Threshold-based extreme value modelling

Nicolas Attalides

A Thesis Submitted for the Degree of
Doctor of Philosophy

in the
Faculty of Mathematical & Physical Sciences
Department of Statistical Science
University College London

March 2015

Declaration of Authorship

I, Nicolas Attalides confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed: ______________________

Date: ______________________

Abstract

There are numerous benefits of analysing and understanding extreme events. More specifically, quantifying the uncertainty of rare environmental extremes has been of great concern for a variety of stakeholders such as insurance companies and governments. What is more, the practical implications of extreme weather events like hurricanes and floods pose a need for engineers to design structures that can be exposed to these conditions and withstand them for many years in the future. It is not surprising therefore that statistical modelling of extremes, in its own right, has been playing an important role in the design process.

This thesis aims to contribute to the extreme value analysis literature primarily in the area concerned with threshold-based extreme value modelling. The major focus is on developing methods for selecting an appropriate threshold and on accounting for the uncertainty in this selection. For much of the thesis, Bayesian methods of inference are used and although the thesis concentrates on environmental applications, the methodology proposed can be applied in a more general context.

We introduce univariate extreme value theory and in particular the statistical methods employed to make inferences using extreme value models. In addition, we examine the intricacies of Bayesian inference and through a simulation study compare different prior distributions based on predictive inferences for future extreme values. For the standard independent and identically distributed (i.i.d.) observations we propose a Bayesian cross-validation method for selecting the threshold and use Bayesian model averaging to combine inferences from different thresholds. We extend this approach to the case where independence is considered as an unrealistic assumption and explore threshold specification in extreme value regression modelling.

Acknowledgements

First and foremost I would like to thank my supervisor Dr. Paul Northrop for his endless support, invaluable advice and guidance throughout my research. I would also like to thank Professor Richard Chandler for his extremely useful suggestions and helpful remarks after my MPhil Upgrade. Secondly I would like to thank UCL and especially all the members of the Statistical Science department which have made me feel very welcome and be part of a big ‘statistician’ family as well as to all my fellow PhD students who have made research fun!

I am also very grateful of the financial support I received from the Engineering and Physical Sciences Research Council which allowed me to attend courses and conferences and meet many interesting people. Finally, I would like to thank my family and friends for their support and encouragement.

I dedicate this thesis to my wife and partner in life, Monika. This would not be possible without your love, never-ending patience and belief in me. Felishiratou.

Contents

List of Figures

List of Tables

List of Acronyms

AR	-	Autoregressive
ARS	-	Adaptive Rejection Sampling
Bin-GP	-	Binomial-Generalised Pareto
BMA	-	Bayesian Model Averaging
CV	-	Cross-Validation
ETT	-	Extremal Types Theorem
EVT	-	Extreme Value Theory
FI	-	Fisher Information
GEV	-	Generalised Extreme Value
GP	-	Generalised Pareto
i.i.d.	-	independent and identically distributed
IMT	-	Information Matrix Test
KF	-	K-Fold
LOO	-	Leave-One-Out
MC	-	Markov Chain
MCMC	-	Markov Chain Monte Carlo
MDI	-	Maximal Data Information
MH	-	Metropolis-Hastings
MLE	-	Maximum Likelihood Estimation
n.i.d.	-	non-independent and identically distributed
NHPP	-	Non-Homogeneous Point Process
p.d.f.	-	probability density function
PC	-	Principal Components
POT	-	Peaks-Over Threshold
PWM	-	Probability Weighted Moments
QR	-	Quantile Regression
RoU	-	Ratio of Uniforms
RRSS	-	Repeated Random Sub-Sampling
UETT	-	Unified Extremal Types Theorem

1 Extreme Value Modelling

Extreme value theory uses asymptotic arguments to suggest models for extreme data. A common practical application of this theory deals with data coming from environmental sources such as rainfall totals, sea wave heights, temperatures etc, when it is of interest to investigate and model the extreme values, in this case the largest values, that these physical phenomena can take.

Thus, the main goal of extreme value modelling is to enable extrapolation, i.e. to infer the stochastic behaviour of a quantity at levels beyond those already observed. A specific example of the practical use of extreme value modelling can be found in marine engineering. The design of marine structures, such as oil platforms, requires information about the most extreme sea conditions likely to be encountered over some future long time period, for example 100, 1000 or even 10,000 years. A common variable used for this purpose is the significant wave height. One can use extreme value theory to suggest models for large significant wave heights, such as the largest value observed over a period of one year or the amounts by which a high threshold is exceeded.

In this chapter we introduce the models involved in extreme value theory. More specifically in section 1.1 we consider the simplest case: univariate independent and identically distributed (i.i.d.) sequences and introduce the two main models, namely, the Generalised Extreme Value (GEV) model (for block maxima) and the Generalised Pareto (GP) model (for threshold excesses). Later in section 1.6 we consider the case for univariate dependent sequences and introduce the K-gaps exponential mixture model (for threshold inter-exceedance times). The theory outlined in this chapter is a summary of fundamental results that are central to this research. We also provide a number of relevant references where the reader can find more details about these results and demonstrate the methodology through simple examples and graphical illustrations. Finally, we conclude this chapter with an outline of the thesis and the topics covered in the remaining chapters.

1.1 Extreme Value Theory for univariate independent sequences

Let us assume that we have a sequence of independent random variables $X_1, \ldots, X_m$ with an identical but unknown distribution function F. Often $\{X_i\}$ is a discrete-time process observed on regular time intervals, such as days. The starting point of extreme value theory is to consider the statistical behaviour of the *block maximum*

$M_n = \max\{X_1, \ldots, X_n\}$, for a block size $n \leqslant m$, as $n \to \infty$.

Under the assumed independence of the X_is the distribution function of M_n is derived as

$$\begin{aligned} P(M_n \leqslant x) &= P(X_1 \leqslant x, \ldots, X_n \leqslant x) \\ &= P(X_1 \leqslant x) \times \cdots \times P(X_n \leqslant x) \\ &= \{F(x)\}^n . \end{aligned}$$

Since the distribution function F is unknown, we investigate the behaviour of $\{F(x)\}^n$ as the block size n increases. The main concern is the fact that the asymptotic distribution of M_n degenerates to a point mass as M_n converges to the upper endpoint, $x^F = \sup\{x : F(x) < 1\}$, of F, that is, for any value of x

$$\lim_{n\to\infty} \{F(x)\}^n = \begin{cases} 1 & \text{if } F(x) = 1, \\ 0 & \text{if } F(x) < 1. \end{cases}$$

The standard approach to steer clear from this problem is to seek a linear normalization

$$M_n^* = \frac{M_n - b_n}{a_n}$$

of M_n, where $a_n > 0$ and b_n are sequences of constants, so that as $n \to \infty$ a non-degenerate limiting distribution results for M_n^*.

The question of importance is "what kinds of limiting distribution for M_n^* are possible"?

1.1.1 Extremal Types Theorem (ETT)

Fisher and Tippett (1928) were the first to describe the asymptotic properties of the normalised block maximum from an unknown distribution function F. The work by Gnedenko (1943) completed in generality this important result, known as the *Extremal Types Theorem*. A detailed proof of this theorem can be found in Leadbetter et al. (1983).

Theorem 1. *Extremal Types Theorem (ETT).*

If there exist sequences of constants $a_n > 0$ and b_n such that

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

where G is a non-degenerate distribution function, then G will belong to one of the following distribution families:

$$\textit{Gumbel:} \quad G(x) = \left\{ \exp\left\{-\exp\left[-\left(\frac{x-b}{a}\right)\right]\right\}, \quad -\infty < x < \infty; \right.$$

$$\textit{Fréchet:} \quad G(x) = \begin{cases} 0, & x \leqslant b, \\ \exp\left\{-\left(\frac{x-b}{a}\right)^{-k}\right\}, & x > b, k > 0; \end{cases}$$

$$\textit{Weibull:} \quad G(x) = \begin{cases} \exp\left\{-\left[-\left(\frac{x-b}{a}\right)^{k}\right]\right\}, & x < b, k > 0, \\ 1, & x \geqslant b, \end{cases}$$

for some location parameter b, scale parameter $a > 0$ and shape parameter k.

Fisher and Tippett (1928) showed that in fact these three distribution families are the only possible limit distributions for the block maximum irrespective of the unknown distribution F of the population. If, for a given F, the Gumbel limit is obtained, we say that F is in the *domain of attraction* of the Gumbel extreme value family, and similarly for the Fréchet and Weibull families.

Historically, when extreme value theory was used to analyse a dataset, a subjective decision was taken a priori as to which of the three families applied. This was a necessary part of the process that was followed by an estimation of the parameters of the distribution that was chosen. However, since each family describes the tail behaviour of the distribution differently there were clear drawbacks with this method:

- it introduced the argument of how the distribution choice should be made;
- it did not allow for uncertainty about the correct distributional choice.

These problems are overcome by combining the three limiting forms into a single family.

1.2 Generalised Extreme Value (GEV) distribution

The Generalised Extreme Value (GEV) distribution was derived by Jenkinson (1955) and von Mises (1964), which uses a parameterisation to integrate the three different distribution families into one. Therefore we can restate theorem 1 as theorem 2.

Theorem 2. *Unified Extremal Types Theorem (UETT).*

If there exist sequences of constants $a_n > 0$ and b_n such that

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

where G is a non-degenerate distribution function, then G is a GEV distribution function

$$G(x) = \exp\left\{-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]_{+}^{-1/\xi}\right\}, \tag{1.1}$$

for parameters $-\infty < \mu < \infty$, $\sigma > 0$ and $-\infty < \xi < \infty$, where $a_+ = \max(a, 0)$.

Thus, the GEV(μ, σ, ξ) distribution (with location parameter μ, scale parameter σ and shape parameter ξ) is defined on $\{x : 1 + \xi(x - \mu)/\sigma > 0\}$. The Gumbel distribution is obtained in the limit as $\xi \to 0$.

1.2.1 Max Stability

An informal proof of theorem 2 centres on the concept of max-stability. Firstly we note that two distributions are of the same type if they differ only in their location and/or scale parameters. A distribution G is said to be *max-stable* if there are constants $a_n > 0$ and b_n such that for every $k = 2, 3 \ldots$,

$$G^k(a_n x + b_n) = G(x),$$

that is, taking block maxima from a distribution function G results in a distribution of the same type of the original distribution G.

It makes sense that if a limiting distribution function G for linearly normalised maxima exists then G must be max-stable: if M_n^* has approximate distribution function G, then M_{nk}^*, $k \geqslant 1$, should have a distribution function of the same type, otherwise convergence has not yet been achieved. In fact, Leadbetter et al. (1983) show that the *only* distribution that is max-stable is the GEV.

1.2.2 Tail behaviour

In practice we assume tentatively that (a) the F from which the data are produced is in the domain of attraction of the GEV family and (b) the value of n is large enough that a GEV distribution is good approximation to the distribution of M_n.

This motivates the use of a GEV distribution as a model to the maxima of large numbers of i.i.d. random variables. In practice F is unknown, so the normalising constants a_n and b_n are also unknown. However, this is not a problem because a_n and b_n appear in the location and scale of the distribution of M_n and are to be estimated anyway. We discuss statistical inference for extreme value models later in section 1.5.

The important benefit of using the GEV model for extreme data analysis, as compared to the historical approach that was described earlier is the fact that it does not need a prior subjective choice of an extreme value family to which the data belong.

The following table summarises informally the tail behaviour of the distribution according to the value of ξ. The value of ξ determines which of the historical extreme value family applies and whether the upper endpoint x^F is finite ($\xi < 0$) or infinite ($\xi \geqslant 0$).

ξ	Tail behaviour	Distribution Family
$\xi = 0$	Exponential upper tail	Gumbel
$\xi > 0$	Heavy upper tail	Fréchet
$\xi < 0$	Finite upper limit	Weibull

Table 1: Tail behaviour for the distribution function F.

1.2.3 Domains of attractions

It is of theoretical interest to consider what properties F must have to be in the domain of attraction of a particular extreme value distribution. Leadbetter et al. (1983) state in section 1.6 the necessary and sufficient conditions for this. They show proves for the sufficiency and provide references for the proves of the necessity. However, for simplicity, here we follow Smith (1987) and restrict attention to absolutely continuous distribution functions F.

We begin with the hazard function $h(x)$ which can be thought of loosely as the instantaneous probability that $X = x$ given that $X \geqslant x$. We define the reciprocal hazard function $\eta(x) = 1/h(x)$ by

$$\eta(x) = \frac{1 - F(x)}{f(x)}, \quad x_F < x < x^F,$$

where x_F is the lower endpoint of the distribution, x^F is the upper endpoint of the distribution and $f(x)$ is the probability density function.

Studying the behaviour of the (reciprocal) hazard function for large x indicates how heavy is the upper tail of F. For example, if F is the distribution function of an exponential random variable, then $\eta(x)$ is constant for all x. In contrast, heavy (light) upper tails produce $\eta(x)$ that increase (decrease) as $x \to x^F$. Therefore the derivative $\eta'(x) = \mathrm{d}\eta(x)/\mathrm{d}x$ of $\eta(x)$ is key.

If $\eta'(x)$ tends to a finite limit ξ as $x \to x^F$ (the von Mises' condition) then F is in the domain of attraction of a GEV distribution with shape parameter ξ. Thus the limiting distribution of M_n is determined by the upper tail of F. Also, suitable normalising constants are given by $b_n = F^{-1}(1-1/n)$ and $a_n = \eta(b_n)$. These results are not of practical use unless we have some knowledge about the tail behaviour of F.

1.3 Generalised Pareto (GP) distribution

The results in section 1.2 relate to block maxima, i.e. the largest of n values. One can argue that since other values in each block are not utilised this method is somewhat wasteful and potentially important information might be lost. A better approach is to use an alternative definition of an extreme value, namely, that an observation is extreme if it exceeds some high threshold u.

Let us assume again that we have a sequence of independent, identically distributed random variables $X_1, \ldots, X_m$ with a distribution function F which is unknown. We introduce a threshold denoted by u. We describe in more detail the various approaches of how an appropriate value of u can be chosen in chapter 3.

Threshold modelling of extremes is based on two aspects: (i) the probability p_u that the threshold u is exceeded, and (ii) the amount by which the threshold is exceeded when it is exceeded. We use the terminology *exceedance* to refer to an X that exceeds u and define the corresponding threshold *excess* by $Z = (X - u) \mid X > u$. Theorem 3 motivates the use of a particular distribution to model the threshold excess Z.

Theorem 3. *Limiting distribution of threshold excesses.*

If theorem 2 holds then as $u \to \infty$, the distribution function of $(X-u) \mid X > u$ is approximately

$$H(z) = 1 - \left[1 + \frac{\xi z}{\sigma_u}\right]_+^{-1/\xi}, \quad z > 0, \tag{1.2}$$

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

This is the distribution function of a Generalised Pareto (GP) distribution (Pickands, 1975) defined on $0 < z < -\sigma_u/\xi$ if $\xi < 0$ and $z > 0$ if $\xi \geqslant 0$ and characterised by a scale parameter σ_u satisfying $\sigma_u > 0$ and a shape parameter ξ satisfying $-\infty < \xi < \infty$. An exponential distribution with rate parameter $1/\sigma_u$ is obtained in the limit as $\xi \to 0$. This result motivates the use of the GP(σ_u, ξ) distribution to model excesses of a high threshold u. In common with the GEV distribution, the shape parameter value determines whether or not x^F is finite as in table 1. Coles (2001, pages 76-77) gives an informal justification of theorem 3 with a more formal proof provided by Leadbetter et al. (1983).

1.3.1 Binomial-Generalised Pareto (Bin-GP) model

This theory motivates the Binomial-Generalised Pareto (Bin-GP) model with parameters (p_u, σ_u, ξ). Under the assumed independence of $X_1, \ldots, X_m$ the number of exceedances of the threshold u (denoted by n_u) has a Bin(m, p_u) distribution.

The Bin-GP model's parameters are related to the GEV parameters (μ, σ, ξ) via

$$\sigma_u = \sigma + \xi(u - \mu) \quad \text{and} \quad p_u = 1 - F(u) \approx \frac{1}{n}\left[1 + \xi\left(\frac{u-\mu}{\sigma}\right)\right]^{-1/\xi},$$

where m is the complete length of the data set, n is the block size that is used to define M_n and the approximate expression for p_u follows from Coles (2001, pages 76-77).

1.4 Gulf of Mexico data

In this section we briefly describe a motivating example of how extreme value analysis can be applied in a practical situation using the GEV and GP distributions. Firstly we introduce the dataset that is used for these analyses.

The data come from Oceanweather's metocean study for the Gulf of Mexico called GOMOS (Oceanweather Inc., 2005). They are hindcasts of a conventional measure of sea surface roughness, *significant wave height* (H_s), defined as the mean of the highest one third of wave heights. The hindcasts are produced by a physical model, calibrated to observed hurricane data, resulting in H_s values on a spatial grid every 30 minutes between September 1900 to September 2005. The motivation for analysing these data comes from marine engineering and more specifically the attempt to develop some design criteria for marine structures, such as oil platforms, where it is of great value to understand better the stochastic behaviour of extreme weather events, such as the hurricanes that occur in the Gulf of Mexico.

Hurricanes are clearly-defined events, so it is straightforward to isolate from raw time series the largest H_s value, the *storm peak significant wave height* H_s^{sp}, for each hurricane event. This eliminates within-event temporal dependence so that H_s^{sp} values from different events can be treated as being independent. The full dataset, which has been analysed by Jonathan and Ewans (2007, 2011), Northrop and Jonathan (2011), consists of hindcast H_s^{sp} values for a 6 × 12 grid of 72 sites in an unnamed location in the Gulf of Mexico. Here we consider a single site (site 31) at the centre of the grid. The data are displayed in figure 1.

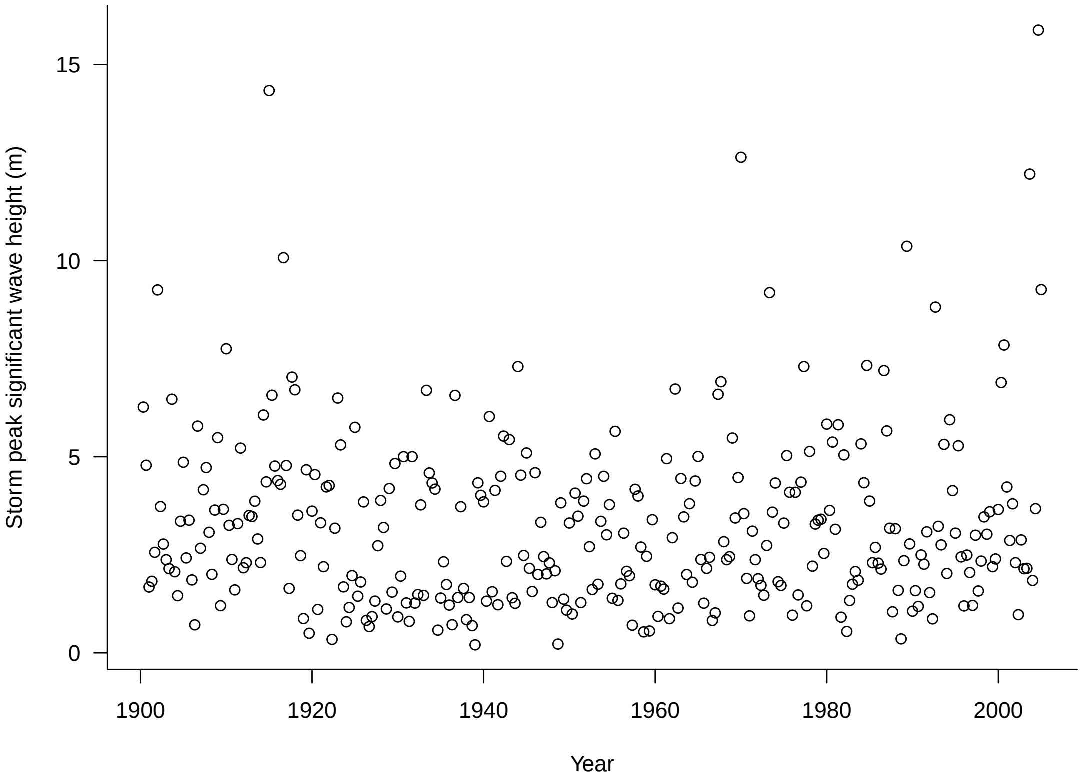

Figure 1: Time series plot of Gulf of Mexico storm peak significant wave heights.

1.4.1 GEV - Block Maxima approach

The first method of extreme value analysis that we describe here is known as the *block maxima* approach. This involves dividing the data into blocks of equal length and then fitting a GEV distribution to the block maxima. A typical choice in environmental applications is a block size equating to one year of observations, which produces annual maxima. It is important to note that the decision about the block size leads to a trade-off between bias and variance. On one hand, deciding on a small block size might violate the asymptotic arguments for the limiting GEV distribution leading to bias. On the other hand, a large block size will provide few points (block maxima) to use for statistical inference and this can result in parameter estimators with high variances. Figure 2 illustrates the blocking procedure using the Gulf of Mexico data.

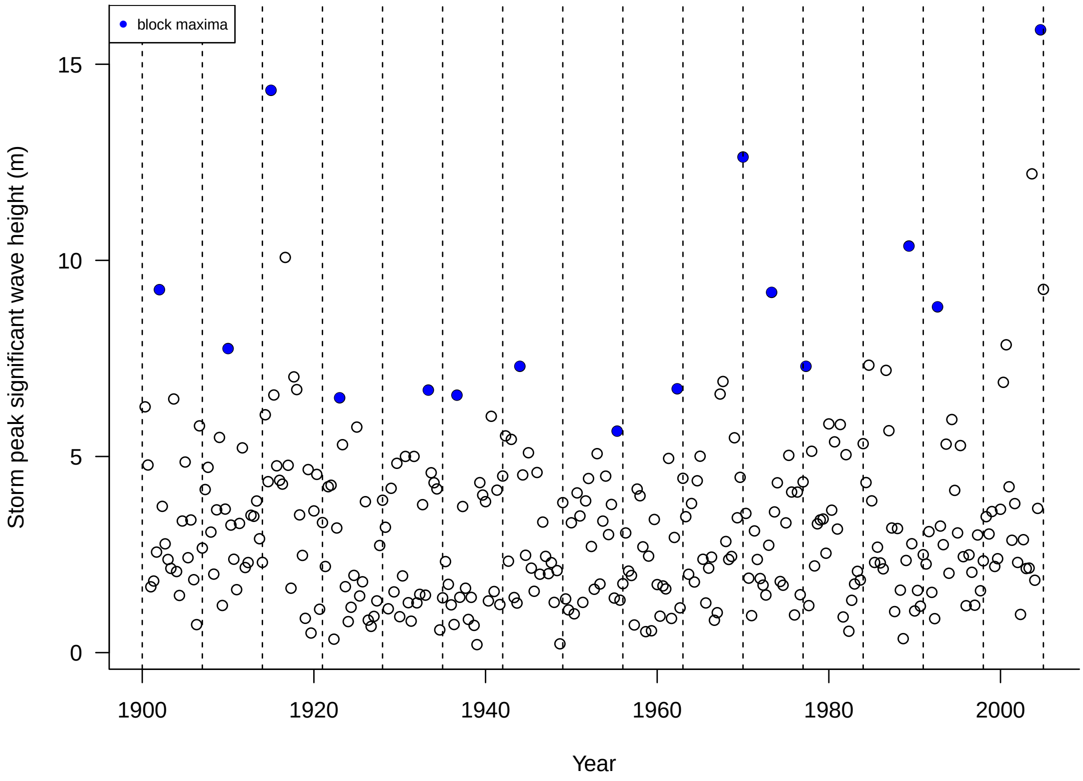

Figure 2: Identification of block maxima.

As this plot is merely illustrative, 15 blocks of 7 years were chosen for convenience. We then treat the block maxima as a random sample from a GEV(μ, σ, ξ) distribution.

1.4.2 GP - Threshold exceedances approach

As an alternative to the block maxima approach, let us assume that a high threshold u is chosen for this dataset. If an observation is higher then u, then this is an exceedance of u and the amount by which the observation exceeds u is the threshold excess. Appealing to theorem 3 suggests the GP distribution as a model for the threshold excesses. This is illustrated in figure 3 below.

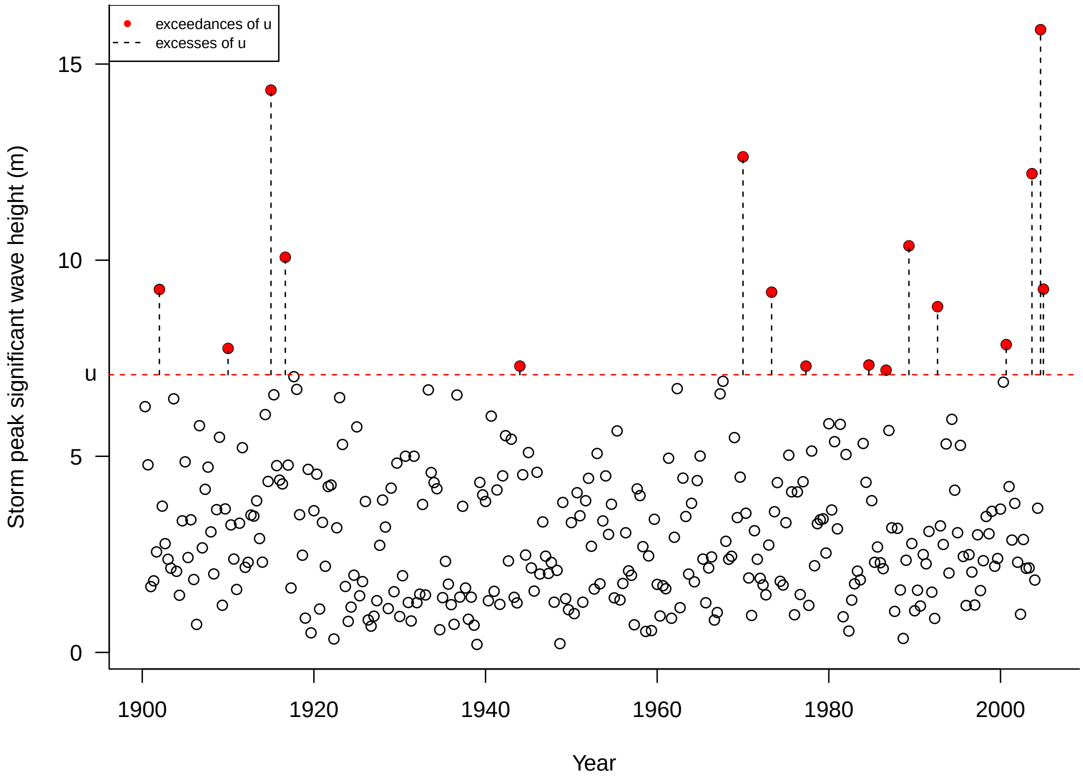

Figure 3: Threshold exceedances and excesses of u.

Here, we have applied a threshold of 7.0798m which corresponds to the 95^{th} quantile of the data. We treat these excesses as a random sample from a GP(σ_u, ξ) distribution.

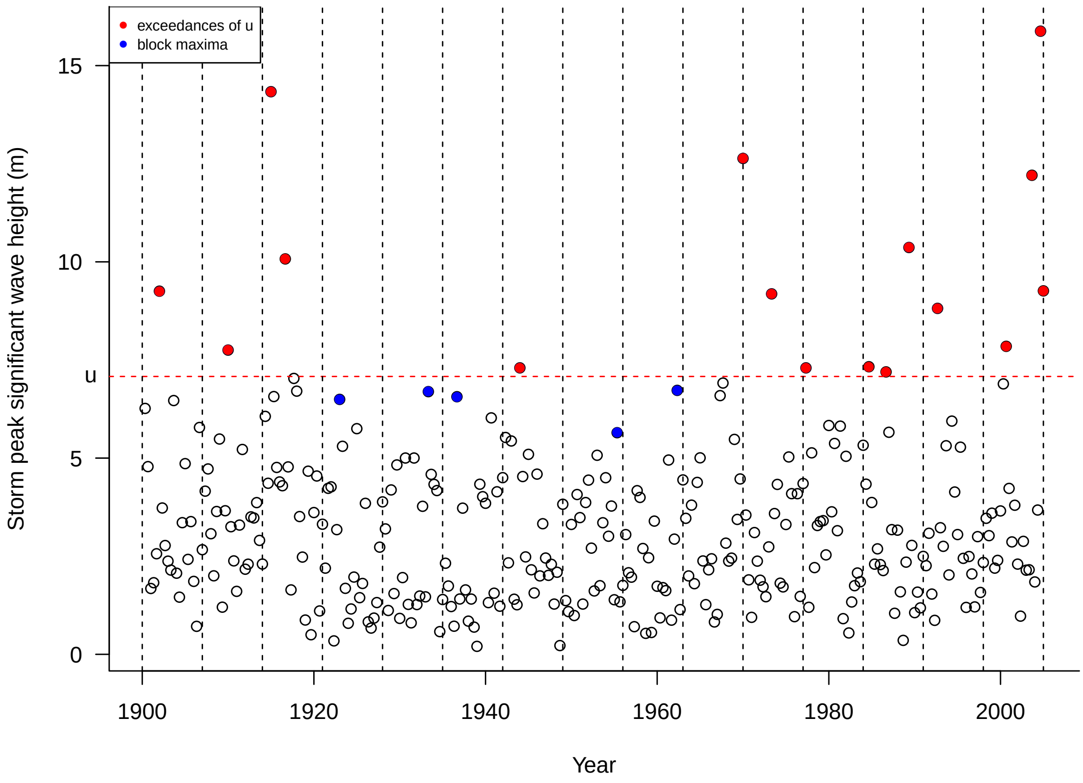

Figure 4: Comparison of block maxima and threshold exceedances.

Figure 4 above illustrates the 'extreme' points through both of the described approaches and shows some important features: (i) that some of the block maxima (shown in blue) are not included in the threshold modelling approach of extreme value analysis, (ii) some second (and third and fourth) largest values in a block are included in the threshold approach and (iii) there are very few exceedances above the threshold from which the inferences about the GP(σ_u, ξ) distribution will be made.

The third point could be addressed by choosing a lower threshold, which will result in more threshold exceedances. However, the solution is not that straightforward because, similarly to the block maxima approach, a bias-variance trade-off is also present in the threshold modelling approach. Choosing too low a threshold leads to bias due to the GP model being inappropriate and too high a threshold results in a small number of exceedances and unnecessarily low estimation precision. We develop new methods for addressing this bias-variance trade-off for a independent and identically distributed stationary process in chapter 3 and for a dependent and identically distributed stationary process in chapter 4.

1.5 Statistical modelling and parameter estimation

One of the tasks that a statistician needs to tackle when having a dataset is to analyse the data and make inferences about the parameters of the supposed random process that generated this data. More specifically, if we are interested in analysing extreme values from a data source, we need to be able to say something about an assumed model and its parameters. By doing so, we can better understand and describe the process and more importantly make inferences on the stochastic behaviour of more extreme observations.

In this section we outline methods of inferences used commonly in extreme value modelling. We concentrate on likelihood-based methods as they can, in principle, be used in a wider variety of modelling situations than competing methods. In particular, they apply more generally than in the simple i.i.d. case we consider initially.

Likelihood and log-likelihood functions

Let us assume that we have a sequence of independent and identically distributed random variables $X_1, \dots, X_m$ with probability density function $f(x_i; \boldsymbol{\theta})$, where the stochastic process of the observed data is characterized by the vector $\boldsymbol{\theta}$, a k-dimensional set of parameters. The joint density of the random variables is defined as the *likelihood function*

$$L(\boldsymbol{\theta}; x_1, \dots, x_m) = \prod_{i=1}^{m} f(x_i; \boldsymbol{\theta}) \quad \text{for} \quad i = 1, \dots, m, \tag{1.3}$$

which is a function defined by the set of unknown parameter vector $\boldsymbol{\theta}$.

Using the fact that the natural logarithm function is monotonic, it is more convenient to work with the *log-likelihood function*

$$\ell(\boldsymbol{\theta}; x_1, \dots, x_m) = \log L(\boldsymbol{\theta}; x_1, \dots, x_m) = \sum_{i=1}^{m} \log f(x_i; \boldsymbol{\theta}). \tag{1.4}$$

Score Function and Fisher Information

The *score function* is defined as the vector (of length k) of the first partial derivatives of the log-likelihood function

$$\mathcal{S}(\boldsymbol{\theta}; X_1, \dots, X_m) = \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}; X_1, \dots, X_m).$$

The score function is itself a vector of random variables and has the following statistical properties. The mean of the score, evaluated at the true set of parameters $\boldsymbol{\theta}$ is found to be zero and its variance is a symmetric $k \times k$ variance-covariance matrix which is called the (expected) *Fisher Information matrix (FI)* and is defined as

$$\mathcal{I}(\boldsymbol{\theta}) = E\left[\left(\frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}; X_1, \dots, X_m)\right)^2\right].$$

Due to the assumed independence and under some regularity conditions the FI matrix can also be written as

$$FI = \mathcal{I}(\boldsymbol{\theta}) = -E\left[\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \ell(\boldsymbol{\theta}; X_1, \dots, X_m)\right].$$

We usually estimate $\mathcal{I}(\boldsymbol{\theta})$ by the observed Fisher information $\mathcal{J}(\boldsymbol{\theta}) = -\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}; x_1, \dots, x_m)$. If the stochastic process of the observed data is characterized by the vector $\boldsymbol{\theta}$ of length k, then matrix FI as defined above will be a $k \times k$ positive semi-definite symmetric matrix.

1.5.1 Maximum Likelihood Estimation

A widely used and flexible approach for parameter estimation is maximum likelihood. The aim of this approach is to obtain the set of parameter estimates for which the joint probability density of the observed data is maximised. In practice, the log-likelihood (1.4) is maximised with respect to $\boldsymbol{\theta}$ to obtain the maximum likelihood estimate (MLE) $\widehat{\boldsymbol{\theta}}$.

On the condition that the log-likelihood is concave, setting the score function $\mathcal{S}(\boldsymbol{\theta}; X_1, \dots, X_m)$ to zero and solving for $\boldsymbol{\theta}$ will result to the vector of maximum likelihood estimators, $\widehat{\boldsymbol{\theta}}$. Furthermore, in regular estimation problems, for a large sample size m it can be shown that, approximately

$$\widehat{\boldsymbol{\theta}} \sim \mathrm{N}(\boldsymbol{\theta}, \mathcal{I}(\boldsymbol{\theta})^{-1}). \tag{1.5}$$

Expressions for the log-likelihood and the Fisher information based on a random sample from a GEV distribution are given in A.1. A.2 gives these expressions for the GP case.

Regularity conditions

Azzalini (1996, page 71) and Davison (2003, page 118) describe the conditions that an estimation problem needs to satisfy in order for maximum likelihood estimation to be *regular*. One regularity condition is that the support of the distribution does not depend on the parameter values. This however, is not the case for either the GEV or GP distribution, so estimating the parameters of these models using maximum likelihood is not automatically a regular estimation problem.

Smith (1985) carried out a theoretical analysis to examine the regularity conditions that are necessary for the estimation to be regular. In cases like the GEV and GP distributions Smith (1985) shows that if $\xi > -1/2$ then maximum likelihood estimation is regular and the resulting maximum likelihood estimator has the usual properties such as (1.5). The reason that this estimation problem is irregular for $\xi \leqslant -1/2$ is that the variance of the score function, $\text{var}[\mathcal{S}(\boldsymbol{\theta}; X_1, \ldots, X_m)]$, does not exist unless $\xi > -1/2$. Smith (1994) extends this result to the regression situation to show that a covariate dependent shape parameter would still need to be greater than $-1/2$ for all values of the covariates.

1.5.2 Bayesian Inference

Let us assume that we have a vector of data $\boldsymbol{x} = (x_1, \ldots, x_m)$ from a sequence of i.i.d. random variables. For example $\boldsymbol{x}$ could represent hindcast storm peak significant wave heights, H_s^{sp} as introduced in 1.4. In maximum likelihood estimation (a *frequentist* method of inference), the parameter vector $\boldsymbol{\theta}$ is viewed as a unknown, but fixed, value to be estimated. In Bayesian inference $\boldsymbol{\theta}$ is viewed as a random variable. A *prior* distribution $\pi(\boldsymbol{\theta})$, representing uncertainty about $\boldsymbol{\theta}$ external to the data $\boldsymbol{x}$, is specified. Prior information about $\boldsymbol{\theta}$, contained in $\pi(\boldsymbol{\theta})$, is combined with information from the data, contained in the likelihood $L(\boldsymbol{x}; \boldsymbol{\theta})$, using Bayes' theorem. This results in a *posterior* distribution $\pi(\boldsymbol{\theta} \mid \boldsymbol{x})$ that is proportional to $L(\boldsymbol{x}; \boldsymbol{\theta})\pi(\boldsymbol{\theta})$ as

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{x}) \propto \pi(\boldsymbol{\theta}) \times L(\boldsymbol{x}; \boldsymbol{\theta}), \tag{1.6}$$

where the normalising constant is $1/\int_\Theta \pi(\boldsymbol{\theta}) L(\boldsymbol{x}; \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}$. Subject to the assumptions made, the posterior distribution gives the distribution of the model parameters conditional on the data observed.

The advantages of a Bayesian analysis (over maximum likelihood estimation) in an extreme value context are:

(a) the $\xi > -1/2$ regularity condition is not required;

(b) information external to the data can be incorporated;

(c) predictive inference, in which predictions of future extreme events account appropriately for model parameter uncertainty, is handled naturally.

Points (b) and (c) are particularly relevant because, by their nature, samples of suitably extreme data can be small. This often results in large uncertainty about model parameters and consequently about extrapolation in the future. Use of prior information can alleviate this problem. Moreover, sample sizes are often small enough that maximum likelihood estimators are far from normally distributed. This means that a frequentist approximation to predictive inference based on normality can be misleading.

In chapters 3 and 4 it is important that we can perform predictive inference reliably, so we take a Bayesian approach. This requires a prior distribution to be specified. In the absence of genuine prior information, the question arises: "Which prior should we use?". We consider this question in chapter 2.

1.5.3 Other methods of inference

There are many methods of inference that have been used in extreme value modelling (see for example Beirlant et al. (2004, Chapter 5), Kotz and Nadarajah (2000), Coles (2001)). Here we briefly describe another popular inference method, namely, the Probability Weighted Moments (PWM), that is often used as an alternative to MLE. This method has been historically used to estimate the parameters of the GEV (Hosking et al., 1985) and the GP (Hosking and Wallis, 1987) distributions and is particularly popular in hydrology and climatology.

Greenwood et al. (1979) introduce the general PWM method which is as follows. For a sequence of i.i.d. random variables $X_1, \ldots, X_m$ with distribution function $F(x)$, the PWM can be found by evaluating

$$E\left[X^p (F(X))^r (1 - F(X))^s\right], \tag{1.7}$$

where p, r and s are real numbers.

In the case of extreme value modelling, the PWM method for obtaining model parameter estimators is limited, similarly to the MLE, to parameter space of the shape parameter. More specifically Hosking et al. (1985) showed that the asymptotic properties of the PWM estimators exist only when $\xi < 0.5$.

Many authors have compared MLE and PWM, including Landwehr et al. (1979), Hosking et al. (1985), Hosking and Wallis (1987), Coles and Dixon (1999), Martins and Stedinger (2000) to name a few. In particular it was found that PWM can perform better for small sample sizes, with the PWM estimators having smaller variance. However, Coles and Dixon (1999) note that the smaller variance of the PWM estimators is partly achieved in place of higher bias as compared to the MLE method due to the parameter space restriction in PWM. Furthermore, it is worth pointing out that the PWM method does not extend easily beyond the i.i.d. case.

1.6 Extreme Value Theory for univariate dependent sequences

So far we have introduced the relevant background theory involving observations from univariate sequences of i.i.d. random variables. In this section we relax the independence property and briefly describe the theory behind non-independent and identically distributed (n.i.d.) random variables, i.e. a stationary sequence of random variables whose joint distribution does not change over time.

It is reasonable to accept, especially in environmental data, that an extreme event can be followed closely by another extreme event and that the assumption of independence is questionable. In fact, it is very common for extreme observations to occur in clusters. The properties of dependent extremes are analysed in detail in Leadbetter et al. (1983, Chapter 3). Chavez-Demoulin and Davison (2012) provide a review of the modelling of the extremes of dependent sequences. We first look at the theory behind block maxima and later concentrate on threshold-based models.

It is not possible to develop a theory for the extremes of dependent sequences that is akin to the UETT of theorem 2, unless some constraint is placed on the form of temporal dependence in the sequence. For example, in the extreme case of perfect dependence, i.e. $X_t = X_1$, for $t = 1, 2, \ldots$, the distribution function of $\max(X_1, \ldots, X_n)$ is identical to that of X_1 (F say) for all n. Therefore, the limiting distribution function of M_n is F, which could be anything. However, to make progress it is only necessary to place a constraint on the strength of long-range dependence at extreme levels: that occurrences of extreme events are approximately independent provided that these events are sufficiently separated in time. More specifically, a sufficient constraint is the following $D(u_n)$ condition (Leadbetter et al., 1983, section 3.2).

Condition $D(u_n)$

Let $X_1, X_2, \ldots$ be a stationary sequence of random variables. The sequence satisfies the $D(u_n)$ condition, if for all $i_1 < \cdots < i_p < j_1 < \cdots < j_q$ with $j_1 - i_p > l$,

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

where A is the event that $X_{i_1} \leqslant u_n, \ldots, X_{i_p} \leqslant u_n$, B is the event that $X_{j_1} \leqslant u_n \ldots, X_{j_q} \leqslant u_n$, there exists a sequence l_n such that $l_n/n \to 0$ as $n \to \infty$ for which $\alpha(n, l_n) \to 0$ as $n \to \infty$.

This is a weak condition (the form of short-term dependence is not restricted) and is plausible for many physical processes. The UETT extends to stationary sequences that satisfy this condition, but the strength of short-term dependence in the extremes of the sequence has an effect on the location, and perhaps the scale, of the limiting GEV distribution, in a way that the following theorem, found in Leadbetter et al. (1983, section 3.3), makes precise.

Theorem 4. *Extremes of dependent sequences.*

Let $X_1, X_2, \ldots$ be a sequence of independent random variables with marginal distribution function F and let $\widetilde{X}_1, \widetilde{X}_2 \ldots$ be a stationary sequence of dependent random variables satisfying $D(u_n)$ condition, with the same marginal distribution function. Let $M_n = \max\{X_1, \ldots, X_n\}$ and $\widetilde{M}_n = \max\left\{\widetilde{X}_1, \ldots, \widetilde{X}_n\right\}$. If, as $n \to \infty$, $P(\{(M_n - b_n)/a_n \leqslant x\} \to G(x)$, for normalising sequences $a_n > 0$ and b_n, then

$$P\left\{(\widetilde{M}_n - b_n)/a_n \leqslant x\right\} \to G^\theta(x) \tag{1.8}$$

for some $0 < \theta \leqslant 1$.

The max-stability of G means that G^θ is a GEV distribution function. Therefore, this theory suggests the GEV distribution as a model for block maxima, as in the independent case.

1.6.1 Extremal index

The quantity θ is known as the *extremal index.* It is the most common measure of the strength of short-term (local) temporal dependence in extremal behaviour. The closer θ is to zero the stronger is the local dependence at extreme levels. For a sequence with $\theta = 1$ there is no local dependence asymptotically but there may be dependence at levels of practical interest.

The extremal index is involved in several characterisations of local extremal dependence based on the extent to which exceedances of a suitably high threshold occur in clusters. Asymptotically (in a sense that we make more precise in section 1.7) the mean number of exceedances in a cluster is given by $1/\theta$ and suitably rescaled times between the last exceedance of one cluster and the first exceedance of the next cluster are exponentially distributed with mean $1/\theta$.

The extremal index is important because it affects extremal inferences. Theorem 4 implies that for large n, $P(\widetilde{M_n} \leqslant x) \approx G^{\theta}(x) = F(x)^{n\theta}$. If we wish to infer from an estimate of F the distribution of the largest value to be observed over some future long time interval then the value of θ matters. Ignoring clustering would lead to overestimation of quantiles of M_n. The extremal index also affects interpretation of extremal inferences because it determines the way in which extreme events (exceedances of some high threshold) occur. Consider two cases, each with the same F. If $\theta = 1$ then threshold exceedances occur singly at some rate λ, say, in time. If $\theta = 1/10$ then threshold exceedances occur in clusters of mean size 10 at a smaller rate $\lambda/10$, i.e. exceedances tend to occur together but there is a larger probability of seeing no such cluster in a given period of time. The difference in behaviour may be important practically.

The presence of local dependence also complicates statistical inference and the selection of an appropriate threshold. Reliable estimation of θ is crucial and many methods have been proposed for achieving this. Of the threshold-based methods we concentrate on those proposed by Ferro and Segers (2003), Süveges (2007) and Süveges and Davison (2010). In section 1.7 we describe the theory underlying these methods and in chapter 4 we use the model proposed by Süveges and Davison (2010) to perform threshold selection. In the next section we outline different general approaches to threshold modelling of serially-dependent extremes.

1.6.2 Threshold-based statistical inference

The presence of local extremal dependence makes threshold-based inferences more difficult than in the independent case. If $\theta < 1$ then there will be some clustering of exceedances at all levels and one cannot eliminate the potential problem of within-cluster dependence between exceedances by setting a high threshold. Although asymptotic theory suggests a GP distribution as a marginal model for (all) threshold excesses it is not appropriate to treat these excesses as independent.

One way round this problem is to extract from the data a set of threshold excesses that *can* be treated as approximately independent. This is achieved by specifying a rule to identify clusters of exceedances, a procedure known as *declustering.* Exceedances greater than a certain number of observations (the *run length*) apart are deemed to be in different clusters, otherwise they are put in the same cluster. Ferro and Segers (2003) automate this process by basing the run length on an estimate of θ. From each cluster the largest excess is extracted, producing a sample of cluster maxima. Then a GP distribution is fitted to these cluster maxima, the *peaks-over-threshold* (POT) approach.

However, Fawcett and Walshaw (2007) demonstrate that, in addition to the loss of statistical precision that results from using only cluster maxima, the declustering process leads to serious bias. They show that it is better to base inferences on *all* threshold excesses: point estimates are based on a likelihood in which the excesses are assumed to be independent, but estimates of parameter uncertainty are adjusted to account for the dependence between these excesses. Fawcett and Walshaw (2012) update this work to incorporate uncertainty about θ in extreme value extrapolations.

An alternative approach is to model within-cluster dependence explicitly (Smith et al., 1997, Fawcett and Walshaw, 2006). This is essential in applications where it is important to gain insight about the nature of this dependence. A common approach is to specify a first-order Markov chain for threshold excesses, based on a particular bivariate extreme value model. However, this raises the issue of which member of the wide class of such models to use. Otherwise, i.e. if it is only necessary to adjust inferences for the strength of local dependence as summarised by θ, then the approach of Fawcett and Walshaw (2012) may be preferable.

It is common to ignore local dependence in extremes at the threshold selection stage, although such dependence can be expected to have an impact. An informal way to include the impact of threshold in terms of local dependence is to choose a threshold above which estimates of the extremal index θ are judged to be insensitive to the threshold. However, in practice estimates of θ often do not stabilise in a clear way as the threshold increases, making this judgement difficult. A more formal model-based approach is developed by Süveges and Davison (2010). In chapter 4 we propose an alternative approach based on the same underlying model, which we describe in the next section.

1.7 K-Gaps exponential mixture model

Let us assume that we have a stationary process $\widetilde{X}_1, \widetilde{X}_2, \ldots,$ with unknown marginal distribution function F. The following theorem is given by Süveges and Davison (2010, page 206). For a sequence of thresholds u_n, introduce the random variable

$$T(u_n) = \min\{k \geqslant 1 : \widetilde{X}_{k+1} > u_n \mid \widetilde{X}_1 > u_n\},$$

for the inter-exceedance times in the sequence $\{\widetilde{X}_i\}$ and the corresponding K-gaps random variable by

$$S^{(K)}(u_n) = \max\{T(u_n) - K, 0\}, \quad K = 0, 1, \ldots.$$

Let $\mathcal{F}_{i,j}(u_n)$ denote the σ-field (Billingsley, 1995, pages 20-21) generated by the events $\widetilde{X}_r \leqslant u_n$, $r = i, \ldots, j$. In simple terms, $\mathcal{F}_{i,j}(u_n)$ defines the possible combinations of the events $\widetilde{X}_r \leqslant u_n$, $r = i, \ldots, j$ that can be assigned probabilities. For any $A \in \mathcal{F}_{1,k}(u_n)$ with $P(A) > 0$, $B \in \mathcal{F}_{k+l,n}(u_n)$ and k, l are integers such that $k = 1, \ldots, n-l$, define

$$\alpha^*(n,l) = \max_k \sup_{A,B} |P(B \mid A) - P(B)|,$$

$\overline{F}(u_n) = 1 - F(u_n)$ and $\widetilde{M}_{r_n} = \max\left\{\widetilde{X}_1, \ldots, \widetilde{X}_{r_n}\right\}$.

Theorem 5. (Süveges and Davison, 2010)

Suppose there exist sequences of integers $\{r_n\}$ and of thresholds $\{u_n\}$ such that as $n \to \infty$, we have $r_n \to \infty$, $r_n\overline{F}(u_n) \to \tau$ and $P(\widetilde{M}_{r_n} \leqslant u_n) \to \mathrm{e}^{-\theta\tau}$ for some $\tau \in (0, \infty)$ and $\theta \in (0,1]$. Moreover, assume that there exists a sequence $l_n = o(n)$ for which $\alpha^*(cr_n, l_n) \to 0$ as $n \to \infty$ for all $c > 0$. Then as $n \to \infty$,

$$P(\overline{F}(u_n)S^{(K)}(u_n) > t) \to \theta \exp(-\theta t), \quad t > 0, \tag{1.9}$$

where the extremal index θ lies in the interval $(0,1]$.

The condition based on $\alpha^*(n,l)$ in theorem 5 is similar to the $D(u_n)$ condition in that it restricts long range dependence at extreme levels. However, it is stronger than the $D(u_n)$ condition because now we are concerned with all combinations of events of the type $\widetilde{X}_i \leqslant u_n$, rather than just $\max_i \widetilde{X}_i \leqslant u_n$. This result motivates an exponential mixture model for the times between exceedances of a high threshold u.

1.7.1 Inter-exceedance times

Let us now suppose that we have N observations from $\widetilde{X}_1, \ldots, \widetilde{X}_m$ that exceed a high threshold u. The *inter-exceedance times* are defined as the times between successive threshold exceedances. Therefore, let $\left\{j_i : \widetilde{X}_{j_i} > u\right\}$ denote the location of an exceedance. Then for $i = 1, \ldots, N-1$ the inter-exceedance times T_i are found by

$$T_i = j_{i+1} - j_i. \tag{1.10}$$

Let $S_i^{(K)} = \max(T_i - K, 0)$ denote the i^{th} K-gap. A pair of exceedances with an inter-exceedance time that is less than the *run parameter* K is deemed to be in the same cluster, and in separate clusters otherwise.

The limiting model (1.9) corresponds to the mixture model

$$\overline{F}(u)S^{(K)} = \begin{cases} 0, & \text{with probability } 1-\theta, \\ W & \text{with probability } \theta \end{cases} \tag{1.11}$$

where W has an exponential distribution with mean $1/\theta$. This generalises the work of Ferro and Segers (2003) who had $K = 0$. The value of K that is appropriate will depend on the dependence structure of the process involved.

It is worth noting here the dual role played by the extremal index θ.

1. The extremal index represents the proportion of non-zero inter-exceedance times and
2. it is the reciprocal of the mean of the distribution of non-zero inter-exceedance times, in other words, it is the rate parameter of an exponential distribution.

Süveges and Davison (2010) use a test to detect misspecification of model (1.11) to inform an appropriate choice of threshold u and run parameter K. In chapter 4 we consider an alternative approach in which, for an appropriate value of K, u is chosen based on the predictive ability of model (1.11) at extreme levels.

1.8 Newlyn data

In this section we briefly describe a motivating example of how extreme value analysis can be applied in a practical situation using the K-gaps exponential mixture model. Firstly we introduce the dataset that is used for this example.

The Newlyn dataset consists of a series of 2894 measurements of sea-surge heights in meters that were taken over the period 1971 - 1976 at a location just off the coast at Newlyn, Cornwall, UK. The data represent the maximum hourly surge heights over periods of 15 hours (see Coles (1991)). Fawcett and Walshaw (2012) used this dataset to estimate the extremal index of the underlying process using several estimators and to make inferences about the extremes of the process. We proceed by first showing the Newlyn data in figure 5 below.

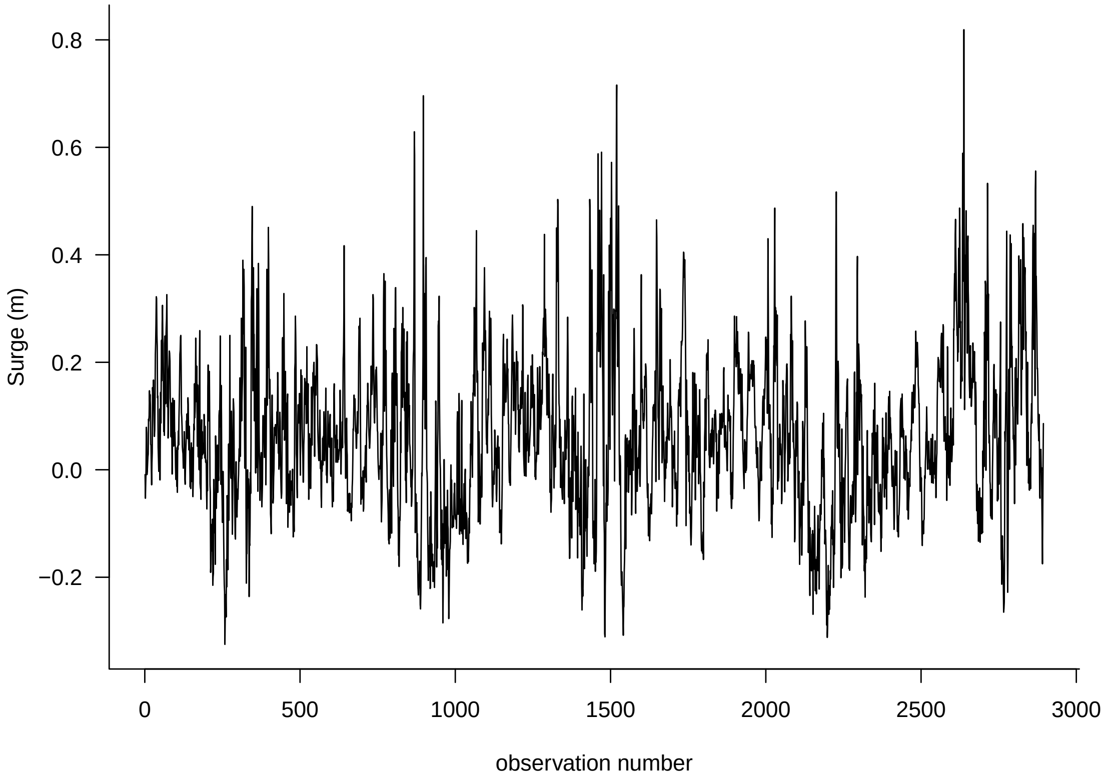

Figure 5: Time series plot of Newlyn sea-surge heights.

In figure 6 we illustrate the procedure of analysing the data using the K-gaps exponential mixture model. For clarity and illustration purposes a small section of 60 observations (around the beginning of 1971) is shown and an 80% sample quantile was selected as threshold.

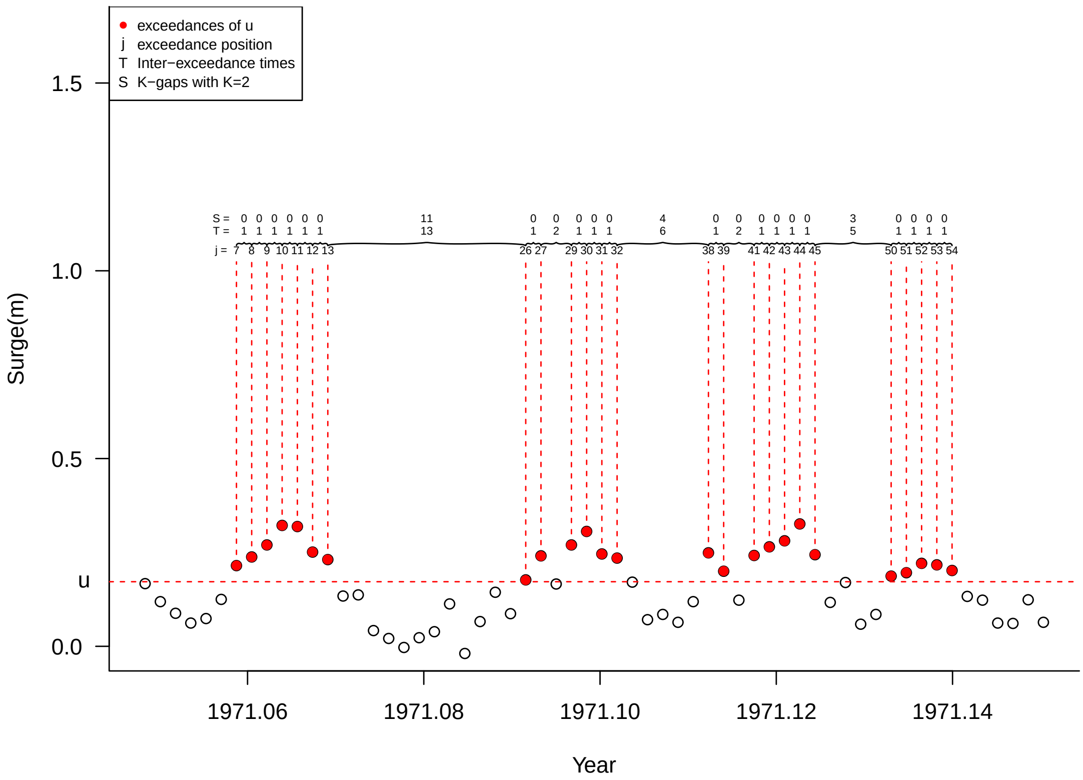

Figure 6: Time series plot of a segment of the Newlyn data illustrating threshold exceedances (red dots), exceedance locations (j), inter-exceedance times (T) and K-gaps (S) for $K = 2$.

1.9 Thesis outline

The aim of this chapter was to introduce univariate extreme value theory and the statistical methods employed to make inferences using extreme value models. Following the question we raised at the end of 1.5.2, chapter 2 considers the use of reference priors in univariate extreme value modelling and present results concerning the propriety of posterior distributions. Furthermore, we consider different methods of performing Bayesian computation, i.e. sampling from the posterior distribution of model parameters and use a simulation study to compare different priors based on predictive inferences for future extreme values. Chapter 3 concerns threshold selection for datasets of independent and identically distributed observations. Bayesian cross-validation is used to compare single thresholds based on predictive ability at extreme levels and we proceed by using Bayesian model averaging to combine inferences from different thresholds. In chapter 4 we extend the approach developed in chapter 3 to the situation where independence is an unrealistic assumption and, in particular, extreme values tend to occur in clusters. Chapter 5 concerns threshold specification in extreme value regression modelling. In the context of a particular model, a theoretical result concerning the optimality of quantile regression is derived. Chapter 6 summarises the main conclusion of this thesis and discusses some possible directions for future research.

2 Bayesian Univariate Extreme Value modelling

We introduced in chapter 1 the background theory related to the analysis of extremes. This involved the various models and approaches to inference in order to estimate the underlying model parameters. From a practical point of view, for example in marine structure design, extreme value analysis is required to provide the design engineers with values that quantify the behaviour of future extremes, of variables such as storm peak significant wave height, over a specified time horizon. One way to view this task is as the prediction of a future extreme observation, such as the largest value M_N to be observed in the next N years, for some large value of N.

Under an *estimative* (or *plug-in* or *nave*) approach, prediction of M_N is based on a model-based distribution into which point estimates of the model parameters are substituted. This is common when using frequentist inference, e.g. when using MLE or PWM. A drawback is that once the model parameters are estimated they are then treated as known, i.e. uncertainty in the values of the model parameters is not incorporated. In contrast, under a *predictive* approach uncertainty in model parameters is incorporated explicitly. In frequentist inference it is possible to try to adjust for parameter uncertainty, e.g. by averaging predictions over the asymptotic (normal) sampling distribution of the parameter estimators, but this distribution may be inappropriate unless the sample size is large, i.e. it may be far from normal. Also, for extreme value models, regularity conditions (on ξ) are required for such asymptotic results to apply (see section 1.5). As we will see in section 2.5 predictive inference is handled naturally under a Bayesian approach, and conditions on the value of ξ are also avoided. For a discussion of the relative merits of estimative and predictive approaches see for example Geisser (1982), Smith (1999), Young and Smith (2005).

The first aim of this chapter is to examine the propriety of the posterior distribution of the model parameter vector $\boldsymbol{\theta}$. This is done by considering a certain type of prior distributions which we discuss in more detail in 2.2 and demonstrate in 2.3 the necessary conditions to yield a proper posterior for each case. In addition we describe two methods of sampling from the posterior distribution in 2.4. We return to the issue of predicting extreme observations in section 2.5 and investigate the choice of prior distribution for this purpose using simulation in 2.6.

2.1 Posterior predictive density

An important part of extreme value analysis is to be able to say something about the probability of future extreme events. Through the Bayesian inference approach we can use the *posterior predictive distribution* to find the probabilities of future extreme observations. Let $x^\dagger$ be a future realisation from the underlying process having probability density function $f(x^\dagger \mid \boldsymbol{\theta})$. Using the posterior distribution $\pi(\boldsymbol{\theta} \mid \boldsymbol{x})$, the posterior predictive density (Aitchison and Dunsmore, 1975) of $x^\dagger$ given $\boldsymbol{x}$ can be found as follows

$$f(x^\dagger \mid \boldsymbol{x}) = \int_\Theta f(x^\dagger \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta} \mid \boldsymbol{x})\,\mathrm{d}\boldsymbol{\theta}. \tag{2.1}$$

However, the prior distribution needs to be specified. In the following sections we direct our attention to the prior distribution $\pi(\boldsymbol{\theta})$, discuss the possible choices available and ultimately investigate the propriety of the resulting posterior distribution.

2.2 Prior distribution

A distinction can be made between *subjective* analyses, in which the prior distribution supplies information from a source such as an expert (Coles and Powell, 1996, Coles and Tawn, 1996, Behrens et al., 2004), from relevant external data (Walshaw, 2000), or from more general experience of the quantity under study (Martins and Stedinger, 2000, 2001), and so-called *objective* analyses (Berger, 2006). In the latter, a prior is constructed using a formal rule, for use when no subjective information is to be incorporated into the analysis. There is disagreement about appropriate terminology for such priors: we follow Kass and Wasserman (1996) in using the term *reference prior*.

2.2.1 Informative priors

The prior, $\pi(\boldsymbol{\theta})$, is a function that describes the distributional behaviour of the model parameter vector $\boldsymbol{\theta}$. A prior which is classified as informative expresses specifically this distributional behaviour. Discussion on the rationale for this choice of prior is present in various research work involving Bayesian inference methodology. More specifically in extreme values analysis Coles and Tawn (1996) use an elicitation scheme for determining the prior distribution by consulting an expert hydrologist. In their analysis of rainfall data they conclude that the use of this type of prior improved their estimates of extremal behaviour. Other examples where prior elicitation has been used within a Bayesian extreme value analysis context are Coles and Powell (1996) and Behrens et al. (2004). In their work Smith and Naylor (1987) used both a seemingly informative and non-informative prior. In fact, this raises an interesting point, that there is a notional middle ground between the two types of priors where some amount of knowledge about the model parameters is known a priori (say from physical characteristics or an expert), but not to the extent where a fully defined prior is constructed. For example de Zea Bermudez and Amaral Turkman (2003) use a weakly informative prior or vague prior for their GP parameter estimation.

In general, proponents of Bayesian inference argue that the fact that an informative prior can be chosen for inference is a huge advantage as this allows for expert knowledge or previous results to be embedded into the inference from the start. This point is reinforced in the context of extremes where extreme data are scarce and prior elicitation allows us to carry out inference using priors that ensure the model is not unrealistic and dismiss any unreasonable or absurd model parameter behaviour.

Whilst we appreciate the benefits of prior elicitation and the valuable inputs that experts can provide prior to the data analysis, one needs to keep in mind that

- the choice of an informative prior is subjective, and
- in most cases where prior elicitation has taken place, it is problem-specific.

It is therefore desirable to propose a Bayesian inference methodology that is not bound to specific problems and allows for a more general extreme value analysis. Therefore, we proceed by discussing the use of non-informative priors.

2.2.2 Non-Informative priors

Contrary to the informative prior, a prior is classified as non-informative when $\pi(\boldsymbol{\theta})$ describes very vaguely and generally the distributional behaviour of the model parameters. In fact, as Kass and Wasserman (1996) argue, the term "non-informative" is somewhat questionable and hard to define since one has to choose the prior and that choice in itself provides some information.

A concern amongst those advocating against the use of informative priors (or Bayesian inference for that matter) is that an informative prior might have the undesired effect of dominating the data resulting in misleading inference about the underlying model. In order to ease this concern the use of non-informative priors typically provides results that are not very different from those obtained through frequentist inference. This is because the likelihood function dominates such a prior in providing information from the data.

The main drawback of these priors is the fact that they are often improper, which in other words, means that they are not proper distributions and fail to integrate to one. However, this might not always cause a problem as the resulting posterior distribution often yields a proper distribution. We examine this aspect of propriety in more detail for the GP and GEV distributions in sections 2.3.1 and 2.3.3 respectively through the use of a specific type of non-informative priors based on formal rules (Kass and Wasserman (1996) provide a comprehensive review of the formal rules that have been proposed) which we call *reference* priors.

2.2.3 Reference priors

Bernardo (1979) introduced reference analysis and was further developed by numerous authors (see for example Bernardo (2005), Berger et al. (2009) and their references) and is one of the most commonly used methods in obtaining objective priors. In this chapter we consider three reference priors that have been used in extreme value analyses:

1. Jeffreys priors (Eugenia Castellanos and Cabras, 2007, Beirlant et al., 2004),
2. maximal data information (MDI) priors (Beirlant et al., 2004), and
3. uniform priors (Pickands, 1994), i.e., independent flat priors on individual parameters.

These priors are *improper*, that is, they do not integrate to a finite number and therefore do not correspond to a proper probability distribution. An improper prior can lead to an improper posterior, which is clearly undesirable. There is no general theory providing simple conditions under which an improper prior yields a proper posterior for a particular model, so this must be investigated case-by-case. Eugenia Castellanos and Cabras (2007) establish that Jeffreys prior for the GP distribution always yields a proper posterior, but no such results exist for the other improper priors we consider. It is important that posterior propriety is established because impropriety may not create obvious numerical problems, for example, Markov Chain Monte Carlo (MCMC) output may appear perfectly reasonable (Hobert and Casella, 1996).

One way to ensure posterior propriety is to use a diffuse proper prior, such as a normal prior with a large variance (Coles and Tawn, 2005, Smith, 2005, Fawcett and

Walshaw, 2006) or by truncating an improper prior (Smith and Goodman, 2000). For example, Coles (2001, chapter 9) uses a GEV(μ, σ, ξ) model for annual maximum sea-levels, placing independent normal priors on μ, $\log \sigma$ and ξ with respective variances $10^4, 10^4$ and 100. However, one needs to check that the posterior is not sensitive to the choice of proper prior and, as Bayarri and Berger (2004) note "…these posteriors will essentially be meaningless if the limiting improper objective prior would have resulted in an improper posterior distribution." In such cases inferences may be sensitive to the diffuseness of the prior, because in the limit as the diffuse prior becomes improper the posterior becomes improper. Therefore, independent uniform priors on separate model parameters are of interest in their own right and represent the limiting case of independent diffuse normal priors.

Let us assume that we have a vector of data $\boldsymbol{x} = (x_1, \ldots, x_m)$ from a sequence of independent and identically distributed random variables $X_1, \ldots, X_m$ with density function $f(x; \boldsymbol{\theta})$, for some parameter vector $\boldsymbol{\theta}$. Furthermore, let us denote the expected Fisher information matrix by $\mathcal{I}(\boldsymbol{\theta})$. We consider the following three reference priors.

Jeffreys priors. Jeffreys "general rule" (Jeffreys, 1961) is

$$\pi_J(\boldsymbol{\theta}) \propto \det(\mathcal{I}(\boldsymbol{\theta}))^{1/2}. \tag{2.2}$$

An attractive property of this rule is that it produces a prior that is invariant to reparameterisation. Jeffreys suggested a modification of this rule for use in location-scale problems. We will follow this modification, which is summarised on page 1345 of Kass and Wasserman (1996). If there is no location parameter then (2.2) is used. If there is a location parameter μ, say, then $\boldsymbol{\theta} = (\mu, \phi)$ and

$$\pi_J(\mu, \phi) \propto \det(\mathcal{I}(\phi))^{1/2}, \tag{2.3}$$

where $\mathcal{I}(\phi)$ is calculated holding μ fixed. In the current context the GP distribution does not have a location parameter whereas the GEV distribution does.

Maximal Data Information (MDI) priors. The MDI priors (Zellner, 1971) are defined as

$$\pi_M(\boldsymbol{\theta}) \propto \exp\left\{\mathrm{E}[\log f(X; \boldsymbol{\theta})]\right\}. \tag{2.4}$$

These are the priors for which the increase in average information, provided by the data via the likelihood function, is maximised. For further information see Zellner (1998).

Uniform priors. Priors that are flat, i.e. equal to a positive constant, say c,

$$\pi_U(\boldsymbol{\theta}) \propto c. \tag{2.5}$$

Flat priors suffer from the problem that they are not automatically invariant to reparameterisation. For example, if we give $\log \sigma$ a uniform distribution then σ is not uniform. Thus, it matters which particular parameterisation is used to define the prior.

2.3 Propriety of posteriors

We continue by looking at the parameterisation of the three reference priors in terms of the GP and GEV distributions and investigate the propriety of the posterior distribution of the relevant model parameters. Proofs of results for the GP distribution can be found in appendices B.2 to B.4 and for the GEV distribution in appendices B.5 to B.8 (Northrop and Attalides, 2015).

2.3.1 GP distribution

Without loss of generality we assume that a certain high threshold is set producing n_u threshold excesses that are ordered: $z_1 < \cdots < z_{n_u}$. Furthermore, for simplicity in our notation we denote the GP scale parameter by σ rather than σ_u. We consider a class of priors of the form $\pi(\sigma, \xi) \propto \pi(\xi)/\sigma, \sigma > 0, \xi \in \mathbb{R}$. In effect we assume *a priori* that σ and ξ are independent and that $\log \sigma$ has an improper uniform prior over the real line.

The posterior distribution for σ and ξ is given by

$$\pi_{GP}(\sigma, \xi \mid \boldsymbol{z}) = C_{n_u}^{-1} \pi(\xi)\, \sigma^{-(n_u+1)} \prod_{i=1}^{n_u} \left[1 + \frac{\xi z_i}{\sigma}\right]_+^{-(1+1/\xi)}, \quad \sigma > 0,\, \xi > -\sigma/z_{n_u},$$

where

$$C_{n_u} = \int_{-\infty}^{\infty} \int_{\max(0, -\xi z_{n_u})}^{\infty} \pi(\xi)\, \sigma^{-(n_u+1)} \prod_{i=1}^{n_u} \left[1 + \frac{\xi z_i}{\sigma}\right]_+^{-(1+1/\xi)} \, \mathrm{d}\sigma \, \mathrm{d}\xi$$

and the inequality $\xi > -\sigma/z_{n_u}$ comes from the constraints $1 + \xi z_i/\sigma > 0$ for $i = 1, \ldots, n_u$ in the likelihood.

Prior distributions

Using (2.2) with $\boldsymbol{\theta} = (\sigma, \xi)$ gives the Jeffreys prior

$$\pi_{J,GP}(\sigma, \xi) \propto \frac{1}{\sigma(1+\xi)(1+2\xi)^{1/2}}, \quad \sigma > 0,\, \xi > -1/2. \tag{2.6}$$

Eugenia Castellanos and Cabras (2007) show that a proper posterior density results for $n_u \geqslant 1$.

Using (2.4) gives the MDI prior

$$\pi_{M,GP}(\sigma, \xi) \propto \frac{1}{\sigma}\,\mathrm{e}^{-(\xi+1)} \quad \sigma > 0,\, \xi \in \mathbb{R}. \tag{2.7}$$

Beirlant et al. (2004, page 447) use this prior but they do not investigate the propriety of the posterior.

Placing independent uniform priors on $\log \sigma$ and ξ, as proposed by Pickands (1994), gives the prior

$$\pi_{U,GP}(\sigma, \xi) \propto \frac{1}{\sigma}, \qquad \sigma > 0,\, \xi \in \mathbb{R}. \tag{2.8}$$

Figure 7 below shows the Jeffreys, MDI and Uniform priors for the GP parameters as functions of ξ (for $\sigma = 1$). The MDI prior increases without limit as $\xi \to -\infty$ and the Jeffreys prior increases without limit as $\xi \downarrow -1/2$.

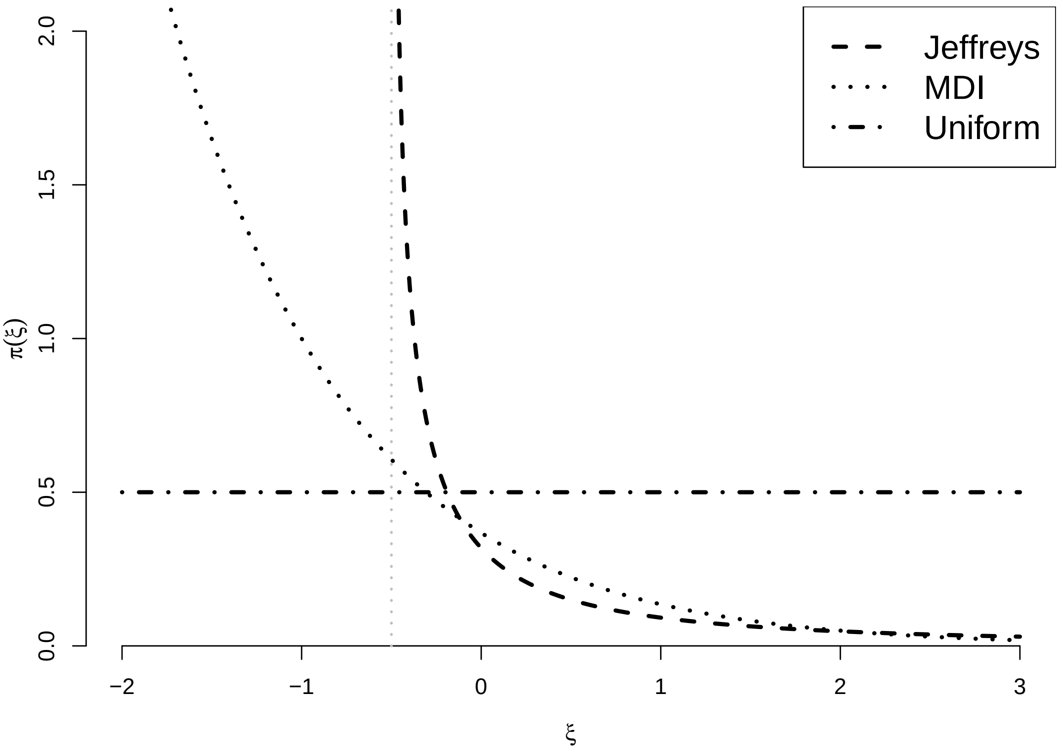

Figure 7: (a) Jeffreys (b) MDI (c) Uniform priors for the GP distribution parameters as a function of ξ, with $\sigma = 1$.

2.3.2 Results for the GP distribution

Theorem 6. *A sufficient condition for the prior $\pi(\sigma, \xi) \propto \pi(\xi)/\sigma, \sigma > 0, \xi \in \mathbb{R}$ to yield a proper posterior density function is that $\pi(\xi)$ is (proportional to) a proper density function.*

The MDI prior (2.7) does not satisfy the condition in theorem 6 because $\exp\{-(\xi + 1)\}$ is not a proper density function on $\xi \in \mathbb{R}$.

Theorem 7. *There is no sample size for which the MDI prior* (2.7) *yields a proper posterior density function.*

The problem with the MDI prior is due to its behaviour for negative ξ so a simple solution is to place a lower bound on ξ *a priori.* This approach is common in extreme value analyses, for example, Martins and Stedinger (2001) constrain ξ to $(-1/2, 1/2)$ *a priori.* We suggest

$$\pi'_{M,GP}(\sigma, \xi) = \frac{1}{\sigma} \mathrm{e}^{-(\xi+1)}, \ \xi \geqslant -1, \tag{2.9}$$

that is, a (proper) unit exponential prior on $\xi + 1$. Any finite lower bound on ξ ensures propriety of the posterior but $\xi = -1$, for which the GP distribution reduces to a uniform distribution on $(0, \sigma)$, seems less arbitrary than other choices as it corresponds to a change in the behaviour of the GP density. For $\xi > -1$, the GP density $f_{GP}(z)$ decreases in z, which is what one anticipates when conducting an extreme value analysis to make inferences about future large, rare values. For $\xi < -1$, $f_{GP}(z)$ increases without limit as it approaches its mode at the upper end point $-\sigma/\xi$, behaviour that is not expected in such analyses.

Corollary to theorem 6. *The truncated MDI prior* (2.9) *yields a proper posterior density function for* $n_u \geqslant 1$.

Theorem 8. *A sufficient condition for the Uniform prior* (2.8) *to yield a proper posterior density function is that* $n_u \geqslant 3$.

2.3.3 GEV distribution

Without loss of generality we take the b block maxima to be ordered: $y_1 < \cdots < y_b$ (where $m = b \times n$, i.e. the total raw data sample size m is a product of the number of blocks b and the block size n). We consider a class of priors of the form $\pi(\mu, \sigma, \xi) \propto \pi(\xi)/\sigma, \sigma > 0, \mu, \xi \in \mathbb{R}$ that is, *a priori* μ, σ and ξ are independent in addition to μ and $\log \sigma$ having improper uniform priors over the real line.

Based on a random sample $y_1, \ldots, y_b$ the posterior distribution for μ, σ and ξ is is given by

$$\pi_{GEV}(\mu, \sigma, \xi \mid \boldsymbol{y}) \propto \sigma^{-(b+1)} \pi(\xi) \exp\left\{ -\sum_{i=1}^{b} \left[1 + \xi \left(\frac{y_i - \mu}{\sigma} \right) \right]_+^{-1/\xi} \right\} \prod_{i=1}^{b} \left[1 + \xi \left(\frac{y_i - \mu}{\sigma} \right) \right]_+^{-(1+1/\xi)} \quad (2.10)$$

If $\xi > 0$ then $\mu - \sigma/\xi < y_1$ and if $\xi < 0$ then $\mu - \sigma/\xi > y_b$.

Prior distributions

Kotz and Nadarajah (2000, page 63) give the Fisher information matrix for the GEV distribution (1.1). Using (2.3) with μ and $\boldsymbol{\phi} = (\sigma, \xi)$ gives the Jeffreys prior

$$\begin{aligned} \pi_{J,GEV}(\mu, \sigma, \xi) &= \frac{1}{\sigma \xi^2} \left\{ [1 - 2\Gamma(2+\xi) + p] \left[\frac{\pi^2}{6} + \left(1 - \gamma + \frac{1}{\xi} \right)^2 - \frac{2q}{\xi} + \frac{p}{\xi^2} \right] \right. \\ & \left. - \left[1 - \gamma + \frac{1}{\xi} - \frac{1}{\xi}\Gamma(2+\xi) - q + \frac{p}{\xi} \right]^2 \right\}^{1/2}, \ \mu \in \mathbb{R}, \sigma > 0, \ \xi > -1/2, \quad (2.11) \end{aligned}$$

where $p = (1{+}\xi)^2\,\Gamma(1{+}2\xi)$, $q = \Gamma(2{+}\xi)\,\{\psi(1+\xi) + (1+\xi)/\xi\}$, $\psi(r) = \partial \log \Gamma(r)/\partial r$ and $\gamma \approx 0.57722$ is Euler's constant. van Noortwijk et al. (2004) give an alternative form for the Jeffreys prior, based on (2.2).

Beirlant et al. (2004, page 435) give the form of the MDI prior:

$$\pi_{M,GEV}(\mu,\sigma,\xi) = \frac{1}{\sigma}\,\mathrm{e}^{-\gamma(\xi+1+1/\gamma)} \propto \frac{1}{\sigma}\,\mathrm{e}^{-\gamma(1+\xi)}, \quad \sigma > 0,\ \mu,\xi \in \mathbb{R}. \tag{2.12}$$

Placing independent Uniform priors on μ, $\log\sigma$ and ξ gives the prior

$$\pi_{U,GEV}(\mu,\sigma,\xi) \propto \frac{1}{\sigma}, \quad \sigma > 0,\ \mu,\xi \in \mathbb{R}. \tag{2.13}$$

Figure 8 below shows the Jeffreys, MDI and Uniform priors for GEV parameters as functions of ξ. The MDI prior increases without limit as $\xi \to -\infty$ and the Jeffreys prior increases without limit as $\xi \to \infty$ and as $\xi \downarrow -1/2$.

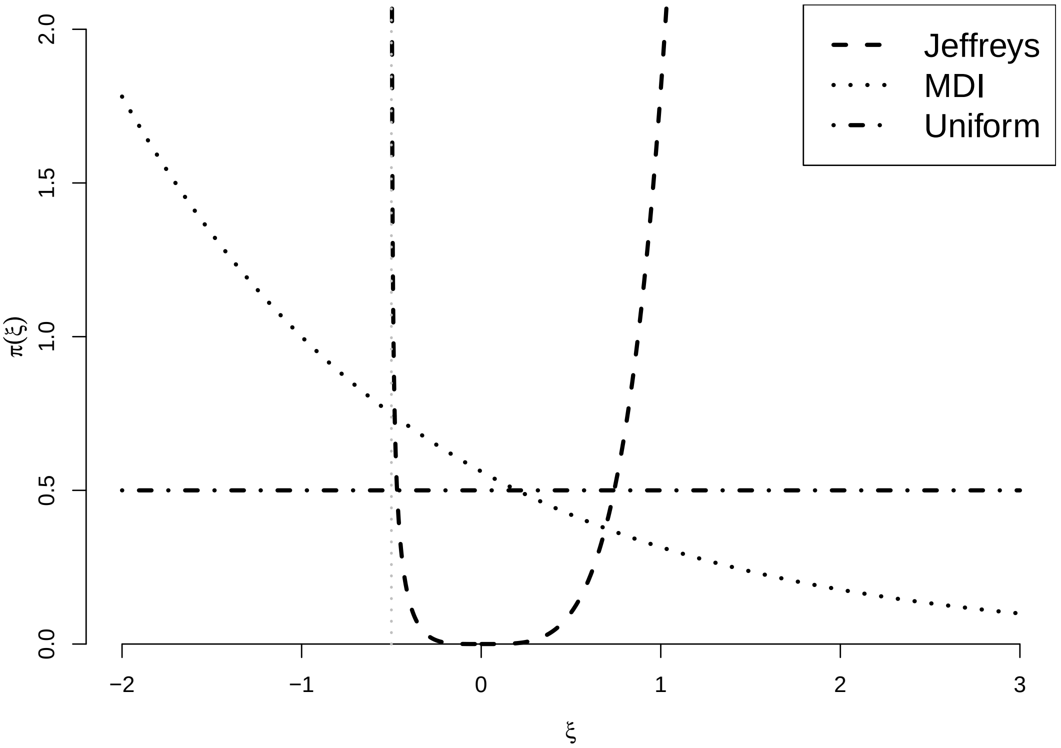

Figure 8: (a) Jeffreys (b) MDI (c) Uniform priors for the GEV distribution parameters as a function of ξ, with $\sigma = 1$.

2.3.4 Results for the GEV distribution

Theorem 9. *For the prior $\pi(\mu,\sigma,\xi) \propto \pi(\xi)/\sigma, \sigma > 0, \mu, \xi \in \mathbb{R}$ to yield a proper posterior density function it is necessary that $b \geqslant 2$ and, in that event, it is sufficient* *that $\pi(\xi)$ is (proportional to) a proper density function.*

Theorem 10. *There is no sample size for which the Jeffreys prior* (2.11) *yields a proper posterior density function.*

Truncation of the independence Jeffreys prior to $\xi \leqslant \xi_+$ would yield a proper posterior density function if $b \geqslant 2$. In this event theorem 9 requires only that $\int_{-1/2}^{\xi_+} \pi(\xi)\, \mathrm{d}\xi$ is finite. From the proof of theorem 10 we have $\pi(\xi) < 2\left[\pi^2/6 + (1-\gamma)^2\right]^{1/2} (1+2\xi)^{-1/2}$ for $\xi \in (-1/2, -1/2+\epsilon)$, where $\epsilon > 0$. Therefore,

$$\begin{aligned} \int_{-1/2}^{-1/2+\epsilon} \pi(\xi)\, \mathrm{d}\xi &< 2\left[\pi^2/6 + (1-\gamma)^2\right]^{1/2} \int_{-1/2}^{-1/2+\epsilon} (1+2\xi)^{-1/2}\, \mathrm{d}\xi, \\ &= 2^{3/2}\left[\pi^2/6 + (1-\gamma)^2\right]^{1/2} \epsilon^{1/2}. \end{aligned}$$

The integral over $(-1/2+\epsilon, \xi_+)$ is also finite. However, the choice of an *a priori* upper limit for ξ may be less obvious than the choice of a lower limit.

Theorem 11. *There is no sample size for which the MDI prior* (2.12) *yields a proper posterior density function.*

As in the GP case, truncating the MDI prior to $\xi \geqslant -1$, that is,

$$\pi'_{M,GEV}(\mu, \sigma, \xi) \propto \frac{1}{\sigma}\, \mathrm{e}^{-\gamma(1+\xi)} \quad \mu \in \mathbb{R}, \sigma > 0,\ \xi \geqslant -1, \tag{2.14}$$

is one way to yield a proper posterior distribution.

Corollary to theorem 9. *The truncated MDI prior* (2.14) *yields a proper posterior density function for $b \geqslant 2$.*

Theorem 12. *A sufficient condition for the Uniform prior* (2.13) *to yield a proper posterior density function is that $b \geqslant 4$.*

2.4 Sampling from the posterior distribution

Having established the conditions to obtain a proper posterior distribution for the GP and GEV model parameters we turn to the Bayesian inference step involving the method of sampling from the posterior. The benefit of sampling variates of the model parameters from the posterior distribution is the fact that we can directly account for the underlying uncertainty of the parameters. This benefit transfers to the later stage of the analysis when predicting future extreme events that account for this uncertainty.

In this section we describe two classical Monte Carlo methods for universal sampling that are well known. The first method involves the popular Markov Chain Monte Carlo (MCMC) and more specifically, we use one possible MCMC method known as the Metropolis-Hastings (MH) algorithm to sample from the posterior. In MCMC a Markov chain is set up such that its equilibrium distribution is the desired posterior distribution. A realisation from the Markov chain is simulated, from some starting value. After a sufficiently large number of time steps (the *burn-in period*) the states of the chain are treated as a dependent values sampled from a distribution that approximates the posterior. The second method is exact and produces values that are sampled independently from the posterior. It is a form of rejection sampling called the generalised Ratio of Uniforms (RoU) method. We continue by first describing the MH algorithm (Metropolis and Ulam, 1949, Metropolis et al., 1953, Hastings, 1970, Liu, 2004) and introduce two special cases of this technique.

2.4.1 Metropolis-Hastings (MH)

Firstly, we need to have a starting value of the Markov chain. Care needs to be taken when choosing this point as it has to ensure that the distributional constraints are not violated. Secondly, we need a *proposal distribution*, which is used to propose candidate values for the chain. A limitation of this method is the fact that the correlation among the generated samples can be high. This means that a smaller number of posterior samples can be accepted, whilst facing the risk of causing the Markov chain to get trapped in local modes resulting to very slow convergence. Additionally, when carrying out this method, one needs to allow a reasonable burn-in period and essentially establish that the Markov chain has converged, a task that can be difficult.

Let us assume that we have a vector of data $\boldsymbol{x} = (x_1, \ldots, x_m)$ and the vector $\boldsymbol{\theta}$ represents the model parameters. Furthermore, let the current state of the Markov chain be $\boldsymbol{\theta}^i$, the starting point of the chain is $\boldsymbol{\theta}^0$ and the candidate values $\boldsymbol{\theta}^*$. The algorithm proceeds by calculating the probability of acceptance of $\boldsymbol{\theta}^*$ as follows

$$p_{\alpha,MH}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^i; \boldsymbol{x}) \quad = \quad \min\left\{ \frac{\pi(\boldsymbol{\theta}^* \mid \boldsymbol{x}) q(\boldsymbol{\theta}^i \mid \boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta}^i \mid \boldsymbol{x}) q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^i)}, 1 \right\}, \qquad (2.15)$$

where, $\pi(\cdot \mid \boldsymbol{x})$ is (up to proportionality) the posterior density given the data and $q(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1)$ is the conditional proposal density for $\boldsymbol{\theta}_2$ given that the current state is $\boldsymbol{\theta}_1$. If the candidate values are accepted, then the chain moves to that point (which becomes the current state) and new candidate values are provided. If $\boldsymbol{\theta}^*$ is rejected, then the chain remains at the current point and new candidate values are provided.

This is repeated for a number of iterations.

Algorithm

The steps to carry out the MH algorithm are as follows:

1. Choose an arbitrary starting point for the Markov chain $\boldsymbol{\theta}^0$.
2. Use the proposal distribution $q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^i)$ to provide a candidate point $\boldsymbol{\theta}^*$, given $\boldsymbol{\theta}^i$.
3. Calculate the probability of acceptance as in (2.15).
4. Decision rule: Accept $\boldsymbol{\theta}^*$ if $p_{\alpha,MH} \geqslant u$, where u is a random value from a Uniform[0,1] distribution, otherwise reject $\boldsymbol{\theta}^*$ and repeat process from step 2.

The proposal distribution is a key component of the algorithm and therefore it matters how this distribution is defined. For that reason we continue by looking at two special cases of this Monte Carlo technique.

Random Walk

For the random walk MH algorithm, the proposal distribution *depends* on the current state of the chain. In other words, $\boldsymbol{\theta}^*$ is a stochastic jump from the current state $\boldsymbol{\theta}^i$. A common proposal distribution for this method is a multivariate Normal distribution centred on the current state $\boldsymbol{\theta}^i$ and with a suitable covariance matrix such as the (scaled) inverse observed information matrix for $\boldsymbol{\theta}$, evaluated at $\boldsymbol{\theta}^i$. Scaling the covariance matrix appropriately can be important to produce a chain that moves rapidly around the posterior distribution, see Bennett et al. (1996, chapter 19) for details. As this particular proposal density is symmetric about $\boldsymbol{\theta}^i$ we have $q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^i) = q(\boldsymbol{\theta}^i \mid \boldsymbol{\theta}^*)$, so that (2.15) simplifies.

Independence Sampler

The second type is the independence sampler MH algorithm (Tierney, 1994). Contrary to random walk, the proposal distribution is *independent* of the current state of the chain and the probability of acceptance is defined as

$$p^{ind}_{\alpha,MH}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^i; \boldsymbol{x}) \quad = \quad \min\left\{\frac{\pi(\boldsymbol{\theta}^* \mid \boldsymbol{x})q(\boldsymbol{\theta}^i)}{\pi(\boldsymbol{\theta}^i \mid \boldsymbol{x})q(\boldsymbol{\theta}^*)}, 1\right\}. \tag{2.16}$$

A common proposal distribution for the independence sampler is a multivariate Normal, centred on the empirical maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$ and with the (scaled) inverse observed information matrix for $\boldsymbol{\theta}$, evaluated at $\widehat{\boldsymbol{\theta}}$, as its covariance matrix.

2.4.2 Ratio of uniforms (RoU)

The ratio-of-uniforms (RoU) method is a sampling technique proposed by Kinderman and Monahan (1977) and offers itself as an alternative method for generating variates from a density function (see for example Ripley (1987)). Benefits of this method, in comparison to MCMC, are the fact that the generated samples are independent and sampled from the exact posterior. This is particularly useful for simulation studies involving analyses of many simulated datasets and comes in handy later on in section 2.6 where, for the GP case we need to sample from a bivariate posterior distribution and wish to do this efficiently. Therefore, we use the following extension of the conventional RoU method.

Multivariate generalised RoU with relocation

Wakefield et al. (1991) adapt the conventional (univariate) RoU method in three ways: a tuning parameter r (see theorem 13) and a relocation parameter $\boldsymbol{\mu}$ are introduced and the method is generalised to an arbitrary number of dimensions. The former adaptations enable fine tuning of the method to increase its efficiency. The RoU algorithm is justified by the following result which we give for the bivariate case, where we have in mind the posterior distribution for GP parameters $\boldsymbol{\theta} = (\sigma, \xi)$ based on modelling a random sample $\boldsymbol{z} = (z_1, \ldots, z_{n_u})$ of threshold excesses.

Theorem 13. *Let h be a positive integrable function over a subset $\mathcal{X}$ of $\mathbb{R}^2$. Suppose that the variables (U, V_σ, V_ξ) are uniformly distributed over the region*

$$\mathcal{A}_h(r, \boldsymbol{\mu}) = \left\{ (u, v_\sigma, v_\xi) : 0 \leqslant u \leqslant \left[h\left(\frac{v_\sigma}{u^r} + \mu_\sigma, \frac{v_\xi}{u^r} + \mu_\xi \right) \right]^{\frac{1}{2r+1}} \right\}, \tag{2.17}$$

where $r \geqslant 0$ and $\boldsymbol{\mu} = (\mu_\sigma, \mu_\xi)$. Then $(V_\sigma/U^r + \mu_\sigma, V_\xi/U^r + \mu_\xi)$ has density $h/\int h$.

In practice it is usually not possible directly to sample uniformly over $\mathcal{A}_h(r, \boldsymbol{\mu})$. Therefore, a rejection method is employed in which $\mathcal{A}_h(r, \boldsymbol{\mu})$ is enclosed within a bounding region, over which it is easy to simulate uniformly. Simulated points that lie in $\mathcal{A}_h(r, \boldsymbol{\mu})$ are accepted; otherwise they are rejected. Provided that over $\mathcal{X}$, h, $x_1^{2r+1}[h(x_1 + \mu_\sigma, x_2)]^r$ and $x_2^{2r+1}[h(x_1, x_2 + \mu_\xi)]^r$ are bounded, a simple bounding cuboid can be used. In the current context this is

$$0 \leqslant u \leqslant u^+(r, \boldsymbol{\mu}), \quad v_\sigma^-(r, \boldsymbol{\mu}) \leqslant v_\sigma \leqslant v_\sigma^+(r, \boldsymbol{\mu}), \quad v_\xi^-(r, \boldsymbol{\mu}) \leqslant v_\xi \leqslant v_\xi^+(r, \boldsymbol{\mu}),$$

where,

$$\begin{aligned}
u^+(r, \boldsymbol{\mu}) &= u^+(r) = \sup_{\sigma,\xi} \, [h(\sigma, \xi)]^{\frac{1}{2r+1}} , \\
v_\sigma^-(r, \boldsymbol{\mu}) &= \inf_{\sigma \leqslant 0} \sigma \, [h(\sigma + \mu_\sigma, \xi + \mu_\xi)]^{\frac{r}{2r+1}} , \\
v_\sigma^+(r, \boldsymbol{\mu}) &= \sup_{\sigma \geqslant 0} \sigma \, [h(\sigma + \mu_\sigma, \xi + \mu_\xi)]^{\frac{r}{2r+1}} , \\
v_\xi^-(r, \boldsymbol{\mu}) &= \inf_{\xi \leqslant 0} \xi \, [h(\sigma + \mu_\sigma, \xi + \mu_\xi)]^{\frac{r}{2r+1}} , \\
v_\xi^+(r, \boldsymbol{\mu}) &= \sup_{\xi \geqslant 0} \xi \, [h(\sigma + \mu_\sigma, \xi + \mu_\xi)]^{\frac{r}{2r+1}} .
\end{aligned}$$

Now suppose that $h(\sigma, \xi) = L(\sigma, \xi; \boldsymbol{z}) \, \pi(\sigma, \xi)$, so that $h(\sigma, \xi) \propto \pi(\sigma, \xi \mid \boldsymbol{z})$. Theorem 13 justifies the following algorithm for sampling from $\pi(\sigma, \xi \mid \boldsymbol{z})$.

1. Generate values of u, v_σ and v_ξ independently from
   (a) $U \ \sim U\left(0, u^+(r)\right)$,
   (b) $V_\sigma \sim U\left(v_\sigma^-(r, \boldsymbol{\mu}), v_\sigma^+(r, \boldsymbol{\mu})\right)$,
   (c) $V_\xi \sim U\left(v_\xi^-(r, \boldsymbol{\mu}), v_\xi^+(r, \boldsymbol{\mu})\right)$.
2. If $u \leqslant \left[h\left(\frac{v_\sigma}{u^r} + \mu_\sigma, \frac{v_\xi}{u^r} + \mu_\xi\right)\right]^{\frac{1}{2r+1}}$ then accept the candidate $\boldsymbol{\theta}^* = (v_\sigma/u^r, v_\xi/u^r)$. Otherwise, reject $\boldsymbol{\theta}^*$ and repeat step 1.

An arbitrary candidate is accepted with probability

$$p_{\alpha,RoU}(r, \boldsymbol{\mu}) \quad = \quad \frac{1}{2r+1} \, \frac{\int\int h(\sigma, \xi) \, d\sigma \, d\xi}{u^+(r) \left[v_\sigma^+(r, \boldsymbol{\mu}) - v_\sigma^-(r, \boldsymbol{\mu})\right] \left[v_\xi^+(r, \boldsymbol{\mu}) - v_\xi^-(r, \boldsymbol{\mu})\right]} \tag{2.18}$$

Wakefield et al. (1991) recommend the general strategy of mode relocation, i.e. moving the mode of h to zero by setting $\boldsymbol{\mu}$ to be equal to the mode of h (here the *maximum a posteriori* estimate of $\boldsymbol{\theta}$), and $r = 1/2$. This is supported by some exact theoretical results for certain special cases (e.g. normal densities and symmetric unimodal densities) and practical experience sampling from various posterior distributions. We expect $\pi(\sigma, \xi \mid \boldsymbol{z})$ to be unimodal but it will tend to be asymmetric for small numbers of threshold excesses. However, we have found that this strategy is sufficiently efficient for our purposes, even when carrying out simulation studies involving datasets with sample sizes as small as 25. A problem can be encountered if the Jeffreys prior (2.6) is used, because, particularly for small sample sizes, it is possible for the posterior to be unbounded as $\xi \downarrow -1/2$. This means that the region $\mathcal{A}_h(r, \boldsymbol{\mu})$ in (2.17) cannot be bounded as described above. However, this is of no real concern as we will see in section 2.6 that other priors are preferable to the Jeffreys prior in the GP case.

2.5 Prediction of extreme observations

In an extreme value analysis the main focus is often the estimation of extreme quantiles called *return levels*. Let us assume that we have a vector of data $\boldsymbol{x} = (x_1, \ldots, x_m)$ from a sequence of independent and identically distributed random variables $X_1, \ldots, X_m$. Let $m = N \times n_y$, where now N is the time horizon in years and n_y is the mean number of observations per year.

Let M_N denote the largest value observed over a time horizon of N years. We proceed by first defining a quantity of interest that has been commonly used throughout extreme value analysis (see for example Coles (2001)), based on the random variable M_1 (the annual maximum). The *N-year return level* x^N is defined as the value exceeded by M_1 with probability $1/N$, or equivalently

$$P(M_1 \leqslant x^N) = 1 - 1/N, \tag{2.19}$$

For a high threshold u and under a Bin-GP model (introduced in section 1.3.1) with parameter vector $\boldsymbol{\theta} = (p_u, \sigma, \xi)$ we have that for $x > u$, the distribution function of M_1 is

$$P(M_1 \leqslant x) = F_{M_1}(x; \boldsymbol{\theta}) = F(x; \boldsymbol{\theta})^{n_y} = \left\{1 - p_u \left[1 + \xi \left(\frac{x-u}{\sigma}\right)\right]^{-1/\xi}\right\}^{n_y}, \tag{2.20}$$

where $F(x; \boldsymbol{\theta})$ is the distribution function of the random variable X. Deciding on the value for N and solving (2.19) using (2.20) for x^N will provide the desired return level. Typical values of N such as 100, 1000, 10,000 come from a variety of fields, for example, off-shore engineering design criteria for marine structures.

A related approach defines the quantity of interest as the random variable M_N, rather than particular quantiles of M_1. Similarly to (2.20), for $x > u$, the distribution function for M_N is

$$P(M_N \leqslant x) = F_{M_N}(x; \boldsymbol{\theta}) = F(x; \boldsymbol{\theta})^{Nn_y} = \left\{1 - p_u \left[1 + \xi \left(\frac{x-u}{\sigma}\right)\right]^{-1/\xi}\right\}^{Nn_y} \tag{2.21}$$

For large N ($N = 100$ is sufficient), x^N is approximately equal to the 37% quantile of the distribution of M_N (Cox et al., 2002). In an estimative approach, based on a point estimate of $\boldsymbol{\theta}$, the value of x^N is below the median of M_N. What is more, a common interpretation of x^N is the level exceeded on average once every N years. However, for large N (again $N = 100$ is sufficient) and under an assumption of independence at extreme levels, x^N is exceeded 0, 1, 2, 3, 4 times with respective approximate probabilities of 37%, 37%, 18%, 6% and 1.5%. It may be more instructive to examine directly the distribution of M_N, rather than very extreme quantiles of the annual maximum M_1.

The relationship between these two approaches is less clear under a predictive approach, in which uncertainty about $\boldsymbol{\theta}$ is incorporated into the calculations. Here we consider the case where this is achieved using a Bayesian posterior distribution for $\boldsymbol{\theta}$. The N-year *(posterior) predictive return level* x_P^N is the solution of

$$P(M_1 \leqslant x_P^N \mid \boldsymbol{x}) = \int F_{M_1}(x_P^N; \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \mid \boldsymbol{x})\ \mathrm{d}\boldsymbol{\theta} = 1 - 1/N.$$

The predictive distribution function of M_N is given by

$$P(M_N \leqslant x \mid \boldsymbol{x}) = F_{M_N}(x; \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \mid \boldsymbol{x})\ \mathrm{d}\boldsymbol{\theta}. \tag{2.22}$$

As noted by Smith (2003, section 1.3), accounting for parameter uncertainty tends to lead to larger estimated probabilities of extreme events, that is, x_P^N tends to be greater than an estimate $\widehat{x}^N$ based on, for example, the MLE. The strong non-linearity of $F_{M_1}(x; \boldsymbol{\theta})$ for large x, and the fact that it is bounded above by 1, mean that averages of $F_{M_1}(x; \boldsymbol{\theta})$ over areas of the parameter space relating to the extreme upper tail of M_1 tend to be smaller than point values near the centre of such areas. This phenomenon is less critical when working with the distribution of M_N because now central quantiles of M_N also have relevance, not just particular extreme tail probabilities.

For a given value of N, we estimate $P(M_N \leqslant x \mid \boldsymbol{x})$ using a sample $\boldsymbol{\theta}_j, j = 1, \ldots, n_{\boldsymbol{\theta}}$ from the posterior density $\pi(\boldsymbol{\theta} \mid \boldsymbol{x})$ to give

$$\widehat{P}(M_N \leqslant x \mid \boldsymbol{x}) = \frac{1}{n_{\boldsymbol{\theta}}} \sum_{j=1}^{n_{\boldsymbol{\theta}}} F(x; \boldsymbol{\theta}_j)^{n_y N}. \tag{2.23}$$

The solution $\widehat{x}_P^N$ of $\widehat{P}(M_1 \leqslant \widehat{x}_P^N \mid \boldsymbol{x}) = 1 - 1/N$ provides an estimate of x_P^N.

2.6 Simulation study: priors for GP parameters

A Bayesian approach to predictive inference requires a prior distribution to be specified. In the absence of genuine prior information we may wish, in the first instance, to use a reference prior, but which one should we use? In section 2.6 we carry out a simulation study to compare the performance of GP priors described in section 2.2.3 in the context of predicting M_N using a Bin-GP model. We also illustrate the undesirable consequences of ignoring parameter uncertainty under an estimative approach.

We use the Jeffreys prior $p_u \sim \text{Beta}(1/2, 1/2)$ for the Binomial parameter p_u. This leads to a $\text{Beta}(n_u + 1/2, m - n_u + 1/2)$ posterior for p_u, where m is the total sample size and n_u is the number of threshold excesses. For the GP parameters we initially consider three prior distributions:

1. Jeffreys prior, $\pi_{J,GP}(\sigma, \xi)$ as shown in (2.6).
2. Truncated MDI prior, $\pi'_{M,GP}(\sigma, \xi)$ as shown in (2.9).
3. Uniform (or flat) prior, $\pi_{U,GP}(\sigma, \xi)$ as shown in (2.8).

Motivated by findings presented later in this section we generalise the truncated MDI prior to MDI(a):

$$\pi'_{M,GP}(\sigma, \xi; a) \propto \frac{1}{\sigma} a \, \mathrm{e}^{-a(\xi+1)} \quad \sigma > 0, \, \xi \geqslant -1, a > 0 \tag{2.24}$$

and also include this prior to our comparison. Figure 9 below shows the Jeffreys, truncated MDI, generalised truncated MDI (for $a = 0.6$) and Uniform priors for the GP parameters as functions of ξ (for $\sigma = 1$).

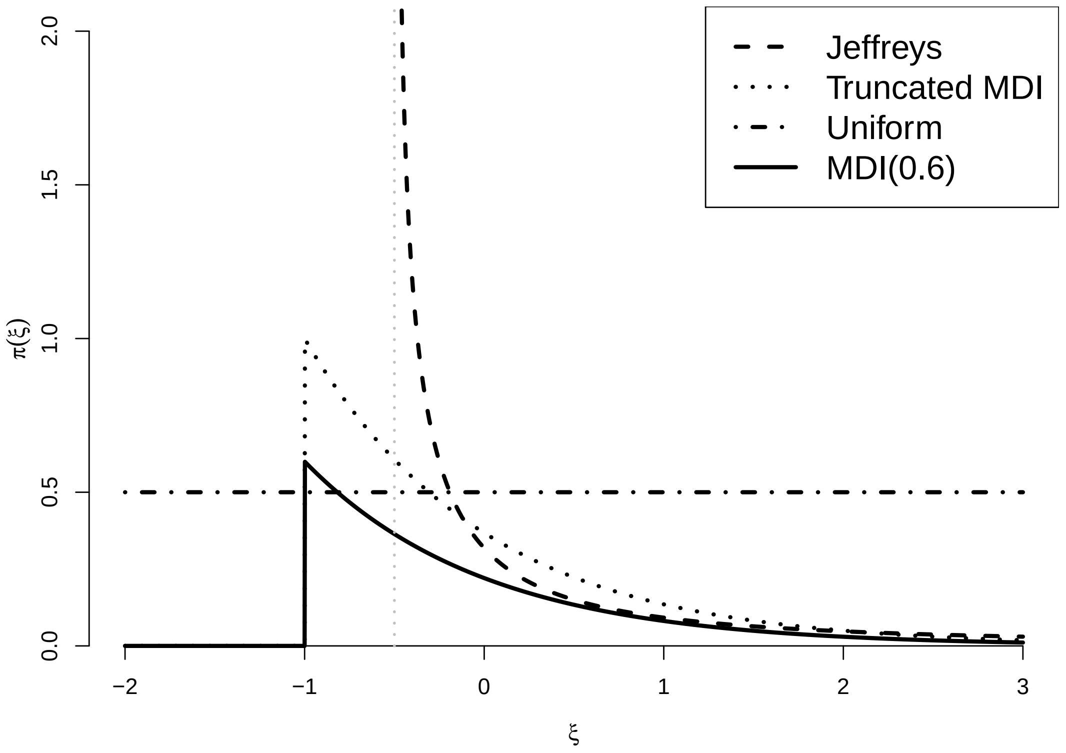

Figure 9: (a) Jeffreys (b) truncated MDI (c) generalised truncated MDI(0.6) (d) Uniform priors for the GP distribution parameters as a function of ξ, with $\sigma = 1$.

The Jeffreys prior (2.6) is unbounded as $\xi \downarrow -1/2$. If there are small numbers of threshold excesses this can result, particularly if $\xi < 0$, in a bimodal posterior distribution, with one mode at the boundary $\xi = -1/2$. This seems undesirable and makes sampling from the posterior more difficult. In the simulation study we also find that, notwithstanding these issues, the Jeffreys prior results in poorer predictive performance than the truncated MDI (2.9) and Uniform (2.8) priors.

Let $x^\dagger$ be a future N-year maximum, sampled from a distribution with distribution function $F(x;\boldsymbol{\theta})^{n_y N}$ (i.e. the distribution function of M_N). If the posterior predictive distribution function (2.22) is the same as that of $x^\dagger$ then $P(M_N \leqslant x^\dagger \mid \boldsymbol{x})$ has a Uniform distribution on $(0,1)$. In practice this can only hold approximately. The closeness of the approximation under repeated sampling provides a basis for comparing different prior distributions. Performance of an estimative approach based on the MLE $\widehat{\boldsymbol{\theta}}$ can be assessed using $F(x^\dagger;\widehat{\boldsymbol{\theta}})^{n_y N}$.

Simulation scheme

For a given prior distribution and given values of N and n_y the simulation scheme is as follows:

1. Simulate a dataset $\boldsymbol{x}_{\text{sim}}$ of $n_{\boldsymbol{x}}$ independent observations from a Bin-GP(p_u, σ, ξ) distribution.

2. Sample $\boldsymbol{\theta}_j, j = 1, \ldots, n_{\boldsymbol{\theta}}$ from the posterior distribution $\pi(\boldsymbol{\theta} \mid \boldsymbol{x}_{\text{sim}})$.

3. Simulate an observation $x^{\dagger}$ from the distribution of M_N as follows:

   (a) Simulate n_u, where $n_u \sim \text{Bin}(N n_y, p_u)$.

   (b) Simulate $x^{\dagger}$ from the distribution of $\max(X_1, \ldots, X_{n_u})$, where $X_i \overset{iid}{\sim}$ GP(σ, ξ), $i = 1, \ldots, n_u$.

4. Use (2.23) to evaluate $\widehat{P}(M_N \leqslant x^{\dagger} \mid \boldsymbol{x}_{\text{sim}})$.

Steps 1. to 4. are repeated 10,000 times, providing a putative sample of size 10,000 from a U$(0, 1)$ distribution. In the frequentist approach step 4 is $F(x^{\dagger}; \widehat{\boldsymbol{\theta}})^{n_y N}$. For this simulation study we produce samples of size $n_{\boldsymbol{\theta}} = 1,000$ from the posterior distribution $\pi(\boldsymbol{\theta} \mid \boldsymbol{x}_{\text{sim}})$ using the generalised ratio-of-uniforms method of Wakefield et al. (1991), following their suggested strategy of relocating the mode of $\pi(\boldsymbol{\theta} \mid \boldsymbol{x}_{\text{sim}})$ to the origin and using $r = 1/2$.

We assess the closeness of the U(0,1) approximation graphically (Geweke and Amisano, 2010), by comparing the proportion of simulated values in each U(0,1) decile to the null value of 0.1. To aid the assessment of departures from this value we superimpose approximate pointwise 95% tolerance intervals based on the number of points within each decile having a Bin$(10000, 0.1)$ distribution, i.e. $0.1 \pm 1.96\,(0.1 \times 0.9/10000)^{1/2} = 0.1 \pm 0.006$. We use $p_u \in \{0.1, 0.5\}$, $\sigma = 1$ and $\xi \in \{0.1, -0.2\}$. These values for ξ are chosen based on estimates from real storm peak significant wave heights data, such as the Gulf of Mexico data (introduced in section 1.4), to reflect the kind of tail behaviours of immediate interest to us.

The plots in figures 10 to and 13 are based on simulated datasets of length $n_{\boldsymbol{x}} = 500$ and $n_y = 10$, i.e. 50 years of data with a mean of 10 observations per year, for future time horizons of lengths $N = 100, 1,000, 10,000$ and $100,000$ years. In all figures it is evident that the estimative approach based on the MLE produces too few values in deciles 2 to 9 and too many in deciles 1 and 10. Failing to take account of parameter uncertainty produces distributions that tend to be too concentrated, resulting in underprediction of large values of $x^{\dagger}$ (the percentage of $x^{\dagger}$ for which $\widehat{P}(M_N \leqslant x^{\dagger} \mid \boldsymbol{x}_{\text{sim}})$ is in the top decile is much greater than 10%) and overprediction of small values of $x^{\dagger}$ (the percentage of $x^{\dagger}$ for which $\widehat{P}(M_N \leqslant x^{\dagger} \mid \boldsymbol{x}_{\text{sim}})$ is in the bottom decile is much greater than 10%). These departures increase with N, i.e. with the degree of extrapolation. The Bayesian predictive approaches perform much better, with much smaller departures from the desired performance (shown in the "control" plots in figures 11 to 13).

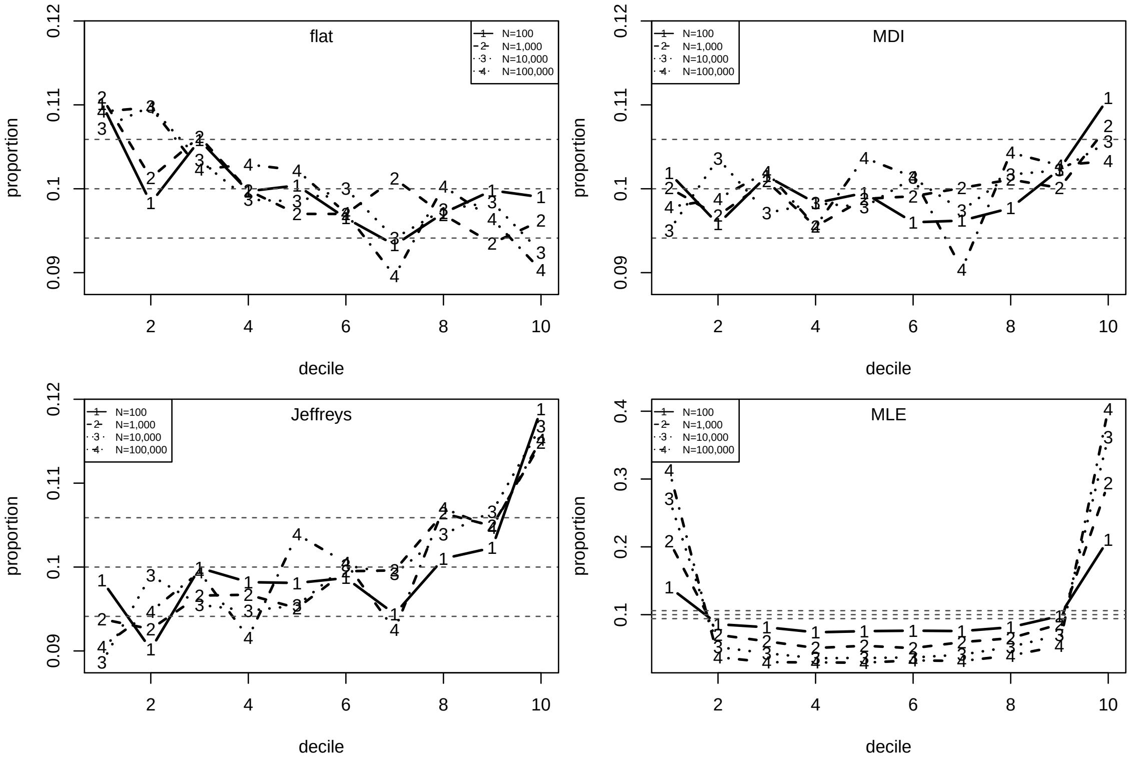

Figure 10: Proportions of simulated values of $\widehat{P}(M_N \leqslant x^\dagger \mid \boldsymbol{x}_{\text{sim}})$ falling in U(0,1) deciles for $\xi = 0.1$ and $p_u = 0.5$ and for $N = 100, 1,000, 10,000$ and $100,000$. 95% tolerance limits are superimposed.

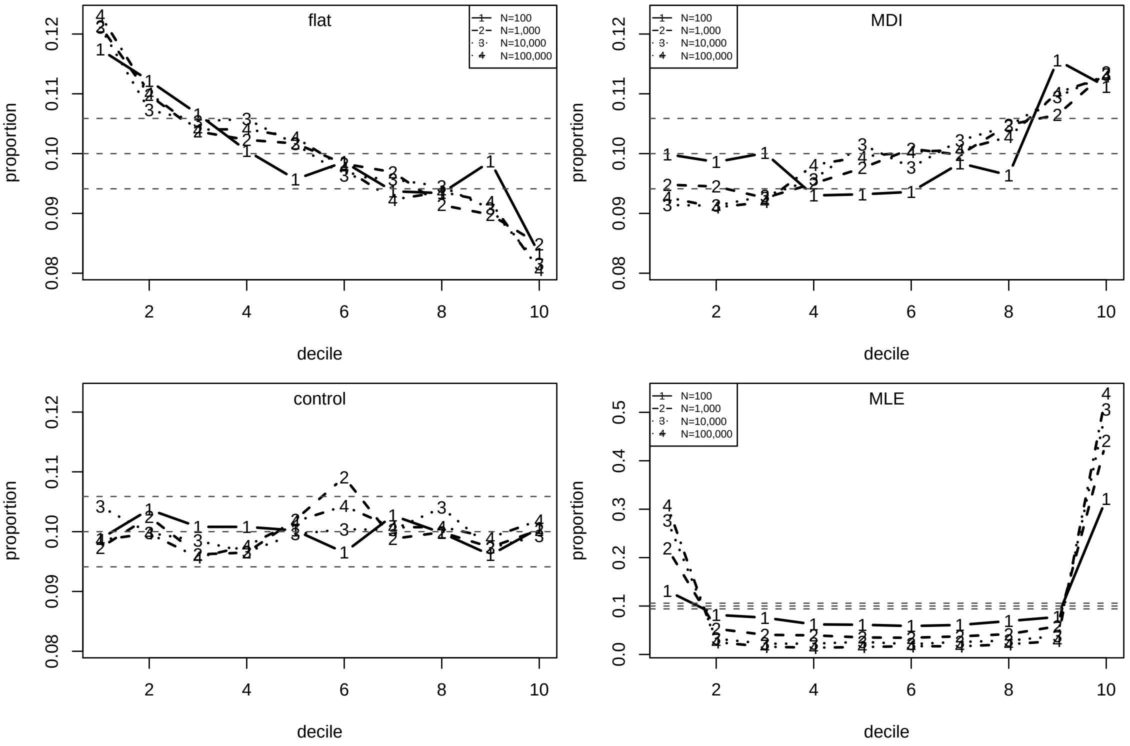

Figure 11: Proportions of simulated values of $\widehat{P}(M_N \leqslant x^{\dagger} \mid \boldsymbol{x}_{\text{sim}})$ falling in U(0,1) deciles for $\xi = 0.1$ and $p_u = 0.1$ and for $N = 100, 1,000, 10,000$ and $100,000$. 95% tolerance limits are superimposed. The control plot is based on random U(0,1) samples.

When the flat prior is used there is a general tendency to overpredict large values and small values. This is perhaps to be expected because the prior weight is constant for all ξ and a large sample of excesses may be required to downweight sufficiently the posterior density of very large values of ξ. The departures from the desired behaviour are more pronounced for $\xi = 0.1$ than $\xi = -0.2$ and for smaller p_u, i.e. a smaller expected number of threshold excesses. The MDI prior performs better, but shows a tendency towards the opposite departures, i.e. underprediction of large values and small values, in some cases. The underprediction of large values is evident for $p_u = 0.1$ (figures 11 and 13), but otherwise the MDI prior performs well. Results for the Jeffreys prior are shown only for $\xi = 0.1$ and $p_u = 0.5$ (figure 10). This is because, for $\xi = -0.2$ and/or $p_u = 0.1$, some of the simulated datasets result in a posterior that is unbounded as $\xi \downarrow -1/2$, preventing the use of the ratio-of-uniforms method (see section 2.4.2). In figure 10 the Jeffreys prior exhibits similar general behaviour to the MDI prior (i.e. underprediction of large values) but the departures are greater. This is to be expected from the shapes of these priors (see figure 9): for $\xi > -1/2$ the Jeffreys prior places greater weight on smaller values of ξ than the MDI prior.

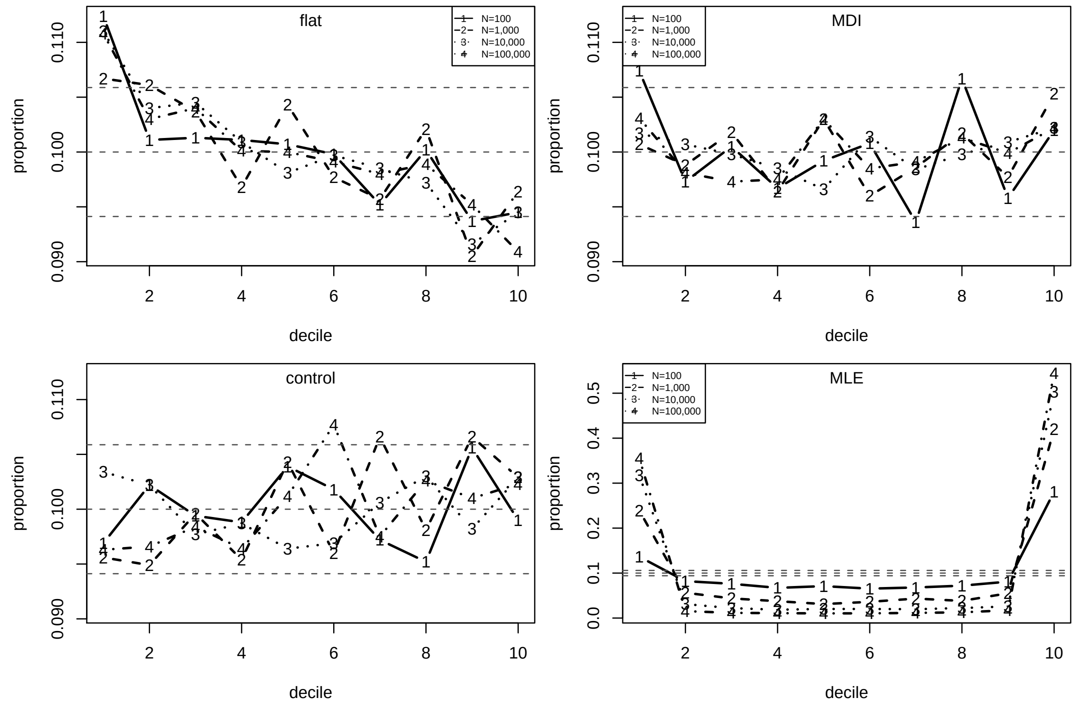

Figure 12: Proportions of simulated values of $\widehat{P}(M_N \leqslant x^\dagger \mid \boldsymbol{x}_{\mathrm{sim}})$ falling in U(0,1) deciles for $\xi = -0.2$ and $p_u = 0.5$ and for $N = 100, 1,000, 10,000$ and $100,000$. 95% tolerance limits are superimposed. The control plot is based on random U(0,1) samples.

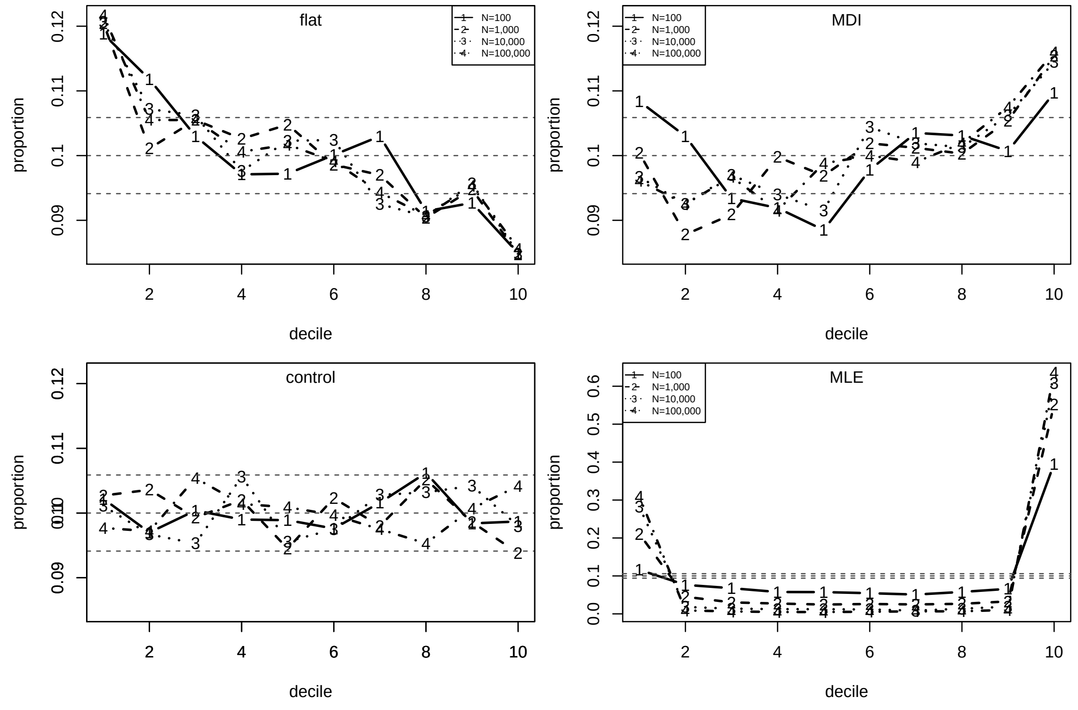

Figure 13: Proportions of simulated values of $\widehat{P}(M_N \leqslant x^\dagger \mid \boldsymbol{x}_{\mathrm{sim}})$ falling in U(0,1) deciles for $\xi = -0.2$ and $p_u = 0.1$ and for $N = 100, 1,000, 10,000$ and $100,000$. 95% tolerance limits are superimposed. The control plot is based on random U(0,1) samples.

These results suggest that, in terms of predicting M_N for large N, the truncated MDI prior performs better than the flat (Uniform) prior and the Jeffreys prior. However, a prior for ξ that is in some sense intermediate between the flat prior and the truncated MDI prior could possess better properties. To explore this further we consider the generalised truncated MDI(a) prior (2.24) for $0 < a \leqslant 1$. Letting $a \to 0$ produces a flat prior for ξ on the interval $[-1, \infty)$. In order to explore a range of values for a quickly, we reuse the posterior samples based on the priors $\pi_U(\sigma, \xi)$ and $\pi'_M(\sigma, \xi)$. We use an importance sampling ratio estimator to estimate $P(M_N \leqslant x^\dagger \mid \boldsymbol{x}_{\text{sim}})$ twice, once using $\pi_U(\sigma, \xi \mid \boldsymbol{x}_{\text{sim}})$ as the importance sampling density $q(\boldsymbol{\theta})$ and once using $\pi'_M(\sigma, \xi \mid \boldsymbol{x}_{\text{sim}})$. We calculate an overall estimate of $P(M_N \leqslant x^\dagger \mid \boldsymbol{x}_{\text{sim}})$ using a weighted mean of the two estimates, with weights equal to the reciprocal of the estimated variances of the estimators (Davison, 2003, page 603).

Figure 14 shows plots based on the truncated MDI(0.6) prior. This value of a has been selected based on plots for $a \in \{0.1, 0.2, \ldots, 0.9\}$. We make no claim that this is optimal, just that it is a reasonable compromise between the flat and the truncated MDI priors, providing relatively good predictive properties for the cases we have considered.

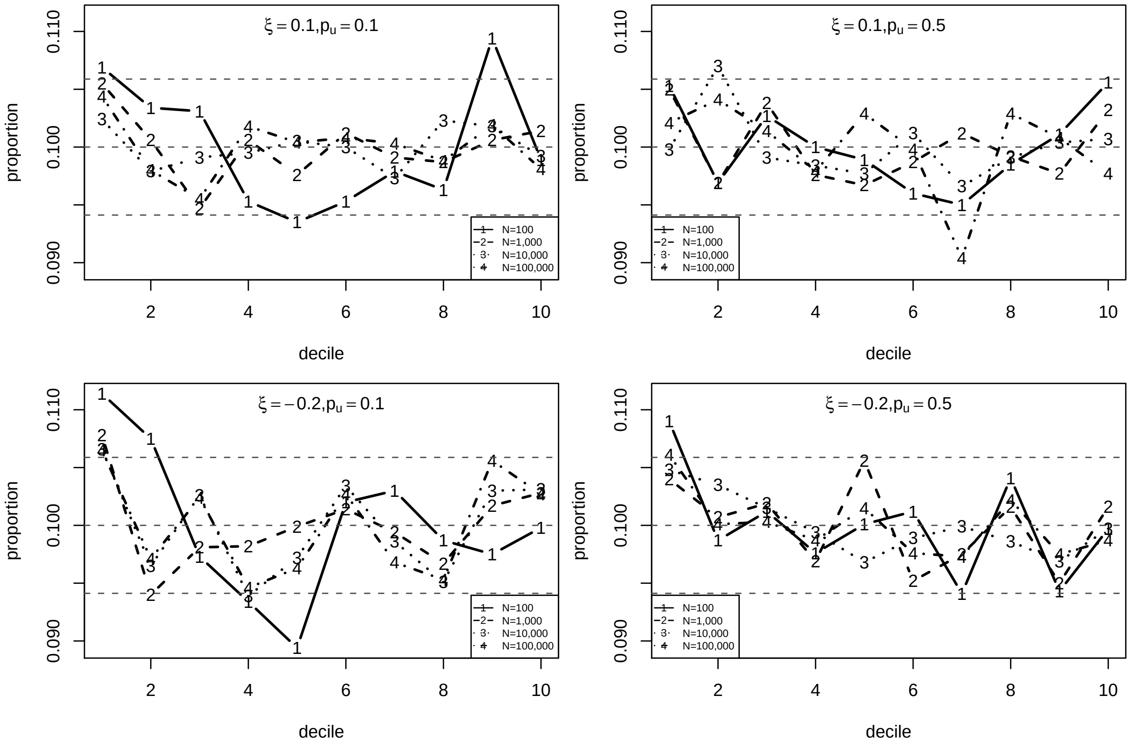

Figure 14: Proportions of simulated values of $\widehat{P}(M_N \leqslant x^\dagger \mid \boldsymbol{x}_{\text{sim}})$ falling in U(0,1) deciles for different combinations of ξ and p_u under the truncated MDI(0.6) prior. Separate lines are drawn for $N = 100, 1,000, 10,000$ and $100,000$. 95% tolerance limits are superimposed.

With our main research focus being the topic of threshold selection in extreme value modelling the results and conclusions presented in this chapter are of particular use for our work in chapter 3. Within a Bayesian framework, we employ a cross-validation methodology and use predictive inference to develop two novel threshold selection strategies. The important link between the two chapters lies on the fact that model parameter uncertainty is incorporated in the threshold selection by using a reference prior with good predictive performance, namely the generalised truncated MID(0.6) prior.

3 Threshold selection in the IID case

In this chapter we direct our focus solely on threshold based extreme value modelling, introduced in section 1.3. More precisely we investigate the important task of selecting a threshold on which subsequent inference is based through the GP distribution. We believe that this is an important step and that there is scope to improve upon existing threshold selection methods. We propose a novel method of selecting the threshold using cross-validation. We have considered an estimative approach using the MLE, but concentrate on a Bayesian predictive approach. The motivation behind our method is to address *directly* the issue of bias-variance trade-off (see sections 1.4.2 and the introduction of section 3.3). In section 3.4 we return to the Gulf of Mexico data that was introduced in section 1.4 to demonstrate the results of our method. Furthermore, in section 3.5 we use Bayesian model averaging approach to extend the idea of selecting a single “best” threshold to one that accounts for uncertainty in the choice.

As we are proposing a new graphical diagnostic tool for selecting the threshold, it is useful to briefly describe the classical graphical diagnostics tools that have been commonly used for extreme value threshold modelling. Our method aims to improve on these existing tools by removing as much as possible the arbitrariness that this choice involves and at the same time, account directly for the bias-variance trade-off underlying this choice. We work within the context of the GP distribution to model threshold excesses and the parameters of interest are the scale, σ and shape, ξ.

3.1 Classical methods

Scarrott and MacDonald (2012) offer an extensive review of threshold selection methods and argue that an advantage of the classical graphical diagnostic methods is the fact that they enable the practitioners to study the data prior to selecting the threshold. However, a serious concern when using these graphical methods is that once the threshold level is chosen, it is fixed and treated as known. In other words, uncertainty about the threshold is completely ignored in the subsequent inference. Alternatively, goodness-of-fit tests can be a crude method of making this decision which relies on statistical properties of estimators, however one needs to make an arbitrary choice of the significance level which in turn affects the power of these tests.

3.1.1 Mean residual life plot

The mean residual life plot was introduced by Davison and Smith (1990). For thresholds that are sufficiently high for the GP model for threshold excesses to apply, the mean threshold excess is linear in the threshold. The aim of a *mean residual life* plot is to aid the identification of a threshold u above which the graph appears to be linear, taking into account sampling variability summarised by confidence intervals. We use the Gulf of Mexico dataset to illustrate this graphical diagnostic tool in figure 15.

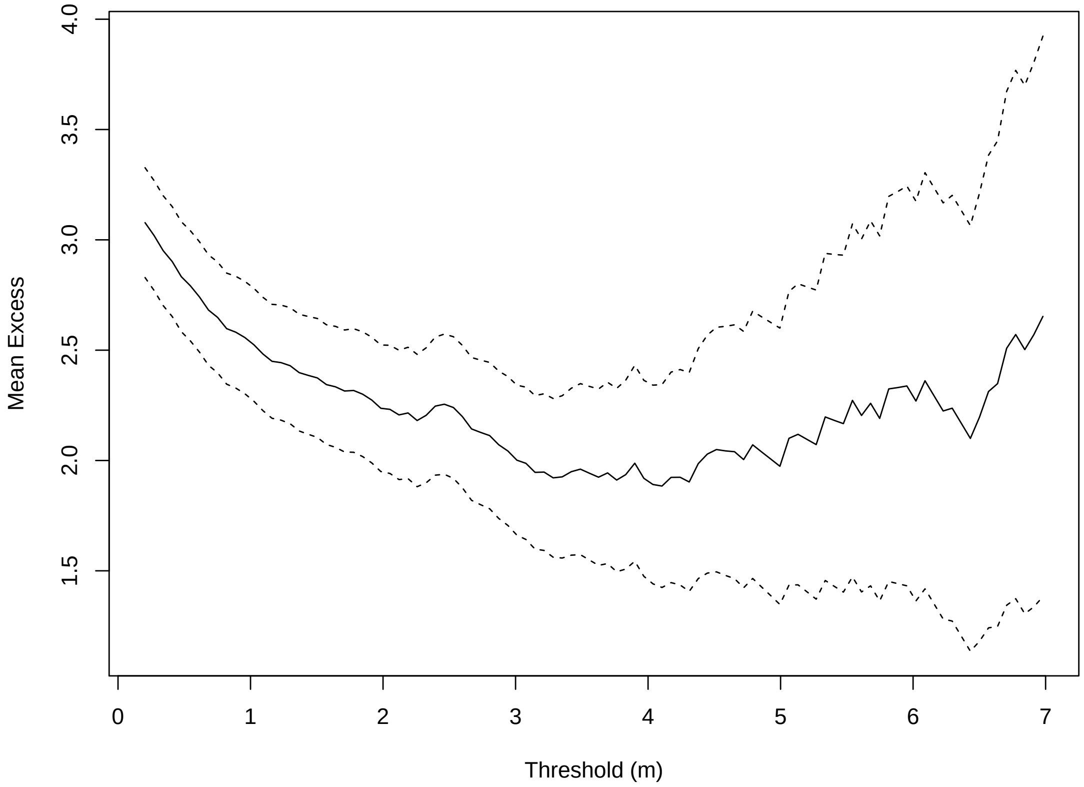

Figure 15: Mean residual life plot for Gulf of Mexico storm peak significant wave height. Dashed lines are the 95% confidence intervals.

In this example, we might judge that the solid line is approximately linear from a threshold of 4m upwards. However, this is a somewhat subjective choice and different viewers of the plot may choose quite different thresholds. This is typical. Also, this method does not generalise easily to more general modelling situations. More examples of the use of the mean residual life plot can be found in Coles (2001), Beirlant et al. (2004) among others.

3.1.2 Parameter stability plot

A more popular graphical method for determining the threshold is the parameter stability plot. This is done by fitting the GP distribution for a set of increasing thresholds with the objective of identifying the threshold at which the parameter estimates appear to stabilise. One can allow for sampling variability by superimposing confidence intervals for the estimates. The GP scale parameter is modified (to $\sigma^* = \sigma - \xi u$) so that both σ^* and ξ are constant, for thresholds over which a GP model applies. Figure 16 shows a parameter stability plot for the Gulf of Mexico dataset.

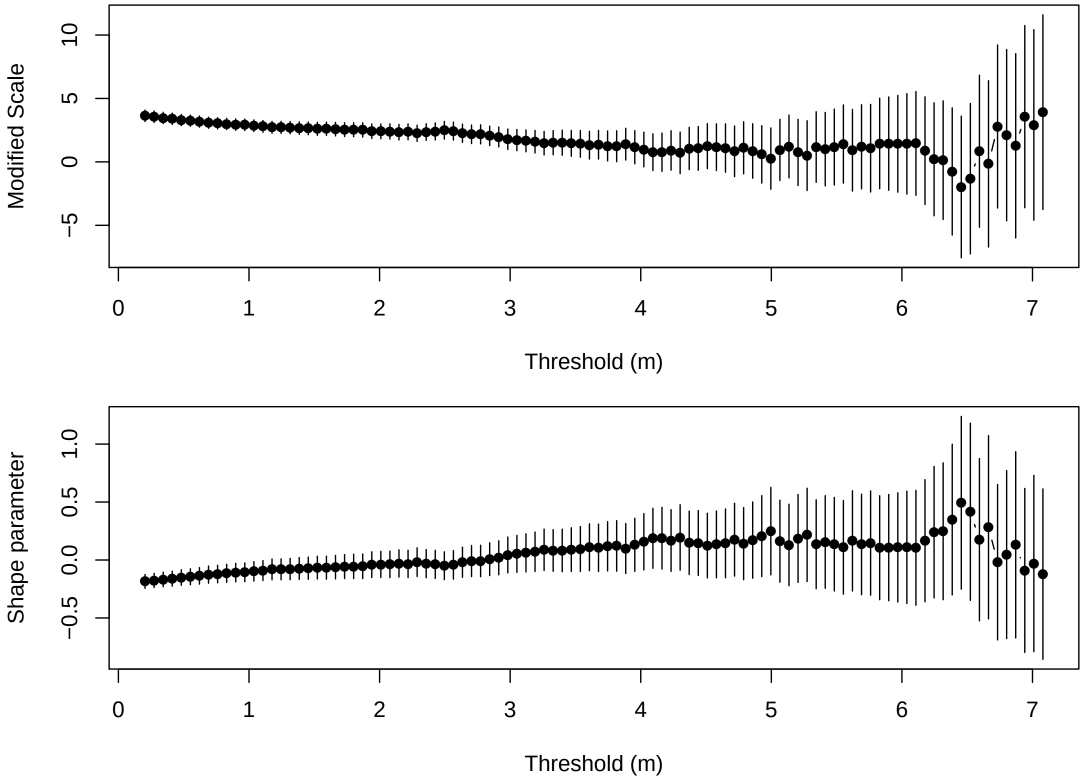

Figure 16: Parameter stability plots for Gulf of Mexico storm peak significant wave height. Vertical lines are the 95% confidence intervals.

Again, we might judge that the parameter estimates stabilise for thresholds close to 4m. This graphical method is generally preferred over the mean residual life plot as it demonstrates more clearly the parameter sensitivity with respect to the threshold. In addition, parameter uncertainty is directly illustrated in the plots and the method can generalise more easily than the mean residual life plot. However, this diagnostic tool shares the same drawback as the previous plot, that is the difficulty in agreeing the threshold level, since stability could be interpreted differently from one practitioner to another.

These two diagnostics plots are a quick and easy way of studying the data prior to any extreme value analysis and in some cases point to the direction of an appropriate threshold level. However, as Scarrott and MacDonald (2012) explain, it is very common when using these diagnostics that statisticians do not agree on a single value of the threshold, but that a number of thresholds can be identified as appropriate. This is the case, for example, in the analysis of river flow rates from the river Nidd in Davison and Smith (1990). A variety of threshold levels could be used which ultimately lead to very different inferences on future extreme events.

Adding to the issue of contradicting conclusions, another drawback of these graphical diagnostic tools is the fact that the subsequent extreme value analysis does not account for the fact that the threshold was subjectively chosen, i.e. threshold uncertainty is completely ignored. The importance of getting the "right" choice for the level of the threshold relates to the issue of a trade-off between bias and variance. Effectively a threshold that is set too low might violate the asymptotic arguments for the GP(σ, ξ) model (see theorem 3) and lead to bias, whereas a threshold that is set too high will provide a small number of excesses leading to high variance in the estimators of σ and ξ. This is something that we aim to address directly using our proposed method described in 3.2.

3.1.3 Other threshold selection methods

Scarrott and MacDonald (2012) provides a thorough review of the various threshold selection methods. As we have seen, one category of method assesses stability of model parameter estimates (particularly the GP shape parameter ξ) with threshold. Estimators of ξ based on functions of order statistics can be used, see for example, Drees et al. (2000). More recently motivated by subasymptotic (or penultimate) extreme value theory Wadsworth and Tawn (2012) formalise the parameter stability plot method and use a likelihood-based test to assess whether the shape parameter remains constant above a certain threshold. Northrop and Coleman (2014) develop this approach further providing a test with improved computational performance as compared to Wadsworth and Tawn (2012). Other categories of method are goodness-of-fit tests (Davison and Smith, 1990, Dupuis, 1998); approaches that minimize the asymptotic mean-squared error of estimators of ξ or of extreme quantiles, under particular assumptions about the form of the upper tail of H (Hall and Welsh, 1985, Hall, 1990, Ferreira et al., 2003, Beirlant et al., 2004); specifying a model below the threshold (Wong and Li, 2010, MacDonald et al., 2011, Wadsworth and Tawn, 2012). In the latter category, the threshold above which the GP model is assumed to hold is treated as a model parameter and *threshold uncertainty* is incorporated by averaging inferences over a posterior distribution of model parameters. In contrast, our earlier discussions have been on a *single threshold approach* in which threshold level is viewed as a tuning parameter, whose value is selected prior to the main analysis and is treated as fixed and known when subsequent inferences are made.

In the first instance our aim is to develop a likelihood-based method for selecting a single threshold that addresses the bias-variance trade-off based on the main purpose of the modelling, i.e. prediction of extremal behaviour. We make use of a data-driven method commonly used for the assessment of predictive performance: cross-validation. Later, in section 3.5, we consider how this method could be adapted to take into account uncertainty in the choice of a single threshold.

3.2 Cross-validation (CV)

In the early 1930's researchers noticed (Larson, 1931) that when evaluating the statistical performance of an algorithm using the same dataset that was initially used for training it returned overoptimistic results. In order to deal with this issue, the procedure of cross-validation came about from the works of Mosteller and Tukey (1968), Stone (1974), Geisser (1975). We continue this section by introducing the methodology behind the classical approaches in applying cross-validation as it forms a key role in our proposed method for selecting the threshold. Arlot and Celisse (2010) is a useful source for a more in-depth analysis on cross-validation procedures.

Methodology

Suppose that we have a vector of raw (unthresholded) data $\boldsymbol{x} = (x_1, \ldots, x_m)$ assumed to have been sampled randomly from a common distribution. The first step involves splitting the dataset into two sub-samples

1. the training sample $\boldsymbol{x_t}$, of size n_t and
2. the validation sample $\boldsymbol{x_v}$ of size n_v,

where $m = n_t + n_v$.

The next step uses the training sample to train a prediction algorithm. Then predictive ability of this algorithm is evaluated based on the errors that it makes when predicting the validation sample. In the context of evaluating the performance of a statistical model, the validation sample is considered as “new” data from the underlying model and *validation* process takes place by assessing the model's prediction error. Repeating this process for several data splits and averaging the prediction error over the number of splits results in the technique known as *cross-validation.*

3.2.1 Leave-one-out (LOO)

The leave-one-out cross-validation approach deals with validation samples consisting of a single observation. In other words, $n_v = 1$ and therefore the training sample is of size $n_t = m - 1$. LOO cross-validation is carried out by subsequently leaving out each observation from the dataset. Thus, for a sample of size m this process is repeated m times and the prediction error can be found by

$$\widehat{\mathcal{L}}^{LOO} = \frac{1}{m}\sum_{i=1}^{m} \mathcal{E}_i \quad i = 1, \dots, m, \tag{3.1}$$

where $\mathcal{E}_i$ represents the prediction error when observation i is treated as the validation sample. Figure 17 shows the leave-one-out cross validation approach for the Gulf of Mexico dataset where observation x_{45} (shaded in blue) was set as the validation sample.

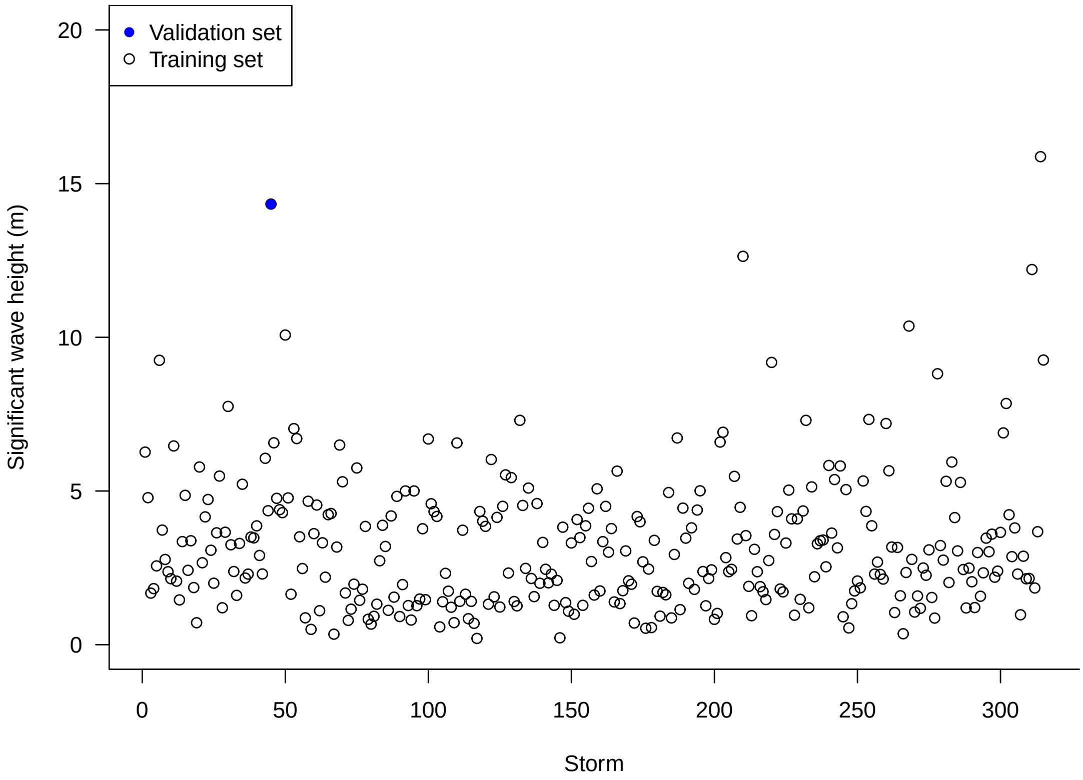

Figure 17: Leave-one-out cross-validation for Gulf of Mexico storm peak significant wave height. ● validation sample ; ○ training sample.

Once observation x_{45} is removed from the dataset, the GP distribution is fitted to excesses of a threshold u in the training sample and the model parameter estimates for σ and ξ can be calculated using, for example, MLE. This process is shown below in figure 18.

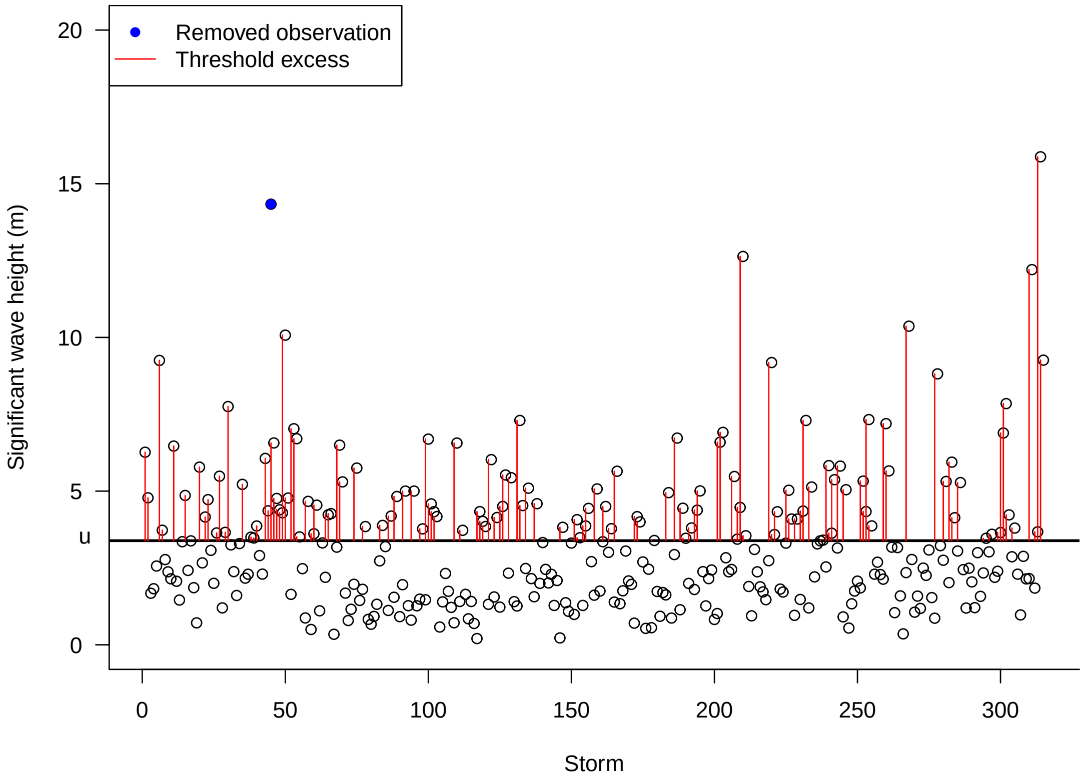

Figure 18: Leave-one-out cross-validation for Gulf of Mexico storm peak significant wave height. ● validation sample ; ○ training sample ; — threshold excesses.

3.2.2 K-fold (KF)

A common extension to the LOO cross-validation is the approach known as *K-fold* cross-validation. This method was first introduced by Geisser (1975) and involves the splitting of the dataset into K approximately equal and mutually exclusive sub-samples or folds. Typical choices of K are 5 or 10 and it is clear that when $K = m$, the method is equivalent to leave-one-out cross-validation. For this approach, the validation sample consists of one of the sub-samples of size $n_v = m/K$ and the remaining dataset forms the training sample. K-fold cross validation takes place by repeating the validation process K times. Just as in LOO cross-validation, the performance of a statistical model is assessed by measuring it's prediction error and is found by

$$\widehat{\mathcal{L}}^{KF} = \frac{1}{K}\sum_{i=1}^{K}\mathcal{E}_i, \quad i = 1, \ldots, K, \tag{3.2}$$

where $\mathcal{E}_i$ represents the prediction error when fold i is treated as the validation sample. Figure 19 shows the K-fold cross validation approach for the Gulf of Mexico dataset for $K = 10$ and where the first fold $[x_1, x_{31}]$ was set as the validation sample, although, in the current context where $x_1, \ldots, x_m$ are i.i.d. there is no requirement for the validation data to consist of contiguous observations.

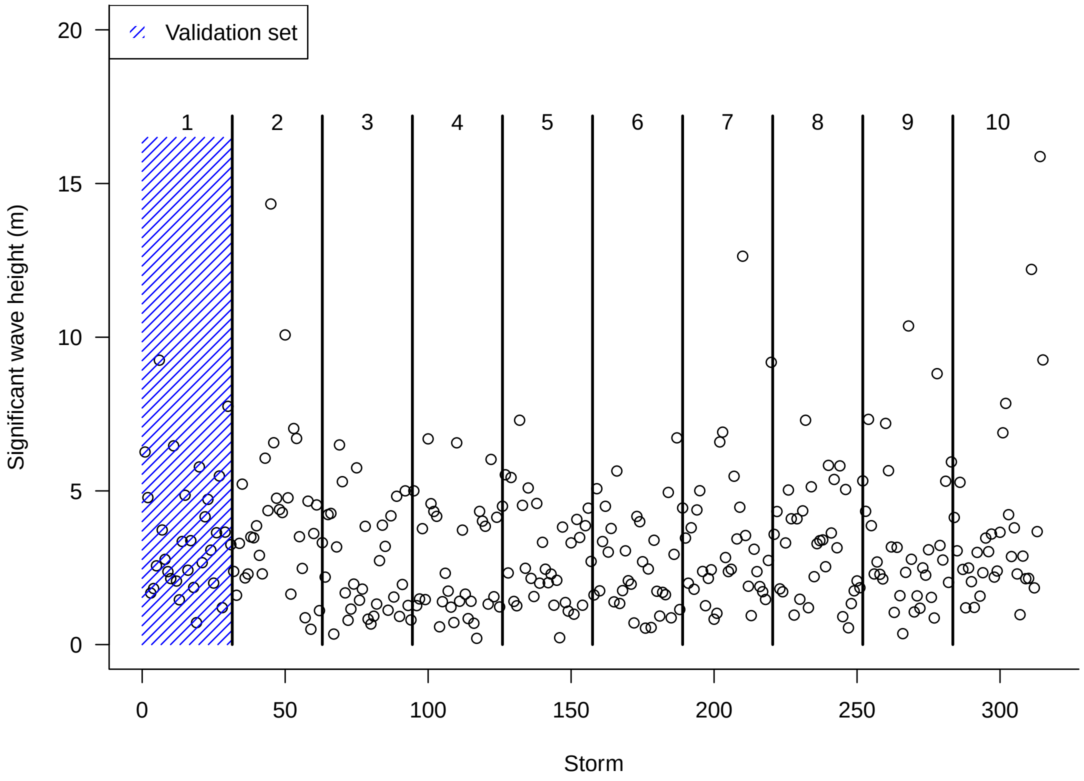

Figure 19: K-fold cross-validation for Gulf of Mexico storm peak significant wave height. Blue shaded area forms the validation sample ; Unshaded area forms the training sample.

Just as in the LOO cross-validation method, once the first fold is removed from the dataset, the GP distribution is fitted to threshold excesses in the training sample and the model parameter estimates for σ and ξ can be calculated using for example MLE.

3.2.3 Repeated random sub-sampling (RRSS)

The repeated random sub-sampling cross-validation approach (Picard and Cook, 1984) is somewhat similar to that of K-fold cross-validation. The limitation of the K-fold method is the fact that the prediction error is measured only K times due to the uniquely defined folds. This however, can be resolved by the RRSS approach which allows for the prediction error to be measured more than K times. This is because we can choose any random sample of size n_t to be the training sample (if $n_t = m - 1$ then we have LOO, if $n_t = m - m/K$ then we have KF). For a validation sample of size $n_v = m/K$, this method allows us to repeat the validation process up to $\binom{m}{n_v}$ times. Let K_{RRSS} be the number of times we repeat the RRSS procedure, then the statistical model's prediction error can be found by

$$\widehat{\mathcal{L}}^{RRSS} = \frac{1}{K_{RRSS}} \sum_{i=1}^{K_{RRSS}} \mathcal{E}_i, \quad i = 1, \ldots, K_{RRSS}, \tag{3.3}$$

where $\mathcal{E}_i$ represents the prediction error when the sub-sample i is treated as the validation sample. Figure 20 shows one example of an RRSS split of the data for the Gulf of Mexico dataset for a validation sample of size $n_v = 31$.

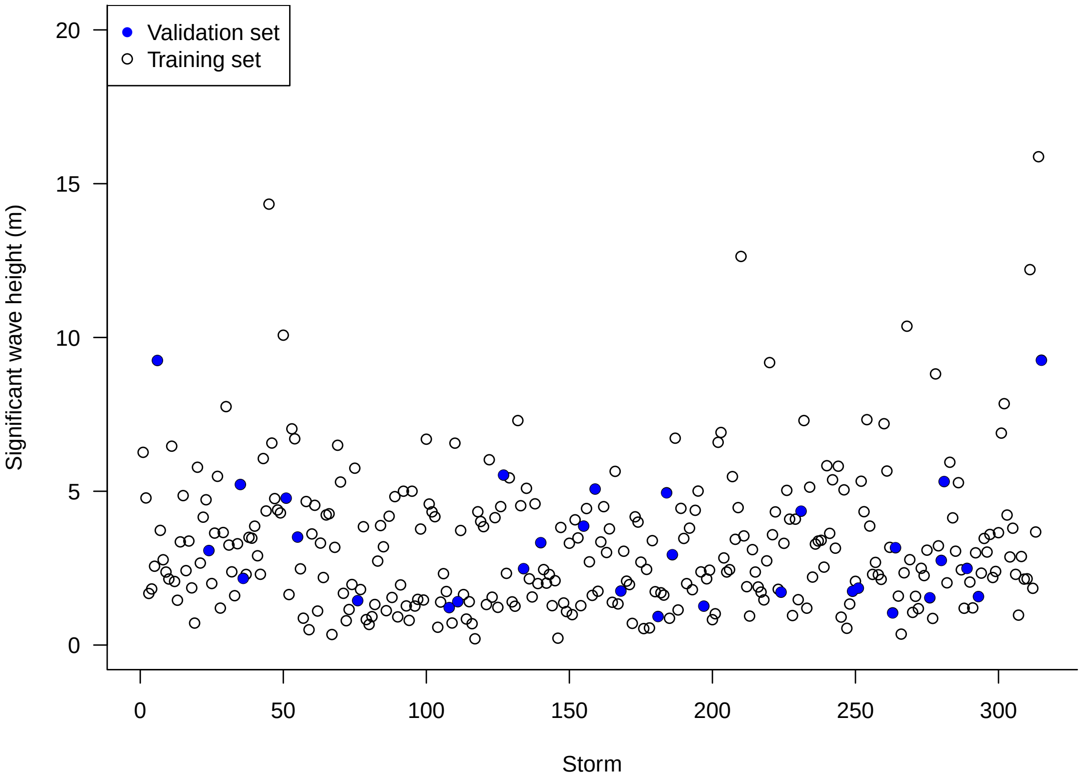

Figure 20: Repeated random sub-sampling cross-validation for Gulf of Mexico storm peak significant wave height. ● validation sample ; ○ training sample.

Once the random sub-sample is removed from the dataset, the GP distribution is fitted to the threshold excesses in the training sample and the model parameter estimates for σ and ξ can be calculated using for example MLE.

3.2.4 Choice of CV method

There is no general theory to indicate which form of CV is best: this will depend on the specific application (Arlot and Celisse, 2010). Here we use LOO CV. This method is *exhaustive*, i.e. all possible training sets of size $m-1$ are used and has the property that each observation is used as a validation observation exactly once. LOO CV is often avoided because it can be more computationally-intensive than other methods. In section 3.3.2 we reduce the computational intensity using an approximation.

For our purposes LOO CV is attractive because it will (a) provide, for each candidate threshold, an estimate of a quantity by which the discrepancy between model and (extreme) data can be measured (see section 3.3.3); and (b) prove useful in dealing with threshold uncertainty by providing weights for a weighted average of inferences from different thresholds (see section 3.5).

3.3 Threshold selection using cross-validation

In this section we direct our focus to threshold based extreme value modelling for the Bin-GP model and utilise the LOO cross-validation technique in order to address the issue of selecting an appropriate threshold. Our aim is to construct a somewhat automatic threshold selection procedure, which improves the classical methods described earlier by lessening the obscurity of where exactly a suitable threshold should be selected.

Bias-Variance trade-off

A well known issue within the threshold based extreme value modelling analysis is that of a bias-variance trade-off. On one hand, a threshold that is set too low might violate the asymptotic arguments of the GP model leading to bias. On the other hand, a threshold that is set too high will provide a small number of threshold excesses leading to high variance in the estimators of the model parameters.

We explained earlier that the classical threshold selection methods do not directly address this issue. For example, using the parameter stability plots (see figure 16), one aims to choose the lowest threshold for which the parameter estimates appear to have stabilised and goodness-of-fit approaches seek the lowest threshold for which the GP model is not rejected by an hypothesis test. Both approaches could be characterised as seeking a threshold that minimises the variance part of the trade-off subject to some subjective and inexplicit constraint on the suitability of the GP model. However, one is unaware whether a lower threshold would be more appropriate since it could be possible that the benefit of reducing imprecision outweighs the increase in bias. Faced with this reality, we are proposing to incorporate a cross-validation step in the threshold selection process. This allows us to directly address the issue of bias-variance trade-off by comparing the predictive performance of a range of thresholds.

3.3.1 Cross-validation predictive performance

Suppose that we have a vector of raw (unthresholded) data $\boldsymbol{x} = (x_1, \ldots, x_m)$ from a sequence of independent and identically distributed random variables $X_1, \ldots, X_m$ with parametric density function $F = f(x_i; \boldsymbol{\theta})$ having parameter vector $\boldsymbol{\theta}$. Furthermore, without loss of generality we assume that $x_1 < \cdots < x_m$. A Bin-GP(p_u, σ_u, ξ) model is used at threshold u, where $p_u = P(X > u)$ and (σ_u, ξ) are the parameters of the GP(σ_u, ξ) distribution that models excesses of u. It is clear that this model is defined by the choice of the threshold and we reiterate the fact that it is of great importance to choose an appropriate threshold for this analysis.

Our aim is to quantify the ability of Bin-GP inferences based on threshold u to predict (out-of-sample) at extreme levels. Let us define u as the *training threshold* and introduce a *validation threshold*, v, for $v \geqslant u$, where high values of v are of greatest interest. At threshold u we have a Bin-GP(p_u, σ_u, ξ) model and if $1 + \xi(v - u)/\sigma_u > 0$ then a Bin-GP(p_v, σ_v, ξ) model is implied at threshold v, where

(a) $\sigma_v = \sigma_u + \xi(v - u)$ and

(b) $p_v = P(X > v) = (1 + \xi(v - u)/\sigma_u)^{-1/\xi}\, p_u$.

Otherwise, i.e. if $1 + \xi(v - u)/\sigma_u \leqslant 0$, then $p_v = 0$. For a particular value of v we wish to compare the predictive ability of the implied Bin-GP(p_v, σ_v, ξ) model (which depends on the choice of u) across a range of values of u.

We employ a leave-one-out cross-validation scheme in which $\boldsymbol{x}_{(r)} = \{x_i, i \neq r\}$ forms the training sample and x_r the validation sample. We use, as a measure of predictive performance at validation threshold v when using training threshold u,

$$\widehat{T}_v(u) = \sum_{r=1}^{m} \log \widehat{f}_v(x_r \mid \boldsymbol{x}_{(r)}, u) \tag{3.4}$$

where $\widehat{f}_v(x_r \mid \boldsymbol{x}_{(r)}, u)$ is an estimate of the density of x_r at validation threshold v based on a training threshold u and on training data $\boldsymbol{x}_{(r)}$. Suppose that the parameters $\boldsymbol{\theta} = (p_u, \sigma_u, \xi)$ are known, and the $\{x_i\}$ are conditionally independent given $\boldsymbol{\theta}$. If $p_v > 0$ then

$$f_v(x_r \mid \boldsymbol{x}_{(r)}, u) = (1 - p_v)^{I(x_r \leqslant v)} \left\{p_v g(x_r - v; \sigma_v, \xi)\right\}^{I(x_r > v)}, \quad r = 1, \ldots, m, \tag{3.5}$$

where $g(\cdot)$ is the density of a GP distribution, i.e.

$$g(x_r - v; \sigma_v, \xi) = \frac{1}{\sigma_v}\left[1 + \xi \frac{(x_r - v)}{\sigma_v}\right]_+^{-(1+1/\xi)}. \tag{3.6}$$

Otherwise, i.e. if $p_v = 0$ then

$$f_v(x_r \mid \boldsymbol{x}_{(r)}, u) \;=\; (1 - p_v)^{I(x_r \leqslant v)} p_v^{I(x_r > v)}, \tag{3.7}$$

where $I(\cdot)$ is the indicator function such that $I(x) = 1$ if x is true and $I(x) = 0$ otherwise.

3.3.2 Estimation of cross-validation densities

One way to estimate $f_v(x_r \mid \boldsymbol{x}_{(r)}, u)$ is using the cross-validation *estimative* density $f_v(x_r \mid \widehat{\boldsymbol{\theta}}, u)$, where $\widehat{\boldsymbol{\theta}} = (\widehat{p}_u, \widehat{\sigma}_u, \widehat{\xi})$ is an estimate, perhaps the MLE, of $\boldsymbol{\theta}$ based on $\boldsymbol{x}_{(r)}$ and $\widehat{\sigma}_v = \widehat{\sigma}_u + \widehat{\xi}(v - u)$ and $\widehat{p}_v = (1 + \widehat{\xi}(v - u)/\widehat{\sigma}_u)^{-1/\widehat{\xi}}\widehat{p}_u$. However, this takes no account of the uncertainty associated with estimating $\boldsymbol{\theta}$ using $\boldsymbol{x}_{(r)}$. This is undesirable, because the size of this uncertainty will vary greatly across different thresholds: uncertainty in GP parameters will tend to increase as the threshold is raised. An additional concern is that a point estimate of GP model parameters can correspond to a zero likelihood for a validation observation. This happens when a point estimate of ξ is negative and the validation observation is greater than the estimated upper endpoint $u - \widehat{\sigma}_u/\widehat{\xi}$. In this circumstance $\widehat{T}_v(u)$ suggests that u is 'infinitely bad' and the estimative approach would effectively rule out the threshold u. Accounting for parameter uncertainty alleviates this problem by giving weight to parameter values other than a particular point estimate. One could approximate the effects of parameter uncertainty using large sample estimation theory or bootstrapping (Young and Smith, 2005, chapter 10). However, large sample results may provide poor approximations for high thresholds (small numbers of excesses) and the GP observed information is known to have poor finite-sample properties (Süveges and Davison, 2010). Bootstrapping, of ML or PWM estimates, increases computation time further and is subject to the regularity conditions mentioned in chapter 1.

For these reasons we prefer a predictive approach, implemented in a Bayesian setting, to incorporate parameter uncertainty. Inferences are averaged over a posterior distribution $\pi(\boldsymbol{\theta} \mid \boldsymbol{x}_{(r)})$ of parameters, to reflect differing parameter uncertainties across thresholds. Specifically, the cross-validation *predictive* densities at validation threshold v, based on a training threshold u, are given by

$$f_v(x_r \mid \boldsymbol{x}_{(r)}, u) = \int f_v(x_r \mid \boldsymbol{\theta}, \boldsymbol{x}_{(r)})\, \pi_u(\boldsymbol{\theta} \mid \boldsymbol{x}_{(r)})\, \mathrm{d}\boldsymbol{\theta}, \quad r = 1, \ldots, m, \tag{3.8}$$

where, assuming that the $\{x_i\}$ are conditionally independent given $\boldsymbol{\theta}$,

$$f_v(x_r \mid \boldsymbol{\theta}, \boldsymbol{x}_{(r)}) = f_v(x_r \mid \boldsymbol{\theta}) = (1 - p_v)^{I(x_r \leqslant v)} \left\{ p_v g(x_r - v; [\sigma_v]_+, \xi) \right\}^{I(x_r > v)}. \tag{3.9}$$

A prior distribution $\pi(\boldsymbol{\theta})$ is required for $\boldsymbol{\theta}$. We will use a prior for the GP parameters based on the results of the simulation study in section 2.6.

Let the parameter vector be denoted as $\boldsymbol{\theta} = (p_u, \sigma_u, \xi)$ and $\pi(\boldsymbol{\theta})$ a prior density for $\boldsymbol{\theta}$. Let $\boldsymbol{x}^s$ denote a subset of $\boldsymbol{x}$. The posterior density is $\pi_u(\boldsymbol{\theta} \mid \boldsymbol{x}^s) \propto L(\boldsymbol{\theta}; \boldsymbol{x}^s, u)\pi(\boldsymbol{\theta})$, where

$$L(\boldsymbol{\theta}; \boldsymbol{x}^s, u) = \prod_{i : x_i \in \boldsymbol{x}^s} f_u(x_i \mid \boldsymbol{\theta}), \tag{3.10}$$

and where

$$f_u(x_i \mid \boldsymbol{\theta}) = (1 - p_u)^{I(x_i \leqslant u)} \left\{ p_u g(x_i - u; \sigma_u, \xi) \right\}^{I(x_i > u)}, \tag{3.11}$$

and $g(x_i - u; \sigma_u, \xi)$ is defined in (3.6).

Suppose that we have a sample $\boldsymbol{\theta}_j^{(r)}, j = 1, \ldots, n_{\boldsymbol{\theta}}$ from the posterior $\pi_u(\boldsymbol{\theta} \mid \boldsymbol{x}_{(r)})$. Then a Monte Carlo estimator of $f_v(x_r \mid \boldsymbol{x}_{(r)}, u)$ based on (3.8) is given by

$$\tilde{f}_v(x_r \mid \boldsymbol{x}_{(r)}, u) = \frac{1}{n_{\boldsymbol{\theta}}} \sum_{j=1}^{n_{\boldsymbol{\theta}}} f_v(x_r \mid \boldsymbol{\theta}_j^{(r)}, \boldsymbol{x}_{(r)}). \tag{3.12}$$

Evaluation of estimator (3.12), for $r = 1, \ldots, m$, is computationally intensive because it involves generating samples from m different posterior distributions. To make this approach practical we consider an importance sampling estimator (Gelfand, 1996, Gelfand and Dey, 1994) that enables estimation of $f_v(x_r \mid \boldsymbol{x}_{(r)}, u)$, for $r = 1, \ldots, m-1$, using a single posterior sample only. We rewrite (3.8) as

$$f_v(x_r \mid \boldsymbol{x}_{(r)}, u) = \int f_v(x_r \mid \boldsymbol{\theta}, \boldsymbol{x}_{(r)})\, q_r(\boldsymbol{\theta})\, h(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}, \quad r = 1, \ldots, m, \tag{3.13}$$

where $q_r(\boldsymbol{\theta}) = \pi_u(\boldsymbol{\theta} \mid \boldsymbol{x}_{(r)})/h(\boldsymbol{\theta})$ and $h(\boldsymbol{\theta})$ is an importance sampling density whose support must include that of $\pi_u(\boldsymbol{\theta} \mid \boldsymbol{x}_{(r)})$. In the current context a common choice is $\pi_u(\boldsymbol{\theta} \mid \boldsymbol{x})$ (Gelfand and Dey, 1994, page 511), i.e. the posterior given the entire dataset. However, the support of $\pi_u(\boldsymbol{\theta} \mid \boldsymbol{x})$: $\xi > -\sigma_u/(x_m - u)$, does not contain that of $\pi_u(\boldsymbol{\theta} \mid \boldsymbol{x}_{(m)})$, i.e. $\xi > -\sigma_u/(x_{m-1} - u)$. Therefore we use $h(\boldsymbol{\theta}) = \pi_u(\boldsymbol{\theta} \mid \boldsymbol{x})$ for $r \neq m$, and (3.12) for $r = m$, requiring only two posterior samples.

Suppose that we have a sample $\theta_j, j = 1, \ldots, n_{\boldsymbol{\theta}}$ from the posterior $\pi_u(\boldsymbol{\theta} \mid \boldsymbol{x})$. For $r = 1, \ldots, m-1$ we use the importance sampling ratio estimator

$$\widehat{f}_v(x_r \mid \boldsymbol{x}_{(r)}, u) = \frac{\sum_{j=1}^{n_{\boldsymbol{\theta}}} f_v(x_r \mid \boldsymbol{\theta}_j)\, q_r(\boldsymbol{\theta}_j)}{\sum_{j=1}^{n_{\boldsymbol{\theta}}} q_r(\boldsymbol{\theta}_j)} = \frac{\sum_{j=1}^{n_{\boldsymbol{\theta}}} f_v(x_r \mid \boldsymbol{\theta}_j)/f_u(x_r \mid \boldsymbol{\theta}_j)}{\sum_{j=1}^{n_{\boldsymbol{\theta}}} 1/f_u(x_r \mid \boldsymbol{\theta}_j)}, \tag{3.14}$$

where $q_r(\boldsymbol{\theta}) = \pi_u(\boldsymbol{\theta} \mid \boldsymbol{x}_{(r)})/\pi_u(\boldsymbol{\theta} \mid \boldsymbol{x}) \propto 1/f_u(x_r \mid \boldsymbol{\theta})$.

If we also have a sample $\boldsymbol{\theta}_j^{(m)}, j = 1, \ldots, n_{\boldsymbol{\theta}}$ from the posterior $\pi_u(\boldsymbol{\theta} \mid \boldsymbol{x}_{(m)})$ then

$$\widehat{f}_v(x_m \mid \boldsymbol{x}_{(m)}, u) = \frac{1}{n_{\boldsymbol{\theta}}} \sum_{j=1}^{n_{\boldsymbol{\theta}}} f_v(x_m \mid \boldsymbol{\theta}_j^{(m)}). \tag{3.15}$$

We can now use (3.14) and (3.15) in (3.4) to measure the predictive performance at validation threshold v when using a training threshold u.

3.3.3 Comparing training thresholds

Suppose that we consider k training thresholds $u_1 < \cdots < u_k$, resulting in a set of estimates $\widehat{T}_v(u_1), \ldots, \widehat{T}_v(u_k)$ of predictive performance, and that we wish to select one of these thresholds. The obvious choice is $u^* = \arg\max_u \widehat{T}_v(u)$. Up to an additive constant, $\widehat{T}_v(u)$ provides an estimate of the negated Kullback-Leibler divergence between the Bin-GP model at validation threshold v and the true density (see, for example, Silverman (1986, page 53)). Therefore, u^* has the property that, of the thresholds considered, it has the smallest estimated Kullback-Leibler divergence.

Implementation of this approach requires some subjective inputs: the choice of training thresholds $(u_1, \ldots, u_k)$, validation threshold v and priors for the Binomial-GP parameters; and therefore cannot be fully automated. However, as we discuss below, the extent of this subjectivity can be reduced.

Choice of training thresholds

Choosing $u_1, \ldots, u_k$ is the starting point for many threshold selection methods. These thresholds should span the range of thresholds that are entertained as plausible and u_k should not be so high that it provides little information from which to make inferences about the GP parameters. For example, in a parameter stability plot using a threshold that is too high can cause numerical problems in calculating estimates of parameters and their standard errors. We return to this issue below when discussing choice of prior distributions. No definitive rule exists for limiting u_k but Jonathan and Ewans (2013) suggest that there should be no fewer than 50 threshold excesses. An initial graphical diagnostic such as a parameter stability plot can be useful and can give an indication of a range of thresholds over which bias-variance trade-off is occurring.

Choice of validation threshold

If we are to compare the performances of training thresholds $u_1, \ldots, u_k$, we need $v \geqslant u_k$. The larger v is the fewer excesses of v there are and the smaller the amount of information available from data thresholded at v. This is a heuristic argument for taking $v = u_k$, which is strengthened by the following argument.

Consider two validation thresholds, $v_1 = u_k$ and $v_2 > u_k$. If we validate the inferences from $u_1, \ldots, u_k$ at $v_1 = u_k$, then, for each training threshold, *all* the exceedances of v_1 enter into the GP part of the validation log-likelihood (3.9), and the *sizes* of the threshold exceedances are used. If we validate at v_2, which is greater than u_k, then this is no longer true. It is only exceedances of v_2 for which the size of the exceedance enters the GP part of (3.9), and information from observations between u_k and v_2 enter the validation log-likelihood only in the form that "these observations were between u_k and v_2". Therefore, in moving from v_1 to v_2 we have lost validation information. Moreover, the contributions to (3.9) from observations that lie above v_2 is the same whether we validate at v_1 or v_2. The prediction of an observation $x > v_2$ is unaffected by whether v_1 or v_2 is used and $p_{v_1}\, g(x - v_1; \sigma_{v_1}; \xi) = p_{v_2}\, g(x - v_2; \sigma_{v_2}; \xi)$ in (3.9). Therefore, in moving from v_1 to v_2 we have reduced the validation information from observations that lie between v_1 and v_2 and gained nothing in return, i.e. choosing $v = u_k$ provides the greatest amount of validation information and nothing is gained by choosing $v > u_k$.

If we choose a set of training thresholds $u_1, \ldots, u_k$ as being plausible a priori then we should validate using validation threshold $v = u_k$. If we want to examine sensitivity to validation threshold we could also validate at, say, u_{k-1}, although we should appreciate that this is effectively changing our set of training thresholds to $u_1, \ldots, u_{k-1}$.

Choice of prior distributions

In chapter 2 we compared the predictive performance of three reference priors (Jeffreys, MDI and Uniform) for GP parameters. We found that a generalisation, MDI(0.6), of the MDI prior, given by (2.24) with $a = 0.6$, provided good predictive properties for the cases we considered. Such priors are intended for use in situations where substantial prior information is not available and it is expected that information provided by the data will dominate the posterior distribution (O'Hagan, 2006). If a very high threshold, with few threshold excesses, is chosen then the data may not dominate resulting in such large posterior uncertainty about model parameters, especially the GP shape parameter ξ, that non-negligible posterior probability is placed on unrealistically large values of ξ. This may mean that extreme value extrapolations are unrealistic. We use the MDI(0.6) prior for the GP parameters, bearing in mind the potential problems we have just discussed. We also use the Jeffreys Beta(1/2, 1/2) prior for p_u, as detailed in section 2.6.

We continue this chapter by applying the Bayesian cross-validation threshold selection approach to the Gulf of Mexico dataset with an aim of choosing the single threshold that attains the best (out-of-sample) predictive performance. We sample from posterior distributions for GP parameters using the generalised ratio-of-uniforms method with mode relocation and $r = 1/2$ (section 2.4.2). For analyses involving a single (real) dataset we simulate posterior samples of size 10,000. For the simulation studies, where many datasets are analysed, we simulate posterior samples of size 1,000. Equivalent results were obtained using a random walk Metropolis-Hastings sampler (section 2.4.1), but we prefer to use the ratio-of-uniforms method as it produces independent samples and convergence monitoring is not required.

3.4 Significant wave height data: single threshold

We apply our Bayesian cross-validation scheme to the Gulf of Mexico storm peak significant wave height dataset, to inform the selection of a single threshold. We use training thresholds $\{u_1, \ldots, u_{20}\}$ set at the 0% (sample minimum), 5%, ..., 95% sample quantiles and validation threshold equal to u_{20}. To examine sensitivity to the choice of the highest threshold we also use validation thresholds set at the 80%, 85% and 90% sample quantiles. With a total sample size of 315 observations, there are only approximately 16 excesses of the 95% threshold: using so few observations to train, and to validate, the GP model is optimistic. The 90%/85%/80% thresholds have approximately 32/48/63 excesses so only the 80% threshold obeys the rule-of-thumb that there should be at least 50 excesses.

To facilitate comparison of the training threshold performance for different validation thresholds we define the estimated *threshold weight* associated with training threshold u_i, assessed at validation threshold v, by

$$w_v(u_i) = \exp\{\widehat{T}_v(u_i)\} / \sum_{j=1}^{k} \exp\{\widehat{T}_v(u_j)\}, \tag{3.16}$$

where $\widehat{T}_v(u)$ is defined in (3.4). The ratio $w_v(u_2)/w_v(u_1)$, an estimate of a *pseudo-Bayes factor* (Geisser and Eddy, 1979), is a measure of the relative performance of threshold u_2 compared to threshold u_1. In section 3.5 these weights are used to combine inferences from different training thresholds.

Figure 21 shows the plot of the estimated threshold weights against training threshold based on different validation thresholds. The 80%, 85% and 90% validation thresholds are in reasonable agreement, suggesting training thresholds, in the region of the 55-70% sample quantiles (for which the MLE of $\xi \approx 0.1$). The 95% validation threshold suggests a slightly higher training threshold.

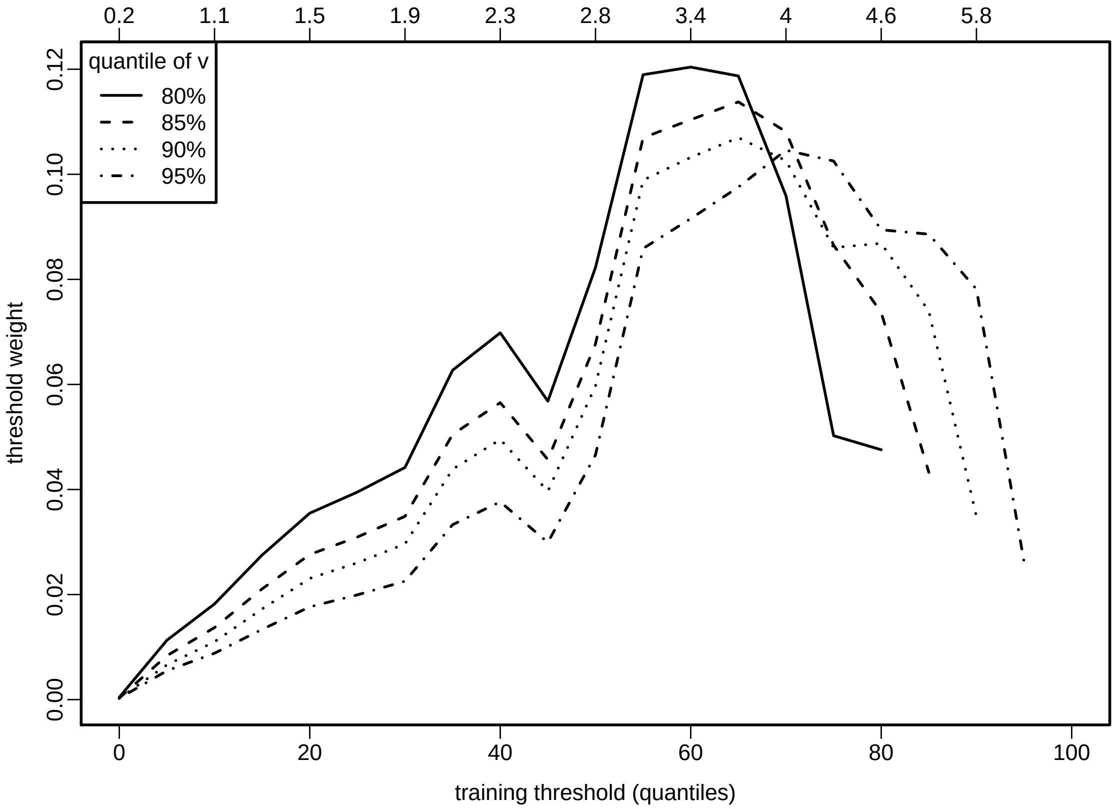

Figure 21: Analysis of Gulf of Mexico significant wave height dataset using an MDI(0.6) prior for the GP parameters. Estimated threshold weights for validation threshold v, by training threshold u. The upper axis gives the significant wave height scale in metres.

The profile of threshold weight with training threshold tends to be flatter for high validation thresholds than for lower validation thresholds, because high validation thresholds contain fewer threshold excesses, with which to compare training thresholds, than low validation thresholds. This is particularly evident if 97.5% and 99% validation thresholds are used (not shown). The Gulf of Mexico dataset contains 315 observations, so the 97.5% and 99% validation thresholds are exceeded by only 8 and 3 observations respectively, too few excesses on which to base an assessment. This conclusion ties in with the rule of thumb suggested earlier that around 50 threshold excesses would be sufficient to carry out meaningful inference. The 80% and 85% validation thresholds provide us with 63 and 48 threshold excesses respectively and it is reassuring to note that both validation thresholds agree on the sample quantile region to select the threshold.

If $v = u_k$ is set at the 80% sample quantile, then the ‘best’ threshold among those considered is the 60% sample quantile. If instead v is set at the 85% sample quantile, then the ‘best’ threshold among those considered is the 65% sample quantile.

Prediction of extreme observations

From a practical point of view the aim is to provide guidance to design engineers on the likely values of future extreme events. This was introduced in section 2.5 where we discussed two approaches, based on (i) the $100(1-1/N)\%$ quantile of the distribution of an annual maximum (the N-year return level), and (ii) (summaries of) the distribution of an N-year maximum M_N. Here we explore how the threshold chosen affects predictive inferences about these quantities for the Gulf of Mexico significant wave height dataset to establish two measures of future extreme events. In section 2.5 we noted that a predictive approach tends to inflate predictive N-year return levels relative to the corresponding median of the predictive distribution of M_N.

Figure 22 shows that the N-year predictive return levels and the medians of the predictive distribution of N-year maxima M_N are close for $N = 100$, where little or no extrapolation is required, but for $N = 1,000$ and $N = 10,000$ the former is greater than the latter. From the 55% training threshold upwards, which includes thresholds that have high estimated training weights, estimates of the median of M_{1000} and M_{10000} from the Gulf of Mexico data are implausibly large, e.g. 31.6m and 56.7m for the 75% threshold. The corresponding estimates of the predictive return levels are even less credible. The problem is that high posterior probability of large positive values of ξ, caused by high posterior uncertainty about ξ, translates into large predictive estimates of extreme quantiles.

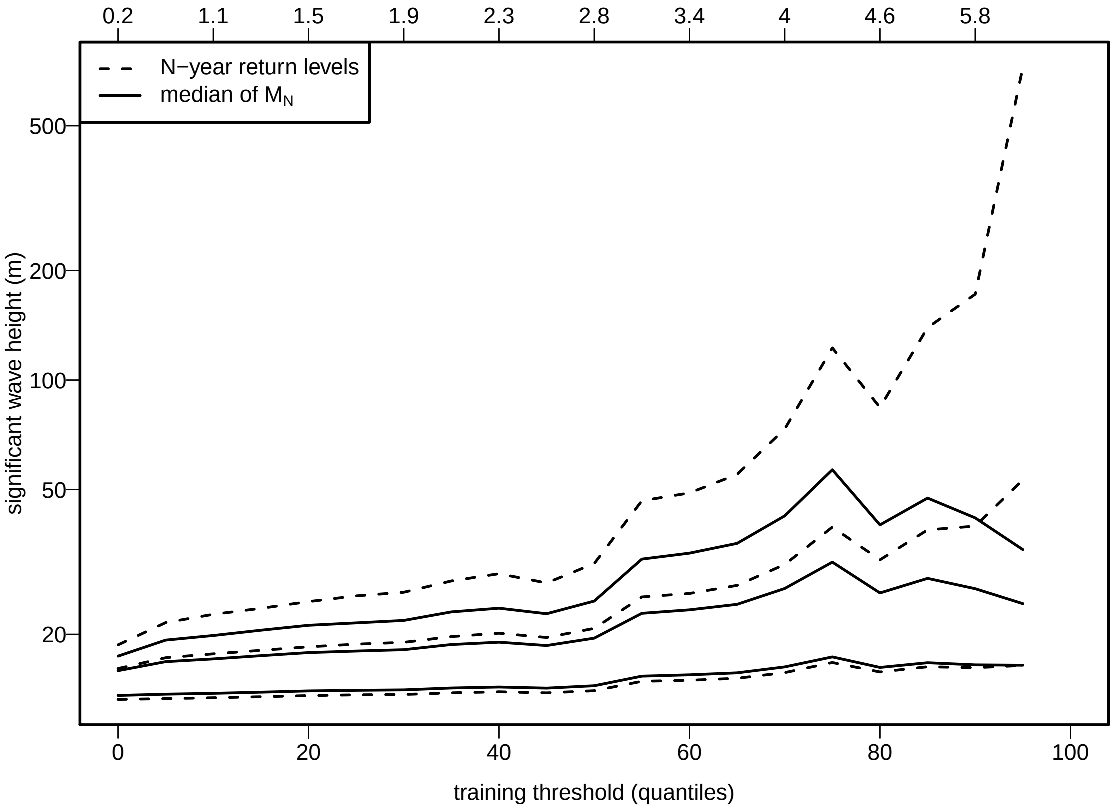

Figure 22: Analysis of Gulf of Mexico significant wave height dataset using an MDI(0.6) prior, for the GP parameters. N-year predictive return levels and medians of the predictive distribution of M_N for $N = 100, 1,000$ and $10,000$ by training threshold u. The upper axis gives the significant wave height scale in metres.

Figure 23 gives two examples of the posterior samples of σ_u and ξ underlying the plots in figures 21 and 22. The marginal posterior distributions of ξ are positively skewed, because for fixed σ_u, ξ is bounded below by $\sigma_u/(x_m - u)$. The higher the threshold the larger the posterior uncertainty and the greater the skewness towards values of ξ that correspond to a heavy-tailed distribution. For the Gulf of Mexico data at the 95% threshold $\widehat{P}(\xi > 1/2) \approx 0.20$ and $\widehat{P}(\xi > 1) \approx 0.05$. This issue is not peculiar to a Bayesian analysis: frequentist confidence intervals for ξ and for extreme quantiles are also unrealistically wide.

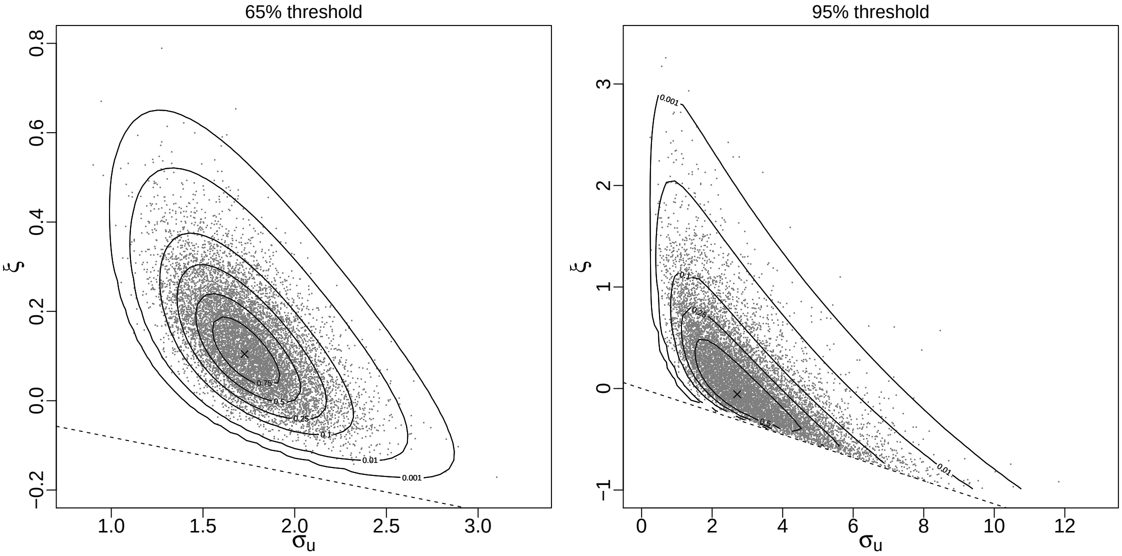

Figure 23: Samples from the posterior distribution $\pi(\sigma_u, \xi \mid \boldsymbol{x})$ using an MDI(0.6) prior, with superimposed (unnormalised) contours of the posterior density. The dashed lines show the support of the posterior distribution, that is, $\xi > \sigma_u/(x_m - u)$. A cross shows the posterior mode. Left: 'best' training threshold (based on the 85% and 90% validation thresholds) at 65% sample quantile. Right: 95% sample quantile training threshold.

Physical considerations suggest that there is a finite upper limit to storm peak significant wave height H_s^{sp} (Jonathan and Ewans, 2013). However, if there is positive posterior probability on $\xi \geqslant 0$ then the implied distribution of H_s is unbounded above and on extrapolation to a sufficiently long time horizon, N_l say, unrealistically large values will be implied. This may not be a problem if N_l is greater than the time horizon of practical interest, that is, the information in the data is sufficient to allow extrapolation over this time horizon. If this is not the case then one should incorporate supplementary data (perhaps by pooling data over space as in Northrop and Jonathan (2011)) or prior information. Jonathan and Ewans (2013) advocate this if there are fewer than 50 threshold excesses. Some practitioners assume that $\xi < 0$ *a priori*, in order to ensure a finite upper limit, but such a strategy may sacrifice performance at time horizons of importance. It is important to appreciate that the extent to which one can hope to extrapolate with realism is limited by the information available and the level at which one can set the threshold.

We now consider another dataset for which these issues are far less of a problem than for the Gulf of Mexico dataset.

3.4.1 North sea significant wave height data

This dataset, from an unnamed location in the North Sea, contains 628 hindcast storm peak significant wave heights from October 1964 to March 1995, restricted to the period October to March within each year. Figure 24 shows a times series plot of the storm peaks and a parameter stability plot for the GP shape parameter ξ.

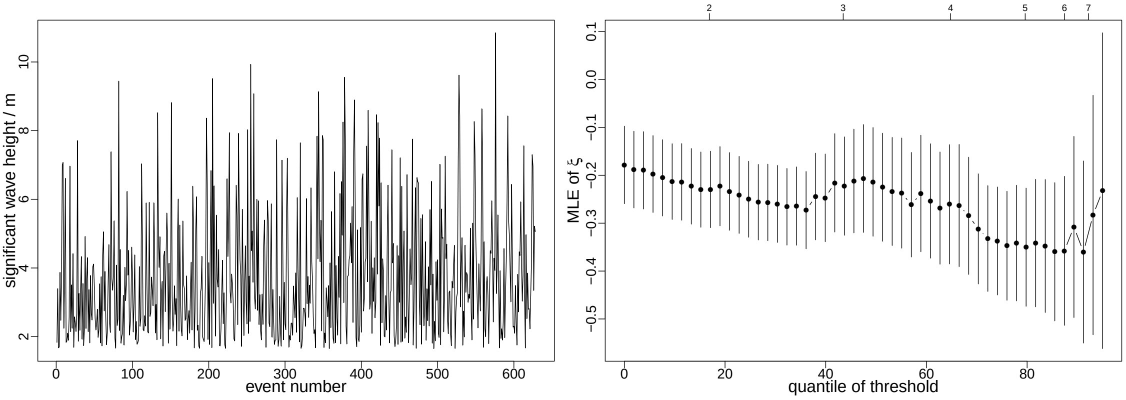

Figure 24: North Sea significant wave height dataset. Left: times series plot. Right: parameter stability plot for the GP shape parameter ξ. The upper axis gives the significant wave height scale in metres.

The main differences between the North Sea and Gulf of Mexico datasets are that (i) the North Sea dataset has approximately double the number of observations, (ii) across all thresholds the MLE of the GP shape parameter ξ is negative for the North Sea dataset, whereas, for higher thresholds, it is positive for the Gulf of Mexico dataset, and (iii) perhaps the MLE of ξ stabilizes at a lower quantile of threshold for the North Sea dataset than for the Gulf of Mexico dataset. The effects of these differences can be seen in figure 25. The Bayesian cross-validation scheme (left hand plot) gives thresholds in the region of the 35% sample quantile the greatest weight, and there is close agreement between the different validation thresholds (effectively different choice of the highest training threshold). In the right hand plot, the estimated medians of the predictive distribution of M_N for $N = 100, 1,000$ and $10,000$ are realistic for all the training thresholds considered. It is only as we approach the very highest thresholds (for example, the 97.5% sample quantile, which has approximately 16 excesses) that the estimated $1,000$-year and $10,000$ year predictive return levels increase rapidly, as the posterior distribution of ξ (see figure 26) becomes very diffuse.

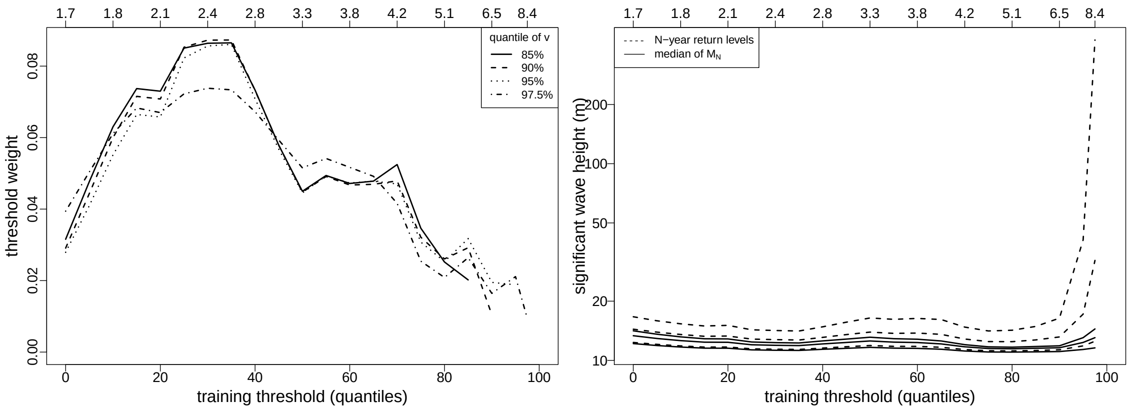

Figure 25: Analysis of North Sea significant wave height dataset using an MDI(0.6) prior for the GP parameters. Left: estimated threshold weights for validation threshold v, by training threshold u. Right: estimated N-year predictive return levels and medians of the predictive distribution of M_N for $N = 100, 1,000$ and $10,000$ by training threshold u. The upper axes gives the significant wave height scale in metres.

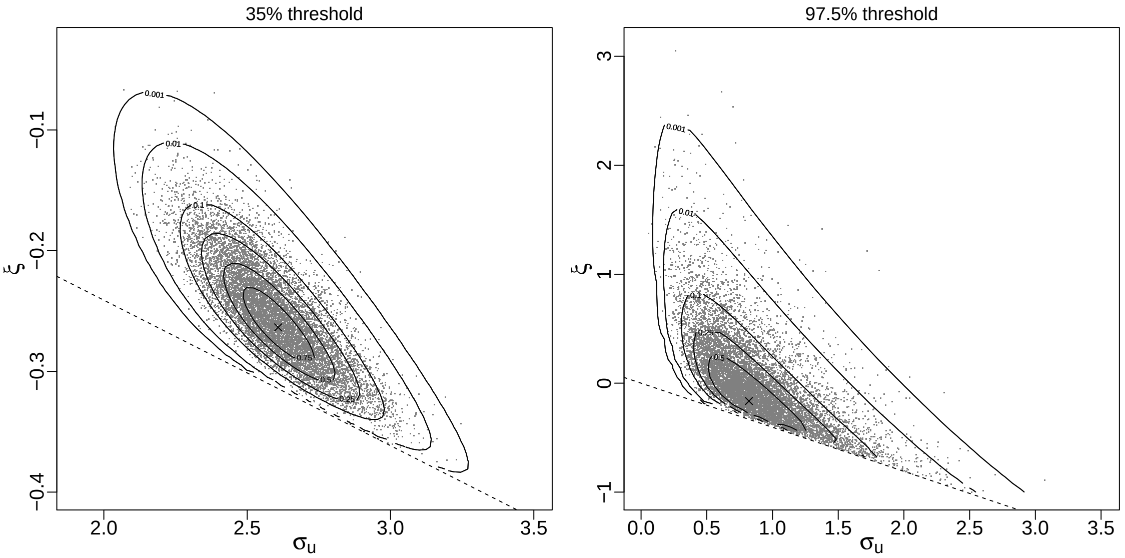

Figure 26: Samples from the posterior distribution $\pi(\sigma_u, \xi \mid \boldsymbol{x})$ using an MDI(0.6) prior, with superimposed (unnormalised) contours of the posterior density. The dashed lines show the support of the posterior distribution, that is, $\xi > \sigma_u/(x_m - u)$. A cross shows the posterior mode. Left: 'best' training threshold (based on the 85%, 90% and 95% validation thresholds) at 35% sample quantile. Right: 95% sample quantile training threshold.

In the following section we investigate how uncertainty about the threshold choice can be accounted for in our cross-validation method for threshold selection.

3.5 Accounting for uncertainty in threshold selection

Traditional statistical analysis within the context of model selection often ignores the uncertainty that is present in the selection process. A technique that is designed to account for this uncertainty is called Bayesian Model Averaging (BMA). It has been successfully applied in a variety of model selection scenarios such as linear regression and generalised linear models with the aim of improving predictive performance. Accounting for uncertainty in the choice of model is achieved by averaging over a number of competing models.

Historically, BMA was first mentioned in the 1960s outside the scope of statistical analysis and gained ground in the 1970s in the economics literature which involved combining predictions from forecasting models. From a statistical point of view, the work of Roberts (1965) involved model averaging in the form of combining opinions from two experts. This was followed by Leamer (1978) who presented the basic concepts for BMA, however as Hoeting et al. (1999) notes this received little attention and "little progress was made until new theoretical developments and computational power enabled researchers to overcome the difficulties related to implementing BMA".

In this section we use BMA (Hoeting et al., 1999, Gelfand and Dey, 1994) to combine inferences based on different thresholds. Consider a set of k training thresholds $u_1, \ldots, u_k$ and a particular validation threshold v. We view the k Bin-GP models associated with these thresholds as competing models. There is evidence that one tends to get better predictive performance by interpolating smoothly between all models entertained as plausible *a priori*, than by choosing a single model (Hoeting et al., 1999, section 7). Suppose that we specify prior probabilities $P(u_i), i = 1, \ldots, k$ for these models. In the absence of more specific prior information, and in common with Wadsworth and Tawn (2012), we use a discrete Uniform prior $P(u_i) = 1/k, i = 1, \ldots, k$. We suppose that the thresholds occur at quantiles that are equally spaced on the probability scale. We prefer this to equal spacing on the data scale because it seems more natural and retains its property of equal spacing under data transformation.

Let $\boldsymbol{\theta}_i = (p_i, \sigma_i, \xi_i)$ be the Bin-GP parameter vector under model u_i, under which the prior is $\pi(\boldsymbol{\theta}_i \mid u_i)$. By Bayes' theorem, the *posterior threshold weights* are given by

$$P_v(u_i \mid \boldsymbol{x}) = \frac{f_v(\boldsymbol{x} \mid u_i)\, P(u_i)}{\sum_{i=1}^{k} f_v(\boldsymbol{x} \mid u_i)\, P(u_i)},$$

where

$$f_v(\boldsymbol{x} \mid u_i) = \int f_v(\boldsymbol{x} \mid \boldsymbol{\theta}_i, u_i)\pi(\boldsymbol{\theta}_i \mid u_i)\ \mathrm{d}\boldsymbol{\theta}_i$$

is the predictive density of $\boldsymbol{x}$ based on validation threshold v under model (training threshold) u_i. However, $f_v(\boldsymbol{x} \mid u_i)$ is difficult to estimate and is improper if $\pi(\boldsymbol{\theta}_i \mid u_i)$ is improper. Following Geisser and Eddy (1979) we use $\prod_{r=1}^{m} f_v(x_r \mid \boldsymbol{x}_{(r)}, u_i) = \exp\{\widehat{T}_v(u_i)\}$ as a surrogate for $f_v(\boldsymbol{x} \mid u_i)$ to give

$$\widehat{P}_v(u_i \mid \boldsymbol{x}) = \frac{\exp\{\widehat{T}_v(u_i)\}\, P(u_i)}{\sum_{j=1}^{k} \exp\{\widehat{T}_v(u_j)\}\, P(u_j)}. \tag{3.17}$$

Let $\boldsymbol{\theta}_{ij}, j = 1, \ldots, n_{\boldsymbol{\theta}}$ be a sample from $\pi(\boldsymbol{\theta}_i \mid \boldsymbol{x})$, the posterior distribution of the GP parameters based on threshold u_i. We calculate a model-averaged estimate of the predictive distribution function of M_N using

$$\widehat{P}_v(M_N \leqslant x \mid \boldsymbol{x}) = \sum_{i=1}^{k} \widehat{P}(M_N \leqslant x \mid \boldsymbol{x}, u_i)\widehat{P}_v(u_i \mid \boldsymbol{x}), \tag{3.18}$$

where, by analogy with (2.23),

$$\widehat{P}(M_N \leqslant x \mid \boldsymbol{x}, u_i) = \frac{1}{n_{\boldsymbol{\theta}}} \sum_{j=1}^{n_{\boldsymbol{\theta}}} F(z; \boldsymbol{\theta}_{ij})^{n_y N}.$$

The solution $\widehat{x}_P^N$ of

$$\widehat{P}_v(M_1 \leqslant \widehat{x}_P^N \mid \boldsymbol{x}) = 1 - 1/N \tag{3.19}$$

provides a model-averaged estimate of the N-year predictive return level, based on validation threshold v.

3.6 Simulation study: single and multiple thresholds

We compare inferences based on a single threshold to those obtained by averaging over many thresholds. The comparisons are based on random samples simulated from three distributions: a unit exponential, a standard normal and a uniform-GP hybrid; chosen to represent qualitatively different extremal behaviour. With knowledge of the simulation model one would be able to choose a suitable threshold, at least approximately. In practice this would not be the case and so it is interesting to see how well the strategies of choosing the ‘best’ threshold u^* (section 3.3.3), and of averaging inferences over different thresholds (section 3.5), compare to this choice and how the estimated weights $\widehat{P}_v(u_i \mid \boldsymbol{x})$ in (3.16) vary over u_i, and with v.

The unit exponential distribution has the property that a GP(1,0) model holds above any threshold. Therefore, choosing the lowest threshold possible is optimal. For the normal distribution the GP model does not hold for any finite threshold, the approximation of a GP model to the truth improving slowly as the threshold increases. The limiting case is the exponential distribution ($\xi = 0$), but at sub-asymptotic levels the effective shape parameter is negative (Wadsworth and Tawn, 2012) and one expects a relatively high threshold to be indicated. The uniform-GP hybrid has a constant density up to its 75% quantile and a GP density with shape parameter 0.1 for excesses of the 75% quantile. Thus, a GP distribution holds only above the 75% threshold.

In each case we simulate 1000 samples each of size 500, representing 50 years of data with an average of 10 observations per year. We set training thresholds at the $50\%, 55\%, \ldots, 95\%$ sample quantiles and validation thresholds at the 95% sample quantile (equal to the largest training threshold). For each sample, and for values of N between 100 and $10,000$, we solve $\widehat{P}_v(M_N \leqslant z \mid \boldsymbol{x}) = 1/2$ for z (see (2.23)) to give estimates of the median of M_N. We show results for three single thresholds: the threshold one might choose based on knowledge of the simulation model; the 'best' threshold u^*, that maximizes the measure $\widehat{T}_v(u)$ of predictive performance (section 3.3.3); and another (clearly sub-optimal) threshold chosen to facilitate further comparisons. We compare these estimates, and a model-averaged estimate based on (3.18) to the true median of M_N, $F_{\text{true}}^{-1}((1/2)^{10N})$, where F_{true} is the true distribution function of the (unthresholded) observations.

The results for the exponential distribution are summarized in figure 27. As expected, all strategies have negligible bias. The model-averaged estimates match closely the behaviour of the optimal strategy (the 50% threshold, as it is the lowest training threshold considered here). The best single threshold results in slightly greater variability, offering less protection than model-averaging against estimates that are far from the truth.

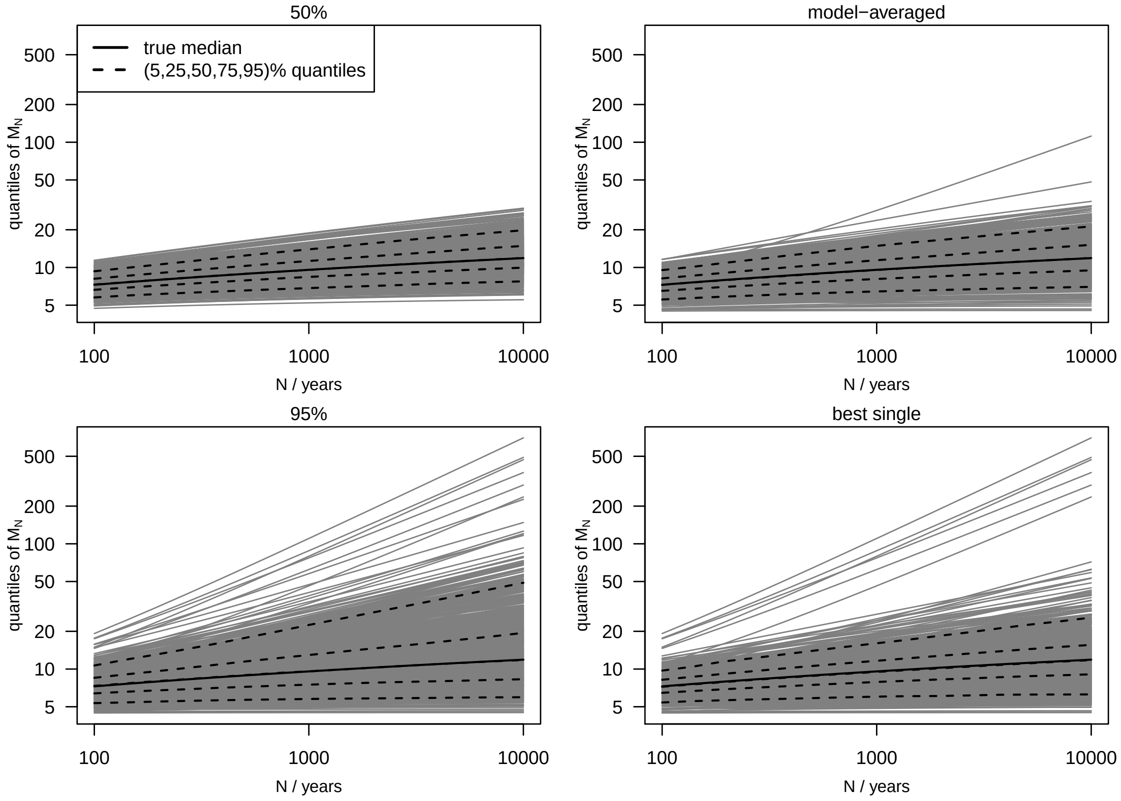

Figure 27: Predictive medians of M_N compared with the true median (solid black lines), for datasets simulated from a unit exponential distribution. The set of training thresholds is the $50\%, 55\%, \ldots, 95\%$ sample quantiles. Grey lines: individual lines for each dataset. Dashed black lines: pointwise 5%, 25%, 50%, 75% and 95% sample quantiles. Threshold strategies: sample median (top left); 95% sample quantile (bottom left); model-averaged estimate (top right); best single threshold (bottom right).

In the normal case (figure 28) the expected underestimation is evident for large N: this is substantial for a 50% threshold but small for a 95% threshold. The CV-based strategies have greater bias than those based on a 95% threshold, because inferences from lower thresholds contribute, but have much smaller variability.

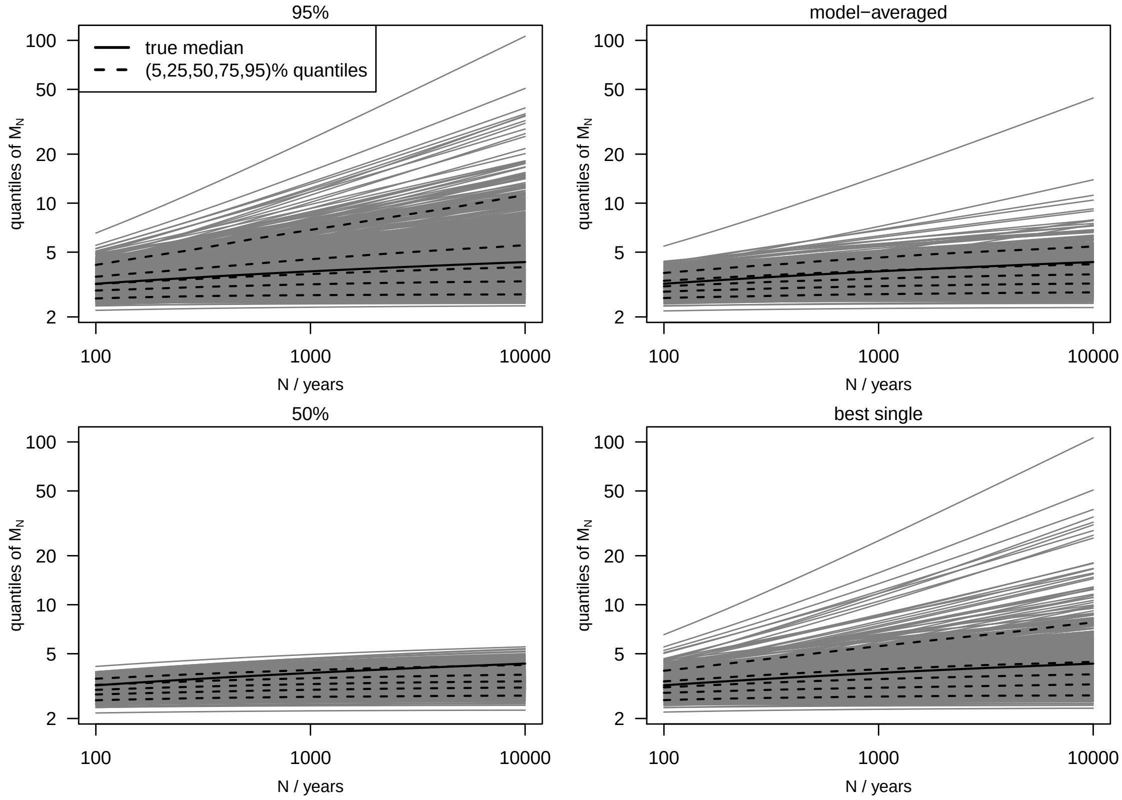

Figure 28: Predictive medians of M_N compared with the true median (solid black lines), for datasets simulated from a standard normal distribution. The set of training thresholds is the $50\%, 55\%, \ldots, 95\%$ sample quantiles. Grey lines: individual lines for each dataset. Dashed black lines: pointwise 5%, 25%, 50%, 75% and 95% sample quantiles. Threshold strategies: 95% sample quantile (top left); sample median (bottom left); model-averaged estimate (top right); best single threshold (bottom right).

Similar findings are evident in figure 29 for the uniform-GP hybrid distribution: contributions from thresholds lower than the 75% quantile produce negative bias but model-averaging achieves lower variability than the optimal 75% threshold.

In all these examples the CV-based strategies seem preferable to a poor choice of a single threshold, and, in a simple visual comparison of bias and variability, are not dominated clearly by a (practically unobtainable) optimal threshold. A more definitive comparison would depend on the problem-dependent losses associated with over- and under-estimation. Using model-averaging to account for threshold uncertainty is conceptually attractive but, compared to the 'best' threshold strategy, it's reduction in variability is at the expense of greater bias. Again the loss function of the problem is relevant to this comparison.

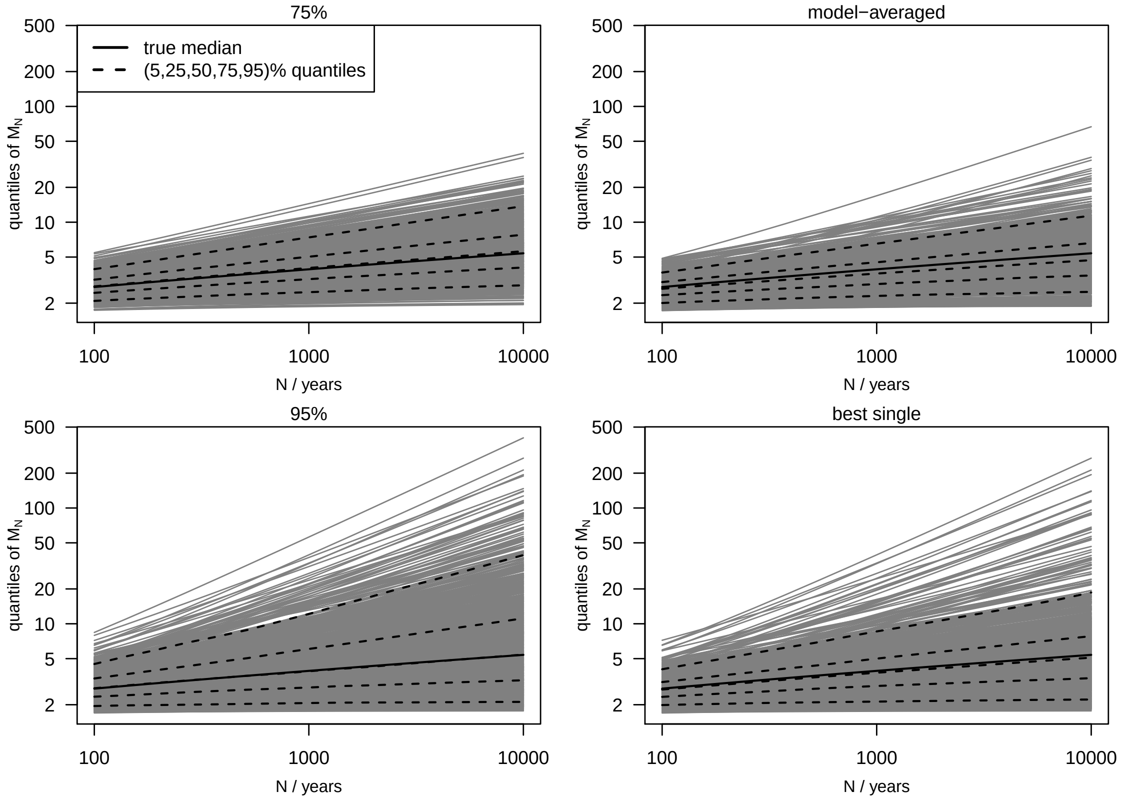

Figure 29: Predictive medians of M_N compared with the true median (solid black lines), for datasets simulated from a hybrid uniform-GP distribution. The set of training thresholds is the $50\%, 55\%, \ldots, 95\%$ sample quantiles. Grey lines: individual lines for each dataset. Dashed black lines: pointwise 5%, 25%, 50%, 75% and 95% sample quantiles. Threshold strategies: 75% sample quantile (top left); 95% sample quantile (bottom left); model-averaged estimate (top right); best single threshold (bottom right).

Figure 30 summarises how the posterior threshold weights vary with training threshold, again based on the 95% validation threshold. There are a few datasets for which the 95% training threshold receives very high weight. These result from samples where the most extreme observations are very large relative to the other observations. The potential for the largest observations to have very strong influence is a well-known feature of extreme value analyses (Davison and Smith, 1990). For the exponential and hybrid examples the mean and median posterior threshold weights behave roughly as one would expect: decreasing in training threshold for the exponential example, and peaking at approximately the 70% quantile (i.e. lower than the 75% quantile) for the uniform-GP example. In the exponential example the best available threshold (the 50% quantile) receives the highest posterior weight with relatively high probability and in the hybrid example this is true of the 70% quantile. It is less clear what to expect for the normal example and the message from the simulation study is less clear. The mean and median posterior weights peak at approximately the 70%–80% quantile. The graph of relative frequency with which each threshold receives the highest posterior weight is relatively flat, with lower thresholds being the 'best' slightly more often than higher thresholds. Given that the GP limit is only attained in the limit as the threshold tends to infinity it may be that much higher thresholds should be explored, requiring much larger simulated sample sizes, such as those used by Wadsworth and Tawn (2012).

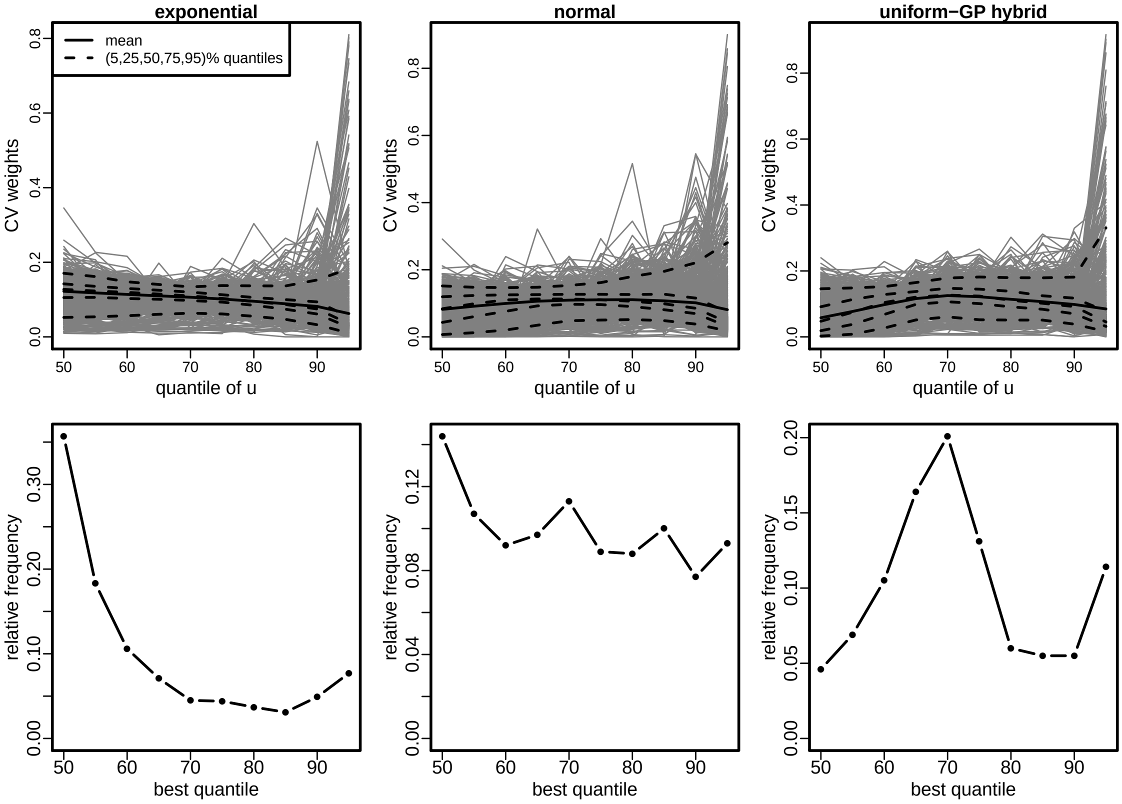

Figure 30: Summaries of CV weights by training threshold where $v = 95\%$. Top: the grey lines give individual lines for each simulated dataset with threshold-specific sample means (solid black line) and sample (5, 25, 50, 75, 95)% quantiles (dashed black lines). Bottom: relative frequency with which each threshold has the largest CV weight. Left: exponential distribution. Middle: normal distribution. Right: uniform-GP hybrid distribution.

In these examples the use of the 95% sample quantile as the largest training threshold, and hence the validation threshold, results in only 25 excesses of the validation threshold. If we lower the largest training threshold, say to the 90% (50 excesses) or 85% (75 excesses) sample quantile then, for the exponential and hybrid example, the locations of peaks in the plots are clearer. Figure 31 summarises the posterior thresholds weights for the 85% quantile case.

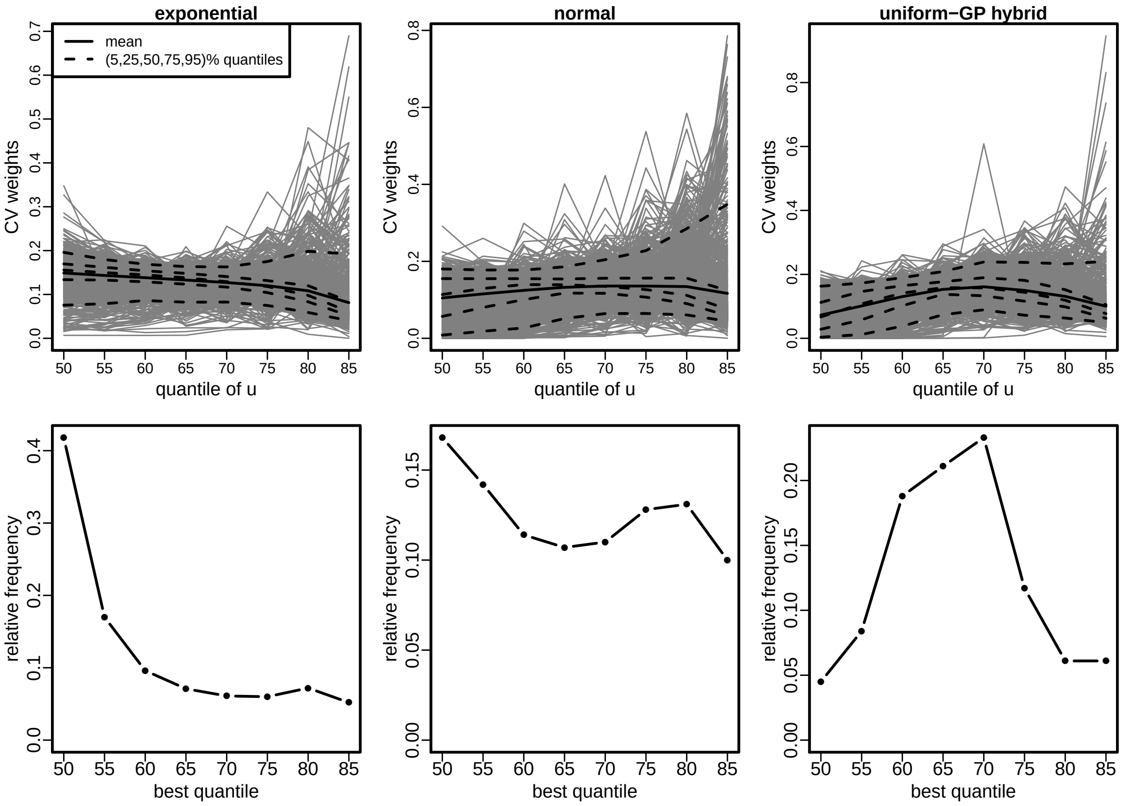

Figure 31: Summaries of CV weights by training threshold where $v = 85\%$. Top: the grey lines give individual lines for each simulated dataset with threshold-specific sample means (solid black line) and sample (5, 25, 50, 75, 95)% quantiles (dashed black lines). Bottom: relative frequency with which each threshold has the largest CV weight. Left: exponential distribution. Middle: normal distribution. Right: uniform-GP hybrid distribution.

3.7 Significant wave height data: threshold uncertainty

We return to the Gulf of Mexico and North Sea significant wave height datasets, using the methodology of section 3.5 to average extreme value inferences obtained from different thresholds. Figure 32 shows plots of estimated predictive N-year return levels and selected (to facilitate comparison with N-year predictive return levels) quantiles of M_N against N, for different validation threshold levels. The set up is the same as in section 3.4, i.e. in the first instance based on training thresholds $\{u_1, \ldots, u_{20}\}$ set at the 0% (sample minimum), 5%, ..., 95% sample quantiles. We use three different validation thresholds, set so that the numbers of excesses are approximately 32, 48 and 64 respectively for each dataset. In cases where a validation threshold is below a training threshold the effect is that inferences from this training threshold get zero weight in the model averaging.

There is lower sensitivity to validation threshold for the North Sea data than the

Gulf of Mexico data. This is partly because the bulk of the posterior probability is on negative values of ξ (which imply a finite upper end point) for the former and on positive values of ξ (which imply an unbounded distribution) for the latter. On extrapolation into the upper tail the uncertainty is generally much smaller in the former case.

Figure 32 shows that the approximate link between predictive N-year return levels and quantiles of the predictive distribution of M_N depends on N: the larger N is the higher is the quantile of M_N to which the N-year predictive return level corresponds approximately.

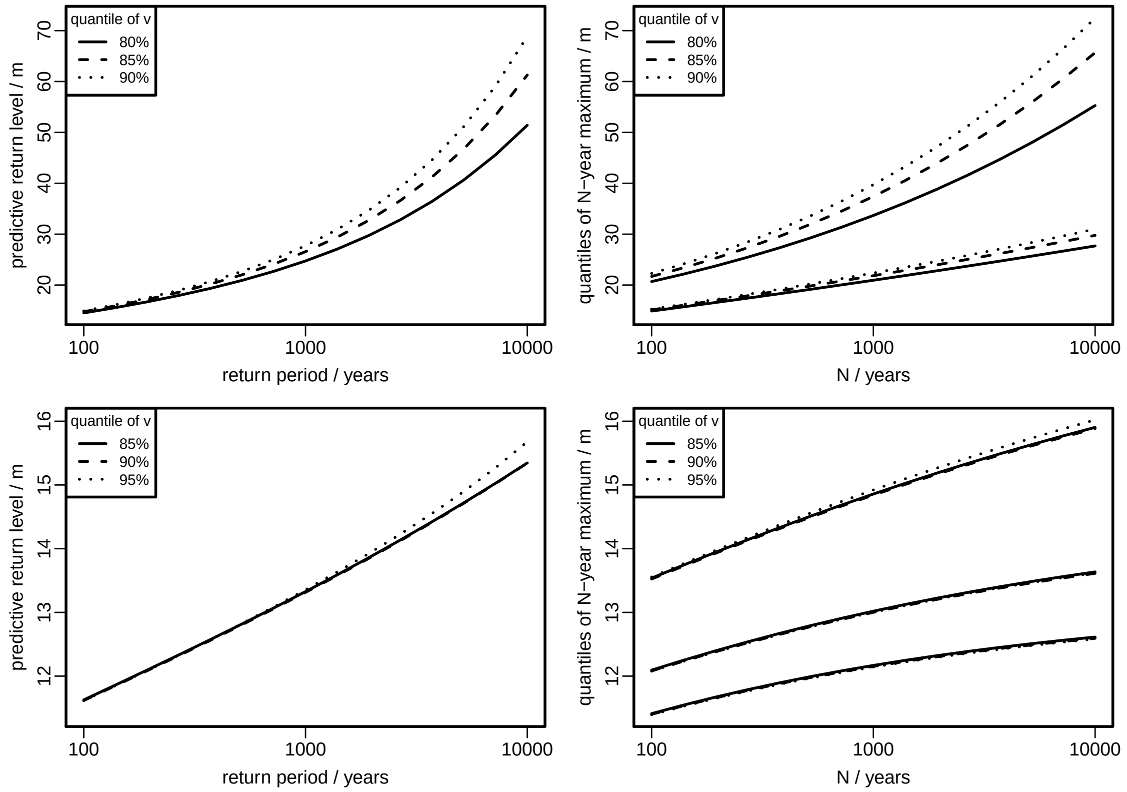

Figure 32: Model-averaged N-year predictive return levels and selected quantiles of the predictive distribution of N-year maximum M_N, based on different validation thresholds. Top: Gulf of Mexico data (50% and 85% quantiles of M_N). Bottom: North sea data (50%, 75% and 95% quantiles of M_N).

For the Gulf of Mexico data the medians of the predictive distribution of M_N are not unrealistic: averaging inferences over thresholds has provided some protection against the very large estimates obtained for some of the individual thresholds (see figures 21 and 25). However, between $N = 1,000$ and $N = 10,000$ the 85% quantiles of M_N and the N-year predictive return levels become unrealistically large.

4 Threshold selection for the NID case

In chapter 3 we directed our focus on extreme value threshold selection process for the i.i.d. case, proposing a new graphical diagnostic tool for selecting the threshold based on a Bin-GP model for threshold exceedances. The methodology uses Bayesian computation to perform predictive inferences. In the absence of genuine prior information about extreme value parameters, we appealed to the results in chapter 2 to suggest a working prior distribution for the parameters of the GP distribution.

In this chapter we extend the cross-validation approach introduced in chapter 3 to the case where the independence assumption of the underlying random variables is assumed to be unrealistic, that is, we investigate the threshold selection process for the n.i.d. case. Following Süveges and Davison (2010) we base threshold selection on a limiting model (the K-gaps model, see section 1.7) for the distribution of threshold *inter*-exceedance times. This model is parameterised in terms of a scalar parameter θ, the extremal index, which measures the strength of temporal dependence at extreme levels. Otherwise, the general idea is the same as in chapter 3: different training thresholds are compared based on the ability to predict out-of-sample inter-exceedance times at a validation threshold v.

Using the K-gaps model, rather than a model for threshold exceedance times *and* threshold excesses, has some advantages. If maximum likelihood estimation is used then the regularity condition ($0 < \theta < 1$) for the K-gaps model is less restrictive than that for the GP model (Süveges and Davison, 2010, page 18). Even in the Bayesian setup considered here the relative simplicity of dealing with a scalar parameter is attractive. While it is necessary that a threshold is judged as suitable in the context of the K-gaps model for threshold inter-exceedance times, its suitability in the context of a model for threshold excesses also matters. However, modelling threshold excesses in the n.i.d. case is more difficult than in the i.i.d. case: see Fawcett and Walshaw (2012) and references therein. We return to this issue in the discussion in chapter 6.

In section 1.7 we introduced the K-gaps exponential mixture model following the work of Süveges and Davison (2010). Recall that, a K-gap is defined as $S = \max(T - K, 0)$, where T is the time between two successive exceedances of a threshold u, and $\overline{F}(u)$ denotes the probability of threshold exceedance. Under the K-gaps limiting mixture model (1.11), a scaled K-gap $\overline{F}(u)S$ is zero with probability $1 - \theta$ (corresponding to $T \leqslant K$), and otherwise it is an exponential variable with rate parameter θ.

For a process sampled at regular time intervals T is an integer and so the scaled K-gaps are discrete: equal to integer multiples of $\overline{F}(u)$ which, in practice, is estimated by the proportion q of observations that exceed u. Thus, there is no finite threshold for which the exponential part of the K-gaps model is true, and therefore the K-gaps model is always misspecified. However, it is still worthwhile to consider for which values of (u, K) the misspecification is sufficiently small for the K-gaps model to be useful. This situation is subtly different from that in chapter 3, where examples do exist (for example, the exponential and hybrid examples in section 3.6) where the Bin-GP is well-specified.

The issues associated with the selection of u are the same as in chapter 3, trading off bias resulting from model misspecification if u is too low with a lack of precision of estimation if u is too high. Süveges (2008, page 55) discusses the effects of making a poor choice of K. If K is too low then the model is misspecified because it wrongly supposes that (non-zero) K-gaps are independent and exponentially distributed, resulting from between-cluster inter-exceedance times, when in fact they result from dependent within-cluster exceedances. If K is too high then too many between-cluster inter-exceedance times are truncated to zero leading to a loss of information and to bias from these times contributing to the point mass at zero part of the mixture model.

Therefore, it is important to choose the pair (u, K) appropriately. In this chapter we seek to inform the choice of u under the assumption that K has been set at an appropriate value. For some processes the value of K is known (see section 4.7 for two such examples). For others, and of course for real datasets, K needs to be chosen empirically. In section 4.6 we use a simple graphical diagnostic tool to make this choice. In future work we will seek to extend the methodology to inform the choice of (u, K), rather than the value of u given an appropriate choice of K.

4.1 Information matrix test (IMT)

Süveges and Davison (2010) describe a maximum likelihood estimator for the extremal index θ and use a model misspecification test known as *information matrix test* (IMT) to reject pairs (u, K) for which the limiting model is judged as invalid. They also assess their ideas through a simulation study and conclude that the IMT is useful in selecting the threshold and run parameter and that it supplements the classical threshold selection approaches.

The null hypothesis for this test (developed by White (1982)) is that the model is well-specified, according to a measure of the distance between the Fisher expected information and the variance of the score vector. Under the null the test statistic follows (asymptotically) a χ^2_1 distribution. Therefore, the rejection of a potential threshold-run pair (u, K) can be carried out through classical hypothesis testing.

Using this method, Fukutome et al. (2014) propose an automatic threshold-run pair selection. They test a large grid of $(u,\ K)$ plausible pairs and retain those that result in a very low IMT value (less than 0.05). Pairs with a value of u that produces fewer than 80 threshold exceedances are discarded because simulations in Süveges and Davison (2010) reveal that such datasets tend to have low power to detect departure from the null. Then, for each pair, they decluster the data to identify a set of approximately independent cluster maxima, and select the pair with the largest number of cluster maxima. Fukutome et al. (2014) demonstrate their method using a dataset of hourly precipitation in Switzerland.

Goodness-of-fit tests are not uncommon in extreme value analysis (Davison and Smith, 1990, Dupuis, 1998, Wadsworth and Tawn, 2012) and benefit from the fact that the assessment of one threshold is not affected directly by tests made at other thresholds. However, this method of threshold selection has a number of drawbacks. Firstly, a subjective choice of a test level needs to be chosen beforehand. Note also that it may be necessary to make some adjustment for the fact that multiple tests are performed on strongly related datasets. Secondly, as the threshold level u increases the power to detect departures from the null hypothesis H_0 decreases. Thirdly, the issue of the bias-variance trade-off does not seem to be accounted for in a clearly defined way. What is more, a complication with this method is the possibility of having two (or more) different thresholds that do not reject the null hypothesis and then it is not clear what is the procedure of choosing a single threshold, hence the need for an automation rule like that proposed by Fukutome et al. (2014).

In section 4.5 we use cross-validation to compare training thresholds based on an extension of the Süveges and Davison (2010) K-gaps likelihood to incorporate information from censored inter-exceedance times, which we describe in the next section.

4.2 Censored inter-exceedance times

Let us now concentrate on the first and last exceedance of the stationary process and in particular the two ‘edges’ of the time series, as depicted by the blue shaded regions in figure 33 using a segment from the Newlyn data. The starting edge contains observations before the first exceedance is observed, and the ending edge contains observations after the last exceedance is observed.

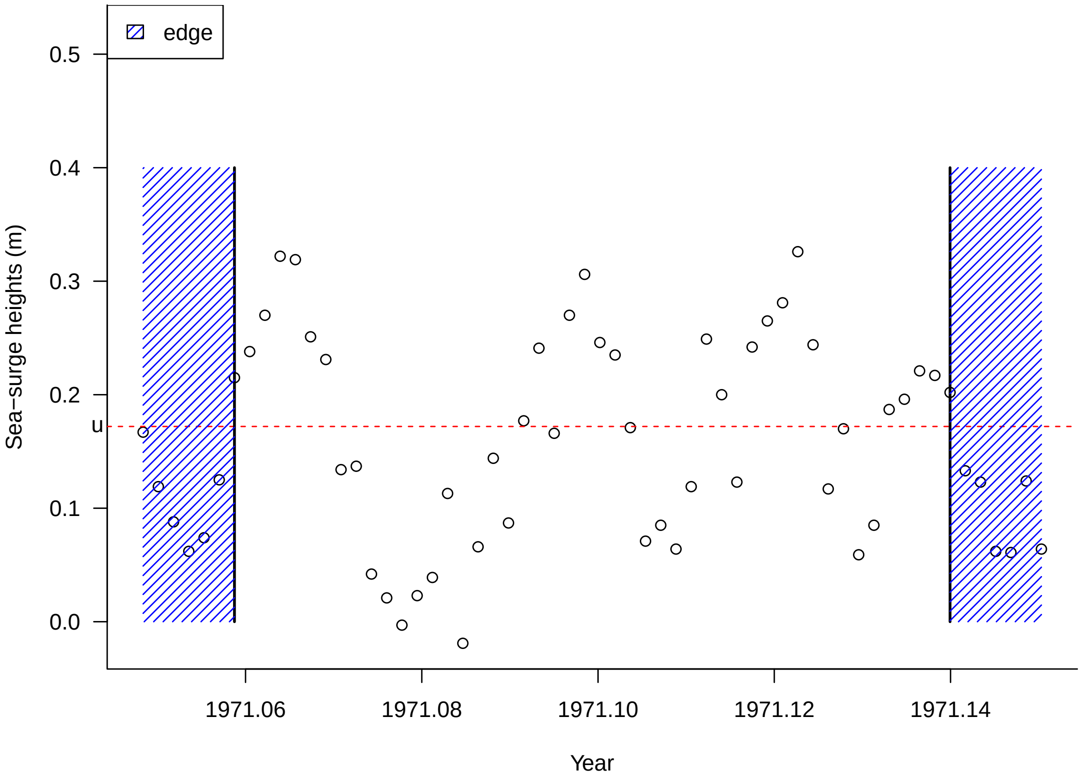

Figure 33: Time series plot of a segment of the Newlyn data illustrating the 'edges' (blue shaded area) at the two ends of the observation period.

Observations before the left hand blue region are not available so the the inter-exceedance time that ends with the first observed exceedance is right-censored: its value is not known but it is bounded below. For example, for the data in figure 33 this inter-exceedance time is no smaller than 6 time units. Similarly, the inter-exceedance time that starts with the last observed exceedance is right-censored. Moreover, depending on the positioning of the largest threshold exceedances, raising the threshold will tend to result in larger edges. Information from censored inter-exceedance times (and hence K-gaps) is not incorporated into the K-gaps likelihood used by Süveges and Davison (2010) for estimating θ and for performing the IMT. In the remainder of this section we consider how to adapt this likelihood to include information from censored K-gaps.

We begin by re-introducing the K-gaps exponential mixture model and stating the likelihood based on a random sample of uncensored K-gaps. Let us assume that we have a stationary process $X_1, X_2, \ldots, X_n$ with unknown marginal distribution function F. Let N be the number of exceedances above a (training) threshold u and the exceedance index $j_i : X_{j_i} > u$, where $i = 1, \ldots, N-1$. The ith inter-exceedance time T_i is given by $T_i = j_{i+1} - j_i$, and for a run parameter K, the i^{th} K-gap is $S_i^{(K)} = \max(T_i - K, 0)$. For notational convenience, from this point forwards we do not make explicit the dependence on K of the K-gaps or any other related quantities. We will only make explicit, when it is necessary, the dependence of K-gaps on the threshold applied to the data. Throughout, we plug-in the estimate $q = (1/n)\sum_{i=1}^{n} I(X_i > u)$ for $\overline{F}(u)$ and assume that K is sufficiently large that K-gaps can be treated as being mutually independent. Following equation (1.11) the likelihood is given by

$$L_K(\theta; S_1, \ldots, S_{N-1}) = (1-\theta)^{N_0}\theta^{2N_1} \exp\left\{-\theta q \sum_{i=1}^{N-1} S_i\right\}, \tag{4.1}$$

where $N_1 = \sum_{i=1}^{N-1} I(S_i > 0)$ is the number of non-zero K-gaps, $N_0 = \sum_{i=1}^{N-1} I(S_i = 0) = N - 1 - N_1$ is the number of zero K-gaps.

We now extend Süveges and Davison (2010) by taking into account information contained in the data about the inter-exceedance times T_0^u and T_N^u that would have been observed if the observation period was extended to include the last exceedance before the observation period started and the next exceedance after the observation period finished. As T_0^u and T_N^u are unknown, we consider their respective censored times, $T_0 = j_1 - 1$ and $T_N = n - j_N$. In other words, $T_0^u \geqslant T_0$ and $T_N^u \geqslant T_N$. Thus, for $i \in \{0, N\}$, the K-gap $S_i^u = \max(T_i^u - K, 0)$ satisfies $S_i^u \geqslant q \max(T_i - K, 0) = qS_i$. Under the limiting K-gaps model,

$$P(S_i^u \geqslant s) = \{\theta \exp(-\theta q s)\}^{I(s>0)}, \quad s \geqslant 0. \tag{4.2}$$

Thus, on incorporating information from the censored K-gaps S_0 and S_N, the likelihood becomes

$$L_K(\theta; S_0, \ldots, S_N) = (1-\theta)^{N_0}\theta^{2N_1+I_0+I_N} \exp\left\{-\theta q \left\{\sum_{i=0}^{N} S_i\right\}\right\}, \tag{4.3}$$

where $I_0 = I(S_0 > 0)$ and $I_N = I(S_N > 0)$.

4.3 The K-gaps maximum likelihood estimation of θ

If K is chosen so that the K-gaps model is well-specified then maximum likelihood estimators of θ can be based on likelihood (4.1) or likelihood (4.3). This approach combines elements of two existing methods of estimating θ: the *runs* estimator (see, for example, Smith and Weissman (1994)) and the likelihood-based estimator in Ferro and Segers (2003). The former is computed as the reciprocal of the mean cluster size after performing runs declustering (see section 1.6.2) with a certain run length. In the current setup this run length is K. Threshold exceedances separated by an inter-exceedance time that is greater than K are judged to be in different clusters; otherwise they are judged to be in the same cluster. The latter is based on a K-gaps likelihood with $K = 0$, for which all K-gaps are positive and therefore enter into the exponential part of the model. As Ferro and Segers (2003) point out, this likelihood is misspecified - in the likelihood the probability that two successive exceedances are in the same cluster is $1-\theta$ but for $K = 0$ this event is never observed in the data. Therefore, it is necessary to modify the likelihood using a positive K.

Let $\boldsymbol{S} = (S_0, \ldots, S_N)$. The log-likelihood based on (4.3) is

$$l_K(\theta; \boldsymbol{S}) = N_0 \log(1-\theta) + [2N_1 + I_0 + I_N] \log\theta - \theta q \left\{ \sum_{i=0}^{N} S_i \right\}. \tag{4.4}$$

Maximizing (4.4) with respect to θ gives the maximum likelihood estimator

$$\widehat{\theta} = -\frac{b}{2a} - \frac{1}{2a}\left(b^2 - 4ac\right)^{1/2}, \tag{4.5}$$

where $a = q\sum_{i=0}^{N} S_i, b = -(N_0 + 2N_1 + I_0 + I_N + a)$ and $c = 2N_1 + I_0 + I_N$.

If $N_0 = 0$, that is, all K-gaps are positive, then $\widehat{\theta} = 1$ as $l_K(\theta; \boldsymbol{S})$ is increasing over $[0, 1]$. If $N_1 = I_0 = I_N = 0$, that is, all K-gaps (or censored K-gaps) are zero then $\widehat{\theta} = 0$ as $l_K(\theta; \boldsymbol{S})$ is decreasing over $[0, 1]$. This corresponds to a degenerate case in which all exceedances are in a single cluster that covers the entire observation period. This could occur in practice if u is very low (so that a high proportion of the observations are exceedances) and/or K is high (so that quite well-separated exceedances are placed in the same cluster). Consider another scenario: that u is so high that all exceedances are in a single cluster surrounded by at least one non-zero censored K-gap. Incorporation of information from censored K-gaps means that the estimator based on (4.3) will give a value in $(0, 1]$, and probably close to 1. However, the estimator based on (4.1) will give an estimate of 0, because seemingly the single cluster covers the entire observation period. Although such extreme cases will probably be avoided in practice it is reassuring to have an estimator that behaves sensibly even in unusual cases. We should avoid using a very high threshold, because with very few exceedances an estimate of θ that is close to 1 will probably result. For such thresholds small distant clusters are likely to occur leading to a misleading estimate of an extremal index close to 1. In the extreme case with only one exceedance we have $\widehat{\theta} = 1$ by definition.

So there is a simple closed-form expression for the extremal index estimator $\widehat{\theta}$ and for inference using the estimative approach through maximum likelihood this is all that is required. However, similarly to the arguments put forward in previous chapters, the main drawback of the estimative approach is that once the model parameter (in this case θ) is estimated it is then treated as known, i.e. uncertainty in the value of the model parameter is not incorporated in the analysis. In contrast, under a predictive approach uncertainty in θ is incorporated explicitly in the analysis. We implement this using Bayesian inference.

4.4 Bayesian inference for the K-gaps mixture model

Suppose that a prior distribution $\pi(\theta)$ is specified for θ. Later, in the absence of a compelling reason to pick a different prior we follow Fawcett and Walshaw (2008) in supposing that *a priori* $\theta \sim U(0,1)$. Therefore, based on K-gaps $\boldsymbol{S} = (S_0, \ldots, S_N)$ the posterior distribution of the extremal index is given by

$$\pi(\theta \mid \boldsymbol{S}) = \frac{(1-\theta)^{N_0}\theta^{2N_1+I_0+I_N}\mathrm{e}^{-\theta V}\pi(\theta)}{\int_0^1 (1-\theta)^{N_0}\theta^{2N_1+I_0+I_N}\mathrm{e}^{-\theta V}\pi(\theta)\ \mathrm{d}\theta}, \tag{4.6}$$

where $V = q\sum_{i=0}^{N} S_i$.

Sampling from posterior (4.6) is easier than sampling from the GP posterior in section 2.4 because it is a 1-dimensional distribution with a simple support: $[0,1]$. Furthermore, under a U(0,1) prior for θ, $\pi(\theta \mid \boldsymbol{S})$ is log-concave, that is, the second derivative of $\log \pi(\theta \mid \boldsymbol{S})$ is negative for all θ (see appendix C.1). Therefore, we are able to use adaptive rejection sampling (ARS) (Gilks and Wild, 1992), an efficient method of sampling from a univariate distribution with a log-concave density function. ARS involves constructing an envelope function for the log density of the target distribution. This function is then used to carry out the conventional rejection sampling (see for example Ripley (1987)). However once a candidate point is rejected the function is updated to move closer to the log density of the target distribution. Consequently each update of the envelope function increases the probability of acceptance of the following candidate point. We implement ARS using R package “ars” (Rodriguez, 2014).

Having established our posterior sampling scheme, we move on to cross-validation which is the key element of our suggested methodology in analysing dependent extremes and more specifically tackling the task of selecting a threshold.

4.5 Threshold selection using cross-validation

The cross-validation methodology that we propose here is very similar to the one introduced in chapter 3. In the current chapter we are using a model for threshold inter-exceedance times, so the main modification is that the data used in carrying out cross-validation are the K-gaps instead of the indicators of threshold exceedances and sizes of threshold excesses. The effect of this is that when performing leave-one-out cross-validation the validation sample corresponds to the data associated with a single K-gap, rather than a single observation from the raw data series.

Suppose that we fix the values of a training threshold u and a run parameter K. The threshold u is applied to a sequence $X_1, \ldots, X_n$ of non-independent and identically distributed random variables, producing N_u exceedances and a vector of $N_u + 1$ K-gaps $\boldsymbol{S}^u = (S_0^u, S_1^u, \ldots, S_{N_u-1}^u, S_{N_u}^u)$. Recall that S_0^u and S_N^u are the values at which the first and last K-gaps are right-censored. As in chapter 3 u is the training threshold, using which inferences from the K-gaps model are made and $v \geqslant u$ is a validation threshold, at which these inferences are validated. Let $\boldsymbol{S}^v$ be the vector of $N_v + 1$ K-gaps based on the N_v threshold exceedances of v. In the interest of clarity we first describe the cross-validation procedure for the special case where $v = u$, as this is somewhat simpler than when $v > u$.

4.5.1 The case $v = u$

Although $v = u$ in this section, we preserve in our notation the distinction between the roles of v and u with a view to the $v > u$ case considered in the next section: u relates to training data and v relates to validation data. In the case when $v = u$ we have $\boldsymbol{S}^v = \boldsymbol{S}^u$ and $N_v = N_u$. We employ a leave-one-out cross-validation scheme in which, for $r = 0, \ldots, N$, $\boldsymbol{S}_{(r)}^u = \{S_i^u, i \neq r\}$ forms the training sample and S_r^v the validation sample. We use the same illustrative plot of a segment from the Newlyn dataset that was used in section 1.8 to show, in figure 34, two examples of validation sample. The blue shaded areas identify S_1^v (in the top plot) and S_7^v (in the bottom plot). Once a validation K-gap has been removed the remainder of the K-gaps constitute the training sample.

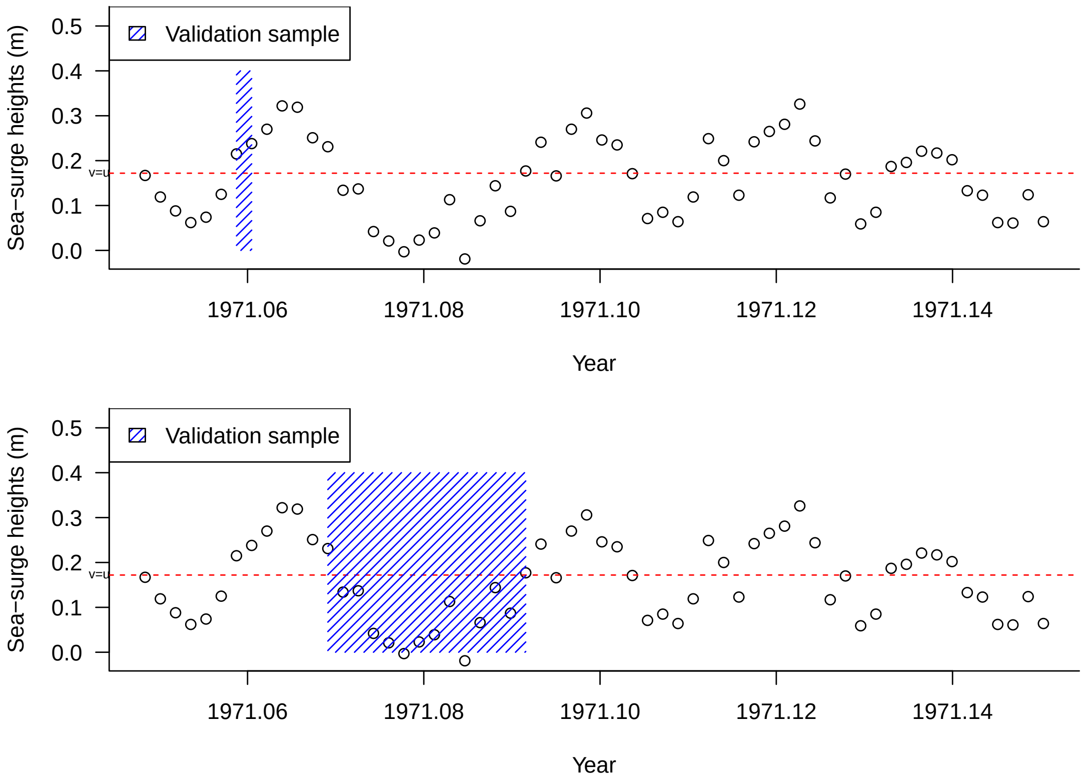

Figure 34: Time series plot of a segment of the Newlyn data illustrating two examples of validation samples denoted by the blue shaded area (using $v = u = 0.172$m and $K = 2$). Top: Validation sample is the 1^{st} K-gap S_1^v with value 0. Bottom: Validation sample is the 7^{th} K-gap S_7^v with value 11.

Let $\pi_u(\theta \mid \boldsymbol{S}^u_{(r)})$ denote the posterior density of θ based on training K-gaps $\boldsymbol{S}^u_{(r)}$. The *cross-validation predictive densities* at validation threshold $v(= u)$, based on a training threshold u, are given by

$$f_v(S_r^v \mid \boldsymbol{S}^u_{(r)}, u) = \int f(S_r^v \mid \theta, \boldsymbol{S}^u_{(r)})\, \pi_u(\theta \mid \boldsymbol{S}^u_{(r)})\, \mathrm{d}\theta, \quad r = 0, \ldots, N_v. \tag{4.7}$$

If the $\{S_i^u\}$ are conditionally independent given θ then the conditional density of S_r^v given $\theta, \boldsymbol{S}^u_{(r)}$ satisfies $f(S_r^v \mid \theta, \boldsymbol{S}^u_{(r)}) = f(S_r^v \mid \theta)$, where

$$f(S_r^v \mid \theta) \quad = \quad \begin{cases} (1-\theta)^{I(S_r^v=0)} \left\{\theta^2 \mathrm{e}^{-\theta q S_r^v}\right\}^{I(S_r^v>0)}, & \text{for } r = 1, \ldots, N_v - 1, \\ \left\{\theta \mathrm{e}^{-\theta q S_r^v}\right\}^{I(S_r^v>0)}, & \text{for } r \in \{0, N_v\}. \end{cases}$$

Suppose that we have a sample $\theta_j^{(r)}, j = 1, \ldots, n_\theta$ from the posterior $\pi_u(\theta \mid \boldsymbol{S}^u_{(r)})$.

Then a Monte Carlo estimator of $f_v(S_r^v \mid \boldsymbol{S}_{(r)}^u, u)$ based on (4.7) is given by

$$\tilde{f}_v(S_r^v \mid \boldsymbol{S}_{(r)}^u, u) = \frac{1}{n_\theta} \sum_{j=1}^{n_\theta} f(S_r^v \mid \theta_j^{(r)}, \boldsymbol{S}_{(r)}^u). \tag{4.8}$$

As in section 3.3.2 we seek to reduce computation time using importance sampling, with the full posterior $\pi_u(\theta \mid \boldsymbol{S}^u)$ as the importance sampling density $h(\theta)$. For a sample $\theta_j, j = 1, \ldots, n_\theta$ from $\pi_u(\theta \mid \boldsymbol{S}^u)$ this gives the estimator

$$\widehat{f}_v(S_r^v \mid \boldsymbol{S}_{(r)}^u, u) = \frac{\sum_{j=1}^{n_\theta} f(S_r^v \mid \theta_j) q_r(\theta_j)}{\sum_{j=1}^{n_\theta} q_r(\theta_j)}, \tag{4.9}$$

where $q_r(\theta) = \pi_u(\theta \mid \boldsymbol{S}_{(r)}^u)/\pi_u(\theta \mid \boldsymbol{S}^u) \propto 1/f(S_r^u \mid \theta)$.

Similarly to equation (3.4) we use

$$\widehat{T}_v(u) = \sum_{r=0}^{N_v} \log \widehat{f}_v(S_r^v \mid \boldsymbol{S}_{(r)}^u, u), \tag{4.10}$$

as a measure of predictive performance at validation threshold v when using training threshold u.

4.5.2 The case $v > u$

We exclude the event that v and u are so close that their exceedances coincide exactly. Therefore, we have fewer exceedances of v than u, that is, $N_v \leqslant N_u$, with the consequence that $\boldsymbol{S}^v \neq \boldsymbol{S}^u$. The general setup of the cross-validation scheme is the same as in the $v = u$ case. However, since different training thresholds are to be compared based on their performance at a validation threshold v, it is the $N_v + 1$ K-gaps $\boldsymbol{S}_{(r)}^v = \{S_i^v, i \neq r\}$ produced by applying the threshold v that define the leave-one-out validation K-gaps.

Let $\boldsymbol{T}^v = \{T_i^v, i = 0, \ldots, N_v\}$ and $\boldsymbol{T}^u = \{T_i^u, i = 0, \ldots, N_u\}$ denote the inter-exceedance times based on thresholds v and u respectively. Each component of $\boldsymbol{T}^v$ is equal to either a member of $\boldsymbol{T}^u$ or the sum of two or more members of $\boldsymbol{T}^u$. Therefore, a given K-gap S_r^v corresponds to a set $\{S_i, i \in r^*\}$ of contiguous K-gaps from $\boldsymbol{S}^u$, where r^* is a subset of $\{0, \ldots, N_u\}$. If S_r^v is taken as a leave-one-out validation sample and the model is trained at threshold $u < v$ then the K-gaps $\{S_i, i \in r^*\}$ should be removed from the full sample of K-gaps. Let $\boldsymbol{S}_{(r^*)}^u$ denote the subset of $\boldsymbol{S}^u$ that remains once $\{S_i, i \in r^*\}$ are removed from the training sample.

We illustrate this in figure 35 where the blue shaded area indicates the validation sample: a single K-gap S_6^v, and the red shaded area indicates the three K-gaps S_{11}^u, S_{12}^u and S_{13}^u that are removed to produce the training sample, so that $r^* = (11, 12, 13)$.

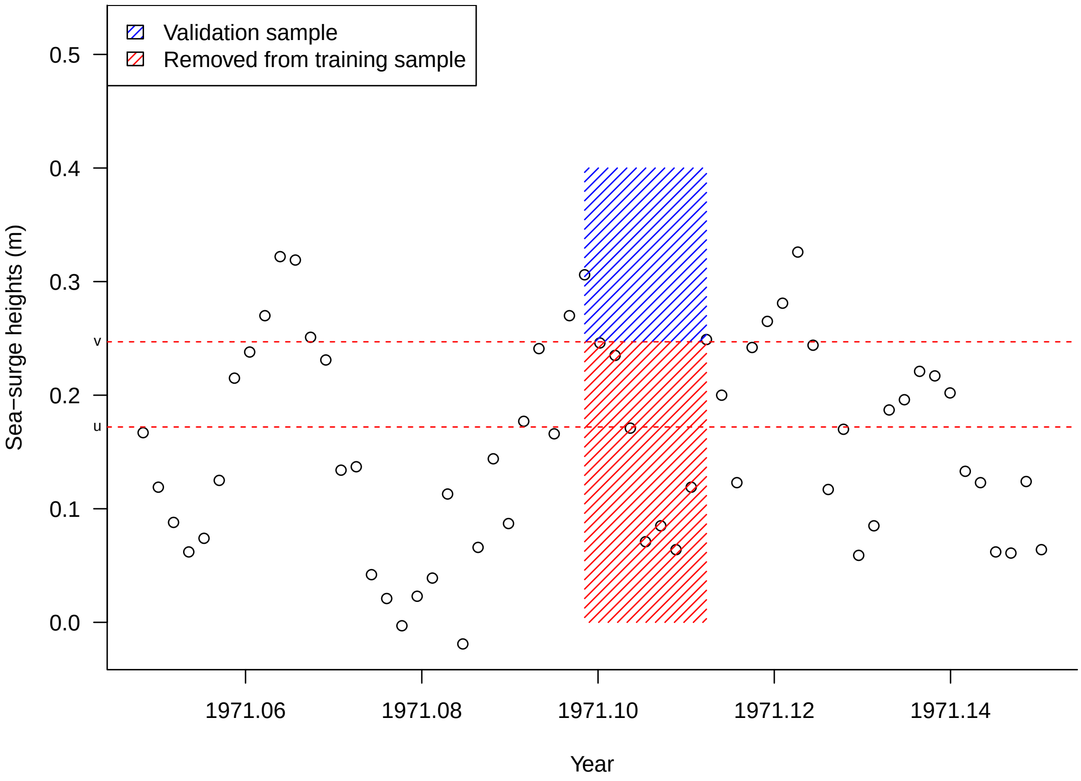

Figure 35: Time series plot of a segment of the Newlyn data illustrating an example of validation sample when $v > u$ denoted by the blue shaded gap. The red shaded area denotes the gaps removed from the training sample.

The equations in section 4.5.1 carry over with only small modification: (r) is replaced with (r^*) throughout and now $q_r(\theta) = \pi_u(\theta \mid \boldsymbol{S}_{(r^*)}^u)/\pi_u(\theta \mid \boldsymbol{S}^u) \propto 1/\prod_{i \in r^*} f(S_i^u \mid \theta)$. We find that the results based on the importance sampling estimator (4.9) differ from those based on the 'brute-force' estimator (4.8) by negligible amounts so we use the former estimator throughout.

4.5.3 Comparing training thresholds

Suppose that for a fixed value of the run parameter K, we consider k training thresholds $u_1 < \cdots < u_k$, resulting in a set of estimates $\widehat{T}_v(u_1), \ldots, \widehat{T}_v(u_k)$, and that we wish to select one of these thresholds. We follow the arguments in section 3.3.3 to choose $u^* = \arg\max_u \widehat{T}_v(u)$. In common with chapter 3 we aim to set $(u_1, \ldots, u_k)$ to cover a range of thresholds over which the bias-variance trade-off seems to be occurring, and to avoid thresholds with small numbers of exceedances. We also standardize the CV measure via (3.16) to produce threshold weights. Following Süveges and Davison (2010) and Fukutome et al. (2014) a rule-of-thumb could be to have no fewer than 80 exceedances of u_k, although this value is chosen based specifically on the their model misspecification test.

In common with chapter 3 we set $v = u_k$. However, the argument for doing this in the context of chapter 3 does not carry over. Firstly, changing the value of v changes the validation data $\boldsymbol{S}^v$ in a way that is more complicated than a simple change of threshold applied to each validation observation. Also, changing v changes the training data $\boldsymbol{S}^u_{(r^*)}$: the higher v is the more K-gaps are removed from $\boldsymbol{S}^u$. Here, we set $v = u_k$ on the grounds that increasing v beyond u_k results in a reduction in validation information.

In the next section we implement our method for dependent extremes using the Newlyn dataset described in section 1.8. The aim is to identify the level of threshold with the 'best' out of sample predictive performance.

4.6 Newlyn

We begin our analysis of the Newlyn dataset with a preliminary graphical investigation of the behaviour of the MLE $\widehat{\theta}$ (equation (4.5)) as we (a) keep the threshold u fixed and vary K, and (b) keep K fixed and vary u. Fawcett and Walshaw (2012) used a mean residual life plot to select a threshold of 0.3m (approximately the 94% sample quantile) to use for modelling threshold excesses with a GP distribution. This threshold was selected with reference to a different aspect of the data than is our focus in this chapter, that is, in reference to a GP distribution for excesses rather than the K-gaps model for extremal dependence. However, it is interesting to see how well a threshold of 0.3m is supported by the data in terms of the latter aspect.

As we are focusing on selection of u for fixed K, the first stage of our analysis involves deciding on the value of K. The parameter stability plots for θ in figures 36 and 37 help us in this decision by illustrating the behaviour of the extremal index estimate as the value of K changes. We have done this for eight different thresholds, more specifically, for the 60% - 95% sample quantiles in steps of 5%. The idea is the same as the parameter stability plots for ξ considered in chapter 3: for a given value of u we look for the smallest value of K above which the estimates of θ are approximately stable. These plots suggest that $K = 6$ is a reasonable choice across all the thresholds considered as all plots seems to show a somewhat stable parameter estimate beyond that point. Since this precise choice is somewhat arbitrary later in this section we investigate informally sensitivity to it by performing analyses with $K = 5$ and $K = 7$.

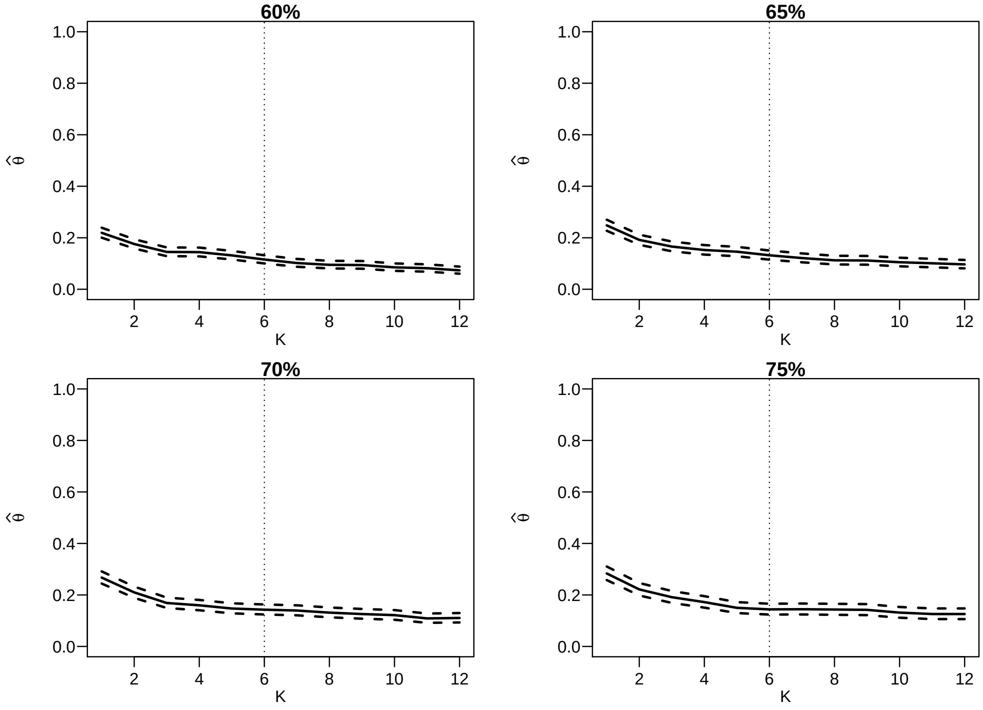

Figure 36: Parameter stability plots for Newlyn sea-surge heights for a range of run parameter K and using the 60%, 65%, 70% and 75% sample quantile for thresholds. The solid lines give the MLEs of θ and the dashed lines give 95% likelihood-based confidence intervals.

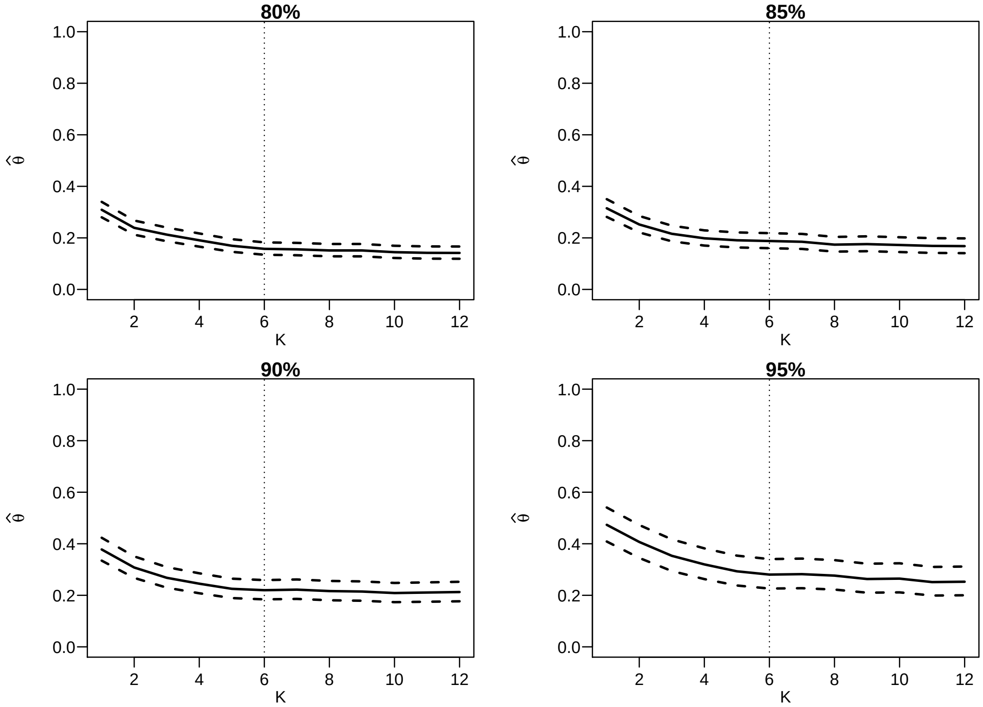

Figure 37: Parameter stability plots for Newlyn sea-surge heights for a range of run parameter K and using the 80%, 85%, 90% and 95% sample quantile for thresholds. The solid lines give the MLEs of θ and the dashed lines give 95% likelihood-based confidence intervals.

We now fix the value of $K = 6$ according to our conclusion based on the figures above and illustrate in figure 38 the parameter stability plot for $\widehat{\theta}$ as we vary the threshold. This is not done to decide on the level of threshold but it can be very helpful in determining the plausible range of training thresholds that we need to use for our proposed cross-validation approach. Note that for very low thresholds close to the sample minimum, estimates of θ close to, or equal to, zero are obtained by definition, because all, or almost all, the sample K-gaps are zero (see section 4.3).

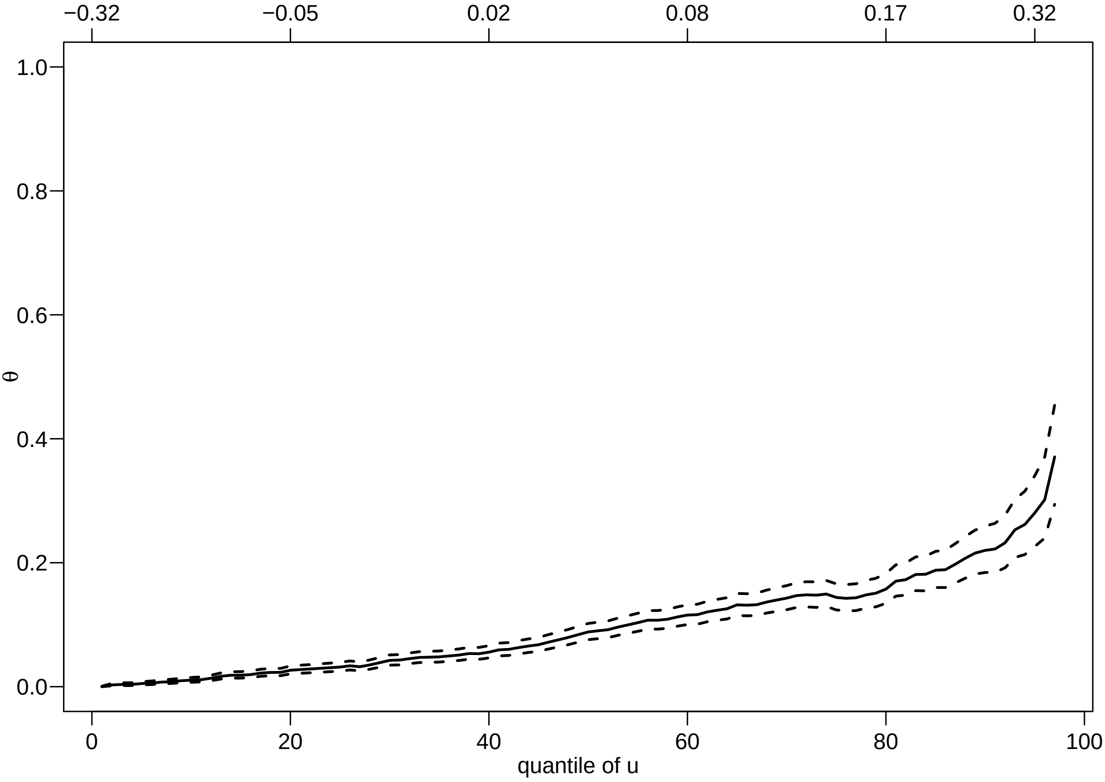

Figure 38: Parameter stability plot for Newlyn sea-surge heights for a range of thresholds. The solid lines give the MLEs of θ and the dashed lines give 95% likelihood-based confidence intervals.

The conclusion from figure 38 is that a range of training thresholds at the 50% - 95% sample quantiles is plausible for our analysis, although we anticipate from the general increase in $\widehat{\theta}$ over this range that thresholds at the lower end of this scale will perform less well than higher thresholds. Note that the 95% sample quantile produces 144 threshold exceedances which is well above the Süveges and Davison (2010) rule-of-thumb of 80 exceedances. Extending the range to an even higher threshold of say the 99% sample quantile results in only 29 exceedances and the misleading behaviour anticipated in section 4.3, that is, estimates of θ that increase sharply towards 1.

We proceed with our analysis, as outlined in section 4.5, of the Newlyn dataset using $K = 6$ for k training thresholds $u_1, \ldots, u_k$ such that u_1 and u_k are the 50% and 95% sample quantile respectively and therefore a validation threshold set at u_k. Furthermore, we employ the IMT statistic (see section 4.1) on this range of thresholds for comparison, using, in the first instance, a significance level of 5%. In figure 39 we show the results from this analysis. It is interesting that for this dataset the two approaches essentially give two different answers. On one hand the IMT statistic of Süveges and Davison (2010) would suggest setting a threshold around the 60% sample quantile as this is the point where the misspecification test results in non-rejection. However, our proposed cross-validation method suggests that the lower thresholds perform relatively poorly in terms of out-of-sample predictions and instead suggests selecting u at a much higher level where the 'best' training threshold is identified at the 93% sample quantile. If the significance level of the IMT were increased to say 10%, with critical value 2.71, then the IMT statistic would suggest both the lower threshold already identified and a higher threshold (at 94% sample quantile) as locations where the misspecification test turns to non-rejection.

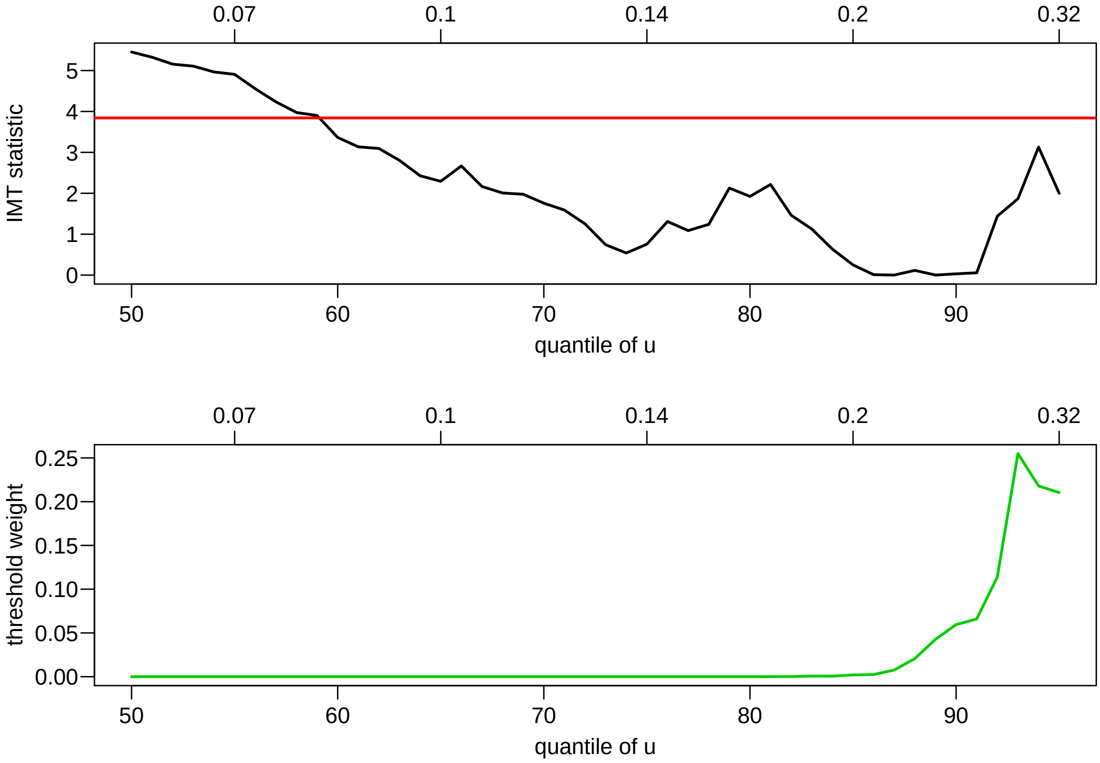

Figure 39: Threshold selection methods for Newlyn sea-surge heights for a range of thresholds between the 50%-95% sample quantiles. Top panel: IMT statistic (the red line represents the 5% significance level critical value of 3.84). Bottom panel: threshold weight for validation threshold at 95% sample quantile.

We should not be surprised that there is not a close correspondence between these two approaches. Firstly, their results are each influenced by somewhat arbitrary choices: the IMT significance level and the value of the highest threshold u_k. Secondly, they are setup for different purposes. The IMT is designed to identify thresholds for which the K-gaps model is not misspecified. It is possible that the null hypothesis that the K-gaps model is well-specified is not rejected over a given range of thresholds, but the estimated strength of extremal dependence is changing noticeably, as seems to be the case in figure 38. Our cross-validatory assessment is not concerned with whether the K-gaps model is well-specified, but with identifying the thresholds for which out-of-sample prediction of K-gaps is better than other thresholds.

In figure 40 we illustrate how the estimated threshold weights vary with the choice of the highest training threshold. There is little sensitivity to this choice until we get to the 97% sample quantile, for which the highest thresholds have much larger estimated threshold weights than the lower thresholds. There are 87 exceedances of this threshold and, as figure 38 shows, the MLE of θ exhibits a sharp rise towards 1 at approximately this threshold. These observations suggest that the 97% sample quantile is perhaps too high to be used as a validation threshold. Additionally, we repeated our cross-validation approach using $K = 5$ and $K = 7$ and produced very similar results (plots not shown).

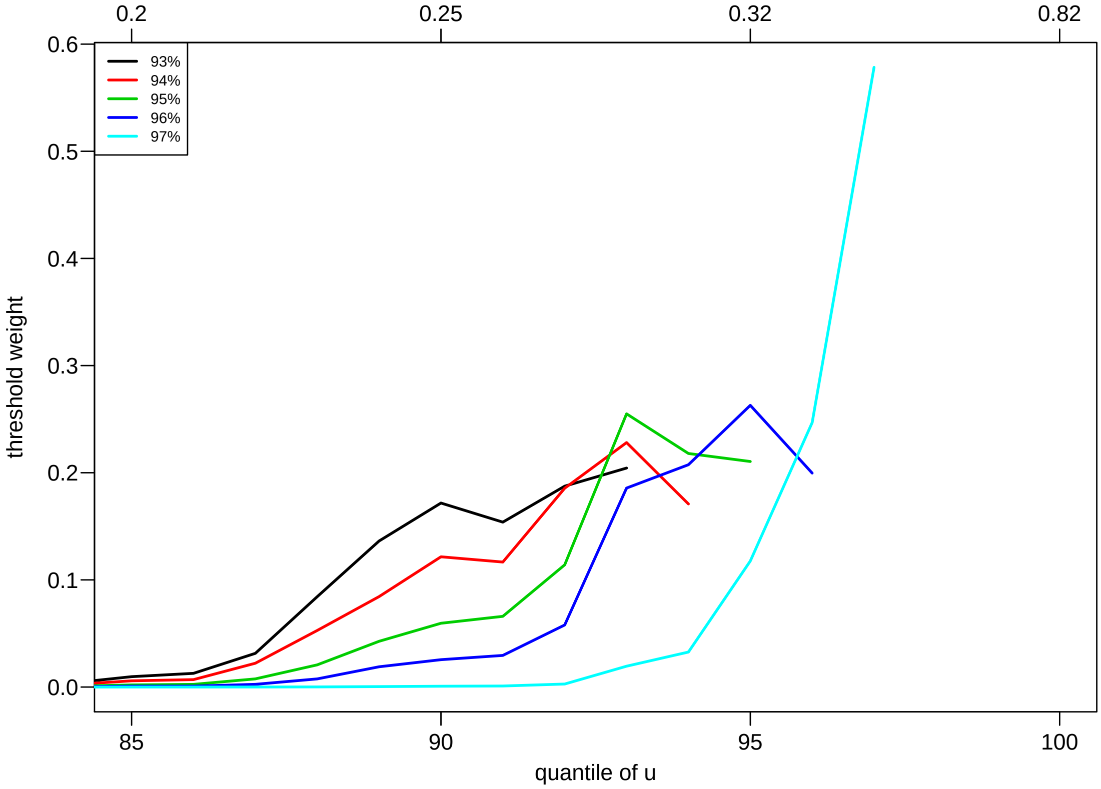

Figure 40: Threshold selection methods for Newlyn sea-surge heights for a range of thresholds between the 85%-97% sample quantiles. Threshold weights for validation thresholds at 93%, 94%, 95%, 96% and 97% sample quantiles.

In the following section we study the behaviour of our proposed threshold selection method, the IMT statistic and the MLE of θ on repeated samples from some example stationary processes.

4.7 Simulation study

This study is based on simulated realisations from the stationary processes used by Süveges and Davison (2010) to illustrate the use of the IMT to choose a threshold and run parameter pair (u, K) for which the K-gaps model is not rejected. The three processes are:

1. an AR(1): $Y_i = \phi Y_{i-1} + Z_i$, with $\phi = 0.7$ and Z_i standard Cauchy and $\theta = 0.3$.
2. an AR(2): $Y_i = \phi_1 Y_{i-1} + \phi_2 Y_{i-2} + Z_i$, with $\phi_1 = 0.95$, $\phi_2 = -0.89$ and Z_i Pareto, with tail index 2 and $\theta = 0.25$.
3. a Markov chain: with Gumbel margins, a symmetric logistic bivariate distribution for consecutive variables and dependence parameter $r = 2$ (Smith, 1992), with $\theta \approx 0.33$.

Süveges (2008) considered the structure of these processes at extreme levels and determined the appropriate run parameter values to be $K = 1$ for the AR(1) process and $K = 6$ for the AR(2) process. This was not possible for the Markov chain so the value of $K = 5$ was suggested from the misspecification tests. Süveges and Davison (2010) use the same values in their work and therefore for each of these processes we generate simulation runs of a sequence of $n = 8,000$ observations and set the above-mentioned values for the run parameter K.

For each process we investigate how the MLE of the extremal index, the IMT statistic and the CV estimated threshold weights behave across a range of thresholds. As before the validation threshold is set at the highest (95%) threshold. The sample size of 8,000 is quite large - with a threshold set at the 95% sample quantile there are 400 exceedances - but we will see that even for relatively low thresholds there is appreciable variability between different simulated datasets. To study how greatly the results vary across different simulated realisations we replicated the simulation at first 5 times (for illustration purposes) and later 100 times to give a fuller appreciation of the effects of sampling variability. In the former case we use training thresholds from 50% to the 95% sample quantiles in steps of 2.5%.

AR(1) process

We can see in the top panel of figure 41 that on average the MLEs of the extremal index are very close to the true value of 0.3 for thresholds around the 90% sample quantile. There is greater variability in the MLEs of θ at the highest thresholds, for which there are smaller numbers of exceedances. The middle panel of figure 41 is a plot of the IMT statistic against threshold. In general the lines behave as expected: the IMT statistics tend to reduce as the threshold is increased. However, variability between the replications results in substantial variability in the point where the test statistics cross the critical values, and two of them cross at in more than one place. The bottom panel of figure 41 is a plot of the threshold weights against threshold. Again, there is appreciable variability in the shape of the plotted curves and the location of the largest threshold weight. For three of the five simulated datasets high weight is given to thresholds near the 90% sample quantile. The light blue curve gives greatest weight near the 80% threshold and the red curve gives greatest weight to much lower thresholds.

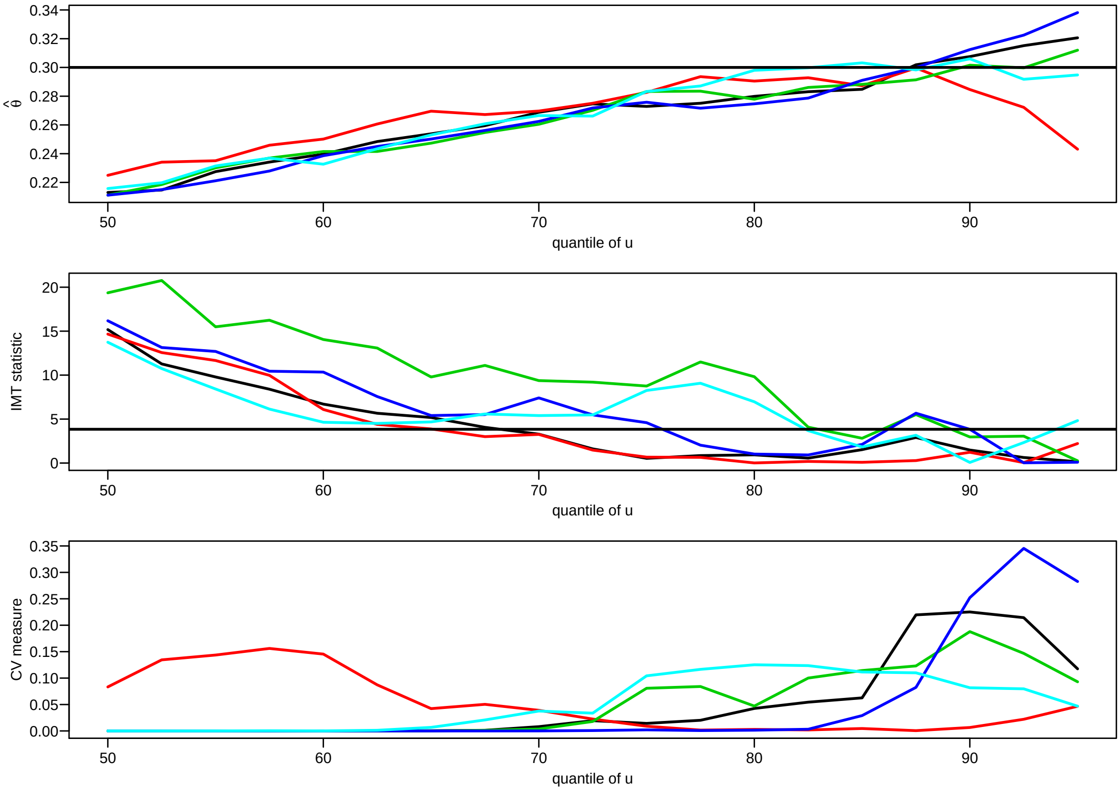

Figure 41: Threshold selection for AR(1) process with fixed $K = 1$. Top panel: MLE against threshold quantile, with horizontal line at the true value of θ. Middle panel: IMT statistic against threshold quantile, with horizontal line at the critical value for a test with significance level of 5%. Bottom panel: threshold weight against threshold quantile. Each coloured line represents a replication of the simulation.

The general behaviour of the threshold weights can be explained by comparing the top and bottom panels of figure 41. Firstly, consider the dark blue curve in the top panel, for which the MLE of θ continues to increase with threshold once it has passed through the true value of 0.3. So, at the validation threshold the K-gaps data produce a point estimate of θ that is greater than at any lower threshold. Therefore, it is not surprising that relatively high thresholds achieve the greater threshold weights than lower thresholds. Interestingly, the 92.5% training threshold achieves the greatest threshold weight, the result of trading increased bias for reduced variance. For this simulation run the IMT first dips below the critical value at the 77.5% threshold and then crosses the critical value again at the 90% threshold.

Now consider the red curve in the top panel. After passing through the true value of θ the MLE decreases as the threshold increases. Now, at the validation threshold the K-gaps data produce a point estimate of θ that is similar to the point estimate near the 55% training threshold. Consequently, thresholds near the 55% threshold achieve high threshold weights. For this simulation run the IMT crosses the critical value at the 65% threshold and remains below it from there on.

In common with any simulation study we are in the artificial position of knowing the true value of the parameter: $\theta = 0.3$. Using a significance level of 5% and choosing a threshold where the IMT crosses the critical values tends to result in underestimation of θ in four of the five cases (the light blue curve is the exception). Using a higher significance level would reduce the underestimation but we wouldn't know this in practice when θ is unknown. Apart from the red curve, which results in underestimation of θ, the high threshold weights tend to correspond to point estimates of θ that are perhaps closer to the true value of θ, although we shouldn't read too much into this.

The results for the AR(2) process and the Markov chain exhibit some features that are similar to the results for the AR(1) process, so in the following we comment only on aspects that are different.

AR(2) process

Firstly, we note that in the first plot of figure 42 (top panel) when the threshold is set at the sample median the MLEs of the extremal index are very close to zero suggesting very strong dependence. The IMT statistic shown in the middle panel of figure 42 displays an interesting, and potentially misleading, feature. It would seem that for all replications a low threshold near the sample median would result in non-rejection of the misspecification test. This could mislead someone in selecting a threshold at that level which would greatly underestimate $\widehat{\theta}$ as well as suggest that there is very strong dependence in the underlying process.

The reason for this feature is that with a low threshold and $K = 6$ most of the K-gaps are zero, that is, clusters tend to contain large numbers of exceedances. This means that the sample K-gaps are predominantly from the point mass at zero part of the K-gap mixture model with few sample K-gaps from the exponential part of the model. The IMT statistic is formulated to test the K-gaps model by assessing departure from the random split between zero and non-zero K-gaps and the exponential distribution for non-zero K-gaps, both of which are controlled by the same parameter: θ. With very few non-zero K-gaps we would expect this test to have low power to detect departure and this seems to be the case. Consider the extreme case where all K-gaps are zero. Inspection of the expression for the IMT statistic in the appendix of Süveges and Davison (2010) shows that in this event the IMT statistic is identically zero.

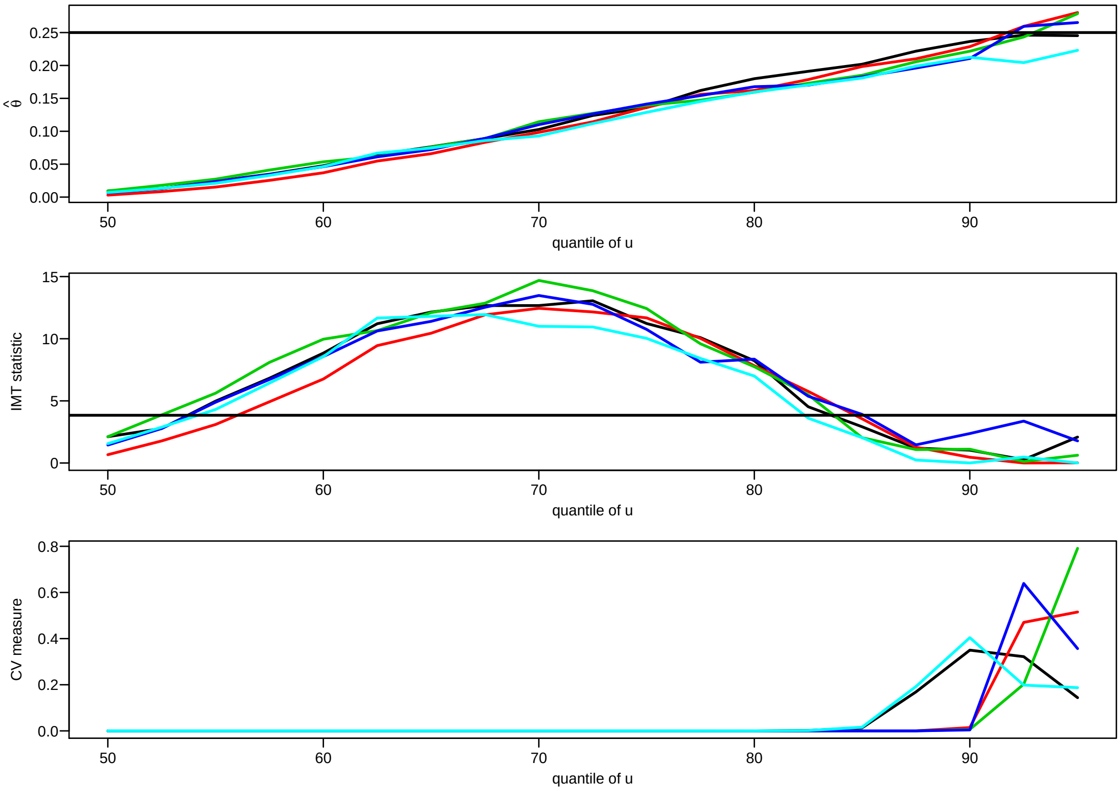

Figure 42: Threshold selection for AR(2) process with fixed $K = 6$. Top panel: MLE against threshold quantile, with horizontal line at the true value of θ. Middle panel: IMT statistic against threshold quantile, with horizontal line at the critical value for a test with significance level of 5%. Bottom panel: threshold weight against threshold quantile. Each coloured line represents a replication of the simulation.

As the threshold increases, a different story is told. The IMT is rejected until a threshold near the 85% sample quantile. If one were to dismiss the suggested low thresholds near the sample median (or not consider them in the first place) and select a threshold near the 85% quantile this would still tend to result in underestimation of the extremal index, albeit less serious underestimation.

In the bottom panel of figure 42 we see that the estimated threshold weights suggest thresholds that are somewhat higher than those suggested by the IMT and that thresholds below the 85% quantile achieve virtually no weight. The thresholds with high weight correspond to estimates of θ that are closer to the true value of θ than the thresholds suggested by the IMT. By this judgement the cross-validatory threshold selection seems to perform better than the IMT. However, increasing the significance level used for the IMT would improve its performance in this respect.

Markov chain (MC)

The results here are similar to those for the AR(2). The main difference is that the IMT behaves as one would hope: it decreases as the threshold increases. Otherwise, the IMT tends to suggest a slightly lower threshold than in the AR(2), approximately the 80% quantile rather than the 85% quantile, and the profiles of estimated threshold weights also seem to shift down by 5 percentage points in the level of the quantile.

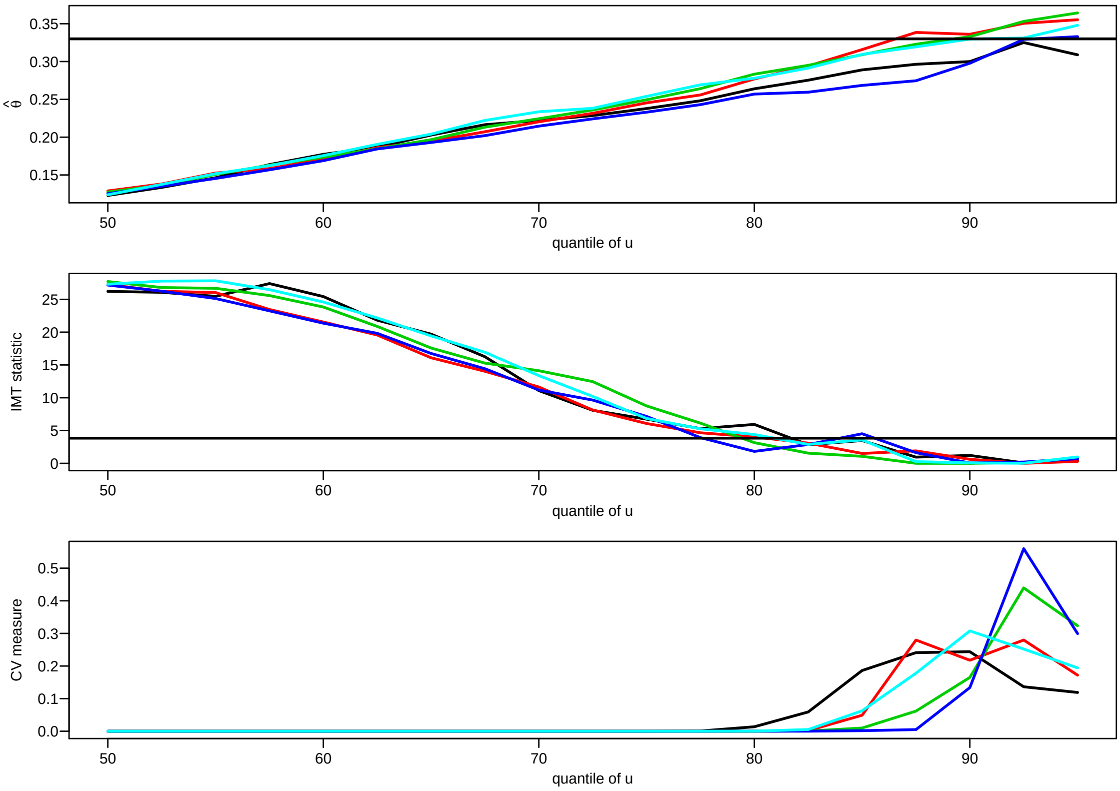

Figure 43: Threshold selection for Markov chain with fixed $K = 5$. Top panel: MLE against threshold quantile, with horizontal line at the true value of θ. Middle panel: IMT statistic against threshold quantile, with horizontal line at the critical value for a test with significance level of 5%. Bottom panel: threshold weight against threshold quantile. Each coloured line represents a replication of the simulation.

We now continue by repeating our analysis using 100 replications. With a larger number of replications we can gain a clearer picture of the properties of the threshold selection methods using cross-validation and using the IMT and of the MLE of θ. We truncate the range of thresholds for the AR(2) and Markov chain examples to the 75% to 95% sample quantiles because of the negligible threshold weight achieved by thresholds lower than the 75% quantile.

Figure 44 shows the results for the cross-validation method. There is a lot of variability between the replications (shown in grey lines), however the summary statistic lines make it easier to draw general conclusions. For the AR(1) the 85% sample quantile has the highest average threshold weight and is selected as the 'best' threshold most frequently. For the AR(2) process, the 92.5% and 95% sample quantiles perform well, as do the 90%, 92.5% and 95% quantiles for the Markov chain. The thresholds achieving high weight are somewhat higher than the thresholds at which the average value of the IMT crosses the 5% significance level critical value (see figure 45).

Figure 46 allows us to investigate what values for $\widehat{\theta}$ would be obtained for each replication for a range of training thresholds. It is interesting that for the AR(1) process, on average, θ is underestimated for all training thresholds considered. However, the threshold suggested by our CV approach would result to smaller underestimation error compared to the one produced by the lower threshold suggested from the IMT approach. It is reassuring to note that for the AR(2) process, on average our method suggests the threshold that estimates the model parameter very close to the true value, whereas the IMT approach would tend to underestimate it. Finally for the Markov chain, again the range of thresholds considered would tend to underestimate, on average, the extremal index. However, the underestimation error is minimised for the highest thresholds which is what our CV approach is suggesting. This compares positively against the IMT approach that would produce a much larger underestimation error.

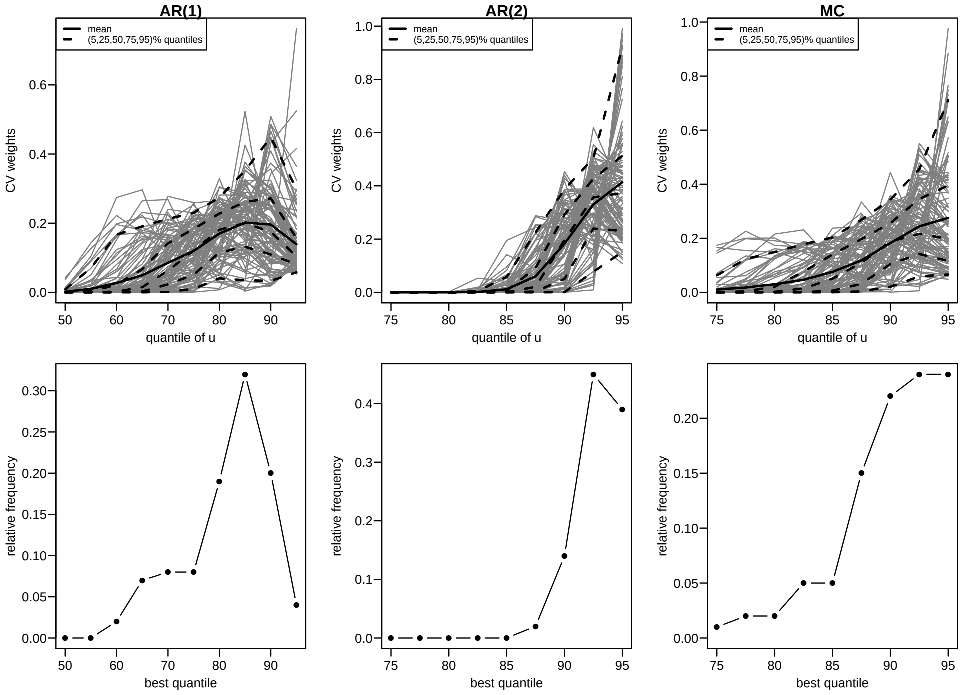

Figure 44: Summaries of threshold weights by training threshold. Top: the grey lines give individual lines for each simulated dataset with threshold-specific sample means (solid black line) and sample (5, 25, 50, 75, 95)% quantiles (dashed black lines). Bottom: relative frequency with which each threshold has the largest CV weight. Left: AR(1) process. Middle: AR(2) process. Right: MC process.

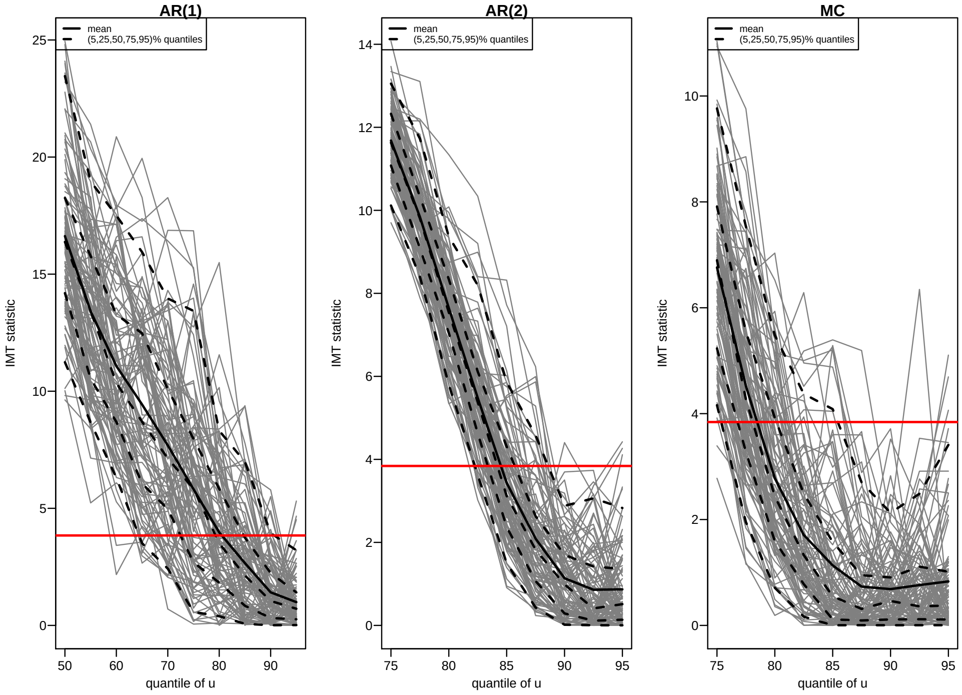

Figure 45: Summaries of IMT statistics by training threshold. The grey lines give individual lines for each simulated dataset with threshold-specific sample means (solid black line) and sample (5, 25, 50, 75, 95)% quantiles (dashed black lines). The horizontal red line is at the critical value for a test with significance level of 5%. Left: AR(1) process. Middle: AR(2) process. Right: MC process.

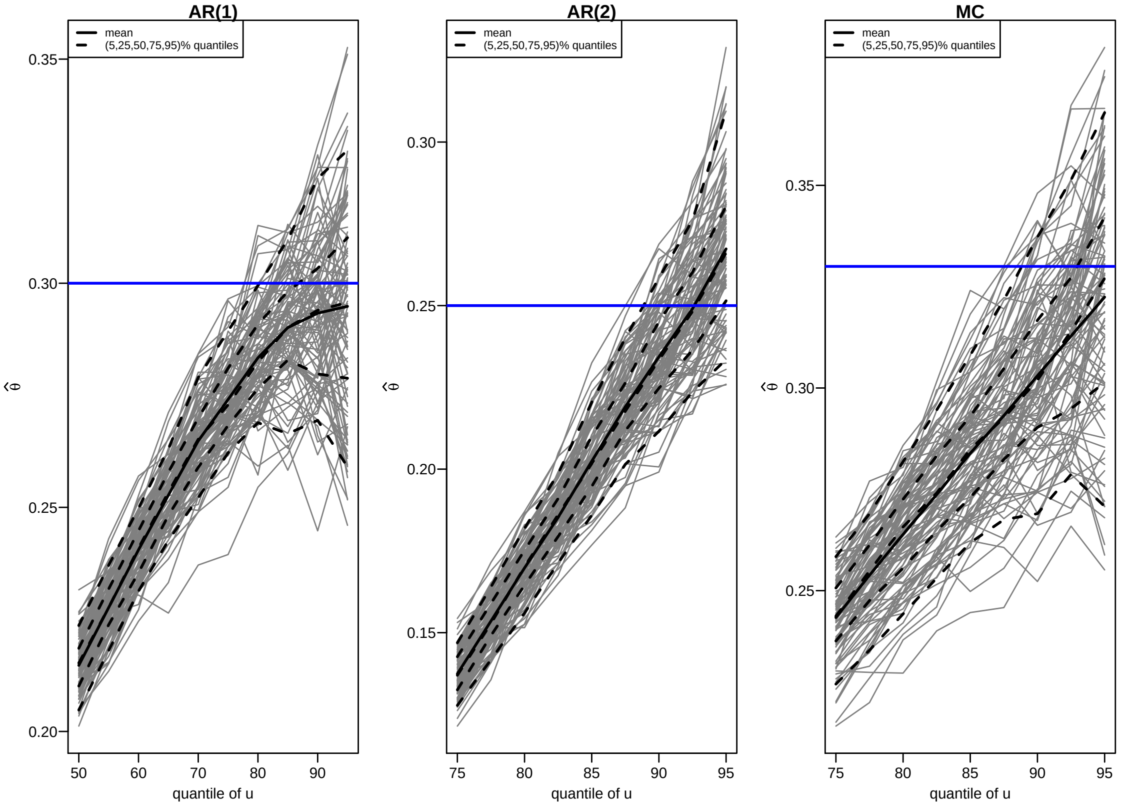

Figure 46: Summaries of MLEs by training threshold. The grey lines give individual lines for each simulated dataset with threshold-specific sample means (solid black line) and sample (5, 25, 50, 75, 95)% quantiles (dashed black lines). Left: AR(1) process. Middle: AR(2) process. Right: MC process.

It is natural when dealing with real datasets, such as the Newlyn sea-surge heights, that there will be some degree of uncertainty regarding threshold selection and in making inferences about the extremal index. In this section we have compared the performances of the IMT and the CV approach in terms of the resulting MLE of θ. In this regard the CV approach performed better than the IMT for the three processes we have considered. However, this is based on arbitrary choices of the significance level used for the IMT and the highest threshold considered in the CV approach. The fact that the methods rely on different choices means that we should not read too much in to this result.

These methods also have different goals: the IMT examines model misspecification and the CV approach assesses out-of-sample prediction. It is our contention that the CV approach tackles more directly the bias-variance trade-off involved in threshold selection. We know that the K-gaps model is at best an approximation to the truth. Therefore, we prefer the idea of basing threshold selection on out-of-sample predictive ability rather than a test of whether a model that is known to be wrong fails to reject that model. The main challenge in implementing the CV approach is the choice of the highest training threshold (at which the validation threshold is set). Although there is no definitive way to decide this, the same general considerations from chapter 3 apply: this threshold should be high enough to observe the bias-variance trade-off in action, but low enough that the adverse consequences of having too few threshold exceedances is avoided.

We continue in the next chapter by returning to the independent case for observations but this time we relax the ‘identically distributed’ assumption by introducing covariate effects in the location parameter.

5 Thresholds for non-stationary extremes

We have introduced earlier in chapter 1 the GEV and GP models that are commonly used in extreme value modelling. This chapter describes an approach in which the extremes of a sequence of independent random variables are represented by a point process. The resulting non-homogeneous point process model (which we refer to as the NHPP model) is a combination of the GP and GEV modelling approaches in the sense that (a) the number and extent of threshold excesses are modelled, and (b) the model is parameterised in terms of the GEV parameters μ, σ and ξ. Thus, in contrast with the GP model, the parameters of the NHPP model do not depend on the threshold. This has the advantage that non-stationary models with covariate effects can be included more naturally in the NHPP model than in the GP model, a point we return to in section 5.2.

We begin by briefly introducing the concept of a point process and the two-dimensional non-homogeneous Poisson process model. We then relax the assumption that the data are realised values of identically distributed random variables, by introducing a covariate effect only in the location parameter of the model, in other words, we investigate a sequence of extremes with a linear trend. Furthermore, we discuss current methods that deal with the topic of non-stationary extremes, focusing on threshold modelling, and point out their strengths and weaknesses. We later suggest the type of thresholds to use for non-stationary extremes where our aim is to use regression modelling to identify (a method for selecting) the optimal threshold. We conclude this chapter by discussing the limitations of our approach and outlining possible extensions.

5.1 Point processes and the NHPP model

A point process model can be used, for example, to describe the process by which point events, such as earthquakes, occur in time. The books by Resnick (1987) and Snyder and Miller (1991) offer themselves as good reference points for a more detail study on this topic.

The simplest point process is the one-dimensional homogeneous Poisson process, in which points occur randomly in, say, time. This process has the property that the number of points that occur in a time interval has a Poisson distribution, with a mean that is proportional to the length of the time interval. Let $N(s_1, s_2)$ be the number of points that occur in the time interval $(s_1, s_2]$. Then, for any $s_1 < s_2$, $N(s_1, s_2)$ has a Poisson distribution with mean $\lambda(s_2 - s_1)$, where λ is the *rate* or *intensity* of the Poisson process. In addition, the number of points in disjoint time intervals are independent, that is, for any disjoint intervals (s_1, t_1) and (s_2, t_2), $N(s_1, t_1)$ and $N(s_2, t_2)$ are independent. A process satisfying these two properties is a Poisson process and the homogeneity results from λ being constant.

Allowing the intensity λ to be a function of time t, i.e. $\lambda(t)$, results in a one-dimensional *non-homogeneous* Poisson process. This can be extended to higher dimensions, i.e. having points occurring in d-dimensional space. Let $\mathcal{X} \subset \mathbb{R}^d$, $A \subset \mathcal{X}$ and let $N(A)$ be the number of points in A. A d-dimensional process is a non-homogeneous Poisson process $\mathcal{P}$ on $\mathcal{X}$ with non-negative intensity measure Λ if:

- $N(A) \sim \text{Poisson}(\Lambda(A))$ for all $A \subset \mathcal{X}$;
- if A and B are disjoint sets then $N(A)$ and $N(B)$ are independent.

Thus the expected number of points in A is given by the *intensity measure* $\Lambda(A)$, which is related to the *intensity (density) function* $\lambda(v)$ by

$$\Lambda(A) = \int_A \lambda(v) \, \mathrm{d}v. \tag{5.1}$$

From this point onwards we focus on a specific Poisson process and therefore we are only interested in the two-dimensional non-homogeneous Poisson process defined as $\mathcal{P}$.

Recall that in chapters 2 and 3 we assumed that $X_1, \ldots, X_n$ is a sequence of i.i.d. random variables with unknown distribution function F. We define a sequence of two-dimensional point process on $A = [0, 1] \times \mathbb{R}$ by

$$\mathcal{P}_n = \left\{ \left(\frac{i}{n+1}, \frac{X_i - b_n}{a_n} \right) : i = 1, \ldots, n \right\},$$

where $a_n > 0$ and b_n are sequences of constants for a sample size n. The scaling $i/(n+1)$ maps observation number, which we think of as time, to (0,1) and the scaling $(X_i - b_n)/a_n$ is the one introduced in section 1.2 to produce a GEV limit for $M_n = \max(X_1, \ldots, X_n)$.

The following theorem motivates the use of a two-dimensional non-homogeneous Poisson process as a model for the locations of the points $(i/(n+1), X_i)$ for which X_i exceeds a high threshold.

Theorem 14. (Pickands, 1971)

If there exist sequences of constants $a_n > 0$ and b_n such that

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

for a non-degenerate distribution function G with lower and upper endpoints w_0 and w_1 respectively, then $\mathcal{P}_n \to \mathcal{P}$, where $\mathcal{P}$ is a non-homogeneous Poisson process on $[0,1] \times (w_0, w_1)$, with intensity measure

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

on $A_x = (a,b) \times (x, w_1)$, where $0 \leqslant a < b \leqslant 1$ and $w_0 < x < w_1$.

Note that the limiting non-homogeneous Poisson process applies on a region where the scaled Xs are greater than the lower endpoint w_0. Thus, for finite n we hope that this limiting process applies approximately above some high threshold u that can be applied to the data, since $w_0 < u < w_1$.

Graphical illustration of NHPP model

Here we have simulated four samples of random exp(1) variables with increasing sample size. In addition we use $a_n = 1$ and $b_n = \log(n)$ and introduce a threshold u which is greater than the lower endpoint w_0 of the limiting distribution. Figure 47 helps to demonstrate the point process graphically and illustrate the asymptotic theory for the two-dimensional NHPP model. The pattern of points above the (arbitrary) threshold at -2 converges to a two-dimensional Poisson process with intensity measure $\Lambda(A_u)$.

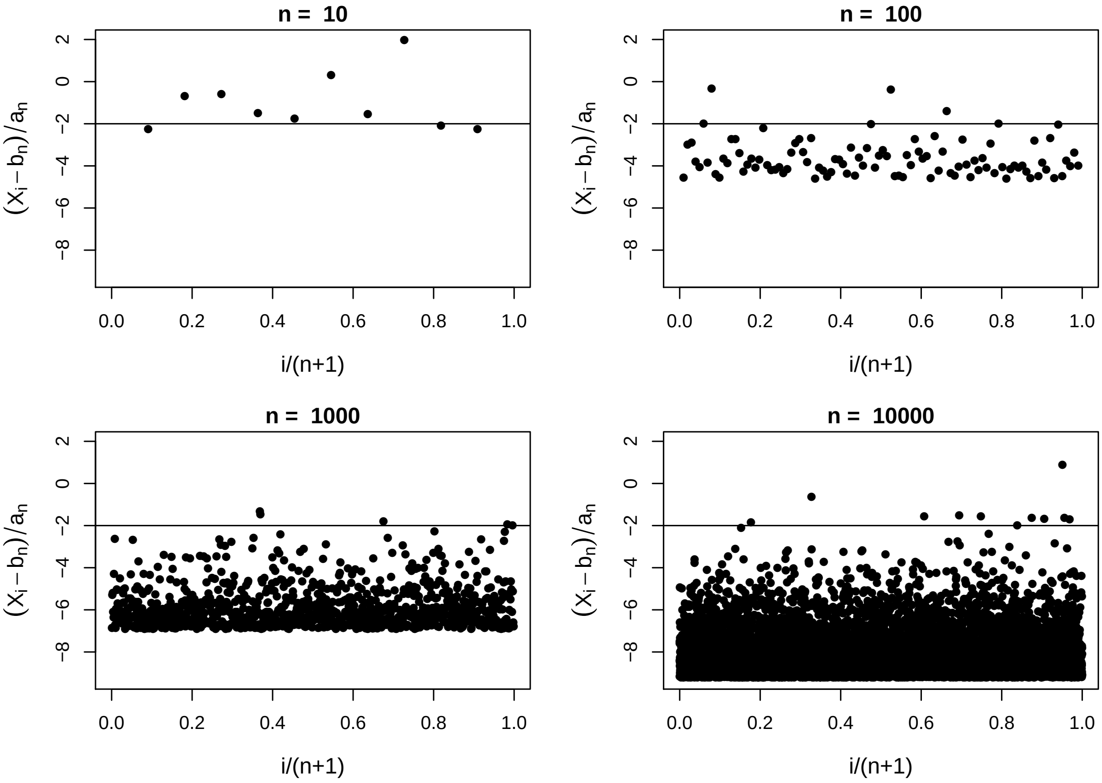

Figure 47: Point process representation with varying sample size and with a high threshold: top left ($n = 10$), top right ($n = 100$), bottom left ($n = 1000$), bottom right ($n = 10000$).

Later we clarify the connection between the NHPP model and the GEV and GP distributions, where the reader can refer to the plots in figure 48 for the two-dimensional regions that are relevant for the Poisson intensity measure of each model. Figure 48 demonstrates graphically the two-dimensional non-homogeneous Poisson process for:

- the case for exceedances above a high threshold u (see plot A) and
- the case for block maxima (see plot B).

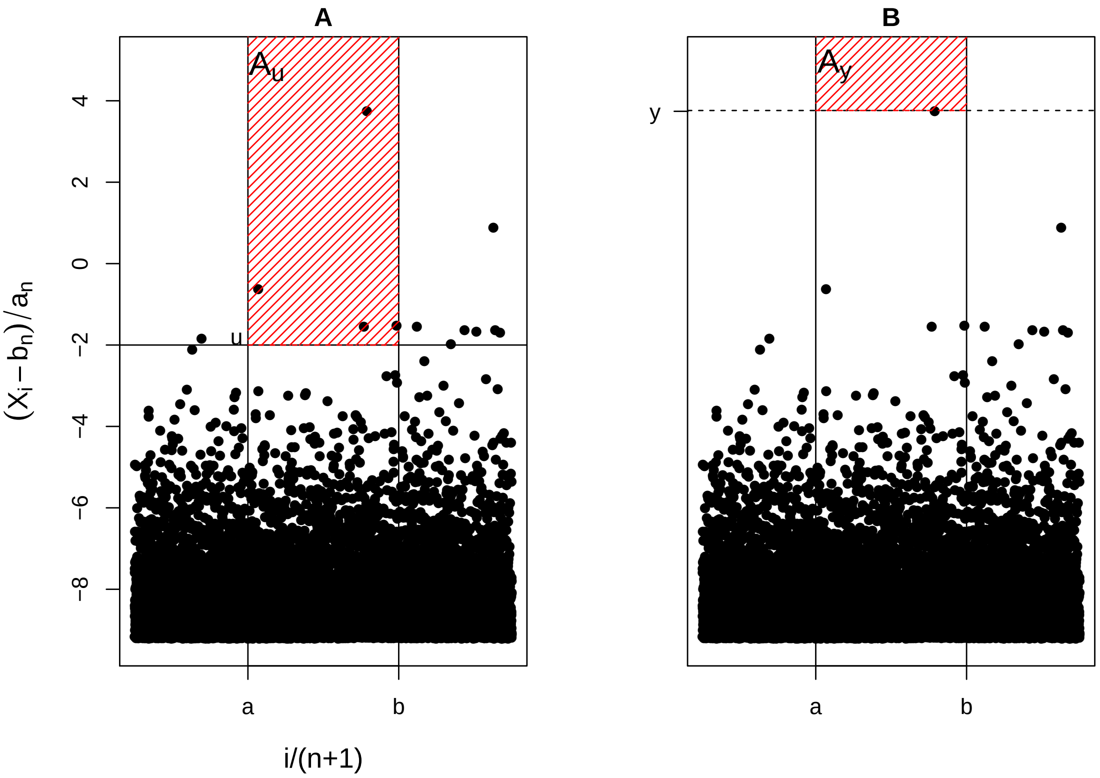

Figure 48: Point process representation ($n = 10000$) for different two-dimensional areas. Plot A: point process according to exceedances above a threshold. Plot B: point process according to block maxima.

The implication of theorem 14 is that, for a sufficiently high threshold u, an approximate model for the process $\mathcal{P}_n$ of threshold exceedances on $A_u = (a, b) \times (u, w_1)$ is provided by a two-dimensional non-homogeneous Poisson process with intensity measure

$$\Lambda(A_u) \;=\; (b-a)\left[1+\xi\left(\frac{u-\mu}{\sigma}\right)\right]_+^{-1/\xi}. \tag{5.2}$$

Plot A of figure 48 illustrates this graphically. Note that here the parameters (μ, σ, ξ) relate to an implied GEV distribution for the maximum value over the time period for which the process is observed. We will adjust this parameterisation later.

Informal justification

An informal justification of theorem 14 uses the fact that by construction the X_i are mutually independent and considers an approximation to the Bin-GP model introduced previously in section 1.3.1. We take $a = 0$ and $b = 1$. Under the Bin-GP model, the number of points, $\mathcal{P}_n(A_u)$, that occur in $A_u = [0, 1] \times (u, \infty)$ has a

Binomial distribution:

$$\mathcal{P}_n(A_u) \sim \text{Binomial}(n, p_u),$$

where the probability p_u that a randomly chosen point of $\mathcal{P}_n$ falls in A_u is given by

$$p_u \;=\; P\left(\frac{X_i - b_n}{a_n} > u\right) \approx \frac{1}{n}\left[1{+}\xi\left(\frac{u{-}\mu}{\sigma}\right)\right]_+^{-1/\xi}.$$

As $n \to \infty$ the distribution of $\mathcal{P}_n(A_u)$ converges to a Poisson limit with intensity measure np_u. Therefore we are moving from a discrete to a continuous process with the properties that

$$\mathcal{P}_n(A_u) \sim \text{Poisson}(np_u),$$

where

$$np_u = \Lambda(A_u) = \left[1{+}\xi\left(\frac{u{-}\mu}{\sigma}\right)\right]_+^{-1/\xi}, \tag{5.3}$$

and, for non-overlapping A_u and A_v, $\mathcal{P}_n(A_u)$ and $\mathcal{P}_n(A_v)$ are independent by assumption. Thus the intensity measure $\Lambda(A_x) = -\log G(x)$ where (following (1.1))

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

GEV and GP models from the NHPP model

We briefly show the connection of the NHPP model with the GEV and GP distributions that were introduced in chapter 1. Under the NHPP model, and letting $M_n = \max\{X_1, \ldots, X_n\}$, we have

$$\begin{aligned} P\left(\frac{M_n - b_n}{a_n} \leqslant y\right) &= P\left(\mathcal{P}_n(A_y) = 0\right) \\ &= \exp\{-\Lambda(A_y)\} = \exp\left\{-\left[1 + \xi\left(\frac{y-\mu}{\sigma}\right)\right]_+\right\}, \end{aligned}$$

where the first equality leads from the fact that under the block maxima approach no points are observed in the region A_y (see plot B in figure 48, noting that in the argument above $a = 0$ and $b = 1$). For the case of threshold excesses, the NHPP model also gives the GP distribution. For suitably high threshold u,

$$\begin{aligned} P\left\{\left(\frac{X_i - b_n}{a_n} > x\right) \;\middle|\; \left(\frac{X_i - b_n}{a_n} > u\right)\right\} &= \frac{\Lambda(A_x)}{\Lambda(A_u)} \\ &= \left[1 + \frac{\xi z}{\sigma_u}\right]_+^{-1/\xi}, \end{aligned}$$

where $z = x - u$ and $\sigma_u = \sigma + \xi(u - \mu)$.

Threshold invariant parameters

The model parameters of the Bin-GP model (introduced in 1.3.1) are p_u, σ_u and ξ. It is important to notice here that two of these parameters, namely p_u and σ_u, depend on the chosen threshold u. If the underlying process does not involve any covariate effects then inference on extremes using the above-mentioned distribution is quite simple and estimation can easily be done. However, if covariate effects are to be modelled then having a threshold-dependent parameterisation is a disadvantage, because the functional relationship between σ_u and the covariates will, in general, depend of the threshold. Eastoe and Tawn (2009) discuss this issue in detail.

The biggest advantage in using the NHPP representation to model threshold excesses is the fact that the parameters μ, σ and ξ are threshold-invariant. This makes it easier to work with a covariate-dependent threshold rather than a constant threshold. We return to this in section 5.2.2.

Likelihood for the NHPP model

Suppose that we observe n_y years of data with n observations per year, giving a total of $m = n_y n$ observations, $x_1, \ldots, x_m$ observed at (scaled) times $t_i = i/(m+1)$. It is common to parameterise the NHPP in terms of the parameters, hereafter denoted $\boldsymbol{\theta} = (\mu, \sigma, \xi)$, of the implied GEV distribution of *annual* maxima. This is achieved by redefining the intensity measure (5.2) as

$$\Lambda(A_u) \;=\; n_y(b-a)\left[1+\xi\left(\frac{u-\mu}{\sigma}\right)\right]_+^{-1/\xi},$$

for some high threshold u. Taking $a = 0$ and $b = 1$, so that $A_u = [0,1] \times (u, \infty)$ leads (see, for example, Coles (2001, page 134)) to the likelihood function

$$\begin{aligned} L(\boldsymbol{\theta}; \boldsymbol{x}) &= \exp\left\{-\Lambda(A_u; \boldsymbol{\theta})\right\} \prod_{i:x_i>u} \lambda(t_i, x_i; \boldsymbol{\theta}), \\ &\propto \exp\left\{-n_y \left[1+\xi\left(\frac{u-\mu}{\sigma}\right)\right]_+^{-1/\xi}\right\} \prod_{i:x_i>u} \sigma^{-1}\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]_+^{-(1+1/\xi)} \end{aligned} \tag{5.4}$$

where the intensity function

$$\lambda(t, x; \boldsymbol{\theta}) = n_y\, \sigma^{-1} \left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-(1+1/\xi)} \tag{5.5}$$

satisfies (5.1), i.e. it is such that $\int_0^1 \int_u^\infty \lambda(t, x; \boldsymbol{\theta})\, \mathrm{d}x\, \mathrm{d}t = \Lambda(A_u)$.

The first term in (5.4) is the rate of threshold exceedance. The second part is the contribution from the exceedances, since the observed x_i values are above the threshold u. Maximum likelihood inference for the NHPP model is subject to the same regularity conditions discussed in chapter 1.

NHPP model assumptions

So far in this chapter we have assumed that $X_1, \ldots, X_m$ are independent and identically distributed. We investigated the topic of dependence in extremes previously in chapter 4. In the remainder of the current chapter the assumption of independence is retained, but we relax the second assumption, by allowing the parameters of the NHPP model to depend on the values of covariates via a regression model. This may be necessary to reflect a temporal trend or seasonal effect. We argue that in this case it is natural to allow the threshold to depend on the covariates. What is more, we focus on the case where covariate effects are present in the location parameter μ and consider how we could set a covariate-dependent threshold in this case.

5.2 Regression modelling

Notation change

We change our notation to conform with the traditional notation used in a regression setting: $Y_1, \ldots, Y_m$ denote the values of a response variable and $x_1, \ldots, x_m$ the corresponding values of a (for the moment, scalar) covariate. A standard approach in this situation is to model the effect of covariate x on the extremal behaviour of Y by allowing the parameters of an extreme value model for Y to depend on x. For example, in the NHPP model we could specify a functional form $\mu(x)$ for the GEV location parameter. In effect we appeal to standard extreme value arguments conditional on the value of the covariate. Furthermore, we assume throughout that conditional on their respective covariate values, the responses $Y_1, \ldots, Y_m$ are independent.

Again we suppose that we observe n_y years of data with n observations per year, giving a total of $m = n_y n$ paired observations, $(Y_1, x_1), \ldots, (Y_m, x_m)$, observed at (scaled) times $t_i = i/(m+1), i = 1, \ldots, m$. The NHPP intensity function (5.5) can be extended by allowing any of its parameters to be covariate-dependent. Here we do this only for the location parameter, considering the simple case where $\mu(x) = \mu_0 + \mu_1 x$, so that

$$\lambda(t, y; \boldsymbol{\theta}) = n_y \, \sigma^{-1} \left[1 + \xi \left(\frac{y - \mu_t}{\sigma} \right) \right]_+^{-(1+1/\xi)}, \tag{5.6}$$

where $\mu_t = \mu(x_t)$ and $\boldsymbol{\theta} = (\mu_1, \mu_0, \sigma, \xi)$. We have ordered the parameters to separate the regression parameter μ_1 from the marginal parameters (μ_0, σ, ξ). We also allow the threshold $u(x)$ to be covariate-dependent, with threshold u_i applying at time t_i with associated covariate value x_i. Under this setup the likelihood is

$$L(\boldsymbol{\theta}; \boldsymbol{y}) = \exp\left\{ - \int_0^1 \int_{u_t}^{\infty} \lambda(t, y; \boldsymbol{\theta}) \, \mathrm{d}y \; \mathrm{d}t \right\} \prod_{i: y_i > u_i} \lambda(t_i, y_i; \boldsymbol{\theta}), \tag{5.7}$$

where $\mu_{t_i} = \mu_i = \mu(x_i)$ and $u_t = u(x_t)$. In practice, and noting that $n_y/m = 1/n$, the integral in (5.7) is approximated by

$$\frac{1}{m} \sum_{i=1}^{m} \int_{u_i}^{\infty} \lambda(t_i, y) \, \mathrm{d}y = \frac{1}{n} \sum_{i=1}^{m} \left[1 + \xi \left(\frac{u_i - \mu_i}{\sigma} \right) \right]_+^{-1/\xi}. \tag{5.8}$$

In sections 5.2.1 and 5.2.2 we discuss some key issues for threshold-based extreme value regression modelling, namely the choice of model, the functional form of the threshold and setting a threshold that is appropriate for all observations. We consider only parametric regression effects, but non-parametric approaches are possible (Chavez-Demoulin and Davison, 2005, Butler et al., 2007).

5.2.1 Threshold-based extreme value regression modelling

Threshold-based extreme value regression modelling dates back to Davison and Smith (1990), who include covariates in the parameters p_u and σ_u of the Bin-GP model, and Smith (1989), who includes covariates in the parameters μ, σ and ξ of the NHPP model. Despite the presence of non-stationarity Davison and Smith (1990) use a constant threshold. Many authors have adopted their approach to the extent that Eastoe and Tawn (2009) described the use of a Bin-GP model with a constant threshold as the “standard”.

It is increasingly accepted (Chavez-Demoulin et al., 2011) that it is usually preferable to use the NHPP rather than the Bin-GP in a regression situation. The dependence of the GP parameter σ_u on the threshold u means that the functional relationship between σ_u and the covariates depends on the value of u. In contrast the parameters of the NHPP are threshold invariant, so if a particular parametric form is used for $\mu(x)$ then this form applies for any threshold.

In a regression situation there are strong arguments (which we discuss in section 5.2.2) against using a constant threshold, and some authors have used non-constant thresholds. Smith (1989) uses a seasonal threshold by applying a different threshold within each month of the year. Similarly, Coles (2001) sets a smooth seasonal threshold by trial-and-error with the aim of achieving an approximately constant rate of threshold exceedance. Eastoe and Tawn (2009) seek to remove regression effects, using a Box-Cox regression model of all the data, so that a constant threshold can reasonably be applied in an extreme value analysis of the residuals from this model. Northrop and Jonathan (2011) set a covariate-dependent threshold by fitting a quantile regression model to estimate a given high conditional quantile of the response as a function of covariates.

5.2.2 Covariate-dependent thresholds

In common with the stationary case considered in previous chapters, setting a threshold involves a bias-variance trade-off (see section 1.4.2 and the introduction of section 3.3). The difference is that now the distribution of the response Y may depend on the value of a covariate x. Therefore, a threshold that is appropriate for one value of x may be inappropriate (too high or too low) for another value of x. This is illustrated in the left hand side of figure 49, where data have been simulated such that x has a linear effect on the location of Y. Using the constant threshold depicted, which is exceeded by 10% of the observations, might be appropriate for the larger values of x but it is too high for small values of x, as it lies above all the observations for which $x < 0.4$. Using a lower (constant) threshold could avoid this problem but it may result in a threshold being applied for large x that is too low for the NHPP model to be applicable.

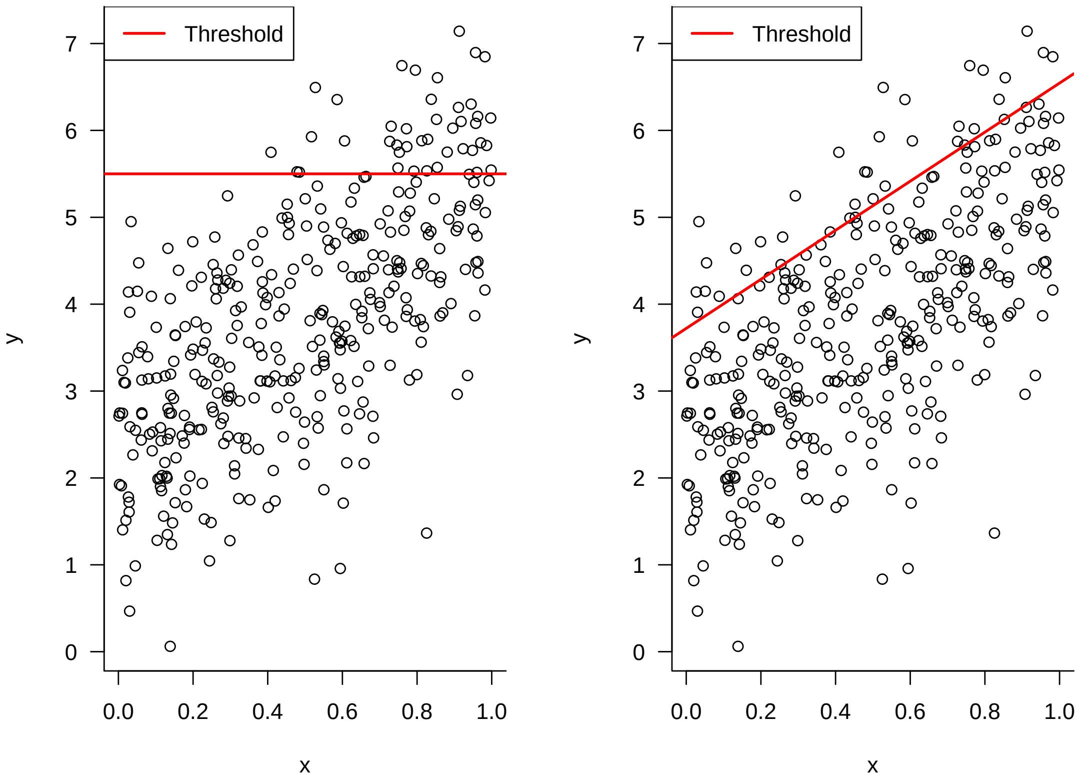

Figure 49: Non-stationary data (x has a linear effect on the location of Y) with a constant threshold (left) and a covariate-dependent threshold (right).

In the right hand side of figure 49 a covariate-dependent threshold is shown, which mimics the linear effect of x on Y. This has been achieved using quantile regression (QR) (Koenker and Bassett, 1978) (see D.1) to estimate the 90% conditional quantile of Y as a function of x, which in the current example is linear in x. Thus, the threshold is of the form $u(x) = u_0 + u_1 x$. The aim is to set a threshold for which (conditional on x) the probability $p(x)$ of threshold exceedance is approximately 0.1 for all values of x. The strategy of seeking a threshold for which $p(x)$ is constant avoids the problems resulting from the use of a constant threshold. It is a logical approach in the current context, where x affects the location of Y only, and seems to be at least a good starting point more generally.

Once a suitable threshold is specified we would model the data in figure 49 using an NHPP model in which $\mu(x) = \mu_0 + \mu_1 x$ and σ and ξ are constant. Setting a threshold using QR is an attempt to set u_1 to be close to μ_1. A further desirable consequence of using the covariate-dependent threshold is that excesses occur over a wider range of values of x than the constant threshold. We would expect to achieve greater precision in estimating μ_1 than with the constant threshold, for which exceedances cover a narrower range of values of x. We might also expect that a threshold with the 'correct' gradient, i.e. with $u_1 = \mu_1$, would be better than one for which $u_1 \neq \mu_1$.

In section 5.3 we study this particular aspect in detail.

5.3 Theoretical study

We consider setting a threshold of the form $u(x) = u_0 + u_1 x$ for an NHPP model in which it is assumed that $\mu(x) = \mu_0 + \mu_1 x$ and σ and ξ are constant. We focus specifically on the effect of the value of the gradient u_1 on the precision of estimation of μ_1, *under the assumption that the NHPP model holds for all the thresholds that are compared.* We do this because we are interested in exploring whether setting $u_1 = \mu_1$ is an optimal strategy when the thresholds compared are sufficiently high such that bias from misspecification of the NHPP model is negligible. However, such biases are important practically and would need to be taken into account if we wished to develop a threshold selection method like those in chapters 3 and 4.

Consider the following data-generating process,

$$Y_i \mid X_i = x_i \sim \text{GEV}(\mu^d(x_i), \sigma^d, \xi), \quad \text{for} \quad i = 1, \ldots, m, \tag{5.9}$$

where $\mu^d(x_i) = \mu_0^d + \mu_1 x_i$ and the superscript d denotes daily parameters. The GEV distribution is chosen because the NHPP model will hold approximately provided that the thresholds we investigate are high. Without loss of generality, we assume that the covariate values have been mean-centred, i.e. that $\sum_{i=1}^m x_i = 0$.

Consider a fixed covariate value x and let $Z = \max(Y_1(x), \ldots, Y_n(x))$, that is, the maximum in a year in which the covariate X is equal to x, throughout. If $Y_1(x), \ldots, Y_n(x)$ are independent then $P(Z \leqslant y \mid X = x) = P(Y \leqslant y \mid X = x)^n$, and setting

$$\mu_0^d = \mu_0 + \sigma\left(n^{-\xi} - 1\right)/\xi \quad \text{and} \quad \sigma^d = \sigma n^{-\xi}. \tag{5.10}$$

means that $Z \mid X = x \sim GEV(\mu_0 + \mu_1 x, \sigma, \xi)$. Thus we have related the daily parameters to the annual parameters $\boldsymbol{\theta} = (\mu_1, \mu_0, \sigma, \xi)$ of the NHPP. For later purposes we state the p.d.f. of daily values parameterised in terms of the annual parameters:

$$f_{Y|X=x}(y; \boldsymbol{\theta}) = \frac{1}{\sigma}\frac{1}{n}\left[1+\xi\left(\frac{y-\mu(x)}{\sigma}\right)\right]_+^{-(1+1/\xi)} \exp\left\{-\frac{1}{n}\left[1+\xi\left(\frac{y-\mu(x)}{\sigma}\right)\right]_+^{-1/\xi}\right\} \tag{5.11}$$

where $\mu(x) = \mu_0 + \mu_1 x$.

Fisher information matrix

We derive the (expected) Fisher information matrix based on the approximate NHPP likelihood defined by (5.7) and (5.8) under the data-generating process (5.9). We require that $\xi > -1/2$ for the Fisher information to exist. Up to an additive constant, the negated log-likelihood is

$$-\ell(\boldsymbol{\theta}; \boldsymbol{y}) \;=\; \frac{1}{n}\sum_{i=1}^{m} g(\boldsymbol{\theta}) + \sum_{i=1}^{m} \delta(y_i > u(x_i))\, h(\boldsymbol{\theta}; y_i),$$

where $\delta(x) = 1$ if x is true and is 0 otherwise,

$$g(\boldsymbol{\theta}) \;=\; \left[1+\xi\left(\frac{u(x_i)-\mu(x_i)}{\sigma}\right)\right]_+^{-1/\xi} \tag{5.12}$$

is a function of $\boldsymbol{\theta}$ and $x_1, \ldots, x_m$ but does not involve the responses $Y_1, \ldots, Y_m$, and

$$h(\boldsymbol{\theta}; Y_i) \;=\; \log\sigma + \left(1+\frac{1}{\xi}\right)\log\left[1+\xi\left(\frac{y_i-\mu(x_i)}{\sigma}\right)\right]_+ \tag{5.13}$$

is a function of $\boldsymbol{\theta}$, $x_1, \ldots, x_m$ and $Y_1, \ldots, Y_m$.

The Fisher information matrix I for $\boldsymbol{\theta} = (\mu_1, \mu_0, \sigma, \xi) = (\theta_1, \theta_2, \theta_3, \theta_4)$ contains the elements $I_{jk} = -\mathrm{E}(\partial^2 l/\partial\theta_j\partial\theta_k)$, for $j, k \in \{1, 2, 3, 4\}$. These elements are derived in appendix D.4, by extending the Fisher information for the stationary case, which is derived in appendix D.3. We partition I as

$$I = \begin{pmatrix} I_{11} & I_1^T \\ I_1 & I_M \end{pmatrix}, \tag{5.14}$$

where $I_1 = (I_{21}, I_{31}, I_{41})^T$ and I_M is the Fisher information matrix for the marginal parameters (μ_0, σ, ξ).

5.3.1 Comparing thresholds

We wish to compare thresholds of the form $u(x) = u_0 + u_1 x$, with different values of u_1, in terms of the efficiency of the MLE of the regression parameter μ_1. A major contributor to the efficiency of parameter estimation is the overall level of the threshold, which is determined by both u_0 and u_1. Lower thresholds will tend to result in greater efficiency than higher thresholds. To make the comparison between different values of u_1 fair, we set u_0 in order to keep constant a measure of the overall level of threshold.

The particular measure we use is the determinant of I_M, $\det I_M$. In earlier work we used an alternative measure, the expected proportion of threshold exceedances, obtaining qualitatively similar numerical results. However, we concentrate on $\det I_M$ because it is a standard measure of total information in multi-parameter problems: it quantifies the information about the marginal parameters (μ_0, σ, ξ) that the data are expected to contain. It also turned out to be easier to work with algebraically.

In the theory of experimental design, a design that maximises the determinant of a Fisher information matrix is *D-optimal* (Atkinson et al., 2007, Chapter 10). Similarly, $\det I$ quantifies the expected information about $(\mu_1, \mu_0, \sigma, \xi)$. Suppose that $\det I_M$ is held constant across all thresholds of the form $u(x) = u_0 + u_1 x$. The value of $\det I$ will depend on u_1. We call the threshold that maximises $\det I$ a *D-optimal threshold.*

In the current context, a D-optimal threshold minimises the asymptotic variance $\text{var}(\widehat{\mu}_1)$ of the MLE of μ_1. Using the block inversion result from appendix D.5 gives the asymptotic variance of $\widehat{\mu}_1$ as

$$\text{var}(\widehat{\mu}_1) = \left(I^{-1}\right)_{11} = (I_{11} - I_1^T I_M^{-1} I_1)^{-1}, \tag{5.15}$$

and therefore the precision of $\widehat{\mu}_1$ is given by

$$\text{prec}(\widehat{\mu}_1) = I_{11} - I_1^T I_M^{-1} I_1. \tag{5.16}$$

Schur's determinant identity (D.5) shows that

$$\det I = \det I_M \det\left(I_{11} - I_1^T I_M^{-1} I_1\right) = \det I_M \,\text{prec}(\widehat{\mu}_1). \tag{5.17}$$

As $\det I_M$ is kept constant, maximising $\det I$ and minimising $\text{var}(\widehat{\mu}_1)$ are equivalent.

To carry out this study we need to explore how the elements of I behave as u_1 varies and $\det I_M$ is kept constant. An interesting aspect of this study is that the covariate values $\boldsymbol{x} = (x_1, \ldots, x_m)$ have an impact. In section 5.3.2 we consider the special case where $\boldsymbol{x}$ are symmetric (about 0). This occurs, for example, if the covariate is time and it is sampled regularly. In section 5.3.3 we consider the case where $\boldsymbol{x}$ are not symmetric.

5.3.2 Symmetric covariate values

We start with the case $u_1 = \mu_1$, for which the thresholds $u(x_i) = u_0 + u_1 x_i, i = 1, \ldots, m$ all lie at the same level, $100q\%$ say, of conditional quantile of $Y_i \mid X = x_i$. We set u_0 to achieve a particular value of q and calculate $d_0 = \det I_M$ in this instance. We then vary u_1, each time altering u_0 so that $\det I_M = d_0$, and calculate the full Fisher information I. Without loss of generality we have used $(\mu_1, \mu_0, \sigma, \xi) = (1, 0, 1, -0.2)$, $m = n = 365$ and $q = 0.95$ to produce the following results: the findings apply to all cases. The covariate values $x_1, \ldots, x_m$ are equally-spaced on the interval $[-1/2, 1/2]$.

In figures 50 and 51 we plot the elements of the Fisher information I against u_1. The diagonal elements of I_M (left hand side of figure 50) all have the properties that (a) they are maximised when $u_1 = \mu_1 = 1$, and (b) the plots are symmetric about $u_1 = \mu_1$. The off-diagonal elements (right hand side of figure 50) exhibit similar behaviour except that (owing to the fact that these elements can be negative) property (a) becomes that the absolute value of each element is maximised when $u_1 = \mu_1$.

The plot on the top left of figure 51 shows that I_{11} behaves in the same way as I_{22}, i.e. its value is symmetric about the location of its maximum at $u_1 = \mu_1$. This element of I is of primary interest because it summarises the expected information about μ_1. The other plots show that the elements I_{21}, I_{31}, I_{41} are monotonic (but on close inspection not quite linear) in u_1 and that they are equal to zero only when $u_1 = \mu_1$.

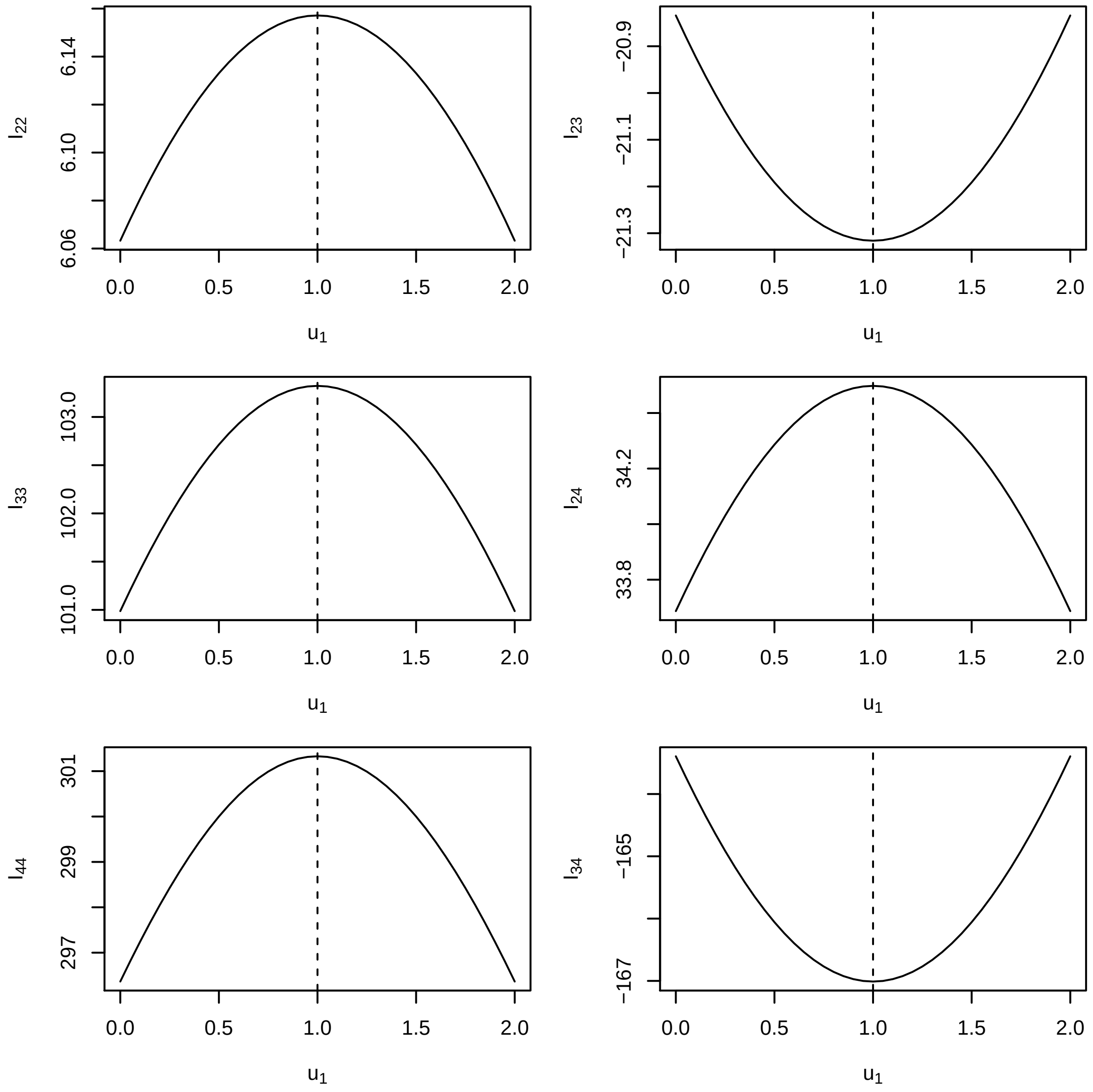

Figure 50: Individual elements of I_M against u_1 using symmetric covariate values.

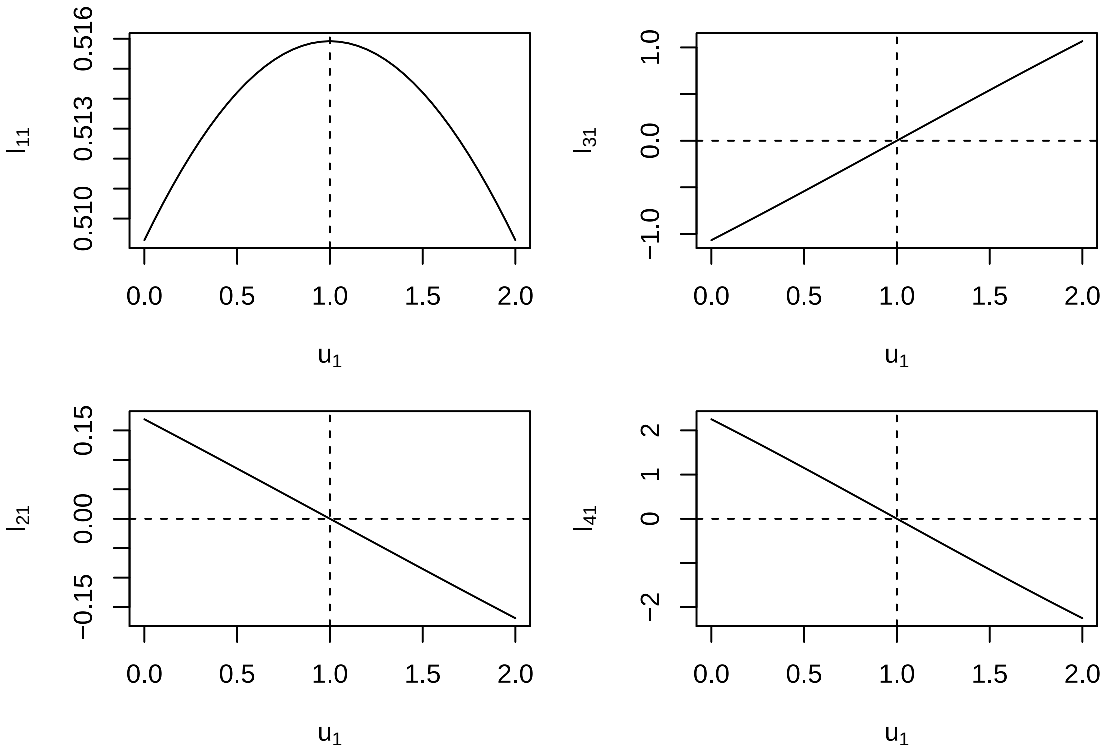

Figure 51: I_{11} and individual elements of I_1 against u_1 using symmetric covariate values.

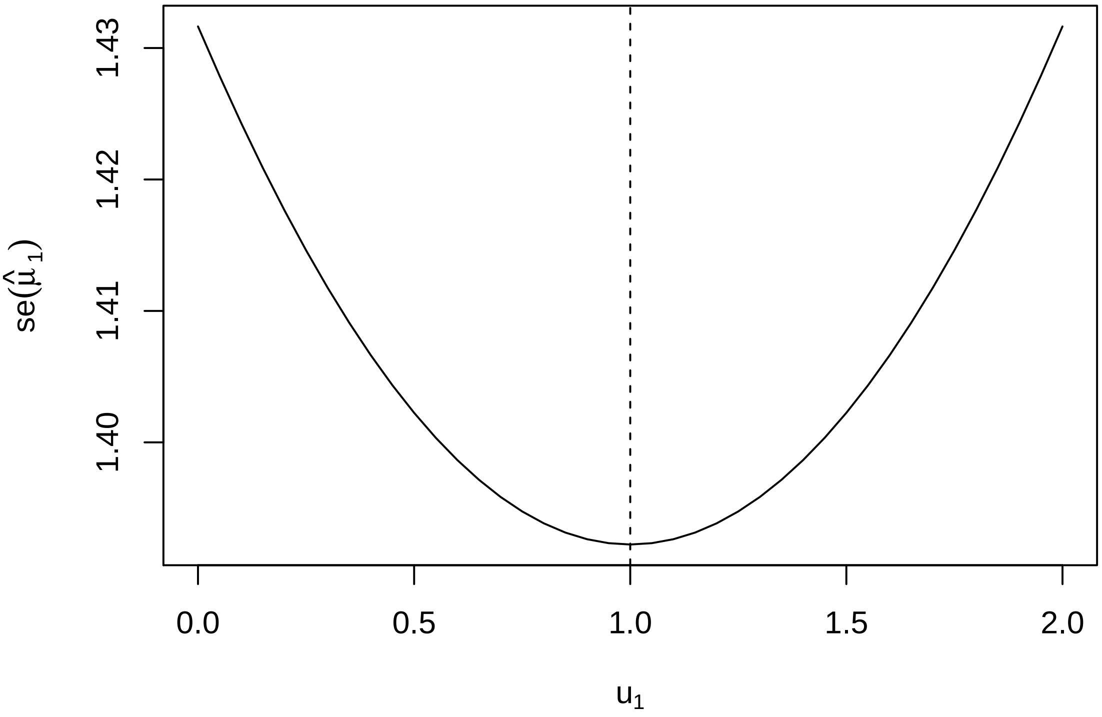

Figure 52: Asymptotic standard error of $\widehat{\mu}_1$ against u_1 using symmetric covariate values.

Figure 52 shows that the asymptotic standard error of $\widehat{\mu}_1$ (based on (5.15)) is minimised when $u_1 = \mu_1$. This is a general result. The proof relies on showing that the behaviour exhibited in figures 50 and 51 by the elements of I is general.

Properties of elements of *I*

We prove that under the setup described in section 5.3.1 the Fisher information matrix I has the following properties.

- **Property 1**: If $u_1 = \mu_1$ then $I_1 = (I_{21}, I_{31}, I_{41})^T = (0, 0, 0)^T$
- **Property 2**: The absolute values of the elements of I_M are maximised when $u_1 = \mu_1$.
- **Property 3**: If $x_1, \ldots, x_m$ are symmetric about 0 then I_{11} is maximised when $u_1 = \mu_1$.

Proof of property 1. From appendix D.4 we have that for $k \in \{2, 3, 4\}$

$$I_{k1} \quad = \quad \sum_{i=1}^{m} x_i f_k(x_i), \tag{5.18}$$

for some $f_k(x_i)$ that depends on x_i only through $u(x_i) - \mu(x_i)$. If $u_1 = \mu_1$ then $u(x_i) - \mu(x_i) = u_0 - \mu_0$, which does not depend on x_i. Therefore, I_{k1} is proportional to $\sum_{i=1}^{m} x_i$, which is equal to 0.

Proof of property 2. See appendix D.6.

Proof of property 3. From appendix D.4 we have

$$I_{22} = \sum_{i=1}^{m} f_2(x_i) \qquad \text{and} \qquad I_{11} = \sum_{i=1}^{m} x_i^2 f_2(x_i).$$

If there are any covariate values of zero then these do not contribute to I_{11}. The other values occur in n_p pairs with paired covariate values $(-c_j, c_j), j = 1, \ldots, n_p$, say. Therefore,

$$I_{11} = \sum_{j=1}^{n_p} c_j^2 \left\{ f_2(-c_j) + f_2(c_j) \right\} = \sum_{j=1}^{n_p} c_j^2 t_j.$$

Each t_j is the contribution to I_{11} in a case where $m = 2$. The proof of property 2 holds when $m = 2$. Therefore, each of $t_j, j = 1, \ldots, n_p$ are maximised when $u_1 = \mu_1$.

Therefore I_{11} is maximised when $u_1 = \mu_1$. Also, when $u_1 = \mu_1$, $f_2(x_i)$ does not depend on x_i so the maximised values of I_{11} and I_{22} satisfy $I_{11} = I_{22}\,(1/m)\sum_{i=1}^{m} x_i^2$.

Proof of optimality of $u(x) = u_0 + \mu_1 x$ for symmetric covariate values

Provided that the regularity condition $\xi > -1/2$ holds, and that the covariate values $x_1, \ldots, x_m$ are not all identical, then I_M is positive definite and I_M^{-1} is also positive definite (see D.5). Therefore,

$$\text{prec}(\widehat{\mu}_1) = I_{11} - I_1^T I_M^{-1} I_1 \leqslant I_{11} \tag{5.19}$$

with equality only when $I_1 = (0, 0, 0)^T$, which occurs when $u_1 = \mu_1$ (Property 1). Property 3 shows that, in the special case where $x_1, \ldots, x_m$ are symmetric about 0, I_{11} is maximised when $u_1 = \mu_1$ and so $\text{prec}(\widehat{\mu}_1)$ is maximized when $u_1 = \mu_1$.

5.3.3 Asymmetric covariate values

We consider the effect of asymmetry in the covariates on the optimal value of u_1. The proofs of properties 1 and 2 in section 5.3.2 do not require that the covariate values are symmetric and therefore continue to hold. However, property 3 does not necessarily hold when the covariate values are asymmetric. Therefore, $u_1 = \mu_1$ may not be optimal.

We illustrate the effect of skewness in the covariate values with an example. Again we use $(\mu_1, \mu_0, \sigma, \xi) = (1, 0, 1, -0.2)$, $m = n = 365$ and $q = 0.95$, but now $x_1, \ldots, x_m$ are positively skewed, with a sample skewness coefficient of 0.64. Figure 53 shows that now the asymptotic standard error of $\widehat{\mu}_1$ is minimised for a value of u_1 that is less than μ_1.

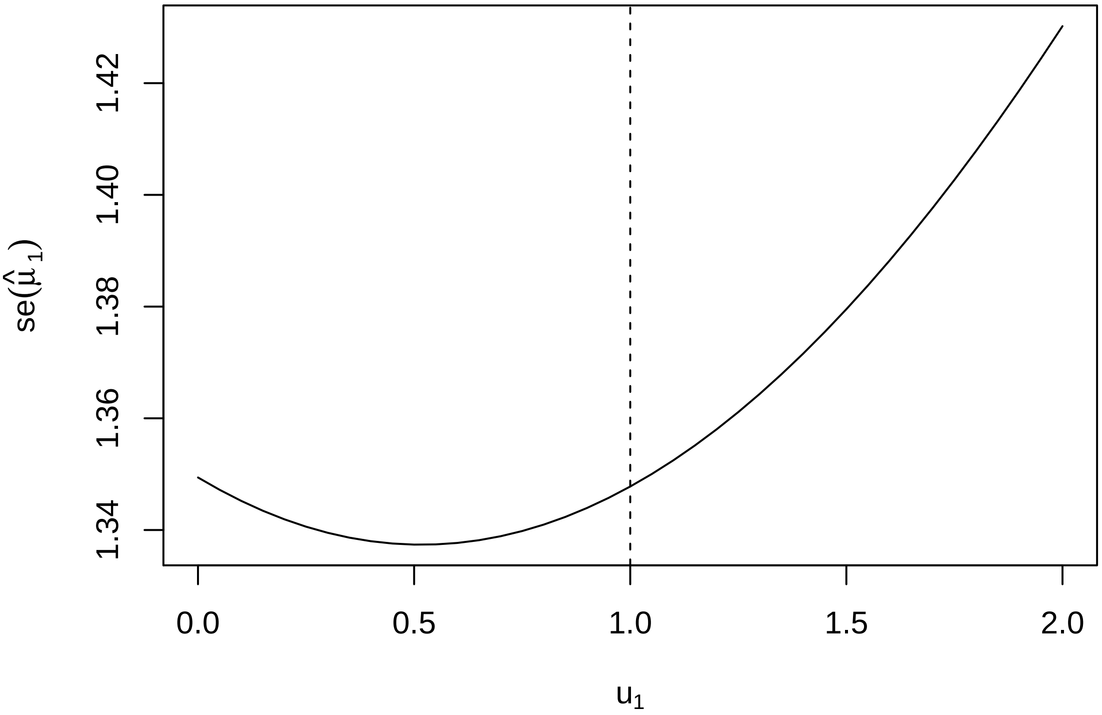

Figure 53: Asymptotic standard error of $\widehat{\mu}_1$ against u_1 using asymmetric covariate values.

In this example the optimal threshold has a lower gradient than the gradient μ_1 in the data-generating process. When the covariate values are positively skewed there are many small covariate values and fewer large values. If $u_1 = \mu_1$ the probability of threshold exceedance is the same for all responses. However, due to the covariate skewness when $u_1 = \mu_1$ we expect fewer threshold exceedances associated with large covariate values and more threshold exceedances associated with small covariate values. To compensate for the expected relative lack of threshold exceedances for large covariate values, a 'locally' lower threshold (than would be the case for $u_1 = \mu_1$) is suggested for large x_i values and a 'locally' high threshold for small x_i values. The resulting optimal threshold has a lower gradient than μ_1. This argument suggests that for negatively skewed covariate values we should find that the optimal value of u_1 is greater than μ_1 and indeed this is what we observe (plots not shown).

The result in section 5.3.2 supports the use of quantile regression to set an extreme value regression threshold in the current example. We have ignored the effects of model misspecification bias. If $u_1 = \mu_1$ this bias is constant across different covariate values. This is not the case when $u_1 \neq \mu_1$. The numerical work summarised in section 5.3.3 shows that, in terms of precision of estimation of μ_1, aiming to set $u_1 = \mu_1$ may not be optimal. However, if changes in bias that result from deviating from $u_1 = \mu_1$ are taken into account we expect that the optimal value of u_1 would change.

5.3.4 Extension to multiple covariates

We extend the model so that the location parameter is linear in p covariates. For each set of covariate values $\boldsymbol{x}_i = (x_{1i}, \ldots, x_{pi}), i = 1, \ldots, m$ the location $\mu(x)$ of the GEV distribution for annual maxima is given by

$$\mu(\boldsymbol{x}_i) = \mu_0 + \mu_1\, x_{1i} + \cdots + \mu_p\, x_{pi}.$$

We write the Fisher expected information matrix for $(\mu_1, \ldots, \mu_p, \mu_0, \sigma, \xi)$ as

$$M = \begin{pmatrix} M_{11} & M_{12}^T \\ M_{12} & M_{22} \end{pmatrix},$$

where

- M_{11} is the $p \times p$ sub-matrix of M relating to the regression parameters $\mu_1, \ldots, \mu_p$;
- M_{22} is the 3×3 sub-matrix relating to the marginal parameters (μ_0, σ, ξ);
- M_{12} is a $3 \times p$ matrix, with jth column, for $j = 1, \ldots, p$, being the vector

$$\left(-E\left[\sum_{i=1}^{m} x_{ji} \frac{\partial^2 \ell(\boldsymbol{\theta}; y_i)}{\partial \mu_0 \partial \mu_j}\right], -E\left[\sum_{i=1}^{m} x_{ji} \frac{\partial^2 \ell(\boldsymbol{\theta}; y_i)}{\partial \mu_j \partial \sigma}\right], -E\left[\sum_{i=1}^{m} x_{ji} \frac{\partial^2 \ell(\boldsymbol{\theta}; y_i)}{\partial \mu_j \partial \xi}\right] \right).$$

Here M_{22} is analogous to I_M in (5.14), M_{11} to I_{11} and M_{12} to I_1.

Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)$. The asymptotic variance matrix of the maximum likelihood estimator of $\boldsymbol{\mu}$ is given by the $p \times p$ upper left submatrix M^{11} of M^{-1}. This is given by

$$M^{11} = (M/M_{22})^{-1} = \left(M_{11} - M_{12}^T M_{22}^{-1} M_{12}\right)^{-1},$$

where M/M_{22} is the Schur complement of M_{22} in M (see D.5).

While keeping $\det M_{22}$ constant, we seek to minimize $\det(M^{11})$, or equivalently to maximize $\det(M/M_{22})$, which is given by

$$\det(M/M_{22}) = \det\left(M_{11} - M_{12} M_{22}^{-1} M_{12}^T\right) = \frac{\det M}{\det M_{22}},$$

(see D.5). This is analogous to the definition of D_S-optimality in experimental design (Atkinson et al., 2007), with the marginal parameters (μ_0, σ, ξ) viewed as nuisance parameters and $\mu_1, \ldots, \mu_p$ as the parameters of interest. As $\det M_{22}$ is kept constant our approach is equivalent to maximizing $\det M$, the criterion for D-optimality.

Suppose that for the set of covariate values $\boldsymbol{x}_i$ the threshold is of the following form

$$u(\boldsymbol{x}_i) = u_0 + u_1\, x_{1i} + \cdots + u_p\, x_{pi}.$$

Without loss of generality suppose that all the covariates are mean-centred at zero, that is, $\sum_{i=1}^{m} x_{ji} = 0$, for $j = 1, \ldots, p$.

We show that a property like Property 1 of section 5.3.2 holds. By analogy with (5.18), for $r = 1, 2, 3$ and $j = 1, \ldots, p$, the (r, j) element of M_{12} can be written as

$$(M_{12})_{rj} = \sum_{i=1}^{m} x_{ji} f_r(\boldsymbol{x}_i),$$

where $f_r(\boldsymbol{x}_i)$ depends only on $\boldsymbol{x}_i$ only through $u(\boldsymbol{x}_i) - \mu(\boldsymbol{x}_i)$. If $u_i = \mu_i$, for $i = 1, \ldots, p$, then $u(\boldsymbol{x}_i) - \mu(\boldsymbol{x}_i) = u_0 - \mu_0$, which does not depend on $\boldsymbol{x}_i$. Therefore, for each r and j, $(M_{12})_{rj}$ is proportional to $\sum_{i=1}^{m} x_{ji}$, which is equal to 0, and M has the following property.

Property 1*. If $u_i = \mu_i$, for $i = 1, \ldots, p$ then M_{12} is a zero matrix.

Similarly, for $j = 1, \ldots, p$ and $k = 1, \ldots, p$, the (j, k) element of M_{11} can be written as

$$(M_{11})_{jk} = \sum_{i=1}^{m} x_{ji} x_{ki}\, f_{jk}(\boldsymbol{x}_i),$$

where $f_{jk}(\boldsymbol{x}_i)$ depends only on $\boldsymbol{x}_i$ only through $u(\boldsymbol{x}_i) - \mu(\boldsymbol{x}_i)$. In the case $j = k$, property 3 of section 5.3.2 applies to $(M_{11})_{jj}$, for $j = 1, \ldots, p$: if $x_{j1}, \ldots, x_{jm}$ are symmetric about 0 then $(M_{11})_{jj}$ is maximized when $u_j = \mu_j, j = 1, \ldots, p$.

Suppose, further, that $\sum_{i=1}^{m} x_{ji} x_{ki} = 0$, for all j and k, that is, the covariates are mutually orthogonal. Now, if $u_j = \mu_j, j = 1, \ldots, p$ then M_{11} is diagonal, with diagonal elements that are as large as they could be. Inequality (D.7) in appendix D.5 shows that, for given diagonal elements, the determinant of M_{11} (given by the product of these elements) is maximised when M_{11} is diagonal. Therefore, $\det M_{11}$ is maximised when $u_j = \mu_j, j = 1, \ldots, p$, leading to the following property.

Property 3*: If $x_{j1}, \ldots, x_{jm}$ are symmetric about 0 for all j and $\sum_{i=1}^{m} x_{ji} x_{ki} = 0$, for all j and k, then $\det M_{11}$ is maximised when $u_j = \mu_j, j = 1, \ldots, p$.

The matrix determinant inequality (D.9) in appendix D.5 shows that

$$\det(M/M_{22}) = \det(M_{11} - M_{12}M_{22}^{-1}M_{12}^{T}) \leqslant \det M_{11}, \tag{5.20}$$

with equality if and only if M_{12} is a zero matrix, which occurs when $u_j = \mu_j, j = 1, \ldots, p$ (Property $1^{\star}$). This is an extension of (5.19) to the multi-dimensional case. Property $3^{\star}$ shows that if the covariates are orthogonal and each is symmetric about zero then $\det(M/M_{22})$ is maximized when $u_j = \mu_j, j = 1, \ldots, p$.

Orthogonality of the covariates is not necessary for this result to hold. Suppose that the covariates $x_1, \ldots, x_p$ are orthogonalised using principle components analysis (Jolliffe, 2002). We replace the original covariates $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ with principal components (PCs) $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p)$. Let S be the sample covariance matrix of $\boldsymbol{x}$. The jth PC is given by $\boldsymbol{z}_j = \boldsymbol{\alpha}_j^T \boldsymbol{x}$,where $\boldsymbol{\alpha}_j$ is an eigenvector of S corresponding to its jth eigenvalue. The PCs are linear combinations of the original covariates that are orthogonal. If we use p PCs $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p$ as covariates we have not changed the model, but re-expressed it in terms of covariates that are orthogonal. If each of the original covariates are symmetric about 0 then linear combinations of them are also symmetric about zero. For $\det(M/M_{22})$ to be maximised when $u_j = \mu_j, j = 1, \ldots, p$ it is sufficient that each of the covariates is symmetric about zero.

We have investigated the scenario where covariates are present only in the location parameter. It is of interest to extend the theory to the case where there are covariates in the scale and/or shape parameter, but this is a non-trivial exercise. For example, consider the case where there are covariate effects in the scale $\sigma(\boldsymbol{x})$ and ξ is constant. In this case, setting a threshold $u(\boldsymbol{x})$ for which the probability of exceedance is constant equates to $[u(\boldsymbol{x}) - \mu(\boldsymbol{x}]/\sigma(\boldsymbol{x})$ being constant with respect to $\boldsymbol{x}$. For such a threshold, the information matrices $K_1, \ldots, K_m$ associated with individual observations are not proportional. This is because the individual elements of K_i contain multiplicative factors in which $\sigma(\boldsymbol{x})$ enters in different ways. Therefore, Minkowski's determinant inequality ((D.10) in appendix D.5), which was used to prove Property 2, cannot be used.

6 Conclusions

The primary aim of this thesis is to contribute to areas of extreme value analysis that involve selecting a high threshold, that is, threshold-based extreme value modelling.

In chapter 1 we introduced models involved in extreme value theory for univariate i.i.d. sequences and univariate dependent sequences. In the former case this theory suggests a Bin-GP model for the occurrence of threshold exceedances and the magnitude of threshold excesses. In the latter case threshold exceedances occur in clusters with a mixture model (the K-gaps exponential mixture model) governing whether two successive exceedances are in the same cluster or from different clusters. In chapter 2 we studied Bayesian inference for the GP distribution, comparing reference priors for the GP parameters. Having established conditions sufficient for these improper priors to result in proper posterior distributions, we identified, through a simulation study, priors that result in reliable predictive inference of quantities of interest in an extreme value analysis.

One of these priors is used in chapter 3 where we proposed a new approach for selecting an extreme value threshold for i.i.d. sequences. This method employs a cross-validation technique through a Bayesian predictive approach. The main motivation behind our method is that, in contrast with many existing methods, it seeks to address directly the bias-variance trade-off associated with selecting a threshold. The outcome is a graphical diagnostic tool that allows the user to identify the 'best' threshold by choosing the one that provides the best cross-validatory performance. In addition, we used a Bayesian model averaging approach to extend the idea of selecting a single 'best' threshold to one that accounts for uncertainty in the choice. We illustrated our method using storm peak significant wave height data from the Gulf of Mexico and from the North Sea. Through a simulation study we assessed the performance of the method and concluded that using model averaging to account for threshold uncertainty produces less variability compared to the single 'best' threshold strategy, at the expense of greater bias.

Moving away from the i.i.d. case, in chapter 4 we relaxed the assumption of independence to consider threshold selection for stationary dependent sequences, based on the K-gaps exponential mixture model for inter-exceedance times. We employed the same general strategy as chapter 3, adapting the cross-validation scheme to operate with K-gaps rather than raw observations. We augmented the K-gaps likelihood to account for censored inter-exceedance times occurring at the start and end of the observation period, to use all available information and to avoid the unrealistic behaviour that can occur for very high thresholds. We applied the method to sea-surge heights from Newlyn obtaining results that are in agreement with the threshold choice made by Fawcett and Walshaw (2012) using the same data. In a simulation study our method was at least competitive with, and perhaps outperformed, the model misspecification testing approach of Süveges and Davison (2010), although a definitive comparison is hampered by the fact that the K-gaps model is not true at any threshold.

In chapter 5 we relaxed the i.d. part of i.i.d. to consider the case where the distribution of the extremes of a response variable are related to covariates. We introduced the NHPP regression model and compared theoretically different (linear in the covariates) thresholds for the case when linear covariate effects are present in the location of the distribution. In particular, we were interested in the performance of thresholds with the same regression coefficients as the covariate effects, as this relates to the use of quantile regression to set a threshold. The utility of the results in this chapter are limited somewhat by the fact that we considered only the precision of estimation of model parameters, ignoring the issue of bias. However, an interesting finding is that quantile regression is optimal if the covariate values are symmetric, but not necessarily so otherwise.

Our proposed CV method for selecting the threshold is a useful addition to the existing threshold selection methods and graphical diagnostic tools. It can be used both to choose a single threshold (which is then treated as fixed and known) or to account for uncertainty in this choice by averaging extreme value inferences over several thresholds, weighting thresholds with better predictive performance more heavily than those with poorer performance. In contrast with other methods that perform the latter function, our general approach uses standard unmodified extreme value models. This makes it more amenable to extension to other settings, as we explain below.

Informative priors

One way that our work in chapters 2 and 3 can be extended is by considering informative or weakly-informative priors for the GP model parameters. We have used a particular reference prior, chosen because it results in better predictive performance than other reference priors. Of course, the basis on which this prior was chosen takes no account of the practicalities of real extreme value problems, about which at least some prior information is likely to exist. As pointed out in section 2.2.1, there are arguments in favour of using an informative prior in an extreme value analysis. The main aim of extreme value modelling is to enable extrapolation beyond the range of observed data. In practice, the extent to which realistic extrapolation can be made is limited by the amount of information provided by data and by the prior and in section 3.4 we observed and discussed the unrealistic extrapolations (see figure 22) that can result if too little information is available. This issue is not specific to our methodology - it is pertinent to any extreme value analysis, Bayesian or not - and our approach can trivially accommodate either a fully-informative prior or a weakly-informative prior, guided by an expert. The latter type of prior is intended merely to prevent unrealistic inferences, downweighting *a priori* parameter values corresponding to unrealistically large events, while allowing information from the data to be influential if it conflicts with this. One possibility is to place a prior on the GP shape parameter ξ that downweights large (e.g. $\xi > 1$) values. Gelman (2006) argues that a half-Cauchy prior provides a suitable weakly-informative prior for a standard deviation. By extension, perhaps a Cauchy prior for ξ with location zero merits some investigation.

Modelling threshold excesses in the dependence case

The work in chapter 4 considers selecting the threshold for the case of dependent and identically distributed sequences. Here we describe ways in which this work can be extended.

We have used the K-gaps model, parameterised by the extremal index, to select a threshold u. First we fixed K at a value judged to be appropriate and then we compared the performance of different thresholds. As, in practice, there is no definitive choice of K it would be better to be able to compare different (u, K) *pairs*. Within the context of our CV approach this implies defining a range of plausible training values for K, say $K_1 < \cdots K_h$ to combine with training thresholds $(u_1, \ldots, u_k)$ to form a grid of $(u_i, K_j), i = 1, \ldots, k, j = 1, \ldots, h$ pairs. The performance of each pair is judged on its ability to predict leave-one-out validation K-gaps, produced using a validation threshold $v = u_k$ and a validation value $K_v = K_h$ of K. It is necessary that the validation is based on fixed values of v and K_v that are no smaller than the largest training threshold u and the largest training K, respectively. Once the grid of (u, K) values has been chosen, performing the cross-validation to search for the 'best' combination of of threshold and K is straightforward and mirrors the model misspecification testing approach of Süveges and Davison (2010). Once more, a key issue is the setting of appropriate values of u_k and K_h.

In chapter 4 we did not investigate the uncertainty involving the choice of u, that is, we did not average inferences about extreme quantiles over u like we did in chapter 3.

The reason for this is that it is less trivial to perform such inferences. As explained in section 1.6.2, even once a threshold has been selected, there is more than one way to proceed.

One possibility is to use the chosen run parameter K to decluster the data to form a set of cluster maxima (Ferro and Segers, 2003, Süveges, 2008), which are treated as independent and modelled using the methods in chapter 3 based on a Bin-GP model. We could use the threshold chosen using the K-gaps model but there is no guarantee that this threshold performs well in the context of the Bin-GP model. This could be checked by performing the threshold selection methodology of chapter 3 on the cluster maxima, but then what would we do if this analysis suggested a rather different threshold than the K-gaps analysis? Perhaps the best declustering-based option is to perform threshold selection using the K-gaps model and the Bin-GP model for cluster maxima simultaneously. For fixed (u, K) inferences about inter-exceedance times and the marginal distribution of cluster maxima can proceed separately and an overall measure of cross-validatory performance equal to the sum of (3.4) and (4.10) can be used.

However, Fawcett and Walshaw (2007, 2012) demonstrate that it is preferable to model a GP distribution using *all* threshold excesses in the raw data, rather than only cluster maxima of declustered data. Using cluster maxima throws away information about the marginal distribution of extremes and the process of declustering can introduce non-negligible bias into inferences. However, opting to use all the excesses raises the question: "what is the likelihood in this case?". As our CV approach uses a Bayesian analysis, having a likelihood is crucial. It is clear that the 'independence' likelihood used in chapter 3 is wrong, but one could adjust it for the presence of local extremal dependence using the method of Chandler and Bate (2007). Fawcett and Walshaw (2012) use an equivalent approach to adjust estimates of uncertainty about model parameters and extreme quantiles. Thus, a CV approach using this adjusted likelihood for the raw data could replace the independence likelihood for cluster maxima discussed in the previous paragraph. Even though declustering is no longer performed the K-gaps modelling is still useful to inform the choice of threshold and necessary because inferences about the extremal index θ are required in order to make inferences about future extreme values.

Another alternative is to use a model that incorporates explicit parametric assumptions for the nature of the temporal extremal dependence, in addition to those already made about the marginal distribution of extremes. This is necessary if the form of the dependence influences a quantity of interest, for example, if we want to make inferences about the duration of an extreme event. One possibility is a Markov chain model based on a bivariate threshold excess model (Smith et al., 1997, Fawcett and Walshaw, 2006). However, there are many such models and many other approaches are possible. Nevertheless, whichever model is used there is scope for the cross-validatory threshold selection methodology developed in this thesis to be used provided that one can estimate the predictive density of the validation data. To alleviate the problem of extreme value model uncertainty one could consider averaging inferences over both thresholds and a set of candidate models, weighting threshold-model combinations according to their cross-validatory predictive performance.

Regression effects

An interesting extension of the work in chapter 3 is to use CV to perform threshold selection in the setting of chapter 5, that is, for an extreme value regression model with responses assumed to be conditionally independent given their respective covariate values. The same general principles can be applied, comparing training thresholds $u_1(\boldsymbol{x}), \ldots, u_k(\boldsymbol{x})$ based on predictive ability at some validation threshold, say $v(\boldsymbol{x}) = u_k(\boldsymbol{x})$, but implementation is less straightforward: models, and covariate-dependent thresholds, are more complicated and have more parameters and there is the potential to compare thresholds in terms of their form, in addition to their overall level. The latter issue can be simplified by using quantile regression to set thresholds at estimates of the $100\tau\%$ conditional quantile, so that threshold selection reduces to a choice of τ. We need to ensure that no $u_i(\boldsymbol{x})$ exceeds $v(\boldsymbol{x})$ in the range of the data, for example, by using a constrained version of quantile regression that avoids crossing of fitted quantile curves for different τ (Bondell et al., 2010).

Simple parametric regression effects on extreme value parameters, like those considered in chapter 5, could be used. However, in some applications such models may not be sufficiently flexible to represent these effects and non-parametric regression may be preferable. One possibility is to use a flexible family of functions, for example cubic splines (Chavez-Demoulin and Davison, 2005, Jonathan et al., 2014), avoiding over-fitting by penalising roughness in the fitted regression curves. A, perhaps more simple, alternative is local-likelihood regression (Ramesh and Davison, 2002, Butler et al., 2007), where simple parametric effects (perhaps constant or linear in covariate) are estimated at each covariate value $\boldsymbol{x}_0$ of interest and a kernel function weights more heavily contributions from observations the closer their covariate value is to $\boldsymbol{x}_0$. Consider the case where there is a single scalar covariate. Then the kernel function will involve a *bandwidth* h that controls how the weights decay with distance from $\boldsymbol{x}_0$: the smaller h the more local to $\boldsymbol{x}_0$ an observation needs to be to influence the fit at $\boldsymbol{x}_0$. The choice of a suitable value of h is crucial and involves a bias-variance trade-off: a small value of h would produce estimators with high variance but low bias and the opposite is true for high values of h. Like thresholds, different values of h could be compared based on cross-validatory performance. In the locally-constant case training thresholds could be set at local estimates of the $100(\tau_1, \ldots, \tau_k)\%$ quantiles, for some set of non-exceedance probabilities $(\tau_1, \ldots, \tau_k)$. For the locally-linear case quantile regression could be used to set thresholds at local estimates of conditional quantiles of the response given the covariate.

Multivariate extremes

So far, we have considered only cases where the extremes of a single variable are of interest. In some applications it is important to model the joint behaviour of the extremes of different variables, in addition to their marginal extremal behaviours. This is achieved using multivariate extreme value models.

A threshold-based example is a bivariate threshold excess model (see, for example, Coles (2001, chapter 8)), where for pairs of random variables (X, Y) the aim is to model the joint distribution of (X, Y) on regions where both X and Y exceed a threshold u_x and u_y respectively. However, even in the bivariate case, multivariate extreme value theory generates a very wide class of extreme value models, meaning that many different bivariate extreme value models can be specified. A key issue is the distinction between the cases of asymptotic dependence, where there is a positive probability that the very largest values of X and Y occur at the same time and asymptotic independence, where this probability is zero. Some dependence will typically be observed in data. Then the crucial issue is to infer how the dependence changes as the levels of interest increases, that is, as we move into the upper tails of the distributions of X and Y, and, in particular, whether the variables are dependent or independent asymptotically. An area of current research is to develop modelling frameworks that are sufficiently flexible to be able to capture different forms of dependence and, beyond the bivariate case, to allow different types of dependence between different pairs of variables. Perhaps the most useful current threshold-based approach (Heffernan and Tawn, 2004, Keef et al., 2013) is one based on the conditional distribution of a vector given that one of its components exceeds some threshold.

For a given multivariate extreme value model the general principles of our CV approach could be applied to inform threshold selection. Of course, in more than one dimension the implementation is more challenging: more complicated models with more parameters and a threshold to set for each variable. As in the explicit modelling of temporal dependence of threshold excesses, we could also average inferences over different multivariate extreme value models. Indeed Bayesian Model Averaging has been used recently by Sabourin et al. (2013) to combine inferences from different trivariate extreme value models for block maxima.

A CHAPTER 1

A.1 Likelihood-based results

The probability density function of a GEV(μ, σ, ξ) random variable is

$$f_{GEV}(y;\boldsymbol{\theta}) = \begin{cases} \dfrac{1}{\sigma}\exp\left\{-\left[1+\xi\left(\dfrac{y-\mu}{\sigma}\right)\right]_+^{-1/\xi}\right\}\left[1+\xi\left(\dfrac{y-\mu}{\sigma}\right)\right]_+^{-1-1/\xi}, & \xi \neq 0, \\ \dfrac{1}{\sigma}\exp\left\{-\left(\dfrac{y-\mu}{\sigma}\right)-\exp\left[-\left(\dfrac{y-\mu}{\sigma}\right)\right]\right\}, & \xi = 0, \end{cases} \tag{A.1}$$

where $\boldsymbol{\theta} = (\mu, \sigma, \xi)$.

The log-likelihood for a random sample $\boldsymbol{Y} = Y_1, \ldots, Y_b$ from a GEV(μ, σ, ξ) distribution is given by

$$\begin{aligned} \ell_{GEV}(\boldsymbol{\theta};\boldsymbol{y}) &= \sum_{i=1}^{b}\log f_{GEV}(y_i;\boldsymbol{\theta}) \\ &= \begin{cases} -b\log\sigma - \left(1+\dfrac{1}{\xi}\right)\displaystyle\sum_{i=1}^{b}\log\left[1+\xi\left(\dfrac{y_i-\mu}{\sigma}\right)\right]_+ - \displaystyle\sum_{i=1}^{b}\left[1+\xi\left(\dfrac{y_i-\mu}{\sigma}\right)\right]_+^{-1/\xi}, & \xi \neq 0, \\ -b\log\sigma - \displaystyle\sum_{i=1}^{b}\left(\dfrac{y_i-\mu}{\sigma}\right) - \displaystyle\sum_{i=1}^{b}\exp\left\{-\left(\dfrac{y_i-\mu}{\sigma}\right)\right\}, & \xi = 0. \end{cases} \end{aligned} \tag{A.2}$$

The expected Fisher Information matrix for $\boldsymbol{\theta}$ based on a single observation is given by

$$FI_{GEV} = \begin{pmatrix} -E\left[\dfrac{\partial^2\ell(\boldsymbol{\theta};\boldsymbol{Y})}{\partial\mu^2}\right] & -E\left[\dfrac{\partial^2\ell(\boldsymbol{\theta};\boldsymbol{Y})}{\partial\mu\partial\sigma}\right] & -E\left[\dfrac{\partial^2\ell(\boldsymbol{\theta};\boldsymbol{Y})}{\partial\mu\partial\xi}\right] \\ -E\left[\dfrac{\partial^2\ell(\boldsymbol{\theta};\boldsymbol{Y})}{\partial\mu\partial\sigma}\right] & -E\left[\dfrac{\partial^2\ell(\boldsymbol{\theta};\boldsymbol{Y})}{\partial\sigma^2}\right] & -E\left[\dfrac{\partial^2\ell(\boldsymbol{\theta};\boldsymbol{Y})}{\partial\sigma\partial\xi}\right] \\ -E\left[\dfrac{\partial^2\ell(\boldsymbol{\theta};\boldsymbol{Y})}{\partial\mu\partial\xi}\right] & -E\left[\dfrac{\partial^2\ell(\boldsymbol{\theta};\boldsymbol{Y})}{\partial\sigma\partial\xi}\right] & -E\left[\dfrac{\partial^2\ell(\boldsymbol{\theta};\boldsymbol{Y})}{\partial\xi^2}\right] \end{pmatrix}. \tag{A.3}$$

For $\xi \neq 0$ the FI_{GEV} components can be expressed in terms of the gamma function

$\Gamma(t) = \int_0^\infty x^{t-1} \exp(-x)\,\mathrm{d}x$ and the digamma function $\psi(t) = \mathrm{d}\log\Gamma(t)/\mathrm{d}t$ as

$$\begin{aligned}
-E\left[\frac{\partial^2 \ell(\boldsymbol{\theta};\boldsymbol{Y})}{\partial\mu^2}\right] &= \frac{p}{\sigma^2}\\
-E\left[\frac{\partial^2 \ell(\boldsymbol{\theta};\boldsymbol{Y})}{\partial\sigma^2}\right] &= \frac{\{1-2\Gamma(2+\xi)+p\}}{\sigma^2\xi^2}\\
-E\left[\frac{\partial^2 \ell(\boldsymbol{\theta};\boldsymbol{Y})}{\partial\xi^2}\right] &= \frac{\{\pi^2/6+(1-\gamma+1/\xi)^2-2q/\xi+p/\xi^2\}}{\xi^2}\\
-E\left[\frac{\partial^2 \ell(\boldsymbol{\theta};\boldsymbol{Y})}{\partial\mu\partial\sigma}\right] &= -\frac{\{p-\Gamma(2+\xi)\}}{\sigma^2\xi}\\
-E\left[\frac{\partial^2 \ell(\boldsymbol{\theta};\boldsymbol{Y})}{\partial\mu\partial\xi}\right] &= \frac{\{q-p/\xi\}}{\sigma\xi}\\
-E\left[\frac{\partial^2 \ell(\boldsymbol{\theta};\boldsymbol{Y})}{\partial\sigma\partial\xi}\right] &= \frac{[1-\gamma+\{1-\Gamma(2+\xi)\}/\xi-q+p/\xi]}{\sigma\xi^2}
\end{aligned}$$

where,

$$\begin{aligned}
p &= (1+\xi)^2\Gamma(1+2\xi),\\
q &= \Gamma(2+\xi)\{\psi(1+\xi)+(1+\xi)/\xi\}
\end{aligned}$$

and $\gamma = -\psi(1) \approx 0.5772157$ is Euler's constant. Prescott and Walden (1980) give the expected information matrix for a parameterisation of the GEV distribution in which the shape parameter $k = -\xi$. For $\xi = 0$ the expressions given above become

$$\begin{aligned}
-E\left[\frac{\partial^2 \ell(\boldsymbol{\theta};\boldsymbol{Y})}{\partial\mu^2}\right] &= \frac{1}{\sigma^2}\\
-E\left[\frac{\partial^2 \ell(\boldsymbol{\theta};\boldsymbol{Y})}{\partial\sigma^2}\right] &= \frac{\pi^2/6+(1-\gamma)^2}{\sigma^2}\\
-E\left[\frac{\partial^2 \ell(\boldsymbol{\theta};\boldsymbol{Y})}{\partial\xi^2}\right] &= \pi^2/6-\pi^2\gamma/2+\gamma^2-\gamma^3-2\zeta(3)+2\gamma\zeta(3)+\pi^2\gamma^2/4+\gamma^4/4+3\pi^4/80,\\
-E\left[\frac{\partial^2 \ell(\boldsymbol{\theta};\boldsymbol{Y})}{\partial\mu\partial\sigma}\right] &= \frac{\gamma-1}{\sigma^2}\\
-E\left[\frac{\partial^2 \ell(\boldsymbol{\theta};\boldsymbol{Y})}{\partial\mu\partial\xi}\right] &= \frac{\pi^2/6+\gamma^2-2\gamma}{2\sigma}\\
-E\left[\frac{\partial^2 \ell(\boldsymbol{\theta};\boldsymbol{Y})}{\partial\sigma\partial\xi}\right] &= \frac{4\gamma+4\zeta(3)+\pi^2\gamma+2\gamma^3-\pi^2-6\gamma^2}{4\gamma}
\end{aligned}$$

where $\zeta(x) = \sum_{i=1}^\infty 1/i^x$ is the Riemann Zeta function and $\zeta(3) \approx 1.2020569$.

A.2 Summary of results for GP distribution

We use the results from section 1.3 to show that the random variable Z has the following distribution function

$$F_{GP}(z) = \begin{cases} 1 - \left[1 + \dfrac{\xi z}{\sigma_u}\right]_+^{-1/\xi}, & \xi \neq 0, \\ 1 - \exp\left(-\dfrac{z}{\sigma_u}\right), & \xi = 0, \end{cases} \tag{A.4}$$

where $\sigma_u > 0$ and $\xi \in \mathbb{R}$.

A.3 Likelihood-based results

The probability density function of a GP(σ_u, ξ) random variable is

$$f_{GP}(z; \boldsymbol{\theta}) = \begin{cases} \dfrac{1}{\sigma_u}\left[1 + \dfrac{\xi z}{\sigma_u}\right]_+^{-1-1/\xi}, & \xi \neq 0, \\ \dfrac{1}{\sigma_u}\exp\left(-\dfrac{z}{\sigma_u}\right), & \xi = 0, \end{cases} \tag{A.5}$$

where now $\boldsymbol{\theta} = (\sigma_u, \xi)$.

The log-likelihood for a random sample $\boldsymbol{Z} = (Z_1, \ldots, Z_{n_u})$ from a GP(σ_u, ξ) distribution is given by

$$\begin{aligned} \ell_{GP}(\boldsymbol{\theta}; \boldsymbol{z}) &= \sum_{i=1}^{n_u} \log f_{GP}(z_i; \boldsymbol{\theta}) \\ &= \begin{cases} -n_u \log \sigma_u - (1 + 1/\xi) \displaystyle\sum_{i=1}^{n_u} \log \left[1 + \xi\left(\frac{z_i}{\sigma_u}\right)\right]_+, & \xi \neq 0, \\ -n_u \log \sigma_u - \sigma_u^{-1} \displaystyle\sum_{i=1}^{n_u} z_i, & \xi = 0. \end{cases} \end{aligned} \tag{A.6}$$

The expected Fisher Information matrix for $\boldsymbol{\theta}$ based on $\boldsymbol{Z}$ is given by

$$\begin{aligned} FI_{GP} &= \begin{pmatrix} -E\left[\dfrac{\partial^2 \ell(\boldsymbol{\theta};\boldsymbol{Z})}{\partial \sigma_u^2}\right] & -E\left[\dfrac{\partial^2 \ell(\boldsymbol{\theta};\boldsymbol{Z})}{\partial \sigma_u \partial \xi}\right] \\ -E\left[\dfrac{\partial^2 \ell(\boldsymbol{\theta};\boldsymbol{Z})}{\partial \sigma_u \partial \xi}\right] & -E\left[\dfrac{\partial^2 \ell(\boldsymbol{\theta};\boldsymbol{Z})}{\partial \xi^2}\right] \end{pmatrix} \\ &= \frac{n_u}{\sigma_u^2(1+\xi)(1+2\xi)} \begin{pmatrix} 1+\xi & \sigma_u \\ \sigma_u & 2\sigma_u^2 \end{pmatrix}. \end{aligned} \tag{A.7}$$

A.4 Summary of results for the K-gaps mixture model

For a sequence of dependent random variables $\widetilde{X}_1, \widetilde{X}_2, \ldots, \widetilde{X}_n$ and a suitably high threshold u we consider the behaviour of the scaled K-gap $Z = \overline{F}(u)S^{(K)}$. We have used that $\overline{F}(u) = P\left(\widetilde{X} > u\right)$, which is estimated by $q = (1/n)\sum_{i=1}^{n} I(\widetilde{X}_i > u)$ and $S^{(K)} = \max(T-K, 0)$, where T is the inter-exceedance time and K is a tuning parameter. Furthermore, without loss of generality, we let $N-1$ be the total number of K-gaps and define $N_1 = \sum_{i=1}^{N-1} I(S_i > 0)$ to be the number of non-zero K-gaps and $N_0 = \sum_{i=1}^{N-1} I(S_i = 0) = N - 1 - N_1$ the number of zero K-gaps.

The probability density function of Z is

$$f_Z(z) = (1-\theta)^{I(z=0)}(\theta^2 \mathrm{e}^{-\theta z})^{I(z>0)} \quad \text{for} \quad z \geqslant 0, \tag{A.8}$$

where I is the indicator function and the condition $z > 0$ is essentially dependent on $s > 0$ since $\overline{F}(u) > 0$ is always positive for an exceedance.

For the model parameter θ and $\boldsymbol{Z} = (Z_1, \ldots, Z_{N-1})$, assuming independence of $\boldsymbol{Z}$, the log-likelihood function for the K-gaps exponential mixture model is

$$l_K(\theta;\boldsymbol{Z}) = N_0 \log(1-\theta) + 2N_1 \log\theta - \theta q \left\{\sum_{i=1}^{N-1} Z_i\right\}. \tag{A.9}$$

The expected Fisher Information for the K-gaps exponential mixture model is

$$\begin{aligned} FI_K &= -E\left[\frac{\partial^2 \ell(\theta;\boldsymbol{Z})}{\partial\theta^2}\right] \\ &= E\left[\sum_{i=1}^{N-1}\left\{\frac{I(Z_i=0)}{(1-\theta)^2} + \frac{2I(Z_i>0)}{\theta^2}\right\}\right] && \text{(A.10)} \\ &= (N-1)\left[\frac{1}{1-\theta} + \frac{2}{\theta}\right]. && \text{(A.11)} \end{aligned}$$

B CHAPTER 2

B.1 Moments of a GP distribution

We give some moments of the GP distribution for later use. Suppose that $Z \sim GP(\sigma, \xi)$, where $\xi < 1/r$. Then (Giles et al., 2011)

$$\mathrm{E}(Z^r) \;=\; \frac{r!\,\sigma^r}{\displaystyle\prod_{i=1}^{r}(1-i\xi)}, \qquad r = 1, 2, \ldots. \tag{B.1}$$

Now suppose that $\xi < 0$. Then, for a constant $a > \xi$, and using the substitution $x = -\xi v/\sigma$, we have

$$\begin{aligned}
\mathrm{E}(Z^{-a/\xi}) &= \int_0^{-\sigma/\xi} v^{-a/\xi}\frac{1}{\sigma}\left(1+\frac{\xi v}{\sigma}\right)^{-(1+1/\xi)} \mathrm{d}v, \\
&= (-\xi)^{a/\xi-1}\sigma^{-a/\xi}\int_0^1 x^{-a/\xi}(1-x)^{-(1+1/\xi)}\,\mathrm{d}x, \\
&= (-\xi)^{a/\xi-1}\sigma^{-a/\xi}\frac{\Gamma(1-a/\xi)\Gamma(-1/\xi)}{\Gamma(1-(a+1)/\xi)},
\end{aligned} \tag{B.2}$$

where we have used integral number 1 in section 3.251 on page 324 of Gradshteyn and Ryzhik (2007), namely

$$\int_0^1 x^{\mu-1}(1-x^\lambda)^{\nu-1}\,\mathrm{d}x = \frac{1}{\lambda}\mathrm{Beta}\left(\frac{\mu}{\lambda},\nu\right) = \frac{\Gamma(\mu/\lambda)\Gamma(\nu)}{\Gamma(\mu/\lambda+\nu)} \qquad \lambda > 0, \nu > 0, \mu > 0,$$

with $\lambda = 1, \mu = 1 - a/\xi$ and $v = -1/\xi$.

In the following proofs we use the generic notation $\pi(\xi)$ for the component of the prior relating to ξ: the form of $\pi(\xi)$ varies depending on the prior being considered.

B.2 Proof of theorem 6 and its corollary

This trivial extension of the proof of theorem 1 in Eugenia Castellanos and Cabras (2007). Suppose $n_u = 1$, with an observation z. The normalizing constant C of the posterior distribution is given by

$$\begin{aligned}
C_1 &= \int_{-\infty}^{0}\pi(\xi)\int_{-\xi z}^{\infty}\sigma^{-2}(1+\xi z/\sigma)^{-(1+1/\xi)}\,\mathrm{d}\sigma\mathrm{d}\xi + \int_0^\infty \pi(\xi)\int_0^\infty \sigma^{-2}(1+\xi z/\sigma)^{-(1+1/\xi)}\,\mathrm{d}\sigma\mathrm{d}\xi, \\
&= \frac{1}{z}\int_{-\infty}^{\infty}\pi(\xi)\,\mathrm{d}\xi.
\end{aligned}$$

If the latter integral is finite, that is, $\pi(\xi)$ is proportional to a proper density function, then the posterior distribution is proper for $n_u = 1$ and therefore, by successive iterations of Bayes' theorem, it is proper for $n_u \geqslant 1$.

The corollary follows directly. □

B.3 Proof of theorem 7

Let $A(\xi) = \mathrm{e}^{-\xi}$ and $B(\sigma,\xi) = \sigma^{-(n_u+1)} \prod_{i=1}^{n_u} (1+\xi z_i/\sigma)^{-(1+1/\xi)}$. Then, from (2.6) we have

$$\begin{aligned} C_N &= \int_{-\infty}^{\infty} A(\xi) \int_{\max(0,-\xi z_{n_u})}^{\infty} B(\sigma,\xi)\ \mathrm{d}\sigma \mathrm{d}\xi, \\ &= \int_{-\infty}^{-1} A(\xi) \int_{-\xi z_{n_u}}^{\infty} B(\sigma,\xi)\ \mathrm{d}\sigma\mathrm{d}\xi + \int_{-1}^{0} A(\xi) \int_{-\xi z_{n_u}}^{\infty} B(\sigma,\xi)\ \mathrm{d}\sigma\mathrm{d}\xi + \int_{0}^{\infty} A(\xi) \int_{0}^{\infty} B(\sigma,\xi)\ \mathrm{d}\sigma\mathrm{d}\xi. \end{aligned}$$

The latter two integrals converge for $n_u \geqslant 1$. However, the first integral diverges for all samples sizes. For $\xi < -1$, $(1+\xi z/\sigma)^{-(1+1/\xi)} > 1$ when z is in the support $(0, -\sigma/\xi)$ of the GP(σ,ξ) density. Therefore $B(\sigma,\xi) > \sigma^{-(n_u+1)}$. Thus, the first integral above satisfies

$$\begin{aligned} \int_{-\infty}^{-1} A(\xi) \int_{-\xi z_{n_u}}^{\infty} B(\sigma,\xi)\ \mathrm{d}\sigma\ \mathrm{d}\xi &> \int_{-\infty}^{-1} A(\xi) \int_{-\xi z_{n_u}}^{\infty} \sigma^{-(n_u+1)}\ \mathrm{d}\sigma\ \mathrm{d}\xi, \\ &= \int_{-\infty}^{-1} A(\xi) \left[-\frac{1}{n_u} \sigma^{-n_u} \right]_{-\xi z_{n_u}}^{\infty} \mathrm{d}\xi, \\ &= \int_{-\infty}^{-1} A(\xi) \frac{1}{n_u} \left[-\xi z_{n_u} \right]^{-n_u}\ \mathrm{d}\xi, \\ &= \frac{1}{n_u z_{n_u}^{n_u}} \int_{1}^{\infty} v^{-n_u} \mathrm{e}^{v}\ \mathrm{d}v, \end{aligned}$$

where $v = -\xi$. This integral is divergent for all $n_u \geqslant 1$, so there is no sample size for which the posterior is proper. □

B.4 Proof of theorem 8

We need to show that C_3 is finite. We split the range of integration over ξ so that $C_3 = I_1 + I_2 + I_3$, where

$$I_1 = \int_{-\infty}^{-1} \int_{-\xi z_3}^{\infty} B(\sigma,\xi)\ \mathrm{d}\sigma\, \mathrm{d}\xi, \quad I_2 = \int_{-1}^{0} \int_{-\xi z_3}^{\infty} B(\sigma,\xi)\ \mathrm{d}\sigma\, \mathrm{d}\xi, \quad I_3 = \int_{0}^{\infty} \int_{0}^{\infty} B(\sigma,\xi)\ \mathrm{d}\sigma\, \mathrm{d}\xi$$

and $B(\sigma,\xi) = \sigma^{-4} \prod_{i=1}^{3} (1+\xi z_i/\sigma)^{-(1+1/\xi)}$. For convenience we let $\rho = \xi/\sigma$.

B.4.1 Proof that I_1 is finite

We have $\xi < -1$ and so $-(1+1/\xi) < 0$, $\rho < 0$ and $0 < 1+\rho z_i < 1$ for $i = 1, 2, 3$. Noting that $-\rho z_3 < 1$ gives

$$\begin{aligned}(1+\rho z_1)(1+\rho z_2)(1+\rho z_3) &> (-\rho z_3+\rho z_1)(-\rho z_3+\rho z_2)(1+\rho z_3),\\ &= (-\rho)^2(z_3-z_1)(z_3-z_2)(1+\rho z_3),\\ &= (-\xi)^2\sigma^{-2}(z_3-z_1)(z_3-z_2)(1+\rho z_3). \quad \text{(B.3)}\end{aligned}$$

Therefore,

$$\prod_{i=1}^{3}\left(1+\frac{\xi z_i}{\sigma}\right)^{-(1+1/\xi)} < (-\xi)^{-2(1+1/\xi)}\sigma^{2(1+1/\xi)}\left[(z_3-z_2)(z_3-z_1)\left(1+\frac{\xi z_3}{\sigma}\right)\right]^{-(1+1/\xi)}.$$

Thus,

$$I_1 \leqslant \int_{-\infty}^{-1}(-\xi)^{-2(1+1/\xi)}\left[(z_3-z_2)(z_3-z_1)\right]^{-(1+1/\xi)} I_{1\sigma}\,\mathrm{d}\xi,$$

where

$$\begin{aligned}I_{1\sigma} &= \int_{-\xi z_3}^{\infty}\sigma^{-4}\sigma^{2(1+1/\xi)}\left(1+\frac{\xi z_3}{\sigma}\right)^{-(1+1/\xi)}\mathrm{d}\sigma,\\ &= z_3^{-1}\int_{0}^{-1/\xi z_3} v^{-2/\xi}\frac{1}{z_3^{-1}}\left(1+\frac{\xi v}{z_3^{-1}}\right)^{-(1+1/\xi)}\mathrm{d}v,\\ &= (-\xi)^{2/\xi-1}z_3^{-(1-2/\xi)}\,\frac{\Gamma(1-2/\xi)\Gamma(-1/\xi)}{\Gamma(1-3/\xi)},\end{aligned}$$

where $v = 1/\sigma$ and the last line follows from (B.2) with $a = 2$ and $\sigma = z_3^{-1}$. Therefore,

$$\begin{aligned}I_1 &\leqslant \int_{-\infty}^{-1}(-\xi)^{-3}\left[(z_3-z_2)(z_3-z_1)\right]^{-(1+1/\xi)} z_3^{-(1-2/\xi)}\frac{\Gamma(1-2/\xi)\Gamma(-1/\xi)}{\Gamma(1-3/\xi)}\,\mathrm{d}\xi,\\ &= \left[z_3(z_3-z_2)(z_3-z_1)\right]^{-1}\int_{-\infty}^{-1}(-\xi)^{-3}\left(1-\frac{z_2}{z_3}\right)^{-1/\xi}\left(1-\frac{z_1}{z_3}\right)^{-1/\xi}\frac{\Gamma(1-2/\xi)\Gamma(-1/\xi)}{\Gamma(1-3/\xi)}\,\mathrm{d}\xi,\\ &= \left[z_3(z_3-z_2)(z_3-z_1)\right]^{-1}\int_{0}^{1}x\left(1-\frac{z_2}{z_3}\right)^{x}\left(1-\frac{z_1}{z_3}\right)^{x}\frac{\Gamma(1+2x)\Gamma(x)}{\Gamma(1+3x)}\,\mathrm{d}x,\\ &= \left[z_3(z_3-z_2)(z_3-z_1)\right]^{-1}\int_{0}^{1}\left(1-\frac{z_2}{z_3}\right)^{x}\left(1-\frac{z_1}{z_3}\right)^{x}\frac{\Gamma(1+2x)\Gamma(1+x)}{\Gamma(1+3x)}\,\mathrm{d}x, \quad \text{(B.4)}\end{aligned}$$

where $x = -1/\xi$ and we have used the relation $\Gamma(1+x) = x\,\Gamma(x)$. The integrand in (B.4) is finite over the range of integration so this integral is finite and therefore I_1 is finite.

B.4.2 Proof that I_2 is finite

We have $-1 < \xi < 0$, so $-(1+1/\xi) > 0$ and $(1+\xi z/\sigma)^{-(1+1/\xi)} < 1$ and decreases in z over $(0, -\sigma/\xi)$. Therefore,

$$\begin{aligned}
I_2 &= \int_{-1}^{0}\int_{-\xi z_3}^{\infty} \sigma^{-4}\prod_{i=1}^{3}\left(1+\frac{\xi z_i}{\sigma}\right)^{-(1+1/\xi)} \mathrm{d}\sigma\ \mathrm{d}\xi,\\
&\leqslant \int_{-1}^{0}\int_{-\xi z_3}^{\infty} \sigma^{-4}\left(1+\frac{\xi z_3}{\sigma}\right)^{-(1+1/\xi)} \mathrm{d}\sigma\ \mathrm{d}\xi,\\
&= \int_{-1}^{0} z_3^{-1}\int_{0}^{-1/\xi z_3} v^2\frac{1}{z_3^{-1}}\left(1+\frac{\xi v}{z_3^{-1}}\right)^{-(1+1/\xi)} \mathrm{d}v\ \mathrm{d}\xi,\\
&= z_3^{-1}\int_{-1}^{0}\frac{2z_3^{-2}}{(1-\xi)(1-2\xi)}\ \mathrm{d}\xi,\\
&= 2z_3^{-3}\int_{-1}^{0}\left\{\left(\frac{1}{2}-\xi\right)^{-1}-(1-\xi)^{-1}\right\}\ \mathrm{d}\xi,\\
&= 2z_3^{-3}\ln(3/2),
\end{aligned}$$

where the integral over v follows from (B.1) with $r = 2$ and $\sigma = z_3^{-1}$.

B.4.3 Proof that I_3 is finite

We have $\xi > 0$ so $-(1+1/\xi) < 0$. Let $g_n = (\prod_{i=1}^{n} z_i)^{1/n}$. Mitrinović (1964, page 130):

$$\prod_{k=1}^{n}(1+a_k) \geqslant (1+b)^n, \qquad a_k > 0; \quad \prod_{k=1}^{n} a_k = b^n, \tag{B.5}$$

with $a_k = \xi z_k/\sigma$ and $b = \xi g_3/\sigma$ gives

$$\prod_{i=1}^{3}\left(1+\frac{\xi z_i}{\sigma}\right)^{-(1+1/\xi)} \leqslant \left(1+\frac{\xi g_3}{\sigma}\right)^{-3(1+1/\xi)},$$

and therefore

$$\begin{aligned}
I_3 &= \int_{0}^{\infty}\int_{0}^{\infty} \sigma^{-4}\prod_{i=1}^{3}\left(1+\frac{\xi z_i}{\sigma}\right)^{-(1+1/\xi)} \mathrm{d}\sigma\,\mathrm{d}\xi,\\
&\leqslant \int_{0}^{\infty}\int_{0}^{\infty} \sigma^{-4}\left(1+\frac{\xi g_3}{\sigma}\right)^{-3(1+1/\xi)} \mathrm{d}\sigma\,\mathrm{d}\xi,\\
&= \int_{0}^{\infty}\beta\int_{0}^{\infty} v^2\frac{1}{\beta}\left(1+\frac{\alpha v}{\beta}\right)^{-(1+1/\alpha)} \mathrm{d}v\,\mathrm{d}\xi,
\end{aligned}$$

where $v = 1/\sigma$, $\alpha = 1/(2 + 3/\xi)$ and $\beta = \alpha/\xi g_3 = 1/(3 + 2\xi)g_3$. For $\xi > 0$, $\alpha < 1/2$ so using (B.1) with $r = 2, \sigma = \beta$ and $\xi = \alpha$ gives

$$\begin{aligned} I_3 &\leqslant \int_0^\infty \beta \frac{2\beta^2}{(1-\alpha)(1-2\alpha)} \,\mathrm{d}\xi, \\ &= \frac{2}{3} g_3^{-3} \int_0^\infty \frac{1}{(\xi+3)(2\xi+3)} \,\mathrm{d}\xi, \\ &= \frac{2}{9} g_3^{-3} \int_0^\infty \left(\frac{1}{\xi + 3/2} - \frac{1}{\xi+3} \right) \,\mathrm{d}\xi, \\ &= \frac{2}{9} g_3^{-3} \ln 2. \end{aligned}$$

The normalizing constant C_3 is finite, so $\pi_{U,GP}(\sigma, \xi)$ yields a proper posterior density for $n_u = 3$ and therefore does so for $n_u \geqslant 3$. □

B.5 Proof of theorem 9 and its corollary

Throughout the following proofs we define $\delta_i = y_i - y_1, i = 2, \ldots, b$.

We make the parameter transformation $\phi = \mu - \sigma/\xi$. Then the posterior density for (ϕ, σ, ξ) is given by

$$\pi(\phi, \sigma, \xi) = K_b^{-1} \pi(\xi) |\xi|^{-b(1+1/\xi)} G_b(\phi, \sigma),$$

where

$$G_b(\phi, \sigma) = \sigma^{b/\xi - 1} \left\{ \prod_{i=1}^{b} |y_i - \phi|^{-(1+1/\xi)} \right\} \exp\left\{ -|\xi|^{-1/\xi} \sigma^{1/\xi} \sum_{i=1}^{b} |y_i - \phi|^{-1/\xi} \right\}$$

and, if $\xi > 0$ then $\phi < y_1$ and if $\xi < 0$ then $\phi > y_b$.

We let $w = |\xi|^{-1/\xi} \sum_{i=1}^{b} |y_i - \phi|^{-1/\xi}$ and $v = \sigma^{1/\xi}$. The normalizing constant K_b is given by

$$\begin{aligned}
K_b &= \int_{-\infty}^{\infty} \int \int_0^{\infty} \pi(\xi)|\xi|^{-b(1+1/\xi)} G_b(\phi,\sigma)\ \mathrm{d}\sigma\ \mathrm{d}\phi\ \mathrm{d}\xi, \\
&= \int_{-\infty}^{\infty} \pi(\xi)|\xi|^{-b(1+1/\xi)} \int \left\{ \prod_{i=1}^{b} |y_i-\phi|^{-(1+1/\xi)} \right\} \int_0^{\infty} \sigma^{b/\xi-1} \exp\left\{-w\sigma^{1/\xi}\right\}\ \mathrm{d}\sigma\ \mathrm{d}\phi\ \mathrm{d}\xi, \\
&= \int_{-\infty}^{\infty} \pi(\xi)|\xi|^{-b(1+1/\xi)} \int \left\{ \prod_{i=1}^{b} |y_i-\phi|^{-(1+1/\xi)} \right\} \int_0^{\infty} v^{b-1} \exp\{-wv\}\,|\xi|\ \mathrm{d}v\ \mathrm{d}\phi\ \mathrm{d}\xi, \\
&= \int_{-\infty}^{\infty} \pi(\xi)|\xi|^{-b(1+1/\xi)} \int \left\{ \prod_{i=1}^{b} |y_i-\phi|^{-(1+1/\xi)} \right\} \Gamma(b) w^{-b}\,|\xi|\ \mathrm{d}\phi\ \mathrm{d}\xi, \\
&= \int_{-\infty}^{\infty} \pi(\xi)|\xi|^{-b(1+1/\xi)} \int \left\{ \prod_{i=1}^{b} |y_i-\phi|^{-(1+1/\xi)} \right\} (b-1)!|\xi|^{b/\xi+1} \left\{ \sum_{i=1}^{b} |y_i-\phi|^{-1/\xi} \right\}^{-b} \mathrm{d}\phi\ \mathrm{d}\xi, \\
&= (b-1)! \int_{-\infty}^{\infty} \pi(\xi)|\xi|^{1-b} \int \left\{ \prod_{i=1}^{b} |y_i-\phi|^{-(1+1/\xi)} \right\} \left\{ \sum_{i=1}^{b} |y_i-\phi|^{-1/\xi} \right\}^{-b} \mathrm{d}\phi\ \mathrm{d}\xi, \qquad \text{(B.6)}
\end{aligned}$$

For $b = 1$ the integral $\int_{\phi:\xi(y_1-\phi)>0} |y_1-\phi|^{-1}\ \mathrm{d}\phi$ is divergent so if $b = 1$ the posterior is not proper for any prior in this class.

Now we take $b = 2$ and for clarity consider the cases $\xi > 0$ and $\xi < 0$ separately, with respective contributions K_2^+ and K_2^- to K_2. For $\xi > 0$, using the substitution $u = (y_1-\phi)^{-1}$ in (B.6) gives

$$\begin{aligned}
K_2^+ &= \int_0^{\infty} \pi(\xi)\,\xi^{-1} \int_{-\infty}^{y_1} \frac{(y_1-\phi)^{-(1+1/\xi)}(y_2-\phi)^{-(1+1/\xi)}}{\{(y_1-\phi)^{-1/\xi}+(y_2-\phi)^{-1/\xi}\}^2}\ \mathrm{d}\phi\ \mathrm{d}\xi, \\
&= \int_0^{\infty} \pi(\xi)\,\xi^{-1} \int_0^{\infty} \frac{(1+\delta_2 u)^{-(1+1/\xi)}}{\{1+(1+\delta_2 u)^{-1/\xi}\}^2}\ \mathrm{d}u\ \mathrm{d}\xi, \\
&= \frac{1}{2}\delta_2^{-1} \int_0^{\infty} \pi(\xi)\ \mathrm{d}\xi,
\end{aligned}$$

the final step following because the u-integrand is a multiple $(\xi\delta_2^{-1})$ of a shifted log-logistic density function with location, scale and shape parameters of $0, \xi\delta_2^{-1}$ and ξ respectively, and the location of this distribution equals the median. For $\xi < 0$ an analogous calculation using the substitution $v = (y_b-\phi)^{-1}$ in (B.6) gives

$$K_2^- = \frac{1}{2}\delta_2^{-1} \int_{-\infty}^{0} \pi(\xi)\ \mathrm{d}\xi.$$

Therefore,

$$K_2 = K_2^+ + K_2^- = \frac{1}{2}\delta_2^{-1} \int_{-\infty}^{\infty} \pi(\xi)\ \mathrm{d}\xi.$$

Thus, K_2 is finite if $\int_{-\infty}^{\infty} \pi(\xi)\, \mathrm{d}\xi$ is finite, and the result follows.

The corollary follows directly. □

B.6 Proof of theorem 10

The crucial aspects are the rates at which $\pi(\xi) \to \infty$ as $\xi \downarrow -1/2$ and as $\xi \to \infty$. The component $\pi(\xi)$ of (2.11) involving ξ can be expressed as

$$\pi^2(\xi) = \frac{1}{\xi^4}(T_1 + T_2), \tag{B.7}$$

where

$$T_1 = \left[\frac{\pi^2}{6} + (1-\gamma)^2\right](1+\xi)^2\,\Gamma(1+2\xi), \tag{B.8}$$

$$T_2 = \frac{\pi^2}{6} + \left[2(1-\gamma)(\gamma + \psi(1+\xi)) - \frac{\pi^2}{3}\right]\Gamma(2+\xi) - [1+\psi(1+\xi)]^2\,[\Gamma(2+\xi)]^2 \tag{B.9}$$

Firstly, we derive a lower bound for $\pi(\xi)$ that holds for $\xi > 3$. Using the duplication formula (Abramowitz and Stegun, 1972, page 256; 6.1.18)

$$\Gamma(2z) = (2\pi)^{-1/2}\, 2^{\,2z-1/2}\, \Gamma(z)\, \Gamma(z+1/2),$$

with $z = 1/2 + \xi$ in (B.8) we have

$$T_1 = \left[\frac{\pi^2}{6} + (1-\gamma)^2\right](1+\xi)^2\, \pi^{-1/2} 2^{2\xi}\, \Gamma(1/2+\xi)\, \Gamma(1+\xi).$$

We note that

$$\Gamma(1/2+\xi) = \frac{\Gamma(3/2+\xi)}{1/2+\xi} > \frac{\Gamma(1+\xi)}{1/2+\xi} = \frac{2\Gamma(1+\xi)}{1+2\xi} > \frac{\Gamma(1+\xi)}{1+\xi},$$

where for the first inequality to hold it is sufficient that $\xi > 1/2$; and that, for $\xi > 3$, $2^{2\xi} > (1+\xi)^3$. Therefore,

$$T_1 > \left[\frac{\pi^2}{6} + (1-\gamma)^2\right] \pi^{-1/2}\, (1+\xi)^4\, [\Gamma(1+\xi)]^2. \tag{B.10}$$

Completing the square in (B.9) gives

$$T_2 = -\left\{[1+\psi(1+\xi)]\,\Gamma(2+\xi) + f(\xi)\right\}^2 + [f(\xi)]^2 + \pi^2/6,$$

where

$$f(\xi) = \frac{\pi^2/6 - (1-\gamma)(\gamma + \psi(1+\xi))}{1+\psi(1+\xi)} = \frac{\pi^2/6 + (1-\gamma)^2}{1+\psi(1+\xi)} - (1-\gamma)$$

and $[f(\xi)]^2 + \pi^2/6 > 0$.

For $\xi > 0$, $\psi(1+\xi)$ increases with ξ and so $f(\xi)$ decreases with ξ. Therefore, for $\xi > 3$, $f(\xi) < f(3) \approx 0.39$ and

$$T_2 \;\; > \;\; -\left\{[1+\psi(1+\xi)]\,\Gamma(2+\xi) + f(3)\right\}^2.$$

For $\xi > 0$, we have $\psi(1+\xi) < \ln(1+\xi) - (1+\xi)^{-1}/2$ (Qiu and Vuorinen, 2004, theorem C) and $\ln(1+\xi) \leqslant \xi$ (Abramowitz and Stegun, 1972, page 68; 4.1.33). Therefore, noting that $\Gamma(2+\xi) = (1+\xi)\,\Gamma(1+\xi)$ we have

$$T_2 \;\; > \;\; -\left\{(1+\xi)^2\,\Gamma(1+\xi) - \frac{1}{2}\Gamma(1+\xi) + f(3)\right\}^2.$$

For $\xi > 3$, $f(3) - \Gamma(1+\xi)/2 < 0$ so

$$T_2 > -(1+\xi)^4\,[\Gamma(1+\xi)]^2. \tag{B.11}$$

Substituting (B.10) and (B.11) in (B.7) gives, for $\xi > 3$,

$$\begin{aligned} \pi^2(\xi) \;\; &> \;\; \frac{(1+\xi)^4}{\xi^4}\left\{\left[\frac{\pi^2}{6} + (1-\gamma)^2\right]\pi^{-1/2} - 1\right\}[\Gamma(1+\xi)]^2, \\ &> \;\; c[\Gamma(1+\xi)]^2, \\ &> \;\; c(1+\xi)^{2(\lambda\xi-\gamma)} \end{aligned}$$

where $c = (4/3)^4\{[\pi^2/6 + (1-\gamma)^2]\pi^{-1/2} - 1\} \approx 0.0913$ and the final step uses the inequality $\Gamma(x) > x^{\lambda(x-1)-\gamma}$, for $x > 0$ (Alzer, 1999), where $\lambda = (\pi^2/6 - \gamma)/2 \approx 0.534$. Thus, a lower bound for the ξ component of the Jeffreys prior (2.11) is given by

$$\pi(\xi) \;\; > \;\; c^{1/2}(1+\xi)^{\lambda\xi-\gamma}, \qquad \text{for } \xi > 3. \tag{B.12}$$

[In fact, numerical work shows that this lower bound holds for $\xi > -1/2$.]

Let K_b^+ denote the contribution to K_b for $\xi > 3$. Using the substitution $u =$

$(y_1 - \phi)^{-1}$ in (B.6) gives

$$K_b^+ = (b-1)! \int_3^\infty \pi(\xi)\, \xi^{1-b} \int_0^\infty u^{b-2} \frac{\displaystyle\prod_{i=1}^{b} (1+\delta_i u)^{-(1+1/\xi)}}{\left\{1 + \displaystyle\sum_{i=2}^{b} (1+\delta_i u)^{-1/\xi}\right\}^b} \, \mathrm{d}u \, \mathrm{d}\xi. \qquad \text{(B.13)}$$

For $\xi > 0$ we have $1 + \sum_{i=2}^{b} (1+\delta_i u)^{-1/\xi} \leqslant b$ and $\prod_{i=1}^{b} (1+\delta_i u)^{-(1+1/\xi)} \geqslant (1 + \delta_b u)^{-(b-1)(1+1/\xi)}$. Applying these inequalities to (B.13) gives

$$\begin{aligned} K_b^+ &\geqslant b^{-b}(b-1)! \int_3^\infty \pi(\xi)\, \xi^{1-b} \int_0^\infty u^{b-2} (1+\delta_b u)^{-(b-1)(1+1/\xi)} \, \mathrm{d}u \ \mathrm{d}\xi, \\ &= b^{-b}(b-1)! \int_3^\infty \pi(\xi)\, \xi^{1-b} \beta \int_0^\infty u^{b-2} \frac{1}{\beta} \left(1 + \frac{\alpha u}{\beta}\right)^{-(1+1/\alpha)} \mathrm{d}u \ \mathrm{d}\xi, \qquad \text{(B.14)} \end{aligned}$$

where $\beta = \alpha/\delta_b$ and $\alpha = [b-2+(b-1)/\xi]^{-1}$ and $0 < \alpha < (b-2)^{-1}$. The u-integrand is the density function of a GP(β, α) distribution and so, using (B.1) with $r = b-2$, the integral over u is given by

$$(b-2)!\, \beta^{b-2} \prod_{i=1}^{b-2} \frac{1}{1 - i\alpha} = (b-2)!\, \xi^{b-2} \delta_b^{2-b} \prod_{i=1}^{b-2} \frac{1}{(b-2-i)\xi + b - 1}. \qquad \text{(B.15)}$$

Substituting (B.15) into (B.14) gives

$$\begin{aligned} K_b^+ &\geqslant b^{-b}(b-1)!(b-2)!\, \delta_b^{1-b} \int_3^\infty \frac{1}{(b-2)\xi + b - 1} \prod_{i=1}^{b-2} \frac{1}{(b-2-i)\xi + b - 1} \, \pi(\xi) \ \mathrm{d}\xi, \\ &= b^{-b}(b-1)!(b-2)!\, \delta_b^{1-b} \int_3^\infty \prod_{i=0}^{b-2} \frac{1}{(b-2-i)\xi + b - 1} \, \pi(\xi) \ \mathrm{d}\xi, \\ &= b^{-b}(b-1)!(b-2)!\, \delta_b^{1-b} (b-1)^{1-b} \int_3^\infty \prod_{i=0}^{b-2} \frac{1}{1 + \frac{i}{b-1}\xi} \, \pi(\xi) \ \mathrm{d}\xi, \\ &> C(b) \int_3^\infty \frac{1}{(1+\xi)^{b-2}} \, \pi(\xi) \ \mathrm{d}\xi, \end{aligned}$$

where $C(b) = b^{-b}(b-1)!(b-2)!\, \delta_b^{1-b}(b-1)^{1-b}$. Applying (B.12) gives

$$K_b^+ \ > \ C(b)\, c^{1/2} \int_3^\infty (1+\xi)^{2-b+\lambda\xi-\gamma} \ \mathrm{d}\xi.$$

For any sample size b the integrand $\to \infty$ as $\xi \to \infty$. Therefore, the integral diverges and the result follows. □

Now we derive an upper bound for $\pi(\xi)$ that applies for ξ close to $-1/2$. We note that for $-1/2 < \xi < 0$ we have $\Gamma(1+2\xi) = \Gamma(2+2\xi)/(1+2\xi) < (1+2\xi)^{-1}$. From (B.7) we have

$$\pi^2(\xi) \;=\; \left[\frac{\pi^2}{6} + (1-\gamma)^2\right]\left(\frac{1+\xi}{\xi^2}\right)^2 \Gamma(1+2\xi) + \frac{T_2}{\xi^4},$$

where $T_2 \to -3.039$ as $\xi \downarrow -1/2$. Noting that $(1+\xi)^2/\xi^4 \to 4$ as $\xi \downarrow -1/2$ shows that $\pi(\xi) < 2\left[\pi^2/6 + (1-\gamma)^2\right]^{1/2}(1+2\xi)^{-1/2}$ for $\xi \in (-1/2, -1/2+\epsilon)$, for some $\epsilon > 0$.

[In fact numerical work shows that $\epsilon \approx 1.29$.]

B.7 Proof of theorem 11

We show that the integral K_b^-, giving the contribution to the normalising constant from $\xi < -1$, diverges. From the proof of theorem 9 we have

$$K_b^- \;= (b-1)! \int_{-\infty}^{-1} \mathrm{e}^{-\gamma(1+\xi)}\,(-\xi)^{1-b} \int_{y_b}^{\infty} \left\{\prod_{i=1}^{b} |y_i - \phi|^{-(1+1/\xi)}\right\} \left\{\sum_{i=1}^{b} |y_i - \phi|^{-1/\xi}\right\}^{-b} \mathrm{d}\phi \; \mathrm{d}\xi.$$

For $\xi < -1$ we have $-(1+1/\xi) < 0$ and $-1/\xi > 0$. Therefore, for $i = 2, \ldots, b$, $(\phi - y_i)^{-(1+1/\xi)} > (\phi - y_1)^{-(1+1/\xi)}$ and $(\phi - y_i)^{-1/\xi} < (\phi - y_1)^{-1/\xi}$, and thus the ϕ-integrand is greater than $b^{-b}(\phi - y_1)^{-b}$. Therefore,

$$\begin{aligned} K_b^- \;&>\; (b-1)! \int_{-\infty}^{-1} \mathrm{e}^{-\gamma(1+\xi)}\,(-\xi)^{1-b} \int_{y_b}^{\infty} b^{-b}(\phi - y_1)^{-b} \; \mathrm{d}\phi \; \mathrm{d}\xi, \\ &=\; (b-1)!\, b^{-b}(b-1)^{-1}(y_b - y_1)^{1-b} \int_{-\infty}^{-1} \mathrm{e}^{-\gamma(1+\xi)}\,(-\xi)^{1-b} \; \mathrm{d}\xi, \\ &=\; (b-2)!\, b^{-b}(y_b - y_1)^{1-b}\mathrm{e}^{-\gamma} \int_{1}^{\infty} x^{1-b}\,\mathrm{e}^{\gamma x} \; \mathrm{d}x, \end{aligned}$$

where $x = -\xi$. For all samples sizes b this integral diverges so the result follows. □

B.8 Proof of theorem 12

We need to show that K_4 is finite. We split the range of integration over ξ in (B.6) so that $K_4 = J_1 + J_2 + J_3$, with respective contributions from $\xi < -1$, $-1 \leqslant \xi \leqslant 0$ and $\xi > 0$.

B.8.1 Proof that J_1 is finite

We use the substitution $u = (\phi - y_1)^{-1}$ in (B.6) to give

$$\begin{aligned}
J_1 &= 3! \int_{-\infty}^{-1} (-\xi)^{-3} \int_{y_4}^{\infty} \left\{ \prod_{i=1}^{4} (\phi - y_i)^{-(1+1/\xi)} \right\} \left\{ \sum_{i=1}^{4} (\phi - y_i)^{-1/\xi} \right\}^{-b} \, \mathrm{d}\phi \, \mathrm{d}\xi, \\
&= 3! \int_{-\infty}^{-1} (-\xi)^{-3} \int_{0}^{1/\delta_4} u^2 \prod_{i=2}^{4} (1 - \delta_i u)^{-(1+1/\xi)} \left\{ 1 + \sum_{i=2}^{4} (1 - \delta_i u)^{-1/\xi} \right\}^{-4} \, \mathrm{d}u \, \mathrm{d}\xi.
\end{aligned}$$

A similar calculation to (B.3) gives

$$\prod_{i=2}^{4} (1 - \delta_i u)^{-(1+1/\xi)} \leqslant u^{-2(1+1/\xi)} \left\{ \prod_{i=2}^{3} (\delta_4 - \delta_i) \right\}^{-(1+1/\xi)} (1 - \delta_4 u)^{-(1+1/\xi)}.$$

Noting also that $1 + \sum_{i=2}^{4} (1 - \delta_i u)^{-1/\xi} \geqslant 1$ we have

$$\begin{aligned}
J_1 &\leqslant 3! \int_{-\infty}^{-1} (-\xi)^{-3} \left\{ \prod_{i=2}^{3} (\delta_4 - \delta_i) \right\}^{-(1+1/\xi)} \int_{0}^{1/\delta_4} u^{-2/\xi} (1 - \delta_4 u)^{-(1+1/\xi)} \, \mathrm{d}u \, \mathrm{d}\xi, \\
&= 3! \int_{-\infty}^{-1} (-\xi)^{-3} \left\{ \prod_{i=2}^{3} (\delta_4 - \delta_i) \right\}^{-(1+1/\xi)} \beta \int_{0}^{1/\delta_4} u^{-2/\xi} \frac{1}{\beta} \left(1 + \frac{\xi u}{\beta} \right)^{-(1+1/\xi)} \, \mathrm{d}u \, \mathrm{d}\xi, \\
&= 3! \int_{-\infty}^{-1} (-\xi)^{-3} \left\{ \prod_{i=2}^{3} (\delta_4 - \delta_i) \right\}^{-(1+1/\xi)} \delta_b^{2/\xi - 1} \frac{\Gamma(1 - 2/\xi)\Gamma(-1/\xi)}{\Gamma(1 - 3/\xi)} \, \mathrm{d}\xi,
\end{aligned}$$

where $\beta = -\xi/\delta_4$ and the last line follows from (B.2) with $a = 2$ and $\sigma = \beta$.

Therefore,

$$\begin{aligned}
J_1 &\leqslant 3! \int_{-\infty}^{-1} (-\xi)^{-3} (y_4 - y_1)^{2/\xi - 1} \prod_{i=2}^{3} (y_4 - y_i)^{-(1+1/\xi)} \frac{\Gamma(1 - 2/\xi)\Gamma(-1/\xi)}{\Gamma(1 - 3/\xi)} \, \mathrm{d}\xi, \\
&= 3! \prod_{i=1}^{3} (y_4 - y_i)^{-1} \int_{-\infty}^{-1} (-\xi)^{-3} \left(\prod_{i=2}^{3} \frac{y_4 - y_i}{y_4 - y_1} \right)^{-1/\xi} \frac{\Gamma(1 - 2/\xi)\Gamma(-1/\xi)}{\Gamma(1 - 3/\xi)} \, \mathrm{d}\xi, \\
&= 3! \prod_{i=1}^{3} (y_4 - y_i)^{-1} \int_{0}^{1} x \left(\prod_{i=2}^{3} \frac{y_4 - y_i}{y_4 - y_1} \right)^{x} \frac{\Gamma(1 + 2x)\Gamma(x)}{\Gamma(1 + 3x)} \, \mathrm{d}x, \\
&= 3! \prod_{i=1}^{3} (y_4 - y_i)^{-1} \int_{0}^{1} \left(\prod_{i=2}^{3} \frac{y_4 - y_i}{y_4 - y_1} \right)^{x} \frac{\Gamma(1 + 2x)\Gamma(1 + x)}{\Gamma(1 + 3x)} \, \mathrm{d}x, \qquad \text{(B.16)}
\end{aligned}$$

where $x = -1/\xi$ and we have used the relation $\Gamma(1 + x) = x\,\Gamma(x)$. The integrand in (B.16) is finite over the range of integration so this integral is finite and therefore J_1 is finite.

B.8.2 Proof that J_2 is finite

Using the substitution $u = (\phi - y_1)^{-1}$ in (B.6) gives

$$J_2 = 3! \int_{-1}^{0} (-\xi)^{-3} \int_{0}^{1/\delta_4} u^2 \prod_{i=2}^{4} (1 - \delta_i u)^{-(1+1/\xi)} \left\{ 1 + \sum_{i=2}^{4} (1 - \delta_i u)^{-1/\xi} \right\}^{-4} \mathrm{d}u \, \mathrm{d}\xi.$$

For $-1 \leqslant \xi \leqslant 0$ we have $-(1 + 1/\xi) \geqslant 0$. Noting that $0 < 1 - \delta_i u < 1$ gives

$$\prod_{i=2}^{4} (1 - \delta_i u)^{-(1+1/\xi)} \leqslant (1 - \delta_4 u)^{-(1+1/\xi)}.$$

Noting also that $1 + \sum_{i=2}^{4} (1 - \delta_i u)^{-1/\xi} \geqslant 1$ we have

$$\begin{aligned} J_2 &\leqslant 3! \int_{-1}^{0} (-\xi)^{-3} \int_{0}^{1/\delta_4} u^2 (1 - \delta_4 u)^{-(1+1/\xi)} \, \mathrm{d}u \, \mathrm{d}\xi, \\ &= 3! \int_{-1}^{0} (-\xi)^{-3} \beta \int_{0}^{1/\delta_4} u^2 \frac{1}{\beta} \left(1 + \frac{\xi u}{\beta} \right)^{-(1+1/\xi)} \mathrm{d}u \, \mathrm{d}\xi, \\ &= 3! \delta_4^{-3} \int_{-1}^{0} \frac{2}{(1 - \xi)(1 - 2\xi)} \, \mathrm{d}\xi, \\ &= 12 (y_4 - y_1)^{-3} \ln(3/2) \end{aligned}$$

where $\beta = -\xi/\delta_4$ and the penultimate line follows from (B.2) with $r = 2$ and $\sigma = \beta$.

B.8.3 Proof that J_3 is finite

Using the substitution $u = (y_1 - \phi)^{-1}$ in (B.6) gives

$$\begin{aligned} J_3 &= 3! \int_{0}^{\infty} \xi^{-3} \int_{-\infty}^{y_1} \left\{ \prod_{i=1}^{4} (y_i - \phi)^{-(1+1/\xi)} \right\} \left\{ \sum_{i=1}^{4} (y_i - \phi)^{-1/\xi} \right\}^{-4} \mathrm{d}\phi \, \mathrm{d}\xi, \\ &= 3! \int_{0}^{\infty} \xi^{-3} \int_{0}^{\infty} u^2 \prod_{i=2}^{4} (1 + \delta_i u)^{-(1+1/\xi)} \left\{ 1 + \sum_{i=2}^{4} (1 + \delta_i u)^{-1/\xi} \right\}^{-4} \mathrm{d}u \, \mathrm{d}\xi. \end{aligned}$$

Noting that for $\xi > 0$ we have $-(1 + 1/\xi) < 0$, using (B.5) with $a_k = \delta_k u$ gives

$$\prod_{i=2}^{4} (1 + \delta_i u)^{-(1+1/\xi)} \leqslant (1 + gu)^{-3(1+1/\xi)},$$

where $g = (\delta_2\delta_3\delta_4)^{1/3}$. Noting also that $1 + \sum_{i=2}^{4}(1+\delta_i u)^{-1/\xi} \geqslant 1$ we have

$$\begin{aligned} J_3 &\leqslant 3! \int_0^\infty \xi^{-3} \int_0^\infty u^2(1+gu)^{-3(1+1/\xi)} \, \mathrm{d}u \, \mathrm{d}\xi, \\ &\leqslant 3! \int_0^\infty \xi^{-3}\beta \int_0^\infty u^2 \frac{1}{\beta}\left(1+\frac{\alpha u}{\beta}\right)^{-(1+1/\alpha)} \mathrm{d}u \, \mathrm{d}\xi, \end{aligned}$$

where $\alpha = \xi/(2\xi+3)$ and $\beta = \alpha/g$. Therefore, (B.1) with $r = 2$, $\sigma = \beta$ and $\xi = \alpha$ gives

$$\begin{aligned} J_3 &\leqslant 3! \int_0^\infty \xi^{-3}\beta \frac{2\beta^2}{(1-\alpha)(1-2\alpha)} \, \mathrm{d}\xi, \\ &= 4g^{-3} \int_0^\infty \frac{1}{(\xi+3)(2\xi+3)} \, \mathrm{d}\xi, \\ &= \frac{4}{3}g^{-3} \int_0^\infty \left(\frac{1}{\xi+3/2} - \frac{1}{\xi+3}\right) \mathrm{d}\xi, \\ &= \frac{4}{3}g^{-3}\ln 2. \end{aligned}$$

The normalizing constant K_4 is finite, so $\pi_{U,GEV}(\mu,\sigma,\xi)$ yields a proper posterior density for $b = 4$ and therefore does so for $b \geqslant 4$. □

C CHAPTER 4

C.1 Log-concavity of posterior distribution of the extremal index

In order to use the ARS method to sample from the posterior density of the extremal index, $\pi(\theta \mid \boldsymbol{S})$ needs to satisfy the condition that it is log-concave for all possible values of θ. In other words, we need to show that

$$\frac{\partial^2 \log \pi(\theta \mid \boldsymbol{S})}{\partial \theta^2} < 0 \quad \text{for} \quad 0 \leqslant \theta \leqslant 1.$$

The posterior density follows from (4.6)

$$\pi(\theta \mid \boldsymbol{S}) \propto (1-\theta)^{N_0}\theta^{2N_1+I_0+I_N}\mathrm{e}^{-\theta V},$$

where $V = q\sum_{i=0}^{N} S_i$.

Taking the log and differentiating with respect to θ twice it can be shown that

$$\frac{\partial^2 \log \pi(\theta \mid \boldsymbol{S})}{\partial \theta^2} = -\frac{N_0}{(1-\theta)^2} - \frac{2N_1 + I_0 + I_N}{\theta^2} < 0 \quad \text{for} \quad 0 \leqslant \theta \leqslant 1,$$

since it is not possible for both N_0 and $2N_1+I_0+I_N$ to be non-positive. Therefore, the ARS method could be used to sample from $\pi(\theta \mid \boldsymbol{S})$. In the case where $N_0 = 0$, we can sample directly from the gamma density with shape parameter $2N_1+I_0+I_N+1$ and rate parameter V.

D CHAPTER 5

D.1 Quantile regression

Quantile regression is attributed to Koenker and Bassett (1978), who extend quantile estimation to the situation where covariates are present. Quantiles of the conditional distribution of a response variable are estimated, expressed as a function of covariates.

Let us assume that we have a random sample $\{y_1, y_2, \ldots, y_m\}$ of size m from a random variable Y that has a cumulative distribution function $F_Y(y) = P(Y \leqslant y)$. The $100\tau\%$ quantile of F_Y is defined as

$$Q(\tau) = F_y^{-1}(\tau) = \inf\left\{y : F_Y(y) \geqslant \tau\right\}, \quad \text{where} \quad 0 < \tau < 1.$$

In classical linear regression, solving the problem of minimizing a sum of squared residuals results to the sample mean. Similarly, solving the problem of minimizing a sum of absolute residuals results to the median, which intuitively makes sense due to symmetry. Therefore, for the symmetrical case of minimizing the sum of absolute residuals ($\tau = 0.5$), the median solves for the scalar κ in

$$Q(0.5) = \text{argmin}_\kappa \sum_{i=1}^{m} |y_i - \kappa|.$$

However, for other quantiles, the $100\tau\%$ quantile of F_Y is found by solving the asymmetric problem

$$Q(\tau) = \text{argmin}_\kappa \sum_{i=1}^{m} \rho_\tau\left(y_i - \kappa\right),$$

where

$$\rho_\tau(z) = z(\tau - 1_{\{z<0\}}).$$

and $1_{\{z<0\}}$ is 1 if $z < 0$ and is 0 otherwise. Now let y^τ denote the conditional $100\tau\%$ quantile of F_Y such that it has the following linear form

$$y^\tau = x_i^{'}\beta_\tau + e_i, \quad i = 1, \ldots, m,$$

where the $100\tau\%$ quantile of $e_i = 0$. By replacing the scalar κ with the parametric function y^τ we can obtain estimates for the conditional quantiles. Therefore the regression parameter is estimated by solving

$$\widehat{\beta}_\tau = \text{argmin}_{\beta_\tau} \sum_{i=1}^{m} \rho_\tau \left(y_i - x_i^{'}\beta_\tau \right)$$

or equivalently by minimizing

$$\min_{\beta_\tau} \left\{ (1-\tau) \sum_{y_i < y_i^\tau} (y_i^\tau - y_i) + \tau \sum_{y_i > y_i^\tau} (y_i - y_i^\tau) \right\},$$

with respect to β_τ. The method for solving this minimization problem is by linear programming. The R package quantreg (Koenker, 2011) can easily be used in order to estimate the quantile regression parameter.

D.2 Observed information for the stationary NHPP model

Up to an additive constant, the negated log-likelihood for $\boldsymbol{\theta} = (\mu, \sigma, \xi)$ based on observations $\boldsymbol{Y} = (Y_1, \ldots, Y_m)$ is

$$-\ell(\boldsymbol{\theta}; \boldsymbol{Y}) \;=\; \frac{1}{n} \sum_{i=1}^{m} g(\boldsymbol{\theta}) + \sum_{i=1}^{m} \delta(Y_i > u)\, h(\boldsymbol{\theta}; Y_i)$$

where $\delta(x) = 1$ if x is true and is 0 otherwise,

$$g(\boldsymbol{\theta}) \;=\; \left[1 + \xi \left(\frac{u - \mu}{\sigma} \right) \right]_+^{-1/\xi} \quad \text{and} \quad h(\boldsymbol{\theta}; Y_i) \;=\; \log \sigma + \left(1 + \frac{1}{\xi} \right) \log \left[1 + \xi \left(\frac{Y_i - \mu}{\sigma} \right) \right]_+ .$$

The observed information matrix is

$$J(\boldsymbol{\theta}) = -\begin{pmatrix} \dfrac{\partial^2 \ell(\boldsymbol{\theta};\boldsymbol{Y})}{\partial \mu^2} & \dfrac{\partial^2 \ell(\boldsymbol{\theta};\boldsymbol{Y})}{\partial \mu \partial \sigma} & \dfrac{\partial^2 \ell(\boldsymbol{\theta};\boldsymbol{Y})}{\partial \mu \partial \xi} \\[2ex] \dfrac{\partial^2 \ell(\boldsymbol{\theta};\boldsymbol{Y})}{\partial \mu \partial \sigma} & \dfrac{\partial^2 \ell(\boldsymbol{\theta};\boldsymbol{Y})}{\partial \sigma^2} & \dfrac{\partial^2 \ell(\boldsymbol{\theta};\boldsymbol{Y})}{\partial \sigma \partial \xi} \\[2ex] \dfrac{\partial^2 \ell(\boldsymbol{\theta};\boldsymbol{Y})}{\partial \mu \partial \xi} & \dfrac{\partial^2 \ell(\boldsymbol{\theta};\boldsymbol{Y})}{\partial \sigma \partial \xi} & \dfrac{\partial^2 \ell(\boldsymbol{\theta};\boldsymbol{Y})}{\partial \xi^2} \end{pmatrix}.$$

Second-order partial derivatives

For convenience we drop the i subscripts from Y_i and consequently $h(\boldsymbol{\theta}; Y_i)$ and define:

$$g_j \;=\; [1+\xi w]_+^{-(j+1/\xi)} \quad \text{and} \quad h_j = [1+\xi V]_+^{-j}, \quad \text{for} \quad j = 1, 2.$$

where $w = (u-\mu)/\sigma$ and $V = (Y-\mu)/\sigma$.

Derivatives of $g(\boldsymbol{\theta})$

$$\frac{\partial^2 g(\boldsymbol{\theta})}{\partial \mu^2} = \frac{1}{\sigma^2}(\xi+1)\left[1+\xi\left(\frac{u-\mu}{\sigma}\right)\right]_+^{-\left(2+\frac{1}{\xi}\right)} = \frac{1}{\sigma^2}(\xi+1)g_2.$$

$$\begin{aligned}\frac{\partial^2 g(\boldsymbol{\theta})}{\partial \sigma^2} &= \frac{1}{\sigma^2}\left(\frac{u-\mu}{\sigma}\right)^2(\xi+1)\left[1+\xi\left(\frac{u-\mu}{\sigma}\right)\right]_+^{-\left(2+\frac{1}{\xi}\right)} - \frac{2}{\sigma^2}\left(\frac{u-\mu}{\sigma}\right)\left[1+\xi\left(\frac{u-\mu}{\sigma}\right)\right]_+^{-\left(1+\frac{1}{\xi}\right)} \\ &= \frac{1}{\sigma^2}w^2(\xi+1)g_2 - \frac{2}{\sigma^2}wg_1.\end{aligned}$$

$$\begin{aligned}\frac{\partial g(\boldsymbol{\theta})}{\partial \xi} &= \left[1+\xi\left(\frac{u-\mu}{\sigma}\right)\right]_+^{-1/\xi}\left\{\frac{1}{\xi^2}\log\left[1+\xi\left(\frac{u-\mu}{\sigma}\right)\right]_+ - \frac{1}{\xi}\left(\frac{u-\mu}{\sigma}\right)\left[1+\xi\left(\frac{u-\mu}{\sigma}\right)\right]_+^{-1}\right\} \\ &= g(\boldsymbol{\theta})\,k(\boldsymbol{\theta}), \quad \text{where} \quad k(\boldsymbol{\theta}) = \left\{\frac{1}{\xi^2}\log\left[1+\xi w\right] - \frac{1}{\xi}w\left[1+\xi w\right]^{-1}\right\}.\end{aligned}$$

Therefore

$$\frac{\partial^2 g(\boldsymbol{\theta})}{\partial \xi^2} = \frac{\partial g(\boldsymbol{\theta})}{\partial \xi}k(\boldsymbol{\theta}) + g(\boldsymbol{\theta})\frac{\partial k(\boldsymbol{\theta})}{\partial \xi},$$

where

$$\begin{aligned}\frac{\partial k(\boldsymbol{\theta})}{\partial \xi} &= -\frac{2}{\xi^3}\log\left[1+\xi\left(\frac{u-\mu}{\sigma}\right)\right]_+ + \frac{2}{\xi^2}\left(\frac{u-\mu}{\sigma}\right)\left[1+\xi\left(\frac{u-\mu}{\sigma}\right)\right]_+^{-1} + \frac{1}{\xi}\left(\frac{u-\mu}{\sigma}\right)^2\left[1+\xi\left(\frac{u-\mu}{\sigma}\right)\right]_+^{-2} \\ &= -\frac{2}{\xi^3}\log\left[1+\xi w\right] + \frac{2}{\xi^2}w\left[1+\xi w\right]^{-1} + \frac{1}{\xi}w^2\left[1+\xi w\right]^{-2}.\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 g(\boldsymbol{\theta})}{\partial \mu \partial \sigma} &= \frac{1}{\sigma^2}(\xi+1)\left(\frac{u-\mu}{\sigma}\right)\left[1+\xi\left(\frac{u-\mu}{\sigma}\right)\right]_+^{-1/\xi-2} - \frac{1}{\sigma^2}\left[1+\xi\left(\frac{u-\mu}{\sigma}\right)\right]_+^{-1/\xi-1} \\ &= \frac{1}{\sigma^2}(\xi+1)wg_2 - \frac{1}{\sigma^2}g_1.\end{aligned}$$

Using $\frac{\partial g(\boldsymbol{\theta})}{\partial \xi} = g(\boldsymbol{\theta})k(\boldsymbol{\theta})$ and differentiating $g(\boldsymbol{\theta})$ again with respect to μ,

$$\frac{\partial^2 g(\boldsymbol{\theta})}{\partial \mu \partial \xi} = \frac{\partial g(\boldsymbol{\theta})}{\partial \mu}k(\boldsymbol{\theta}) + g(\boldsymbol{\theta})\frac{\partial k(\boldsymbol{\theta})}{\partial \mu},$$

where

$$\frac{\partial g(\boldsymbol{\theta})}{\partial \mu} = \frac{1}{\sigma}\left[1+\xi\left(\frac{u-\mu}{\sigma}\right)\right]_+^{-\left(1+\frac{1}{\xi}\right)} = \frac{1}{\sigma}g_1$$

and

$$\frac{\partial k(\boldsymbol{\theta})}{\partial \mu} = -\frac{1}{\sigma}\left(\frac{u-\mu}{\sigma}\right)\left[1+\xi\left(\frac{u-\mu}{\sigma}\right)\right]_+^{-2} = -\frac{1}{\sigma}w\left[1+\xi w\right]^{-2}.$$

Therefore

$$\begin{aligned}\frac{\partial^2 g(\boldsymbol{\theta})}{\partial \mu \partial \xi} &= \frac{1}{\sigma}\left[1+\xi\left(\frac{u-\mu}{\sigma}\right)\right]_+^{-\left(1+\frac{1}{\xi}\right)}\left\{\frac{1}{\xi^2}\log\left[1+\xi\left(\frac{u-\mu}{\sigma}\right)\right]_+ - \left(1+\frac{1}{\xi}\right)\left(\frac{u-\mu}{\sigma}\right)\left[1+\xi\left(\frac{u-\mu}{\sigma}\right)\right]_+^{-1}\right\} \\ &= \frac{1}{\sigma}g_1\left\{\frac{1}{\xi^2}\log\left[1+\xi w\right] - \left(1+\frac{1}{\xi}\right)w\left[1+\xi w\right]^{-1}\right\}.\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 g(\boldsymbol{\theta})}{\partial \sigma \partial \xi} &= \left(\frac{u-\mu}{\sigma}\right)\frac{\partial^2 g(\boldsymbol{\theta})}{\partial \mu \partial \xi}, \quad \text{since} \quad \frac{\partial g(\boldsymbol{\theta})}{\partial \sigma} = \left(\frac{u-\mu}{\sigma}\right)\frac{\partial g(\boldsymbol{\theta})}{\partial \mu} \\ &= w\frac{\partial^2 g(\boldsymbol{\theta})}{\partial \mu \partial \xi}.\end{aligned}$$

Derivatives of $h(\boldsymbol{\theta}; Y)$

$$\frac{\partial^2 h(\boldsymbol{\theta};Y)}{\partial \mu^2} = -\frac{\xi}{\sigma^2}(\xi+1)\left[1+\xi\left(\frac{Y-\mu}{\sigma}\right)\right]_+^{-2} = -\frac{\xi}{\sigma^2}(\xi+1)h_2.$$

$$\begin{aligned}\frac{\partial^2 h(\boldsymbol{\theta};Y)}{\partial \sigma^2} &= -\frac{1}{\sigma^2}+\frac{1}{\sigma^2}(\xi+1)\left(\frac{Y-\mu}{\sigma}\right)\left[1+\xi\left(\frac{Y-\mu}{\sigma}\right)\right]_+^{-1}+\frac{1}{\sigma^2}(\xi+1)\left(\frac{Y-\mu}{\sigma}\right)\left[1+\xi\left(\frac{Y-\mu}{\sigma}\right)\right]_+^{-2}\\ &= -\frac{1}{\sigma^2}+\frac{1}{\sigma^2}(\xi+1)Vh_1+\frac{1}{\sigma^2}(\xi+1)Vh_2.\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 h(\boldsymbol{\theta};Y)}{\partial \xi^2} &= \frac{2}{\xi^3}\log\left[1+\xi\left(\frac{Y-\mu}{\sigma}\right)\right]_+ - \frac{2}{\xi^2}\left(\frac{Y-\mu}{\sigma}\right)\left[1+\xi\left(\frac{Y-\mu}{\sigma}\right)\right]_+^{-1}\\ &\quad -\left(\frac{1}{\xi}+1\right)\left(\frac{Y-\mu}{\sigma}\right)^2\left[1+\xi\left(\frac{Y-\mu}{\sigma}\right)\right]_+^{-2}\\ &= \frac{2}{\xi^3}\log[1+\xi V]-\frac{2}{\xi^2}Vh_1-\left(\frac{1}{\xi}+1\right)V^2h_2.\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 h(\boldsymbol{\theta};Y)}{\partial \mu \partial \sigma} &= -\frac{\xi}{\sigma^2}(\xi+1)\left(\frac{Y-\mu}{\sigma}\right)\left[1+\xi\left(\frac{Y-\mu}{\sigma}\right)\right]_+^{-2}+\frac{1}{\sigma^2}(\xi+1)\left[1+\xi\left(\frac{Y-\mu}{\sigma}\right)\right]_+^{-1}\\ &= -\frac{\xi}{\sigma^2}(\xi+1)Vh_2+\frac{1}{\sigma^2}(\xi+1)h_1 = \frac{1}{\sigma^2}(\xi+1)h_2.\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 h(\boldsymbol{\theta};Y)}{\partial \mu \partial \xi} &= -\frac{1}{\sigma}\left\{\left[1+\xi\left(\frac{Y-\mu}{\sigma}\right)\right]_+^{-1}-(\xi+1)\left(\frac{Y-\mu}{\sigma}\right)\left[1+\xi\left(\frac{Y-\mu}{\sigma}\right)\right]_+^{-2}\right\}\\ &= -\frac{1}{\sigma}\left\{h_1-(\xi+1)Vh_2\right\}.\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 h(\boldsymbol{\theta};Y)}{\partial \sigma \partial \xi} &= \left(\frac{Y-\mu}{\sigma}\right)\frac{\partial^2 h(\boldsymbol{\theta};Y)}{\partial \mu \partial \xi}, \quad \text{since} \quad \frac{\partial h(\boldsymbol{\theta};Y)}{\partial \sigma} = \frac{1}{\sigma}+\left(\frac{Y-\mu}{\sigma}\right)\frac{\partial h(\boldsymbol{\theta};Y)}{\partial \mu}\\ &= V\frac{\partial^2 h(\boldsymbol{\theta};Y)}{\partial \mu \partial \xi}.\end{aligned}$$

D.3 Expected information for the stationary NHPP model

The expected information matrix for the stationary NHPP model is given by $I_M = I_M(\boldsymbol{\theta}) = \mathrm{E}[J(\boldsymbol{\theta})]$. Note that $g(\boldsymbol{\theta})$ does not involve the response data. Therefore element (j,k) of I_M is given by

$$\begin{aligned} -E\left[\frac{\partial^2 \ell(\boldsymbol{\theta};\boldsymbol{Y})}{\partial\theta_j\partial\theta_k}\right] &= \frac{1}{n}\sum_{i=1}^{m}\frac{\partial^2 g(\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k} + \sum_{i=1}^{m} E\left[\delta(Y_i>u)\frac{\partial^2 h(\boldsymbol{\theta};Y_i)}{\partial\theta_j\partial\theta_k}\right], \\ &= \frac{m}{n}\frac{\partial^2 g(\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k} + m\,E\left[\delta(Y_i>u)\frac{\partial^2 h(\boldsymbol{\theta};Y)}{\partial\theta_j\partial\theta_k}\right]. \end{aligned}$$

We need to evaluate $E\left[\delta(Y>u)\,\partial^2 h(\boldsymbol{\theta};Y)/\partial\theta_j\partial\theta_k\right]$, for $j,k\in\{1,2,3\}$, where, using a stationary version of (5.11) by setting $\mu_1=0$ and letting $\mu=\mu_0$, Y has p.d.f.

$$f_Y(y) \;=\; \frac{1}{\sigma}\frac{1}{n}\left[1+\xi\left(\frac{y-\mu}{\sigma}\right)\right]_+^{-(1+1/\xi)} \exp\left\{-\frac{1}{n}\left[1+\xi\left(\frac{y-\mu}{\sigma}\right)\right]_+^{-1/\xi}\right\}.$$

Let

$$E_{ab} \;=\; \mathrm{E}\left\{\delta(Y>u)\,V^a\,h_b\right\} = \int_{y>u}\left(\frac{y-\mu}{\sigma}\right)^a\left[1+\xi\left(\frac{y-\mu}{\sigma}\right)\right]^{-b} f_Y(y)\;\mathrm{d}y.$$

We make the transformation $r=(1/n)[1+\xi(y-\mu)/\sigma]_+^{-1/\xi}$, leading to

$$E_{ab}=\xi^{-a}\int_0^{\beta}\left[(rn)^{-\xi}-1\right]^a\,(rn)^{b\xi}\,\exp\{-r\}\,dr, \tag{D.1}$$

where

$$\beta=\frac{1}{n}\left[1+\xi\left(\frac{u-\mu}{\sigma}\right)\right]_+^{-1/\xi}. \tag{D.2}$$

We will need

$$\begin{aligned} E_{00} &= \gamma(1,\beta)=1-\exp(-\beta), \\ E_{01} &= n^{\xi}\,\gamma(1+\xi,\beta), \\ E_{02} &= n^{2\xi}\,\gamma(1+2\xi,\beta), \\ E_{11} &= \xi^{-1}\left\{\gamma(1,\beta)-n^{\xi}\gamma(1+\xi,\beta)\right\}, \\ E_{12} &= \xi^{-1}\left\{n^{\xi}\gamma(1+\xi,\beta)-n^{2\xi}\gamma(1+2\xi,\beta)\right\}, \\ E_{22} &= \xi^{-2}\left\{\gamma(1,\beta)-2n^{\xi}\gamma(1+\xi,\beta)+n^{2\xi}\gamma(1+2\xi,\beta)\right\}, \end{aligned} \tag{D.3}$$

where

$$\gamma(s,\beta)=\int_0^{\beta} t^{s-1}\exp(-t)\;\mathrm{d}t.$$

is the lower incomplete gamma function. Similarly,

$$\phi = \mathrm{E}\left\{\delta(Y > u)\, \log\left[1 + \xi\left(\frac{Y-\mu}{\sigma}\right)\right]\right\} = -\xi\, \gamma(1,\beta)\, \log n - \xi\, \gamma'(1,\beta),$$

where $\gamma'(s,\beta) = \partial\gamma(s,\beta)/\partial s = \int_0^\beta t^{s-1}e^{-t}\log s \; \mathrm{d}t$. Therefore,

$$\begin{aligned}
E\left[I\left(Y > u\right)\frac{\partial^2 h(\boldsymbol{\theta};Y)}{\partial\mu^2}\right] &= -\sigma^{-2}\xi(1+\xi)E_{02}, \\
E\left[I\left(Y > u\right)\frac{\partial^2 h(\boldsymbol{\theta};Y)}{\partial\sigma^2}\right] &= -\sigma^{-2}\left\{E_{00} - (1+\xi)E_{11} - (1+\xi)E_{12}\right\}, \\
E\left[I\left(Y > u\right)\frac{\partial^2 h(\boldsymbol{\theta};Y)}{\partial\xi^2}\right] &= 2\xi^{-3}\phi - 2\xi^{-2}E_{11} - \xi^{-1}(1+\xi)E_{22}, \\
E\left[I\left(Y > u\right)\frac{\partial^2 h(\boldsymbol{\theta};Y)}{\partial\mu\partial\sigma}\right] &= \sigma^{-2}(1+\xi)E_{02}, \\
E\left[I\left(Y > u\right)\frac{\partial^2 h(\boldsymbol{\theta};Y)}{\partial\mu\partial\xi}\right] &= -\sigma^{-1}\left\{E_{01} - (1+\xi)E_{12}\right\}, \\
E\left[I\left(Y > u\right)\frac{\partial^2 h(\boldsymbol{\theta};Y)}{\partial\sigma\partial\xi}\right] &= -\sigma^{-1}\left\{E_{11} - (1+\xi)E_{22}\right\}.
\end{aligned} \tag{D.4}$$

D.4 Expected information for the non-stationary NHPP model

First, we consider the case where $\mu(x) = \mu_0 + \mu_1 x$ and $u(x) = u_0 + u_1 x$. Recall from (5.14) that the Fisher information for $\boldsymbol{\theta} = (\mu_1, \mu_0, \sigma, \xi)$ is given by

$$I = \begin{pmatrix} I_{11} & I_1^T \\ I_1 & I_M \end{pmatrix}.$$

The Fisher information I_M for the marginal parameters $\boldsymbol{\psi} = (\mu_0, \sigma, \xi)$ can be inferred from section D.3, noting that, for observation i, μ has been replaced by $\mu(x_i) = \mu_0 + \mu_1 x_i$. Since $\partial\mu(x_i)/\partial\mu_0 = 1$, quantities involving derivatives of μ_0 can be inferred from the corresponding quantities involving derivatives of μ given in section D.3. Element (k,l) of I_M is given by

$$\frac{1}{n}\sum_{i=1}^{m}\frac{\partial^2 g(\boldsymbol{\theta})}{\partial\psi_k\partial\psi_l} + \sum_{i=1}^{m} E\left[\delta(Y_i > u)\frac{\partial^2 h(\boldsymbol{\theta};Y_i)}{\partial\psi_k\partial\psi_l}\right], \tag{D.5}$$

where now $g(\boldsymbol{\theta}) = [1+\xi w_i]_+^{-1/\xi}$, where $w_i = [u(x_i) - \mu(x_i)]/\sigma$, depends on i, although this is not explicit in the notation. The second derivatives of $g(\boldsymbol{\theta})$ are given by replacing w by w_i and g_j by $[1+\xi w_i]_+^{-(j+1/\xi)}$ in the expressions in (D.3) and (D.4) in section D.2. The expectations in the second term of (D.5) are given by the expressions given at the end of section D.3, but with β replaced by

$$\beta_i = \frac{1}{n}\left[1+\xi\left(\frac{u(x_i)-\mu(x_i)}{\sigma}\right)\right]_+^{-1/\xi}. \tag{D.6}$$

in E_{ab} and ϕ.

To derive the elements of I that involve derivatives with respect to μ_1, we note that $\partial\mu(x_i)/\partial\mu_1 = x_i$. Let ℓ_i be the contribution to the log-likelihood corresponding to observation Y_i. The relevant second derivatives of ℓ_i are given by

$$\begin{aligned} \frac{\partial^2 \ell_i}{\partial\mu_1 \partial\psi_k} &= x_i \, \frac{\partial^2 l_i}{\partial\mu(x_i)\partial\psi_k}, \quad k = 1, 2, 3, \\ \frac{\partial^2 \ell_i}{\partial\mu_1^2} &= x_i^2 \, \frac{\partial^2 l_i}{\partial\mu(x_i)^2}. \end{aligned}$$

Let $I(i)$ denote the contribution to I from Y_i and $I_{kl}(i)$ the (k, l) element of this matrix. Then,

$$I_{k1} = \sum_{i=1}^{m} x_i \, I_{k1}(i), \quad k = 2, 3, 4,$$

and

$$I_{11} = \sum_{i=1}^{m} x_i^2 \, I_{22}(i).$$

All the elements of $I(i)$ depend on x_i only through $u(x_i) - \mu(x_i)$.

Let $I_M(i)$ denote the contribution to I_M from Y_i. Then I is given by

$$I = \sum_{i=1}^{m} T_i \, I_M(i) \, T_i^T$$

where

$$T_i = \begin{pmatrix} 1 & 0 & 0 \\ x_i & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

To extend to the model in which there are p covariates in location, i.e.

$$\mu(\boldsymbol{x}_i) = \mu_0 + \mu_1 x_{1i} + \mu_2 x_{2i} + \ldots + \mu_p x_{pi},$$

we use

$$T_i = \begin{pmatrix} 1 & 0 & 0 \\ x_{1i} & 0 & 0 \\ \vdots & \vdots & \vdots \\ x_{pi} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

D.5 Matrix theory

We summarise some standard results related to matrices, for use in chapter 5.

Positive definiteness

A symmetric $n \times n$ real matrix M is said to be positive definite if $z^T M z > 0$, for all non-zero real $1 \times n$ vectors z. A positive definite matrix M is invertible and this inverse is also positive definite (Horn and Johnson, 1990).

In the following M is a $(p+k) \times (p+k)$ Fisher information matrix and is thus symmetric and positive definite.

Schur complements

Suppose that M is partitioned as

$$M = \begin{pmatrix} M_{11} & M_{12}^T \\ M_{12} & M_{22} \end{pmatrix},$$

where M_{22} is a $k \times k$ nonsingular matrix with $1 \leqslant k < n$. The Schur complement of M_{22} in M is given by

$$M/M_{22} = M_{11} - M_{21}^T M_{22}^{-1} M_{21},$$

and Schur's determinant identity (Yan, 2009) is

$$\det M = \det M_{22} \det M/M_{22}.$$

The information submatrices M_{11} and M_{22} are positive definite as is M_{22}^{-1}.

Block inversion

The block inversion technique (see Yan (2009) or Boyd and Vandenberghe (2004, page 650)) allows us to invert M using

$$M^{-1} = \begin{pmatrix} (M_{11} - M_{21}^T M_{22}^{-1} M_{21})^{-1} & -M_{11}^{-1} M_{21}^T (M_{22} - M_{21} M_{11}^{-1} M_{21}^T)^{-1} \\ -(M_{22} - M_{21} M_{11}^{-1} M_{21}^T)^{-1} M_{21} M_{11}^{-1} & (M_{22} - M_{21} M_{11}^{-1} M_{21}^T)^{-1} \end{pmatrix}.$$

Hadamard's determinant inequality (Mirsky, 1955, page 417)

If $A = \{a_{ij}\}$ is an $n \times n$ symmetric positive definite matrix then

$$\det A \leqslant a_{11} a_{22} \cdots a_{nn}, \tag{D.7}$$

with equality if and only if A is diagonal.

Matrix determinant inequality (Yan, 2009, Lemma 1.4)

Let A and B be two symmetric positive semi-definite matrices of the same size. Then

$$\det(A + B) \geqslant \det(A) + \det(B). \tag{D.8}$$

Another matrix determinant inequality

$$\det(M_{11} - M_{21}^T M_{22}^{-1} M_{21}) \leqslant \det(M_{11}), \tag{D.9}$$

with equality if and only if M_{21} is a $r \times p$ zero matrix.

Proof: Let y be a non-zero $1 \times p$ vector. Thus $x = y M_{21}^T$ is a $1 \times r$ vector and

$$y M_{21}^T M_{22}^{-1} M_{21} y^T = x M_{22}^{-1} x^T \geqslant 0,$$

with equality if and only if x is a zero vector, that is, if M_{21} is a zero matrix.

Therefore, if M_{21} is non-zero, $M_{21}^T M_{22}^{-1} M_{21}$ is positive definite.

Substituting $A = M_{11} - M_{21}^T M_{22}^{-1} M_{21}$ and $B = M_{21}^T M_{22}^{-1} M_{21}$ in (D.8) gives

$$\begin{aligned} \det(A + B) &= \det(M_{11}) \\ &\geqslant \det(M_{11} - M_{21}^T M_{22}^{-1} M_{21}) + \det(M_{21}^T M_{22}^{-1} M_{21}) \\ &\geqslant \det(M_{11} - M_{21}^T M_{22}^{-1} M_{21}), \end{aligned}$$

with equality if and only if M_{21} is a zero matrix. ■

D.6 Proof of property 2

We consider how, for fixed $\det I_M$, the elements of the Fisher expected information I_M vary as the form of the threshold is varied. Without loss of generality we consider the case $\mu_1 = 0$. Suppose that a constant threshold v is set at a given high quantile of the marginal distribution of Y, resulting in $\det I_M = (md)^3$, say. We will need the following result.

A generalized Minkowski Determinant Inequality. Let $K_1, \ldots, K_m$ be $d \times d$ real (symmetric) positive definite matrices. Then

$$[\det(K_1 + \cdots + K_m)]^{1/d} \geqslant [\det K_1]^{1/d} + \cdots + [\det K_m]^{1/d}, \qquad \text{(D.10)}$$

with equality if and only if $K_i = c_i K_1$, $i = 2, \ldots, m$ for some constants $c_i \geqslant 0$, that is, the matrices $K_1, \ldots, K_m$ are *proportional*. This follows directly by applying repeatedly the original ($m = 2$) Minkowski Determinant Inequality (Horn and Johnson, 1990, page 482).

For $m \geqslant 2$, let

$$I_M = K_1 + \cdots + K_m,$$

where K_i is the contribution to I_M from observation i. If the threshold is constant then

$$K_1 = \cdots = K_m = K$$

and we have equality in (D.10), with $\det K = d^3$, producing $\det I_M = \det(mK) = (md)^3$.

Suppose that one threshold, say v_1 is increased, while $v_2, \ldots, v_m$ are decreased (at a common rate) such that $\det I_M$ remains equal to $(md)^3$. The pairwise ratios of the elements of I_M are not constant with respect to threshold, that is, changing the threshold does not result in a simple scaling of I_M. This means that K_2 is no longer proportional to K_1 and we have strict inequality in (D.10), giving

$$[\det K_1]^{1/3} + [\det((m-1)K_2)]^{1/3} < [\det(mK)]^{1/3}.$$

Therefore, as we deviate from a constant threshold, $\det K_1$ decreases more quickly from $\det K$ than $(m-1)\det(K_2)$ increases from $\det K$. The absolute values of the elements of each K_i are strictly decreasing in u. Thus, the elements of K_1 decrease in absolute value more quickly than the elements of $(m-1)K_2$ increase and so the elements of I_M decrease in absolute value as v_1 increases. Repeating this process, that is, increasing v_j while decreasing $v_i, i > j$ at a common rate will always result in a decrease in the absolute values of the elements of I_M. Therefore, for a given value of $\det I_M$, the absolute values of the elements of I_M are maximized when a constant threshold is used.

References

Abramowitz, M. and I. A. Stegun (1972). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* New York: Dover Publications.

Aitchison, J. and I. R. Dunsmore (1975). Statistical prediction analysis.

Alzer, H. (1999). Inequalities for the gamma function. *Proceedings of the American Mathematical Society 128*(1), pp. 141–147.

Arlot, S. and A. Celisse (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys 4*, 40–79.

Atkinson, A. C., A. N. Donev, and R. D. Tobias (2007). *Optimum experimental designs, with SAS.* Oxford: Oxford University Press.

Azzalini, A. (1996). *Statistical Inference Based on the likelihood.* Chapman & Hall/CRC.

Bayarri, M. J. and J. O. Berger (2004). The interplay of Bayesian and frequentist analysis. *Statistical Science 19*(1), pp. 58–80.

Behrens, C. N., H. F. Lopes, and D. Gamerman (2004). Bayesian analysis of extreme events with threshold estimation. *Statistical Modelling 4*(3), pp. 227–244.

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

Bennett, J. E., A. Racine-Poon, and J. C. Wakefield (1996). MCMC for nonlinear hierarchical models. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice*, pp. 339–357. London: Chapman & Hall.

Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian analysis 1*(3), pp. 385–402.

Berger, J. O., J. M. Bernardo, and D. Sun (2009). The formal definition of reference priors. *The Annals of Statistics 37*(2), pp. 905–938.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society. Series B (Methodological) 41*(2), pp. 113–147.

Bernardo, J. M. (2005). Reference analysis. In D. K. Dey and C. R. Rao (Eds.), *Bayesian Thinking Modeling and Computation*, Volume 25 of *Handbook of statistics*, pp. 17–90. Amsterdam: Elsevier.

Billingsley, B. (1995). *Probability and Measure.* New York: John Wiley & Sons.

Bondell, H. D., B. J. Reich, and H. Wang (2010). Noncrossing quantile regression curve estimation. *Biometrika 97*(4), pp. 825–838.

Boyd, S. and L. Vandenberghe (2004). *Convex Optimization.* Cambridge: Cambridge University Press.

Butler, A., J. E. Heffernan, J. A. Tawn, and R. A. Flather (2007). Trend estimation in extremes of synthetic North Sea surges. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 56*(4), pp. 395–414.

Chandler, R. E. and S. B. Bate (2007). Inference for clustered data using the independence loglikelihood. *Biometrika 94*(1), pp. 167–183.

Chavez-Demoulin, V. and A. C. Davison (2005). Generalized additive modelling of sample extremes. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 54*(1), pp. 207–222.

Chavez-Demoulin, V. and A. C. Davison (2012). Modelling time series extremes. *REVSTAT-Statistical Journal 10*(1), pp. 109–133.

Chavez-Demoulin, V., A. C. Davison, and L. Frossard (2011). Discussion of 'Threshold modelling of spatially-dependent non-stationary extremes with application to hurricane-induced wave heights' by P.J. Northrop and P. Jonathan. *Environmetrics 22*(7), pp. 810–812.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values.* London: Springer.

Coles, S. G. (1991). *Modelling extreme multivariate events.* Ph. D. thesis, University of Sheffield.

Coles, S. G. and M. J. Dixon (1999). Likelihood-based inference for extreme value models. *Extremes 2*(1), pp. 5–23.

Coles, S. G. and E. A. Powell (1996). Bayesian methods in extreme value modelling: a review and new developments. *International Statistical Review 64*(1), pp. 119–136.

Coles, S. G. and J. A. Tawn (1996). A Bayesian analysis of extreme rainfall data. *Applied Statistics 45*(4), pp. 463–478.

Coles, S. G. and J. A. Tawn (2005). Bayesian modelling of extreme surges on the UK east coast. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 363*(1831), pp. 1387–1406.

Cox, D. R., V. S. Isham, and P. J. Northrop (2002). Floods: some probabilistic and statistical approaches. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences 360*(1796), pp. 1389–1408.

Davison, A. C. (2003). *Statistical models.* New York: Cambridge University Press.

Davison, A. C. and R. L. Smith (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society. Series B (Methodological) 52*(3), pp. 393–442.

de Zea Bermudez, P. and M. A. Amaral Turkman (2003). Bayesian approach to parameter estimation of the generalized Pareto distribution. *Test 12*(1), pp. 259–277.

Drees, H., L. de Haan, and S. Resnick (2000). How to make a Hill plot. *The Annals of Statistics 28*(1), pp. 254–274.

Dupuis, D. J. (1998). Exceedances over high thresholds: A guide to threshold selection. *Extremes 1*(3), pp. 251–351.

Eastoe, E. F. and J. A. Tawn (2009). Modelling non-stationary extremes with application to surface level ozone. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 58*(1), pp. 25–45.

Eugenia Castellanos, M. and S. Cabras (2007). A default Bayesian procedure for the generalized Pareto distribution. *Journal of Statistical Planning and Inference 137*(2), pp. 473–483.

Fawcett, L. and D. Walshaw (2006). Markov chain models for extreme wind speeds. *Environmetrics 17*(8), pp. 795–809.

Fawcett, L. and D. Walshaw (2007). Improved estimation for temporally clustered extremes. *Environmetrics 18*(2), pp. 173–188.

Fawcett, L. and D. Walshaw (2008). Bayesian inference for clustered extremes. *Extremes 11*(3), 217–233.

Fawcett, L. and D. Walshaw (2012). Estimating return levels from serially dependent extremes. *Environmetrics 23*(3), pp. 272–283.

Ferreira, A., L. de Haan, and L. Peng (2003). On optimising the estimation of high quantiles of a probability distribution. *Statistics 37*(5), pp. 401–434.

Ferro, C. A. T. and J. Segers (2003). Inference for clusters of extreme values. *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 65*(2), pp. 545–556.

Fisher, R. A. and L. H. C. Tippett (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical Proceedings of the Cambridge Philosophical Society*, Volume 24, pp. 180–190.

Fukutome, S., M. A. Liniger, and M. Süveges (2014). Automatic threshold and run parameter selection: a climatology for extreme hourly precipitation in switzerland. *Theoretical and Applied Climatology*, pp. 1–14.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association 70*(350), pp. 320–328.

Geisser, S. (1982). Aspects of the predictive and estimative approaches in the determination of probabilities. *Biometrics 38*, pp. 75–85.

Geisser, S. and W. F. Eddy (1979). A predictive approach to model selection. *Journal of the American Statistical Association 74*(365), pp. 153–160.

Gelfand, A. E. (1996). Model determination using sampling-based methods. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice*, pp. 145–161. London: Chapman & Hall.

Gelfand, A. E. and D. K. Dey (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B (Methodological) 56*(3), pp. 501–514.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis 1*(3), pp. 515–533.

Geweke, J. and G. Amisano (2010). Comparing and evaluating Bayesian predictive distributions of asset returns. *International Journal of Forecasting 26*(2), pp. 216–230.

Giles, D. E., H. Feng, and R. T. Godwin (2011). Bias-corrected maximum likelihood estimation of the parameters of the generalized Pareto distribution. Econometrics Working Papers EWP1105, Department of Economics, University of Victoria.

Gilks, W. R. and P. Wild (1992). Adaptive rejection sampling for Gibbs sampling. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 41*(2), pp. 337–348.

Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une serie aleatoire. *The Annals of Mathematics 44*(3), pp. 423–453.

Gradshteyn, I. S. and I. W. Ryzhik (2007). *Table of Integrals, Series and Products.* New York: Academic Press.

Greenwood, J. A., J. M. Landwehr, N. C. Matalas, and J. R. Wallis (1979). Probability weighted moments: Definition and relation to parameters of several distributions expressable in inverse form. *Water Resources Research 15*(5), pp. 1049–1054.

Hall, P. (1990). Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems. *Journal of Multivariate Analysis 32*(2), pp. 177–203.

Hall, P. and A. H. Welsh (1985). Adaptive estimates of parameters of regular variation. *The Annals of Statistics 13*(1), pp. 331–341.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika 57*(1), pp. 97–109.

Heffernan, J. E. and J. A. Tawn (2004). A conditional approach for multivariate extreme values (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66*(3), pp. 497–546.

Hobert, J. P. and G. Casella (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association 91*(436), pp. 1461–1473.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999). Bayesian model averaging: A tutorial. *Statistical Science 14*(4), pp. 382–401.

Horn, R. A. and C. R. Johnson (1990). *Matrix Analysis.* Cambridge: Cambridge University Press.

Hosking, J. R. M. and J. R. Wallis (1987). Parameter and quantile estimation for the Generalized Pareto distribution. *Technometrics 29*(3), pp. 339–349.

Hosking, J. R. M., J. R. Wallis, and E. F. Wood (1985). Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics 27*(3), pp. 251–261.

Jeffreys, H. (1961). *Theory of Probability.* Oxford: Oxford University Press.

Jenkinson, A. F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quarterly Journal of the Royal Meteorological Society 81*(348), pp. 158–171.

Jolliffe, L. T. (2002). *Principal Components Analysis.* New York: Springer.

Jonathan, P. and K. Ewans (2007). Uncertainties in extreme wave height estimates for hurricane-dominated regions. *Journal of Offshore Mechanics and Arctic Engineering 129*(4), pp. 300–305.

Jonathan, P. and K. Ewans (2011). A spatiodirectional model for extreme waves in the Gulf of Mexico. *Journal of Offshore Mechanics and Arctic Engineering 133*(1), pp. 011601.

Jonathan, P. and K. Ewans (2013). Statistical modelling of extreme ocean environments for marine design: A review. *Ocean Engineering 62*(0), pp. 91–109.

Jonathan, P., D. Randell, Y. Wu, and K. Ewans (2014). Return level estimation from non-stationary spatial data exhibiting multidimensional covariate effects. *Ocean Engineering 88*(0), pp. 520–532.

Kass, R. E. and L. Wasserman (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association 91*(435), pp. 1343–1370.

Keef, C., I. Papastathopoulos, and J. A. Tawn (2013). Estimation of the conditional distribution of a multivariate variable given that one of its components is large: Additional constraints for the heffernan and tawn model. *Journal of Multivariate Analysis 115*(0), 396–404.

Kinderman, A. J. and J. F. Monahan (1977). Computer generation of random variables using the ratio of uniform deviates. *ACM Trans. Math. Softw. 3*(3), pp. 257–260.

Koenker, R. (2011). *quantreg: Quantile Regression.* R package version 4.67.

Koenker, R. and G. Bassett (1978). Regression quantiles. *Econometrica 46*(1), pp. 33–50.

Kotz, S. and S. Nadarajah (2000). *Extreme value distributions: theory and applications.* London: Imperial College Press.

Landwehr, J. M., N. C. Matalas, and J. R. Wallis (1979). Probability weighted moments compared with some traditional techniques in estimating Gumbel parameters and quantiles. *Water Resources Research 15*(5), pp. 1055–1064.

Larson, S. C. (1931). The shrinkage of the coefficient of multiple correlation. *Journal of Educational Psychology 22*(1), pp. 45–55.

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

Leamer, E. E. (1978). *Specification searches.* New York: Wiley.

Liu, J. S. (2004). *Monte Carlo strategies in scientific computing.* New York: Springer.

MacDonald, A., C. J. Scarrott, D. Lee, B. Darlow, M. Reale, and G. Russell (2011). A flexible extreme value mixture model. *Computational Statistics & Data Analysis 55*(6), pp. 2137–2157.

Martins, E. S. and J. R. Stedinger (2000). Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data. *Water Resources Research 36*(3), pp. 737–744.

Martins, E. S. and J. R. Stedinger (2001). Generalized maximum likelihood Pareto-Poisson estimators for partial duration series. *Water Resources Research 37*(10), pp. 2551–2557.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics 21*(6), pp. 1087–1092.

Metropolis, N. and S. Ulam (1949). The Monte Carlo method. *Journal of the American statistical association 44*(247), pp. 335–341.

Mirsky, L. (1955). *An introduction to linear algebra.* Oxford: Oxford University Press.

Mitrinović, D. A. (1964). *Elementary inequalities.* Groningen: Noordhoff.

Mosteller, F. and J. W. Tukey (1968). Data analysis, including statistics. In G. Lindzey and E. Aronson (Eds.), *Handbook of Social Psychology*, Volume 2. Addison-Wesley.

Northrop, P. J. and N. Attalides (2015). Posterior propriety in Bayesian extreme value analyses using reference priors. *Statistica Sinica.* To appear.

Northrop, P. J. and C. L. Coleman (2014). Improved threshold diagnostic plots for extreme value analyses. *Extremes 17*(2), pp. 289–303.

Northrop, P. J. and P. Jonathan (2011). Threshold modelling of spatially dependent non-stationary extremes with application to hurricane-induced wave heights. *Environmetrics 22*(7), pp. 799–809.

Oceanweather Inc. (2005). GOMOS – USA Gulf of Mexico Oceanographic Study, Northern Gulf of Mexico Archive.

O'Hagan, A. (2006). Science, subjectivity and software (comment on articles by Berger and by Goldstein). *Bayesian Analysis 1*(3), pp. 445–450.

Picard, R. R. and R. D. Cook (1984). Cross-validation of regression models. *Journal of the American Statistical Association 79*(387), pp. 575–583.

Pickands, J. (1971). The two-dimensional Poisson process and extremal processes. *Journal of Applied Probability 8*(4), pp. 745–756.

Pickands, J. (1975). Statistical inference using extreme order statistics. *Annals of Statistics 3*(1), pp. 119–131.

Pickands, J. (1994). Bayes quantile estimation and threshold selection for the generalized Pareto family. In J. Galambos, J. Lechner, and E. Simiu (Eds.), *Extreme Value Theory and Applications*, pp. 123–138. Springer.

Prescott, P. and A. T. Walden (1980). Maximum likelihood estimation of the parameters of the generalized extreme-value distribution. *Biometrika 67*(3), pp. 723–724.

Qiu, S.-L. and M. Vuorinen (2004). Some properties of the gamma and psi functions, with applications. *Mathematics of computation 74*(250), pp. 723–742.

Ramesh, N. I. and A. C. Davison (2002). Local models for exploratory analysis of hydrological extremes. *Journal of Hydrology 256*(1-2), pp. 106–119.

Resnick, S. I. (1987). *Extreme Values, Regular Variation and Point Processes.* New York: Springer.

Ripley, B. D. (1987). *Stochastic simulation.* New York: Wiley.

Roberts, H. V. (1965). Probabilistic prediction. *Journal of the American Statistical Association 60*(309), pp. 50–62.

Rodriguez, P. P. (2014). *ars: Adaptive Rejection Sampling.*

Sabourin, A., P. Naveau, and A.-L. Fougres (2013). Bayesian model averaging for multivariate extremes. *Extremes 16*(3), pp. 325–350.

Scarrott, C. and A. MacDonald (2012). A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT - Statistical Journal 10*(1), pp. 33–60.

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

Smith, E. (2005). *Bayesian Modelling of Extreme Rainfall Data.* Ph. D. thesis, University of Newcastle upon Tyne.

Smith, R. L. (1985). Maximum likelihood estimation in a class of non-regular cases. *Biometrika 72*(1), pp. 67–90.

Smith, R. L. (1987). Approximations in extreme value theory. Technical Report 205, University of North Carolina, Department of Statistics.

Smith, R. L. (1989). Extreme value analysis of environmental time series: An application to trend detection in ground-level ozone. *Statistical Science 4*(4), pp. 367–377.

Smith, R. L. (1992). The extremal index for a Markov chain. *Journal of Applied Probability 29*(1), pp. 37–45.

Smith, R. L. (1994). Nonregular regression. *Biometrika 81*(1), pp. 173–183.

Smith, R. L. (1999). Bayesian and frequentist approaches to parametric predictive inference. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Eds.), *Bayesian Statistics 6*, pp. 589–612. Oxford University Press.

Smith, R. L. (2003). Statistics of extremes, with applications in environment, insurance and finance. In B. Finkenstädt and H. Rootzén (Eds.), *Extreme Values in Finance, Telecommunications and the Environment*, pp. 1–78. London: Chapman & Hall CRC.

Smith, R. L. and D. J. Goodman (2000). Bayesian risk analysis. In P. Embrechts (Ed.), *Extremes and Integrated Risk Management*, pp. 235–251. London: Risk Books.

Smith, R. L. and J. C. Naylor (1987). A comparison of maximum likelihood and Bayesian estimators for the three- parameter weibull distribution. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 36*(3), pp. 358–369.

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

Smith, R. L. and I. Weissman (1994). Estimating the extremal index. *Journal of the Royal Statistical Society. Series B (Methodological) 56*, pp. 515–528.

Snyder, D. L. and M. I. Miller (1991). *Random Point Processes in Time and Space*. New York: Springer.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological) 36*(2), pp. 111–147.

Süveges, M. (2007). Likelihood estimation of the extremal index. *Extremes 10*(1-2), pp. 41–55.

Süveges, M. (2008). *Statistical analysis of clusters of extreme events*. Ph. D. thesis, École Polytechnique Fédérale de Lausanne.

Süveges, M. and A. C. Davison (2010). Model misspecification in peaks over threshold analysis. *The Annals of Applied Statistics 4*(1), pp. 203–221.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics 22*(4), pp. 1701–1728.

van Noortwijk, J. M., H. J. Kalk, and E. H. Chbab (2004). Bayesian estimation of design loads. *Heron 49*(2), pp. 189–205.

von Mises, R. (1964). La distribution de la plus grande de n valeurs. In *Selected papers of Richard von Mises, Volume II*, pp. 271–294. American Mathematical Society, Providence, RI.

Wadsworth, J. and J. Tawn (2012). Likelihood-based procedures for threshold diagnostics and uncertainty in extreme value modelling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74*(3), pp. 543–567.

Wakefield, J. C., A. E. Gelfand, and A. F. M. Smith (1991). Efficient generation of random variates via the ratio-of-uniforms method. *Statistics and Computing 1*(2), pp. 129–133.

Walshaw, D. (2000). Modelling extreme wind speeds in regions prone to hurricanes. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 49*(1), pp. 51–62.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica 50*(1), pp. 1–25.

Wong, T. S. T. and W. K. Li (2010). A threshold approach for peaks-over-threshold modelling using maximum product of spacings. *Statistica Sinica 20*(3), pp. 1257–1272.

Yan, Z. (2009). Schur complements and determinant inequalities. *Journal of Mathematical Inequalities 3*(2), pp. 161–167.

Young, G. A. and R. L. Smith (2005). *Essentials of statistical inference.* Cambridge: Cambridge University Press.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics.* New York: John Wiley & Sons, Ltd.

Zellner, A. (1998). Past and recent results on maximal data information priors. *Journal of Statistical Research 32*(1), pp. 1–22.