REVSTAT – Statistical Journal
Volume 10, Number 1, March 2012, 33–60

A REVIEW OF EXTREME VALUE THRESHOLD ESTIMATION AND UNCERTAINTY QUANTIFICATION

Authors: Carl Scarrott
– Mathematics and Statistics Department, University of Canterbury,
New Zealand
carl.scarrott@canterbury.ac.nz

Anna MacDonald
– Mathematics and Statistics Department, University of Canterbury,
New Zealand
anna.macdonald@pg.canterbury.ac.nz

Abstract:

- The last decade has seen development of a plethora of approaches for threshold estimation in extreme value applications. From a statistical perspective, the threshold is loosely defined such that the population tail can be well approximated by an extreme value model (e.g., the generalised Pareto distribution), obtaining a balance between the bias due to the asymptotic tail approximation and parameter estimation uncertainty due to the inherent sparsity of threshold excess data. This paper reviews recent advances and some traditional approaches, focusing on those that provide quantification of the associated uncertainty on inferences (e.g., return level estimation).

Key-Words:

- *extreme value threshold selection; graphical diagnostics; mixture modelling; rule of thumb; threshold uncertainty.*

AMS Subject Classification:

- 62G32, 62G07, 62G30, 62E20.

1. INTRODUCTION

This paper reviews the key historical threshold estimation approaches for extreme value applications, and the latest developments. The focus is on approaches which provide not only threshold estimation but also uncertainty quantification for the threshold itself and subsequent inferences for quantities like return levels, though we also discuss some exceptions for key developments which provide threshold estimation, but not uncertainty quantification. There is a certain focus on recently developed mixture model type approaches, as these deal naturally with both estimation and formal uncertainty quantification. The aim is to be all encompassing, a near-impossible task, so we apologise for any omissions.

The classical asymptotically motivated model for excesses above a high threshold is the generalised Pareto distribution (GPD). Pickands (1975) and Balkema & de Haan (1974) showed that if there is a non-degenerate limiting distribution for appropriately linearly rescaled excesses of a sequence of independent and identically distributed observations $X_1, ..., X_n$ above a threshold u, then the limiting distribution will be a GPD. In applications, the GPD is used as a tail approximation to the population distribution from which a sample of excesses $x - u$ above some suitably high threshold u are observed. The GPD is parameterised by scale and shape parameters $\sigma_u > 0$ and ξ, and can equivalently be specified in terms of threshold excesses $x - u$ or, as here, exceedances $x > u$, as

$$(1.1) \qquad G(x \mid u, \sigma_u, \xi) \;=\; \Pr\bigl(X < x \mid X > u\bigr) \;=\; \begin{cases} 1 - \left[1 + \xi\left(\dfrac{x-u}{\sigma_u}\right)\right]_+^{-1/\xi}, & \xi \neq 0\,, \\[2ex] 1 - \exp\left[-\left(\dfrac{x-u}{\sigma_u}\right)\right]_+, & \xi = 0\,, \end{cases}$$

where $y_+ = \max(y, 0)$. When $\xi < 0$ there is an upper end point, so $u < x < u - \sigma_u/\xi$. Implicitly underlying the GPD is a third parameter required for estimation of quantities like return levels, the proportion of threshold excesses $\phi_u = \Pr(x > u)$, used to calculate the unconditional survival probability:

$$(1.2) \qquad \Pr\bigl(X > x\bigr) \;=\; \phi_u\Bigl[1 - \Pr\bigl(X < x \mid X > u\bigr)\Bigr]\,.$$

This representation is often referred to as a Poisson-GPD, as it explicitly accounts for the Poisson rate of excess events. Smith (1989) and Davison & Smith (1990) consider statistical aspects of a Poisson point process (PPP) representation of the classical extreme value models, details of which are not provided for brevity. The principle benefits of the PPP representation is that it can be parameterised in terms of location μ, scale σ and shape ξ which are independent of the threshold, which can simplify extension to nonstationary or random effects type models, and the excess rate ϕ_u is incorporated as a function of these parameters.

The GPD satisfies a threshold stability property: for any higher threshold $v > u$, the subsequent excesses also follow a GPD with the same shape but shifted scale $\sigma_v = \sigma_u + \xi(v - u)$. The subscript u on σ_u makes the threshold dependence explicit, although in the limiting case $\xi = 0$ this disappears. The 'modified scale' reparameterisation $\sigma^* = \sigma_v - \xi v$ is constant above u, i.e., once the GPD provides an adequate tail approximation. The shape and scale parameter can be orthogonalised following Cox & Reid (1987) with the preferred form usually $(\tilde{\sigma}_u, \xi)$, where $\tilde{\sigma}_u = \sigma_u(1 + \xi)$, as the shape is often a key parameter of interest.

Traditionally, the threshold was chosen before fitting, giving the so-called *fixed threshold approach*. Threshold choice involves balancing bias and variance. The threshold must be sufficiently high to ensure that the asymptotics underlying the GPD approximation are reliable, thus reducing the bias. However, the reduced sample size for high thresholds increases the variance of the parameter estimates. Threshold choice is practically equivalent to estimation of the k^{th} upper order statistic $X_{(n-k+1)}$ from the ordered sequence $X_{(1)}, \ldots, X_{(n)}$, called the 'tail fraction' below. Formally, to ensure tail convergence, as $n \to \infty$ the order $k \to \infty$ but at a reduced rate so that $k/n \to 0$ (the so called intermediate sequence of order statistics of Leadbetter *et al.* (1983)), i.e., as the sample size grows, the quantile level of the threshold increases at a faster rate.

2. SUMMARY OF ESTIMATION APPROACHES

The classical fixed threshold modelling approach uses graphical diagnostics, essentially assessing aspects of the model fit, to make an *a priori* threshold choice. Some of the commonly used diagnostics and related statistics are described in §2.1. A benefit of this approach is that it requires practitioners to graphically inspect the data, comprehend their features and assess the model fit, when choosing the threshold. A key drawback with these approaches is they can require substantial expertise and can be rather subjective, as will be seen below. Further, application of this approach when there are many datasets (e.g., different stock returns series in finance applications) is time-consuming. In this situation, it is common for practitioners to assume a constant quantile level across all series, determined by some assessment of fit across all or a subset of the datasets. In some applications the threshold is pre-determined by physical considerations, e.g., government target level for pollution concentrations. Some simple rules of thumb for threshold selection are detailed in §3.

The drawback with fixed threshold approaches is that once the threshold has been chosen it is treated as fixed, so the associated subjectivity and/or uncertainty is ignored in subsequent inferences. Further, it is frequently observed in applications that there is more than one suitable threshold with different inferred tail behaviours, which will be ignored when fixing the threshold. An informal approach to overcoming these problems is to evaluate the sensitivity of the inferences (e.g., parameters or quantiles) to different threshold choices. There has been a proliferation of new approaches to estimate the threshold more objectively and/or formally account for the threshold uncertainty. In §4 we summarise the large literature on tail fraction estimation including further graphical diagnostics, which mostly use asymptotic optimality-based arguments under various population distribution assumptions, and §5 outlines resampling based approaches which typically require weaker assumptions.

Direct comparison of the GPD likelihood for different thresholds is complicated by the varying sample sizes. Recently, various extreme value mixture models have been developed to overcome this problem. These mixture models typically approximate the entire distribution function, so have a fixed sample size for each threshold considered. Traditionally, the 'bulk' of the distribution below the threshold was ignored, as they were not supported by the tail asymptotics. Further, from a practical viewpoint the extreme and non-extreme events are often caused by different driving forces, so the latter will provide little information about the tails. These mixture models have a rather ad-hoc development, often motivated by their applications or by the underlying properties of the population distribution (e.g., bounds on the support, multi-modality). The guiding principle in their development is to choose a physically sensible model for the bulk distribution, for the application at hand, along with an appropriate tail/threshold model. The most widely applicable mixture models are carefully defined to ensure that the bulk and tail fits are not too influenced by each other. There are a range of mixture models with different assumptions for bulk, tail and threshold components, which have loosely been classified into parametric, semiparametric and nonparametric estimators for the bulk distribution in §6.1–6.3 below.

§7 describes approaches outside these general categories, e.g., robust estimation. A nice short review of a subset of the threshold estimation approaches outlined below is provided by de Zea Bermudez & Kotz (2010).

2.1. Graphical diagnostics

Coles (2001) outlines the common graphical diagnostics for threshold choice:

- Mean residual life (or mean excess) plot;
- Threshold stability plot(s);
- A suite of the usual distribution fit diagnostics (e.g., probability plots, quantile plots, return level plots, empirical and fitted density comparison).

The mean residual life (MRL) plot introduced by Davison & Smith (1990) uses the expectation of the GPD excesses, $E(X - u \mid X > u) = \sigma_u/(1-\xi)$, as a diagnostic, defined for $\xi < 1$ to ensure the mean exists. For any higher $v > u$ the expectation becomes

$$E\big(X - v \mid X > v\big) \;=\; \frac{\sigma_u + \xi v}{(1-\xi)}$$

which is linear in v with gradient $\xi/(1-\xi)$ and intercept $\sigma_u/(1-\xi)$. Examples of the behaviour of the MRL function for various distributions are given by Beirlant *et al.* (2004). Empirical estimates of the sample mean excesses are typically plotted against a range of thresholds, along with Wald type interval estimates, though bootstrap or similar estimates would generally be more appropriate for small tail samples. The threshold is chosen to be the lowest level where all the higher threshold based sample mean excesses are consistent with a straight line, once the sample uncertainty is accounted for. Coles (2001) acknowledges that the interpretation of such plots can be challenging.

Figure 1 gives an example of a MRL plot for the Fort Collins total daily precipitation data from the `extRemes` package in `R` (Gilleland *et al.*, 2010).

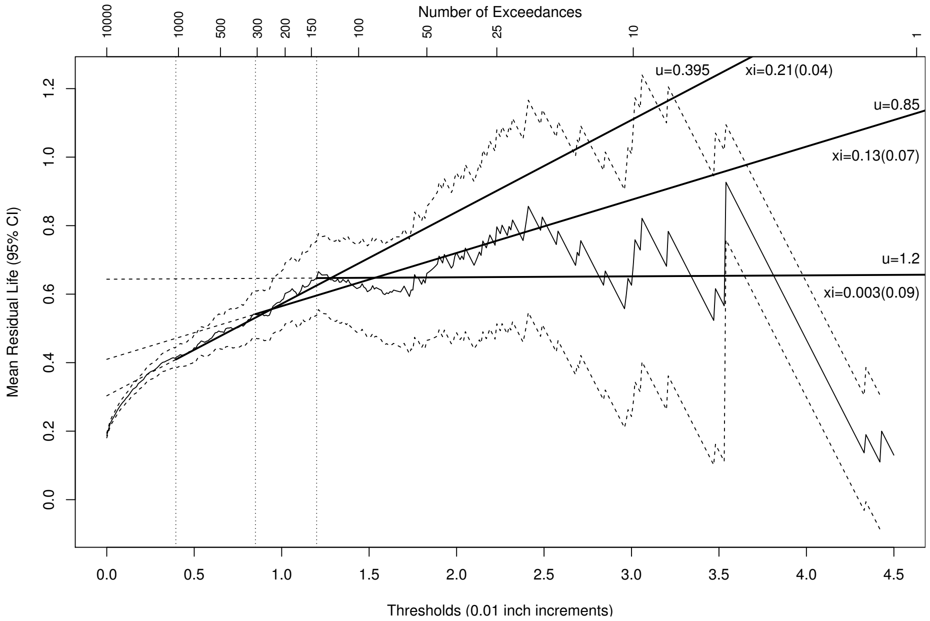


Figure 1: Mean residual life plot for the Fort Collins precipitation data. Solid jagged line is empirical MRL with approximate pointwise Wald 95% confidence intervals as dashed lines. The MRL implied by maximum likelihood (ML) parameter estimates for thresholds $u = 0.395$, 0.85 and 1.2 inches are the upper, middle and lower straight lines respectively. Vertical dashed lines mark these thresholds.

A simplistic analysis of the marginal distribution is considered here, ignoring the obvious seasonality and dependence structure, as we just want to demonstrate the diagnostic plots. The ML estimates of the parameters for three threshold choices provide three fitted MRL straight lines. The increasing variance of the MRL for high thresholds leads to wide confidence intervals, which must be taken into account when assessing the threshold choice.

The threshold of $u = 0.395$ inches suggested by Katz *et al.* (2002) gives over 1000 exceedances, and is justified by the empirical MRL becoming close to linear above this level, and below this level a curved MRL is observed indicating a bias due to the GPD asymptotics breaking down. However, above 1.2 inches strong deviations from the fitted MRL, shown by the upper straight line, are observed with higher thresholds exhibiting a lighter tail. The upper straight line is also close to the pointwise upper interval limit around 1.7 inches (where there is still a reasonable amount of data), and above 2.7 inches the line lies outside the intervals (although the sample sizes here are small so the interval estimates are less reliable). A very different picture arises when considering the highest threshold of $u = 1.2$, with 138 exceedances, which could be justified on the same grounds: above this threshold the MRL is consistent with a straight line (in fact a horizontal line indicating $\xi \approx 0$) and before this level the MRL could be considered to exhibit bias-related curvature. Both these thresholds are consistent with the general guidelines for threshold choice using the MRL plot, but they provide very different tail extrapolations ($\xi = 0.21$ and 0.003 respectively).

The threshold stability plots shown in Figure 2 also do not provide firm conclusions. At a threshold of $u = 0.395$ the shape parameter appears to reach a plateau, compared to lower thresholds which exhibit the bias-related curvature. However, as with the MRL plot, inconsistencies are observed between the estimated shape parameter at this level and higher thresholds around $u = 1.2$ and 2.3–2.4. The shape parameter reaches another plateau around 0.85–1.8 inches (above which the sample variation is too large to make useful inferences), and the shape parameter for the threshold $u = 0.85$ is essentially contained within all the confidence intervals for higher thresholds. The plateau above a threshold of $u = 1.2$ is also strongly indicated by the shape parameter threshold stability plot, despite the confidence interval being much wider at this level. The corresponding modified scale threshold stability plot exhibits some similar features, due to the negative dependence with the shape parameter, but is rather more challenging to interpret for high thresholds due to the large sample variability.

This example demonstrates the substantial subjectivity in interpreting these diagnostic plots, and the resulting uncertainty. Similar challenges are seen with the River Nidd data, shown in Tancredi *et al.* (2006), and many other examples in the literature. These examples suggests that a more ‘objective’ threshold estimation approach is needed and that uncertainty must be accounted for.

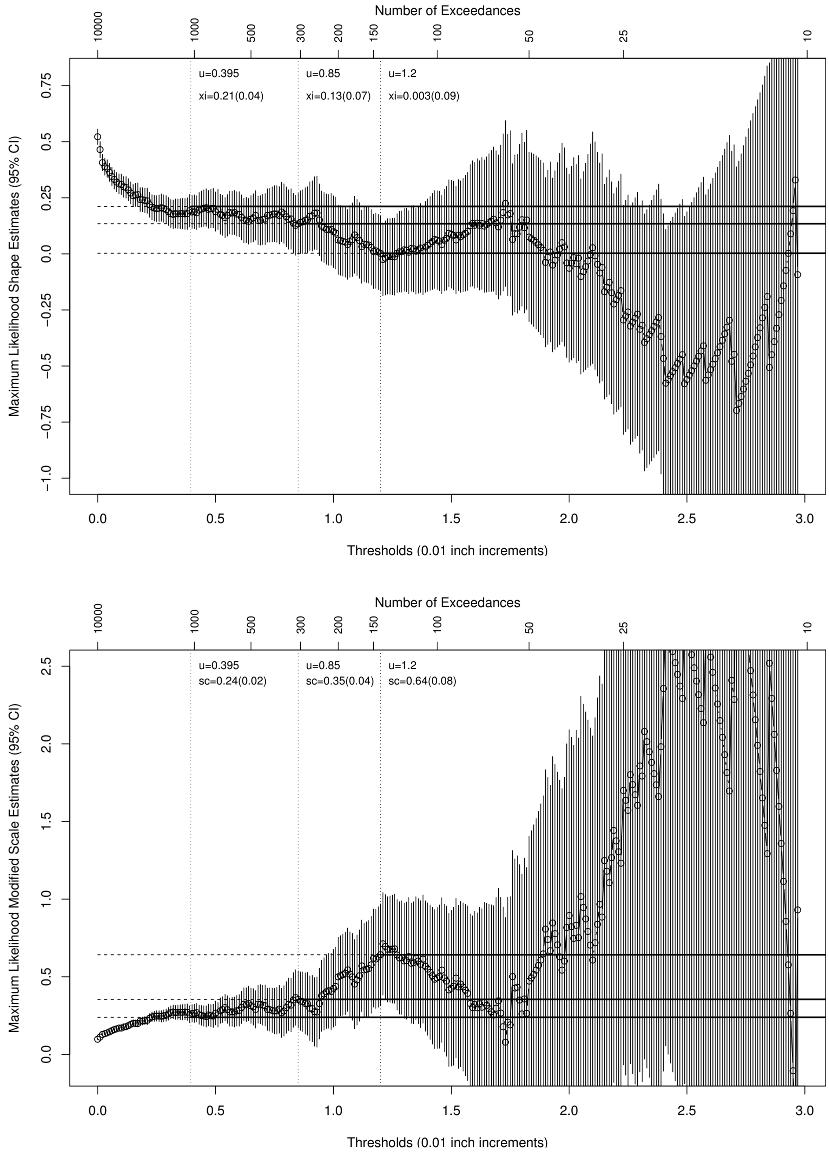


Figure 2: Threshold stability plots for (upper) shape and (lower) modified scale parameters respectively for the Fort Collins precipitation data. Circles are maximum likelihood estimates with vertical lines as approximate pointwise Wald 95% confidence intervals. Three thresholds $u = 0.395$, 0.85 and 1.2 inches are shown by vertical dashed lines with corresponding ML parameter estimates as horizontal lines.

In particular, multiple possible thresholds are indicated in both these examples, so our inference should take this into account. Of course, if the inferences are insensitive to the threshold choice then there is little value to be added from more sophisticated analysis, but in all other cases pre-fixing a single threshold in advance seems inappropriate.

There is a plethora of other diagnostic plots in the literature assessing other features of the model fit at different thresholds, e.g., assessing PPP assumptions or general model goodness of fit statistics considered by Davison & Smith (1990). Under the assumption of Pareto type tails ($\xi > 0$), Hill plots and their many variants, are commonly used, and are discussed in §4.

3. RULES OF THUMB

Leadbetter *et al.* (1983) showed that the threshold sequence (for different sample sizes n), for a population in the domain of attraction of a GPD, is a function of the properties of that distribution. So for a known distribution function F a closed form, or approximation, to the threshold sequence can be derived. For example, a unit exponential population has threshold sequence $u_n = \log(n)$. In the more relevant case of an unknown F there is no general form for the threshold sequence, although some properties of the sequence are known (e.g., order statistic convergence mentioned in §1). Many of the following results follow from fairly general classes of population distributions, such as the Hall (1982) class, which have certain specific tail properties.

Given the general order statistic convergence properties, various rules of thumb have been derived in the literature. Simple fixed quantile rules, like the upper 10% rule of DuMouchel (1983), are inappropriate from a theoretical viewpoint, though frequently used in practice. Ferreira *et al.* (2003), amongst others, use the square root rule $k = \sqrt{n}$ in their simulation study to deterministically specify the tail fraction, which satisfies the intermediate order statistic convergence property in §1, but the source of a formal derivation of this rule is unknown to us. Ho & Wan (2002) and Omran & McKenzie (1999) use the empirically driven rule $k = n^{2/3}/\log[\log(n)]$ proposed by Loretan & Philips (1994).

Reiss & Thomas (2007) heuristically justify choosing the lowest upper order statistic k to minimize

$$\frac{1}{k}\sum_{i=1}^{k} i^{\beta}\left|\hat{\xi}_i - \text{median}(\hat{\xi}_1, \ldots, \hat{\xi}_k)\right| \tag{3.1}$$

where $\hat{\xi}_i$ is an shape parameter estimator for the tail fraction above upper order statistic i and the tuning parameter satisfies $0 \le \beta \le 0.5$. Minimising (3.1) searches for the tail fraction where the distribution of these estimated shape parameters stabilises (downweighting the small tail fractions when $\beta \neq 0$) and k is chosen as an estimate of the location of the distribution. In practice, automated implementation is this approach is unreliable for small k (despite the weighting by i^{β}), so a minimum value of k is usually used. Further, best practice would validate the selection using a graphical summary of (3.1), similar to the threshold stability plot above, in combination with other diagnostics. Reiss & Thomas (2007) point out there are many sensible extensions using alternative distance metrics or weighting schemes and potentially using smoothers when there is limited data. Neves & Alves (2004) investigated refinements to the choice of β.

4. PROBABILISTIC RESULTS

The Hill (1975) estimator is a classic tail index estimator for the Pareto type distribution ($\xi > 0$), which has a power law form with regularly varying tails,

$$1 - F(x) \approx x^{-\alpha} L(x) \;, \qquad \alpha > 0 \;, \tag{4.1}$$

where $L(x)$ is a slowly varying function, i.e.,

$$\lim_{x\to\infty} \frac{L(tx)}{L(x)} = 1 \;, \qquad t > 0 \;,$$

which allows flexibility in the lower tail but that ensures the power law behavior dominates the upper tail. Clearly, this model does not have such flexible upper tail behaviour as the GPD, but it is an important special case in many applications and since a wide range of techniques has been developed for both tail index and tail fraction estimation, it is worthy of a brief review. Let $X_{(1)}, ..., X_{(n)}$ represent the data in ascending order, where we assume that heavy-tailed but negative data are transformed to be positive. The Hill estimator for the tail index $\xi = \alpha^{-1}$ based on the $k+1$ upper order statistics,

$$H_k \;=\; \frac{1}{k} \sum_{i=1}^{k} \log X_{(n-i+1)} - \log X_{(n-k)} \;,$$

is the ML estimator for relative excesses from a strict Pareto tail, i.e., with a constant for the slowly varying function in (4.1). Beirlant *et al.* (2004, §4.2) outline alternative derivations of H_k.

The results from the Hill estimator are critically dependent on the tail fraction chosen. The Hill plot, explored by Drees *et al.* (2000), is another graphical diagnostic for prior determination of the tail fraction, which plots the Hill estimator for a range of values of k against either k, the tail fraction, or the corresponding threshold. The value of k is chosen as the largest value (i.e., lowest threshold) such that the Hill estimator has stabilised. Despite its rather different formulation, the Hill plot is essentially the MRL plot of the log-transformed data, so suffers from many of the same benefits and drawbacks, and has been referred to as the Hill horror plot by Resnick (1997). The Hill estimator can exhibit substantial bias if the slowly varying component decays slowly in the limit (Beirlant *et al.*, 2004). The Hill estimator is not shift invariant, although invariant estimators have been proposed by Fraga Alves (2001), Drees (1995, 1998) and Pereira (1994). The latter two issues are not of concern here, as we focus on threshold estimation.

The Hill plot for the Fort Collins data shown in Figure 3 provides a rather different picture of threshold choices than do Figures 1 and 2. The low threshold of $u = 0.395$ inches is suggested to be inappropriate due to the Hill function not having stabilised. The Hill function is rather more stable around the mid-range threshold of $u = 0.85$, but is unstable around the high threshold of $u = 1.2$ inches. So again, the Hill plot gives a different interpretation, thus redemonstrating the expertise required in using these diagnostic plots; see also Drees *et al.* (2000) and references therein. Drees *et al.* (2000) show that the 'altplot' of Resnick & Starica (1997), which is a simple modification to the Hill plot, by essentially using a log-scale on the k-axis, is beneficial for cases where the slowly varying function is non-constant but also state this is not a panacea for threshold determination.

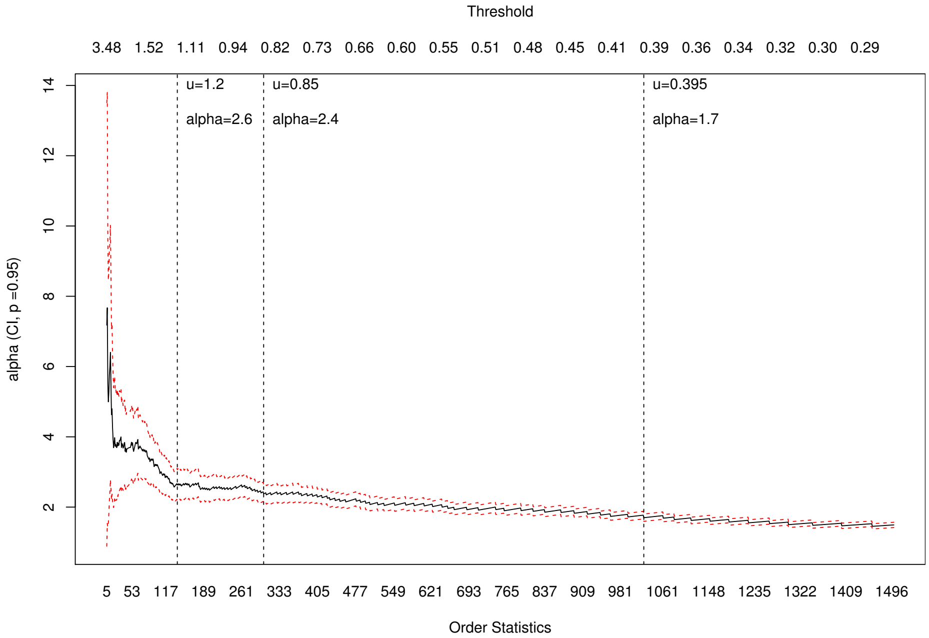


Figure 3: Hill plot of α of regularly varying tail model for the Fort Collins precipitation data as solid line with approximate 95% confidence intervals as dashed lines. The three thresholds $u = 0.395$, 0.85 and 1.2 used in Figure 1 are shown by the vertical dashed lines, with the corresponding Hill estimates.

Hill (1975) also suggested using the fact that the log spacings between the order statistics should be exponentially distributed under the Pareto tail assumption in choosing the tail fraction, by choosing the minimum k such that $\{\log(X_{(n-i+1)}) - \log(X_{(n-k)}) : i = 1, ..., k\}$ does not fail a exponentiality test (e.g., a Kolmogorov–Smirnov or similar test). However, in their work to find an optimal tail fraction Hall & Welsh (1985) showed that this tends to overestimate the tail fraction in large samples. Guillou & Hall (2001) extend the idea of Hill (1975) to derive a plug-in estimator by applying the hypothesis test on an accumulation of the log spacings. Goegebeur *et al.* (2008) further considered a kernel based goodness of fit statistic of the tail fit in the Pareto type tail case, extending Hill (1975), but taking advantage of the relationship between the specific kernel statistics and bias in the asymptotic mean square error of the Hill estimator. Pickands (1975) suggested choosing the tail fraction for the more general GPD case by minimising the distance between empirical and GPD distribution function estimators, with the latter using the Pickands' estimator of the parameters. This idea was extended by Gonzalo & Olmo (2004) using a weighted distance measure, which includes the Pickands' measure and the Kolmogorov–Smirnov statistic as special cases.

Various authors have investigated automated approaches to determining the tail fraction by, for example, minimising the mean square error of estimators of properties of the tail distribution, such as the tail index (Beirlant *et al.*, 1999), the quantiles (Ferreira *et al.*, 2003), or the tail probabilities (Hall & Weissman, 1997), for which optimal asymptotic results typically require second order assumptions in addition to (4.1), see Beirlant *et al.* (2004, §4.7). For example, the plug-in estimator for the tail fraction of Hall (1982) requires prior knowledge of the parameters of the restricted Hall class of tail behaviours. Dekkers (1993) considered an extension of the Hall (1982) estimator to the GPD case using an additional moment-based estimator incorporating the usual Hill estimator, allowing $\xi \leq 0$. Under assumed constraints on these parameters, Hall & Welsh (1985) were able to find a simple adaptive plug-in estimator for the tail fraction, but, the parameters are rather difficult to estimate. The approach of Feuerverger & Hall (1999) suffers similarly from the prescriptive form of the second order characteristics.

Even if the assumptions underlying these approaches are appropriate, their major drawback is that they do not account for the threshold uncertainty on subsequent inferences. Moreover the finite sample properties of such estimators are not well understood. The need to estimate parameters of the unknown population distribution F prior to determination of the optimal tail fraction (e.g., tail index or second order characteristics) has led to the development of bootstrap approaches or algorithmic estimators. As noted by Drees *et al.* (2000), both replace the problem of threshold determination with that of other characteristics — bootstrap sample sizes or stopping rules, respectively.

5. COMPUTATIONAL APPROACHES

Hall (1990) first proposed a resampling based method for estimation of the (asymptotically) optimal tail fraction, by minimising the mean square error (MSE) of the Hill estimator. This approach was applied to the Hall class of Pareto type tails, with a further restriction on the power law parameter of the slowly varying component. A sub-sampling approach is used for overcoming the substantial bias observed when using a standard bootstrap, but it requires an initial estimate of the bias. Gomes & Oliveira (2001) showed the entire procedure is rather sensitive to this estimate. The bias correction uses the relationship between the bootstrap on the full and sub-sample to quantify the bias.

Danielsson *et al.* (2001) overcome much of the restrictiveness of the plug-in estimator of Hall (1990) using a two stage bootstrap procedure for tail index estimation, by minimisation of the asymptotic MSE criterion, which includes automated sample fraction estimation. They avoid the need to know the second order parameter, which is consistently estimated as a consequence of their bootstrap procedure. Two bootstrap estimates of the tail fraction to minimise the MSE of an auxiliary statistic based on the Hill estimator are estimated using two different sample sizes of a particular form. These two estimates are then combined in a closed form to determine the asymptotically optimal tail fraction, and consequently the tail index. The only tuning parameters are the first stage bootstrap resample size and the number of bootstrap samples, the latter being essentially determined by computational capacity. The choice of tuning parameter for the bootstrap sample size at the first stage is in some sense automated by minimisation of a diagnostic criterion over a grid of suitable values. Further, Gomes & Oliveira (2001) have shown that the estimates are robust to the choice of resample size. As this approach is based around the Hill estimator it is restricted to positive shape parameter ($\xi > 0$), though Draisma *et al.* (1990) have considered extension to the GPD case.

Ferreira *et al.* (2003) developed a bootstrap procedure extending that of Danielsson *et al.* (2001), towards optimal estimation of high quantiles (or the upper end-point) by minimisation of the asymptotic MSE. Unlike the similar bootstrap procedure by Hall & Weissman (1997), which is designed for tail probability estimation, they do not require knowledge of the parameters or of second order properties of the distribution for appropriate choice of the bootstrap tuning parameters. They note that its performance is not satisfactory for sample sizes below 2000.

Beirlant *et al.* (1996) consider choosing the tail fraction to provide an optimal linear fit to the Pareto quantile plot (Beirlant *et al.*, 2004, §4.2), which uses the fact that the log transformed Pareto variables are exponential distributed.

They suggest using optimal weights in the regression which require knowledge of the second order parameter, but, estimation of these can be combined in an iterative procedure. However, as with all these approaches, after selection of the appropriate threshold there is no formal assessment of the uncertainty associated with the threshold choice, though some progress has been made by Caers & Dyck (1999) in extending the Hall (1990) approach to account for the tail fraction uncertainty in Pareto power law parameter.

Drees & Kaufmann (1998) provide a hierarchical algorithm to determine the optimal tail fraction for the Hall class. However, this procedure requires prior knowledge of the tail index and second order characteristics (power law of slowly varying function), for which they provide heuristically defined suggestions as part of their simulation study. Beirlant *et al.* (2004, §4.7) reviews comparative studies between this algorithmic, bootstrap based and adaptive Hill based estimators of the tail fraction. In general the restrictive assumptions underlying these approaches hinder their wide applicability.

6. MIXTURE MODELS

The main drawback of most of the aforementioned threshold estimation approaches is that they do not account for the uncertainty associated with the threshold choice. In the last decade, extreme value mixture models have been proposed which encapsulate the usual threshold model in combination with a component intended to capture (some or) all of the non-extreme distribution (henceforth called the 'bulk distribution'). The motivation for ignoring the non-extremal data in early statistical developments in extreme value modelling was mainly that:

1. Extreme and non-extreme events are often physically caused by different underlying processes, which implies there is little information in the bulk distribution for describing the tail behaviour;

2. The GPD is a flexible asymptotically justifiable model for the tail excesses, but classical probability models which combine flexible tail and bulk behaviours are hard to come by and are often application specific;

3. The information content in the sample is typically spread between the high density 'low information' bulk distribution and the low density but 'highly informative' tail observations. The balance between these in terms of influence on the parameters (or related quantities like return levels) is strongly dependent on the model and estimation method. Therefore, when including non-extreme data one has to be careful that the model/estimator provide sufficient relative importance to the tail versus the bulk fit.

The general principle with the mixture models is to combine the flexible threshold model (e.g., GPD), with a suitably flexible and/or application appropriate model of the bulk distribution. The threshold is either implicitly or explicitly defined as a parameter to be automatically estimated, and in most cases the uncertainty associated with the threshold choice can be accounted for naturally in the inferences. In this way, appropriate tail fits can be achieved whilst allowing automated threshold estimation and, provided the bulk distribution model is sufficiently flexible, the bulk and tail fit should not strongly influence each other.

The major drawback of such models is their ad-hoc heuristic definitions, the asymptotic properties of which are still little understood. They have also not had time to be well used in practice and currently there is no readily available software implementation to allow practitioners to gain wider experience. Arguably, the biggest danger with using these models is ensuring that the bulk and tail fits are fairly robust to each other. They cannot be fully disjoint, as they share information, at least about the threshold location, thus ruling out simple application of the EM algorithm. However, robustness of the tail fit to that of the bulk distribution is clearly of major concern. Some discussion of their robustness in this regard will be given, with some intuition, though a fuller comparative study is needed. Another concern with these models is their behaviour at the threshold (e.g., is the fitted density continuous?) and whether attempts to resolve such issues have been successful or have created other problems.

The available mixture models have been broadly classified by the type of bulk distribution models: fully parametric, semiparametric and nonparametric. A general description of the model and estimation approaches as proposed in their original papers will be given, followed by discussion and comparison of their properties and areas for future developments. Many of the common features of the mixture models are summarised in §6.1.

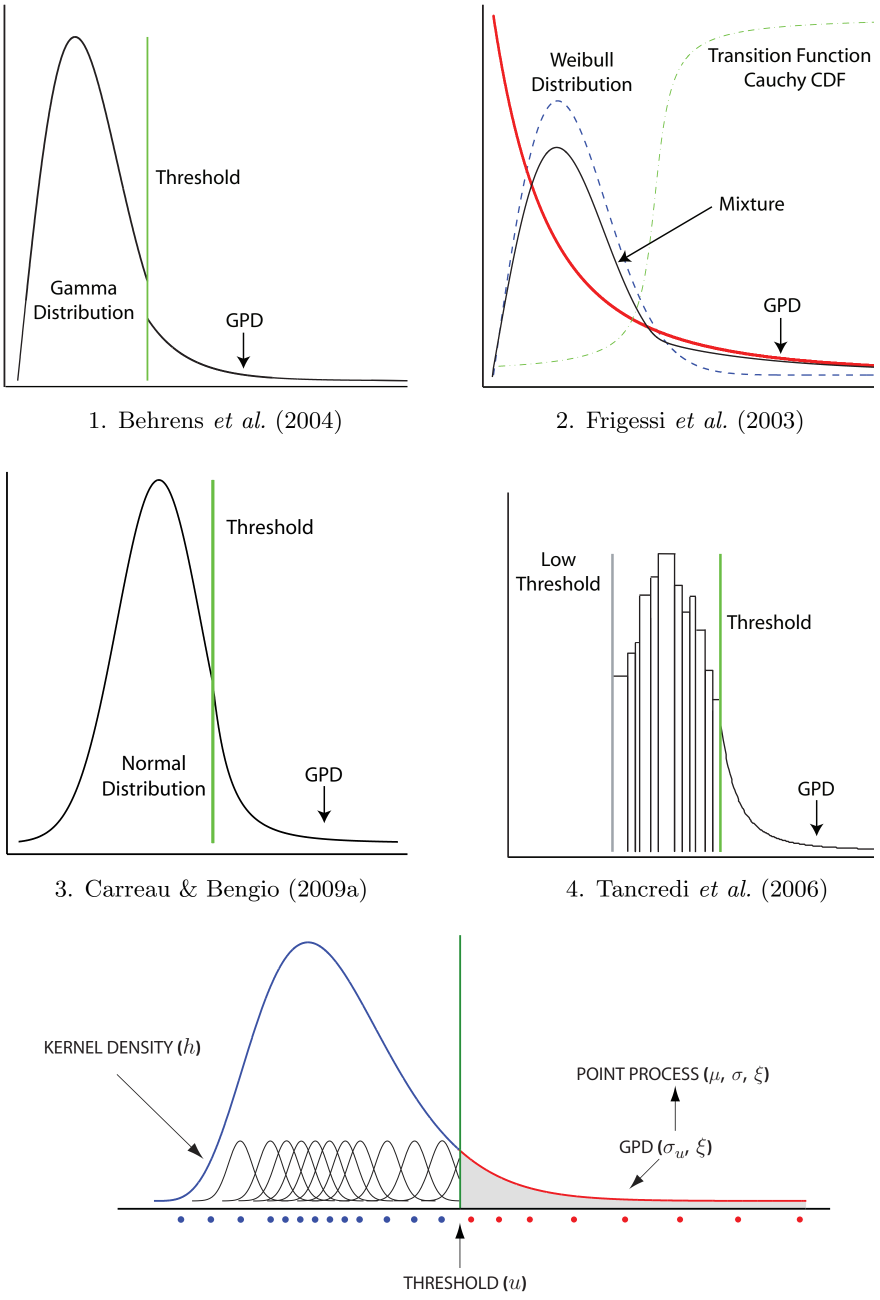
6.1. Parametric bulk models

One of the simplest extreme value mixture models is the spliced distribution consisting of two parametric components:

- A parametric model for the bulk distribution below the threshold — gamma in Behrens *et al.* (2004) and normal in Carreau & Bengio (2009a);
- A threshold tail model (GPD in these proposals) above the threshold.

These components are spliced together at the threshold, which is treated as a parameter to be estimated. Behrens *et al.* (2004) also mooted inclusion of other parametric, semiparametric and nonparametric possibilities below the threshold, but these were not developed further.

The normal spliced with GPD tail developed by Carreau & Bengio (2009a) shown in Figure 4.3, named the 'hybrid Pareto' model, was further developed to include constraints on the parameters to ensure continuity up to the first derivative of the density. However, due to its poor performance in practice, this model was extended to a mixture of hybrid Pareto distributions, and is discussed in §6.2.



1. Behrens *et al.* (2004)

2. Frigessi *et al.* (2003)

3. Carreau & Bengio (2009a)

4. Tancredi *et al.* (2006)

5. MacDonald *et al.* (2011a)

Figure 4: Schematic representations of some mixture models in the literature.

Figure 4.1 illustrates the Behrens *et al.* (2004) model. The discontinuity at the threshold highlights the lack of continuity constraint. This model was specifically designed and evaluated for distributions with a notable discontinuity at the threshold, the relevance of which for applications is debatable. They also explicitly mention that in the case of a smooth transition at the threshold this type of model (and more importantly the estimation procedure) struggles. Some evaluation of performance in the latter situation was reported by Behrens *et al.* (2002). In the case of a strong discontinuity this would lead to a discernible kink in mean excess function at the threshold, so it would be easy to choose a threshold (with little uncertainty) using traditional graphical diagnostics. In the more realistic case of a smooth transition, the traditional diagnostics would also be harder to interpret.

Bayesian inference was used by Behrens *et al.* (2004) with sensible prior forms for the bulk, tail and threshold parameters. However, posterior sampling for their approach is rather inefficient, as they have not accounted for the GPD scale dependence on the threshold. Further, they treat the threshold and tail parameters as independent in the prior. These problems could be easily overcome by suitable transformation of the GPD scale parameter or use of the PPP representation of Smith (1989). The lack of independence of the threshold and GPD scale parameters also makes comprehension of posterior statistics challenging. Orthogonalisation of the GPD scale and shape parameters, or an adaptive posterior sampling scheme (Roberts & Rosenthal, 2009), may help with convergence. Cabras & Castellanos (2010) carried out a comparative study of this simple mixture model and a semiparametric alternative, discussed below.

The major benefits and drawbacks of this parametric approach compared to the semi- and nonparametric alternatives in §§6.2, 6.3, are common to other modelling situations, so are not discussed in detail. In short, if the parametric model is 'correct' then the parametric approach will usually provide the most efficient inferences, but it will suffer when the model is misspecified. There are currently no published results on the performance under model misspecification, and in particular the robustness of the tail fit. In general, the coarse split of information from the bulk and tail data will afford some robustness. However, if the bulk fit is poor, then this will influence the location of the threshold, which will impact the tail fit.

The most beneficial property of this mixture model approach is to provide an objective estimate of a suitable threshold, that provides the best fit to the data (according to the fitting metric used). However, the lack of a continuity (or higher order) constraint at the threshold causes a specific issue with this type of model and for similar approaches described below. The threshold adds an extra degree of freedom, which has a strong localised effect on the fitted distribution function; see Figure 6 of MacDonald *et al.* (2011a). Therefore, if the upper tail of the sample density has spurious peaks or troughs, due to natural sample variability, then the threshold will often be drawn to those locations. This feature is frequently seen in the posterior threshold samples (e.g., modes where sharp changes in the sample density are observed) for this model and others below. However, this type of feature also causes kinks in the MRL and deviations in the threshold stability/Hill plots, so is actually capturing what practitioners would interpret from these traditional graphical diagnostics as well.

A further beneficial property of these mixture models when using Bayesian inference, or sampling based frequentist approaches, is that the entire parameter posterior distribution is available. It is frequently observed in real life applications from the traditional graphical diagnostics that multiple suitable thresholds are plausible, which is naturally accounted for in the inference for these mixture models. The multiple threshold choices show themselves as a bi- or multi-modal posteriors for the threshold, and sometimes other tail parameters; see Scarrott & MacDonald (2010), Figure 4 of Behrens *et al.* (2004) and Figure 2b of Tancredi *et al.* (2006) for example. Computational Bayesian inference approaches are particularly beneficial in this situation to avoid the optimisation challenges associated with likelihood inference when there are multiple modes.

Posterior predictive inference (PPI) is often appropriate in applications where the return levels for future events are of interest; see Coles & Powell (1996). For such mixture models, PPI has a secondary benefit. Even if the underlying density model is potentially discontinuous at the threshold, the PPI will integrate over all posterior threshold possibilities and often provides practically continuous density estimates. See further discussion and examples in do Nascimento *et al.* (2011) and MacDonald *et al.* (2011a).

The first approach in the literature which attempts to give a continuous transition between the bulk and tail models, but unfortunately can fail in applications, is provided by Frigessi *et al.* (2003); see Figure 4.2. Their model is appropriate when there is a lower bound on the support, if the upper tail is of interest (or vice versa). Without loss of generality we assume a zero lower bound. The GPD tail is defined over the whole range of support. The bulk distribution is also defined over the whole range of support, but is presumed to have a light upper tail (e.g., a Weibull tail). A dynamic weight function, such as the cumulative distribution function (cdf) of some smooth unimodal distribution, is then applied to these two components, with highest weight given to the bulk distribution function at low ranges in the support and high weight given to the GPD in the upper tail. A normalization constant ensures that the density has unit integral. The weight function allows the bulk model to dominate the lower tail and the GPD to dominate the upper tail (especially as the bulk model has a light upper tail), but permits a smooth transition. There is no explicit threshold, but the threshold could be estimated by the point at which the relative contribution of the weighted bulk model is sufficiently small compared to the weighted GPD tail model. Frigessi *et al.* (2003) point out that if the transition function is the Heaviside function, then this model includes the Behrens *et al.* (2004) mixture model as a special case.

Frigessi *et al.* (2003) use the Weibull for the bulk model, the GPD for tail and the location-scale Cauchy cdf for the transition function, thus giving a six parameter model. ML estimation is used throughout. However, parameter identifiability can be challenging due to multiple modes, which cause problems for simple black box optimisation schemes. More sophisticated optimisation schemes (e.g., allowing multiple starting points) would likely be beneficial, or alternatively a Bayesian inference MCMC based sampling scheme would be easily implemented. Vrac & Naveau (2007) apply this mixture model to rainfall-runoff modelling.

The idea of a smooth transition is sensible, but two problems occur in practice (MacDonald, 2012). Consider the Cauchy cdf, which has a single parameter controlling the spread of the transition from the bulk to tail model. A quick but smooth transition (approximating the Heaviside function in the previous model) is achieved when the Cauchy scale is almost zero. Therefore, the supposed smooth transition can be lost in application, as the 'localised degree of freedom' of the threshold discussed on page 49 rears its ugly head again. Vrac & Naveau (2007) also noted a fast transition in some applications.

Another problem is the lack of robustness of the tail fit to that of the bulk. Although the weight function controls the relative contributions, because all three components of the model are defined over the whole range of support they all potentially contribute to the fit over entire range. In particular, the GPD reaches its pole at zero, so even though the weight applied to the GPD may be low close to the lower bound if it is non-zero then the relative contribution of the GPD to the lower tail fit can be high compared to the bulk model. Thus, the lower tail fit can impact on the upper tail fit, which is undesirable. Further, the fit in the upper tail is also impacted by the GPD and bulk models as, although at asymptotic levels the GPD will dominate, at sub-asymptotic levels of typically of interest in applications both the Weibull and GPD may contribute to the tail. Thus, the fit to the bulk can again affect the tail fit. This feature can be particularly problematic when there is an exponential or short upper tail.

As this model includes the Behrens *et al.* (2004) model as a special case, most of the benefits and drawbacks carry over. Of particular note is the threshold and GPD scale dependence, which complicates the inference.

Zhao *et al.* (2010) and Mendes & Lopes (2004) introduce a variant on the Behrens *et al.* (2004) mixture model. They propose a mixture with a normal distribution (as an example) for the bulk, with both tails represented by separate threshold models (a so called two-tail model). Neither consider the threshold dependence of the GPD scale parameter.

Zhao *et al.* (2010) apply their model to financial applications where both the gain and loss risks are of interest. The upper and lower thresholds are estimated in tandem with the other parameters in a Bayesian framework. Thus their approach, like that of Behrens *et al.* (2004), allows for automated threshold choice and uncertainty quantification. Zhao *et al.* (2011) and Zhao (2010) consider the testing of asymmetry of the gain and losses tails, by comparing the model fit between a general parameterisation of both tails and the two tails having a common shape parameter.

Prior to application of their two tail model Mendes & Lopes (2004) robustly standardise the data to make 'well defined' tails, though it is unclear what is meant by this phrase, as they apply the same linear rescaling to all the observations, thus not providing any separation of the bulk and tail data. A robust nonlinear transformation which pushes out the tail data and shrinks in the bulk observations would be needed to achieve this goal. After standardising the data Mendes & Lopes (2004) propose a rather lengthy estimation procedure:

1. Select a grid of candidate tail fractions for both tails, and estimate the corresponding empirical threshold (quantile) estimates);
2. Fit a standard normal or Student-t distribution to the robustified data and use it to find robust estimates of the two tail fractions;
3. Use L-moment estimators to robustly fit the GPD to both tails;
4. Calculate the likelihood for the combined two-tail mixture model with these robust estimates of the tail fractions, thresholds and GPD parameters;
5. The final selection of the tail fractions is that which maximises the likelihood.

This procedure provides an objective automated threshold estimate, but the threshold uncertainty is ignored and it is not obvious how to evaluate the overall uncertainty.

6.2. Semiparametric bulk models

Carreau & Bengio (2009a) propose the hybrid Pareto model depicted in Figure 4.3, which constrains the bulk and tail densities to have continuous zeroth and first derivatives to ensure a smooth transition at the threshold. The five parameters are thus reduced to three. Carreau & Bengio (2009a) chose to vary the mean and location of the normal and the GPD shape parameter, from which the threshold and GPD scale can be derived.

The fundamental idea underlying the hybrid Pareto to ensure smoothness at the threshold is sensible. Unfortunately, the model performs poorly in practice, seemingly because these two constraints are rather strong, as noted by Carreau & Bengio (2009a). Exploration of the single constraint of continuity would confirm this. Carreau & Bengio (2009a) used a finite mixture of these hybrid Pareto densities as their final model to overcome this. This gives a distribution of thresholds across the mixture components. Further, the tail is approximated by multiple GPD's. Asymptotically the mixture component with the heaviest tail behaviour will dominate, but the sub-asymptotic behaviour is influenced by all the tail components. Carreau & Bengio (2009a) suggest the threshold for the dominant tail component could be used as a proxy for the threshold choice if this is of interest.

This model bypasses the threshold estimation problem, but it has been included because the final estimates in some sense encapsulate the uncertainty associated with the threshold estimation. Although Carreau & Bengio (2009a) call their mixture of hybrid Pareto's model a nonparametric model, it is categorised here as a semiparametric approach as the number of degrees of freedom used in estimation will generally be small, growing slowly with the sample size, compared to the nonparametric mixtures outlined below.

Maximum likelihood estimation is used throughout, including for the number of mixture components, although a neural network learning approach is implemented by Carreau & Bengio (2009b) and Carreau *et al.* (2009) in an extension to nonstationary and bivariate modelling situations. Carreau & Bengio (2009a) point out the the interrelation between the parameters means that the standard EM algorithm cannot be used, as also indicated by Behrens *et al.* (2004) and MacDonald *et al.* (2011a) for related mixture models.

The performance of the mixture of hybrid Pareto models in situations where the tail behaviour is exponential or short tailed needs exploration. In these cases, the normal bulk model in each component of the mixture could adequately approximate the tail, so the implied threshold would be rather variable and it is unclear how this would affect the inferences.

A major benefit of the mixture of hybrid Pareto's model compared to the other mixtures is the required continuity in the zeroth and first derivatives at the threshold. This will reduce the local degree of freedom problem of the threshold in the other mixtures mentioned above, but will introduce greater flexibility, so it is not clear whether the uncertainty associated with this will be reduced, particularly in the mixture of hybrid Paretos. A drawback with this mixture is the parameter dependence, which means there may be a lack of robustness of the tail fit to that of the bulk, as with the dynamically weighted mixture of Frigessi *et al.* (2003). However, the flexibility may override this lack of robustness, as with the fully nonparametric mixture models of Tancredi *et al.* (2006) and MacDonald *et al.* (2011a) outlined below. It would be interesting to investigate both of these issues in future research.

Cabras & Castellanos (2010) consider another semiparametric bulk model spliced with a GPD upper tail. The bulk distribution is approximated by an equi-spaced binning of the data, followed by fitting a Poisson log-link GLM to the counts, with a polynomial smoother for the mean parameter. A profile likelihood for all the observations is obtained by combining the GPD tail model for observations above the threshold and Poisson GLM density estimator below the threshold. However, the fitted density for the bulk distribution is assumed to be fixed, so the likelihood is maximised with respect to only the threshold and GPD parameters. Appropriate scalings are applied to ensure that the density integrates to unity. Bayesian inference is used, but the threshold dependence of the GPD scale parameter is ignored.

do Nascimento *et al.* (2011) extended the Behrens *et al.* (2004) model by defining the bulk distribution as a weighted mixture of gamma densities. They use Bayesian inference, conditional on a fixed number of the gamma components, which is chosen using a BIC- or DIC-based statistic. This ensures that specific parametric forms or constraints such as unimodality are not imposed, yielding a flexible model for the bulk distribution. This mixture model is unlike that of the hybrid model described by Carreau & Bengio (2009a), as it relies on a single GPD for tail estimate and thus requires that only one threshold be estimated. do Nascimento *et al.* (2011) also showed that the use of posterior predictive inference practically eliminates any discontinuity at the threshold, even though the individual posterior samples will likely exhibit a discontinuity.

The major benefit of these semiparametric mixtures over the parametric ones is that they provide reasonably flexible models for the bulk without using up as many degrees of freedom as the nonparametric alternatives discussed below. Thus they combine some of the benefits of both these approaches. Of course, it is not clear whether they provide sufficient flexibility to be robust to model misspecification, particularly robustness of the tail fit to the bulk.

6.3. Nonparametric bulk models

Tancredi *et al.* (2006) were the first to propose an extreme value mixture model combining a nonparametric estimator for the bulk distribution spliced with an extreme value tail model. They were also the first to overcome the dependence of the GPD scale parameter on the threshold by using the PP representation for the tail excesses. Their bulk model is the mixture of uniforms density estimator of Robert (1998) and Robert & Casella (2010), providing a piecewise linear approximation to the cdf below the threshold; see Figure 4.4. The nonparametric mixture is defined between the upper threshold and a lower threshold that is definitely too low (which could be the lower bound on the range of support).

A benefit of this approach is that it fixes the sample size being included in the inference, but potentially excludes data which is uninformative about the tail behaviour.

Bayesian inference using MCMC is implemented. A hierarchy determines whether the uniform widths should be the same or should be allowed to vary and to what degree. However, as there is an unknown number of uniform densities, the parameter space varies in dimension and a reversible jump algorithm is required (Robert & Casella, 2010), whose implementation can be challenging.

MacDonald *et al.* (2011a) use a kernel density estimator for the bulk distribution, spliced at the threshold with a PPP tail model shown in Figure 4.5, following a cruder version with a GPD tail model (Scarrott & MacDonald, 2010). Nonparametric kernel density estimators using symmetric kernels work well with populations with unbounded support, or at least a proper tail before the lower boundary. For populations with bounded support, a boundary-corrected kernel density estimator is used as an alternative by MacDonald *et al.* (2011b). The kernel density estimator assumes a particular kernel such as the normal density, which is centred at each datapoint, and is parameterised by a single bandwidth. Such kernel density estimators can approximate most smooth densities (Silverman, 1986). If the lower tail is heavy then MacDonald *et al.* (2011a) also showed that a mixture model, with both tails replaced by threshold models, can be used to provide flexibility and robustness of the fits between the two tails and the bulk.

MacDonald *et al.* (2011a) utilise standard cross-validation likelihood to choose the bandwidth, combined with the likelihood for PPP tail model, including the threshold, to give a full likelihood for all the observations. The combined likelihood is used in a Bayesian inference framework, with posterior predictive inference used for all the key quantities of interest.

The major benefit of these nonparametric approaches as compared to the parametric approaches is that the tail fit is robust to the bulk fit, as demonstrated by simulation in Tancredi *et al.* (2006). MacDonald *et al.* (2011b) and MacDonald (2012) use sensitivity curves to show the robustness of the tail fit to that of the bulk and vice-versa. The main drawback with the mixture of uniforms for the bulk distribution of Tancredi *et al.* (2006) is the computational complexity, noted by Thompson *et al.* (2009), and in particular the difficulty of ensuring convergence. The nonparametric kernel density estimator is computationally simpler, but the cross-validation likelihood can be burdensome for large samples, and the overall computing time can be reduced by other forms of subsampling or by using an alternative penalty function. Further, both nonparametric mixture models could be fit in two stages: (1) fit the nonparametric density estimator, followed by (2) put this in a combined likelihood with the tail model and carry out the tail inferences ignoring the uncertainty associated with the bulk parameter estimation, as in the profile likelihood approach of Cabras & Castellanos (2010).

The lack of sensitivity of the bulk parameter estimation to the tail parameters would mean that little would be lost while substantially reducing the computational burden.

7. OTHER APPROACHES

Dupuis (2000) proposed a robust procedure for GPD fitting, including statistics to guide threshold choice. The optimal bias robust estimation (OBRE) procedure weights each observation between 0 and 1 depending on how consistent they are with the GPD model, with high weights indicating a good fit. A test statistic (or p-value) for the weights under the null hypothesis that the GPD is the correct model is also provided, along with average weights from simulations under the fitted models for further guidance. She suggests trying out a sensible set of thresholds and choose the lowest threshold such that the weights are all 'sufficiently close to one'.

This procedure essentially replaces the threshold selection problem with that of selecting the tuning parameter which controls the cut-off of closeness to unity. Dupuis (2000) suggested that this new threshold choice is easier to automate. The OBRE also requires specification of the bound on the influence function, which balances efficiency and robustness of the estimator (too high a bound leads to a lack of robustness but higher efficiency as more information is utilised, and vice versa). Some guidance is required to provide a suitable choice for this bound, but this adds to concerns whether this this approach can be fully automated. The principle drawback with this approach is that after choosing the threshold, the OBRE procedure reduces to a fixed threshold approach.

Thompson *et al.* (2009) recently developed an automated procedure for threshold estimation and uncertainty quantification. They set a uniformly spaced grid of possible threshold values (between the median and 98% empirical quantile). For each potential threshold the GPD is fitted (using ML estimation) and the differences in the modified scale parameters for neighbouring thresholds is calculated. They assume asymptotic normality of these scale differences, and use the fact that they are centred around mean zero if the GPD is the correct model. They pool these differences and treat them as a sample of normal random variables. A forward selection procedure (increasing the threshold increment by increment) is then applied, until the Pearson χ^2 test under the assumption of normality and mean zero is not rejected. At this point the threshold is said to be consistent with a GPD. A simple bootstrap approach is used to quantify the threshold uncertainty on the final inferences.

de Zea Bermudez *et al.* (2001) use a Bayesian hierarchical model for tail estimation, essentially averaging over all possible values of tail fraction. The GPD is used to describe the tail excesses conditional on the tail fraction, with the prior distribution defined at the next level of the hierarchy. The tail fraction 'parameter' k at the next level in the hierarchy is described by a doubly truncated Poisson distribution with parameters for the mean and truncation points chosen to be relatively uninformative, but still ensuring a suitable tail fraction. The next hierarchy level describes the priors for the three tail fraction parameters, followed by their hyperparameters at the final level. They consider only the Gumbel case (exponential upper tail). In order to avoid the specification of the bulk model, as used in the extreme value mixture models outlined in §6, they use an approximate posterior, assuming that the contribution to the likelihood from the lower order statistics can be safely ignored and only the conditional contributions for the upper tail fraction contribute. Therefore, for each random effect for the tail fraction k in the hierarchy, there is a differing amount of data contributing to the approximate likelihood (and therefore posterior). The properties of such an estimator (e.g., asymptotics, relative weighting of data points) are unclear.

Cabras & Morales (2007) propose using a sequential outlier detection method to identify points above (below) some upper (lower) threshold which appear to differ in character from the tail(s) of some assumed known parametric model for the population distribution. The partial posterior predictive distribution is used to sequentially estimate the likelihood of observing the pairs of the most outlying (upper and lower) order statistics under the assumed population model, ignoring those that have already been dropped. This procedure continues until the probability of observing both in the pair is sufficiently high. The thresholds are then defined by the last upper and lower order statistics that were dropped. Of course, this requires specification of the outlyingness tolerance, which the user must specify *a priori*, though the authors claim this will be straightforward to elicit. In some ways this approach is related to the two-tail models considered by Zhao *et al.* (2010) and Mendes & Lopes (2004) discussed above. No threshold uncertainty quantification is applied.

ACKNOWLEDGMENTS

Anna MacDonald received support from both a New Zealand TEC Top Achievers Doctoral Scholarship under the Bright Futures Scheme and a Canterbury Scholarship from the University of Canterbury. We wish to thank the editors for the invitation to prepare this review. Much of this paper was written whilst Carl Scarrott was on sabbatical at Lancaster University, UK, whose support is acknowledged, particularly discussions with Jonathan Tawn and Jennifer Wadsworth.

REFERENCES

Balkema, A.A. & de Haan, L. (1974). Residual life time at great age, *Ann. Prob.*, **2**, 792–804.

Behrens, C.N.; Lopes, H.F. & Gamerman, D. (2002). *Bayesian analysis of extreme events with threshold estimation*, Technical report, Laboratório de Estatística, Universidade Federal do Rio de Janeiro.

Behrens, C.N.; Lopes, H.F. & Gamerman, D. (2004). Bayesian analysis of extreme events with threshold estimation, *Statist. Mod.*, **4**, 227–244.

Beirlant, J.; Dierckx, G.; Goegebeur, Y. & Matthys, G. (1999). Tail index estimation and an exponential regression model, *Extremes*, **2**, 177–200.

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

Beirlant, J.; Vynckier, P. & Teugels, J.L. (1996). Tail index estimation, Pareto quantile plots and regression diagnostics, *J. Am. Statist. Assoc.*, **91**, 1659–1667.

Cabras, S. & Castellanos, M.E. (2010). *An objective Bayesian approach for threshold estimation in the peaks over the threshold model*, Technical Report TR2010.10, Análisis de riesgo.

Cabras, S. & Morales, J. (2007). Extreme value analysis within a parametric outlier detection framework, *Appl. Stoch. Mod. Bus. Ind.*, **23**, 157–164.

Caers, J. & Dyck, J.V. (1999). Nonparametric tail estimation using a double bootstrap method, *Comp. Statist. Data Anal.*, **29**, 191–211.

Carreau, J. & Bengio, Y. (2009a). A hybrid Pareto model for asymmetric fat-tailed data: the univariate case, *Extremes*, **12**, 53–76.

Carreau, J. & Bengio, Y. (2009b). A hybrid Pareto mixture for conditional asymmetric fat-tailed distributions, *IEEE Trans. Neur. Net.*, **20**, 1087–1101.

Carreau, J.; Naveau, P. & Sauquet, E. (2009). A statistical rainfall-runoff mixture model with heavy-tailed components, *Water Resourc. Res.*, **45**(W10437).

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

Coles, S.G. & Powell, E.A. (1996). Bayesian methods in extreme value modelling: A review and new developments, *Int. Statist. Rev.*, **64**, 119–136.

Cox, D. & Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion), *J. R. Statist. Soc. B*, **49**, 1–39.

Danielsson, J.; de Haan, L.; Peng, L. & de Vries, C.G. (2001). Using a bootstrap method to choose the sample fraction in tail index estimation, *J. Mult. Anal.*, **76**, 226–248.

Davison, A.C. & Smith, R.L. (1990). Models for exceedances over high thresholds (with discussion), *J. R. Statist. Soc. B*, **52**, 237–254.

de Zea Bermudez, P.; Amaral Turkman, M.A. & Turkman, K.F. (2001). A predictive approach to tail probability estimation, *Extremes*, **4**, 295–314.

de Zea Bermudez, P. & Kotz, S. (2010). Parameter estimation of the generalized Pareto distribution – Part II, *J. Statist. Plan Inf.*, **140**, 1374–1388.

Dekkers, A.L.M. (1993). Optimal choice of sample fraction in extreme-value estimation, *J. Mult. Anal.*, **47**, 173–195.

do Nascimento, F.F.; Gamerman, D. & Lopes, H.F. (2011). A semiparametric Bayesian approach to extreme value estimation, *Statist. Comput.* pp. Early View: 10.1007/s11222–011–9270–z.

Draisma, G.; de Haan, L.; Peng, L. & Pereira, T.T. (1999). A bootstrap-based method to achieve optimality in estimating the extreme-value index, *Extremes*, **2**, 367–404.

Drees, H. (1995). Refined Pickands estimators of the extremal index, *Ann. Statist.*, **23**, 2059–2080.

Drees, H. (1998). Optimal rates of convergence for estimates of the extreme value index, *Ann. Statist.*, **26**, 434–448.

Drees, H.; de Haan, L. & Resnick, S.I. (2000). How to make a Hill plot, *Ann. Statist.*, **28**, 254–274.

Drees, H. & Kaufmann, E. (1998). Selecting the optimal sample fraction in univariate extreme value estimation, *Stoch. Proc. Appl.*, **75**, 149–172.

DuMouchel, W.H. (1983). Estimating the stable index α in order to measure tail thickness: A critique, *Ann. Statist.*, **11**, 1019–1031.

Dupuis, D.J. (2000). Exceedances over high thresholds: A guide to threshold selection, *Extremes*, **1**, 251–261.

Ferreira, A.; de Haan, L. & Peng, L. (2003). On optimising the estimation of high quantiles of a probability distribution, *Statistics*, **37**, 401–434.

Feuerverger, A. & Hall, P.G. (1999). Estimating a tail exponent by modelling departure from a Pareto, *Ann. Statist.*, **27**, 760–781.

Fraga Alves, M.I. (2001). A location invariant Hill-type estimator, *Extremes*, **4**(3), 199–217.

Frigessi, A.; Haug, O. & Håvard, R. (2003). A dynamic mixture model for unsupervised tail estimation without threshold selection, *Extremes*, **5**, 219–235.

Gilleland, E.; Katz, R.W. & Young, G. (2010). *extRemes: Extreme value toolkit.* R package version 1.62.

Goegebeur, Y.; Beirlant, J. & de Wet, T. (2008). Linking Pareto-tail kernel goodness-of-fit statistics with tail index at optimal threshold and second order estimation, *Revstat*, **6**, 51–69.

Gomes, M.I. & Oliveira, O. (2001). The bootstrap methodology in statistics of extremes: Choice of the optimal sample fraction, *Extremes*, **4**, 331–358.

Gonzalo, J. & Olmo, J. (2004). Which extreme values are really extreme? *J. Fin. Econ.*, **2**, 349–369.

Guillou, A. & Hall, P.G. (2001). A diagnostic for selecting the threshold in extreme value analysis, *J. R. Statist. Soc. B*, **63**, 293–305.

Hall, P.G. (1982). On some simple estimates of an exponent of regular variation, *J. R. Statist. Soc. B*, **44**, 37–42.

Hall, P.G. (1990). Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems, *J. Mult. Anal.*, **32**, 177–203.

Hall, P.G. & Weissman, I. (1997). On the estimation of extreme tail probabilities, *Ann. Statist.*, **25**, 1311–1326.

Hall, P.G. & Welsh, A.H. (1985). Adaptive estimates of the parameters of regular variation, *Ann. Statist.*, **13**, 331–341.

Hill, B.M. (1975). A simple general approach to inference about the tail of a distribution, *Ann. Statist.*, **3**, 1163–1174.

Ho, A.K.F. & Wan, A.T.K. (2002). Testing for covariance stationarity of stock returns in the presence of structural breaks: An intervention analysis, *Appl. Econ. Lett.*, **9**, 441–447.

Katz, R.W.; Parlange, M.B. & Naveau, P. (2002). Statistics of extremes in hydrology, *Adv. Water Res.*, **25**, 1287–1304.

Leadbetter, M.R.; Lindgren, G. & Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Series*, Springer, London.

Loretan, M. & Philips, P.C.B. (1994). Testing the covariance stationarity of heavy tailed time series: an overview of the theory with applications to several financial datasets, *J. R. Statist. Soc. D*, **1**, 211–248.

MacDonald, A. (2012). *Extreme value mixture modelling with medical and industrial applications*, PhD thesis, University of Canterbury, New Zealand.

MacDonald, A.; Scarrott, C.J.; Lee, D.; Darlow, B.; Reale, M. & Russell, G. (2011a). A flexible extreme value mixture model, *Comp. Statist. Data Anal.*, **55**, 2137–2157.

MacDonald, A.; Scarrott, C.J. & Lee, D.S. (2011b). Boundary correction, consistency and robustness of kernel densities using extreme value theory. Submitted. Available from: `http://www.math.canterbury.ac.nz/ c.scarrott`.

Mendes, B. & Lopes, H.F. (2004). Data driven estimates for mixtures, *Comp. Statist. Data Anal.*, **47**, 583–598.

Neves, C. & Alves, M.I.F. (2004). Reiss and Thomas' automatic selection of the number of extremes, *Comp. Statist. Data Anal.*, **47**, 689–704.

Omran, M.F. & McKenzie, E. (1999). Testing for covariance stationarity in the uk all-equity, *J. R. Statist. Soc. D*, **48**, 361–369.

Pereira, T.T. (1994). *Second order behaviour of domains of attraction and the bias of generalized Pickands' estimator.* In "Extreme Value Theory and Applications III", Proceedings of Gaithersburg Conference, 1993 (J. Galambos, L. Lechner and E. Simiu, Eds.), vol. 866, pp. 165–177.

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

Reiss, R.-D. & Thomas, M. (2007). *Statistical Analysis of Extreme Values: With Applications to Insurance, Finance, Hydrology and Other Fields*, Birkhauser, Boston.

Resnick, S.I. (1997). Heavy tail modeling and teletraffic data, *Ann. Statist.*, **25**, 1805–1869.

Resnick, S.I. & Starica, C. (1997). Smoothing the Hill estimator, *Adv. Appl. Prob.*, **29**, 271–293.

Robert, C.P. (1998). *Discretization and MCMC Convergence Assessment*, Lecture Notes in Statistics, vol. 135, Springer, New York.

Robert, C.P. & Casella, G. (2010). *Monte Carlo Statistical Methods*, Springer, New York.

Roberts, G.O. & Rosenthal, J.S. (2009). Examples of adaptive MCMC, *J. Comp. Graph. Statist.*, **18**, 349–367.

Scarrott, A. & MacDonald, C.J. (2010). Risk assessment for critical temperature exceedance in nuclear reactors, *J. Risk Rel.*, **224**.

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

Smith, R.L. (1989). Extreme value analysis of environmental time series: An example based on ozone data (with discussion), *Statist. Sci.*, **4**, 367–393.

Tancredi, A.; Anderson, C.W. & O'Hagan, A. (2006). Accouting for threshold uncertainty in extreme value estimation, *Extremes*, **9**, 87–106.

Thompson, P.; Cai, Y.; Reeve, D. & Stander, J. (2009). Automated threshold selection methods for extreme wave analysis, *Coastal Engineering*, **56**, 1013–1021.

Vrac, M. & Naveau, P. (2007). Stochastic downscaling of precipitation: From dry events to heavy rainfalls, *Water Resourc. Res.*, **43**(W07402).

Zhao, X. (2010). *Extreme value modelling with application in finance and neonatal research*, PhD thesis, University of Canterbury, New Zealand.

Zhao, X.; Scarrott, C.J.; Oxley, L. & Reale, M. (2011). *Let the tails speak for themselves": Bayesian extreme value mixture modelling for estimating VaR*, Submitted. Available from: `http://www.math.canterbury.ac.nz/∼c.scarrott`.

Zhao, X.; Scarrott, C.J.; Reale, M. & Oxley, L. (2010). Extreme value modelling for forecasting the market crisis, *Appl. Fin. Econ.*, **20**, 63–72.