $\omega^{(m)}$ is the set of lags at which the strongest form of extremal dependence occurs. If $\omega^{(m)}$ consists of one element only, or all $m-1$ elements, then the asymptotic approximation (2.2.11) provides an upper bound, lower bound respectively, on the estimate of $\theta(u,m)$ for all u.

As expression (2.2.11) involves both marginal and dependence features it is helpful to gain some understanding of how $\theta(u,m) \to 1$ as $u \to x_+$. Consider the ratio $\{1-\theta(u+v,m)\}/\{1-\theta(u,m)\}$ as $u \to x_+$. Since $u \to x_+$, we cannot keep $v > 0$ constant and so scale the excess v accordingly. Specifically let $u = b_n$ and $v = a_n x$, where $a_n > 0$ and b_n are the normalising constants defined in the limiting relationship (2.2.1), then as $n \to \infty$,

$$\frac{1-\theta(a_n x + b_n, m)}{1-\theta(b_n,m)} \sim \frac{\mathcal{L}^{(m)}\left(\frac{1}{\bar{F}(a_n x + b_n)}\right)\bar{F}(a_n x + b_n)^{\zeta^{(m)}}}{\mathcal{L}^{(m)}\left(\frac{1}{\bar{F}(b_n)}\right)\bar{F}(b_n)^{\zeta^{(m)}}}.$$

Now, since b_n must be taken as the $(1-1/n)th$ quantile of F, as $n \to \infty$, we have both

$$\frac{\bar{F}(a_n x + b_n)}{\bar{F}(b_n)} \sim n[\bar{F}(a_n x + b_n)] \to -\log G(x)$$

where G is the GEV distribution function (2.2.2), and

$$\frac{\mathcal{L}^{(m)}\left(\frac{1}{\bar{F}(a_n x + b_n)}\right)}{\mathcal{L}^{(m)}\left(\frac{1}{\bar{F}(b_n)}\right)} \sim \frac{\mathcal{L}^{(m)}(-n/\log G(x))}{\mathcal{L}^{(m)}(n)} \to 1,$$

so that

$$\frac{1-\theta(a_n x + b_n, m)}{1-\theta(b_n,m)} \to \{-\log G(x)\}^{\zeta^{(m)}}$$

This ratio is a decreasing function in the unscaled excesses x. As $x \to 0$ the ratio tends to $\left[1-\frac{\xi\mu}{\psi}\right]_+^{-\zeta^{(m)}/\xi}$ and as $x \to x_+$ it tends to zero. If $\bar{\chi}^{(m)} = 0$ the ratio decreases at the same rate as $-\log G(x)$.

When $\bar{\chi}^{(m)} = 1$ the process is asymptotically dependent at some lag so $\theta < 1$ and although we do not have an asymptotic expansion for $\theta(u,m)$ in this case the asymptotic bounds for $\theta(u,m)$ as $u \to x_+$ discussed above still hold, i.e.

$$1 - \sum_{\omega \in \omega^{(m)}} \mathcal{L}_\omega\left(\frac{1}{\bar{F}(u)}\right) < \theta(u,m) < 1 - \max_{\omega \in \omega^{(m)}} \mathcal{L}_\omega\left(\frac{1}{\bar{F}(u)}\right).$$

So for $\bar{\chi}^{(m)} = 1$ we are only able to bound the sub-asymptotic extremal index using the pairwise dependence structure models of Ledford and Tawn (2003). If we were willing to make assumptions about higher order dependence then like Latham (2006) we could determine $\theta(u, m)$ more precisely as the higher order conditional probabilities in the expression for $1 - \theta(u, m)$ might be at least as big as the first order conditional probabilities used here.

In Figure 2.2 we show two estimates of the sub-asymptotic extremal index $\theta(u, m)$ made using the asymptotic characterisation (2.2.11). One estimate is made using empirical estimation of the relevant probabilities $\Pr(X_\omega > u | X_0 > u)$ and the other by estimating the parameters $\bar{\chi}^{(m)}$ and $\mathcal{L}^{(m)}(\cdot)$ at a 99% threshold and then 'plugging these estimates in' to equation (2.2.11) to obtain $\theta(u, m)$ as a function of the threshold u. Note that, following Ledford and Tawn (2003), the slowly varying function $\mathcal{L}^{(m)}(\cdot)$ is assumed to be constant in this estimate. The plot shows that, within the range of the data, the model and empirical estimates are very close. Both estimates coming from the asymptotic approximation are seen to be upper bounds for $\theta(u, m)$, as estimated using the runs method, with the bound being reasonably tight over all the values of u that are considered, and the bound getting closer to $\theta(u, m)$ for large u.

2.3 Derivation of the sub-asymptotic distribution

In this section we derive the distribution for the cluster maxima of threshold exceedances for an asymptotically independent series with $\theta(x, m) \neq 1$ for thresholds x within the range of the data. In what follows, motivated by the asymptotic conditions $D(u_n)$ and $D^{(m)}(u_n)$, we assume that above some (pre-selected) threshold level u exceedances separated by at least $m - 1$ consecutive non-exceedance values are independent. Choice of threshold u and run length m will depend on the particular process; in this section we assume these are known and address their choice in Section 2.5 and 2.6. Following the assumption that arbitrary exceedances of u follow a GP distribution, a model for the distribution of the sub-asymptotic block maxima is given by Hsing *et al.* (1996), Kratz and Rootzén (1997) and Bortot and Tawn (1998) to be

$$\Pr(M_{1,n} < z) = \{G(z)\}^{\theta(z,m)} = \exp\left\{-\theta(z,m)\left[1+\xi\left(\frac{z-\mu}{\psi}\right)\right]^{-1/\xi}\right\} \tag{2.3.1}$$

where G is the GEV(μ, ψ, ξ) distribution function. This form of distribution is appropriate due to the fast convergence of the maxima of independent GP variables to the GEV form relative to the convergence of the dependence structure, hence the extremal index in equation (2.2.3) is replaced by the sub-asymptotic extremal index in equation (2.3.1). Since we are no longer dealing with the limit distribution the normalising constants a_n and b_n of equation (2.2.3) are absorbed into the location (μ) and scale (ψ) parameters, therefore these parameters do not take the same values as their equivalents in equation (2.2.3). The shape parameter does remain the same.

By the assumption of distribution (2.3.1) for the unnormalised block maxima, it follows that we must also alter the point process P_n given in equation (2.2.4). Motivated by the asymptotic case described in Section 2.2, we assume that, given n, u and m, the series $\{X_i\}$ has r_n independent clusters separated by at least $m-1$ consecutive non-exceedances. The definitions of clusters (C_j) and the times (T_j) and sizes (Y_j) of cluster maxima therefore follow from Section 2.2. We denote by P_n^* the process P_n in which the sizes of the cluster maxima are not normalised so that

$$P_n^* = \left\{\left(\frac{T_j}{n+1}, Y_j\right) : Y_j = \max(X_i : i \in C_j),\ T_j = (i \in C_j : X_i = Y_j),\ 1 \le j \le r_n\right\}. \tag{2.3.2}$$

Following the limit result of equation (2.2.5), we assume that, as u approaches x_+, the point process P_n^* is well-approximated by a Poisson process on $A = [0,1] \times$

$[u, \infty)$, with intensity function given by

$$\lambda^*(t, x) = \theta(x, m)g(x), \quad \text{for } 0 < t < 1 \text{ and } x > u \tag{2.3.3}$$

where g is a function we derive subsequently. This intensity function is independent of time since the series $\{X_i\}$ is stationary and cluster maxima are assumed independent. The intensity function (2.3.3) is assumed to factorise in a similar way to the limiting form (2.2.5), *i.e.* with the first term relating to the short-range dependence structure and the second term to marginal features, but here the extremal index is replaced by the sub-asymptotic extremal index.

We can now obtain the function g. First we follow the steps taken in result (2.2.7) and link the block maxima and point process approaches to give, for $v > u$,

$$\Pr(M_{1,n} \leq v) = \exp\{-\Lambda^*(A)\} = \exp\left\{-\int_v^\infty \theta(x, m)g(x)\ dx\right\} \tag{2.3.4}$$

where $\Lambda^*(\cdot)$ is the integrated intensity of P_n^*. Combining this with the assumed distribution of the block maxima (2.3.1) and taking the logarithm of both sides, we obtain that for all v

$$\int_v^\infty \theta(x, m)g(x)\ dx = \theta(v, m)\left[1 + \xi\left(\frac{v - \mu}{\psi}\right)\right]_+^{-1/\xi}. \tag{2.3.5}$$

Then on differentiating equation (2.3.5) with respect to v we obtain the following expression for the function g in terms of the extreme value parameters (μ, ψ, ξ) and the sub-asymptotic extremal index,

$$g(v) = \frac{1}{\psi}\left[1 + \xi\left(\frac{v - \mu}{\psi}\right)\right]_+^{-1/\xi - 1} - \frac{\theta'(v, m)}{\theta(v, m)}\left[1 + \xi\left(\frac{v - \mu}{\psi}\right)\right]_+^{-1/\xi}, \quad \text{for } v > u \tag{2.3.6}$$

where $\theta'(v, m) = \partial\theta(v, m)/\partial v$.

Now, by comparison with relationship (2.2.8) which links the limiting point process and POT approaches, the conditional distribution for the cluster maxima is, for $v > 0$,

$$\Pr(Y > u + v | Y > u) = \frac{\int_{u+v}^{x+} \theta(x, m) g(x) \, \mathrm{d}x}{\int_{u}^{x+} \theta(x, m) g(x) \, \mathrm{d}x}. \tag{2.3.7}$$

From (2.3.5) we find that the conditional distribution for cluster maxima, in terms of the sub-asymptotic extremal index and the GP distribution parameters, is

$$\begin{aligned} \Pr(Y > u + v | Y > u) &= \frac{\theta(u+v, m) \left[1 + \xi \left(\frac{u+v-\mu}{\psi}\right)\right]_+^{-1/\xi}}{\theta(u, m) \left[1 + \xi \left(\frac{u-\mu}{\psi}\right)\right]_+^{-1/\xi}} \\ &= \frac{\theta(u+v, m)}{\theta(u, m)} \left[1 + \frac{\xi v}{\sigma_u}\right]_+^{-1/\xi} \end{aligned} \tag{2.3.8}$$

where $\sigma_u = \psi + \xi(u - \mu)$.

Next we use the sub-asymptotic extremal index model (2.2.11) of Ledford and Tawn (2003) to approximate the terms $\theta(x, m)$ and $\theta'(x, m)$ that are needed for evaluating expressions (2.3.6) and (2.3.8). Let $\kappa_u = \Pr(X > u)$ be the probability of a threshold exceedance and assume that the arbitrary threshold exceedances follow a $\mathrm{GP}(\sigma_u, \xi)$ distribution. The marginal probability that an arbitrary random variable from the original series exceeds $x > u$ is then

$$\bar{F}(x) = \Pr(X > x | X > u) \Pr(X > u) = \kappa_u \bar{W}_u(x - u) \tag{2.3.9}$$

where $\bar{W}_u$ is the survivor function for the GP distribution with parameters (σ_u, ξ). Finally, we follow Ledford and Tawn (2003) in modelling the slowly varying function $\mathcal{L}_\omega(\cdot)$ as a constant, so that $\mathcal{L}_\omega(x) = c_\omega$ for all x. Then combining results (2.2.11) and (2.3.9), we have

$$\frac{\theta(u+v, m)}{\theta(u, m)} \approx \frac{1 - c^{(m)} \kappa_u^{\zeta^{(m)}} \bar{W}_u(v)^{\zeta^{(m)}}}{1 - c^{(m)} \kappa_u^{\zeta^{(m)}}} \tag{2.3.10}$$

where $c^{(m)} = \mathcal{L}^{(m)}(x) = \sum_{\omega \in \omega^{(m)}} c_\omega$. Combining this with equation (2.3.8) therefore gives an expression for the conditional distribution of the cluster maxima.

An interesting special case is that of an IID process. In this case, as $u \to \infty$, $\bar{\chi}_\omega = 0$ and $c_\omega = 1$ for all lags $\omega = 1, \ldots, (m-1)$, so that $c^{(m)} = m - 1$. In fact, any method for selecting m should give $m = 1$, but we suppose that this is not the case and for some reason we are using $m > 1$. We show that we still get a correct approximation up to first order. Using result (2.3.10) in this case we can approximate the ratio of extremal indices $\theta(u+v, m)$ and $\theta(u, m)$ by

$$\begin{aligned}\frac{\theta(u+v,m)}{\theta(u,m)} &\approx \frac{1-(m-1)\kappa_u \bar{W}_u(v)}{1-(m-1)\kappa_u} \\ &= \left[1-(m-1)\kappa_u \bar{W}_u(v)\right]\left[1+(m-1)\kappa_u\right] + o(\kappa_u)\end{aligned}$$

since $\kappa_u \to 0$ as $u \to \infty$, allowing us to use a binomial series expansion on the denominator. Finally, combining this with equation (2.3.8) we get the following approximation for the distribution of the cluster maxima; for $v > 0$,

$$\Pr(Y > u+v | Y > u) \;=\; \bar{W}_u(v)\left[1+(m-1)\kappa_u\left(1-\bar{W}_u(v)\right)\right] + o(\kappa_u).$$

We shall now look in more generality at the distribution attained for the cluster maxima using expressions (2.3.8) and (2.3.10). Under asymptotic independence, by taking a binomial series expansion of the denominator in equation (2.3.10) we can further approximate the ratio of the extremal indices $\theta(u+v,m)$ and $\theta(u,m)$ as follows;

$$\begin{aligned}\frac{\theta(u+v,m)}{\theta(u,m)} &= \left[1-c^{(m)}\kappa_u^{\zeta^{(m)}}\bar{W}_u(v)^{\zeta^{(m)}}\right]\left[1+c^{(m)}\kappa_u^{\zeta^{(m)}}\right] + o(\kappa_u^{\zeta^{(m)}}) \\ &= 1 + c^{(m)}\kappa_u^{\zeta^{(m)}}\left[1-\bar{W}_u(v)^{\zeta^{(m)}}\right] + o(\kappa_u^{\zeta^{(m)}}) \qquad (2.3.11)\end{aligned}$$

where the approximations hold since the higher order terms are of order $o(\kappa_u^{\zeta^{(m)}})$ if $\bar{\chi}^{(m)} < 1$. Similarly, for $x > u$,

$$\theta'(x,m) = \frac{c^{(m)}\kappa_u^{\zeta^{(m)}}\zeta^{(m)}}{\sigma_u}\{\bar{W}_u(x-u)\}^{\zeta^{(m)}+\xi}.$$

Notice that when $\bar{\chi}^{(m)} = 1$, *i.e.* asymptotic dependence at some lag, then $\zeta^{(m)} = 0$ and hence $\theta'(x, m) = 0$, so we have stability of the sub-asymptotic extremal index. In this case the function g simplifies to a single term and intensity (2.3.3) is identical to the limiting intensity (2.2.6).

In theory the modelling of the cluster maxima can proceed using the point process approach with intensity function (2.3.3) or modelling the excesses over threshold through distribution (2.3.8). We focus on the latter to enable comparisons with the peaks over thresholds method. Combining equations (2.3.8) and (2.3.11), the survivor function for the cluster maxima above a threshold u takes the form, for $v > 0$,

$$\begin{aligned}\bar{W}_u^*(v) = \Pr(Y > u + v | Y > u) \approx \\ \bar{W}_u(v) \left\{1 + c^{(m)} \kappa_u^{(1-\bar{\chi}^{(m)})/(1+\bar{\chi}^{(m)})} \left(1 - \bar{W}_u(v)^{(1-\bar{\chi}^{(m)})/(1+\bar{\chi}^{(m)})}\right)\right\}.\end{aligned} \quad (2.3.12)$$

Distribution (2.3.12) consists of two terms: the first term, $\bar{W}_u(v)$, is the generalised Pareto distribution of the peaks over thresholds method, and the second term accounts for instability of $\theta(x, m)$ for $x > u$. Specifically, observe that in the case of asymptotic dependence, where $\bar{\chi}^{(m)} = 1$, the distribution (2.3.12) immediately collapses to the GP distribution for all thresholds (u).

Distribution (2.3.12) offers an extension to the generalised Pareto distribution for modelling the cluster maxima of sub-asymptotic threshold exceedances. To understand how this distribution differs from the generalised Pareto distribution, consider how its second term depends on the threshold u through the marginal probability κ_u of exceeding u and also through $\bar{W}_u(v)$. As $u \to x_+$, $\kappa_u \to 0$ so it appears that the second term disappears as u gets larger. However we need to be careful as we cannot increase u and keep v constant, so we must scale the excess v accordingly. Specifically if $\bar{\chi}^{(m)} < 1$, as $u \to x_+$

$$\begin{aligned}\bar{W}_u^*(\sigma_u v) &= \bar{W}_u(\sigma_u v) \left\{1 + c^{(m)} \kappa_u^{(1-\bar{\chi}^{(m)})/(1+\bar{\chi}^{(m)})} \left(1 - \bar{W}_u(\sigma_u v)^{(1-\bar{\chi}^{(m)})/(1+\bar{\chi}^{(m)})}\right)\right\} \\ &= (1 + \xi v)_+^{-1/\xi} + O(\kappa_u^{(1-\bar{\chi}^{(m)})/(1+\bar{\chi}^{(m)})}) \qquad (2.3.13)\end{aligned}$$

which shows that the second term acts as a bias correction term for the GP distribution and it disappears as the threshold is increased to the upper endpoint.

Finally we need to ensure that distribution (2.3.12) is a proper by forcing the density to be non-negative, which places an upper bound on the choice of threshold for some combinations of the dependence parameters $(c^{(m)}, \bar{\chi}^{(m)})$. This bound takes the form

$$0 < \kappa_u < \left(\frac{1 + \bar{\chi}^{(m)}}{c^{(m)}(1 - \bar{\chi}^{(m)})} \right)^{(1+\bar{\chi}^{(m)})/(1-\bar{\chi}^{(m)})} \leq 1. \tag{2.3.14}$$

Note that in practise it might not be appropriate to take the first order term in the binomial series expansion of the expression for the ratio of $\theta(u+v, m)$ and $\theta(u, m)$ as described in equation (2.3.11). This will be the case if, for instance, we believe that $c^{(m)}\kappa_u^{\zeta^{(m)}}$ is not converging to zero sufficiently quickly. In this case, in an analogue to equation (2.3.12), we can approximate the distribution of the cluster maxima as follows; for $v > 0$,

$$\bar{W}_u^1(v) \approx \bar{W}_u(v) \left[\frac{1 - c^{(m)}\kappa_u^{(1-\bar{\chi}^{(m)})/(1+\bar{\chi}^{(m)})} \bar{W}_u(v)^{(1-\bar{\chi}^{(m)})/(1+\bar{\chi}^{(m)})}}{1 - c^{(m)}\kappa_u^{(1-\bar{\chi}^{(m)})/(1+\bar{\chi}^{(m)})}} \right]. \tag{2.3.15}$$

2.4 Inference for the sub-asymptotic model

We now describe how to fit either form of our proposed distribution to a data set by using likelihood inference to estimate the parameters $(c^{(m)}, \bar{\chi}^{(m)}, \sigma_u, \xi)$. Since the overlap between the information used to estimate the dependence parameters $(c^{(m)}, \bar{\chi}^{(m)})$ and that used to estimate the GP parameters (σ_u, ξ) is small, we suggest a two stage fitting procedure for parameter estimation.

First a lag m is selected so that all threshold exceedances that are separated by at least $m - 1$ consecutive non-exceedances are assumed to be independent. One diagnostic for this selection is to plot the graph of $\bar{\chi}_\tau$ against τ for a much larger range of τ than that at which you would expect to see extremal dependence. The lag m is then selected so that $\bar{\chi}_\tau$ is approximately zero for $\tau \geq m$ and is greater than zero for $\tau < m$ (see Figure 2.7).

Given the lag m, the first step then comprises of estimating the dependence parameters $(c^{(m)}, \bar{\chi}^{(m)})$, using the full data set and the censored pseudolikelihood approach of Ledford and Tawn (2003) as follows. First transform the observations to unit Fréchet margins using the empirical distribution function and probability integral transform. We denote this transformed series by $X_1^*, \ldots, X_n^*$ and the series of the minima of all pairs of the transformed variables at lag $\tau = 1, ..., m-1$ by $\mathbf{T}^\tau = \{\min(X_i^*, X_{i+\tau}^*) : i = 1, ..., n-\tau\}$. For each lag τ, this series has survivor function given in equation (2.2.10) and the maximum likelihood estimates (MLEs) of the dependence parameters $(\hat{c}_\tau, \hat{\bar{\chi}}_\tau)$ are obtained from the pseudolikelihood

$$PL(c_\tau, \bar{\chi}_\tau) = \prod_{i:T_i^\tau < u_\tau} \left[1 - c_\tau u_\tau^{-2/(1+\bar{\chi}_\tau)}\right] \prod_{i:T_i^\tau \geq u_\tau} \left[-\left(\frac{2c_\tau}{1+\bar{\chi}_\tau}\right) T_i^{\tau(-2/(1+\bar{\chi}_\tau)-1)}\right] \tag{2.4.1}$$

where u_τ is a threshold selected for the series $\mathbf{T}^\tau$, which need not correspond to the original modelling threshold u. By the invariance property of the maximum likelihood we have $\hat{\bar{\chi}}^{(m)} = \max_{1 \leq \tau \leq m-1}(\hat{\bar{\chi}}_\tau)$ and $\hat{c}^{(m)} = \{\hat{c}_\tau : \hat{\bar{\chi}}_\tau = \hat{\bar{\chi}}\}$.

The second step of the inference procedure then uses the excesses over the modelling threshold (u) of the cluster maxima, denoted $\{Y_j - u | Y_j > u : j = 1, ..., r_n\}$, where the cluster maxima Y_j are defined in Section 2.3. The dependence parameters are fixed at the best estimates obtained in the first step and κ_u (the marginal probability of observing an exceedance in the original data) is estimated by the empirical probability of the original series $\{X_i\}$ exceeding the threshold. The likelihood used to obtain the MLE's of the parameters $(\hat{\sigma}_u, \hat{\xi})$ is the product of either the densities $w_u^*(v) = d/dv\left\{1 - \bar{W}_u^*(v)\right\}$ or $w_u^1(v) = d/dv\left\{1 - \bar{W}_u^1(v)\right\}$, depending on which approximation seems appropriate for the data set in question, taken over the series $\{Y_j - u | Y_j > u\}$.

It remains to be shown when either form of our proposed distribution should be used in preference to the GP distribution as a model for the cluster maxima. In an extreme value analysis we are most interested in extreme quantile estimation and so one assessment of which distribution to use would look at which one best estimates the underlying quantiles of the cluster maxima, see Section 2.5. However the underlying quantiles are unknown and so a simpler measure to calculate is the expected error in using the GP distribution to model the cluster maxima, under the assumption that the true distribution is in fact one of our proposed distributions. For the distribution $\bar{W}_u^*$ given by equation (2.3.12) we denote this measure by D, so that

$$D = \int_0^\infty \left(\bar{W}_u^*(v) - \bar{W}_u(v)\right)^2 w_u^*(v)dv. \tag{2.4.2}$$

The closed form of the measure D is given by

$$D = \left[K^{(m)}\right]^2 \left[\frac{1}{3} - \frac{C^{(m)} + K^{(m)}}{2 + \bar{\chi}^{(m)}} + \frac{C^{(m)} + 4K^{(m)}}{5 + \bar{\chi}^{(m)}}\right] \tag{2.4.3}$$

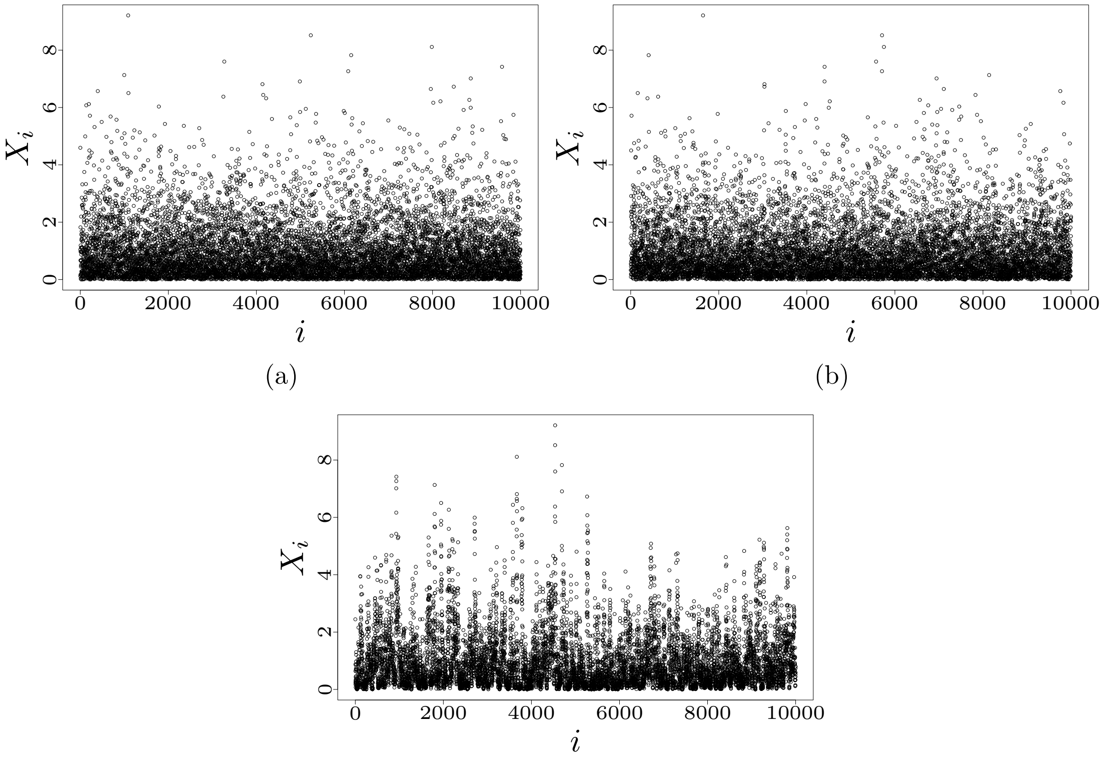
where $K^{(m)} = c^{(m)}\kappa_u^{\zeta^{(m)}}$ and $C^{(m)} = (1+K^{(m)})(1+\bar{\chi}^{(m)})$. To evaluate this statistic it is only necessary to estimate the dependence parameters $(c^{(m)}, \bar{\chi}^{(m)})$. For fixed $\bar{\chi}^{(m)}$, D increases with $c^{(m)}$. For fixed $c^{(m)}$, D is biggest when $\bar{\chi}^{(m)}$ is small and positive, *i.e.* when the series is nearly independent or displays some positive dependence. The measure D decreases as $\bar{\chi}^{(m)} \to \pm 1$ *i.e.* as the dependence, which may be either positive or negative association, becomes stronger. We have an equivalent distance measure for the distribution $\bar{W}_u^1$ given in equation (2.3.15), which we denote by D_1 and which takes the form

$$D_1 = \frac{(K^{(m)})^2}{(1 - K^{(m)})^3} \left[\frac{1 - K^{(m)}}{3} - \frac{1 + \bar{\chi}^{(m)} + K^{(m)}}{2 + \bar{\chi}^{(m)}} + \frac{1 + \bar{\chi}^{(m)} + 4K^{(m)}}{5 + \bar{\chi}^{(m)}}\right]. \tag{2.4.4}$$

Choice of the critical D (D_1) value, used to decide whether or not the GP distribution is a sufficiently close fit to model the data will depend on the context, although the simulation results of the following section may provide some guidance.

2.5 Simulation study

The simulation study detailed below compares the goodness-of-fit of the GP and both forms of our distribution, $\bar{W}_u^*$ and $\bar{W}_u^1$, to the true distribution of the cluster maxima of an asymptotically independent, first-order Markov process, at two different thresholds. Since the process exhibits non-negative dependence and is first-order the dependence between the variables $(X_i, X_{i+\tau})$ for $\tau \geq 2$ will be weaker than that for the variables at lag 1, hence we only need consider the dependence structure obtained from the distribution of (X_i, X_{i+1}) and so drop references to lags in the subsequent discussion. In what follows we fit the GP distribution $\bar{W}_u$ and compare this to the two forms of our proposed distribution, W_u^* given in equation (2.3.12) and W_u^1 given in equation (2.3.15).



(a)

(b)

(c)

Figure 2.3: Simulated runs, length 10000, of the (a) AR(0.1), (b) AR(0.5) and (c) AR(0.9) processes, each with standard exponential margins.

The process that we shall consider is the Gaussian autoregressive (AR) process with parameter ρ, which has standard normal margins and dependence structure given by $(X_i, X_{i+1}) \sim \text{BVN}(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is a 2×2 matrix with 1's on the diagonal and the correlation coefficient ρ on the off-diagonal. We consider various parameterisations of this process given by three values of the dependence parameter, $\rho = 0.1, 0.5$ and 0.9. The series is transformed to exact standard exponential margins using the empirical distribution function before modelling, for the reason given in Section 2.1. A single simulation of each parameterisation of the process is given in Figure 2.3.

For each parameterisation of the process, 500 data sets of length 10000 were simulated. Each data set was declustered at the 90% and 99% quantile thresholds and the GP and our distribution were fitted to the cluster maxima of both thresholds, using, respectively, the POT method and the two step fitting procedure described in Section 2.4. A 90% threshold was used to estimate the dependence parameters $(c, \bar{\chi})$, regardless of the declustering threshold used.

Figure 2.4, which shows estimates of the parameter $\bar{\chi}$ for each parameterisation of the process, as well as histograms of the estimated extremal indices $\theta(u)$ when $\kappa_u = 0.1$. As the correlation coefficient ρ increases so does the strength of the sub-asymptotic dependence. The estimated GP parameters (σ_u, ξ) were almost identical under both distributions for each parameterisation, but this does not necessarily mean that the two distributions model the underlying distribution of the cluster maxima equally well. To investigate this we compare estimates of the quantiles of the cluster maxima made using fits of the different distributions.

To judge how well the estimated distributions fit the cluster maxima, simulations are used to approximate the true distribution, since this distribution cannot be obtained analytically. For each process, a run of length one million was simulated and declustered above the 90% and 99% quantile thresholds. Since the run length is so long, we considered the empirical distributions of these cluster maxima to be the true distribution. We compare the quantiles of these true distributions with those of the estimated distributions. The *pth* quantile of the GP distribution is the solution of $\bar{W}_u(v_p) = 1 - p$ which has closed form $v_p = \frac{\sigma_u}{\xi}\left[(1 - v_p)^{-\xi} - 1\right]$.

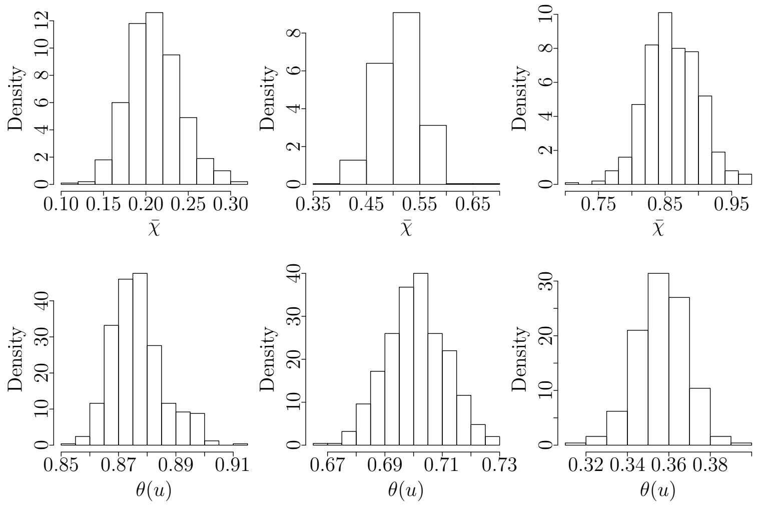


Figure 2.4: Estimates made using a 90% threshold, of the dependence parameter $\bar{\chi}$ (top) and the extremal index $\theta(u)$ where $\kappa_u = 0.1$ for each of the 500 simulations of the AR(0.1) (left), AR(0.5) (centre) and AR(0.9) (right) processes.

For both forms of our proposed distribution numerical methods are needed to solve the equations $\bar{W}_u^*(v_p) = 1 - p$ and $\bar{W}_u^1(v_p) = 1 - p$, which have no closed form.

For each data set and each threshold we estimated a range of quantiles of the cluster maxima using both of the fitted distributions. We assess the goodness of fit of the two distributions by comparing the bias and root mean square error (RMSE) of each estimated quantile. Suppose that a given quantile q of the cluster maxima has true value q^T. The bias, b, and the RMSE, r, of the quantile estimates $\{\hat{q}_j : j = 1, \ldots, n\}$ obtained from n simulations are then given by

$$b = \frac{1}{n}\sum_{j=1}^{n}(\hat{q}_j - q^T) \quad \text{and} \quad r = \left[\frac{1}{n}\sum_{j=1}^{n}(\hat{q}_j - q^T)^2\right]^{1/2}.$$

We estimated 1000 quantiles between 0.001 and 0.999. Figure 2.5 shows the bias and RMSE of the estimated quantiles under all three distributions and both threshold choices for the AR(0.1) process. For the higher quantiles, which we are likely to be most interested in for an extreme value analysis, the estimates of the quantiles from our distribution $\bar{W}_u^1$ show the smallest bias, regardless of the threshold used, although the difference is much less at the higher threshold. Similarly, the RMSE is lower for the distribution $\bar{W}_u^1$ than for either the distribution $\bar{W}_u^*$ or the GP distribution, but again the differences are markedly reduced at the higher threshold. This decrease in differences as the threshold is increased is unsurprising, since both of our distributions tend to the GP as the threshold approaches its upper limit, see equation (2.3.13).

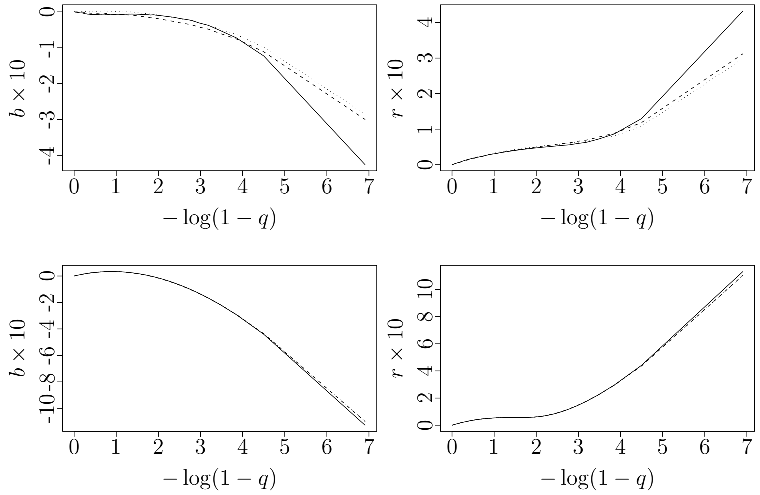


Figure 2.5: Results for the AR(0.1) process. Bias (b), left, and RMSE (r), right, for estimated quantiles q. Estimates were made using the GP (full line) and our model (dashed line - $\bar{W}_u^*$ and dotted line - $\bar{W}_u^1$) fits to the cluster maxima of both the 90%, top, and 99%, bottom, thresholds.

Overall though the AR process study showed no great improvements in using either form of our distribution instead of the GP distribution. Figure 2.6 shows the bias and RMSE for the 90% threshold models of the AR process with parameter values $\rho = 0.5$ and $\rho = 0.9$. When $\rho = 0.5$ both forms of our distribution have smaller bias and RMSE in the upper tail of the distribution, but there is a tradeoff as they both have larger bias and RMSE in the body and lower tail of the distribution. For $\rho = 0.9$, both forms of our distribution show greater bias and RMSE throughout the distribution. Note that the plots emphasise the difference, both in bias and in RMSE, between the full form of our distribution $\bar{W}_u^1$ and the approximated form $\bar{W}_u^*$; these differences are considerably larger than the equivalent one between the asymptotic GPD and the approximation $\bar{W}_u^*$. This suggests that taking the full form of our distribution may be more appropriate here.

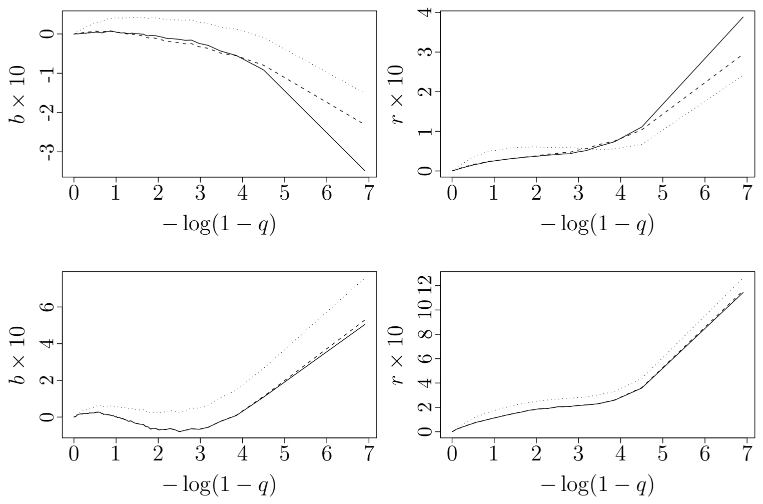


Figure 2.6: Results for AR(0.5), top, and AR(0.9), bottom, processes. Bias (b), left, and RMSE (r), right, for quantiles (q) estimated from both the GP (full line) and our model (dashed line - $\bar{W}_u^*$ and dotted line - $\bar{W}_u^1$) fits to the cluster maxima of 90% thresholds.

The distance measures D (D_1) are also fairly non-informative in this case. At the 90% threshold the median values of D (D_1) across the 500 simulated data sets are, for $\rho = 0.1$, 0.000279 (0.000366), for $\rho = 0.5$, 0.000519 (0.00106) and for $\rho = 0.9$, 0.000162 (0.00218). At the 99% threshold the equivalent values are, for increasing ρ, 0.0000138 (0.0000145), 0.000116 (0.000157) and 0.000114 (0.000541). We see that as the threshold increases the distance between the GP and the two forms of our distribution decreases and that there is a bigger distance between the GP and the full form of our distribution $\bar{W}_u^1$ than between the GP and the approximate form of our distribution $\bar{W}_u^*$. Note that all the distances are very small, but that they increase as the strength of the sub-asymptotic dependence increases, *i.e.* as $\rho \to 1$.

This simulation study shows that for the Gaussian first-order AR process there is little to be gained from using either form of our distribution over the asymptotically motivated GP distribution as a model for the distribution of the cluster maxima of exceedances of sub-asymptotic thresholds. This is especially evident as the threshold increases, in which case the gain from using our proposed distribution over the GP distribution becomes much less. Further, examination of various other dependence structures given, for example, by the BB6 and Morgernstern copulas (Joe, 1997), and the lower tail of the bivariate extreme value distribution, none of which are shown, suggests that this is the case for a wide range of asymptotically independent dependence structures. However because there are some evident differences in fit between the GP distribution and the two forms of our model shown in Figures 2.5 and 2.6 we suggest that it is still worth investigating the fit of all three distributions for any particular example, especially in the case of large sample size where bias is more important than variance.

2.6 Ozone Analysis

We now return to the initial problem of the most suitable model for the cluster maxima of the ozone data set, introduced in Section 2.1, by comparing the differences in fit between the GP and our distributions. The ozone data is shown on both the original and transformed scales in Figure 2.1. The transformation removes the non-stationarity before modelling the cluster maxima. In order to make the minimum number of modelling assumptions in doing this we follow a local, nonparametric approach, the two stages of which are as follows.

The daily data are first standardised within years, to remove annual linear trends in mean or variance. Let X_{ki} be the observation made on the ith day ($i = 1, \ldots, 365$) of the kth year ($k = 1, \ldots, Y$) where Y is the total number of years. Then we calculate the standardised series $Z_{ki} = (X_{ki} - m_k)/s_k$, where m_k and s_k are, respectively, the sample mean and standard deviation of the data in year k. For each day of the year i, the standardised data are then pooled across years k. We assume that this standardised pooled series is stationary within a window over any short time interval $[i - h, i + h]$ for some h. By moving this window across days i and using the empirical distribution of the within window data, we can transform the data observed on the central day i in the window to standard uniform margins. The inverse probability integral transform is then used to transform to any desired margins. The size of the window (equivalently the value of h) depends entirely on the structure of the non-stationarity in the data set. After an exploratory investigation, we found that for the ozone data, a window of two months ($h = 30$ days) produced satisfactory results.

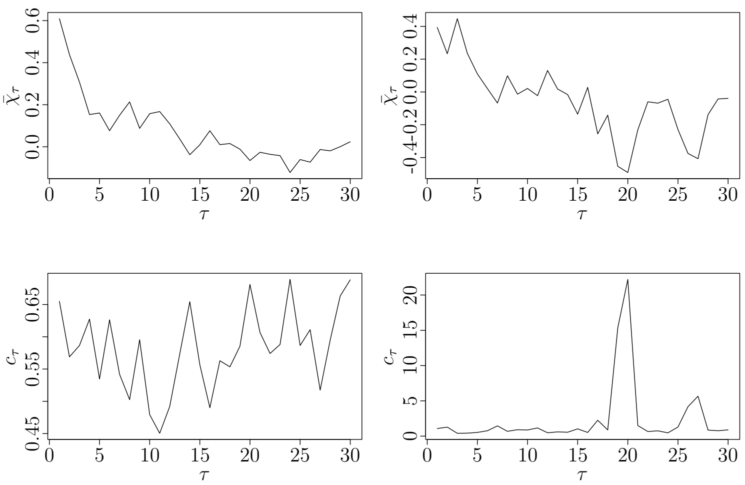


Figure 2.7: Estimates of $\bar{\chi}_\tau$ (top) and c_τ (bottom) using the lags $\tau = 1, ..., 30$. The left hand plots use the 90% quantile for the threshold and the right hand plots the 99% quantile.

We declustered the transformed data above fifteen declustering thresholds, uniformly spread between the 70- and 95% quantiles, using the runs estimator with a run length of 3. For each threshold, we compared the fit of the three distributions $\bar{W}_u$, $\bar{W}_u^*$ and $\bar{W}_u^1$ to the cluster maxima by calculating the two distance measures D and D_1 given in equations (2.4.3) and (2.4.4). For each declustering threshold, we also considered both 90- and 99% quantiles as thresholds for estimating the parameters $(\bar{\chi}, c)$. Figure 2.7 shows estimates of $(\bar{\chi}_\tau, c_\tau)$ obtained using the lags $\tau = 1, \ldots, 30$ and both 90- and 99% thresholds. These plots show the data getting closer to independence for any lag greater than two or three days, since then $\bar{\chi}$ is getting close to zero, with near independence achieved from a lag of about 10 days. For smaller lags there is weak positive association. These conclusions hold regardless of the threshold chosen. Further, by combining these results with plots showing the stability of the behaviour of the extremal index across various run lengths (not shown) and the fact that the persistence of high levels of ozone is due to the persistence of external conditions, such as the weather (which in the

UK may last for several days) rather than the lifetime of the individual ozone molecules (which is a number of hours) we believe that the use of a run length of 3 may be justified.

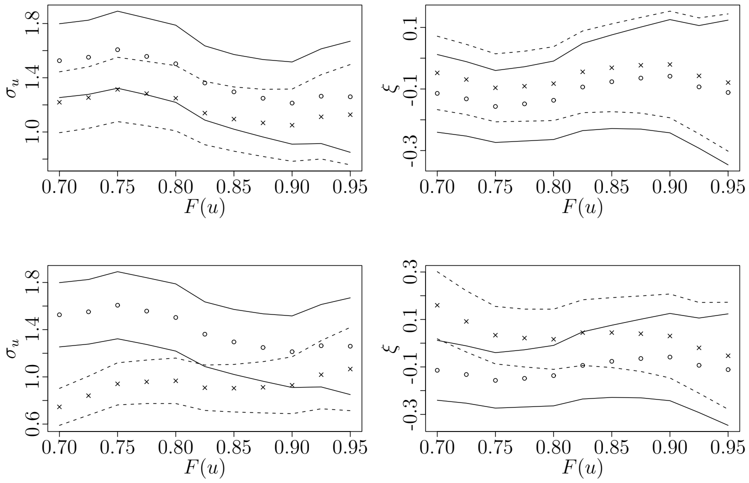


Figure 2.8: Fitted parameters σ_u (left) and ξ (right) for the distribution of the cluster maxima over a range of declustering thresholds (70-95% quantiles) using the GP (circles) and our distribution $\bar{W}_u^1$ (crosses). Also shown are estimated 95% confidence intervals; full lines for the GP and dashed lines for our distribution. Two thresholds, 90% (top) and 99% (bottom) quantiles were used to estimate the parameters $(\bar{\chi}, c)$.

Figure 2.8 shows estimates of the parameters (σ_u, ξ) obtained by fitting the distributions $\bar{W}_u$ and $\bar{W}_u^1$. Parameter estimates obtained from fitting the distribution $\bar{W}_u^*$ (not shown) lie somewhere in-between the estimates from the other two models; this is to be expected given that this distribution is an asymptotic approximation to $\bar{W}_u^1$, as discussed in Section 2.3. The 95% confidence intervals show that, although the point estimates of the parameters are quite different under the two models, these differences are small compared to the uncertainty, especially at the higher thresholds. For both thresholds the D- and D_1-statistics are very small; the D-statistic is less than 7×10^{-4} for the 90% threshold and 0.004 for the 99% threshold. Further, as the declustering threshold increases, so D decreases and, as we expect, the difference between the fitted distributions diminishes.

Finally we consider what happens if we try to extrapolate the fit of either one of our distributions to higher thresholds. Specifically, we fit all three of the distributions to the cluster maxima at the 80% threshold, also using an 80% threshold level to estimate the dependence parameters $(\bar{\chi}, c)$. We then use the threshold stability property of the GP distribution outlined in Sections 2.1 to look at the goodness of fit of these fitted models to the cluster maxima of the 90- and 99% thresholds. To demonstrate this we use quantile-quantile (QQ) plots; for the distributions $\bar{W}_u$ and $\bar{W}_u^1$ these are shown in Figure 2.9, plots for $\bar{W}_u^*$ (not shown) lie somewhere in between the two. We see that the models fitted under both distributions appear to fit well and as we increase the threshold the difference in fit is negligible, as our distribution approaches the GP distribution.

For this data set the difference between the two distributions at all thresholds is so small that it seems unnecessary to use the distribution introduced in this paper despite the instability of $\theta(u, m)$ found in Figure 2.2. On the evidence given in Figures 2.7 and 2.8 we believe that the GP distribution is an adequate approximation for the distribution of the cluster maxima of the transformed ozone data set.

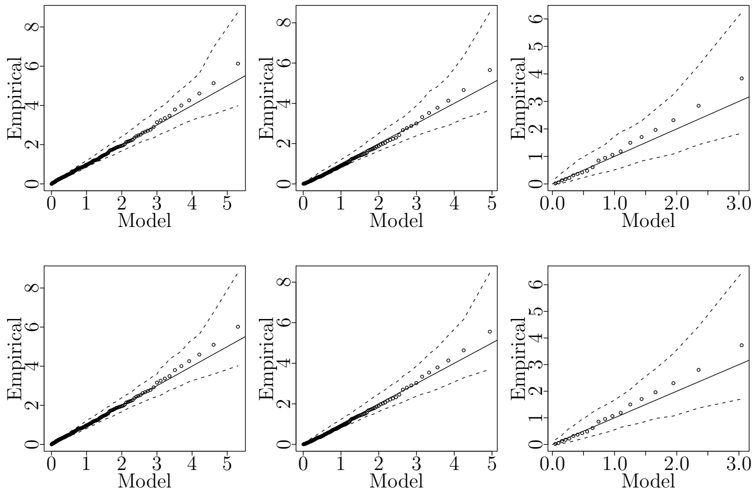


Figure 2.9: QQ plots for GP $\bar{W}_u$ (top) and our $\bar{W}_u^1$ (bottom) distributions fitted to the cluster maxima of the 80% threshold (left). An 80% threshold was also used to estimate the parameters $(\bar{\chi}, c)$. Plots in the centre and right show, respectively, goodness of fit of the distributions fitted at the 80% threshold to the cluster maxima of the 90- and 99% thresholds. 95% confidence intervals are given by dashed lines, with the 45° line giving exact agreement between the model and the data.

References

Beirlant, J., Goegebeur, Y., Segers, J. and Teugels, J. (2004) *Statistics of Extremes: Theory and Applications*, John Wiley & Sons Ltd, Chichester, UK.

Bortot, P. and Tawn, J.A. (1998) Models for the extremes of Markov chains, *Biometrika* **85**:4, 851-867.

Chernick, M.R., Hsing, T. and McCormick, W.P. (1991) Calculating the Extremal Index for a class of stationary sequences. *Adv. Appl. Prob.* **23**, 835-850.

Coles, S., Heffernan, J. and Tawn, J. (1999) Dependence Measures for Extreme Value Analyses *Extremes* **2**:4, 339-365.

Colls, J. (2002) *Air Pollution, Second Edition*, Spon Press, London, UK.

Davison, A.C. and Smith, R.L. (1990) Models for Exceedances over High Threshold *JRSSB* **52**:3, 393-442.

Ferro, C.A.T. and Segers, J. (2003) Inference for clusters of extreme values. *JRSSB* **65**:2, 545-556.

Hall, P. and Tajvidi, N. (2000) Nonparametric Analysis of Temporal Trend When Fitting Parametric Models to Extreme-Value Data. *Statistical Science* **15**:2, 153-167.

Hsing, T, Hüsler, J and Reiss, R. (1996). The extremes of triangular array of normal random variables. *Ann. Appl. Probab.*, **6**, 671-686.

Kratz, M.-F. and Rootzén, H. (1997). On the rate of convergence for extremes of mean square differentiable stationary normal processes. *J. Appl. Probab.*, **34**, 908-923.

Joe, H. (1997) *Multivariate Models and Dependence Concepts* Chapman and Hall, London, UK.

Küchenhoff, H. and Thamerus, M. (1996) Extreme Value Analysis of Munich air pollution data. *Environmental and Ecological Statistics* **3**, 127-141.

Latham, M. (2006) Statistical methodology for the extreme values of dependent processes. *PhD thesis, Lancaster University.*

Leadbetter, M.R. (1991) On a basis for ‘Peaks over Threshold’ modeling. *Statis-*

tics and Probability Letters **12**, 357-362.

Leadbetter, M.R., Lindgren, G. and Rootzén, H. (1983) *Extremes and Related Properties of Random Sequences and Process*, Springer-Verlag, New York.

Ledford, A.W. and Tawn, J.A. (1996) Statistics for near independence in multivariate extreme values. *Biometrika* **83**:1, 169-187.

Ledford, A.W. and Tawn, J.A. (2003) Diagnostics for dependence within time series extremes. *JRSSB* **65**:2, 521-543.

O'Brien, G.L. (1987) Extreme values for stationary and Markov sequences. *Ann. Prob.* **15**:1, 281-291.

Pickands, J. (1971) The two-dimensional Poisson process and extremal processes. *J. Appl. Prob.* **8**, 745-756.

Ramesh, N.I. and Davison, A.C. (2002) Local models for exploratory analysis of hydrological extremes. *Journal of Hydrology* **256**, 106-119.

Smith, R.L. (1989) Extreme Value Analysis of Environmental Time Series: An Application to Trend Detection in Ground-Level Ozone. *Statistical Science* **4**:4, 367-393.

Smith, R.L., Tawn, J.A. and Coles, S.G. (1997) Markov chain models for threshold exceedances. *Biometrika* **84**:2, 249-268.

Smith, R.L. and Weissman, I. (1994) Estimating the Extremal Index. *JRSSB* **56**:3, 515-528.

Chapter 3

Modelling non-stationary extremes

3.1 Introduction

Statistical methods for modelling extremes of stationary sequences have received much attention and though different methods for inference do exist the modelling strategies are basically identical (Coles, 2001; Bierlant *et al.*, 2004; and de Haan and Ferreira, 2006). Specifically local maxima which exceed a high threshold are modelled by a parametric model which is motivated by the asymptotic theory of extreme values of independent and identically distributed random variables.

In many cases however an analysis of the extremes of a series is required where there is clear non-stationarity in the series. This is especially common in environmental data sets. The focus of this paper is the analysis of the ozone (O_3) data set shown in Figure 3.1, which consists of the daily maxima of hourly concentrations of surface-level ozone. These data were measured at a monitoring site in central Reading, UK, which is part of an automatic air quality monitoring network run on behalf of the UK government.

Features of the apparent non-stationarity of the ozone data can be explained by accounting for the underlying mechanisms driving the ozone generation process.

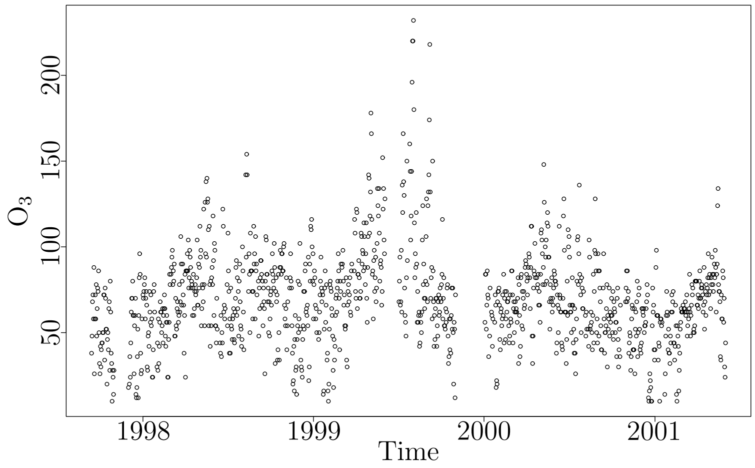


Figure 3.1: Daily maxima of hourly ozone (O_3) concentrations observed in central Reading from 13/09/97 until 01/06/01. Measurements of O_3 are in μmg^{-3}.

Surface-level ozone is a secondary pollutant, meaning that it is formed in the atmosphere, and its level depends on the concentrations of precursors, principally nitric oxide (NO) and nitrogen dioxide (NO_2), and various hydrocarbons. Various human activities cause increases beyond ambient levels in these precursors. This is primarily due to the combustion of fossil fuels and consequently the precursors show seasonal trends (combustion tends to increase in colder weather). Further, the chemical reactions involved in the synthesis of ozone depend on meteorological conditions. The key reactions are photochemical and so sunlight is an important factor, as are temperature, wind speed and wind direction. Thus it is natural to incorporate such covariates into an analysis of the ozone data in order to attempt to explain the non-stationarity.

The statistical analysis of ozone data has been much investigated in recent years (Thompson *et al.*, 2001). They suggest four potential motivations for the statistical study of ozone data sets which are: forecasting high levels in order to give out public health warnings; identifying trends in high ozone levels, possibly in response to legislation regulating pollution emissions; understanding the underlying mechanisms of the process; and recognising health impacts. We also suggest that, given the current scientific, political and economic interest in ascertaining the impact of human activities on the environment, a further motivation is to assess changes in ozone levels due to such activities, either directly, through changing emission patterns, or indirectly, through climate change. Extreme value methods are particularly suited to analyses concerned with questions relating to the first two and the last of these factors. Specifically we are interested in explaining the changes in extreme ozone levels conditional on the covariates relating to the precursor concentrations and meteorological conditions and in summarising the marginal distribution of extreme ozone levels under current conditions and for scenarios corresponding to future changes in emission patterns and climate change.

Let $\{Y_t\}$ be a process with associated covariates $\{\mathbf{X}_t\}$. The simplest approach to predict future extreme levels of the marginal distribution of $\{Y_t\}$ is to model the extremes using methods for stationary sequences. We can then estimate $100(1-p)\%$ quantile (termed the *marginal return level*), denoted by y_p such that $\Pr(Y_t > y_p) = p$, for p close to zero; under stationarity this is exceeded on average approximately once every $1/p$ observations. However, if $\{Y_t\}$ is non-stationary such a direct approach is subject to unbounded and unquantifiable bias. Furthermore, it does not allow the identification of trends or covariate relationships which are required for deriving the distribution of future extreme ozone levels under scenarios of change.

An alternative approach is to model the extremes conditional on the covariates. We focus on modelling the non-stationarity in the marginal distribution of $\{Y_t\}$ through the conditional distribution of $Y_t|\mathbf{X}_t = \mathbf{x}_t$. The return level is then most naturally defined as a quantile of Y_t conditionally on the vector of covariates $\mathbf{X}_t$. These $100(1-p)\%$ *conditional return levels*, denoted $y_{p,t}$, satisfy

$$\Pr(Y_t > y_{p,t}|\mathbf{X}_t = \mathbf{x}_t) = p. \tag{3.1.1}$$

However, if interest is in the behaviour of Y_t alone then we can integrate out the covariates as follows

$$\Pr(Y_t > y) = \int_{\mathbf{x}_t} \Pr(Y_t > y|\mathbf{X}_t = \mathbf{x}_t) f_{\mathbf{X}_t}(\mathbf{x}_t) d\mathbf{x}_t, \tag{3.1.2}$$

where $f_{\mathbf{X}_t}(\cdot)$ is a model for the joint density of the covariates $\mathbf{X}_t$ at time t, and obtain the *(marginal) return level* y_p given by $\Pr(Y_t > y_p) = p$. Under the assumption that the observed covariates form a representative sample from the joint distribution in some specified period of interest then, in the absence of any prior information, the joint distribution can be estimated empirically then the marginal return level is the solution to the equation

$$p = \frac{1}{n}\sum_{t=1}^{n} \Pr(Y_t > y_p | \mathbf{X}_t = \mathbf{x}_t), \qquad (3.1.3)$$

where n is the size of the sample of covariates. Different models for $f_{\mathbf{X}_t}(\cdot)$ can be proposed to account for future emission and climate change scenarios. The resulting change in the marginal return level under a change scenario from that given by the use of equation (3.1.3) gives a single measure of how a particular scenario might affect extreme ozone concentration levels.

The standard method (Davison and Smith, 1990) of analysis for modelling the extremes of a non-stationary process retains the use of a constant high threshold and introduces covariates into the threshold exceedance rate and the parameters of the extreme value model for the threshold exceedances. In this paper, we present a case against such a modelling approach and introduce an alternative strategy. The novel step in the alternative strategy is first to attempt to model the non-stationarity in the whole data set. This non-stationarity is then removed from the data, a technique referred to as pre-processing, and the extremes of the pre-processed data are modelled using the standard approach. Critical to our approach is that if pre-processing is successful the extremes of the preprocessed series will have had most, if not all, of the non-stationarity of $\{Y_t\}$ removed and thus a simple extreme value analysis of the preprocessed series can be conducted.

Using the entire data set to model the extremes seems a departure from the usual extreme value techniques, which capitalise on the general theory of extreme values of stationary sequences that allows inference on the tails of a distribution to be made independently of the main distribution body. However such theory does not directly extend to the extremes of sequences whose underlying distribution is conditional on covariates. Provided that a reasonable model for the non-stationarity in the entire data set is used at the pre-processing step, we believe that our proposed strategy often will provide a better description of the non-stationarity of the extremes, a clearer scientific interpretation, a more appropriate identification of the extreme values, easier threshold selection, reduced threshold sensitivity, and improved covariate model selection and efficiency of inference for covariate effects and extremal properties.

In Section 3.2 we review existing methods for modelling the extremes of both stationary and non-stationary processes. We then introduce our proposed approach in Section 3.3, as well as an 'in-between' approach, termed the varying threshold method. Results from a simulation study are shown in Section 3.4, in which the efficiency and ability to select the correct covariate models of the standard and pre-processing methods are compared. We show results of an analysis of the ozone data in Section 3.5 comparing the various methods and assessing the impact of mis-specification of the covariate model in the pre-processing step. Our reasons for such a study are motivated by the fact that we do not have available all the precursor and meteorological covariates necessary to account fully for the known mechanisms behind ozone generation; so we assess the impact of using a scientific and data-based rationale for model building. We conclude with a comparison of the standard, pre-processing and varying threshold methods in Section 3.6 which summarises the findings in the paper and justifies our claimed benefits for the pre-processing approach.

Throughout the paper we assume that the extreme events of either $\{Y_t\}$ or $\{Y_t|\mathbf{X}_t = \mathbf{x}_t\}$ are temporally independent. However when evaluating confidence intervals of estimates in the ozone application we use a block bootstrap to account for any temporal dependence.

3.2 The standard approach

3.2.1 Stationary processes

Suppose that the process of interest $\{Y_t\}$ is stationary with univariate marginal distribution F which has upper endpoint x^F. We define the extremes of $\{Y_t\}$ to be the exceedances of a high threshold u, $u < x^F$. As u tends to x^F, Pickands (1975) showed that, if the distribution of the excesses, $Y_t - u$, of u, scaled as a function of u, converges to a non-degenerate limiting distribution, that distribution must be the generalised Pareto distribution (GPD). This motivates the use of the GPD as a statistical model for the excesses of a high, fixed threshold u. The conditional survivor function for the exceedances of u under the assumption that excesses follow a GPD(ψ_u, ξ) model is, for $y > 0$

$$\Pr(Y > y + u \,|\, Y > u) = \left[1 + \frac{\xi y}{\psi_u}\right]_+^{-1/\xi} \tag{3.2.1}$$

where $a_+ = \max\{0, a\}$ and $\psi_u > 0$ and ξ are scale and shape parameters respectively. An additional parameter of the tail model is $\phi_u = \Pr(Y > u)$ which determines the rate of exceedance of the threshold. The theoretical justification for this model requires that ϕ_u is small, since, unless F itself is GPD, the approximation to the tail of F by the GPD holds only as u tends to x^F. This threshold approach, popularised by Davison and Smith (1990), models the size and rate of occurrence of the observations which exceed the threshold.

An important property of the GPD is that of *threshold-stability*. Suppose that the conditional distribution of the exceedances of u is a GPD(ψ_u, ξ). Then for any level v, $u < v < x^F$, the conditional distribution of the exceedances of v is a GPD(ψ_v, ξ) distribution, where $\psi_v = \psi_u + \xi(v - u)$. This result shows that the form of the distribution of the threshold exceedances, including the shape parameter, is invariant to the selection of a higher threshold.

There are a range of methods for inference for the GPD and tail models when the data are assumed to be stationary. To avoid dependence in exceedances most often a *peaks over threshold* (POT) analysis is used, where only cluster maxima data are used in the GPD fit and confidence interval evaluation (Davison and Smith, 1990). One alternative is to fit the GPD using all exceedances and explicitly model the dependence between exceedances in a cluster (Smith, Tawn and Coles, 1997). A second alternative is to fit the GPD using all exceedances, falsely assuming that these are independent, and then account for dependence in the confidence interval evaluation by using block bootstrap methods (Buishand, 1993), to be discussed in Section 3.5.

We shall use likelihood inference throughout. Using all the threshold exceedances, under the assumption of independence of extreme events the likelihood function for the stationary model is

$$L(\psi_u, \xi, \phi_u) = \prod_{t=1}^{n} (1-\phi_u)^{1-I[y_t>u]} \left(\phi_u \psi_u^{-1} \left[1 + \frac{\xi(y_t - u)}{\psi_u} \right]_+^{-1/\xi - 1} \right)^{I[y_t>u]} \tag{3.2.2}$$

where $I[y_t > u]$ is the indicator function taking the value 1 if $y_t > u$ and zero otherwise. The maximum likelihood estimate (MLE) for the rate parameter is $\hat{\phi}_u = n_u/n$, where n_u is the number of exceedances of the threshold u. The MLE's of the GPD parameters are found by numerical optimisation. For a stationary series, assuming $p < \hat{\phi}_u$ which implies that $y_p > u$, the estimated marginal return level is

$$\hat{y}_p = u + \frac{\hat{\psi}_u}{\hat{\xi}} \left[\left(\frac{\hat{\phi}_u}{p} \right)^{\hat{\xi}} - 1 \right]. \tag{3.2.3}$$

3.2.2 Non-stationary processes

Now suppose that the process $\{Y_t\}$ is non-stationary and has an associated sequence of covariates $\{\mathbf{X}_t\}$. The first full proposal for extending the GPD to non-stationary cases was given by Davison and Smith (1990) with an associated proposal made by Smith (1989). They suggest continuing to model the exceedances of a fixed high threshold u and to account for the non-stationarity of the exceedances by allowing the parameters of the GPD to be modelled as functions of the covariates. Thus they model the rate of exceedance by $\phi_u(\mathbf{x}) = \Pr(Y > u|\mathbf{X} = \mathbf{x})$ and the distribution of excesses by a GPD$(\psi_u(\mathbf{x}), \xi(\mathbf{x}))$, *i.e.* for $y > 0$

$$\Pr[Y > y + u|Y > u, \mathbf{X} = \mathbf{x}] = \left[1 + \frac{\xi(\mathbf{x})y}{\psi_u(\mathbf{x})}\right]_+^{-1/\xi(\mathbf{x})}. \tag{3.2.4}$$

Under the assumption of temporal independence the likelihood function takes the form

$$\prod_{t=1}^{n}[1 - \phi_u(\mathbf{x}_t)]^{1-I[y_t>u]} \left[\phi_u(\mathbf{x}_t)\psi_u(\mathbf{x}_t)^{-1}\left[1 + \xi(\mathbf{x}_t)\frac{y_t - u}{\psi_u(\mathbf{x}_t)}\right]_+^{-1/\xi(\mathbf{x}_t)-1}\right]^{I[y_t>u]} \tag{3.2.5}$$

Initially linear covariate models were used, with a log-link for the rate and scale parameters, *e.g.* Davison and Smith (1990), Smith and Shively (1995) and Coles (2001), although more recent studies have considered the use of additive or fully nonparametric models *e.g.* Hall and Tajvidi (2000), Davison and Ramesh (2000) and Chavez-Demoulin and Davison (2005). In this paper we treat $\log \psi_u$, ξ and logit ϕ_u as linear functions of covariates, so for vectors of coefficients $\boldsymbol{\psi}_u$, $\boldsymbol{\xi}$ and $\boldsymbol{\phi}_u$,

$$\log \psi_u(\mathbf{x}) = \boldsymbol{\psi}_u'\mathbf{x}, \quad \xi(\mathbf{x}) = \boldsymbol{\xi}'\mathbf{x}, \quad \text{and} \quad \text{logit } \phi_u(\mathbf{x}) = \boldsymbol{\phi}_u'\mathbf{x}. \tag{3.2.6}$$

One disadvantage of this model is that it does not retain the threshold-stability property of the GPD as discussed for the stationary case. In order to retain this property in the non-stationary model the functional form of the scale parameter must satisfy, for any $v > u$,

$$\psi_v(\mathbf{x}) = \psi_u(\mathbf{x}) + (v - u)\xi(\mathbf{x}). \tag{3.2.7}$$

If different covariates were included in the scale $\psi_u(\mathbf{x})$ and shape $\xi(\mathbf{x})$ parameters this would obviously lead to inconsistency between the covariates included in $\psi_u(\mathbf{x})$ and those included in $\psi_v(\mathbf{x})$ for all $v > u$. This fundamental property of the standard model does not seem to have been identified before and it rather undermines the use of such models as it implies their form of covariate selection in the parameters is non-invariant to threshold choice. It could be argued that as $\xi(\mathbf{x})$ is often constant then the implications of constraint (3.2.7) are not problematic. However, even then constraint (3.2.7) implies that $\psi_u(\mathbf{x})$ cannot retain the same functional form unless it is constant or a linear function, with the latter being inconsistent with the log link formulation shown in equation (3.2.6).

The conditional return levels of equation (3.1.1) can be found in a similar manner to the return levels in the stationary case. The conditional return level when $\phi_u(\mathbf{x}_t) \leq p$ must be below the threshold so the only available information is that it is censored by $y_{p,t} \leq u$. However for observations where $\phi_u(\mathbf{x}_t) > p$ we have, for $y_{p,t} > u$

$$y_{p,t} = u + \frac{\psi_u(\mathbf{x}_t)}{\xi(\mathbf{x}_t)} \left[\left(\frac{\phi_u(\mathbf{x}_t)}{p} \right)^{\xi(\mathbf{x}_t)} - 1 \right]. \tag{3.2.8}$$

Under the assumption of stationarity in the covariate distribution then equation (3.1.3) can be used for finding the marginal return level y_p. Because we make no distributional assumption on the data below the threshold, we cannot estimate y_p when $y_p \leq u$. For $y_p > u$ then equation (3.1.3) gives

$$\begin{aligned} p &= \frac{1}{n}\sum_{t=1}^{n} \Pr(Y_t > y_p | \mathbf{X}_t = \mathbf{x}_t, Y_t > u) \Pr(Y_t > u | \mathbf{X}_t = \mathbf{x}_t) \\ &= \frac{1}{n}\sum_{t=1}^{n} \phi(\mathbf{x}_t) \left[1 + \xi(\mathbf{x}_t) \frac{y_p - u}{\psi(\mathbf{x}_t)} \right]_+^{-1/\xi(\mathbf{x}_t)}. \end{aligned} \tag{3.2.9}$$

To find the MLE $\hat{y}_p$, replace the parameters in equation (3.2.9) by their MLE's and solve numerically.

3.3 Pre-processing Methods

3.3.1 Full pre-processing model

A common approach for handling non-stationarity in a time series is to pre-process (or *pre-whiten*) the full data series before fitting a model for a stationary series (Chatfield, 2004). Essentially we propose this as the basis for modelling extreme values of a non-stationary process. Our pre-processing approach involves first fitting a model for the covariate effect on the underlying distribution of the process $\{Y_t\}$. In some contexts an established model, based on a scientific or data-based rationale, may already exist. In the absence of such a model a flexible statistical model could be fitted. Specifically, we propose a Box-Cox location-scale model of the form

$$\frac{Y_t^{\lambda(\mathbf{x}_t)} - 1}{\lambda(\mathbf{x}_t)} = \mu(\mathbf{x}_t) + \sigma(\mathbf{x}_t) Z_t \tag{3.3.1}$$

where $\{Z_t\}$ are assumed to be approximately stationary, and λ, μ and $\log(\sigma)$ are linear functions of the covariates. We do not include previous values of $\{Y_t\}$ as covariates since we assume that, conditionally on the covariates, the $\{Y_t\}$ process has independent events and also for consistency with the standard method, where we know of no examples of using previous values of the process as covariates for the current value.

We shall assume that the body of the distribution of the derived series $\{Z_t\}$ is stationary and can be modelled using its empirical distribution $\tilde{F}_Z$. However, we do not use the stationary model of Section 3.2.1 for the extremes of $\{Z_t\}$ as the extreme values of $\{Y_t\}$ may have a different form of non-stationarity than for all of $\{Y_t\}$ or our Box-Cox location-scale model may not fully capture all the covariate effects, so the extreme values of $\{Z_t\}$ may not behave like extreme values of a stationary series. Instead, we model the extreme values of $\{Z_t\}$ using the methods for non-stationary extremes in Section 3.2.2, *i.e.* with a fixed threshold u_z. Let $\phi_{z,u}(\mathbf{x}_t)$ be the rate of exceedance of u_z by Z_t, and define the GPD scale and shape parameters by $\psi_{z,u}(\mathbf{x}_t)$ and $\xi_z(\mathbf{x}_t)$ respectively. Thus the full pre-processing model comprises a GPD$(\psi_{z,u}(\mathbf{x}_t), \xi_z(\mathbf{x}_t))$ for threshold exceedances and the empirical distribution of the transformed process $\{Z_t\}$, $\tilde{F}_Z$, below this level. To estimate return levels, we therefore use the GPD if $\phi(\mathbf{x}_t) > p$, otherwise we use the empirical distribution $\tilde{F}_Z$. Critical to our use of the standard method of analysis for the extremes of $\{Z_t\}$ is that we believe most, if not all, of the non-stationarity of $\{Y_t\}$ will have been removed, or at least simplified, so that the majority of problems identified in Section 3.2.2 concerning the lack of threshold stability will have been alleviated.

Inference for this model then follows a two-step procedure; the first step is to estimate the Box-Cox location-scale parameters $(\lambda(\mathbf{x}_t), \mu(\mathbf{x}_t), \sigma(\mathbf{x}_t))$. There are many possible ways to do this, but we suggest assuming that the underlying distribution is Gaussian since it is then straightforward to use likelihood inference to estimate the Box-Cox and location-scale parameters and it is robust to observations in the tails. The second step is to model the tail of the approximately stationary series $\{Z_t\}$ using the approach for non-stationary series discussed in Section 3.2.2.

The conditional and marginal return levels defined for a non-stationary series in equations (3.1.1) and (3.1.2) can easily be obtained under the pre-processing approach. We start with the conditional return levels. Since

$$p = \Pr(Y_t > y_{p,t} | \mathbf{X}_t = \mathbf{x}_t) = \Pr\left(\mu(\mathbf{x}_t) + \sigma(\mathbf{x}_t) Z_t > \frac{y_{p,t}^{\lambda(\mathbf{x}_t)} - 1}{\lambda(\mathbf{x}_t)} \middle| \mathbf{X}_t = \mathbf{x}_t\right)$$

we can first find the conditional return levels $z_{p,t}$ for the transformed series $\{Z_t\}$ and then back transform these to give

$$y_{p,t} = \{\lambda(\mathbf{x}_t)[\mu(\mathbf{x}_t) + \sigma(\mathbf{x}_t) z_{p,t}] + 1\}^{1/\lambda(\mathbf{x}_t)}.$$

Unlike in the standard method, if $\phi_{z,u}(\mathbf{x}_t) \leq p$ the conditional return levels $z_{p,t}$ can be estimated using $\tilde{F}_Z$. If $\phi_{z,u}(\mathbf{x}_t) > p$ the conditional return levels $z_{p,t}$ can be estimated using expression (3.2.8).

Let $z_p(\mathbf{x}_t)$ be the transformation under equation (3.3.1) of the marginal return level y_p. Then y_p is the solution to the equation

$$\begin{aligned} p &= \frac{1}{n}\left[\sum_{t\in T}\Pr(Z_t > z_p(\mathbf{x}_t)|\mathbf{X}_t = \mathbf{x}_t, Z_t > u_z)\Pr(Z_t > u_z|\mathbf{X}_t = \mathbf{x}_t)\right. \\ &\quad \left. + \sum_{t\notin T}\Pr(Z_t > z_p(\mathbf{x}_t)|\mathbf{X}_t = \mathbf{x}_t)\right] \\ &= \frac{1}{n}\left[\sum_{t\in T}\left(\phi_{z,u}(\mathbf{x}_t)\left[1+\xi_z\frac{z_p(\mathbf{x}_t)-u_z}{\psi_{z,u}(\mathbf{x}_t)}\right]_+^{-1/\xi_z}\right) + \sum_{t\notin T} 1 - \tilde{F}_Z(z_p(\mathbf{x}_t))\right] \end{aligned}$$

where $T = \{t : z_p(\mathbf{x}_t) > u_z\}$ is the set of all times where the transformed marginal return level exceeds the threshold u_z so that the GPD model for exceedances holds.

3.3.2 Varying threshold approach

An alternative method that is ‘in-between’ the standard and pre-processing methods is to use a time (and/or covariate) varying threshold to define the extremes on the original scale. This can be seen as an extension to the already popular approach of splitting data into seasons to allow for different thresholds in different seasons (see Smith, 1989, Küchenhoff and Thamerus, 1996 and Heffernan and Tawn, 2004, for examples with ozone data), which allows a continuously varying threshold. Such a threshold may be obtained from the pre-processing method by transforming the constant threshold u_z back to the original scale to give the varying threshold

$$u(\mathbf{x}_t) = \{\lambda(\mathbf{x}_t)[\mu(\mathbf{x}_t) + \sigma(\mathbf{x}_t)u_z] + 1\}^{1/\lambda(\mathbf{x}_t)}. \qquad (3.3.2)$$

The excesses of this threshold can then be modelled using the method for non-stationary extremes outlined in Section 3.2.2. Estimates of both conditional and marginal return levels are obtained in the same way as for the standard method. Specifically, as in the standard method and unlike the pre-processing method, we cannot make estimates of either return level below the threshold.

A further disadvantage of this method compared to the pre-processing method is that the GPD parameters fitted under the varying threshold method are likely to have more covariates than in the pre-processing model, making it more difficult to fit the model. This can be seen by considering the simplest case where the extremes of $\{Z_t\}$ are stationary, *i.e.* $Z_t | Z_t > u_z \sim \text{GPD}(\psi_{z,u}, \xi_z)$. By a change of variable, the distribution of the exceedances of the varying threshold given in equation (3.3.2) is then, for $y > 0$

$$\Pr(Y_t \geq y + u(\mathbf{x}_t) | Y_t > u(\mathbf{x}_t), \mathbf{X}_t = \mathbf{x}_t) = \left[1 + \frac{\xi_z \left\{[y + u(\mathbf{x}_t)]^{\lambda(\mathbf{x}_t)} - u(\mathbf{x}_t)^{\lambda(\mathbf{x}_t)}\right\}}{\psi_{z,u}\lambda(\mathbf{x}_t)\sigma(\mathbf{x}_t)}\right]_+^{-1/\xi_z}. \tag{3.3.3}$$

For general $\lambda(\mathbf{x}_t)$, this is not a GPD and so any attempt to model the exceedances $Y_t - u(\mathbf{x}_t) | Y_t > u(\mathbf{x}_t)$ using a GPD model is likely to result in a poor fit. Suppose that the Box-Cox parameter $\lambda(\mathbf{x}_t)$ is equal to 1; in this case equation (3.3.3) simplifies to a GPD with shape parameter ξ_z and scale parameter $\psi_{z,u}\sigma(\mathbf{x}_t)$. Now $\sigma(\mathbf{x}_t)$ needs to be estimated for both varying threshold and pre-processing methods, however we can see that the pre-processing method will give the more efficient estimate of $\sigma(\mathbf{x}_t)$ as it uses all the data $\{Y_t\}$, not only those $\{Y_t\}$ which are exceedances of $u(\mathbf{x}_t)$.

3.4 Theoretical and simulation study

To avoid over-complicating matters we do not consider the varying threshold method in this section, preferring to compare the standard method with the full proposed alternative; we shall return to the varying threshold approach in Section 3.5. We first illustrate the increased efficiency of the pre-processing method over the standard method and then show, under the assumption that the correct form of the covariate-response relationship is known, that the pre-processing method is more likely to select the model with the correct covariates than the standard method.

The non-stationary process $\{Y_t\}$ is obtained under the location-scale transformation

$$Y_t = \mu(X_t) + \sigma(X_t)Z_t \tag{3.4.1}$$

where the location and scale parameters $\mu(X_t)$ and $\log \sigma(X_t)$ are functions of a time-varying covariate X_t and $\{Z_t\}$ is an IID sequence of random variables with Gumbel marginal distribution and a scale parameter k. By varying k we assess the impact of the signal to noise ratio on each of the methods. The distribution function of $Y_t|X_t$ is

$$\Pr(Y_t \leq y|X_t) = \exp\left\{-\exp\left[-\left(\frac{y-\mu(X_t)}{k\sigma(X_t)}\right)\right]\right\}, \quad -\infty < y < \infty. \tag{3.4.2}$$

Further the upper tail of the distribution of $Y_t|X_t$ converges asymptotically to an Exponential($k\sigma(X_t)$) distribution, as $u \to \infty$

$$\Pr(Y_t > y + u|Y_t > u, X_t) \sim \exp\left\{-\frac{y}{k\sigma(X_t)}\right\}, \quad y > 0. \tag{3.4.3}$$

We consider two models for each of the parameters $\mu(X_t)$ and $\sigma(X_t)$, each containing either a linear or a cyclic trend, with coefficients $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$. The cyclic trend is given by a first-order Fourier series;

1. $\mu(X_t) = \mu_0 + \mu_1 \frac{t}{n+1}$, $\sigma(X_t) = 1$,

2. $\mu(X_t) = 0$, $\log \sigma(X_t) = \sigma_0 + \sigma_1 \frac{t}{n+1}$,

3. $\mu(X_t) = \mu_0 + \mu_1 \cos(\frac{2\pi t}{N}) - \mu_2 \sin(\frac{2\pi t}{N})$, $\sigma(X_t) = 1$,

4. $\mu(X_t) = 0$, $\log \sigma(X_t) = \sigma_0 + \sigma_1 \cos(\frac{2\pi t}{N}) - \sigma_2 \sin(\frac{2\pi t}{N})$

where n in the total number of observations and N is the number of observations in each of the cycles generated by the Fourier series.

3.4.1 Efficiency

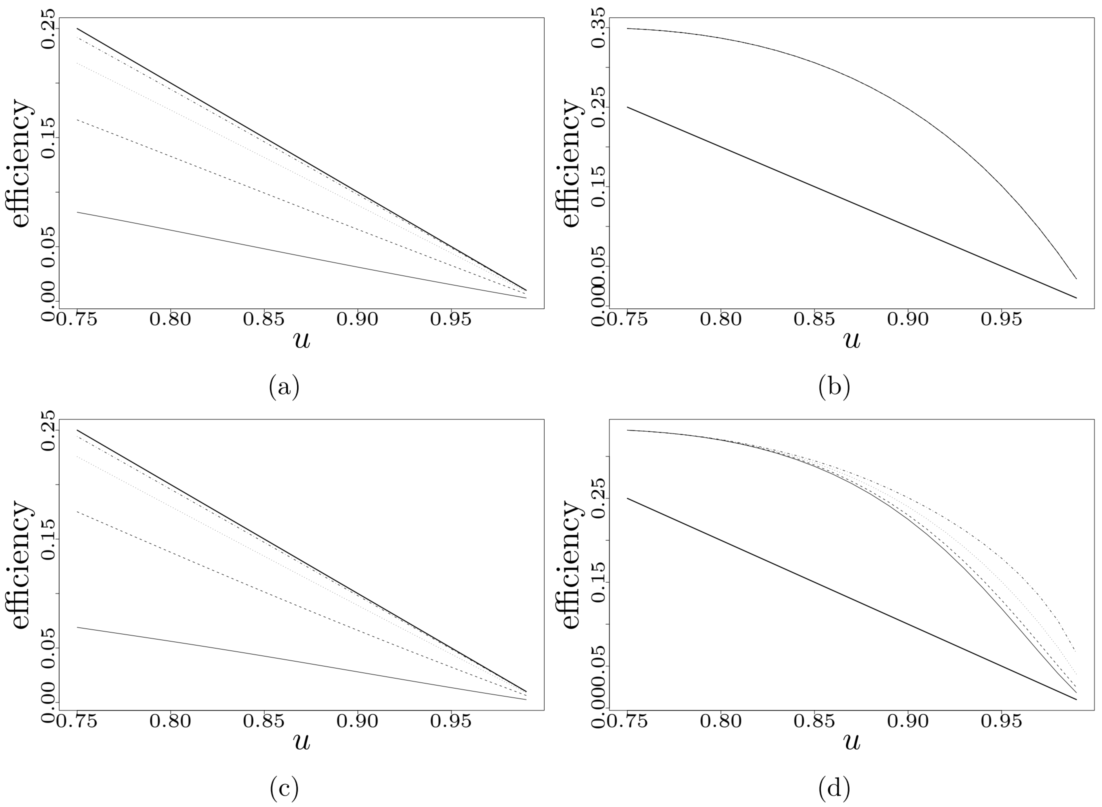
In showing the decrease in efficiency caused by using the standard rather than pre-processing method we use the exact probability of exceeding u given by equation (3.4.2) rather than estimating the rate parameter $\phi_u(X_t)$. Also, following the asymptotic result in equation (3.4.3) we model the threshold exceedances using an Exponential distribution, *i.e.* a GPD with $\xi(X_t) = 0$. The likelihood for this model under the standard method is

$$L_0(\boldsymbol{\mu}, \boldsymbol{\sigma}, k) = \prod_{t=1}^{n} [1 - \Pr(Y_t > u|X_t)]^{1-I[y_t>u]} \left[\Pr(Y_t > u|X_t)\frac{1}{k\sigma(x_t)} \exp\left\{-\left(\frac{y_t - u}{k\sigma(x_t)}\right)\right\}\right]^{I[y_t>u]} \tag{3.4.4}$$

By construction, the only trends in the process are through the mean or variance. To estimate the covariate coefficients using the pre-processing method we should therefore only need to estimate the location and scale parameters by fitting the regression model with likelihood

$$L_1(\boldsymbol{\mu}, \boldsymbol{\sigma}, k) = \prod_{t=1}^{n} \frac{1}{k\sigma(x_t)} \exp\left\{-\left(\frac{y_t - \mu(x_t)}{k\sigma(x_t)}\right)\right\} \exp\left\{-\exp\left[-\left(\frac{y_t - \mu(x_t)}{k\sigma(x_t)}\right)\right]\right\}. \tag{3.4.5}$$

Efficiency here is measured as the ratio of the asymptotic variances of the MLE for the trend parameter under the pre-processing method to that under the standard method. The required variances can be obtained from the inverse of the expected information matrix, details of the calculations of these are in Appendix A. Figure 3.2 shows efficiency results for each of the four models for a range of Gumbel scale parameters k and a range of thresholds. In all cases the efficiency of the standard method compared to the pre-processing method decreases towards zero as the threshold increases. The efficiency gain is less when there is non-stationarity in the scale than when there is non-stationarity in the location.



(a) (b)

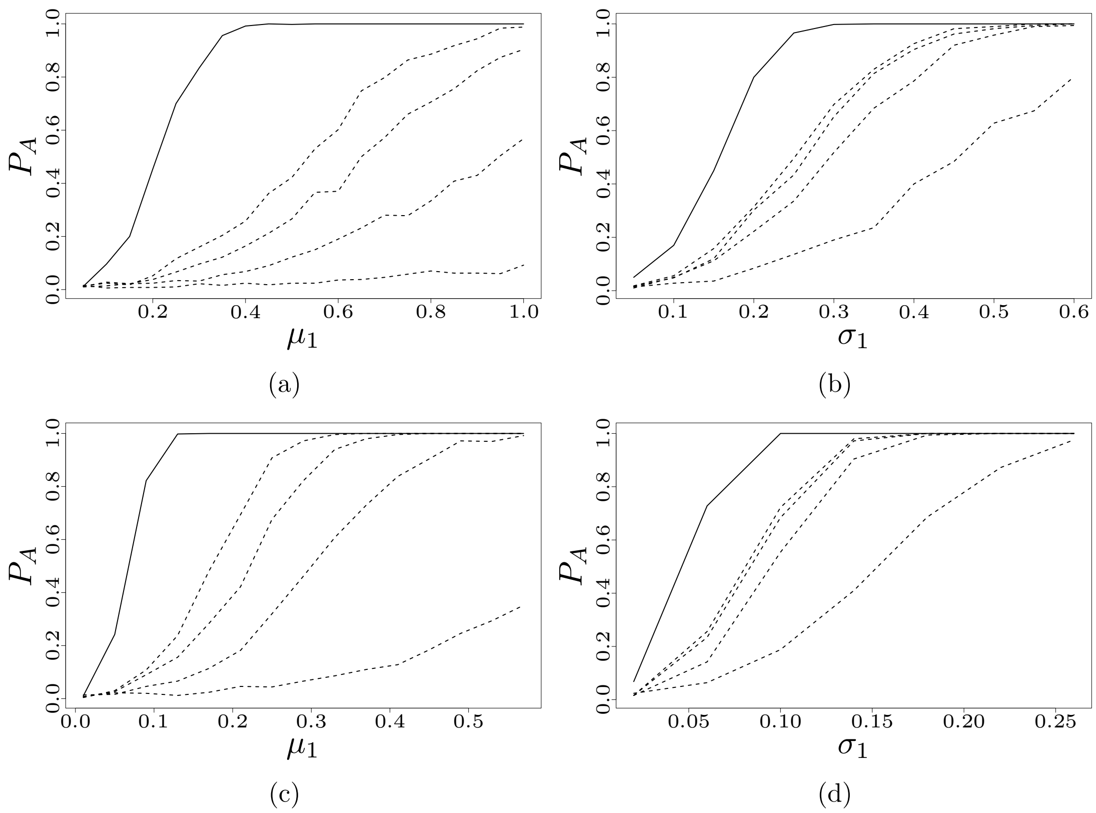
(c) (d)

Figure 3.2: Efficiency of standard method compared to pre-processing method. Efficiency is shown for the time covariate coefficient in the linear trend models (a) and (b) and for the coefficient of the cosine term in the cyclic models (c) and (d) for thresholds between the 75- and 99% quantiles. Four values of the Gumbel scale parameter k are shown, $k = 0.5$ (full line), $k = 1$ (dashed line), $k = 2$ (dotted line) and $k = 5$ (dash-dot line). The thick line shows the proportion of the full data set exceeding the threshold.

Increasing the Gumbel scale parameter k (equivalently decreasing the signal to noise ratio) increases the efficiency of the standard approach, except in the case of a linear trend in the scale where there is no change in the efficiency. For both location trends the maximum efficiency of the standard method seems to tend to the proportion of data exceeding the threshold, as k increases. The efficiency of the standard method relative to the pre-processing method is greater when the trend is observed in the scale parameter than when it is observed in the location. The reason for this difference is that the scale parameter appears in all parts of the standard method likelihood (3.4.4), whereas the location parameter contributes to the rate part only.

3.4.2 Model Selection

Evaluating efficiency as above assumes that the correct covariate model has been selected. We next consider the likelihood of this happening under the two methods. Given a data set we use the likelihoods given in equations (3.4.4) and (3.4.5) to fit both the null model, with no covariates, and the correct covariate model, under both approaches. We use the likelihood ratio statistic to decide whether or not to accept the correct model.



(a)

(b)

(c)

(d)

Figure 3.3: Proportion of covariate models selected correctly, P_A, under standard (dashed lines) and pre-processing (full lines) methods. 85-, 90-, 95- and 99% thresholds were used for the standard method (left to right in each plot). Models are (a) linear trend in mean, (b) linear trend in scale, (c) cyclic trend in mean and (d) cyclic trend in scale. Gumbel scale parameter is $k = 1$ in all cases. Proportions were estimates using 500 simulated data sets.

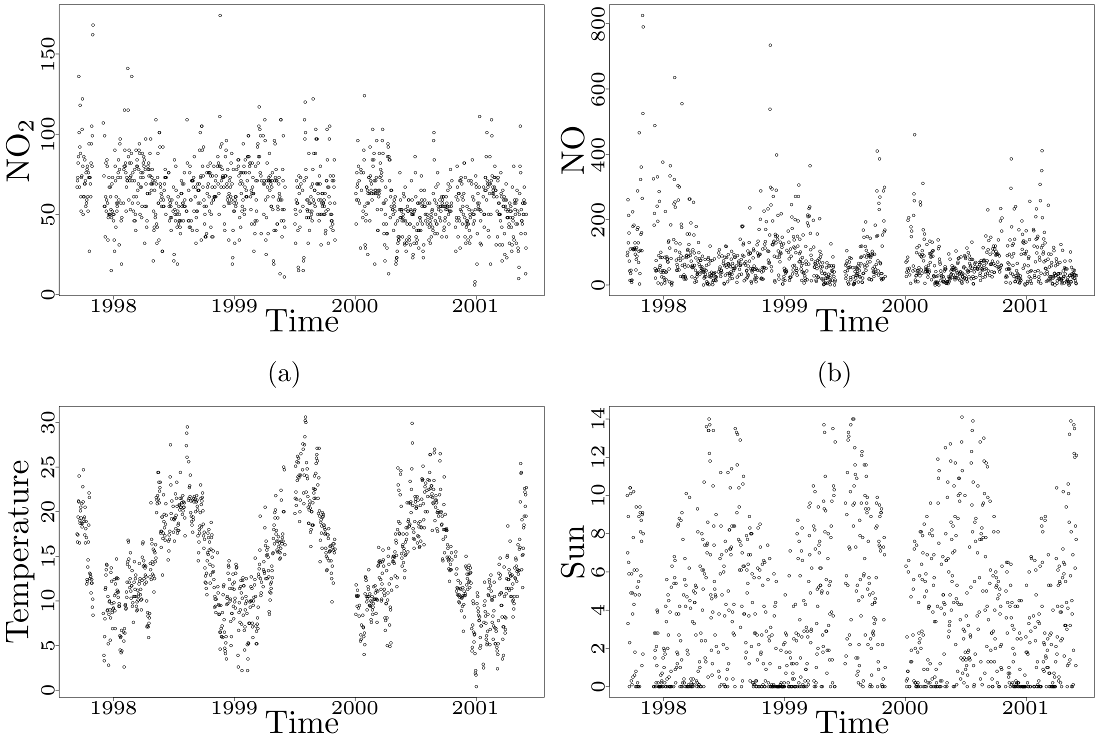
Figure 3.3 shows results for each of the four models considered. In the linear models we consider a range of values for the coefficient of the covariate. For the cyclic models we always take the coefficient of the sine term to be the negative of the cosine term and so we just vary the value of this coefficient. For each model and set of parameter values we simulated 500 data sets of length 1825 (equivalently 5 years of data) and calculated the proportion of these for which the correct model was selected. For the standard method we considered 85-, 90-, 95- and 99% thresholds. These plots clearly show that in all cases the pre-processing method has a higher probability of picking out the correct model, especially for very low values of the trend coefficients. This seems to confirm our intuition that a multiple regression model under the pre-processing method is more likely to correctly identify response-covariate relationships than a multiple regression model under the standard method.

3.5 Ozone data Analysis

3.5.1 Background

We now discuss various methods of analysing the ozone data set shown in Figure 3.1. Throughout, we assume that any missing data is missing at random (for example due to machine failure) and is therefore non-informative. We begin with a naive approach and assume that the data are stationary. Standard diagnostic plots, for example, mean residual life and threshold-shape plots (Coles, 2001) suggest a 90% quantile threshold, $u = 100$, should be sufficient and the QQ plot shown in Figure 3.9(a) shows that the GPD fitted to the exceedances of this threshold fits reasonably well.

However, the data clearly do not satisfy the assumption of stationarity, so that, although the model appears to fit the observations well given the QQ plot in Figure 3.9(a), we would not trust it in making predictions. Further, the stationary model is of no use in helping us to estimate trends in the data since it does not allow us to build in the known physical mechanisms involved in ozone generation. Neither can we use it to make predictions of extreme ozone levels under any forecast changes in the variables involved in these underlying mechanisms; for example, we might be especially interested in the likely impact on ozone levels of various climate change scenarios. To address these issues we instead fit a model with covariates following each of the three methods presented in Sections 3.2.2 and 3.3.



(a)

(b)

(c)

(d)

Figure 3.4: Daily maxima of (a) NO$_2$ (μgm^{-3}), (b) NO (μgm^{-3}) and (c) temperature (°C). Also (d) daily aggregate sunshine (hours).

The precursor chemicals involved in the production of ozone are well known, and it is further known that this process is dependent on meteorological conditions (see Section 3.1). Selection of potential covariates should be driven by this information. As potential covariates in this study we have maximum daily measurements of two precursors, NO and NO$_2$, and two meteorological variables, temperature (daily maxima) and sunshine (daily aggregate), as shown in Figure 3.4. Since they are likely to be related, we allow a first order interaction ($\times$) between temperature and sunshine. The meteorological covariates, obtained from the UK Meteorological Office, come from a site located 2km away from the air pollution monitoring site, this is close enough to be considered representative of conditions at the air pollution site. As additional covariates, we consider indicator functions for each year and for each season, defined as winter (December-February), spring

(March-May), summer (June-August) and autumn (September-November). The yearly indicators show whether the ozone levels display any long-term trends over and above that which is accounted for by trends in the covariates; and thus allow for more subtle trends than linearity. The seasonal indicators allow for any seasonal trend due to missing covariates, such as volatile organic compounds (VOC's), road traffic indicators or proximity to point sources.

Standard threshold diagnostic plots are no longer informative for non-stationary data, instead one can attempt to fit the covariate models over a range of thresholds and look for consistencies in fit. However, using the standard method, we experienced difficulties in fitting covariate GPD models to the exceedances of a range of thresholds, since the numerical routine used to maximise the likelihood frequently failed to converge without a great deal of tuning. This problem did not occur when the pre-processing and varying threshold methods were used. We show results using the 90% threshold for all methods. This guarantees the same number of exceedances are used in each method, thus ensuring a fair comparison of the methods. Both constant and varying 90% thresholds are shown in Figure 3.5(a).

The likelihoods used for model fitting, see equations (3.2.2) and (3.2.5), require the assumption that the data are independent, which in practice is unlikely to be the case. One way to account for any dependence is to use all the data to obtain point estimates for the parameters, but to then use a block bootstrap scheme to estimate confidence intervals for the estimates. The scheme that we propose involves resampling the pre-processed series $\{Z_t\}$, assuming that this is approximately stationary. For this example, a block length of 5 days was chosen to minimise dependence between blocks as this period is more than sufficient to ensure independence between pollutants, since molecules of ozone, NO and NO_2 react within minutes of being present in the atmosphere, but importantly was long enough to ensure independence between the meteorological variables, since climate events may last several days. The re-sampled series is then back-transformed to the original scale using the parameters fitted to the original data. The resulting bootstrapped sample can be modelled using either of the standard, pre-processing or varying threshold methods and the results used to obtain a sampling distribution of the model parameters or return levels under each method.

3.5.2 Results

Selection of the actual covariate model is non-trivial, since the mechanisms controlling the ozone process are themselves exceedingly complex (Thompson *et al.*, 2001). We want a model with a minimum number of covariates which reflects the scientific understanding of the process and well represents the data. We selected the covariates in stages; for example, for the Box-Cox parameter $\lambda(\mathbf{x}_t)$, we compared models with $\lambda(\mathbf{x}_t) = 1$, $\lambda(\mathbf{x}_t) = \lambda$ and $\lambda(\mathbf{x}_t) = \lambda(\mathbf{x}_t)$. Similarly we first attempted to fit a model without the yearly and seasonal indicators. Finally, we used a mixture of forward and backward selection with a significance level of 1% to decide which covariates to include in each of the parameters, see McCullagh and Nelder (1989). In the GPD models we follow a standard procedure and fix the shape parameter as constant, since the amount of information required to estimate this well as a function of covariates is too great. We do not claim that the model given below is the definitive model for surface-level ozone concentrations, merely that it is one that seems plausible; from both a scientific and statistical perspective.

Following the exploratory analysis described above, we chose to fix the Box-Cox parameter at $\lambda = 0.5$, since this maximised the profile likelihood for λ across a range of interpretable values (*e.g.* $\lambda = -0.5,\ 0,\ 0.5$ and 1). Taking $\lambda = 0.5$ we model the mean and scale in the pre-processing model as functions of the square roots of NO and NO_2, since the relationship between these and the square root of ozone seemed closer to linearity than that between the square root of ozone and both NO and NO_2 on their observed scale. Further, it is standard procedure to model chemicals on comparable scales; hence for the standard and varying threshold methods, where we model ozone on the observed scale we also retain NO and NO_2 on their observed scales.

	Scientific		Data-based	
	$\mu(\mathbf{x}_t)$	$\log\sigma(\mathbf{x}_t)$	$\mu(\mathbf{x}_t)$	$\log\sigma(\mathbf{x}_t)$
Constant	7.63	-0.404	7.70	-0.366
$\sqrt{\text{NO}}$	-0.232	0.0443	-0.256	0.0520
$\sqrt{\text{NO}_2}$	0.148		0.211	
Temp	0.00748	0.0130		
Sun	-0.00668	-0.0345	0.0949	
Temp×Sun	0.00685	0.00226	N/A	N/A
I[1998]	0.724		0.702	
I[1999]	0.958		0.571	
I[2000]	0.306	-0.124		
I[Spring]	0.587		N/A	N/A
$\sqrt{\text{NO}}$×I[Summer2]	N/A	N/A		-0.0335
Temp×I[Summer2]	N/A	N/A		0.0199
I[Summer2,1999]	N/A	N/A	0.466	-0.326
Sun×I[Summer2,1999]	N/A	N/A	0.144	0.0618

Table 3.1: Maximum likelihood estimates of significant covariates in location and scale parameters for the two pre-processing models. N/A refers to covariates not fitted in that model, as opposed to blank entries which show covariates which were not significant.

The MLE's of the best fitting location-scale parameters selected under this procedure are shown in Table 3.1, under the heading 'Scientific'. The MLE's for the best fitting GPD and rate parameters are, using the standard method

$$\begin{aligned}
\text{logit}\ \phi_u(\mathbf{x}_t) &= -8.81 - 0.0127\text{NO} + 0.0881\text{Temp} + 0.0433\text{Sun} \\
&\quad +0.0127\text{Temp} \times \text{Sun} + 3.51\text{I}[1998] + 4.44\text{I}[1999] \\
&\quad +2.89\text{I}[2000] + 1.60\text{I}[\text{Spring}] \\
\log\psi_u(\mathbf{x}_t) &= 1.27 + 0.0647\text{Temp} + 0.0636\text{Sun} + 0.541\text{I}[1999] \\
\xi &= -0.438,
\end{aligned}$$

using the pre-processing method,

$$\phi_{z,u}(\mathbf{x}_t) = 0.100, \quad \psi_{z,u}(\mathbf{x}_t) = 0.510, \quad \xi_z = -0.227,$$

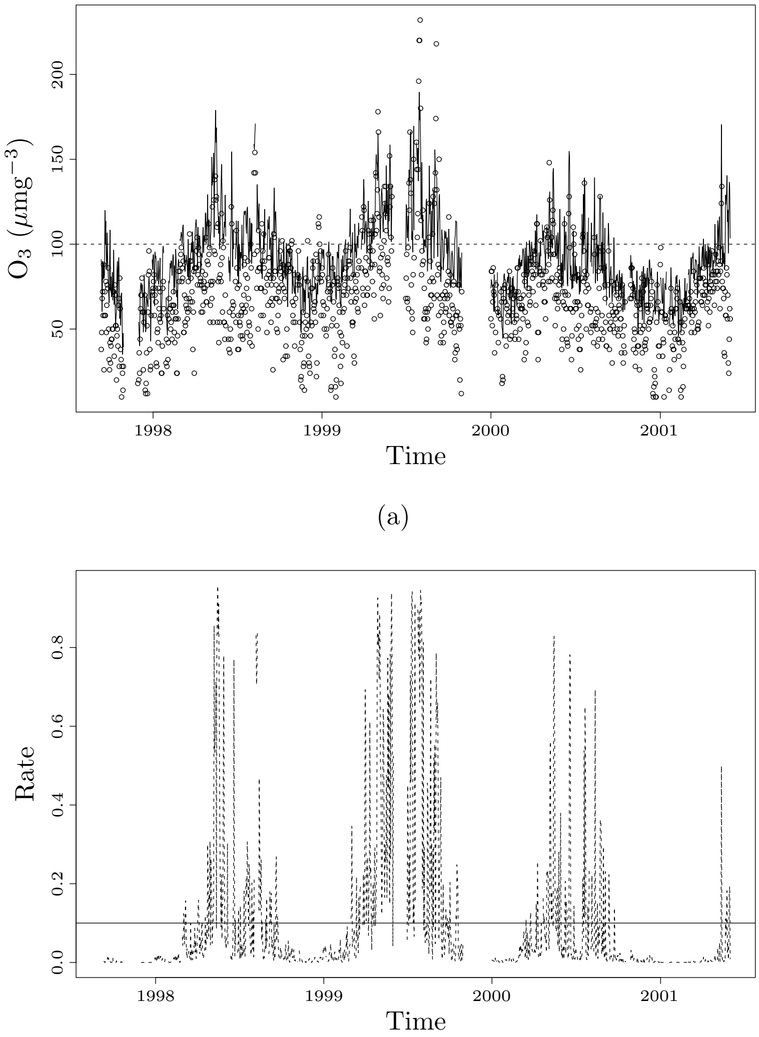
and using the varying threshold method,

$$\phi_{u(x_t)}(\mathbf{x}_t) = 0.100, \quad \log \psi_{u(x_t)}(\mathbf{x}_t) = 1.74 + 0.0474\text{Temp} + 0.353\text{I}[1999], \quad \xi = -0.279.$$

The results of the pre-processing model fit shown in Table 3.1 confirm that each of the precursors and meteorological covariates are important in describing the ozone process. For example, from the fitted location parameter, we see that high NO levels correspond to low ozone levels (plots confirm that NO tends to peak in winter, when ozone levels are at there lowest), whereas there is some positive relationship between NO_2 and ozone. Similarly increases in both temperature and the interaction between sunshine and temperature lead to higher ozone levels. Ozone levels seem to be higher, on average and given the values of the precursors and meteorological covariates, in the years 1998, 1999 and 2000, with the greatest increase in 1999, and also during the spring (March-May). This is likely to be due to the presence of some missing covariates, such as VOC's or traffic volume. Note that the standard method doesn't pick up all the covariate relationships found using the pre-processing method; for example the level of NO_2 is not significant in the standard model.

The functional forms of the rate and scale parameters found using the pre-processing and varying threshold methods are much simpler than those found using the standard method. Specifically, for the pre-processing method, there is no evidence of any covariate effects in either the rate or scale parameters. Results, not shown here, suggest that these findings hold for a range of thresholds. A consequence of the simplicity of the GPD model for the pre-processing method is that, under the stationarity assumption, threshold choice can be improved by using standard methods (Coles, 2001).

Figure 3.5(b) shows a plot of the estimated rate parameter for the standard method. Compared to the constant rate parameter for the other methods, this shows considerable variation. Specifically, over the summer periods, the probability of observing an exceedance under the standard method is extremely high, at least



(a)

(b)

Figure 3.5: (a) Reading ozone data with constant (dashed line) and varying (full line) 90% thresholds and (b) estimated rate parameters for the exceedances of the constant (dashed line) and varying (full horizontal line) 90% thresholds.

50% for most days; suggesting that these observations are not extreme at all. In contrast, the pre-processing method has a higher threshold over the same periods that the standard method has an increased rate parameter and so by accounting for the underlying mechanisms in determining the threshold the pre-processing method is ensured a constant rate of exceedance. PP (not shown) and QQ plots (shown for the pre-processing method only in Figure 3.6) suggest that all three models fit the exceedances reasonably well.

We now compare the ability of each of the models to predict return levels.

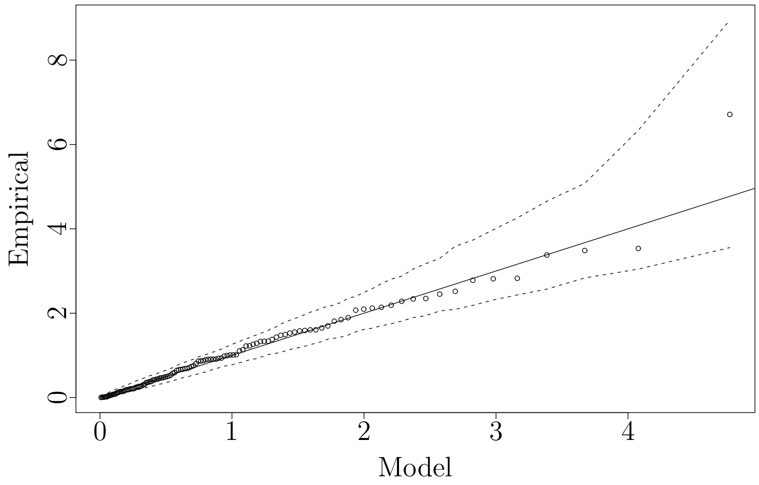
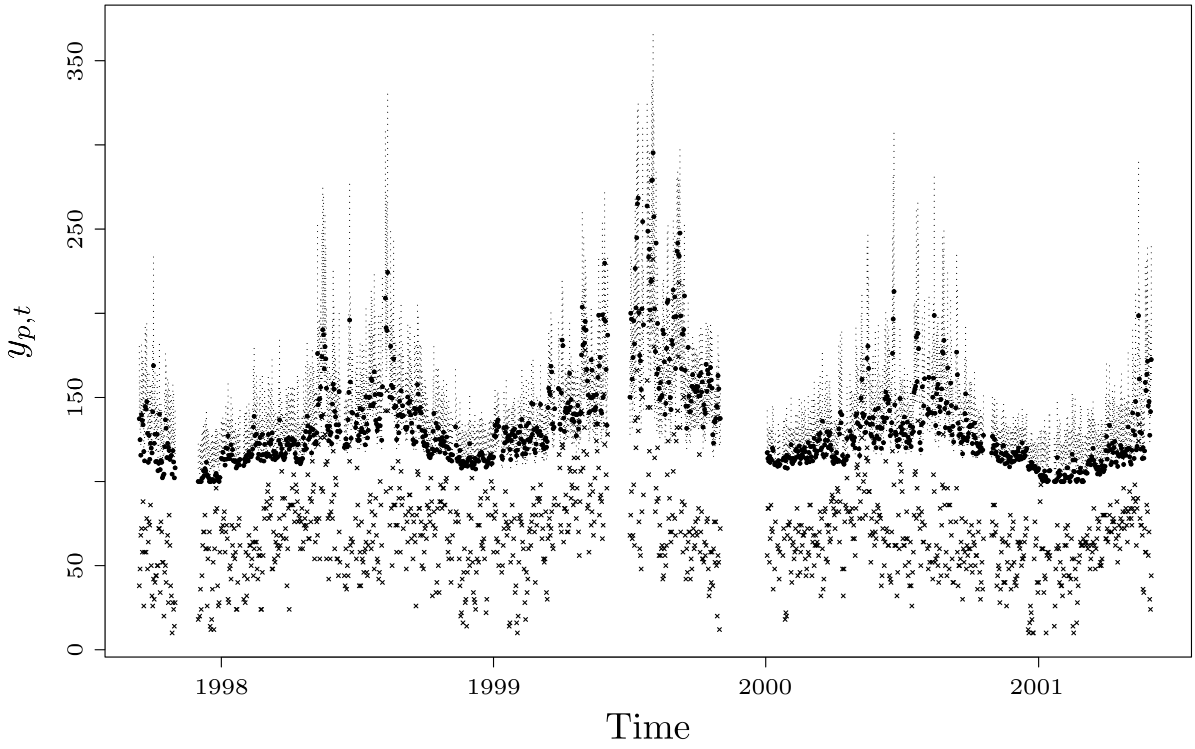


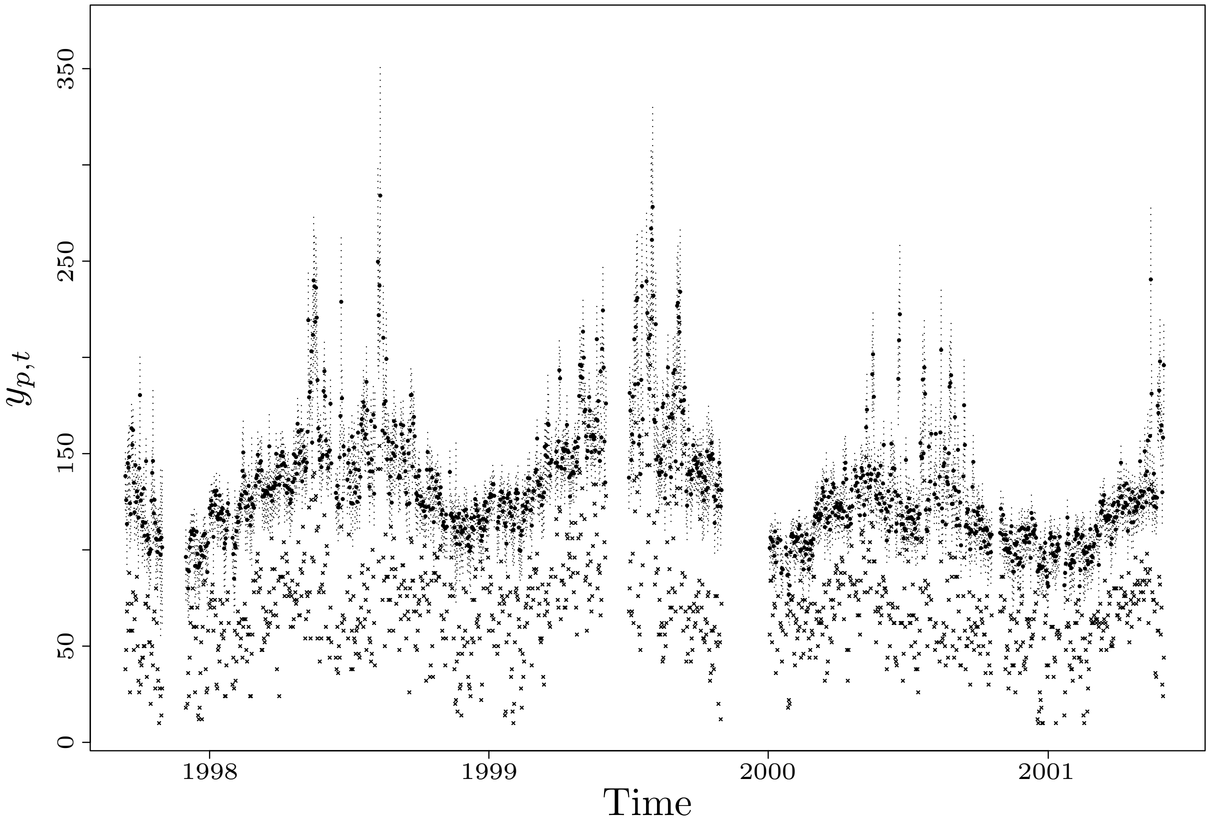
Figure 3.6: QQ plot to show the goodness of fit of the GPD model for the 90% threshold exceedances fitted using the pre-processing method. The plot is shown on the standard exponential scale.

Recall that, for the standard method, if $\phi_u(\mathbf{x_t}) < p$ we know only that $y_{p,t} \leq u$. In this case, by taking $y_{p,t} = u$ we obtain falsely high point estimates and falsely narrow confidence intervals. However by choosing p small enough we minimise the occurrence of this. We look at the conditional return level $y_{p,t}$ where $p = (365n)^{-1}$; if identical values for $\mathbf{x}_t$ were observed each day for n years we would expect $y_{p,t}$ to be exceeded once.

Figure 3.7 shows the 10-year conditional return levels. The plots show that the estimates using either the pre-processing or varying threshold approach follow the pattern of the observed data more closely than the estimates made under the standard approach. During the summer the standard method seems to under-estimate the return levels, relative to either of the other methods, whereas during the winter it over-estimates the return levels, relative to the other methods. Further, many of the point estimates from the standard method fall just outside the 95% confidence intervals, especially during the winter months. This suggests that the GPD model with parameters as functions of covariates might not be a good model for some vectors of covariates, especially for those covariates for which the associated ozone level was a non-exceedance. Figure 3.8 shows boxplots of the 95% confidence interval widths for the three methods. Even taking into account the falsely

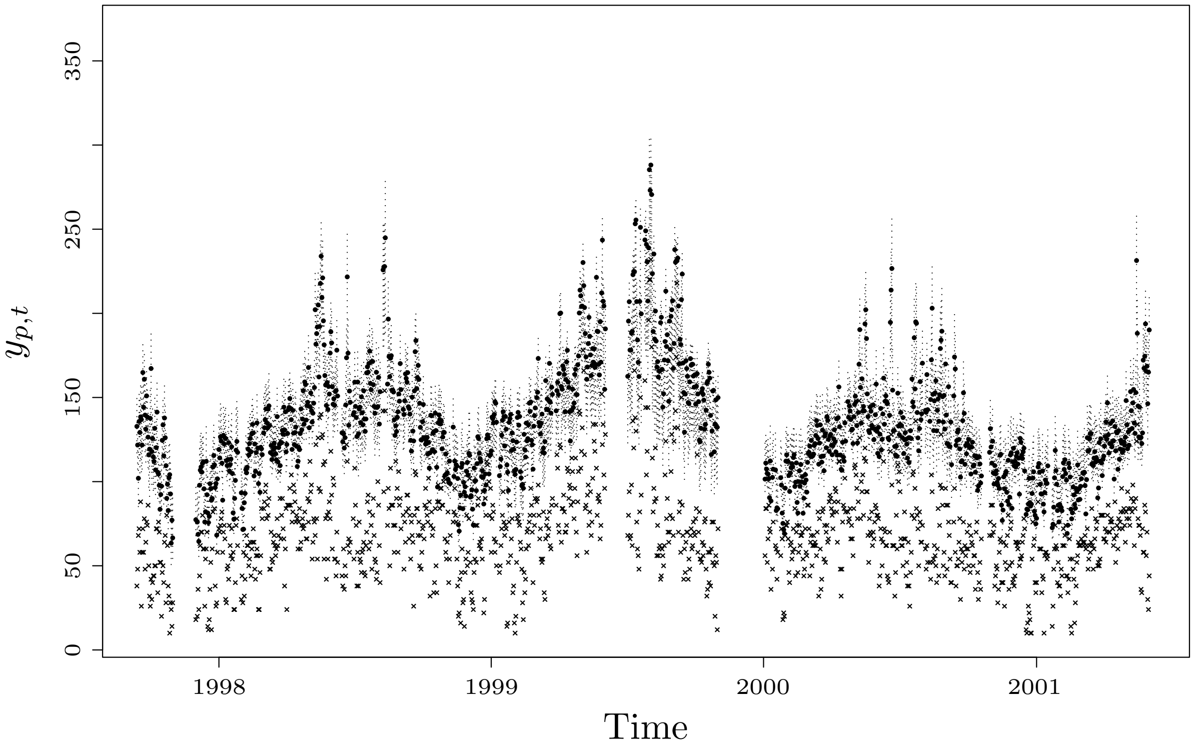


(a)



(b)

Figure 3.7: 10-year conditional return levels, point estimates (dots) and 95% confidence intervals (vertical lines), for (a) standard, (b) pre-processing (scientific model) and (c) varying threshold methods. Exceedances are shown by crosses.



(c)

Figure 3.7 (continued)

narrow confidence intervals for some of the estimates under the standard method the scientific pre-processing model seems to have narrower confidence intervals.

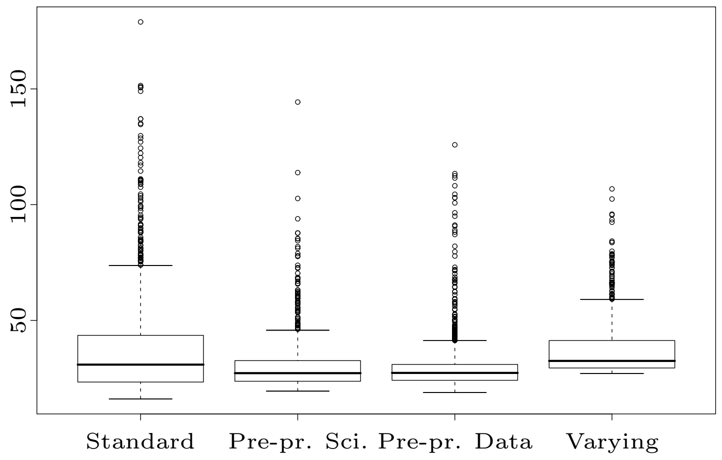


Figure 3.8: Box plots to summarise the 95% confidence interval widths for 10-year conditional return levels estimated under the standard, pre processing (scientific and data-based models) and varying threshold methods.

3.5.3 Covariate mis-specification

All three methods for modelling non-stationary extremes are susceptible to missing or mis-specified covariates. To understand the effect that missing covariates might have on the model output, we fit a second pre-processing model to the ozone data using *data-based* covariates; mostly this means using indicator functions to mimic unobserved covariates. The key to this model is that there were unusually high levels of ozone in 1999 which are not explained by any of the available precursors and meteorological covariates (see Figures 3.1 and 3.4). Exploratory analyses (not shown) suggested that the best way to model this was using a two season model with summer defined to be the period April-September. We then allow second-order interactions between season and all precursors and meteorological covariates and yearly indicators, as well as third order interactions between season, the yearly indicator for 1999 and the physical covariates. Significant covariates were selected in the a similar way as for the previous models.

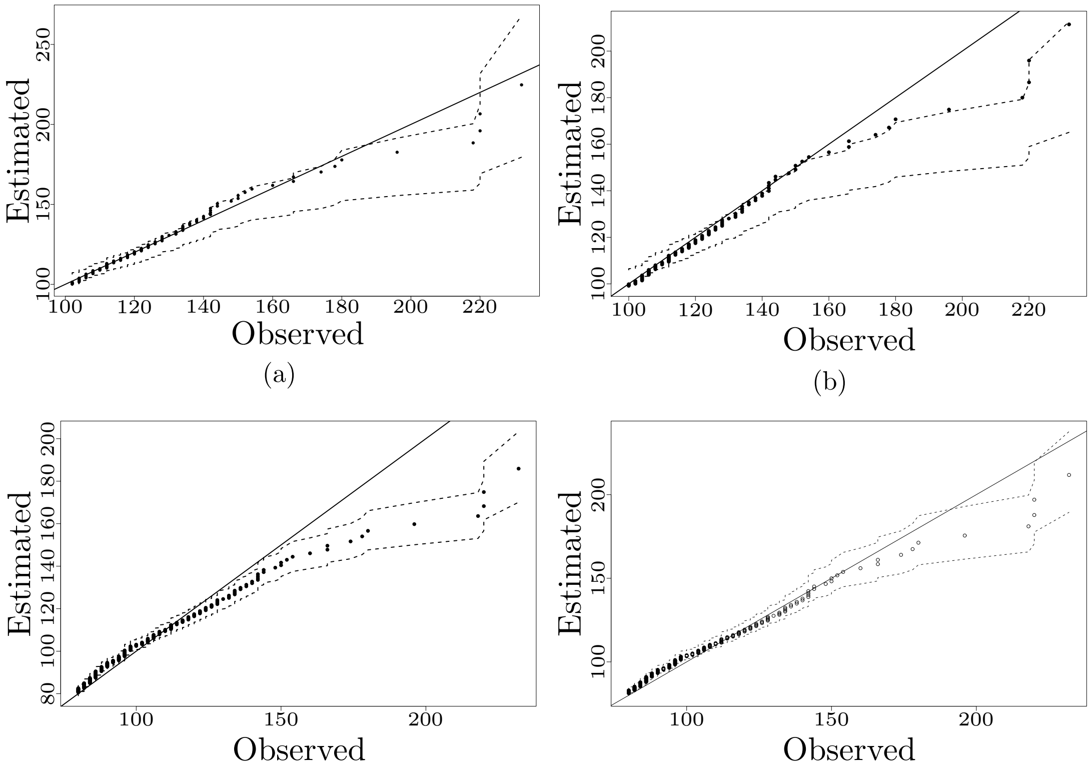
We follow the scientific model and select $\lambda = 0.5$. Then the MLE's for the location and scale parameters under this new model are shown in Table 3.1, under the heading 'Data-based'. This model has fewer covariates than its scientific counterpart, but it is less interpretable; for example, there is no reason that the positive relationship between sunshine and mean ozone level should have increased in the summer of 1999 or that the variance should also have decreased at this time, other than that these covariates are acting as dummy variables for missing covariates that actually caused these changes in mean and variance. The MLE's for the tail parameters under this model are

$$\phi_{z,u}(\mathbf{x}_t) = 0.100, \quad \psi_{z,u}(\mathbf{x}_t) = 0.532, \quad \xi_z = -0.293.$$

There are a number of similarities between the two pre-processing models; primarily the series $\{Z_t\}$ is close to stationarity regardless of which set of covariates is used to model $\mu(\mathbf{x}_t)$ and $\sigma(\mathbf{x}_t)$ and so it is not necessary to fit covariates in the tail parameters in either case. Specifically, the rate parameter is constant, whereas we found that when using the standard method the rate parameter varied considerably through time, even when data-based covariates are used.

The point estimates of the tail parameters of the two pre-processing models are very similar, especially when estimation uncertainty is taken into account. Many of the covariates common to both location or scale parameters also have similar coefficients; for example, from Table 3.1, NO has a similarly sized positive effect on the mean level under both models.

We also estimate 10-year conditional return levels, and associated 95% confidence intervals, for this data-based pre-processing model. We found that the largest differences between the point estimates from the two pre-processing models occur during the summer, when the scientific model over-estimates the return levels compared to the data-based model. Plots (not shown) of a similar comparison between standard models fitted using the scientific and data-based covariates show similar differences. However, the conditional return levels estimated from the pre-processing models show a much closer fit to the data than estimates from models fitted using the standard method, regardless of the covariates used. We found little difference between the uncertainty associated with the estimates from different covariate models from a particular method (see the box plot of confidence interval widths in Figure 3.8 for a comparison of the different models under the pre-processing method), but there was consistently lower uncertainty from the pre-processing method than from the standard method, regardless of the covariates used.



(a)

(b)

(c)

(d)

Figure 3.9: QQ plots of estimated marginal return levels against observed data using models fitted using the (a) stationary, (b) standard and (c) and (d) pre-processing methods. Plot (c) shows the scientific and (d) the data-based pre-processing models. Dotted lines show 95% bootstrapped confidence intervals. Perfect fit is identified by the 45° line. Plots (a) and (b) shows only the top 10% of the observed data, whereas plots (c) and (d) show the top 30%.

Finally we consider the estimation of the marginal return levels for the observed period and under future covariate conditions. For the observed period, marginal return levels estimated using the stationary GPD, standard and both pre-processing models are compared in the form of QQ-plots in Figure 3.9. Because we consider the observed period only we can take the empirical distribution as an estimate of the joint distribution of the covariates. These plots appear to show that there are considerable similarities between the different estimation methods in terms of marginal return levels over this period. Figure 3.9 shows that using the pre-processing method we can make estimates below the threshold as we have estimates below 100. We have found that the marginal return levels were particularly difficult to estimate, under both standard and pre-processing methods, with the best estimates coming from the data-driven, rather than scientifically motivated, covariate models; this is confirmed by the plots in Figure 3.9.

The QQ plots in Figure 3.9 suggest that the non-stationary models lead to estimated marginal return levels which show a poorer fit to the data than those estimated under the stationary model. This is probably due to the use of the empirical joint distribution for the covariates, see equation (3.1.3). By modelling the covariates, especially in the tails, we might expect better estimates. The idea of modelling covariates is explored to some extent in Chapter 4, where marginal return levels are obtained by simulation, but further work is required on this subject. Despite the difficulty in estimating marginal return levels, a non-stationary model may still be preferable to a stationary one since it allows us to use modelled response-covariate relationships to make predictions, especially of conditional return levels, given future covariate values. Clearly such prediction does rely on the usual assumption that the covariate-response relationships hold for the new covariate values.

If interest is in future covariate scenarios we believe that there are distinct advantages in the pre-processing scientific model. The stationary GPD model is clearly inappropriate if the covariate scenario is much different from that observed. The superior ability of the pre-processing method relative to the standard method to capture the under-lying data generating mechanisms is likely to lead to improved marginal return level estimation under a greater range of covariate scenarios. Further, because the covariates in the pre-processing data-based model were determined specifically to fit the data across the time period of observation (*e.g.* using the interactions between the indicator variables for years and seasons and the other covariates) we could not use this model to estimate marginal return levels for future covariate scenarios. The pre-processing scientific model does contain yearly indicators, which were included in an attempt to identify trends, clearly some assumption is required to extend this inference to future covariate scenarios.

3.6 Discussion and comparison

The pre-processing method for modelling extremes of non-stationary processes introduced in this paper seems to show several improvements on alternative methods. First we discuss the varying threshold approach discussed in Section 3.3.2; this can be seen as an extension of splitting data into 'seasons' and separately modelling the data within seasons. The disadvantage of this approach compared to the full pre-processing approach is that, because the excesses are modelled on the original scale, it is impossible to distinguish the covariate effects found in the GPD parameters between those which affect the centre of the distribution and those which affect the tails. This also implies that there are likely to be more significant covariates in the GPD parameters with less data from which to estimate their form, suggesting a numerically complex model-fitting situation. Results from the ozone data analysis in Section 3.5 also suggest that there is a greater uncertainty in predictions made under this approach compared to those made under the pre-processing approach.

We now compare the pre-processing and standard method. Primarily, the pre-processing approach is better because the reasons for non-stationarity in a data set are often intricately tied up with the mechanisms generating the underlying process, so modelling non-stationarity in the underlying process is more likely to capture the appropriate form of non-stationarity. In some contexts a model for the underlying process may already exist. The simulation study in Section 3.4 confirms the benefits of exploiting the full structure of the mechanism, showing that the pre-processing method is more likely to correctly select the covariate model than the standard method. Further, the threshold exceedances under the pre-processing method are those that are extreme when we have taken into account the covariate relationship. This is not necessarily true for the standard method making the theoretical justification for the standard method seem weak.

The pre-processing method also produces a simpler and more efficient model fit; if there is no difference between the covariate effects in the body and extremes of the process $\{Y_t\}$ then the pre-processing approach which uses all the data to estimate these effects is bound to be more efficient as the rate and GPD parameters $\phi_{z,u}(\mathbf{x}_t)$, $\psi_{z,u}(\mathbf{x}_t)$ and $\xi_z(\mathbf{x}_t)$ will then be independent of the covariates. Alternately, if there is a different covariate effect in the body and extremes of $\{Y_t\}$ then the standard model confounds these whereas the pre-processing model allows separate estimation of each and so gives a clearer scientific interpretation of the covariate effects.

We believe that for hypothesis testing in the extreme value components of the models the pre-processing method changes the strategy of covariate model fitting to be scientifically rational. The standard method rejects covariates from the model if there is not significant evidence for their inclusion based on the extreme value data and in this case, as Figure 3.3 suggests, rejection of the covariate will often arise. In contrast the pre-processing method essentially tests whether there is significant evidence from the extreme values for a departure of the covariate form from that estimated using the body of the data.

The above discussion assumes that there is no mis-specification in the selected covariate form for the pre-processing stage. We anticipate that gross mis-specification would be identified by standard diagnostic tools as the pre-processed series will not appear stationary. Consider instead a small level of mis-specification. As the varying threshold method differs from the standard method only in which extreme events are selected for analysis then the varying threshold should be better as it will use data that are generally more appropriate. As the pre-processing method only differs from the varying threshold method by its choice of covariate forms, if the same covariates are available for each analysis we see no reason why the pre-processing method should perform worse than the varying threshold method.

The pre-processing approach also has the advantage of being computationally simpler since the extremes of the transformed process $\{Z_t\}$ are far closer to stationarity than those of the original process $\{Y_t\}$. It follows that threshold selection is easier in the pre-processing approach as tools for threshold selection for stationary extremes are likely to be suitable for $\{Z_t\}$ extremes but not $\{Y_t\}$ extremes. It also follows that, since it is more likely to be independent of the covariates, the GPD scale parameter is more likely to satisfy the threshold stability property discussed in Section 3.2.2.

Further work is necessary to investigate how to model the GPD parameters in the presence of covariates in order to retain the threshold stability property, so that it holds even if the pre-processing method suggests covariates are necessary. Our initial ideas for how this problem should be addressed are as follows. Smith (1989) showed that there is a connection between the generalised extreme value (GEV) distribution parameters and the GPD parameters through a general point process result for extreme values. All the parameters of the GEV distribution are threshold invariant so exploiting this link may help. Given this property, it may appear preferable to use the GEV model instead of the GPD model for the extreme value modelling in this paper. However, we feel it is important to parametrise the non-stationary models through the GPD formulation, as this leads to orthogonal parameters for the rate and excess distribution.

References

Beirlant, J., Goegebeur, Y., Segers, J. and Teugels, J. (2004). *Statistics of Extremes: Theory and Applications.* John Wiley & Sons, Chichester.

Buishand, T. A. (1993). Rainfall depth-duration-frequency curves; a problem of dependent extremes. In *Statistics for the Environment* (eds. V. Barnett and K. F. Turkman), 183-197, John Wiley & Sons, Chichester.

Chatfield, C. (2004). *The Analysis of Time Series - an Introduction*, 6th edition. Chapman and Hall, Florida.

Chavez-Demoulin, V. and Davison, A.C. (2005). Generalized additive modelling of sample extremes. *Appl. Statist.* **54**, 207-222.

Coles, S.G. (2001). *An Introduction to Statistical Modelling of Extreme Values.* Springer-Verlag, London.

Davison, A.C. and Ramesh, N.I. (2000). Local likelihood smoothing of sample extremes. *J. Roy. Statist. Soc. B.* **62**, 191-208.

Davison, A.C. and Smith, R.L. (1990). Models for exceedances over high Thresholds. *J. Roy. Statist. Soc. B.* **52**, 393-442.

de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory: an Introduction*, Springer, Berlin.

Hall, P. and Tajvidi, N. (2000). Nonparametric analysis of temporal trend when fitting parametric models to extreme-value data. *Statist. Sci.* **15**, 153-167.

Heffernan, J.E. and Tawn, J.A. (2004). A conditional approach for multivariate extreme values (with discussion). *J. Roy. Statist. Soc. B.* **66**, 497-546.

Küchenhoff, H. and Thamerus, K. (1996). Extreme value analysis of Munich air pollution data. *Env. and Ecol. Stat.* **3**, 127-141.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*, 2nd edition. Chapman and Hall, Florida.

Pickands, J. (1975). Statistical inference using extreme order statistics. *Ann. Statist.* **3**, 119-131.

Smith, R.L. (1989) Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone. *Statist. Sci.* **4**, 367-393.

Smith, R.L. and Shively, T.S. (1995). Point process approach to modeling trends in tropospheric ozone based on exceedances of a high threshold. *Atm. Env.* **29**, 3489-3499.

Smith, R. L., Tawn, J. A. and Coles, S. G. (1997). Markov chain models for threshold exceedances. *Biometrika*, **84**, 249-268.

Thompson, M.L., Reynolds, J., Cox, L.H., Guttorp, P. and Sampson, P.D. (2001). A review of the statistical methods for the meteorological adjustment of tropospheric ozone. *Atm. Env.* **35**, 617-630.

Appendix A

A.1 Simulation Study

In Chapter 3 we used a simulation study to illustrate the increased efficiency of the proposed pre-processing method for modelling non-stationary extremes over the existing method. The simulation study involves a process with known mean and variance covariate structure and is defined in equation (3.4.1). Comparison of the efficiency of the two methods requires calculation of the expected information matrices; details of these calculations are given below, first for the general process and then for the specific examples considered. We invert the expected information matrices using numerical methods to obtain the covariance matrices for the parameters.

A.1.1 Details of efficiency calculations

Given a realisation $\{Y_1, \ldots, Y_n\}$ of the process (3.4.1) the likelihood function for the existing method is given in equation (3.4.4) and for the pre-processing method in equation (3.4.5). Let l_0 be the log-likelihood function corresponding to equation (3.4.4) and write

$$z = \frac{u - \mu(X_t)}{k\sigma(X_t)} \quad \text{and} \quad r = \frac{\exp\{-2z\}\exp\{\exp(-z)}{1 - \exp\{-\exp(-z)\}}\}.$$

Then the elements of the expected information matrix for the existing approach are as follows;

$$
\begin{aligned}
E\left[-\frac{\delta^2 l_0}{\delta\mu_j\delta\mu_i}\right] &= \sum_{t=1}^{n}\left(\frac{1}{k^2\sigma^2(X_t)}r\right)\frac{\delta\mu}{\delta\mu_j}\frac{\delta\mu}{\delta\mu_i},\\
E\left[-\frac{\delta^2 l_0}{\delta\sigma_j\delta\mu_i}\right] &= \sum_{t=1}^{n}\left(\frac{u-\mu(X_t)}{k^2\sigma^3(X_t)}r\right)\frac{\delta\sigma}{\delta\sigma_j}\frac{\delta\mu}{\delta\mu_i},\\
E\left[-\frac{\delta^2 l_0}{\delta k\delta\mu_i}\right] &= \sum_{t=1}^{n}\left(\frac{u-\mu(X_t)}{k^3\sigma^2(X_t)}r\right)\frac{\delta\mu}{\delta\mu_i},\\
E\left[-\frac{\delta^2 l_0}{\delta\sigma_j\delta\sigma_i}\right] &= \sum_{t=1}^{n}\left(\frac{(u-\mu(X_t))^2}{k^2\sigma^4(X_t)}r+\frac{1-\exp\{-\exp(-z)\}}{\sigma^2(X_t)}\right)\frac{\delta\sigma}{\delta\sigma_j}\frac{\delta\sigma}{\delta\sigma_i},\\
E\left[-\frac{\delta^2 l_0}{\delta k\delta\sigma_i}\right] &= \sum_{t=1}^{n}\left(\frac{(u-\mu(X_t))^2}{k^3\sigma^3(X_t)}r+\frac{1-\exp\{-\exp(-z)\}}{k\sigma(X_t)}\right)\frac{\delta\sigma}{\delta\sigma_i},\\
E\left[-\frac{\delta^2 l_0}{\delta k^2}\right] &= \sum_{t=1}^{n}\left(\frac{(u-\mu(X_t))^2}{k^4\sigma^2(X_t)}r+\frac{1-\exp\{-\exp(-z)\}}{k^2}\right)\frac{\delta\sigma}{\delta\sigma_i}.
\end{aligned}
$$

Evaluation of the expected information matrix therefore requires a value for the threshold u. For a given probability p notice that the p^{th} quantile must satisfy

$$
\begin{aligned}
p &= \Pr(Y > u)\\
&= \int_{\mathbf{x}_t}\Pr(Y_t > u|\mathbf{X}_t=\mathbf{x}_t)\Pr(\mathbf{X}_t=\mathbf{x}_t)\,d\mathbf{x}_t\\
&\approx \frac{1}{n}\sum_{t=1}^{n}\Pr(Y_t > u|\mathbf{X}_t=\mathbf{x}_t)\\
&= \frac{1}{n}\sum_{t=1}^{n}\exp\left\{-\exp\left[-\left(\frac{u-\mu(X_t)}{\sigma(X_t)}\right)\right]\right\}
\end{aligned}
$$

where the approximation comes by using the empirical distribution for the covariates. It follows that, as $n \to \infty$ and $p \to 0$, numerical solution of this equation gives a value for u in terms of the model parameters.

Now let l_1 be the log-likelihood function corresponding to the likelihood for the pre-processing method in equation (3.4.5). The elements of the expected information matrix under this approach are as follows;

$$\begin{aligned}
E\left[-\frac{\delta^2 l_1}{\delta\mu_j\delta\mu_i}\right] &= \frac{1}{k^2}\sum_{t=1}^{n}\frac{1}{\sigma^2(X_t)}\frac{\delta\mu}{\delta\mu_j}\frac{\delta\mu}{\delta\mu_i},\\
E\left[-\frac{\delta^2 l_1}{\delta\sigma_j\delta\mu_i}\right] &= \frac{(\gamma-1)}{k}\sum_{t=1}^{n}\frac{1}{\sigma^2(X_t)}\frac{\delta\sigma}{\delta\sigma_j}\frac{\delta\mu}{\delta\mu_i},\\
E\left[-\frac{\delta^2 l_1}{\delta k\delta\mu_i}\right] &= \frac{(\gamma-1)}{k^2}\sum_{t=1}^{n}\frac{1}{\sigma^2(X_t)}\frac{\delta\mu}{\delta\mu_i},\\
E\left[-\frac{\delta^2 l_1}{\delta\sigma_j\delta\sigma_i}\right] &= (1+\frac{\pi^2}{6}+\gamma(\gamma-2))\sum_{t=1}^{n}\frac{1}{\sigma^2(X_t)}\frac{\delta\sigma}{\delta\sigma_j}\frac{\delta\sigma}{\delta\sigma_i},\\
E\left[-\frac{\delta^2 l_1}{\delta k\delta\sigma_i}\right] &= \frac{1}{k}(1+\frac{\pi^2}{6}+\gamma(\gamma-2))\sum_{t=1}^{n}\frac{1}{\sigma(X_t)}\frac{\delta\sigma}{\delta\sigma_i},\\
E\left[-\frac{\delta^2 l_1}{\delta k^2}\right] &= \frac{n}{k^2}(1+\frac{\pi^2}{6}+\gamma(\gamma-2)).
\end{aligned}$$

Here $\gamma = 0.577215\ldots$ is Euler's constant.

A.1.2 Examples

The previous section gave the contributions to the expected information matrix by each of the parameters for the general process described in equation (3.4.1). Here we give the expected matrices for each of the four examples considered in Section 3.4.

1. $\mu(X_t) = \mu_0 + \mu_1\frac{t}{n+1}$, $\sigma(X_t) = 1$
2. $\mu(X_t) = 0$, $\log\sigma(X_t) = \sigma_0 + \sigma_1\frac{t}{n+1}$,
3. $\mu(X_t) = \mu_0 + \mu_1\cos(\frac{2\pi t}{N}) - \mu_2\sin(\frac{2\pi t}{N})$, $\sigma(X_t) = 1$,
4. $\mu(X_t) = 0$, $\log\sigma(X_t) = \sigma_0 + \sigma_1\cos(\frac{2\pi t}{N}) - \sigma_2\sin(\frac{2\pi t}{N})$

For the existing method the information matrices are as follows;

1. Let

$$\begin{aligned}
r &= \frac{\exp\{-2(u-\mu(X_t))/k\}\exp\{-\exp[-(u-\mu(X_t))/k]\}}{1-\exp\{-\exp[-(u-\mu(X_t))/k]\}} \quad \text{and} \\
v &= 1-\exp\left\{-\exp\left(-(u-\mu(X_t))/k\right)\right\}
\end{aligned}$$

then

$$I = \frac{1}{k^2}\begin{bmatrix} \sum_{t=1}^n r & \sum_{t=1}^n tr & \frac{1}{k}\sum_{t=1}^n (u-\mu(X_t))r \\ \sum_{t=1}^n tr & \sum_{t=1}^n t^2 r & \frac{1}{k}\sum_{t=1}^n (u-\mu(X_t))tr \\ \frac{1}{k}\sum_{t=1}^n (u-\mu(X_t))r & \frac{1}{k}\sum_{t=1}^n (u-\mu(X_t))tr & \frac{1}{k^2}\sum_{t=1}^n \big((u-\mu(X_t))^2 r \\ & & +(2uk+2-k^2)v\big) \end{bmatrix}.$$

2. Let

$$\begin{aligned}
r &= \frac{\exp\{-2u/k\sigma(X_t)\}\exp\{-\exp[-u/k\sigma(X_t)]\}}{1-\exp\{-\exp[-u/k\sigma(X_t)]\}} \quad \text{and} \\
s &= u^2 r + k^2\sigma^2(X_t)(1-\mathrm{e}^{-\mathrm{e}^{-u/k\sigma(X_t)}})
\end{aligned}$$

then

$$I = \begin{bmatrix} \sum_{t=1}^n \frac{t^2}{\sigma^2(X_t)k^2}s & \sum_{t=1}^n \frac{t}{\sigma^2(X_t)k^3}s \\ \sum_{t=1}^n \frac{t}{\sigma^2(X_t)k^3}s & \sum_{t=1}^n \frac{1}{\sigma^2(X_t)k^4}s \end{bmatrix}.$$

3. Let $C = \cos\frac{2\pi t}{N}$, $S = \sin\frac{2\pi t}{N}$ and

$$\begin{aligned}
z &= \frac{u-\mu(X_t)}{k}, \\
r &= \frac{\exp\{-2z\}\exp\{-\exp(-z)\}}{1-\exp\{-\exp(-z)\}} \quad \text{and} \\
v &= 1-\exp\left\{-\exp(-z)\right\}
\end{aligned}$$

then

$$I = \frac{1}{k^2}\begin{bmatrix} \sum_{t=1}^n r & \sum_{t=1}^n Cr & \sum_{t=1}^n Sr & \frac{1}{k}\sum_{t=1}^n (u-\mu(X_t))r \\ \sum_{t=1}^n Cr & \sum_{t=1}^n C^2r & \sum_{t=1}^n CSr & \frac{1}{k}\sum_{t=1}^n Cr \\ \sum_{t=1}^n Sr & \sum_{t=1}^n CSr & \sum_{t=1}^n S^2r & \frac{1}{k}\sum_{t=1}^n Sr \\ \frac{1}{k}\sum_{t=1}^n (u-\mu(X_t))r & \frac{1}{k}\sum_{t=1}^n Cr & \frac{1}{k}\sum_{t=1}^n Sr & \frac{1}{k^2}\sum_{t=1}^n \{(u-\mu(X_t))^2 r \\ & & & +(2uk+2-k^2)v\} \end{bmatrix}.$$

4. Let C and S be as above and

$$r = \frac{\exp\{-2u/k\sigma(X_t)\}\exp\{-\exp(-u/k\sigma(X_t))\}}{1-\exp\{-\exp(-u/k\sigma(X_t))} \text{ and } s = u^2r + k^2\sigma^2(X_t)(1-\mathrm{e}^{-\mathrm{e}^{-z}})$$

then

$$I = \frac{1}{k^2}\begin{bmatrix} \sum_{t=1}^n \frac{C^2}{\sigma^2(X_t)}s & \sum_{t=1}^n \frac{CS}{\sigma^2(X_t)}s & \frac{1}{k}\sum_{t=1}^n \frac{C}{\sigma^2(X_t)}s \\ \sum_{t=1}^n \frac{CS}{\sigma^2(X_t)}s & \sum_{t=1}^n \frac{S^2}{\sigma^2(X_t)}s & \frac{1}{k}\sum_{t=1}^n \frac{S}{\sigma^2(X_t)}s \\ \frac{1}{k}\sum_{t=1}^n \frac{C}{\sigma^2(X_t)}s & \frac{1}{k}\sum_{t=1}^n \frac{S}{\sigma^2(X_t)}s & \frac{1}{k^2}\sum_{t=1}^n \frac{1}{\sigma^2(X_t)}s \end{bmatrix}.$$

Under the pre-processing method the information matrices are as follows;

1.

$$I = \frac{1}{k^2}\begin{bmatrix} n & \sum_{t=1}^n t & n(\gamma-1) \\ \sum_{t=1}^n t & \sum_{t=1}^n t^2 & (\gamma-1)\sum_{t=1}^n t \\ n(\gamma-1) & (\gamma-1)\sum_{t=1}^n t & n(1+\frac{\pi^2}{6}+\gamma(\gamma-2)) \end{bmatrix}.$$

2.

$$I = (1+\frac{\pi^2}{6}+\gamma(\gamma-2))\begin{bmatrix} \sum_{t=1}^n t^2 & \frac{1}{k}\sum_{t=1}^n t \\ \frac{1}{k}\sum_{t=1}^n t & \frac{n}{k^2} \end{bmatrix}.$$

3. Let C and S be as above, then

$$I = \frac{1}{k^2}\begin{bmatrix} n & \sum_{t=1} nC & \sum_{t=1} nS & n(\gamma-1) \\ \sum_{t=1} nC & \sum_{t=1} nC^2 & \sum_{t=1} nCS & (\gamma-1)\sum_{t=1} nC \\ \sum_{t=1} nS & \sum_{t=1} nCS & \sum_{t=1} nS^2 & (\gamma-1)\sum_{t=1} nS \\ n(\gamma-1) & (\gamma-1)\sum_{t=1} nC & (\gamma-1)\sum_{t=1} nS & n(1+\frac{\pi^2}{6}+\gamma(\gamma-2)) \end{bmatrix}.$$

4.

$$I = (1+\frac{\pi^2}{6}+\gamma(\gamma-2))\begin{bmatrix} \sum_{t=1}^{n} C^2 & \sum_{t=1}^{n} CS & \frac{1}{k}\sum_{t=1}^{n} C \\ \sum_{t=1}^{n} CS & \sum_{t=1}^{n} S^2 & \frac{1}{k}\sum_{t=1}^{n} S \\ \frac{1}{k}\sum_{t=1}^{n} C & \frac{1}{k}\sum_{t=1}^{n} S & \frac{n}{k^2} \end{bmatrix}.$$

Chapter 4

Models for multivariate extremes

4.1 Introduction

We aim to extend the pre-processing approach for modelling the extreme levels of ozone to a fully multivariate approach by modelling the joint distribution of ozone, NO and NO_2, conditional on associated meteorological covariates. Such an approach allows the estimation of ozone return levels which takes into account the non-stationarity and tail distributions of NO and NO_2. We suggest a conditional (or *hierarchical*) approach, in which we apply the pre-processing method to each pollutant in turn. The pollutants are first ordered in some scientifically meaningful way. Starting with the lowest ranked of the pollutants, level j in the hierarchy consists of modelling the jth pollutant, conditional on the covariates and the preceding pollutants. This conditional approach provides a model from which we can estimate return levels by simulation.

Let $\{Y_{1t}\}$, $\{Y_{2t}\}$ and $\{Y_{3t}\}$ be processes representing NO, NO_2 and ozone levels respectively and let $\{\mathbf{X}_t\}$ denote the associated covariates (more details of these are given in Section 4.4). Our approach is then to model each of these processes by following the pre-processing method introduced in Chapter 3, first fitting a Box-Cox location-scale model and transforming the processes using these estimated parameters and then modelling the extremes of the transformed process using the threshold exceedances approach. From henceforth we shall refer to the combined

Box-Cox and location-scale parameters as the *pre-processing parameters* and in the same way refer to the combined rate and GPD parameters as the *tail parameters.*

We now describe the ordering for our hierarchical model. At the first level we model NO, or $\{Y_{1t}\}$, assuming that the model parameters are functions of the covariates $\{\mathbf{X}_t\}$ only. At the second level, we model NO_2, or $\{Y_{2t}\}$, assuming that the model parameters are functions of both $\{\mathbf{X}_t\}$ and $\{Y_{1t}\}$. Finally at the third level we model ozone, or $\{Y_{3t}\}$, assuming that the model parameters are functions of $\{Y_{1t}\}$, $\{Y_{2t}\}$ and $\{\mathbf{X}_t\}$. The reasons for this choice of ordering are as follows. As discussed at length in Chapter 3, ozone is a secondary pollutant with primary precursors NO and NO_2 so it seems sensible to model ozone levels as a function of both NO and NO_2 levels. Now both NO and NO_2 are primarily released by the combustion of fossil fuels; once in the atmosphere, both may become involved in chemical reactions resulting in the formation of the other. Such reactions take place in a matter of minutes, so it follows that if, as in our case, we are modelling daily maxima the choice to model NO_2 conditional on NO, rather than the other way round, is arbitrary. However, note that the complex and cyclic nature of the chemical reactions taking place means that almost any ordering of these pollutants may be justified; for example, Shi and Harrison (1997) consider modelling NO_2 as a function of the interactions between NO and O_3.

By simulating from this hierarchical model we can estimate, for large N, the N-year return level for each of the pollutants and also examine the values of the remaining pollutants when one of them attains such a value. Suppose that we are interested in one of the three processes $\{Y_{it}\}$, where $i = 1, 2$ or 3. Recall from Chapter 3 that the N-year return level, which we shall denote by $y_{i,N}$, is the level exceeded once every N years, and is defined as

$$\Pr(Y_i > y_{i,N}) = \frac{1}{365N}, \tag{4.1.1}$$

since we have daily data. In Chapter 3 we estimated *marginal* return levels by averaging across the return levels estimated by conditioning on each of the observed vectors of covariates. In this chapter, by estimating the return levels by simulation, we show that we can obtain the marginal return levels directly.

We begin by fitting our model in a likelihood framework, following the approach taken in Chapter 3. However the hierarchical nature of the model suggests that it might be more sensible to carry out the model fit in a Bayesian framework using Markov Chain Monte Carlo (MCMC) techniques. Bayesian inference has several advantages over likelihood inference, specifically in terms of estimating the uncertainty in the parameter estimates, since Bayesian inference results in a posterior distribution, rather than a single point estimate, for each of the model parameters and functions there-of. By straight-forward extension this means that we can estimate posterior distributions of the return levels. This makes estimating uncertainty in our parameter (return level) estimates much simpler than under the likelihood approach where we have to use bootstrap methods. The main disadvantages of Bayesian inference are that it generally requires computationally intensive techniques and that it also requires the placing of a prior probability distribution on each model parameter.

For a generic parameter θ using Bayes theorem, the posterior distribution of θ given observed data y is given by

$$f(\theta|y) = \frac{f(y|\theta)f(\theta)}{\int f(y|\theta)f(\theta)\ \mathrm{d}\theta} \propto f(y|\theta)f(\theta)$$

where $f(y|\theta)$ is the likelihood of the data, $f(\theta)$ is the prior probability distribution on the parameter θ and the integral in the denominator for the exact expression is known as the normalising constant (or *constant of proportionality*). However, in practise and especially for high-dimensional problems, the posterior distribution often has a complex and non-standard form; further it might only be known up to the constant of proportionality. For these reasons, the only way to summarise the posterior distribution is often to simulate from it. However direct simulation is frequently impossible for the very same reasons. A solution to this problem comes from a range of techniques, collectively known as Markov Chain Monte Carlo

(MCMC) methods, which can be used to produce an independent and identically distributed (IID) sample from the posterior distribution for θ, even when it is known only up to a constant of proportionality.

MCMC techniques involve simulating data from a Markov chain which has the required posterior distribution as its stationary distribution. Following an inital period (known as *burn-in*) whilst the chain settles down to its stationary distribution, each simulated value is assumed to have been drawn from the posterior. Due to the ways in which the Markov chains are simulated, there will often be dependence between consecutive draws, so to induce independence in the sample it is normal to save only every τth update, where τ is chosen so that the chain has only weak dependence at this lag. An alternative approach which also aims to achieve independence is to generate many short chains and save only the final update of each. We adopt the former approach.

There are many MCMC algorithms (see Smith and Roberts (1993) and Wilks *et al.*, 1998) each of which has different mechanisms to update the value of the Markov chain. We use two of the simplest; the Metropolis-Hastings random walk algorithm and the Gibbs sampler. The Gibbs sampler is a special case of the Metropolis-Hastings algorithm which is useful when it is possible to simulate directly from the required posterior distribution (conditional on any other parameters).

Coles and Powell (1996) provide a good overview of the application of Bayesian inference and MCMC techniques to extreme value problems. An advantage of a Bayesian approach is that it allows for more complex model structures in a more intuitive way than under likelihood inference. More recent developments in the extremes literature have taken advantage of this; for example Coles and Casson (1999), Fawcett and Walshaw (2006) and Cooley *et al.* (2007) look at modelling spatial extremes. Renard *et al.* (2006) consider modelling non-stationary extremes in a hydrological context using step-change as well as trend models and Tancredi *et al.* (2006) use the Bayesian setting to suggest a way of automatic threshold selection in the threshold exceedances approach discussed in Chapter 3.

The remainder of the chapter is organised as follows. In Section 4.2 we outline more fully the methodology involved in our approach. We discuss two possible methods of inference for our model in Section 4.3; these being either a likelihood or a Bayesian approach. Finally in Section 4.4 we present results of fitting the proposed model to our air pollution data by estimating 10- and 100-year return levels for NO, NO_2 and ozone.

4.2 Methodology

As described in the previous section, our proposed hierarchical model for multivariate extremes is a simple extension of the pre-processing method introduced in Chapter 3. However we introduce some modifications. We allow the model parameters for the response variable Y_i $(i = 1, 2, 3)$ to be a function not only of the covariates but also of the response variables with a lower ordering in the hierarchy *i.e.* the set $S_i = \{Y_j : j < i\}$. This allows us to account for the non-stationarity in the marginal distributions of $\{Y_{1t}\}$, $\{Y_{2t}\}$ and $\{Y_{3t}\}$ and also, through the hierarchical nature of the model, for non-stationarity in their extremal dependence structure. We also allow for a model on *both* tails of each process. The reason for this is that, whilst we are ultimately interested in extreme high values, it is possible that a negative dependence structure between two variables means that it is the lowest values of one variable which affect the highest values of another. Our method is the same at each level of the hierarchy and is described as follows.

Suppose that we are interested in modelling the process $\{Y_{it}\}$ at level i. We first apply the Box-Cox location-scale model as follows.

$$\frac{Y_{it}^{\lambda_i(\mathbf{x}_t, S_{i,t})} - 1}{\lambda_i(\mathbf{x}_t, S_{i,t})} = \mu_i(\mathbf{x}_t, S_{i,t}) + \sigma_i(\mathbf{x}_t, S_{i,t}) Z_{it} \tag{4.2.1}$$

where $S_{i,t} = \{y_{jt} : j < i\}$, which is the empty set for Y_1. As before we assume that, for each i, the $\{Z_{it}\}$ are identically distributed. Further we also assume that both the transformed processes $\{Z_{1t}\}$ and $\{Z_{2t}\}$, given $\mathbf{x}_t$ and $\{Y_{1t}\}$, and the transformed processes $\{Z_{2t}\}$ and $\{Z_{3t}\}$, given $\mathbf{x}_t$, $\{Y_{1t}\}$ and $\{Y_{2t}\}$, are conditionally independent. We take λ_i, μ_i and $\log(\sigma_i)$ to be linear functions of $(\mathbf{x_t}, S_{i,t})$, and we denote the vectors of coefficients for the three parameters by $\boldsymbol{\lambda}_i$, $\boldsymbol{\mu}_i$ and $\boldsymbol{\sigma}_i$.

Assuming that the body of the distribution of the process $\{Z_{it}\}$ is stationary, we model both tails of the process using the threshold exceedances method. That is we select an upper (lower) threshold u_i (u_i^l) and estimate the rate $\phi_{i,u}(\mathbf{x}_t, S_{i,t})$ ($\phi_{i,u}^l(\mathbf{x}_t, S_{i,t})$) of an observation occurring above (below) this level. The observations falling above (below) this threshold are then modelled using a generalised Pareto distribution (GPD) with scale $\psi_{i,u}(\mathbf{x}_t, S_{i,t})$ ($\psi_{i,u}^l(\mathbf{x}_t, S_{i,t})$) and shape $\xi_i(\mathbf{x}_t, S_{i,t})$ ($\xi_i^l(\mathbf{x}_t, S_{i,t})$) parameters. Note the change of notation for the extremes parameters from that used in Chapter 3, since all tail models are fitted to the transformed process we chose to label the parameters not by z which denotes this, but by i to denote the level of the hierarchy to which they belong. We suggest using logit and log link functions for the rate and GPD scale parameter respectively and taking $\text{logit}(\phi_{i,u})$, $\text{logit}(\phi_{i,u}^l)$, $\log(\psi_{i,u})$, $\log(\psi_{i,u}^l)$, ξ_i and ξ_i^l to be linear functions of $(\mathbf{x}_t, S_{i,t})$.

4.2.1 Return levels

To estimate the marginal N-year return levels $y_{i,N}$ we simulate N-years of data directly from the model and take the largest value from this simulation as our estimate of $y_{i,N}$ in a similar manner to that used by Buishand *et al.* (2006) to estimate return levels from their stationary spatial model. Note that, for $M < N$, we can also estimate the M-year return level by taking the N/Mth order statistic of the simulated data.

Simulation of the multivariate response variable $\mathbf{Y} = (Y_1, Y_2, Y_3)$ from the hierarchical model is straight forward. We simulate from the lowest level of the hierarchy first and then work our way up, conditioning on the simulated values from the lower levels to simulate from the higher levels.

However before simulating the response $\mathbf{Y}$ we must first simulate a set of covariates. At this point it is worth noticing that we are interested not just in simulating a single observation $\mathbf{Y}$ but in simulating a whole sequence of observations $\{\mathbf{Y}_t\}$ which has a length of N-years. Because of this the simulation method must account for two things; first, that we have not assumed a probability distribution for the covariates and, second, that the covariates are likely to be non-stationary themselves. Since we have no distribution for the covariates we simply resample with replacement randomly from the observations. In order to retain any seasonal trends in the covariates, for each day of the year, rather than random resampling from *all* the covariates, we resample randomly across observed years from that day only.

Let $\{\mathbf{X}_1^*, \ldots, \mathbf{X}_N^*\}$ denote the simulated covariates, then a sequence $\{Y_{11}^*, \ldots, Y_{1N}^*\}$ of observations from the model for $\{Y_{1t}|\mathbf{X}_t = \mathbf{x}_t^*\}$ is simulated as follows.

1. Simulate a vector $\{U_1, \ldots, U_N\}$ of realisations from N independent uniform(0,1) random variables.

2. For each $i = 1, \ldots, N$,

   (a) If $U_i > 1 - \hat{\phi}_{1,u}(\mathbf{x}_t^*)$ then simulate Z_{1i}^* from the fitted upper tail model $\text{GPD}(\hat{\psi}_{1,u}(\mathbf{x}_t^*), \hat{\xi}_1(\mathbf{x}_t^*))$.

   (b) If $U_i < \hat{\phi}_{1,u}^l(\mathbf{x}_t^*)$, simulate Z_{1i}^* from the fitted lower $\text{GPD}(\hat{\psi}_{1,u}^l(\mathbf{x}_t^*), \hat{\xi}_1^l(\mathbf{x}_t^*))$.

   (c) If $\hat{\phi}_{1,u}^l(\mathbf{x}_t^*) < U_i < 1 - \hat{\phi}_{1,u}(\mathbf{x}_t^*)$ simulate Z_{1i}^* from the empirical distribution of the transformed process *i.e.* resample Z_{1i}^* from the set $\{Z_{1i} : u_1^l \leq Z_{1i} \leq u_1\}$.

3. Back transform the simulated observations to the original scale using equation (4.2.1)

$$Y_{1i}^* = \{\lambda_1(\mathbf{x}_t^*)[\mu_1(\mathbf{x}_t^*) + \sigma_1(\mathbf{x}_t^*)Z_{1i}^*] + 1\}^{1/\lambda_1(\mathbf{x}_t^*)}$$

Using the same steps, we then use these simulated values to simulate the sequence $\{Y_{21}^*, \ldots, Y_{2N}^*\}$ from the model for $\{Y_{2t}|Y_{1t} = y_{1t}^*, \mathbf{X}_t = \mathbf{x}_t^*\}$ and similarly combine our simulated values for the covariates $\mathbf{X}_t$, Y_1 and Y_2 to simulate the sequence $\{Y_{31}^*, \ldots, Y_{3N}^*\}$ from the model for $\{Y_{3t}|Y_{2t} = y_{2t}^*, Y_{1t} = y_{1t}^*, \mathbf{X}_t = \mathbf{x}_t^*\}$.

4.3 Inference

As discussed in Chapter 3 the pre-processing model is fitted in two steps; first the pre-processing parameters are estimated, then the tail model(s) are fitted to the transformed data. The rate and GPD parameters can be estimated independently, which simplifies computational matters. We further simplify matters by choosing to fix the Box-Cox parameter at some constant value, which can then be estimated, for example, by maximising the profile likelihood for λ. Following the results in Chapter 3, we also chose to fix the tail parameters as constant.

4.3.1 Likelihood inference

Likelihood inference for the model parameters is straightforward. Each level of the hierarchy is fitted independently of the rest. Estimating uncertainty in the model parameters and return levels is not quite so straightforward. For example, the rate and GPD parameters depend on the Box-Cox and location-scale parameters, but because of the two-stage nature of the model fit there is no way to take this into account when attempting to estimate their standard errors using standard asymptotic likelihood results; instead we use a block bootstrap method. Note that to estimate uncertainty in the estimated return levels we simulate a new set of covariates for each bootstrapped sample, thus accounting for uncertainty in the covariates as well.

4.3.2 Bayesian model inference

Since the location-scale and tail models are fitted independently to the data at each level of the hierarchy, our Bayesian model specification is the same at each level and so we state it only for the generic level i.

We update the location-scale parameters using a Metropolis within Gibbs step, the rate parameter using a Gibbs step and the GPD parameters using a Metropolis within Gibbs step. Further details of the updating steps used for each parameter, along with the priors used, are given below. First, the general algorithm to generate samples from the posterior distributions of all the parameters at level i is as follows.

1. Simulate the Markov chains whose stationary distributions are the posterior distributions of the pre-processing parameters $\boldsymbol{\mu}_i$ and $\boldsymbol{\sigma}_i$ by iterating the following procedure. Starting at some initial value,

   (a) Jointly update the vector $\boldsymbol{\mu}_i$ given the current values of $\boldsymbol{\mu}_i$ and $\boldsymbol{\sigma}_i$

   (b) Jointly update the value of $\boldsymbol{\sigma}_i$ given the updated value of $\boldsymbol{\mu}_i$ and the current value of $\boldsymbol{\sigma}_i$.

2. Following burn-in, an independent sample is taken from the posterior distributions for the location-scale parameters by taking only every τth update, with τ chosen so that the chain has only weak dependence at this lag.

3. For each of the sample values of $\boldsymbol{\mu}_i$ and $\boldsymbol{\sigma}_i$ found in step 2, calculate the transformed process $\{Z_{it}\}$ using equation (4.2.1). Select the upper (lower) threshold as a pre-specified quantile of this transformed process; where the quantile level is constant across all iterations.

4. Using the threshold obtained for each transformed process from step 3, simulate the Markov chains whose stationary distributions are the posterior distributions of the upper and lower tail parameters by iterating the following procedure. Starting at some initial values,

   (a) Update the rate parameter $\phi_{i,u}$ ($\phi^l_{i,u}$) conditional on its current value.

   (b) Update the GPD scale $\psi_{i,u}$ ($\psi^l_{i,u}$) conditional on the current values of the GPD scale $\psi_{i,u}$ ($\psi^l_{i,u}$) and shape ξ_i (ξ^l_i) parameters.

   (c) Update the GPD shape ξ_i (ξ^l_i) conditional on the updated value of the GPD scale $\psi_{i,u}$ ($\psi^l_{i,u}$) and the current value of the shape ξ_i (ξ^l_i)

parameters.

5. Following burn-in, an independent sample is taken from the posterior distributions for the tail parameters by taking only those updates from the chains simulated in step 4 that are separated by at least one other consecutive update. The samples are pooled across the simulated location-scale parameters.

The priors used in our model fit are as follows. For the first stage we place independent multivariate Gaussian priors on the location and scale coefficients $\boldsymbol{\mu}_i$ and $\boldsymbol{\sigma}_i$, that is

$$\begin{aligned} \boldsymbol{\mu}_i &\sim \mathrm{MVN}(\boldsymbol{\mu}_0, \Sigma_\mu) \\ \boldsymbol{\sigma}_i &\sim \mathrm{MVN}(\boldsymbol{\sigma}_0, \Sigma_\sigma). \end{aligned}$$

With this choice of prior, the posterior distribution of $\boldsymbol{\mu}_i|\boldsymbol{\sigma}_i, \mathbf{X}, S_i, \mathbf{Y}_i$ is also multivariate Gaussian, and so we can update each value in the Markov chain for $\boldsymbol{\mu}_i$ using a Gibbs step. However the posterior distribution of $\boldsymbol{\sigma}_i|\boldsymbol{\mu}_i, \mathbf{X}, S_i, \mathbf{Y}_i$ is more complicated and requires the use of a Metropolis-Hastings random walk for the updates.

When fitting the tail models we assume *a priori* that all the extremes parameters are independent. We place independent Beta(α, β) priors on the rate parameters, so that for $0 < \phi_{i,u}, \phi^l_{i,u} < 1$,

$$\phi_{i,u} \sim \mathrm{Beta}(\alpha, \beta), \quad \phi^l_{i,u} \sim \mathrm{Beta}(\alpha, \beta).$$

For the GPD scale parameters we assume independent Gamma(η,θ) priors; for $\psi_{i,u}, \psi^l_{i,u} > 0$,

$$\psi_{i,u} \sim \mathrm{Gamma}(\eta, \theta), \quad \psi^l_{i,u} \sim \mathrm{Gamma}(\eta, \theta).$$

Finally for the GPD shape parameters we follow Cooley *et al.* (2006) and assume independent improper uniform priors; for $-\infty < \xi_i, \xi_i^l < \infty$,

$$\xi_i \sim 1, \quad \xi_i^l \sim 1.$$

With these priors the posterior distributions of the rate parameters are also independent Beta and so can be updated using a Gibbs step. The GPD parameters each have complicated posteriors with unknown normalising constants and so, like the scale parameter, we update these using a Metropolis-Hastings random walk. We note that whilst Coles and Tawn (1996) suggest placing priors on the quantiles of the distribution for exceedances, rather than on the GPD parameters, here we deal directly with the parameters since there is no extra information available on the quantiles.

Calculation of the posterior distribution of the return levels follows immediately using the samples from the posterior distributions of the model parameters. For each set of parameters drawn from the posterior distributions we can simulate an N-year data set, as discussed in Section 4.2 and so estimate the N-year return level. Carrying this out for all draws from the parameter posteriors gives an estimate of the posterior distribution for the return levels.

4.3.3 MCMC details

Following initial exploration to try to determine the speed of convergence of the chains we ran the chain for the pre-processing parameters for a total of 205000 updates, removed the first 5000 as burn-in and took every 10th update, giving a posterior sample of size 20000. For the tail parameters we ran a chain of length 10000 for every 500th update of the pre-processing parameters, following burn-in. For each chain we removed the first 5000 as burn-in and then took every 100th update to get a posterior sample also of size 20000.

We tried various initial values for each of the chains; for the pre-processing and tail parameters there was little sensitivity to this choice, certainly once we had removed burn-in periods of the size described above. Initial values for the GPD parameters proved a little more difficult, particularly because the parameters seem to be quite strongly negatively correlated so that picking a value of either well out of its likely range had quite drastic effects on the convergence of the chain. We found that starting the scale close to 1 and the shape close to 0 had desirable consequences.

We chose prior parameter values as follows. For the pre-processing parameters we make the priors uninformative, with means $\boldsymbol{\mu}_0 = \boldsymbol{\sigma}_0 = \mathbf{0}$ and covariance matrices $\Sigma_\mu = \Sigma_\sigma = 1000I$ where I is the identity matrix. For the rate parameters we take $\alpha = \beta = 1$ and similarly for the GPD scale parameters we set $\eta = \theta = 1$. As with the initial values, altering these parameter values had little effect. We also considered using an uninformative Gaussian prior, rather than the improper uniform prior, for the GPD shape. This too made no evident difference to the final results.

4.4 Results

The data that we use to demonstrate our hierarchical modelling scheme consist of maximum daily concentrations of hourly measurements of ozone, NO and NO_2. The data are shown in Figure 4.1 and are the same as those used in Chapter 3. The data has been measured at a monitoring station located in central Reading, UK. This site is classified as being in an urban location and the data used here cover the period from September 1997 to June 2001.

We have three meteorological covariates, all of which have the potential to affect the concentration levels of one or more of our response variables. These are wind speed (measured daily at 0900), maximum daily temperature and total daily sunshine. Measurements of these covariates are shown in Figure 4.2 across the same period as the air pollution data. We also consider as covariates the first-order interactions of these variables. Reasons for including sunshine and temperature were discussed in Chapter 3 and are primarily due to their being key factors in

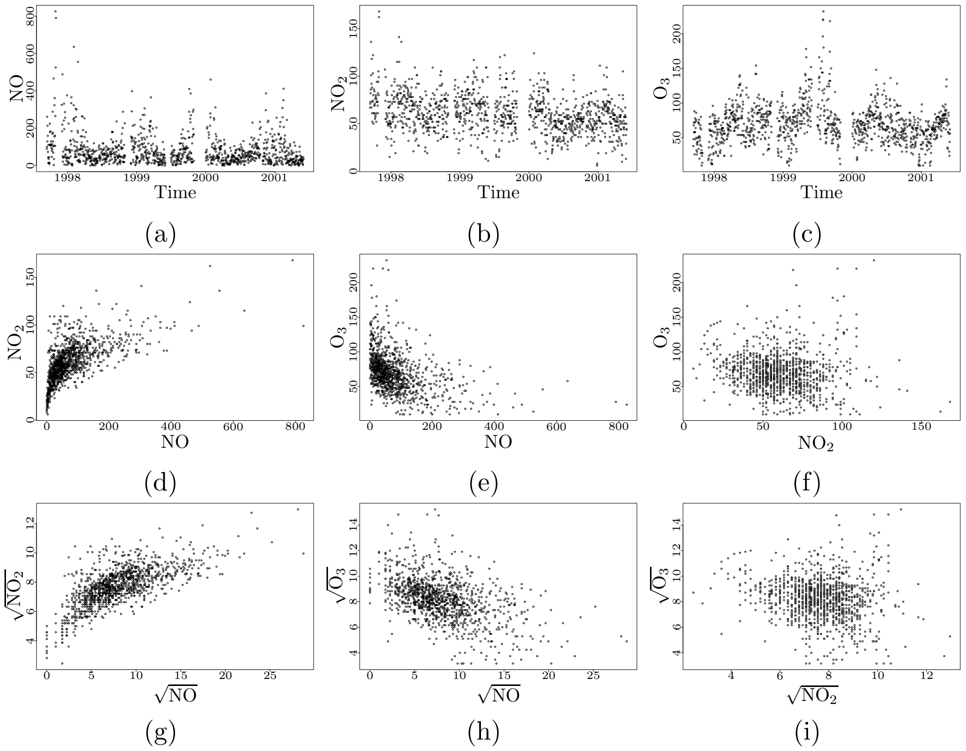
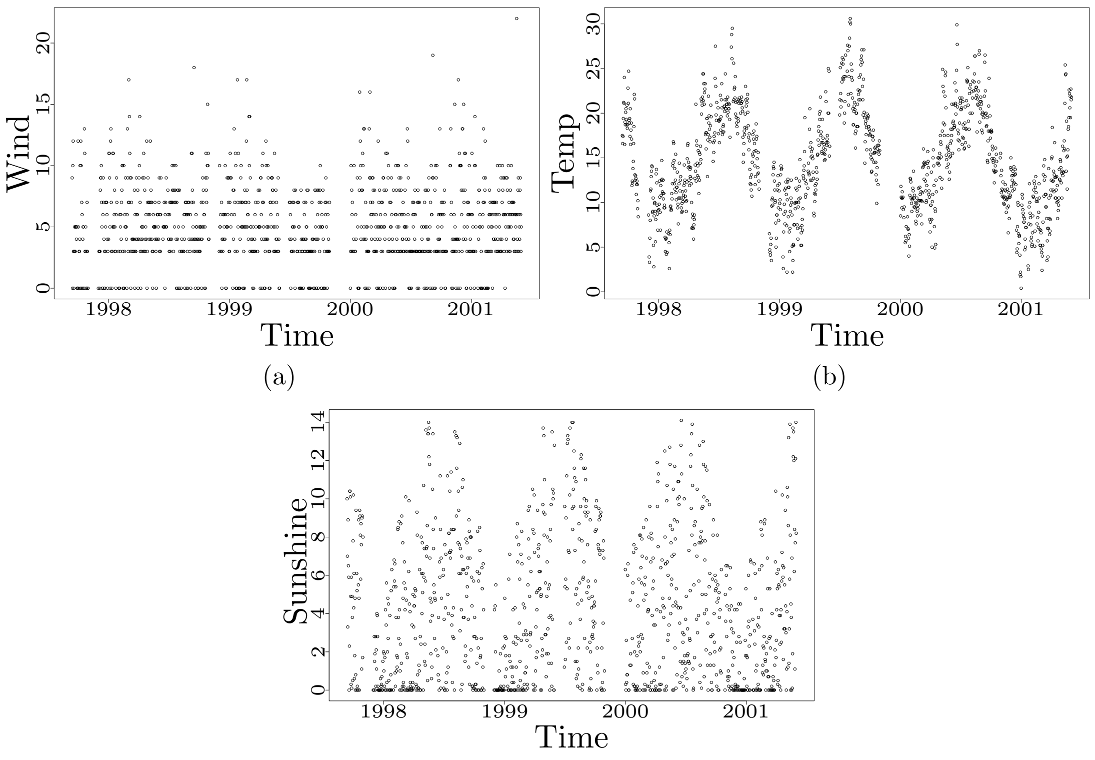


Figure 4.1: Time series plots of maximum daily NO, NO_2 and ozone concentrations (top) with bivariate scatter plots on the original (middle) and square root (bottom) scales. Data were measured in central Reading from September 1997 until June 2001. Measurements are in μmg^{-3}.

the synthesis of ozone. We include wind speed because increases in wind speed cause greater mixing of particles in the atmosphere which in turn leads to faster dispersion of the air pollutants; this is especially relevant for the primary pollutants NO and NO_2.

We also use various time indicators as covariates. The purpose of these is mostly to try to account for physical covariates for which we have no data; examples of such covariates include other air pollutants, such as volatile organic compounds (VOC's), traffic volume and proximity of the site to potential point sources (such as factories). As substitutes for these, we use yearly, seasonal and weekend indicators. The year indicator should pick up long term trends attributable, for example, to successful implementation of legislation to decrease emissions due to the combustion of fossil fuels. We use a year indicator rather than a linear year-on-year trend, as this will allow the detection of more subtle trends. The reason for including a seasonal indicator is evident from the time-series plots of the pollutants (see Fig-



(a) (b)

(c)

Figure 4.2: Time series plots of wind speed, measured daily at 9am (knots), daily maxima temperature (°C) and total daily sunshine (hours), measured in central Reading from September 1997 until June 2001.

ure 4.1) all of which show clear seasonal variation; with peaks in the winter for NO and NO_2 and in the summer for ozone. We use four three-month seasons defined as winter (December-February), spring (March-May), summer (June-August) and autumn (September-November). Finally we use the weekend indicator since there is evidence in the literature of a marked difference between NO and NO_2 levels in the week and those at the weekend, especially in urban areas (*e.g.* Shi and Harrison, 1997). This is due to alterations in the traffic pattern at weekends.

For models fitted using likelihood inference we report maximum likelihood estimates (MLE's) as point estimates for the parameters, whereas we use posterior medians (PM's) for the Bayesian models. To estimate uncertainty we use, respectively, the asymptotic normality property of the MLE, with the standard error estimated using the observed information matrix, and posterior credibility regions.

4.4.1 Individual models

To begin with we consider the fit of the *saturated* model to each of the NO, NO_2 and ozone data sets. By saturated we mean that the location-scale parameters at level i ($i = 1, 2, 3$) of the hierarchy contain all possible covariates $\{\mathbf{X}_t\}$ as well as the responses $\{Y_{jt} : Y_{jt} \in S_i\}$. Unless specified otherwise we fit models only to the upper tail of the transformed data sets.

In Section 4.3 we decided to fix the Box-Cox parameter $\lambda(\mathbf{x}_t)$ to be a constant value λ. For each data set, we select λ_i ($i = 1, 2, 3$) by maximising the profile likelihood for λ_i over a discrete number of parameter values. The profile likelihood is found by maximising the joint likelihood for the full vectors of location-scale coefficients $\boldsymbol{\mu}_i$ and $\boldsymbol{\sigma}_i$ conditional on each possible value of λ_i. The potential values for λ_i are usually chosen to have some meaningful interpretation, for example $\lambda_i = -1, -0.5, 0, 0.5, 1, 2$. In this case, for all three data sets, the profile likelihood is maximised across these values when $\lambda = 0.5$. Further, the plots in Figure 4.1 show that the relationship between both $(\sqrt{\mathrm{NO}}, \sqrt{\mathrm{NO_2}})$ and $(\sqrt{\mathrm{NO}}, \sqrt{\mathrm{O_3}})$ looks closer to being linear than the equivalent relationships on the original scale (see Figure 4.1) or on the log scale (not shown). Given both of these results we shall model the square root of NO, the square root of NO_2 conditional on the square root of NO and the square root of ozone conditional on the square roots of both NO and NO_2.

Tables 4.1-4.3 show point estimates, under both methods of inference, for the location-scale coefficients of the saturated model fitted to, respectively, NO, NO_2 and ozone. We see that in all cases the MLE's are very close to the posterior medians (PM's), in particular they always fall within the 95% posterior credibility regions. Figure 4.3 shows the fitted means $\mu_i(\mathbf{x}_t)$ and transformed processes $\{Z_{it}\}$ for each of the data sets using the PM's given in Tables 4.1-4.3 as point estimates for the location-scale coefficients $\boldsymbol{\mu}_i$ and $\boldsymbol{\sigma}_i$. These plots show both that the mean functions follow the data well and that the transformed processes are considerably closer to stationarity than the original ones. Scatter plots of the transformed

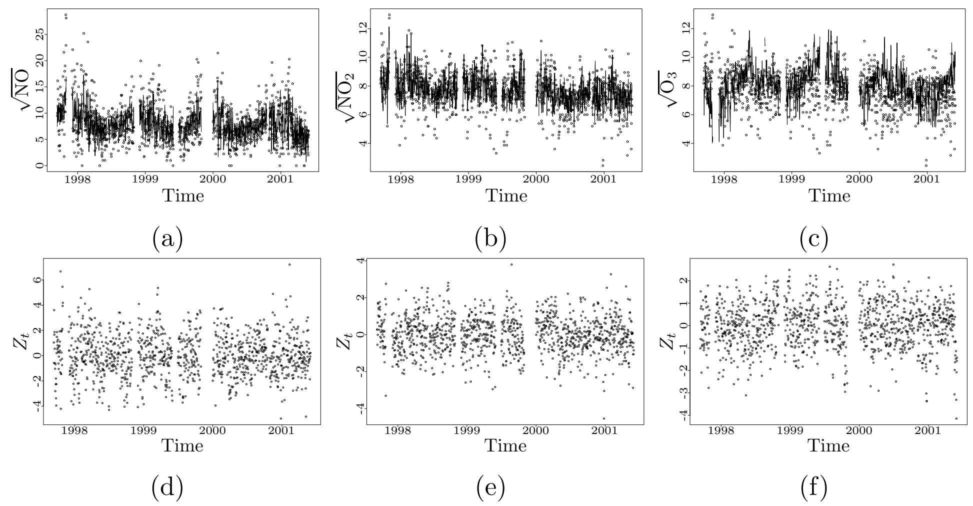


Figure 4.3: Square root of NO, NO_2 and ozone data sets with estimated mean $\mu_i(\mathbf{x}_t)$ (top) and transformed processes $\{Z_{it}\}$ (bottom). Estimates come from fitting the saturated model using Bayesian inference.

processes against each other, shown in Figure 4.4, and against the covariates (not shown) suggest that, under the saturated model, the assumptions that the transformed processes $\{Z_{it}\}$ are independent of each other and of the covariates $\{\mathbf{X}_t\}$ are reasonable.

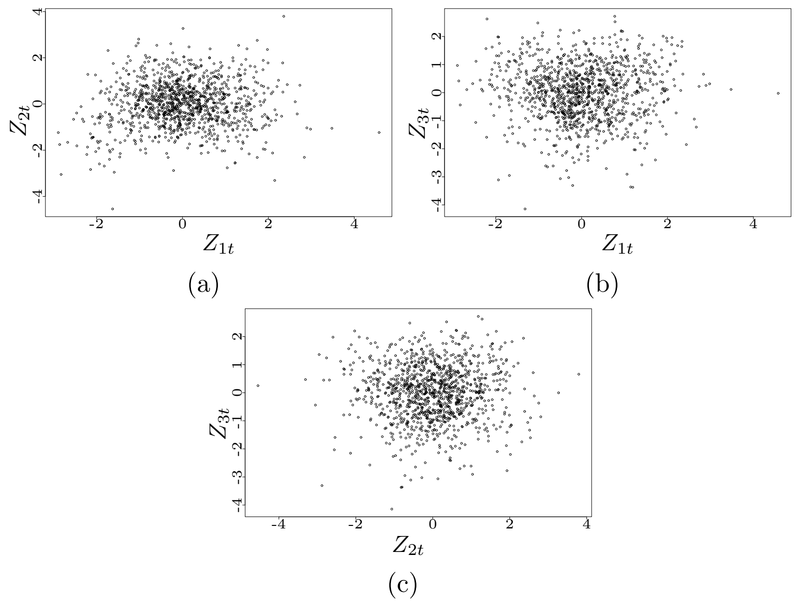
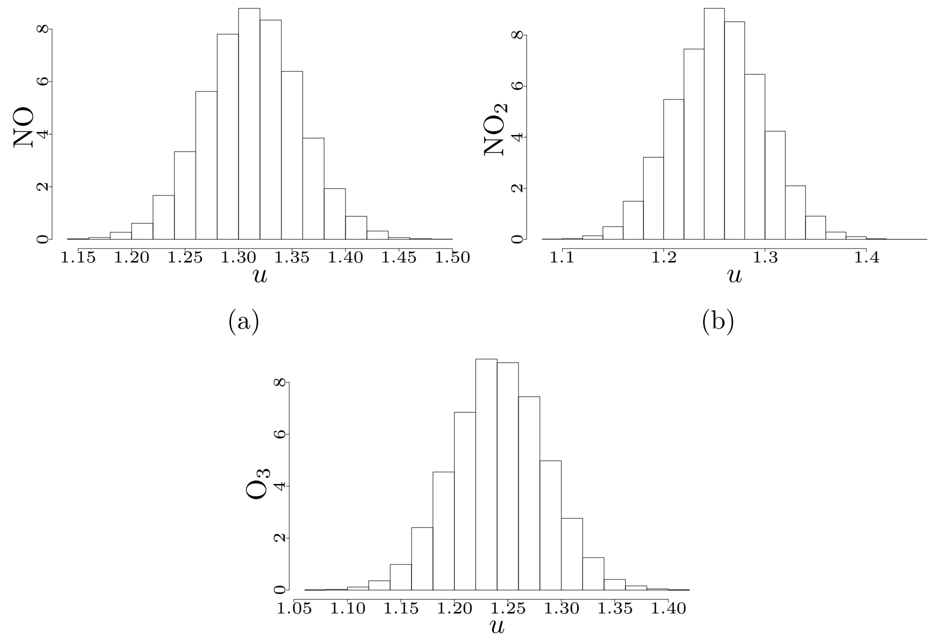


Figure 4.4: Scatter plots of the transformed processes $\{Z_{it}\}$ for $i = 1, 2, 3$ from the saturated model to show their independence. Plot (a) shows transformed NO_2 against transformed NO, plot (b) transformed ozone against transformed NO and plot (c) transformed ozone against NO_2.



(a)

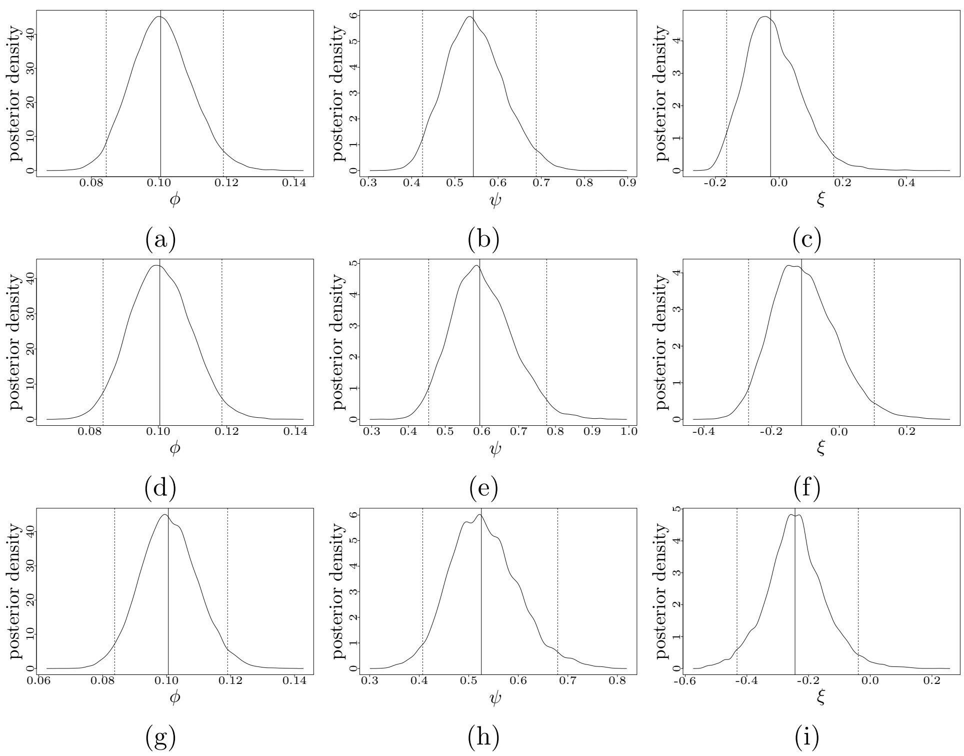
(b)

(c)

Figure 4.5: Histograms of the upper tail thresholds obtained from each draw out of the posterior distribution of the location-scale parameters for (a) NO, (b) NO_2 and (c) ozone. For each draw, thresholds were found by taking the 90% quantile of the transformed data sets given by these values of the location-scale parameters.

For each data set, we select the 90% (10%) quantile of the transformed data as the threshold for our upper (lower) tail model. For each of the three pollutants, Figure 4.5 shows histograms of the upper thresholds obtained by pooling the thresholds given at each draw from the posterior distribution of the location-scale parameters. These show some, but not much, variation; plots for the lower threshold (not shown) were similar. The estimated parameters for both tail models fitted under both methods of inference are shown in Table 4.4. As with the location-scale parameters, the likelihood and Bayesian point estimates for the parameters are very similar and the MLE's all fall within the corresponding 95% posterior credibility regions.

The estimated posterior densities for the upper tail parameters are shown in Figure 4.6. The kernel density estimation in these plots used the default method in R, *i.e.* a Gaussian kernel using Silverman's (1986) associated 'rule of thumb' for choice of bandwidth (0.9 times the minimum of the standard deviation and



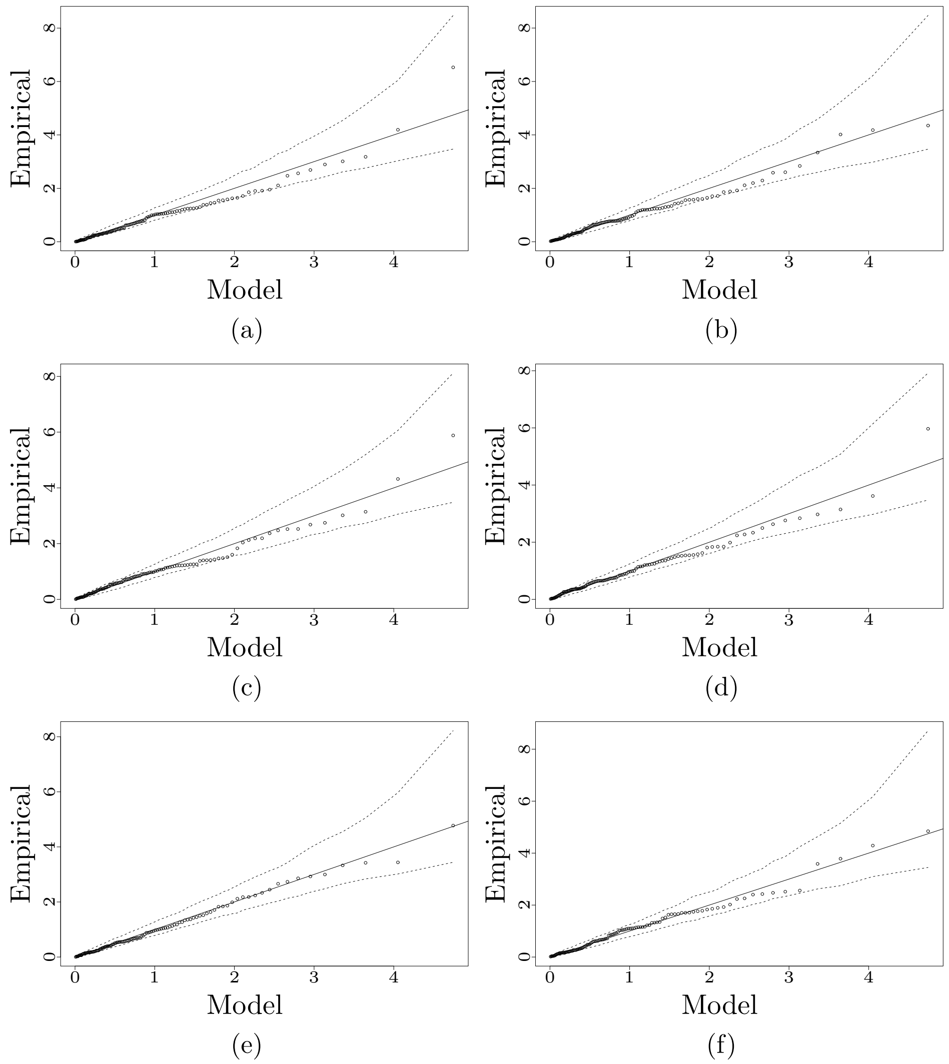
(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 4.6: Estimated posterior distributions for upper tail parameters $\phi_{i,u}$, $\psi_{i,u}$ and $\xi_{i,u}$ using the saturated location-scale model and a 90% threshold. Results are for NO (top), NO$_2$ (middle) and ozone (bottom). Full vertical lines indicate posterior medians and dashed vertical lines 95% posterior credibility regions. To produce these plots, a kernel density estimate was used to smooth the sample histograms.

interquartile range, divided by 1.34, multiplied by the sample size to the power of -0.2). For illustrative purposes this seems sufficient. All of these parameters seem to have posterior distributions whose upper tails are slightly heavier than their lower tails, but they are all reasonably symmetric, especially close to the posterior mode. To demonstrate the goodness of fit of the GPD model to the tails we show, in Figure 4.7, quantile-quantile (QQ) plots for the GPD model fit to both the upper and lower tails again taking the PM's of the GPD parameters as point estimates. Note that the QQ plots are all on the standard exponential scale in order to enable easy cross model comparisons. Under exact agreement between the model and the data, we would expect the points on the QQ plot to lie on the 45° line. In all cases the plots lie fairly closely to this line and they certainly fall within the 95% credibility regions.



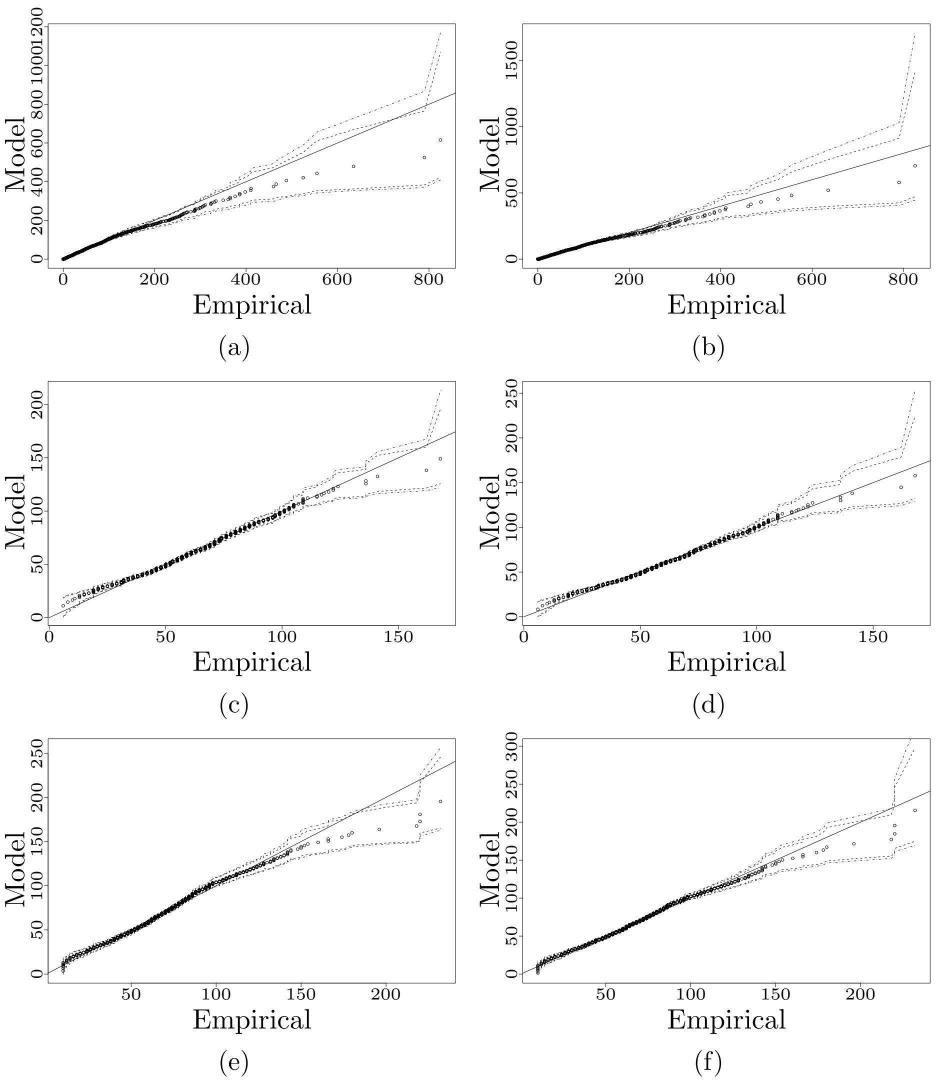
(a) (b)

(c) (d)

(e) (f)

Figure 4.7: QQ plots to show the goodness of fit of the GPD model to the values of the transformed processes falling above the 90% (left) or below the 10% (right) thresholds for NO (top), NO_2 (middle) and ozone (bottom). Point estimates for the GPD parameters were given by the posterior medians and the transformed process was found using the saturated model. Dashed lines show 95% credibility regions and the plots are on the standard exponential scale.

To assess the overall model fit, we plot the observed order statistics against the order statistics obtained by simulating data from the fitted model. We use the simulation methods described in Section 4.2.1 to simulate a number of data sets of the same length n as the observed data. For the likelihood approach we chose to take 500 bootstrapped resamples of the data and then simulate a data set from the model fitted to each of these resamples; in the Bayesian case we simply simulate a new data set from each of the draws from the posterior distribution.



(a) (b)

(c) (d)

(e) (f)

Figure 4.8: QQ plots to show overall model fit. Plots were generated by simulating data from the hierarchical models using likelihood (left) and Bayesian (right) inference. Plots (a) and (b) refer to the NO model, plots (c) and (d) to the NO_2 model and plots (e) and (f) to the ozone model. The green line shows exact agreement between the model and observations whilst the dashed (dashed-dot) lines show 95% (99%) confidence intervals (credibility regions).

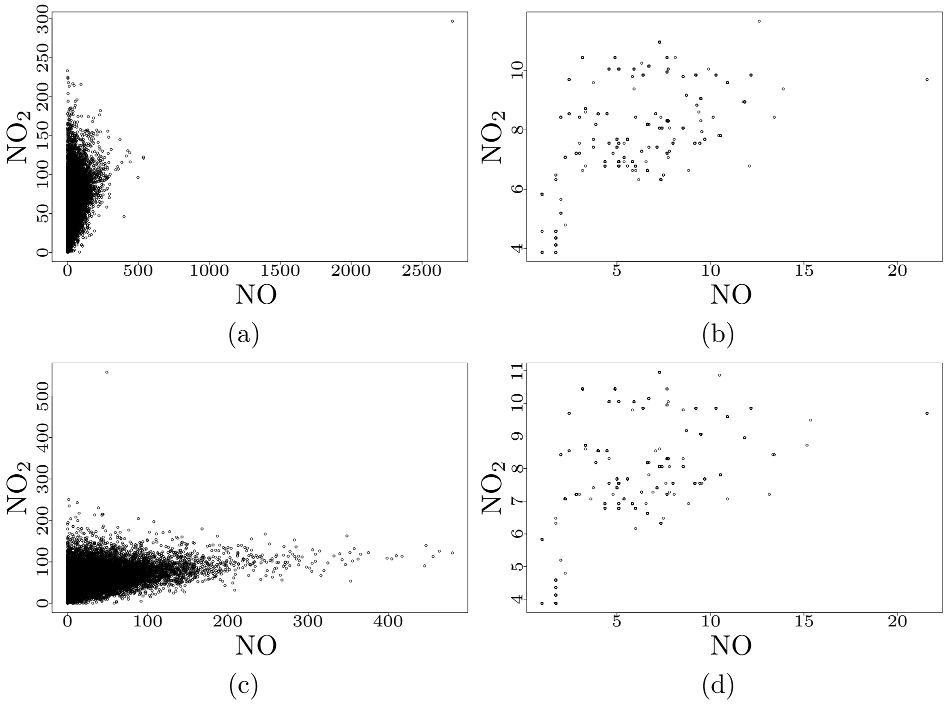
Each simulated data set is first ordered and then, for $i = 1, \ldots, n$, we find the median, $\alpha/2$ and $1 - \alpha/2$ quantiles of the ith order statistic *across* the simulated data sets. We take the medians as point estimates for the order statistics under the fitted model, whereas the $\alpha/2$ and $1 - \alpha/2$ quantiles provide $100(1 - \alpha)\%$ confidence intervals (credibility regions).

The results of this overall measure of fit for the saturated, upper tail only model are shown in Figure 4.8. These show that the model fits under both methods of inference are reasonably good, with a possibility of slight under estimation in the upper tails, particularly for the NO and ozone models. This under-fitting can almost certainly be explained by random variation in the simulated data as the 45° line almost always falls within the 95% confidence intervals (posterior credibility regions). The posterior credibility regions are noticeably wider than the corresponding likelihood confidence intervals. Finally, time series plots (not shown) summarising the simulated data sets show that the simulated data seems to reproduce the seasonal trends in the observed data sets well, which again confirms overall model fit.

4.4.2 Return levels

We compare the estimated return levels from four different models. The first of these is the saturated model, discussed in the previous section. We try fitting this with a GPD model for the upper tail only (Model 1) and then with a GPD model for both tails (Model 2). To specify the next model we apply a forward selection procedure to choose only the most significant covariates in the location and scale parameters (Model 3). We apply the forward selection to the models fitted using the likelihood approach, but also fit the best fitting models using the Bayesian approach. Finally we estimate the return levels using the saturated upper tail model but *without* assuming the hierarchical structure (Model 4). This means that instead of using the response data simulated from the previous levels S_i^* to simulate the data at the current level we use response data that has been resampled from the responses S_i in the same way that we resample the covariates $\mathbf{X}_t$.

Tables 4.5 and 4.6 show the 10- and 100-year return levels respectively for all three pollutants estimated using each of the four models. Note that the return levels for NO are the same under Model 1 as they are under Model 4, since the set up of the model and the way of simulating the data is the same at the first level of the hierarchy for these models. The most obvious feature of the results in these tables is that neither point estimates nor the measures of uncertainty vary much between the models. Whilst this suggests that we can fit the simpler Model 3, it also suggests that the model in which we do not use the hierarchical structure of the data gives as good an estimate as the hierarchical model. This is possibly because the responses are asymptotically independent once we have accounted for the covariates $\mathbf{X}_t$. Also notice that the posterior credibility regions are wider than the confidence intervals.



(a) (b) (c) (d)

Figure 4.9: Estimates of the joint distributions of NO and NO_2 when ozone achieves it's N-year return levels, for $N = 5$ (top) and $N = 10$ (bottom). Estimates come from the Bayesian fits of Model 1 (left) and Model 4 (right). Note the different scales on each of the plots.

Finally in Figure 4.9 we show scatter plots of NO and NO_2, conditional on ozone achieving it's 5- and 10-year return levels. These plots can be used to estimate the joint posterior distribution of NO and NO_2 given that ozone has attained an N-year return level. We show results for the Bayesian fits of both Model 1 and Model 4. Because of the resampling method used in Model 4 we get a poor estimate of the joint distribution because we cannot extrapolate into the distribution tails, further only a small subset of the observed values of NO and NO_2 seem to contribute to the extreme values of ozone; using Model 1 instead gives a much fuller picture of the joint distribution.

4.4.3 Further work

There are several ways in which the work summarised here could be extended. Firstly a simulation study should be run to test the accuracy and efficiency of the estimation of the return levels under the proposed method. It might also be informative to quantify how well the hierarchical method (Models 1-3) models the extremal dependence structure, perhaps under the assumption of different levels of asymptotic (in)dependence, especially when compared to the non-hierarchical model (Model 4). We might also consider the effect that the choice of ordering in the hierarchy has; for example would the estimated return levels be the same if we had modelled NO conditional on NO_2, rather than the other way round?

For the pre-processing model, it might be interesting to consider other more complex methods of pre-processing the data, for example to account for auto-correlation in the residuals which may be due to missing covariates or some mis-specification in the covariate model. For this particular data set it would also be useful if we could repeat the analysis with additional covariates; for example wind direction, levels of VOC's or traffic volume.

	$\boldsymbol{\mu}_1$		$\boldsymbol{\sigma}_1$	
Covariate	Likelihood	Bayesian	Likelihood	Bayesian
Constant	14.0 (0.529)	13.3 (12.3,14.4)	1.23 (0.138)	1.29 (1.02,1.57)
Temperature	-0.328 (0.0341)	-0.272 (-0.339,-0.207)	-0.0293 (0.00914)	-0.0319 (-0.0502,-0.0138)
Sunshine	0.569 (0.0867)	0.533 (0.360,0.784)	0.034 (0.0187)	0.0388 (0.000758,0.0758)
Wind	-0.932 (0.0752)	-0.816 (-0.966,-0.669)	-0.00240 (0.0200)	-0.00922 (-0.0496,0.0295)
Temperature × sunshine	-0.0161 (0.00379)	-0.0166 (-0.0241,-0.00918)	0.000251 (0.000899)	-0.0000274 (-0.00149,0.00215)
Temperature × wind	0.0433 (0.00494)	0.0342 (0.0245,0.0439)	-0.000280 (0.00130)	0.0000247 (-0.00235,0.00287)
Sunshine × wind	-0.0335 (0.00764)	-0.0249 (-0.0400,-0.101)	-0.00274 (0.00183)	-0.00318 (0.00677,0.000466)
$\mathcal{I}$[wkend]	-2.12 (0.168)	-2.13 (-2.46,-1.79)	-0.0926 (0.0488)	-0.0996 (-0.195,-0.00507)
$\mathcal{I}$[1997]	2.38 (0.566)	2.36 (1.24,3.47)	0.608 (0.113)	0.589 (0.366,0.812)
$\mathcal{I}$[1998]	1.38 (0.279)	1.36 (0.811,1.91)	0.279 (0.0759)	0.264 (0.111,0.413)
$\mathcal{I}$[1999]	1.00 (0.280)	0.938 (0.382,1.50)	0.193 (0.0817)	0.186 (0.0224,0.348)
$\mathcal{I}$[2000]	0.675 (0.260)	0.650 (0.133,1.16)	0.0417 (0.0742)	0.0252 (-0.123,0.172)
$\mathcal{I}$[spring]	-1.83 (0.253)	-1.88 (-2.38,-1.37)	-0.0360 (0.0650)	-0.0453 (-0.173,0.843)
$\mathcal{I}$[summer]	-0.631 (0.338)	-0.715 (-1.38,-0.0488)	-0.129 (0.0967)	-0.136 (-0.325,-0.0556)
$\mathcal{I}$[autumn]	0.484 (0.300)	0.466 (-0.147,1.04)	0.0148 (0.0778)	0.0164 (-0.136,0.171)

Table 4.1: Point estimates for location-scale coefficients for the saturated model for $\sqrt{\text{NO}}$ (level 1 in the hierarchy). Both MLE's and asymptotic standard errors (in brackets) and PM's and 95% posterior credibility regions (in brackets) are shown.

	$\boldsymbol{\mu}_2$		$\boldsymbol{\sigma}_2$	
Covariate	Likelihood	Bayesian	Likelihood	Bayesian
Constant	5.18 (0.184)	5.17 (4.80,5.53)	-0.0397 (0.156)	-0.0222 (-0.314,0.291)
$\sqrt{\mathrm{NO}}$	0.211 (0.00857)	0.212 (0.195,0.230)	-0.0391 (00680)	-0.382 (-0.0518,-0.0246)
Temperature	0.0505 (0.0110)	0.0511 (0.0294,0.729)	0.00699 (0.00932)	0.00576 (-0.0126,0.0241)
Sunshine	0.0490 (0.0225)	0.0478 (0.00316,0.0923)	0.000853 (0.0206)	-0.00192 (-0.0427,0.0392)
Wind	0.0766 (0.0204)	0.0750 (0.0350,0.116)	-0.0135 (0.0184)	-0.0137 (-0.0507,0.0227)
Temperature × sunshine	-0.000599 (0.000124)	-0.000584 (-0.00303,0.00150)	0.00174 (0.000967)	0.00190 (-0.0000445,0.00388)
Temperature × wind	-0.00782 (0.00148)	-0.00772 (-0.0106,-0.00481)	-0.000499 (0.00126)	-0.000449 (-0.00297,0.00205)
Sunshine × wind	0.00341 (0.00220)	0.00360 (-0.000657,0.00750)	-0.00113 (0.00199)	-0.00110 (-0.00502,0.00280)
$\mathcal{I}$[wkend]	-0.282 (0.0541)	-0.278 (-0.385,-0.169)	-0.147 (0.0503)	-0.141 (-0.239,-0.0459)
$\mathcal{I}$[1997]	0.290 (0.154)	0.290 (-0.0172,0.602)	0.564 (0.119)	0.566 (0.341,0.811)
$\mathcal{I}$[1998]	0.489 (0.0830)	0.488 (0.328,0.648)	-0.00490 (0.0783)	-0.00415 (-0.161,0.147)
$\mathcal{I}$[1999]	0.420 (0.0849)	0.419 (0.250,0.586)	0.0579 (0.0814)	0.0614 (-0.103,0.221)
$\mathcal{I}$[2000]	0.0983 (0.0798)	0.0982 (-0.0606,0.257)	0.0443 (0.0753)	0.0443 (-0.107,0.191)
$\mathcal{I}$[spring]	0.109 (0.0700)	0.105 (-0.0332,0.244)	0.0803 (0.0712)	0.0807 (-0.0631,0.187)
$\mathcal{I}$[summer]	-0.557 (0.108)	-0.565 (-0.779,-0.353)	-0.00543 (0.0953)	-0.00410 (-0.190,0.187)
$\mathcal{I}$[autumn]	-0.378 (0.0838)	-0.387 (-0.552,-0.223)	0.0212 (0.0844)	0.0213 (-0.144,0.188)

Table 4.2: As Table 4.1, point estimates for location-scale coefficients for the saturated model for $\sqrt{\mathrm{NO}_2}$ (level 2 in the hierarchy).

	$\boldsymbol{\mu}_3$		$\boldsymbol{\sigma}_3$	
Covariate	Likelihood	Bayesian	Likelihood	Bayesian
Constant	6.64 (0.325)	6.36 (5.71,7.01)	0.321 (0.213)	0.395 (-0.0241,0.823)
$\sqrt{\text{NO}}$	-0.184 (0.156)	-0.179 (-0.120,-0.148)	0.0359 (0.00951)	0.0352 (0.0167,0.0536)
$\sqrt{\text{NO}_2}$	0.0957 (0.0382)	0.102 (0.0266,0.180)	-0.0283 (0.241)	-0.0342 (-0.0806,0.135)
Temperature	0.0683 (0.0156)	0.0808 (0.0498,0.112)	-0.0101(0.00993)	-0.0113 (-0.0315,0.00819)
Sunshine	0.0540 (0.0337)	0.0592 (-0.00747,0.126)	-0.108 (0.0215)	-0.106 (-0.147,-0.0628)
Wind	0.181 (0.0293)	0.209 (0.151,0.268)	-0.0417 (0.0205)	-0.0425 (-0.0853,-0.00110)
Temperature × sunshine	0.00671 (0.00172)	0.00643 (0.00297,0.00980)	0.00599 (0.00106)	0.00595 (0.00386,0.00807)
Temperature × wind	-0.0100 (0.00200)	-0.0117 (-0.0157,-0.00773)	0.000499 (0.00136)	0.000597 (-0.00214,0.00343)
Sunshine × wind	-0.00855 (0.00279)	-0.000877 (-0.0143,-0.00337)	0.00425 (0.00201)	0.00403 (0.000132,0.00790)
$\mathcal{I}$[wkend]	-0.071 (0.0716)	-0.0562 (-0.200,0.0848)	-0.0134 (0.0501)	-0.0151 (-0.110,0.0865)
$\mathcal{I}$[1997]	0.0824 (0.186)	0.0600 (-0.312,0.434)	0.0971 (0.114)	0.111 (-0.115,0.335)
$\mathcal{I}$[1998]	0.751 (0.108)	0.733 (0.518,0.944)	0.0189 (0.0800)	0.0220 (-0.142,0.178)
$\mathcal{I}$[1999]	0.995 (0.1112)	0.974 (0.753,1.20)	0.000240 (0.0849)	0.00000601 (-0.166,0.165)
$\mathcal{I}$[2000]	0.384 (0.0997)	0.337 (0.140,0.537)	-0.0800 (0.0782)	-0.0769 (-0.235,0.0703)
$\mathcal{I}$[spring]	0.532 (0.0881)	0.537 (0.363,0.713)	-0.0851 (0.0706)	-0.0791 (-0.214,0.0561)
$\mathcal{I}$[summer]	-0.383 (0.139)	-0.390 (-0.664,-0.110)	-0.0836 (0.0972)	-0.0777 (-0.264,0.110)
$\mathcal{I}$[autumn]	-0.162 (0.114)	-0.173 (-0.399,-0.0596)	(0.0794)	0.0163 (-0.137,0.172)

Table 4.3: As Table 4.1, point estimates for location-scale coefficients for the saturated model for $\sqrt{\text{O}_3}$ (level 3 in the hierarchy).

	Upper tail			Lower tail		
	$\phi_{i,u}$	$\psi_{i,u}$	$\xi_{i,u}$	$\phi^l_{i,u}$	$\psi^l_{i,u}$	$\xi^l_{i,u}$
Likelihood						
NO	0.100	0.476	0.0100	0.100	0.681	-0.332
	(0.00888)	(0.0595)	(0.0828)	(0.00888)	(0.0806)	(0.0789)
NO_2	0.100	0.577	-0.121	0.100	0.615	-0.0731
	(0.00888)	(0.0683)	(0.0735)	(0.00888)	(0.0724)	(0.0717)
O_3	0.1001	0.520	-0.270	0.100	0.657	-0.0635
	(0.00888)	(0.0625)	(0.0795)	(0.00888)	(0.0899)	(0.0999)
Bayesian						
NO	0.101	0.543	-0.0267	0.101	0.688	-0.301
	(0.0844,0.119)	(0.425,0.688)	(-0.165,0.172)	(0.0839,0.119)	(0.536,0.873)	(-0.445,-0.122)
NO_2	0.100	0.594	-0.111	0.101	0.613	-0.0325
	(0.0839,0.119)	(0.455,0.776)	(-0.268,0.103)	(0.0834,0.119)	(0.467,0.796)	(-0.184,0.195)
O_3	0.101	0.525	-0.244	0.101	0.665	-0.0343
	(0.0837,0.119)	(0.407,0.679)	(-0.432,-0.0387)	(0.0839,0.119)	(0.503,0.869)	(-0.214,0.203)

Table 4.4: Estimates for upper (lower) tail rate ϕ_u (ϕ^l_u) and GPD scale ψ_u (ψ^l_u) and shape xi (xi^l) parameters for the saturated model fitted under both the likelihood and Bayesian methods. Point estimates are MLE's with asymptotic standard errors (in brackets) and PM's with 95% posterior credibility regions (in brackets).

	Model 1		Model 2		Model 3		Model 4	
	Likelihood	Bayesian	Likelihood	Bayesian	Likelihood	Bayesian	Likelihood	Bayesian
NO	698 (573,868)	824 (670,1123)	689 (562,895)	824 (669,1123)	666 (548,865)	820 (670,1144)		
NO_2	156 (144,175)	169 (154,198)	156 (142,174)	169 (154,198)	156 (143,173)	167 (152,193)	163 (151,180)	169 (156,192)
O_3	208 (190,236)	237 (211,275)	208 (188,235)	237 (210,276)	213 (190,246)	234 (208,271)	208 (187,232)	234 (209,272)

Table 4.5: Estimated 10-year return levels under the four different models. Point estimates were obtained by simulation with either bootstrapped 95% confidence intervals (in brackets) or 95% posterior credibility intervals (in brackets).

	Model 1		Model 2		Model 3		Model 4	
	Likelihood	Bayesian	Likelihood	Bayesian	Likelihood	Bayesian	Likelihood	Bayesian
NO	1044 (722,1840)	1357 (909,3058)	1027 (711,1912)	1349 (912,3042)	937 (677,1399)	1346 (913,3366)		
NO_2	190 (163,250)	215 (177,369)	188 (158,257)	215 (177,371)	184 (158,229)	210 (174,363)	188 (164,236)	203 (174,317)
O_3	241 (211,294)	292 (241,397)	241 (211,296)	292 (242,399)	248 (215,312)	288 (239,390)	238 (205,288)	290 (238,396)

Table 4.6: As Table 4.5 but estimated 100-year return levels.

References

Buishand, T. A., de Haan, L. and Zhou, C. (2006) On spatial extremes: with application to a rainfall problem. *Pre-print.*

Coles, S. G. and Tawn , J. A. (1996) A Bayesian Analysis of Extreme Rainfall Data. *Appl. Statist.*, **45**:4, 463-478.

Coles, S. G. and Powell, E. A. (1996) Bayesian Methods in Extreme Value Modelling : A Review and New Developments. *International Statistical Review*, **64**:1, 119-136.

Coles, S. G. and Casson, E (1999) Spatial Regression Models for Extremes. *Extremes*, **1**:4, 449-468.

Cooley, D., Naveau, P. and Jomelli, V. (2006) A Bayesian Hierarchical Extreme Value Model for Lichenometry. *Environmetrics*, **17**:555-574.

Cooley, D., Nychka, D. and Naveau, P. (2007) Bayesian Spatial Modelling of Extreme Precipitation Levels Return Levels. *Journal of the American Statistical Association*, *in press.*

Fawcett, L. and Walshaw, D. (2006) A hierarchical model for extreme wind speeds. *Appl. Statist.*, **55**:5, 631-646.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1998) *Markov Chain Monte Carlo in practice.* Chapman and Hall, London.

Renard, B., Lang, M. and Bois, P. (2006) Statistical analysis of extreme events in a non-stationary context via a Bayesian framework: case study with peak-over-threshold data. *Stoch. Environ. Res. Risk Assess. 21*, 97-112.

Shi, J. P. and Harrison, R.M. (1997) Regression modelling of hourly NO_x and NO_2 concentrations in urban air in London. *Atmospheric Environment*, **24**, 4081-4094.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

Smith, A. F. M. and Roberts, G. O. (1993) Bayesian Computation Via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods. *J. R. Statist. Soc. B*, **55**:1, 3-23.

Tancredi, A., Anderson, C. and O'Hagan, A. (2006) Accounting for threshold uncertainty in extreme value estimation. *Extremes*, **9**, 87-106.

Chapter 5

Nonparametric estimation of extremal dependence measures using a limiting conditional representation.

5.1 Introduction

The quantification of dependence is the central issue in probabilistic and statistical methods for multivariate extreme value problems. When estimating the probability of any extreme multivariate event it is vital to make inferences about the extremal dependence structure. A growing literature on this topic illustrates the importance of understanding the properties of the joint tails of multivariate distributions. The range of applied fields on which multivariate statistical extreme value theory is making an impact is expanding and includes to date: environmental impact assessment (Coles and Tawn, 1991, Joe, 1994, de Haan and de Ronde, 1998, Schlather and Tawn, 2003), financial risk management (Embrechts *et al.*, 1997, 2000, Longin, 2000, Stărică, 1999, Poon *et al.*, 2003a, 2003b), internet traffic modelling (Maulik *et al.*, 2002, Resnick and Rootzén, 2002) and sports (Barão and

Tawn, 1999).

The two types of extremal dependence structure are asymptotic dependence and asymptotic independence. For random variables (X_1, X_2) with identical marginal distributions with upper endpoint x^*, following Coles *et al.* (1999) we define

$$\chi = \lim_{x \to x^*} \Pr\{X_2 > x \,|\, X_1 > x\}. \tag{5.1.1}$$

If $\chi > 0$ then we say that X_1 and X_2 are asymptotically dependent, in which case the largest values of both variables tend to occur simultaneously. If $\chi = 0$ then the variables are asymptotically independent and it is impossible to get the largest values of X_1 and X_2 to occur simultaneously, even though (X_1, X_2) can be positively dependent.

Traditionally, attention has focused on extremal dependence models arising from the class of distributions that describe the stochastic behaviour of componentwise maxima data. This is the class of so-called multivariate extreme value distributions (de Haan and Resnick, 1977, Pickands, 1981, and Resnick, 1987). This class provides a rich description of data that are asymptotically dependent, but collapses all asymptotically independent distributions to being treated as independent. Ledford and Tawn (1996, 1997) and Coles *et al.* (1999) pointed out the inadequacies of multivariate extreme value distribution models, and asymptotically dependent distributions more generally, to describe data which are asymptotically independent. Recently, much work has concentrated on developing more general extremal dependence modelling frameworks which can accommodate both asymptotically dependent and asymptotically independent data.

One such approach is offered by the recent work of Heffernan and Tawn (2004), who put forward a new strategy for modelling the joint tails of multivariate distributions. Their approach uses the conditional distribution of the remaining variables given that at least one variable is large. This approach offers a flexible class of models incorporating both asymptotic dependence and asymptotic independence, and allows the modelling of parts of the joint distribution for which not all variables are large.

Despite the development of methods that can accommodate both asymptotically dependent and asymptotically independent data there is still much focus on the asymptotically dependent class of models due to many examples naturally falling in the class. For example many financial variables exhibit asymptotic dependence (Stărică, 1999, Embrechts, 2000 and Poon, *et al.* 2003b); one such example is given in the financial application illustrating our proposed methods in this paper.

In the current paper we exploit the tail representation presented by Heffernan and Tawn (2004) to refine their estimation procedure in the case when the variables can be treated as being asymptotically dependent. We focus on the bivariate case and obtain new nonparametric estimators for the underlying spectral measure and Pickands' dependence function that characterise extremal dependence structure and the bivariate extreme value distribution respectively. We show consistency of these estimators by considering their asymptotic distributions. The performance of the resulting methodology is shown to be competitive with, if not slightly better than, that of the existing estimators which assume the data are asymptotically dependent. However, we believe that a major benefit of our augmentation of the Heffernan and Tawn approach with the methods described in this paper is that they uniquely offer a unified methodology for the analysis of a broad range of dependence structures which extends beyond the class of asymptotic dependence.

The remainder of this paper is structured as follows. In Section 5.2 we introduce the classical point process representation for bivariate extremes and we derive from this the probabilities of various extreme events including the bivariate extreme value distribution. In Section 5.3 we recall the conditional representation of Heffernan and Tawn (2004), and explicitly in closed form express their nonparametric estimator which can be used whether the variables are asymptotically dependent or not. Our proposed nonparametric estimators, obtained under the assumption of asymptotic dependence, are developed in Section 5.4. We also present here theorems on the consistency of the estimators, although proofs are relegated to an appendix. In Section 5.5 we review existing nonparametric estimators of Pickands' dependence function, and in Section 5.6 we compare the performance of the new estimator with leading existing nonparametric estimators. In Section 5.7 we illustrate the use of the new estimator with an application to finance. We finish in Section 5.8 with a discussion and outline how our estimator can be extended to the multivariate case.

5.2 Classical results for bivariate extremes

Let (X_1, X_2) be a vector random variable with unknown distribution function F. We assume the marginal distributions of F to be unit Fréchet. Where the margins are unknown they may be estimated by the empirical distribution function. This is justified by Genest *et al.* (1995) who show that replacing the true margins by their empirical counterparts does not affect the efficiency of dependence parameter estimators. Suppose that (X_{1i}, X_{2i}), $i = 1, \ldots, n$, is a series of independent random variables distributed as (X_1, X_2). Let

$$P_n = \{(X_{1i}/n, X_{2i}/n) : i = 1, \ldots, n\}$$

represent the point process of normalised points (X_{1i}, X_{2i}) on $\mathbb{R}_+^2$. The normalisation by n arises from the max-stability property of unit Fréchet variables. As $n \to \infty$, subject to weak regularity conditions on the tail form of F (Resnick, 1987), P_n converges to an inhomogeneous Poisson process P on $\mathbb{R}_+^2 \backslash \{\mathbf{0}\}$.

A key feature of P is that its intensity measure factorises into functions of pseudo-radial R and angular W components defined by

$$R = ||(X_1, X_2)|| \quad \text{and} \quad W = X_1/R \tag{5.2.1}$$

where $||\cdot||$ is any choice of norm. For ease of exposition, we follow Coles and Tawn (1991) and choose to work henceforth with the L_1 norm so that $R = ||(X_1, X_2)|| = X_1 + X_2$, though others also work with the L_2 (Einmahl *et al.*, 1993) and L_∞ norms (Einmahl *et al.*, 1997 and 2001). With the L_1 norm, the intensity measure of P satisfies

$$\mu(\,\mathrm{d}r \times \,\mathrm{d}w) = \frac{\mathrm{d}r}{r^2} 2\,\mathrm{d}H(w) \tag{5.2.2}$$

where H is a distribution function on the interval $[0, 1]$ satisfying the moment condition

$$\int_0^1 w\,\mathrm{d}H(w) = 1/2. \tag{5.2.3}$$

The angular measure H and its density h, when it exists, are referred to as the spectral measure and the spectral density respectively.

The form of the known function of the radial component in intensity (5.2.2) arises from the choice of unit Fréchet margins. Thus, the dependence structure of extreme observations is characterised entirely by the spectral measure. Specifically, let $N_n(B)$ be the number of occurrences of the event $B \subset \mathbb{R}^2_+ \backslash \{\mathbf{0}\}$ by the process P_n and let $N(B)$ be the equivalent number for the Poisson process P, so that $N_n(B)$ converges in distribution to $N(B)$ and $N(B)$ follows a Poisson distribution with mean

$$\Lambda(B) = \int_B \frac{\mathrm{d}r}{r^2} 2\,\mathrm{d}H(w). \tag{5.2.4}$$

Furthermore, for $C \subset B$, as $n \to \infty$

$$\Pr\{(X_1/n, X_2/n) \in C \,|\, (X_1/n, X_2/n) \in B\} \to \frac{\Lambda(C)}{\Lambda(B)}. \tag{5.2.5}$$

These results illustrate that inference for H is fundamental to all inferences for extreme events.

We illustrate the use of results (5.2.4) and (5.2.5) for examples of events B and C which will be useful in Section 5.4. The simplest such sets are those for which at least one component of (X_1, X_2) exceeds some high level, that is $B_1 = \{(X_1, X_2) : X_1 > x\}$ and $B_2 = \{(X_1, X_2) : X_2 > y\}$. Then $\Lambda(B_1)$ is

$$\Lambda(B_1) = \int_0^1 \int_{x/w}^{\infty} \mu(\,\mathrm{d}r \times \,\mathrm{d}w)\,\mathrm{d}r\,\mathrm{d}w = \int_0^1 \int_{x/w}^{\infty} \frac{\mathrm{d}r}{r^2} 2\,\mathrm{d}H(w) = \frac{2}{x}\int_0^1 w\,\mathrm{d}H(w) = \frac{1}{x} \tag{5.2.6}$$

by the moment condition on H of equation (5.2.3). Similarly, we have $\Lambda(B_2) = \frac{1}{y}$. The second pair of interesting sets are subsets of B_1 and B_2 given by

$$B_1^{(t)} = \left\{(X_1, X_2) : X_1 > x, \frac{X_1}{X_1 + X_2} < t\right\} \text{ and } B_2^{(t)} = \left\{(X_1, X_2) : X_2 > y, \frac{X_1}{X_1 + X_2} < t\right\}.$$

which are created by adding a constraint on the size of the angular coordinate. Then, using the moment condition on H, the integrated intensity for $B_1^{(t)}$ is given by

$$\Lambda(B_1^{(t)}) = \int_0^t \int_{x/w}^{\infty} \frac{\mathrm{d}r}{r^2} 2\,\mathrm{d}H(w) = \frac{2}{x}\int_0^t w\,\mathrm{d}H(w). \tag{5.2.7}$$

Similarly, we have

$$\Lambda(B_2^{(t)}) = \frac{2}{y}\int_0^t (1-w)\,\mathrm{d}H(w) = \frac{2}{y}H(t) - \frac{2}{y}\int_0^t w\,\mathrm{d}H(w). \tag{5.2.8}$$

A consequence of these results for the integrated intensity is that if

$$C_1(t) = \lim_{n\to\infty} \Pr\{(X_1/n, X_2/n) \in B_1^{(t)} \mid (X_1/n, X_2/n) \in B_1\} \tag{5.2.9}$$

and

$$C_2(t) = \lim_{n\to\infty} \Pr\{(X_1/n, X_2/n) \in B_2^{(t)} \mid (X_1/n, X_2/n) \in B_2\} \tag{5.2.10}$$

then,

$$\frac{1}{2}\left[C_1(t) + C_2(t)\right] = H(t), \tag{5.2.11}$$

for $t \in [0, 1]$. Equation (5.2.11) is a new representation for $H(t)$ and is the basis of our statistical estimator in Section 5.4.

A further use of the point process convergence is to derive the bivariate extreme value distribution as the limiting distribution of the componentwise maxima

$$M_{n,1} = \max_{i=1,\ldots,n} X_{1i} \quad \text{and} \quad M_{n,2} = \max_{i=1,\ldots,n} X_{2i}.$$

Specifically consider the event $B_{xy} = \{(X_1, X_2) : X_1 > x \text{ or } X_2 > y\}$, then by the convergence of the process P_n to P, as $n \to \infty$,

$$\begin{aligned} \Pr\{M_{n,1}/n \leq x, M_{n,2}/n \leq y\} &\to \Pr\{N(B_{xy}) = 0\} \\ &= \exp\{-\Lambda(B_{xy})\} \end{aligned} \qquad (5.2.12)$$

where

$$\Lambda(B_{xy}) = \int_0^1 \int_{\min\{x/w, y/(1-w)\}}^{\infty} \frac{\mathrm{d}r}{r^2} 2\,\mathrm{d}H(w) = \int_0^1 2\max\left(\frac{w}{x}, \frac{1-w}{y}\right)\mathrm{d}H(w). \qquad (5.2.13)$$

We denote the limiting distribution by $G(x, y)$, where

$$G(x, y) = \exp\left\{-\int_0^1 2\max\left(\frac{w}{x}, \frac{1-w}{y}\right)\mathrm{d}H(w)\right\} \qquad (5.2.14)$$

which is the bivariate extreme value distribution, see de Haan and Resnick (1977) and Pickands (1981).

A widely used characterisation of the dependence structure of G is the Pickands' dependence function (Pickands, 1981 and Resnick, 1987; Chapter 5), defined as

$$A(t) = 2\int_0^1 \max\{wt, (1-w)(1-t)\}\,\mathrm{d}H(w) \qquad (5.2.15)$$

so that G is given in terms of A as

$$G(x, y) = \exp\left\{-\left(\frac{1}{x} + \frac{1}{y}\right) A\left(\frac{y}{x+y}\right)\right\}.$$

The property that G is a distribution function and the moment condition on H in (5.2.3) requires that A be a convex function on $[0,1]$ satisfying $\max(t, 1-t) \leq A(t) \leq 1$. Noting that

$$H(t) = \frac{1 - A'(1-t)}{2}, \tag{5.2.16}$$

at all points for which A is differentiable, then it is clear that G, A and H are all uniquely determined by the specification of any one of them. Furthermore, for χ as defined in equation (5.1.1), $\chi = 2(1 - A(0.5))$ and so provides a natural simple measure of asymptotic dependence.

If X_1 and X_2 are asymptotically independent, then $\chi = 0$ and H places all of its mass on the endpoints of the interval $[0,1]$, and equivalently $G(x,y) = \exp(-1/x - 1/y)$ and $A(t) = 1$ for $t \in [0,1]$. When X_1 and X_2 are asymptotically dependent, $\chi > 0$ and broadly speaking the larger χ the stronger the asymptotic dependence between X_1 and X_2. The cases of stronger asymptotic dependence arise when H places mass closer to the centre of $[0,1]$ in which case $A(t)$ is closer to the bounding curve $\max(t, 1-t)$ for $t \in [0,1]$.

5.3 Heffernan and Tawn method for bivariate tail estimation

Let $(Y_1, Y_2) = (\log X_1, \log X_2)$ so that the random variable (Y_1, Y_2) has Gumbel margins. In the bivariate case, Heffernan and Tawn (2004) assume the existence of normalising functions $a_{|1}(y_1), a_{|2}(y_2)$ and $b_{|1}(y_1), b_{|2}(y_2)$, which can be chosen such that the residuals $Z_{|1}$ and $Z_{|2}$ defined by

$$Z_{|1} = \frac{Y_2 - a_{|1}(y_1)}{b_{|1}(y_1)} \text{ and } Z_{|2} = \frac{Y_1 - a_{|2}(y_2)}{b_{|2}(y_2)}$$

have non-degenerate limit distributions $D_{|1}$ and $D_{|2}$ such that

$$\begin{aligned} &\lim_{y_1\to\infty} \Pr\{Y_2 \le a_{\,|\,1}(y_1) + b_{\,|\,1}(y_1) z_{\,|\,1} \,|\, Y_1 = y_1\} = D_{\,|\,1}(z_{\,|\,1}), \\ \text{and} \quad &\lim_{y_2\to\infty} \Pr\{Y_1 \le a_{\,|\,2}(y_2) + b_{\,|\,2}(y_2) z_{\,|\,2} \,|\, Y_2 = y_2\} = D_{\,|\,2}(z_{\,|\,2}). \end{aligned}$$

The variables $Z_{\,|\,1}$ and Y_1 (equivalently $Z_{\,|\,2}$ and Y_2) are independent as Y_1 (equivalently Y_2) approaches its limit. Further details of the required normalising constants are given by Heffernan and Tawn (2004) and Heffernan and Resnick (2007).

Using the Heffernan and Tawn model, we derive the conditional probability of being in the general set $B^* \subset \mathbb{R}^2$, given that the first component of (Y_1, Y_2) exceeds some large threshold v, *i.e.* $\Pr\{(Y_1, Y_2) \in B^* \,|\, Y_1 > v\}$. Under the assumption that the Heffernan and Tawn model holds for $Y_1 > v$ and using the limiting independence of $Z_{\,|\,1}$ and Y_1 we have, for large v,

$$\begin{aligned} \Pr\{(Y_1, Y_2) \in B^* \,|\, Y_1 > v\} &= \int_v^\infty \Pr\{(Y_1, Y_2) \in B^* \,|\, Y_1 = y_1\} f_{Y_1\,|\,Y_1>v}(y_1)\,\mathrm{d}y_1 \\ &\approx \int_v^\infty \Pr\{(y_1, a_{\,|\,1}(y_1) + b_{\,|\,1}(y_1) Z_{\,|\,1}) \in B^*\} f_{Y_1\,|\,Y_1>v}(y_1)\,\mathrm{d}y_1 \end{aligned} \tag{5.3.1}$$

where $f_{Y_1\,|\,Y_1>v}$ is the conditional density function of $Y_1 \,|\, Y_1 > v$. Let $\{y_{1j}\}$ be the n_v points whose first component exceeds the threshold v where $n_v = \sum_{k=1}^n I_{\{Y_{1k}>v\}}$ and I is the indicator function. To estimate probability (5.3.1) we first approximate the distribution of $Z_{\,|\,1}$ by the empirical distribution of the n_v residuals $\{z_{\,|\,1,j}\}$ associated with the points $\{y_{1j}\}$. Keeping $y_1 > v$ fixed, we can then estimate the probability that $(y_1, a_{\,|\,1}(y_1) + b_{\,|\,1}(y_1) Z_{\,|\,1})$ lies in B^* by

$$\hat{\Pr}\{(y_1, a_{\,|\,1}(y_1) + b_{\,|\,1}(y_1) Z_{\,|\,1}) \in B^*\} = \frac{1}{n_v} \sum_{i=1}^{n_v} I_{\left[(y_1, a_{\,|\,1}(y_1) + b_{\,|\,1}(y_1) z_{\,|\,1,i}) \in B^*\right]}.$$

Using a similar empirical estimate for the distribution of $Y_1|Y_1 > v$ the required conditional probability that (Y_1, Y_2) is in the set B^* is then estimated by

$$\hat{\Pr}\{(Y_1, Y_2) \in B^* \,|\, Y_1 > v\} \;=\; \frac{1}{n_v^2} \sum_{j=1}^{n_v} \sum_{i=1}^{n_v} I_{\left[(y_{1j}, a_{\,|\,1}(y_{1j}) + b_{\,|\,1}(y_{1j}) z_{\,|\,1,i}) \in B^*\right]}.$$

A similar estimate holds when conditioning on $Y_2 > v$. In practice the normalising functions $a_{\,|\,1}$, $a_{\,|\,2}$, $b_{\,|\,1}$ and $b_{\,|\,2}$ must also be estimated, see Heffernan and Tawn (2004) for details of how to do this.

We use the probability integral transform to transform back to unit Fréchet margins. We first transform the set B^* and the threshold v. Since B^* is an arbitrary set on $\mathbb{R}^2$ it can be transformed to the set $B = \{\exp\{\mathbf{y}\} : \mathbf{y} \in B^*\} \subset \mathbb{R}_+^2 \backslash \{\mathbf{0}\}$. Similarly, the threshold v is transformed to $u = \exp\{v\}$. Since this transformation is strictly monotonic, points with first component exceeding the threshold v on Gumbel margins are the same points for which the first component exceeds the threshold u on Fréchet margins, hence $n_v = n_u = \sum_{k=1}^{n} I_{\{X_{1k} > u\}}$. Thus the estimated probability that (X_1, X_2) is in the set B, given that the first component exceeds the threshold u, is

$$\hat{\Pr}\{(X_1, X_2) \in B \,|\, X_1 > u\} \;=\; \frac{1}{n_u^2} \sum_{j=1}^{n_u} \sum_{i=1}^{n_u} I_{\left[(\exp\{y_{1j}\}, \exp\{a_{\,|\,1}(y_{1j}) + b_{\,|\,1}(y_{1j}) z_{\,|\,1,i}\}) \in B\right]}. \tag{5.3.2}$$

This estimate holds regardless of the extremal dependence structure of the variables (X_1, X_2).

5.4 New nonparametric estimators

In Section 5.3 we used the Heffernan and Tawn model to find an estimate of $\Pr\{(X_1, X_2) \in B \,|\, X_1 > u\}$ as $u \to \infty$. In the special case of asymptotic dependence this estimate can be simplified further since, as shown by Heffernan and Tawn (2004), in this case the normalising functions are given by $a_{\,|\,1}(y_1) = y_1$ and

$b_{\,|\,1}(y_1) = 1$. Thus we obtain the estimate

$$\hat{\Pr}\{(X_1, X_2) \in B \,|\, X_1 > u\} \;=\; \frac{1}{n_u^2} \sum_{j=1}^{n_u} \sum_{i=1}^{n_u} I_{\left[(\exp\{y_{1j}\}, \exp\{y_{1j} + z_{\,|\,1,i}\}) \in B\right]}. \quad (5.4.1)$$

When $B = B_1^{(t)}$ (equivalently, $B = B_2^{(t)}$) the estimate of equation (5.4.1) can be simplified further as follows. Consider the observations for which the indicator function is non-zero, *i.e.* for which $(\exp\{y_{1j}\}, \exp\{y_{1j} + z_{\,|\,1,i}\}) \in B$. For the case $B = B_1^{(t)}$ this expression is equivalent to

$$\exp\{y_{1j}\}/(\exp\{y_{1j}\} + \exp\{y_{1j} + z_{\,|\,1,i}\}) < t$$

which, following multiplication of both the numerator and denominator on the left hand side by $\exp\{y_{1i} - y_{1j}\}$ and transformation by the probability integral transform to the unit Fréchet marginal space, is equivalent to

$$w_i \equiv x_{1i}/(x_{1i} + x_{2i}) < t,$$

using the definition of W given in equation (5.2.1). Hence equation (5.4.1) can be rewritten for $B = B_1^{(t)}$ as

$$\begin{aligned} \hat{\Pr}\{(X_1, X_2) \in B_1^{(t)} \,|\, X_1 > u\} &= \frac{1}{n_u^2} \sum_{j=1}^{n_u} \sum_{i=1}^{n_u} I_{[w_i < t]} \\ &= \frac{\sum_{i=1}^{n} I_{[x_{1i} > u, w_i < t]}}{\sum_{i=1}^{n} I_{[x_{1i} > u]}}. \qquad (5.4.2) \end{aligned}$$

The additional constraint in the indicator function on the numerator of the second expression ensures that we continue to count only those variables whose first component is a threshold exceedance even though the sum is taken over all n variables. A similar expression may be found for $\hat{\Pr}((X_1, X_2) \in B_2^{(t)} \,|\, X_2 > u)$. Since the sets $[X_1 > u]$ and $[(X_1, X_2) \in B_1]$ are equivalent, we can combine these estimates using equation (5.2.11) to obtain our first empirical estimator of the spectral measure

H. Allowing for different marginal thresholds (u_1 and u_2), this estimator is

$$\hat{H}_1(t) = \frac{1}{2}\left\{\frac{1}{\sum_{i=1}^{n} I_{[X_{1i}>u_1]}}\sum_{i=1}^{n} I_{[X_{1i}>u_1 \& W_i<t]} + \frac{1}{\sum_{i=1}^{n} I_{[X_{2i}>u_2]}}\sum_{i=1}^{n} I_{[X_{2i}>u_2 \& W_i<t]}\right\}. \tag{5.4.3}$$

Note that the estimator $\hat{H}_1$ given in equation (5.4.3) can also be written as follows

$$\hat{H}_1(t) = \frac{1}{2}\left[\frac{\hat{\Lambda}(B_1^{(t)})}{\hat{\Lambda}(B_1)} + \frac{\hat{\Lambda}(B_2^{(t)})}{\hat{\Lambda}(B_2)}\right] \tag{5.4.4}$$

where Λ is the integrated intensity function (5.2.4) of the limiting Poisson process P and $\hat{\Lambda}$ is the empirical estimate of Λ.

Analogously, an empirical estimator of the dependence function A in equation (5.2.15) follows naturally from (5.4.3). For $j = 1, 2$, let n_{u_j} be the number of variables whose j^{th} component exceeds the associated marginal threshold u_j. Using the estimator $\hat{H}_1$ of equation (5.4.3) it is clear that each variable (X_{1i}, X_{2i}) has point mass $(m_{1i}^* + m_{2i}^*)/2$ where

$$m_{ji}^* = \frac{1}{n_{u_j}} I_{[X_{ji}>u_j]}, \qquad i = 1, \ldots, n, \ j = 1, 2.$$

It then follows from equation (5.2.15) that the empirical estimator of A is

$$\hat{A}_1(t) = \sum_{i=1}^{n} (m_{1i}^* + m_{2i}^*) \max\{tW_i, (1-t)(1-W_i)\}. \tag{5.4.5}$$

In Theorems 5.4.1 and 5.4.2 we show consistency of the estimators $\hat{H}_1(t)$ and $\hat{A}_1(t)$. Proofs can be found in Appendix B. These theorems show that both estimators are unbiased and have variance tending to zero as sample size increases. For the estimator $\hat{H}_1(t)$ we can also prove asymptotic normality; for the estimator $\hat{A}_1(t)$ this result is more complicated and we do not prove it here. However empirical evidence from simulations suggests that $\hat{A}_1(t)$ is indeed asymptotically normal. In both theorems we assume, for large n, that the process $P_n \equiv P$ on the region $\mathbb{R}_+^2 \backslash \{[0, u_1] \times [0, u_2]\}$.

Theorem 5.4.1 *For any fixed* t*,* $0 \leq t \leq 1$*, as* $n \to \infty$*,*

$$\sqrt{n}\left[\hat{H}_1(t) - H(t)\right] \to \mathrm{N}(0, \sigma_t^2(\mathbf{u})), \tag{5.4.6}$$

where σ^2 *is constant and is derived in Appendix B and* $\mathbf{u} = (u_1, u_2)$*.*

Theorem 5.4.2 *For fixed* t*,* $0 \leq t \leq 1$*, as* $n \to \infty$*,*

$$\mathbb{E}[\hat{A}_1(t)] = A(t) \quad \text{and} \quad \mathrm{Var}(\hat{A}_1(t)) = O(n^{-1}).$$

Full expressions for the asymptotic variances of both estimators are given in Appendix B. Figure 5.1 shows some plots of these variances in the case of the spectral measure taking the form of the logistic distribution; this distribution is characterised by a single parameter α which defines the strength of asymptotic dependence (for further details see Section 5.6). We have assumed a sample size of $n = 100000$ and marginal threshold levels of 99%; in this case Figure 5.1 shows that, for a range of parameter values, the variances of both estimators are very small. As expected, for fixed n, the variances increase as u_1 (u_2) increase, since there are fewer data points for use in the inference. To verify our theoretical variance functions, we also estimated the variances of the estimators by simulation, *i.e.* we simulated a number of data sets with the required form of the spectral measure and applied the estimators $\hat{A}_1(t)$ and $\hat{H}_1(t)$. For each t, we then found the sample variance of the estimates; these are also plotted in Figure 5.1. We see that they are very similar to the theoretical variances. Similar plots (not shown) of both the theoretical and simulated expected values of the estimators showed both to be unbiased.

We now introduce a minor modification to the estimator $\hat{H}_1(t)$, since as it is defined in (5.4.3), the estimator does not satisfy moment condition (5.2.3). We propose a linear tilting of this estimator to give the modified estimator $\hat{H}(t)$ with

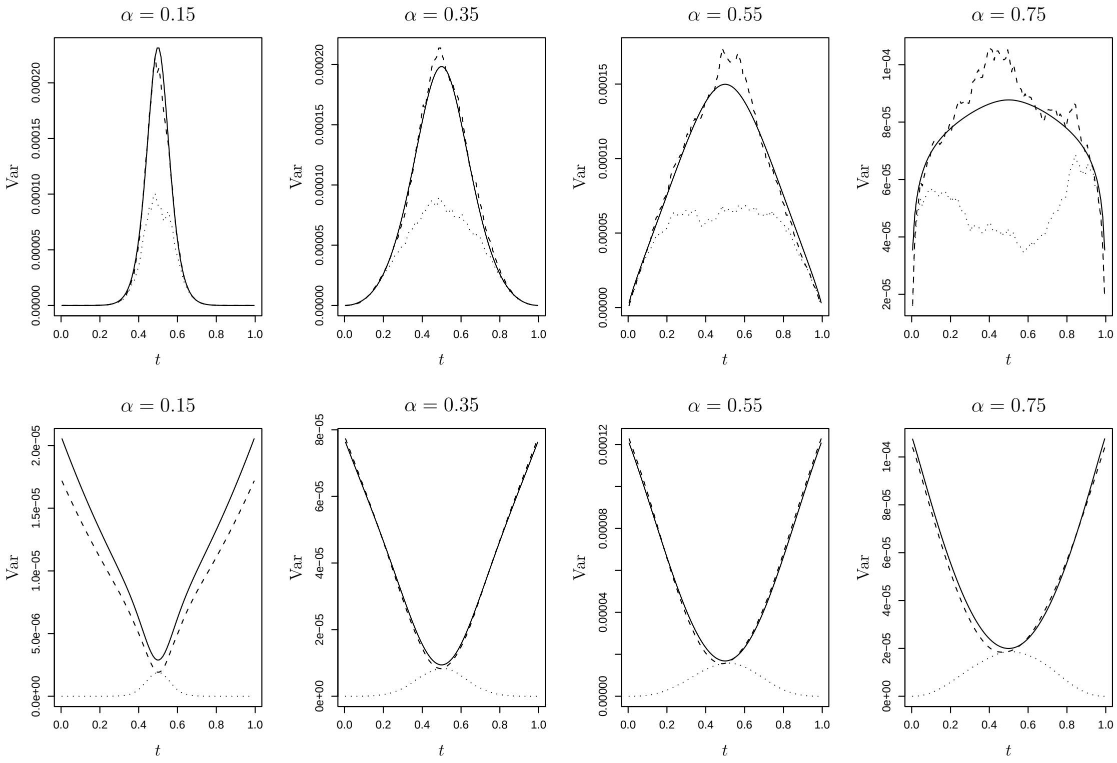


Figure 5.1: Theoretical (full lines) and simulated variances of untilted (dashed lines) $\hat{H}_1(t)$ and $\hat{A}_1(t)$ and tilted (dotted lines) $\hat{H}(t)$ and $\hat{A}(t)$ estimators for the logistic dependence function with parameters $\alpha = 0.15, 0.35, 0.55, 0.75$. Variances for H estimators are shown on the top row and for A estimators on the bottom row. In these plots the sample size is $n = 100000$ and thresholds were fixed at the marginal 99% quantile.

the following property:

$$d\hat{H}(t) = (\tilde{\alpha} + \tilde{\beta}t)\,\mathrm{d}\hat{H}_1(t) \tag{5.4.7}$$

where constants $\tilde{\alpha}$ and $\tilde{\beta}$ are chosen to ensure that $\hat{H}(t)$ satisfies (5.2.3) and has mass 1 on $[0, 1]$. The values of $\tilde{\alpha}$ and $\tilde{\beta}$ that satisfy these constraints are:

$$\tilde{\alpha} = \frac{S - T}{S^2 - T} \quad \text{and} \quad \tilde{\beta} = \frac{2(S - 1)}{S^2 - T} \tag{5.4.8}$$

where

$$S = \int_0^1 w \, \mathrm{d}\hat{C}_1(w) + \int_0^1 w \, \mathrm{d}\hat{C}_2(w) \text{ and } T = 2\left(\int_0^1 w^2 \, \mathrm{d}\hat{C}_1(w) + \int_0^1 w^2 \, \mathrm{d}\hat{C}_2(w)\right), \tag{5.4.9}$$

with $\hat{C}_1(\cdot)$ and $\hat{C}_2(\cdot)$ being empirical estimates of the functions given in equations (5.2.9) and (5.2.10). In practice the moments determining $\tilde{\alpha}$ and $\tilde{\beta}$ are estimated by their sample values, using data above thresholds u_1 and u_2 as appropriate. This modification to $\hat{H}_1(t)$ results in the following estimator $\hat{H}(t)$:

$$\hat{H}(t) = \frac{1}{2}\left\{\frac{1}{\sum_{i=1}^n I_{[X_{1i}>u_1]}} \sum_{i=1}^n (\tilde{\alpha} + \tilde{\beta} W_i) I_{[X_{1i}>u_1 \& W_i<t]} + \right. \\ \left. \frac{1}{\sum_{i=1}^n I_{[X_{2i}>u_2]}} \sum_{i=1}^n (\tilde{\alpha} + \tilde{\beta} W_i) I_{[X_{2i}>u_2 \& W_i<t]}\right\}. \tag{5.4.10}$$

Analogously, we can modify our empirical estimator of the dependence function A given in equation (5.4.5) which follows naturally from (5.4.10). To do this, we simply modify the point mass $m_{1i} + m_{2i}$ to take account of the tilting, so that we now have

$$m_{ji} = \frac{1}{n_{u_j}} (\tilde{\alpha} + \tilde{\beta} W_i) I_{[X_{ji}>u_j]}, \quad i = 1, \ldots, n, \; j = 1, 2.$$

It then follows from equation (5.2.15) that the empirical estimator of A is

$$\hat{A}(t) = \sum_{i=1}^n (m_{1i} + m_{2i}) \max\{tW_i, (1-t)(1-W_i)\}. \tag{5.4.11}$$

It is reasonably straightforward to show that the tilting parameters tend to their asymptotic values as $n \to \infty$, *i.e.* that $\alpha \to 1$ and $\beta \to 0$. From this it follows that, in the limit, $\mathrm{d}\hat{H} \sim \mathrm{d}\hat{H}_1$, and similarly for $\hat{A}(t) \sim \hat{A}_1(t)$, so that the tilted and untilted estimators are asymptotically equivalent.

As with the untilted estimators, we conducted a simulation study to estimate the variances of the tilted estimators; thus for each of the simulated data sets we fitted the tilted estimators and then, for each t, found the sample variance for each of $\hat{H}(t)$ and $\hat{A}(t)$. Plots of these variances are shown, for $n = 100000$ and 99% marginal threshold levels in Figure 5.1. They show that the variances of both of the tilted estimators are much smaller than those of their counterparts, $\hat{H}_1(t)$ and $\hat{A}_1(t)$, which we know to be consistent by Theorems 5.4.1 and 5.4.2; this confirms that tilting only improves the accuracy of the estimators. Further the tilted estimator for the Pickands' dependence function drastically improves the variance estimate at the end-points of the range of t (*i.e.* when t is close to 0 or 1). This is because the tilting forces the estimator to satisfy the conditions $\hat{A}(0) = \hat{A}(1) = 1$ or, equivalently, that the mean of $\hat{H}(t)$ is 0.5.

Implementation of these estimators requires the choice of thresholds u_1 and u_2. We follow the form of diagnostics proposed by Heffernan and Tawn (2004), which check for the stability of the fitted model above the selected threshold. We first check for independence between the angular variables W_i and the conditioning variables, for values of X_{1i} and X_{2i} above their respective thresholds. Thus for a given data set, we plot w_i against x_{1i} for $x_{1i} > u_1$ and against x_{2i} for $x_{2i} > u_2$. Dependence of the w_i's on the x_{1i}'s or x_{2i}'s indicates that either the associated threshold is not sufficiently high, or that the limiting BEV distribution has a spectral measure which puts mass on the endpoints of the interval $[0, 1]$. For proposed thresholds u_1 and u_2, we also check that the estimated dependence functions do not differ greatly when the thresholds are raised still higher, although clearly some changes due to random variation will arise.

5.5 Existing estimators

We concentrate on nonparametric estimators for the dependence functions defined in Section 5.2. For insight into existing parametric estimators see Tawn (1988), Smith *et al.* (1990), Shi *et al.* (1992) and Stephenson and Tawn (2004). We first present details of the estimator offered by Capéraà and Fougères (2000). This estimator was shown by Capéraà and Fougères to perform similarly in terms of L_1 errors to the estimators of Einmahl *et al.* (1993, 1997 and 2001) and Joe *et* *al.* (1992) in the case of strong dependence, and to out-perform these estimators when dependence is weak.

Capéraà and Fougères use the L_1 norm and define R_i and W_i as in equation (5.2.1). They propose an initial estimator of the Pickands dependence function in (5.2.15) to be

$$\bar{A}_1(t) = \frac{2}{k_n} \sum_{i=1}^{n} \max\{tW_i, (1-t)(1-W_i)\} I_{[R_i \geq 1/k_n]}, \tag{5.5.1}$$

where $\{k_n, n \in \mathbb{N}\}$ is a sequence of integers such that $k_n \to \infty$ and $k_n/n \to 0$ as $n \to \infty$. To ensure that their estimate is a consistent convex estimator of A satisfying $\max(t, 1-t) \leq A(t) \leq 1$, Capéraà and Fougères propose the following modification to their estimator

$$\bar{A}(t) = \max\left\{t, 1-t, \bar{A}_1(t) + (2t-1)(1-2\Gamma_n)\right\}. \tag{5.5.2}$$

Here $\Gamma_n = 1/k_n \sum_{i=1}^{n} W_i I_{[R_i \geq 1/k_n]}$. This modification is equivalent to our tilting of the conditional estimator. We obtain the corresponding estimator of H using the relation of equation (5.2.16). For small samples Capéraà and Fougères propose a bias correction amounting to a down-weighting of the contributions of the central W_i's occurring with the R_i's exceeding $1/k_n$. Note that we follow Capéraà and Fougères and down-weight the W_i corresponding to the 30% largest radial order statistics above the threshold. This is consistent with the "empirically optimal" proportions found by Capéraà and Fougères (2000).

The second non-parametric estimator which we shall look at is that proposed by Abdous and Ghoudi (2005). They observe that all existing nonparametric estimators are empirical estimators of the spectral measure H, for data with radial component above a high threshold. This is slightly different to our estimator, in which the data used for estimation are those with at least one component exceeding a marginal threshold (*i.e.* either $X_1 > u_1$ or $X_2 > u_2$), rather than those with a large radial component. The main differences between the existing methods arise from different choices of norm, which influence the precise form of H, and the exact approach taken for its estimation. Further minor differences arise from different methods being adopted to ensure the satisfaction of moment condition (5.2.3) and different approaches to threshold choice.

Abdous and Ghoudi assume general margins, F_j $(j = 1, 2)$. Of the characterisations of A studied by Abdous and Ghoudi we present the kernel-based estimator of Abdous *et al.* (1999) which takes the naïve form

$$\tilde{A}_1(t) = \frac{1}{l_n} \sum_{i=1}^{n} I_{[\zeta_{t,i} \leq l_n/n]}. \tag{5.5.3}$$

where

$$\zeta_{t,i} = \begin{cases} 1 - \max\{F_1(X_{1i})^{1/t}, F_2(X_{2i})^{1/(1-t)}\} & \text{for } t \in (0,1), \\ 1 - F_2(X_{2i}) & \text{for } t = 0, \\ 1 - F_1(X_{1i}) & \text{for } t = 1. \end{cases} \tag{5.5.4}$$

The marginal distribution functions are replaced by their empirical counterparts when unknown. We assume that the margins are identical unit Fréchet distribution functions, using the probability integral transform if necessary.

This estimator, like all the other existing nonparametric estimators, requires a choice of threshold. For the estimators of Capéraà and Fougères (2000) and Abdous *et al.* (1999) and our conditional estimator this is determined by the values of k_n, l_n and u_1 (u_2). One diagnostic developed for this purpose is given by Stărică (1999). However, Abdous and Ghoudi (2004) propose a method of automatic threshold selection. Observing that all the existing estimators presented are approximated by the derivative of a distribution function close to zero, they suggest using local polynomial fitting and kernel estimation to generate an estimator for A. If m is the degree of the polynomial, K the kernel and h the bandwidth used for the kernel estimation, then Abdous and Ghoudi propose updating the estimator $\tilde{A}_1$ of equation (5.5.3), to give

$$\tilde{A}(t) = \frac{1}{n}\sum_{i=1}^{n}\int_{\zeta_{t,i}}^{1} K^{[m]}(v,h)\,\mathrm{d}v. \tag{5.5.5}$$

The function $K^{[m]}(v,h) = e_1^T S_m^{-1}[v,\ldots,v^m]^T K\left(v/h\right)/h$ is an equivalent kernel, where the vector e_1 has value 1 for the first component and zero thereafter and the $(i,j)th$ element of the matrix S is given by the $(i+j)th$ moment of $K\left(v/h\right)/h$.

This estimate depends on the choice of polynomial degree m, kernel K and bandwidth h. Since only points within the bandwidth are used in the estimation procedure, the choice of a bandwidth is equivalent to choosing a threshold. Abdous and Ghoudi suggest automatic selection of the bandwidth (equivalently, threshold) using the L_1-double kernel method, first proposed by Devroye (1989). The estimator is constrained to fulfil the properties of the dependence function using either convex hulls or smoothing splines. Abdous and Ghoudi do not discuss estimation of the spectral measure H. Using equation (5.2.16) we propose using finite differencing methods to obtain an estimate of H given an estimate $\tilde{A}$.

5.6 Simulation study

We carry out two studies to compare the performance of the conditional estimators for dependence functions A and H with that of the Capéraà and Fougères estimator $\bar{A}$ and the Abdous and Ghoudi estimator $\tilde{A}$.

5.6.1 Study design

For data arising from a distribution F in the domain of attraction of a BEV distribution G, the performance of any estimator for H or A is principally driven by two features. First is the rate of convergence of the Poisson process P_n to its limit P as $n \to \infty$. This is determined exclusively by the underlying distribution F; we are not interested in the effect of this feature as it will be the same for all estimators. For comparison with the Capéraà and Fougères estimator we therefore simulate directly from the limiting Poisson process. However, for comparison with the Abdous and Ghoudi estimator, due to its self-selecting threshold feature, we must simulate from the full distribution.

The second feature affecting the performance of the estimators is the rate of convergence of each estimator to its limiting distribution. This is of interest and is driven by the number of points above the estimation threshold(s). Due to the self-selecting threshold, this number is determined within the Abdous and Ghoudi estimator, whereas it is determined by the prior choice of threshold for both the Capéraà and Fougères and our conditional estimators. To allow for a fair comparison between the two models in the Abdous and Ghoudi study, we apply the conditional estimator to exceedances of several thresholds. Such examination of a range of thresholds is regularly employed during the threshold selection component in an extreme value analysis.

For comparison with the Capéraà and Fougères estimator, we simulate 1000 independent replicate data sets from the limiting point process P, and retain a total of m points above the threshold(s) for estimation. To allow a fair comparison of the two methods this means that the (R_i, W_i) points giving the largest M radial order statistics are used for the Capéraà and Fougères method. We then select a threshold $u = u_1 = u_2$ so that exactly M ($M = 50, 200, 1000$ in our case) points lie above either threshold, these are used for estimation with our conditional method. The different definitions of thresholds for the two methods mean that some points will be included in one analysis but not in the other.

For comparison with the Abdous and Ghoudi estimator $\tilde{A}$, we simulate 1000 independent replicate data sets, of 100000 points each, from the full distribution. Three thresholds (empirical 90-, 95- and 99% quantiles) were tested for the conditional estimator. In applying the Abdous and Ghoudi estimator, we followed their choice of the Epanechnikov kernel for kernel smoothing and used first and second order polynomials ($m = 1, 2$) for selection of the optimal bandwidth. To constrain the function to be a dependence function we used convex hulls and then used finite differencing of the estimate of A to estimate H.

We used a range of parametric forms of the spectral density h and a variety of strengths of dependence for each form of h. All of the spectral densities considered place all their mass on the interior of the interval $[0, 1]$. The spectral densities we used are:

Logistic: Gumbel (1960). This is a symmetric density with a single parameter $\alpha \in [0, 1]$. Perfect dependence is obtained in the limit as $\alpha \to 0$ and exact independence is given by $\alpha = 1$. For $\alpha < 0.5$, h is unimodal, whereas for increasingly large values of $\alpha > 0.5$, the density places greater mass towards the ends of the interval $[0, 1]$. The density is given by

$$h(w) = \frac{1}{2}\left(\frac{1}{\alpha} - 1\right) w^{-1-1/\alpha}(1-w)^{-1-1/\alpha}\{w^{-1/\alpha} + (1-w)^{-1/\alpha}\}^{\alpha-2}$$

Hüsler-Reiss: Hüsler-Reiss (1989). This symmetric model has a

single parameter $\lambda > 0$. Perfect dependence and exact independence are obtained as limiting cases as $\lambda \to \infty$ and $\lambda \to 0$ respectively. The density is given by

$$h(w) = \frac{a(w)}{2w^2(1-w)} + \frac{a(1-w)}{2w(1-w)^2}$$

where $a(w) = \lambda\phi(1/\lambda + \lambda/2 \log\{(1-w)/w\})/2 + \lambda^2\phi'(1/\lambda + \lambda/2 \log\{(1-w)/w\})/4$, ϕ is the standard Gaussian density function and ϕ' its first derivative.

Dirichlet: Coles and Tawn (1991). This model has two parameters $\alpha_1 > 0$ and $\alpha_2 > 0$. For $\alpha_1 = \alpha_2$ this model is symmetric and for $\alpha_1 \neq \alpha_2$ it is asymmetric, allowing for nonexchangeability of the variables. Perfect dependence and exact independence are obtained as limiting cases as $\alpha_1 = \alpha_2 \to \infty$ and

$\alpha_1 = \alpha_2 \to 0$ respectively. The density is given by

$$h(w) = \frac{\alpha_1\alpha_2\Gamma(\alpha_1+\alpha_2+1)}{2\Gamma(\alpha_1)\Gamma(\alpha_2)k(w)^3}\left(\frac{\alpha_1 w}{k(w)}\right)^{\alpha_1-1}\left(\frac{\alpha_2(1-w)}{k(w)}\right)^{\alpha_2-1}$$

where $k(w) = \alpha_1 w + \alpha_2(1-w)$.

We used four sets of parameter values to explore the performance of each of the estimators at various levels of dependence within the class of asymptotic dependence. The four parameterisations used correspond to having an $A(0.5) = 0.555, 0.637, 0.732$ and 0.841. We note that a lower value of $A(0.5)$ corresponds to stronger dependence within the class of asymptotic dependence. These choices of $A(0.5)$ correspond to the logistic parameter α taking the values 0.15, 0.35, 0.55 and 0.75.

For each value of α, the equivalent parameter values for the other distributions are as follows. For the Hüsler-Reiss distribution, we take $\lambda = 1/\Phi^{-1}(2^{\alpha-1})$. For the two parameter Dirichlet distribution, a further constraint on the parameters is needed for identifiability. We used three different constraints D1:$\alpha_2 = \alpha_1$, D2:$\alpha_2 = 2\alpha_1$ and D3:$\alpha_2 = 4\alpha_1$. These three constraints allowed us to explore different degrees of nonexchangeability of the variables. For each value of the logistic parameter α, the value of α_1 was found numerically under each of constraints D1, D2 and D3.

5.6.2 Results

We summarise the output of the two studies by looking at the median and the 2.5% and 97.5% quantiles of the sampling distributions of $\hat{A}(t) - A(t)$ and $\bar{A}(t) - A(t)$ (Capéraà and Fougères comparison, see Figure 5.2) and of $\hat{A}(t) - A(t)$ and $\tilde{A}(t) - A(t)$ (Abdous and Ghoudi comparison, see Figure 5.3) for $t \in [0,1]$. We discuss first the Capéraà and Fougères comparison.

For each dependence structure and each distribution a pilot study of 100 simulated data sets with $M = 1000$ was made to observe the proportion of the data sets that were used in both methods. For example, for the logistic distribution with parameter fixed at $\alpha = 0.15, 0.35, 0.55$ and 0.75 the median proportion of the points used in the conditional method that were also used by the Capéraà and Fougères method are 0.558, 0.643, 0.733 and 0.841. Corresponding interquartile ranges are 0.0233, 0.0235, 0.0153 and 0.0185. Results for the remaining distributions are similar. The number of overlapping points depends on the underlying strength of dependence with a greater overlap between the sets of points used under the two methods for weaker dependence structures. There is very little difference in the number of overlapping points between the different distributions once the strength of dependence is fixed.

The value of M appears to have little influence on the relative performance of the two estimators. Results for $M = 1000$ for all of the distributions under each parameterisation are shown in Figure 5.2. Similar plots for $M = 50, 200$ (not shown) have vertical axes with larger ranges as expected, but there is little or no material difference in the shapes of the plotted curves.

For smaller values of α, corresponding to stronger dependence, the Capéraà and Fougères estimator is more variable than the conditional estimator, particularly away from the centre of the interval. For such α the methods are comparable in the very centre of the interval. This is where the effect of the Capéraà and Fougères bias correction is evident. The relatively poor performance of the Capéraà and Fougères estimator appears to be due to the nature of the correction applied in equation (5.5.2), adding $(2t - 1)(1 - 2\Lambda_n)$ to $\bar{A}_1(t)$.

For values of the parameters corresponding to weaker dependence the two methods seem to perform comparably. The Capéraà and Fougères estimator is slightly less variable than the conditional estimator for the very weakest dependence considered. For all parameter values the median lines for both methods are very close so that the main differences between the performances correspond to differences in variability rather than in bias. There is little systematic difference between the output for different underlying distributions.

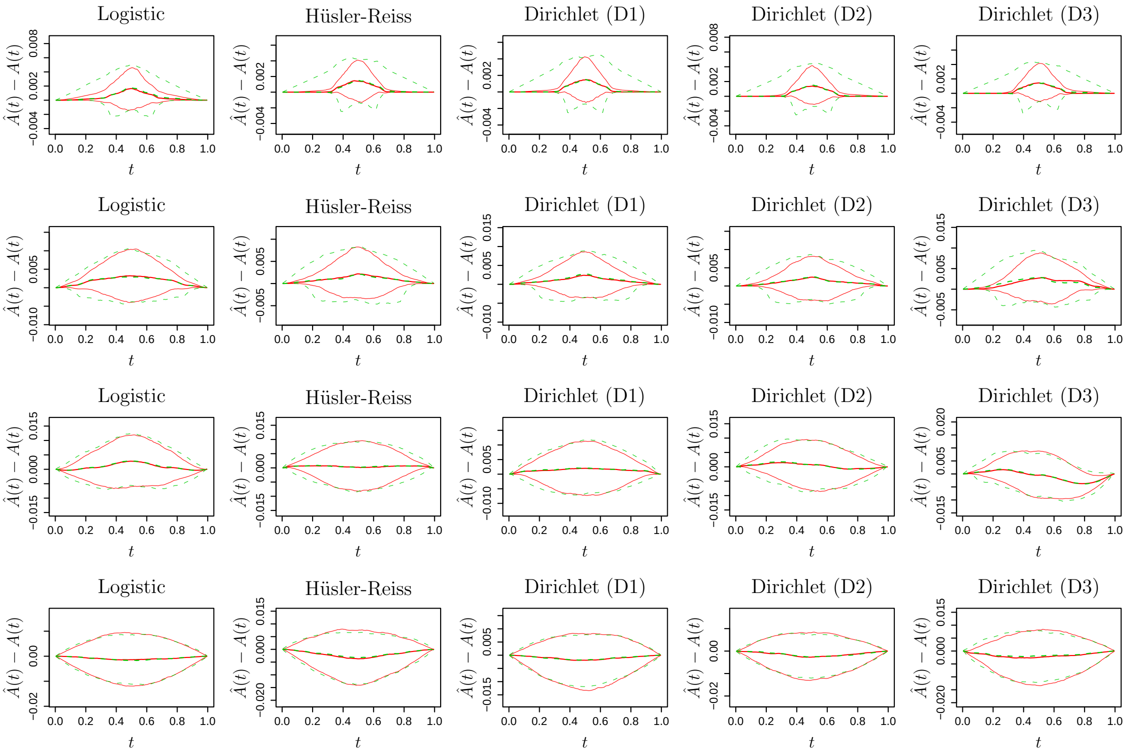


Figure 5.2: Pointwise median and 2.5% and 97.5% quantiles of sampling distribution of proposed estimator $\hat{A}(t) - A(t)$ (solid lines) and the Capéraà and Fougères estimator $\bar{A}(t) - A(t)$ (dashed lines). All plots show output for $M = 1000$ data points used for estimation by both methods. The five columns show left to right Logistic, Hüsler-Reiss, Dirichlet (D1), Dirichlet (D2) and Dirichlet (D3) distributions. The four rows show top to bottom parameter values corresponding to Logistic parameter $\alpha = 0.15, 0.35, 0.55, 0.75$.

We now go on to discuss the relative performances of our conditional estimator and the Abdous and Ghoudi estimator. Results for all of the distributions for each of the four parameterisations are shown in Figure 5.3. This shows that the factor most strongly influencing the relative performance of the estimators is the strength of dependence. The Abdous and Ghoudi estimator shows more bias than the conditional estimator, especially at higher levels of dependence ($\alpha = 0.15, 0.35$). In all cases the Abdous and Ghoudi estimator tends to overestimate the dependence, especially in the midrange of t. This estimator is also much more variable than the conditional estimator at all levels of dependence. The conditional estimator shows a decrease in bias but a corresponding increase in variance as the threshold is increased. This is as we would expect from the standard bias-variance trade-off (higher thresholds approximate the asymptotics better, but use fewer data points).

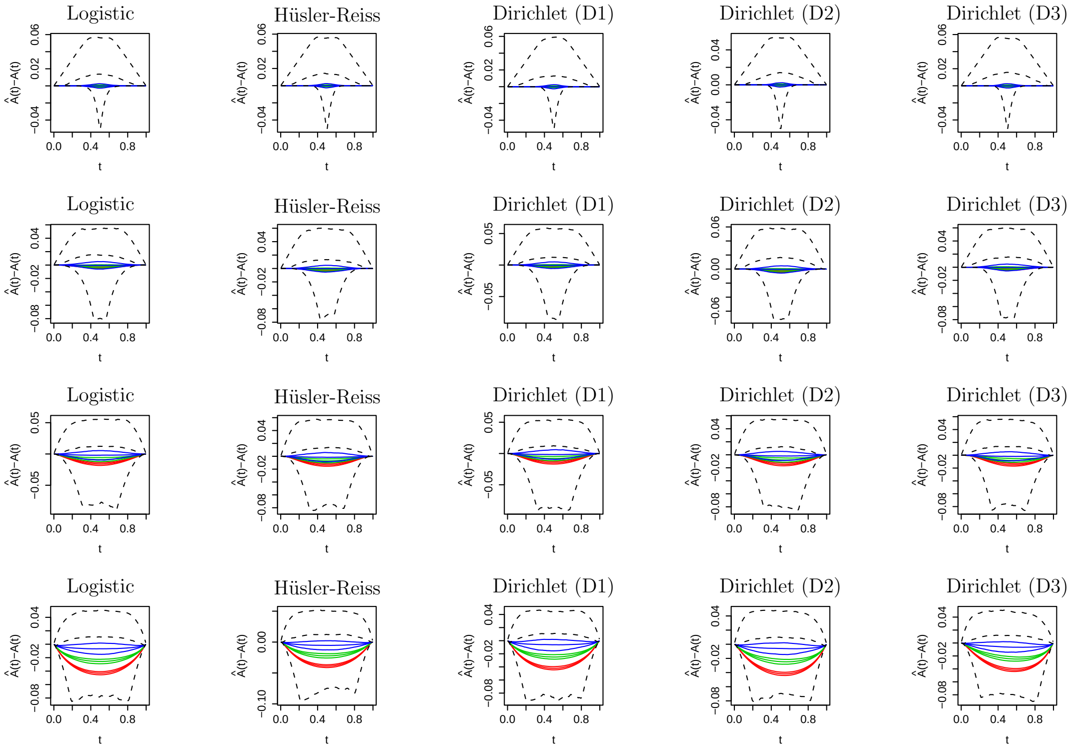


Figure 5.3: Pointwise median and 2.5% and 97.5% quantiles of sampling distribution of proposed estimator $\hat{A}(t) - A(t)$ (solid lines) and the Abdous and Ghoudi estimator $\tilde{A}(t) - A(t)$ (dashed lines). All plots show output from the 1000 replications used for estimation by both methods. The five columns show left to right Logistic, Hüsler-Reiss, Dirichlet (D1), Dirichlet (D2) and Dirichlet (D3) distributions. The four rows show top to bottom parameter values corresponding to Logistic parameter $\alpha = 0.15, 0.35, 0.55, 0.75$.

We also examined the performance of the proposed estimator for the spectral measure $\hat{H}(t)$ given in (5.4.10) relative to the estimator $\tilde{H}(t)$ that follows from the finite differencing of the Abdous and Ghoudi estimator $\tilde{A}(t)$. Our conclusions were very similar to those highlighted by the results of Figure 5.3. This is hardly surprising since the estimates of the two dependence functions are functionally linked. It is interesting to note that the Abdous and Ghoudi estimator performed poorly in estimating H at the ends of the (0,1) interval, consistently overestimating H when t is close to 0 and underestimating it when t is close to 1. This is possibly due to the convex hull and finite differencing techniques which resulted in a step function estimate of H, whereas the conditional estimator returns a smooth estimate. The erroneous placing of mass at the points $t = 0$ and 1 by the

Abdous and Ghoudi estimator is emphasised in the plots of Figure 5.4 which show $\hat{H}(t) - H(t)$ and $\tilde{H}(t) - H(t)$, for the logistic distribution only. The conclusions for both the A and H functions, appear to hold regardless of whether the underlying spectral density is uni- or bi-modal, symmetric or asymmetric.

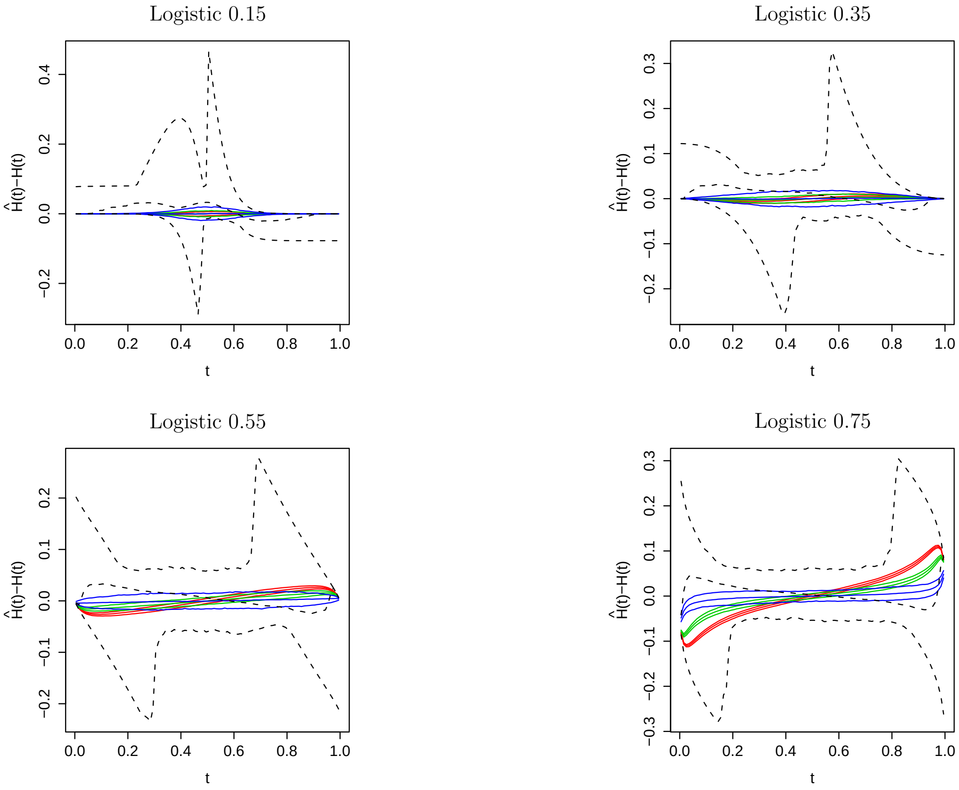


Figure 5.4: Pointwise median and 2.5% and 97.5% quantiles of sampling distribution of proposed estimator $\hat{H}(t) - H(t)$ (solid lines) and the Abdous and Ghoudi estimator $\tilde{H}(t) - H(t)$ (dashed lines). All plots show output from the 1000 replications used for estimation by both methods. The results shown here are for the Logistic distribution, with parameter values $\alpha = 0.15, 0.35, 0.55, 0.75$. Plots for the remaining distributions (not shown) are similar.

We conclude with a further point of interest regarding the automatic bandwidth selection by the Abdous and Ghoudi estimator. Histograms of these bandwidths (not shown) illustrate the wide variation in the bandwidths selected for any given dependence structure, although the range of bandwidths does not seem to vary greatly across distributions. The lower the level of dependence the greater the range of bandwidths selected. For the logistic distribution the median (2.5% and 97.5% quantiles) of the bandwidths selected for the 1000 datasets simulated for the dependence parameters in order of decreasing dependence ($\alpha = 0.15, 0.35, 0.55, 0.75$) were, 0.21 $(0.03, 0.49)$, 0.22 $(0.03, 0.53)$, 0.24 $(0.03, 0.56)$ and 0.31 $(0.03, 0.75)$. Since data exceeding these bandwidths are utilised in the estimation procedure thresholds, it is clear that this estimator favours much lower thresholds than one would intuitively pick for an extreme value analysis. Indeed in all cases the median threshold is lower than any of the thresholds used for the conditional estimator.

5.7 Application to finance data

We now analyse financial indices describing the performance of four national stock exchanges during the years leading up to and following European Economic Monetary Union (EMU) in 1999.

The FTSE 100 is a benchmark index tracking the performance of the London Stock Exchange. We consider data comprising daily values of the FTSE 100 Index on trading days from 1st January 1985 to 12th November 2001, as well as values from US (Standard and Poors 500, equivalently S&P 500), French (CAC 40) and German (DAX 30) indices on the same days. Much of this data was examined in a larger extreme value analysis by Poon *et al.* (2003b). They analysed data going back to the late 1960's but did not focus on the effect of EMU on the extremal behaviour as we do here.

This period is of particular interest as it was during these years that the European currencies preparing for EMU were harmonised. In this analysis, we are interested in the effect of this harmonisation on the extremal properties of the concomitant stock exchange behaviour. As such, we compare the joint extremal behaviour of the German DAX and French CAC; the DAX and the UK FTSE; and the DAX and the US Standard and Poors indices. This gives us three comparisons: the first between two European economies who did join the EMU in 1999; the second between two European economies, one of which did not join the EMU and the third between a European economy joining the EMU and a Western economy outside Europe.

The key dates during these years are: 1st July 1990, preliminary reforms and the beginning of convergence; 1st January 1994, preparation for EMU; 1st January 1999, adoption of the Euro and fixing of exchange rates for countries in EMU. We break the data into four periods accordingly.

As is standard for analyses of such data, we work with daily returns rather than with the raw data (Embrechts *et al.*, 1997). This transformation removes the time trend, giving an approximately stationary time series. We are interested in extreme losses and work with negative returns. The first stage is to transform the negative daily returns to unit Fréchet scale using the rank transform. Denote the negative return variables after transformation to Fréchet scale as $X_{\mathrm{DAX},i}$ *etc.*

The proposed methods are appropriate for data arising as realisations from the asymptotic dependence class of bivariate distributions. Therefore we must verify that our pairs of negative returns are realisations of vector random variables which are members of this class. Clearly it is impossible to ascertain this unequivocally. However, we can check some necessary conditions for membership and a number of such diagnostics exist. Nonparametric estimation of $\chi(u)$ and $\bar{\chi}(u)$ of Coles *et al.* (1999) provides a helpful visual diagnostic for the limiting values of these functions as $u \to 1$. We require the respective limits to be $\chi > 0$ subject to $\bar{\chi} = 1$. Figure 5.5 shows the estimated function $\bar{\chi}(u)$ for the first (1985-1990) and last (1999-2001) of the time periods. These show $\bar{\chi}(u)$ to be tending to 1 as the threshold tends to its limit. Plots of $\chi(u)$ for these data show a positive limit. Similar diagnostic plots of both $\chi(u)$ and $\bar{\chi}(u)$ suggest that the data from the middle time periods is also consistent with the required limiting values ($\bar{\chi} = 1$ and $\chi > 0$).

Equivalently we can estimate the *coefficient of tail dependence* of Ledford and Tawn (1996) for each pair. The coefficient of tail dependence for asymptotically dependence variables is equal to 1, with values less than one indicating asymptotic independence. Let $(X_{85,\mathrm{FTSE},i}, X_{85,\mathrm{DAX},i}); i = 1, \ldots, n_{85}$ denote the pairs of

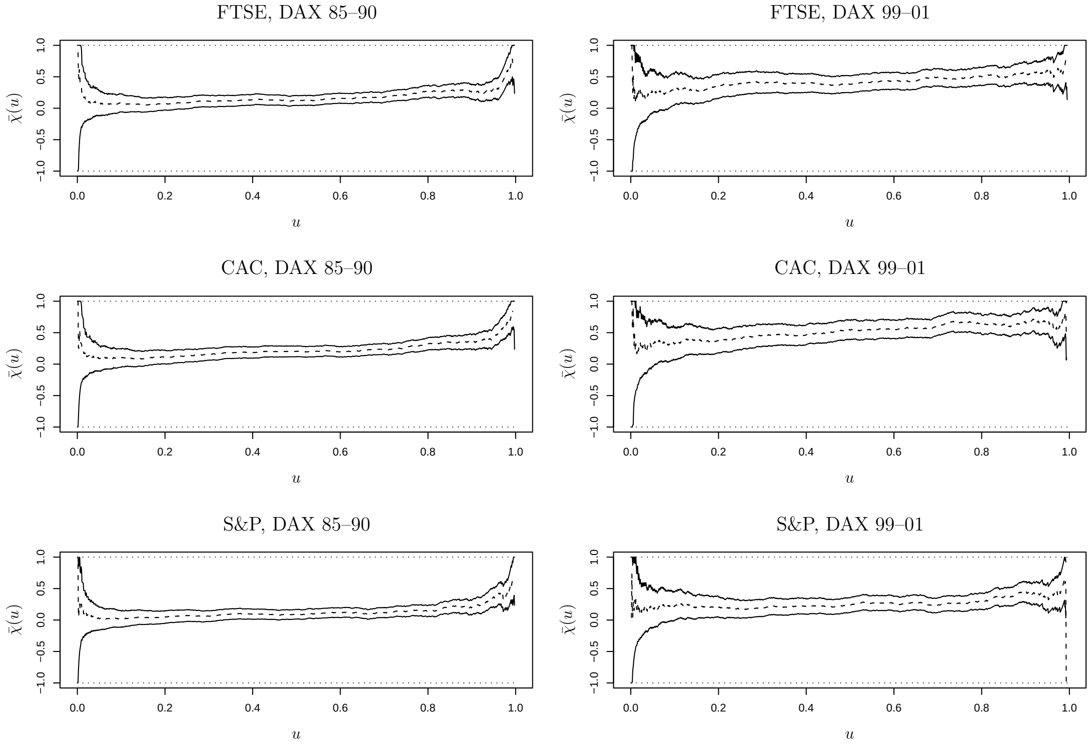


Figure 5.5: Estimates of $\bar{\chi}(u)$ (dashed lines) for financial indices over time periods 1st January 1985 - 30th June 1990 (left-hand side) and 1st January 1999 - 12th November 2001 (right-hand side), with pointwise 95% confidence intervals.

Fréchet transformed negative FTSE and DAX returns during the period 1st January 1985 - 31st June 1990, where n_{85} denotes the number of such returns recorded in this period. The remaining pairs of indices for all four periods are defined analogously. Then let $T_{85,\text{FTSE},\text{DAX},i} = \min(X_{85,\text{FTSE},i}, X_{85,\text{DAX},i}); i = 1, \ldots, n_{85}$. Then the coefficient of tail dependence $\eta_{85,\text{FTSE},\text{DAX}}$ is the shape parameter of the univariate variables $T_{85,\text{FTSE},\text{DAX},i}; i = 1, \ldots, n_{85}$. Standard univariate extreme value techniques lead to inferences on $\eta_{85,\text{FTSE},\text{DAX}}$ and on the coefficients of tail dependence for the other pairs and other periods. We follow Davison and Smith (1990) in adopting a threshold based likelihood approach.

Table 5.1 shows maximum likelihood estimates for the coefficients of tail dependence for each pair. Threshold selection was carried out using standard diagnostics including mean residual life plots and parameter threshold stability plots (Coles, 2001). Table 5.1 shows that all of these pairs exhibit tail behaviour that is consistent with a coefficient of tail dependence equal to 1. This indicates the appropriateness of the proposed methods to describe extremal dependence within the asymptotic dependence class for this data set.

	DAX, FTSE	DAX, FR	DAX, S&P
1st January 1985 – 30th June 1990	0.89 (0.12)	0.99 (0.12)	0.92 (0.12)
1st July 1990 – 31st December 1993	1.01 (0.16)	1.08 (0.16)	0.98 (0.15)
1st January 1994 – 31st December 1998	1.05 (0.13)	0.92 (0.12)	0.87 (0.13)
1st January 1999 – 12th November 2001	0.87 (0.16)	0.81 (0.16)	0.91 (0.17)

Table 5.1: Maximum likelihood estimates of coefficients of tail dependence for pairs of indices in different time periods from January 1985 – November 2001. Standard errors in parentheses.

We estimate the spectral measure and associated dependence function A for each pair of indices in each of the four time periods considered. We use the proposed estimation method of Section 5.4. Estimation uncertainty is assessed using a nonparametric bootstrap in which we sample with replacement from the pairs of variables within each time period to obtain replicate data sets of the same size as the original data sets. The number of bootstrap replicate data sets generated for this analysis was 1000. We estimated the dependence function A for each bootstrap data set. Each estimate was treated as a realisation from the sampling distribution of the estimator for the dependence function.

The estimation threshold is selected by assessing stability of estimates to threshold choice, as described in Section 5.4. For simplicity, the threshold exceedance probability was constrained to be the same for each marginal variable, the selected threshold value being equal to the 0.9 marginal quantile.

Figure 5.6 shows estimated dependence functions for each of the pairs and each of the time periods considered. We look first at dependence between the German DAX and the French CAC. Both of these countries joined the EMU on the 1st January 1999. For this pair the dependence is weakest in the earliest period, similar during the two periods preparing for EMU, and strongest after the exchange rate freeze on the 1st January 1999. Indeed, this final dependence is the strongest between any pair of indices during any period.

Dependence between the UK FTSE and the German DAX in the first period is similar to that between the DAX and the CAC in this period. However despite the dependence between the UK and German indices strengthening over time, the ultimate dependence between these indices in the final period considered is weaker than that of the DAX and CAC.

Dependence between the US Standard and Poors and the German DAX is weakest of all, with little change in dependence occurring during the first three periods and slightly stronger dependence during the last period. We investigated whether this increased dependence in the last period could be explained by the downturn in shares occurring on or around September 11th 2001, which affected markets internationally. There is little evidence from the data to support this hypothesis as although large negative marginal returns are observed on and following this date, these large values do not occur simultaneously.

Poon *et al.* (2003) point out the greater influence of the US stock market on other international markets. They argue that since the US stock markets close later than the European markets, the effect of US activity is liable to be seen in the following day's activity of the European markets. We therefore repeated the analysis, this time comparing US returns with German returns recorded the following day, rather than on the same day as above. This change in approach actually decreased the observed lower tail dependence between the S&P and DAX. Values of these returns from the first three periods were found to be consistent with asymptotic dependence. This was not the case for the final period, for which the estimated coefficient of tail dependence was 0.49 (s.e. 0.13), corresponding to near independence.

Comparing the estimated dependence functions and their pointwise confidence intervals for the different pairs in each time period, we can see that the differences between strength of dependence in the first and last period is significant for all three pairs. The estimated S&P and DAX dependence in the final period is significantly weaker than the CAC and DAX dependence, although the FTSE and DAX dependence is not.

Our results suggest that the harmonisation of European currencies joining the EMU may have had some converging effect on these nations' stock exchanges. All of the indices considered become more strongly dependent on the German DAX. Stronger dependence is seen between European pairs and the strongest dependence of all is observed between the two economies within the EMU. Further comparisons of stock exchange indices could be undertaken to investigate whether this phenomenon occurs more widely than for the limited data set considered here.

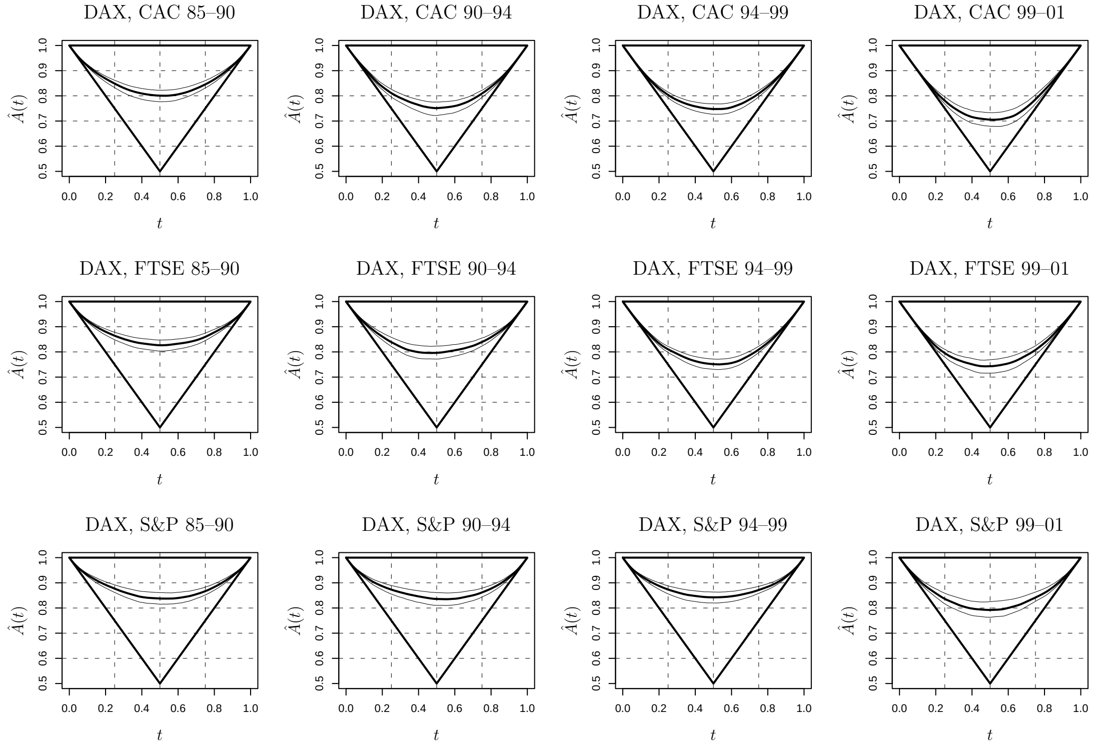


Figure 5.6: Estimated dependence functions (thick solid lines) for financial indices in time period during 1985 - 2001, with pointwise 95% bootstrap based confidence intervals.

5.8 Discussion

We have exploited the model structure used by Heffernan and Tawn (2004) to motivate the new consistent conditional estimator of Section 5.4. The resulting estimator for asymptotically dependent distributions lies within the broader class of models proposed by Heffernan and Tawn. This estimator is thus seen as an important special case of the more flexible modelling strategy which accommodates asymptotic dependence and asymptotic independence as well as negative extremal dependence.

The augmentation of the Heffernan and Tawn approach with the new methods described in this paper therefore offers a unified methodology for the analysis of a broad range of dependence structures. We have demonstrated that the performance of our conditional estimator for the dependence function of a bivariate extreme value distribution is similar to the existing estimator of Capéraà and Fougères (2000) (in the case of weak dependence) and that the conditional estimator out-performs the same estimator in the case of strong dependence. The conditional estimator out-performs the Abdous and Ghoudi (2004) estimator under all types of dependence. These conclusions hold for a variety of underlying distributional forms within the asymptotic dependence class. Further, our estimator is the only one of the estimators which extends outside this family to classes of asymptotic independence.

We have concentrated on obtaining conditional estimators of the spectral measure H and the Pickands' dependence function A for bivariate random variables. We now extend these to the multivariate case. Let $\mathbf{X} = (X_1, \ldots, X_p)$ be the p-dimensional random variable with distribution function F and unit Fréchet margins. To derive an estimate for the spectral measure H, we consider the natural multivariate extension to the Poisson process described in Section 5.2. In this case the pseudo-radial R and angular W_j coordinates are

$$R = ||\mathbf{X}|| \ \text{ and } \ \mathbf{W} = (X_j/R : j = 1, \ldots, p-1)$$

and $W_p = 1 - \sum_{j=1}^{p-1} W_j$. We continue to use the L_1 norm to define R. Then, as $n \to \infty$, the point process $P_n = \{\mathbf{X}_i/n : i = 1, \ldots, n\}$ tends to the Poisson process with intensity measure

$$\mu(\,\mathrm{d}r \times \,\mathrm{d}\mathbf{w}) = \frac{\mathrm{d}r}{r^2} p \,\mathrm{d}H(\mathbf{w})$$

where the spectral measure H is a distribution function on the unit simplex $S_p = \{\mathbf{w} : \sum_{j=1}^p w_j = 1 \; ; \; w_j \geq 0, \; j = 1, \ldots, p\}$. Further the measure satisfies the marginal moment conditions

$$\int_{S_p} w_j \,\mathrm{d}H(\mathbf{w}) = \frac{1}{p}, \qquad j = 1, \ldots, p. \tag{5.8.1}$$

The multivariate extreme value distribution and the multivariate Pickands' dependence function are then defined, respectively, as

$$G(\mathbf{x}) = \exp\left\{ -\int_{S_p} \max_{1\leq j\leq p} (w_j/x_j)p \,\mathrm{d}H(\mathbf{w}) \right\} \tag{5.8.2}$$

and

$$A(\mathbf{t}) = \int_{S_p} \max_{j=1,\ldots,p} \{t_j w_j\} p \,\mathrm{d}H(\mathbf{w}). \tag{5.8.3}$$

Now take $\mathbf{Y} = \log(\mathbf{X})$ to be the transformed random variable with Gumbel margins and let $\mathbf{Y}_{-j}$ denote the random variable $\mathbf{Y}$ with the j^{th} component removed. The Heffernan and Tawn (2004) model discussed in Section 5.3 extends to the multivariate setting to give, conditional on one component of $\mathbf{Y}$ exceeding some high threshold, the distribution of the remaining components as follows:

$$\lim_{y_j\to\infty} \Pr\{\mathbf{Y}_{-j} \leq \mathbf{a}_{\,|\,j}(y_j) + \mathbf{b}_{\,|\,j}(y_j)\mathbf{z}_{\,|\,j} \,|\, Y_j = y_j\} = D_{\mathbf{Z}_{\,|\,j}}(\mathbf{z}_{\,|\,j}), \qquad j = 1, \ldots, p,$$

where $\mathbf{a}_{\,|\,j}(y_j)$ and $\mathbf{b}_{\,|\,j}(y_j)$ are vectors of normalising functions. If $\mathbf{Y}_{-j}$ is asymptotically dependent on Y_j these functions are simply $\mathbf{a}_{\,|\,j}(y_j) = y_j\mathbf{1}$ and $\mathbf{b}_{\,|\,j}(y_j) = \mathbf{1}$. As in the bivariate case we assume such asymptotic dependence in deriving our estimators.

The conditional estimate of H can be found by considering the sets $B_j = \{\mathbf{X} : X_j > x_j\}$ and $B_j^{(\mathbf{t})} = \{\mathbf{X} : X_j > x_j, \mathbf{W} < \mathbf{t}\}$, for $j = 1, \ldots, p$. Following the methods of Section 5.4 we have

$$H(\mathbf{t}) = \frac{1}{p}\sum_{j=1}^{p}\frac{\Lambda(B_j^{(\mathbf{t})})}{\Lambda(B_j)}. \tag{5.8.4}$$

This conditional formula for H leads to the natural empirical estimate

$$\hat{H}_1(\mathbf{t}) = \frac{1}{p}\sum_{j=1}^{p}\left\{\frac{1}{\sum_{i=1}^{n} I_{[X_{ji}>u_j]}}\sum_{i=1}^{n} I_{[X_{ji}>u_j \& \mathbf{W}_i \le \mathbf{t}]}\right\} \tag{5.8.5}$$

where u_j is the threshold used for the j^{th} component.

As in the bivariate case, this first estimate does not satisfy the moment conditions of equation (5.8.1). We introduce a linear tilting, similar to that of equation (5.4.7). This takes the form

$$\mathrm{d}\hat{H}(\mathbf{t}) = \tilde{\boldsymbol{\beta}}(1, t_1, \ldots, t_{p-1})\,\mathrm{d}\hat{H}_1(\mathbf{t}) \tag{5.8.6}$$

where $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_p)^T$. In order to satisfy the moment conditions and ensure a mass of 1 on S_p the constants $\tilde{\boldsymbol{\beta}}$ are obtained by solving the system of equations

$$\tilde{\boldsymbol{\beta}}^T\hat{V} = (p, 1, \ldots, 1)^T$$

where $\hat{V}$ is a $p \times p$ matrix defined by

$$\hat{V}_{i,j} = \begin{cases} \int_{S_p}\sum_k \mathrm{d}\hat{C}_k(t) & \text{if } i=1,\ j=1, \\ \int_{S_p} t_{i-1}\sum_k \mathrm{d}\hat{C}_k(t) & \text{if } i\ge 2,\ j=1, \\ \int_{S_p} t_{j-1}\sum_k \mathrm{d}\hat{C}_k(\mathbf{t}) & \text{if } i=1,\ j\ge 2, \\ \int_{S_p} t_{i-1}t_{j-1}\sum_k \mathrm{d}\hat{C}_k(\mathbf{t}) & \text{if } i\ge 2,\ j\ge 2, \end{cases}$$

in which the $\hat{C}_k(\cdot)$ are estimates of the multivariate extensions to equations (5.2.9) and (5.2.10), *i.e.* for $k = 1, \ldots, p$, they estimate

$$C_k(\mathbf{t}) = \lim_{n\to\infty}\Pr\{\mathbf{X}/n \in B_k^{(\mathbf{t})} \,|\, \mathbf{X}/n \in B_k\}.$$

Hence the modified estimator of equation (5.8.5) is given by

$$\hat{H}(\mathbf{t}) = \frac{1}{p}\sum_{j=1}^{p}\left\{\frac{1}{\sum_{i=1}^{n} I_{[X_{ji}>u_j]}}\sum_{i=1}^{n}\tilde{\boldsymbol{\beta}}\mathbf{W}_i^* I_{[X_{ji}>u_j \& \mathbf{W}_i\leq\mathbf{t}]}\right\} \tag{5.8.7}$$

where $\mathbf{W}_i^* = (1, \mathbf{W}_i)$.

The analogous empirical estimator for the multivariate Pickands' dependence function is found by extension of the method used to obtain the bivariate estimator in equation (5.4.11). The estimator $\hat{H}$ defined in (5.8.7) assigns a point mass of $\sum_{k=1}^{p} m_{ki}$ to each variable $\mathbf{X}_i$, where

$$m_{ji} = \frac{1}{n_{u_j}}\tilde{\boldsymbol{\beta}}\mathbf{W}_i^* I_{[X_{ji}>u_j]}.$$

Using definition (5.8.3) the empirical estimator of the multivariate Pickands' dependence function is then

$$\hat{A}(\mathbf{t}) = \sum_{i=1}^{n}\left\{\max_{j=1,\ldots,p}\{t_j W_{ji}\}\sum_{k=1}^{p} m_{ki}\right\}. \tag{5.8.8}$$

This estimator satisfies all the conditions for a Pickands' dependence function. Further, by the method of their derivation, these estimators for A and H are self consistent with themselves and with G, the multivariate extreme value distribution function (5.8.2). Note also that the bivariate estimators for both A and H given in Section 5.4 arise for any bivariate marginal of the multivariate estimators presented here.

References

Abdous, B., Ghoudi, K. and Khoudraji, A. (1999) Non-parametric estimation of the limit dependence function of multivariate extremes. *Extremes*, **2:3**, 245-268.

Abdous, B. and Ghoudi, K. (2005) Nonparametric estimators of multivariate extreme dependence functions, *J. Nonparametric Statist.*, **17:8**, 915-935.

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.

Capéraà, P. and Fougères, A.-L. (2000) Estimation of a bivariate extreme value distribution. *Extremes*, **3**, 311-329.

Coles, S. G. (2001) *An Introduction to Statistical Modeling of Extreme Values.* Springer.

Coles, S. G., Heffernan, J. E. and Tawn, J. A. (1999) Dependence Measures for Extreme Value Analyses. *Extremes*, **2**, 339-365.

Coles, S. G. and Tawn, J. A. (1991) Modelling extreme multivariate events. *J. Roy. Statist. Soc., B*, **53**, 377–392.

Davison, A. C. and Smith, R. L. (1990) Models for exceedances over high thresholds (with discussion). *J. R. Statist. Soc. B*, **52**, 393-442.

Devroye, L. (1989) The double kernel method in kernel density estimation. *Annales de l'Institut Henri Poincar'e*, **25**, 533-580.

Draisma, G., Drees, H., Ferreira, A. and de Haan, L. (2004). Bivariate tail estimation: dependence in asymptotic independence. *Bernoulli* **10**:2, 251-280.

Einmahl, J., de Haan, L. and Huang, X. (1993) Estimating a multidimensional extreme value distribution. *J. Multivariate Anal.* **47**, 35-47.

Einmahl, J., de Haan L. and Sinha, A.K. (1997) Estimating the spectral measure of an extreme value distribution. *Stoch. Proc. Appl.* **70**, 143-171.

Einmahl, J., de Haan, L. and Piterbarg, V. (2001) Nonparametric estimation of the spectral measure of an extreme value distribution. *Annals of Statistics*, **29**, 1401-1423.

Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997) *Modelling Extreme events for Insurance and Finance*, Springer, Berlin.

Embrechts, P. (Editor) (2000) *Extremes and Integrated Risk Management*, UBS Warburg and Risk Books, London.

Genest, C., Ghoudi, K. and Rivest, L.-P. (1995) A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*,

82, (3), pp. 543-52.

Gumbel, E. J. Bivariate exponential distributions. (1960) *J. Amer. Statist. Assoc.*, **55**, 698–707.

de Haan, L. and Resnick, S. I. (1977) Limit theory for multivariate sample extremes. *Z. Wahrsch. Theor.*, **40**, 317–337.

de Haan, L. and de Ronde, J. (1998) Sea and wind: multivariate extremes at work. *Extremes*, **1**, 7–45.

Heffernan, J. E. and Tawn, J. A. (2004) A conditional approach for multivariate extreme values (with discussion). *J. Roy. Statist. Soc. B*, **66** (3), 497-546.

Heffernan, J.E. and Resnick, S.I. (2007) Limit laws for random vectors with an extreme component. *Ann. Appl. Prob.*, **17** (2), 537-571.

Huang, X. (1992) Statistics of bivariate extreme values. Thesis, Erasmus University Rotterdam, Tinbergen Institute Research Series **22**.

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., Smith, R. L. and Weissman, I. (1992) Bivariate threshold methods for extremes. *J. Roy. Statist. Soc., B*, **54**, 171–183.

Joe, H. (1994) Multivariate extreme–value distributions with applications to environmental data. *Canadian. J. Statist.* **22**, 47–64.

Ledford, A. W. and Tawn, J. A. (1996) Statistics for near independence in multivariate extreme values. *Biometrika*, **83**, 169–187.

Ledford, A. W. and Tawn, J. A. (1997) Modelling dependence within joint tail regions. *J. Roy. Statist. Soc., B*, **59**, 475–499.

Longin, F. (2000) From value at risk to stress testing: the extreme value approach. *J. Bnkng. Finan.*, **24**, 1097-1130, 1130.

Maulik, K., Resnick S. I. and Rootzén H. (2002) Asymptotic independence and a network traffic model. *J. Appl. Probab.* **39** (4), 671-699.

Peng, L. (1999). Estimation of the coefficient of tail dependence in bivariate extremes. *Statist. Probab. Lett.*, **43**, 399-409.

Pickands, J. (1981) Multivariate extreme value distributions. *Proc. 43rd Sess. Int. Statist. Inst.*, 859–878.

Poon, S.-H., Rockinger, M. and Tawn, J. A. (2003a) Extreme-value dependence in financial markets: diagnostics, models and financial implications. To appear in *Review of Financial Studies.*

Poon, S.-H., Rockinger, M. and Tawn, J. A. (2003b) Modelling extreme-value dependence in international stock markets. *Statistica Sinica*, **13** (4), 929-953.

Resnick, S. I. (1987) *Extreme Values, Regular Variation, and Point Processes.* New York: Springer–Verlag.

Resnick, S. I. and Rootzén, H. (2000) Self-similar communication models and very heavy tails. *Ann. Appl. Probab.*, **10**, 753-778.

Schlather, M. and Tawn, J. A. (2003) A dependence measure for multivariate and spatial extreme values: properties and inference. *Biometrika*, **90**, 139–156.

Shi, D., Smith, R. L. and Coles, S. G. (1992) Joint versus marginal estimation for bivariate extremes. Tech. Rep. **2074**, Department of Statistics, University of North Carolina at Chapel Hill.

Smith, R. L., Tawn, J. A. and Yuen, H. K. (1990) Statistics of multivariate extremes. *Int. Statist. Inst. Rev.*, **58**, 47–58.

Stărică, C. (1999) Multivariate extremes for models with constant conditional correlations. *Journal of Empirical Finance*, **6**, 515-553.

Stephenson, A. and Tawn, J. A. (2004) Exploiting occurrence times in likelihood inference for componentwise maxima. *Biometrika*, **92:1**, 213–227.

Tawn, J.A. (1988) Bivariate extreme value theory: models and estimation. *Biometrika* **75**, 387–415.

Appendix B

Recall that in stating both Theorems 5.4.1 and 5.4.2 we assumed that $P_n \equiv P$ on the region $\mathbb{R}_+^2 \backslash \{[0, u_1] \times [0, u_2]\}$ where P is a Poisson process with intensity given by equation (5.2.2). In the following proofs we therefore assume that we have a sequence $\mathbf{X}_1, \ldots, \mathbf{X}_n$ from this limiting process P. We further assume unit Fréchet margins.

B.1 Proof of Theorem 5.4.1

The key here is to write the estimator $\hat{H}_1(t)$ as in equation (5.4.4) as the mean, across the components j, of the proportion of variables for which the jth component exceeds the marginal threshold and the angular coordinate is less than t out of the total number of variables for which the jth component exceeds the marginal threshold. For a data set of size n, we can re-write equation (5.4.4) as follows; let $v = u_1/(u_1 + u_2)$ and consider the sets

$$\begin{aligned}
S_1 &= \{\mathbf{X} : X_1 < u_1, X_2 > u_2, W < \min(t, v)\}, \\
S_2 &= \{\mathbf{X} : X_1 < u_1, X_2 > u_2, \min(t, v) < W < v\}, \\
S_3 &= \{\mathbf{X} : X_1 > u_1, X_2 > u_2, W < t\}, \\
S_4 &= \{\mathbf{X} : X_1 > u_1, X_2 > u_2, W > t\}, \\
S_5 &= \{\mathbf{X} : X_1 > u_1, X_2 < u_2, W > \max(t, v)\}, \\
S_6 &= \{\mathbf{X} : X_1 > u_1, X_2 < u_2, v < W < \max(t, v)\}.
\end{aligned}$$

These sets partition the region of interest $\{\mathbf{X} : X_1 > u_1 \text{ or } X_2 > u_2\}$. For $i = 1, \ldots, 6$ let the number of points $\mathbf{X}_1, \ldots, \mathbf{X}_n$ in the set S_i be N_i. We can then write the estimator $\hat{H}_1(t)$ as the sum of the ratios

$$\hat{H}_1(t) = \frac{1}{2}\left(\frac{N_1 + N_3}{N_1 + N_2 + N_3 + N_4} + \frac{N_3 + N_6}{N_3 + N_4 + N_5 + N_6}\right). \tag{B.1.1}$$

Clearly $\mathbf{N} = (N_i : i = 1, \ldots, 6)$ follows a multinomial$(n, \mathbf{p})$ distribution, where $\mathbf{p} = (p_i : i = 1, \ldots, 6)$ is the vector of probabilities of falling in each set. The mean and covariance structure of $\mathbf{N}$ is therefore given by $\mu_i = \mathbb{E}(\mathrm{N_i}) = np_i$, $\mathrm{var}(N_i) = np_i(1 - p_i)$ and $\mathrm{cov}(N_i, N_j) = -np_ip_j$, where the covariance is defined for $i \neq j$.

We can now derive the asymptotic distribution of the quantity of interest $\sqrt{n}\left[\hat{H}_1(t) - H(t)\right]$ by writing,

$$\sqrt{n}\left[\hat{H}_1(t) - H(t)\right] = \sqrt{n}\left[\hat{H}_1(t) - \frac{p_1 + p_3}{p_1 + p_2 + p_3 + p_4} + \frac{p_3 + p_6}{p_3 + p_4 + p_5 + p_6}\right]$$

Using the expression for $\hat{H}_1(t)$ given in equation (B.1.1) and a first order Taylor-series expansion of the terms in the denominators we can show that this is approximately equal to

$$\begin{aligned} \frac{\sqrt{n}}{2}\Bigg[& \frac{(\mu_2 + \mu_4)(N_1^* + N_3^*) - (\mu_1 + \mu_3)(N_2^* + N_4^*)}{(\mu_1 + \mu_2 + \mu_3 + \mu_4)^2} \\ & + \frac{(\mu_4 + \mu_5)(N_3^* + N_6^*) - (\mu_3 + \mu_6)(N_4^* + N_5^*)}{(\mu_3 + \mu_4 + \mu_5 + \mu_6)^2}\Bigg] \end{aligned} \tag{B.1.2}$$

where $N_i^* = N_i - \mu_i$. By the univariate central limit theorem, as $n \to \infty$ these two ratios each follow a normal distribution with mean zero. Further, by the bivariate central limit theorem, their sum is also normal. Using the variance-covariance properties of the multinomial distribution, the variance of expression (B.1.2) and also therefore the variance term in equation (5.4.6) is given by

$$\sigma_t^2(\mathbf{u}) = \frac{1}{4}\left[\frac{(p_1+p_3)(p_2+p_4)}{(p_1+p_2+p_3+p_4)^3} + \frac{(p_3+p_6)(p_4+p_5)}{(p_3+p_4+p_5+p_6)^3}\right.$$
$$\left. + 2\frac{p_3(p_2+p_4)(p_4+p_5)+p_4(p_1+p_3)(p_3+p_6)}{(p_1+p_2+p_3+p_4)^2(p_3+p_4+p_5+p_6)^2}\right].$$

Finally, in order to evaluate this variance for a particular form of spectral distribution H we can write the probabilities $\mathbf{p}$ in terms of H; writing $v_1 = \min(t, v)$ and $v_2 = \max(t, v)$ and using the intensity given in equation (5.2.2), these are

$$\begin{aligned}
p_1 &= \int_0^{v_1} \frac{1-s}{u_2} 2\,\mathrm{d}H(s) - \int_0^{v_1} \frac{s}{u_1} 2\,\mathrm{d}H(s),\\
p_2 &= \int_{v_1}^{v} \frac{1-s}{u_2} 2\,\mathrm{d}H(s) - \int_{v_1}^{v} \frac{s}{u_1} 2\,\mathrm{d}H(s),\\
p_3 &= \int_0^{v_1} \frac{s}{u_1} 2\,\mathrm{d}H(s) + \int_v^{v_2} \frac{1-s}{u_2} 2\,\mathrm{d}H(s),\\
p_4 &= \int_{v_1}^{v} \frac{s}{u_1} 2\,\mathrm{d}H(s) + \int_{v_2}^{1} \frac{1-s}{u_2} 2\,\mathrm{d}H(s),\\
p_5 &= \int_{v_2}^{1} \frac{s}{u_1} 2\,\mathrm{d}H(s) - \int_{v_2}^{1} \frac{1-s}{u_2} 2\,\mathrm{d}H(s),\\
p_6 &= \int_v^{v_2} \frac{s}{u_1} 2\,\mathrm{d}H(s) - \int_v^{v_2} \frac{1-s}{u_2} 2\,\mathrm{d}H(s).
\end{aligned}$$

We note that, using the relationship between the Pickands dependence function A and the spectral measure H defined in equation (5.2.16), that all of these probabilities can be found in closed form using the result that

$$\int_a^b s2\,\mathrm{d}H(s) = 2bH(b) - 2aH(a) + a + A(1-a) - b - A(1-b).$$

B.2 Proof of Theorem 5.4.2

The proof of Theorem 5.4.2 follows simply from the following lemma and theorem.

Lemma B.2.1 *For a sequence of bivariate random variables $\mathbf{X}_1, \ldots, \mathbf{X}_n$ with unit Fréchet margins, let N_{u_1} be the number of variables in the set $\{\mathbf{X}_i : X_{1i} > u_1\}$ and N_{u_2} be the number of variables in the set $\{\mathbf{X}_i : X_{2i} > u_2\}$. Then, as $n \to \infty$, the mean and covariance structure of $1/N_{u_1}$ and $1/N_{u_2}$ are given by*

$$\mathbb{E}\left[\frac{1}{N_{u_j}}\right] = \frac{u_j}{n}, \quad \mathrm{Var}\left(\frac{1}{N_{u_j}}\right) = \frac{u_j^2(u_j - 1)}{n^3} \quad \text{for } j = 1, 2 \tag{B.2.1}$$

and

$$\mathrm{Cov}\left(\frac{1}{N_{u_1}}, \frac{1}{N_{u_2}}\right) = \frac{u_1 u_2 (u_1 u_2 p_{11} - 1)}{n^3}. \tag{B.2.2}$$

Proof

Consider the four quadrants

$$\begin{aligned} R_{00} &= \{\mathbf{X} : X_1 \le u_1, X_2 \le u_2\}, \\ R_{10} &= \{\mathbf{X} : X_1 > u_1, X_2 \le u_2\}, \\ R_{01} &= \{\mathbf{X} : X_1 \le u_1, X_2 > u_2\}, \\ R_{11} &= \{\mathbf{X} : X_1 > u_1, X_2 > u_2\}. \end{aligned} \tag{B.2.3}$$

Let N_{00}, N_{10}, N_{01} and N_{11} denote the number of points in each of the four regions respectively, and let p_{00}, p_{10}, p_{01} and p_{11} be the probabilities of the variable $\mathbf{X}$ falling in each. Clearly $\mathbf{N} = (N_{00}, N_{10}, N_{01}, N_{11})$ follows a multinomial distribution so that, for example, $\mathbb{E}[N_{00}] = np_{00}$, $\mathrm{Var}(N_{00}) = np_{00}(1 - p_{00})$ and $\mathrm{Cov}(N_{00}, N_{10}) = -np_{00}p_{10}$. Using a normal to binomial approximation, for large n, we can approximate the random variables N_{u_1} and N_{u_2} by

$$N_{u_1} \simeq np_{10} + np_{11} + Z_{10}\sqrt{np_{10}\,(1 - p_{10})} + Z_{11}\sqrt{np_{11}\,(1 - p_{11})}$$

and

$$N_{u_2} \simeq np_{01} + np_{11} + Z_{01}\sqrt{np_{01}\,(1-p_{01})} + Z_{11}\sqrt{np_{11}\,(1-p_{11})},$$

where the random variable $\mathbf{Z} = (Z_{10}, Z_{01}, Z_{11})$ is trivariate normal with standard margins and pairwise covariance given by

$$\mathrm{Cov}(Z_{10}, Z_{11}) = -\left(\frac{p_{10}p_{11}}{(1-p_{10})(1-p_{11})}\right)^{1/2}$$

and similarly for (Z_{10}, Z_{01}) and (Z_{01}, Z_{11}). As $n \to \infty$, we use a binomial series expansion to find first order approximations for the random variables $1/N_{u_j}$, $j = 1, 2$, as a linear function of the random variables Z_{10}, Z_{01} and Z_{11}. From these expressions it is then straightforward to see that the asymptotic expressions for the mean and covariance structure of the random variable $(1/N_{u_1}, 1/N_{u_2})$ are those given in equations (B.2.1) and (B.2.2).

Theorem B.2.1 *For a sequence of bivariate random variables* $\mathbf{X}_1, \ldots, \mathbf{X}_n$ *from the point process* P *with intensity (5.2.2) with unit Fréchet margins, let* $Y_i = m_i^*(\mathbf{X}) \max\{tW_i, (1-t)(1-W_i)\}$ *where* $m_i^*(\mathbf{X}) = m_{1i}^* + m_{2i}^*$ *with* $m_{ji}^* = n_{u_j}^{-1} I_{[X_{ji} > u_j]}$, $i = 1, \ldots, n$ and $j = 1, 2$. *Let* $\mathbf{N} = (N_{00}, N_{10}, N_{01}, N_{11})$ *represent the number of variables in the regions* R_{00}*,* R_{10}*,* R_{01} *and* R_{11} *defined in equation (B.2.3) and let* I_i *and* I_j *be indicator functions denoting which of the quadrants the* i*th and* j*th variables lie in. Then, for fixed* t*, as* $n \rightarrow, \infty$*,*

$$n\mathbb{E}[Y_i] = A(t)$$

$$n^2 \mathrm{Var}(Y_i) \quad \rightarrow \quad u_1(\beta_1^2 + \beta_3) + u_2(\beta_2^2 + \beta_4) + 2u_1u_2(p_{11}\beta_1\beta_2 + \beta_5) - 2A(t)^2 \tag{B.2.4}$$

and

$$\begin{aligned} n^3 \mathrm{Cov}(Y_i, Y_j) \quad \rightarrow \quad & \frac{\beta_1^2}{u_1^2}\left[V_1 - 2u_1^2(1+u_1)\right] + \frac{\beta_2^2}{u_2}\left[V_2 - 2u_2^2(u_2+1)\right] \\ & +2\left\{A_1A_2\left[C - 2u_1u_2\right] + A_1A_3\left[C - u_1u_2(u_1+2)\right]\right. \\ & \left. + A_2A_3\left[C - u_1u_2(u_2+2)\right] + A_3^2\left[C - u_1u_2(u_1+u_2+2)\right]\right\} \end{aligned} \tag{B.2.5}$$

where β_1*,* β_2*,* β_3*,* β_4*,* β_5*,* A_1*,* A_2 *and* A_3 *are defined in the proof and, taking* $v = u_1/(u_1+u_2)$*,*

$$\begin{aligned} V_j &= u_j^3\left(1 - \frac{1}{u_j}\right), \quad j = 1, 2, \\ C &= c(u_1u_2)^{3/2}\left(1 - \frac{1}{u_1}\right)^{1/2}\left(1 - \frac{1}{u_2}\right)^{1/2}, \\ c &= \left(p_{11} - \frac{1}{u_1u_2}\right)\left(\frac{1}{u_1u_2}\left(1 - \frac{1}{u_1}\right)\left(1 - \frac{1}{u_2}\right)\right)^{-1/2} \quad \text{and} \\ p_{11} &= \Pr(\mathbf{X}_i \in R_{11}) = \int_0^v \frac{w}{u_1} 2\, dH(w) + \int_v^1 \frac{1-w}{u_2} 2\, dH(w). \end{aligned}$$

Proof The expectation of Y_i is found by first conditioning on the random variables N_{u_1} and N_{u_2}, and then using the total law of expectation,

$$\mathbb{E}[Y_i] = \mathbb{E}\{\mathbb{E}[Y_i|N_{u_1} = n_{u_1}, N_{u_2} = n_{u_2}]\}.$$

The conditional expectation $\mathbb{E}[Y_i|N_{u_1} = n_{u_1}, N_{u_2} = n_{u_2}]$ is obtained by integrating across the Poisson process intensity of equation (5.2.2), to give

$$\mathbb{E}[Y_i|N_{u_1} = n_{u_1}, N_{u_2} = n_{u_2}] = \frac{\beta_1}{u_1 n_{u_1}} + \frac{\beta_2}{u_2 n_{u_2}}.$$

where we redefine $v_1 = \min(1-t, v)$ and $v_2 = \max(1-t, v)$ to write

$$\begin{aligned}\beta_1 &= \int_0^{v_1} w(1-w)(1-t)2\,\mathrm{d}H(w) + \int_{v_1}^{v} w^2 t2\,\mathrm{d}H(w) \\ &\quad + \int_{v}^{v_2} w(1-w)(1-t)2\,\mathrm{d}H(w) + \int_{v_2}^{1} w^2 t2\,\mathrm{d}H(w)\end{aligned}$$

and

$$\begin{aligned}\beta_2 &= \int_0^{v_1} (1-w)^2(1-t)2\,\mathrm{d}H(w) + \int_{v_1}^{v} w(1-w)t2\,\mathrm{d}H(w) \\ &\quad + \int_{v}^{v_2} (1-w)^2(1-t)2\,\mathrm{d}H(w) + \int_{v_2}^{1} w(1-w)t2\,\mathrm{d}H(w).\end{aligned}$$

Note that $\beta_1+\beta_2 = A(t)$. Undoing the conditioning using the limiting expectations of $1/N_{u_j}$, $j = 1, 2$, given in equation (B.2.1) of Lemma B.2.1 we find that, as $n \to \infty$, the expectation of Y_i is simply $\mathbb{E}[Y_i] = A(t)/n$.

The variance of Y_i is also found by conditioning on N_{u_1} and N_{u_2} and then using the total law of variance,

$$\mathrm{Var}(Y_i) = \mathbb{E}[\mathrm{Var}(Y_i|N_{u_1} = n_{u_1}, N_{u_2} = n_{u_2})] + \mathrm{Var}(\mathbb{E}[Y_i|N_{u_1} = n_{u_1}, N_{u_2} = n_{u_2}]). \tag{B.2.6}$$

This requires the variance and covariance results for $(1/N_{u_1}, 1/N_{u_2})$ given in equations (B.2.1) and (B.2.2). Combining these and taking the highest order terms gives the variance as expressed in equation (B.2.4), where the constants are

$$\begin{aligned}\beta_3 &= \int_0^{v_1} w(1-w)^2(1-t)^2 2\,\mathrm{d}H(w) + \int_{v_1}^{v} w^3 t^2 2\,\mathrm{d}H(w)\\ &\quad + \int_v^{v_2} w(1-w)^2(1-t)^2 2\,\mathrm{d}H(w) + \int_{v_2}^{1} w^3 t^2 2\,\mathrm{d}H(w),\end{aligned}$$

$$\begin{aligned}\beta_4 &= \int_0^{v_1} (1-w)^3(1-t)^2 2\,\mathrm{d}H(w) + \int_{v_1}^{v} (1-w)w^2 t^2 2\,\mathrm{d}H(w)\\ &\quad + \int_v^{v_2} (1-w)^3(1-t)^2 2\,\mathrm{d}H(w) + \int_{v_2}^{1} (1-w)w^2 t^2 2\,\mathrm{d}H(w),\end{aligned}$$

and

$$\begin{aligned}\beta_5 &= \frac{1}{u_1}\int_0^{v_1} w(1-w)^2(1-t)^2 2\,\mathrm{d}H(w) + \frac{1}{u_1}\int_{v_1}^{v} w^3 t^2 2\,\mathrm{d}H(w)\\ &\quad + \frac{1}{u_2}\int_v^{v_2} (1-w)^3(1-t)^2 2\,\mathrm{d}H(w) + \frac{1}{u_2}\int_{v_2}^{1} (1-w)w^2 t^2 2\,\mathrm{d}H(w).\end{aligned}$$

Finally, to obtain the covariance term $\mathrm{Cov}(Y_i, Y_j) = \mathbb{E}[Y_iY_j] - \mathbb{E}[Y_i]\mathbb{E}[Y_j]$ we need to find the expectation of the product of Y_i and Y_j. Since these are not independent, even when we condition on N_{u_1} and N_{u_2}, to simplify matters, we also condition on which of the quadrants R_{00}, R_{10}, R_{01} and R_{11} the variables $\mathbf{X}_i$ and $\mathbf{X}_j$ lie in; this information is denoted by the indicator functions I_i and I_j. By conditioning on the actual location of the variables $\mathbf{X}_i$ and $\mathbf{X}_j$ and repeated application of the total law of expectation we can obtain the expectation of the product of Y_i and Y_j as

$$\mathbb{E}[Y_iY_j] = \mathbb{E}_{\mathbf{I}}\mathbb{E}_{\mathbf{N}^*|\mathbf{I}}\left\{\mathbb{E}[Y_j|\mathbf{N}^*, \mathbf{I}]\mathbb{E}[Y_i|Y_j, \mathbf{N}^*, \mathbf{I}]\right\}$$

where $\mathbf{N}^* = (N_{u_1}, N_{u_2})$ and $\mathbf{I} = (I_i, I_j)$.

The conditional expectations of Y_i and Y_j in this expression are straightforward. By conditioning on both N_{u_j} $(j = 1, 2)$ and the actual position of the two variables, we can find the expectation of Y_i (Y_j) in the region in which $\mathbf{X}_i$ $(\mathbf{X}_j)$ is now known to lie in by evaluating the expectation with respect to the Poisson process intensity (5.2.2) across this region and then normalising by the probability of being in this region. Averaging over the combinations of the indicator variables I_i and I_j then gives

$$\begin{aligned}
\mathbb{E}[Y_iY_j] \quad = \quad & A_1^2\mathbb{E}\left[\frac{1}{N_1^2}\middle|\mathbf{X}_i \in R_{10}, \mathbf{X}_j \in R_{10}\right] + 2A_1A_2\mathbb{E}\left[\frac{1}{N_1}\frac{1}{N_2}\middle|\mathbf{X}_i \in R_{10}, \mathbf{X}_j \in R_{01}\right] \\
& +2A_1A_3\mathbb{E}\left[\frac{1}{N_1^2} + \frac{1}{N_1}\frac{1}{N_2}\middle|\mathbf{X}_i \in R_{10}, \mathbf{X}_j \in R_{11}\right] \\
& +2A_2A_3\mathbb{E}\left[\frac{1}{N_2^2} + \frac{1}{N_1}\frac{1}{N_2}\middle|\mathbf{X}_i \in R_{01}, \mathbf{X}_j \in R_{11}\right] \\
& +A_2^2\mathbb{E}\left[\frac{1}{N_2^2}\middle|\mathbf{X}_i \in R_{01}, \mathbf{X}_j \in R_{01}\right] \\
& +A_3^2\mathbb{E}\left[\frac{1}{N_1^2} + \frac{1}{N_2^2} + 2\frac{1}{N_1}\frac{1}{N_2}\middle|\mathbf{X}_i \in R_{11}, \mathbf{X}_j \in R_{11}\right].
\end{aligned} \tag{B.2.7}$$

The constants A_1, A_2 and A_3 are, respectively, the expectations of Y_i in each of the regions R_{10}, R_{01} and R_{11}, and as such are given by

$$\begin{aligned}
A_1 \quad = \quad & \frac{1}{u_1}\int_v^{v_2} w(1-w)(1-t)2\,\mathrm{d}H(w) - \frac{1}{u_2}\int_v^{v_2}(1-w)^2(1-t)2\,\mathrm{d}H(w) \\
& +\frac{1}{u_1}\int_{v_2}^1 w^2t2\,\mathrm{d}H(w) + \frac{1}{u_2}\int_{v_2}^1 w(1-w)t2\,\mathrm{d}H(w),
\end{aligned}$$

$$\begin{aligned}
A_2 \quad = \quad & \frac{1}{u_2}\int_0^{v_1}(1-w)^2(1-t)2\,\mathrm{d}H(w) - \frac{1}{u_1}\int_0^{v_1} w(1-w)(1-t)2\,\mathrm{d}H(w) \\
& +\frac{1}{u_2}\int_{v_1}^v (1-w)wt2\,\mathrm{d}H(w) + \frac{1}{u_1}\int_{v_1}^v w^2t2\,\mathrm{d}H(w),
\end{aligned}$$

and

$$\begin{aligned} A_3 &= \frac{1}{u_1}\int_0^{v_1} w(1-w)(1-t)2\,\mathrm{d}H(w) - \frac{1}{u_1}\int_{v_1}^{v} w^2(1-t)2\,\mathrm{d}H(w) \\ &\quad +\frac{1}{u_2}\int_v^{v_2}(1-w)^2 t2\,\mathrm{d}H(w) + \frac{1}{u_2}\int_{v_2}^{1} w(1-w)t2\,\mathrm{d}H(w). \end{aligned}$$

Note that if either of $\mathbf{X}_i$ or $\mathbf{X}_j$ lies in R_{00} the contribution to the conditional expectation, and hence also to the overall unconditional expectation, is zero, since in this case $Y_i = 0$ $(Y_j = 0)$.

What remains is to find the conditional expectations of the various functions of N_{u_j} $(j = 1, 2)$ given in equation (B.2.7). These expectations must be worked out for all combinations of $\mathbf{I}$, although in what follows, we consider only the case in which $\mathbf{X}_i$ lies in the region R_{10} and $\mathbf{X}_j$ in R_{01}, since all other cases follow similarly. Conditional on the value of $\mathbf{I}$, the variable $\mathbf{N} = (N_{00}, N_{10}, N_{01}, N_{11})$ follows a multinomial$(n-2, \mathbf{p})$ distribution. So that, in our example, $\mathbb{E}[N_{10}|\mathbf{I}] = (n-2)p_{10}+1$, $\mathrm{Var}(N_{10}|\mathbf{I}) = (n-2)p_{10}(1-p_{10})$ and $\mathrm{Cov}(N_{10}, N_{01}|\mathbf{I}) = -(n-2)p_{10}p_{01}$. We can use these primary results to find, conditional on $\mathbf{I}$, the expectations, variances and covariance of N_{u_1} and N_{u_2} using the same methods as those used in the proof of Lemma B.2.1. In the example considered, we then approximate $1/N_{u_j}$ $(j = 1, 2)$ using a binomial series expansion as follows,

$$\begin{aligned} \frac{1}{N_{u_j}} &= \left\{\frac{n-2}{u_j} + 1 + \left[\frac{n-2}{u_j}\left(1-\frac{1}{u_j}\right)\right]^{1/2} Z_j\right\}^{-1} \\ &= \frac{u_j}{n-2}\left\{1 - \frac{u_j}{n-2} - \left[\left(\frac{u_j}{n-2}\right)^{1/2}\left(1-\frac{1}{u_j}\right)^{1/2} Z_j + \frac{u_j}{n-2}\left(1-\frac{1}{u_j}\right) Z_j^2\right]\right\} \\ &\quad +o(n^{-2}), \end{aligned}$$

where $Z_j \sim \mathrm{N}(0, 1)$ and $\mathrm{Cov}(Z_1, Z_2) = c$, where c is given in Theorem B.2.1.

To find the conditional expectations required in equation (B.2.7) requires the conditional expectation, variance and covariances of $1/N_{u_j}$ $(j = 1, 2)$. To find these we use the moment generating function (mgf) for bivariate normal random variables to find the necessary higher order moments of Z_j $(j = 1, 2)$. The required mgf is

$$M_{\mathbf{Z}}(\mathbf{t}) = \exp\left\{\frac{1}{2}(t_1^2 + 2ct_1t_2 + t_2^2)\right\}$$

where $c = \text{Cov}(Z_1, Z_2)$ as previously. We find that $\mathbb{E}[Z_1^2 Z_2] = \mathbb{E}[Z_1 Z_2^2] = 0$ and $\mathbb{E}[Z_1^2 Z_2^2] = 1 + 2c^2$. Further, using the moment generating function for the normal distribution $\text{Var}(Z_j^2) = 2$ and $\text{Cov}(Z_j, Z_j^2) = 0$, for $j = 1, 2$. Using these results, it is then straightforward to find the moments required in equation (B.2.7). For example,

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{N_{u_1}}\middle|\mathbf{X}_i \in R_{10}, \mathbf{X}_j \in R_{01}\right] &= \frac{u_1}{n-1}\left(1 - \frac{1}{n-2}\right), \\
\text{Var}\left(\frac{1}{N_{u_1}}\middle|\mathbf{X}_i \in R_{10}, \mathbf{X}_j \in R_{01}\right) &= V_1, \\
\text{Cov}\left(\frac{1}{N_{u_1}}, \frac{1}{N_{u_2}}\middle|\mathbf{X}_i \in R_{10}, \mathbf{X}_j \in R_{01}\right) &= C,
\end{aligned}$$

where V_1 and C are defined in Theorem B.2.1.

Proof of Theorem 5.4.2 Using Theorem B.2.1 we can now prove the main result given by Theorem 5.4.2. Using the definition of our estimator for the Pickands' dependence function as $\hat{A}_1(t) = \sum_{i=1}^n Y_i$, we have

$$\mathbb{E}[\hat{A}_1(t)] = n\mathbb{E}[Y_i]$$

and

$$\text{Var}(\hat{A}_1(t)) = n\text{Var}(Y_1) + n(n-1)\text{Cov}(\text{Y}_1, \text{Y}_2)$$

where the expectation and variance of Y_1 and the covariance of Y_1 and Y_2 are defined in Theorem B.2.1. From this we get the required results that, as $n \to \infty$,

$$\mathbb{E}[\hat{A}_1(t)] = A(t)$$

and that

$$\begin{aligned}
\mathrm{Var}(\hat{A}_1(t)) &= \frac{1}{n}\left(u_1(\beta_1^2+\beta_3)+u_2(\beta_2^2+\beta_4)+2u_1u_2(p_{11}\beta_1\beta_2+\beta_5)-2A(t)^2\right) \\
&\quad + n(n-1)\left\{\frac{\beta_1^2}{u_1^2}\left[V_1-\frac{2u_1^2(1+u_1)}{(n-2)^3}\right]+\frac{\beta_2^2}{u_2}\left[V_2-\frac{2u_2^2(u_2+1)}{(n-2)^3}\right]\right. \\
&\quad + 2A_1A_2\left[C-\frac{2u_1u_2}{(n-2)^3}\right]+2A_1A_3\left[C-\frac{u_1u_2(u_1+2)}{(n-2)^3}\right] \\
&\quad \left. + 2A_2A_3\left[C-\frac{u_1u_2(u_2+2)}{(n-2)^3}\right]+2A_3^2\left[C-\frac{u_1u_2(u_1+u_2+2)}{(n-2)^3}\right]\right\} \\
&= O(n^{-1}).
\end{aligned}$$