A Flexible Extreme Value Mixture Model

A. MacDonald[a], C.J. Scarrott[a,*], D. Lee[a], B. Darlow[b], M. Reale[a], G. Russell[c]

[a] *Mathematics and Statistics Department, University of Canterbury, Private Bag 4800, Christchurch, New Zealand*

[b] *Department of Pediatrics, Christchurch School of Medicine and Health Science, University of Otago, Christchurch, New Zealand*

[c] *National Health Service, Imperial College, London, UK*

Abstract

Extreme value theory is used to derive asymptotically motivated models for unusual or rare events, e.g. the upper or lower tails of a distribution. A new flexible extreme value mixture model is proposed combining a nonparametric kernel density estimator for the bulk of the distribution with an appropriate tail model. The complex uncertainties associated with threshold choice are accounted for and new insights into the impact of threshold choice on density and quantile estimates are obtained. Bayesian inference is used to account for all uncertainties and enables inclusion of expert prior information, potentially overcoming the inherent sparsity of extremal data. A simulation study and empirical application for determining normal ranges for physiological measurements for pre-term infants is used to demonstrate the performance of the proposed mixture model.

Keywords: Extreme values, mixture model, kernel density, threshold selection

1. Introduction

Extreme value theory is unlike most traditional statistical theory, which typically examines the "usual" or "average" behaviour of process, in that it is used to motivate models for describing the unusual behaviour or rare events of a process. Practical applications are seen in many fields of endeavour including finance (Embrechts et al., 2003), engineering (Castillo et al., 2004) and environment science (Reiss and Thomas, 2007), where the risk of rare events are of interest. At the heart of extreme value techniques is reliable extrapolation of risk estimates past the observed range of the sample data. Typically, a parametric extreme value model for describing the upper (or lower) tail of the data generating process is proposed, which is fitted to the available extreme value data. The model performance is evaluated by how well is describes the observed tail behaviour of the sample data. If the model provides a good fit then it is used for extrapolation of the quantities of interest, e.g. typically certain high quantiles, with estimation of the associated extrapolation uncertainty.

[*] Corresponding author. Tel.: +64 3 3642600; fax.: +64 3 3642587
Email address: `carl.scarrott@canterbury.ac.nz` (C.J. Scarrott)
URL: `http://www.math.canterbury.ac.nz/∼c.scarrott` (C.J. Scarrott)

2. Classical Extreme Value Model

Davison and Smith (1990) showed that for a sequence of independent and identically distributed observations $\{x_i : i = 1, \ldots, n\}$, under certain mild conditions, the excesses $x - u$ of some suitably high threshold u can be well approximated by a generalised Pareto distribution, denoted by $\mathrm{GPD}(\sigma_u, \xi)$, with:

$$G(x|\xi, \sigma_u, u) = \Pr(X < x|X > u) = \begin{cases} 1 - \left[1 + \xi\left(\dfrac{x-u}{\sigma_u}\right)\right]_+^{-1/\xi} & \xi \neq 0, \\ \\ 1 - \exp\left[-\left(\dfrac{x-u}{\sigma_u}\right)\right]_+ & \xi = 0, \end{cases} \tag{1}$$

where $x > u$, $y_+ = \max(y, 0)$ and ξ and $\sigma_u > 0$ are the shape and scale parameters respectively. The unconditional survival probability is then given by:

$$\Pr(X > x) = \phi_u[1 - \Pr(X < x|X > u)] \tag{2}$$

where ϕ_u is the probability of being above the threshold u. The value of the shape parameter ξ is key in determining the tail extrapolations of the GPD:

- $\xi = 0$ exponential tail, considered in the limit $\xi \to 0$;
- $\xi > 0$ heavier tail than exponential (i.e. power law decay); and
- $\xi < 0$ short tail with finite upper endpoint $u - \sigma_u/\xi$.

An alternative representation of the GPD is available which removes the dependence of the parameters on the threshold, leading to simpler inferences and better mixing of Markov Chain Monte Carlo (MCMC) chains discussed in Section 5. For a sufficiently high threshold u the point process defined by

$$P_n = \left\{\left(\frac{i}{n+1}, X_i\right); i = 1, ..., n\right\}$$

is well approximated by an inhomogeneous Poisson process on the region $A = [0, 1] \times (u, \infty)$ with intensity function on the subregion $B = (t_1, t_2) \times (x, \infty)$ given by:

$$\Lambda(B) = \begin{cases} (t_2 - t_1)\, n_b \left[1 + \xi\left(\dfrac{x-\mu}{\sigma}\right)\right]_+^{-1/\xi} & \xi \neq 0 \\ \\ (t_2 - t_1)\, n_b \exp\left\{-\left(\dfrac{u-\mu}{\sigma}\right)\right\} & \xi = 0 \end{cases} \tag{3}$$

where $x > u$ and the scaling constant n_b is the number of blocks of observations (e.g. number of years of daily data). Modelling extremes using the point process (PP) framework, follows a similar process as that of threshold modelling via the GPD above, with the PP conditional probability model parameterised by (μ, σ, ξ). As with the GPD, application of the PP theory relies on the choice of a suitably high threshold u, above which the asymptotically motivated PP model can provide a reliable approximation.

While the value of n_b can be seen as completely arbitrary, as for any particular choice the impact on the PP parameters is deterministic, it is possible to define n_b in such a way that the three classical extreme value models (namely block maxima or generalised extreme value model, r-largest model and threshold excess GPD model) can be derived as special cases. The GPD is a special case where n_b is set to be the number of threshold exceedances, with the major benefit of the PP parameters (μ, σ, ξ) not being dependent on the threshold as is the case with the GPD. In particular, the shape parameter for both the PP and GPD models are the same and the GPD scale parameter σ_u is related to the PP parameters by $\sigma_u = \sigma + \xi(u - \mu)$. However, the value of n_b will change the properties of the likelihood, so can be adjusted to improve the likelihood properties and therefore computational aspects. The PP parameters corresponding to n_b blocks (ξ_b, σ_b, μ_b) are related to a PP with n_y blocks (ξ_y, σ_y, μ_y) by:

$$\xi_y = \xi_b \qquad \sigma_y = \sigma_b \left(\frac{n_b}{n_y}\right)^{\xi_b} \qquad \mu_y = \mu_b - \frac{\sigma_b}{\xi_b}\left[1 - \left(\frac{n_b}{n_y}\right)^{\xi_b}\right] \tag{4}$$

It is common practice to use properties of the GPD/PP models to aid threshold selection, often using graphical diagnostics. For example, a mean excess plot shows various thresholds plotted against average excess above the threshold. Once a sufficiently high threshold u has been reached then (if the tail follows a GPD) the mean excesses above this threshold $v \geq u$ will be linear as:

$$E(X - v|X > v) = \{\sigma_u + \xi(v - u)\}/(1 - \xi),$$

where $\xi <$, see Embrechts et al. (2003) for details. Threshold selection using these diagnostics frequently requires subjective expert judgement, and for some applications the choice of a suitable threshold u can have a substantial influence on tail extrapolation. General principles to follow (Coles, 2001) are to maximise the amount of data for efficient inference, without selecting too low a threshold such that the asymptotic theory underlying the tail models is invalidated.

Further, traditional inference approaches for the GPD/PP models assume the threshold, once chosen, is a fixed quantity so the estimation uncertainty is not accounted for in further inferences. Automating the threshold choice for efficient application to many datasets has also proven elusive. Dupuis (2000) has recently developed a more robust technique to aid threshold choice which is designed to be easier to automate, but can still require subjective judgement. However, even in this approach the uncertainty associated with threshold choice is not accounted for.

3. Extreme Value Mixture Models

Various mixture models have been proposed for the entire distribution function, simultaneously capturing the bulk of the distribution (typically the main mode) with the flexibility of an extreme value model for the upper and/or lower tails. These mixture models either explicitly include the threshold as a parameter to be estimated, or somewhat bypass this choice by the use of smooth transition functions between the bulk and tail components, thus overcoming the issues associated with threshold choice and uncertainty estimation.

Mendes and Lopes (2004) propose a mixture model where the main mode is assumed to be normal and two separate GPD models are used for the tails. Estimation of the threshold is carried out by either quasi-likelihood procedure or a model fit statistic. Frigessi et al. (2002) proposed a dynamically weighted mixture model, where the weight function varies over the range of support, shifting the weights from a light-tailed density (such as the Weibull) for the main mode, to the GPD which will dominate the upper tail. There is no explicit threshold in this approach, however a threshold could be determined by the point at which the weighted contribution from the Weibull is sufficiently small compared to the GPD. Behrens et al. (2004) present a mixture model that combines a parametric form for the bulk distribution (e.g. Gamma, Weibull or Normal) upto some threshold and a GPD for the tail above this threshold. In their approach, the threshold is explicitly treated as a parameter to be estimated. Recently, Carreau and Bengio (2009) introduced a hybrid Pareto distribution (combination of normal and GPD tail, with density constrained to be continuous upto first derivative) to approximate the distribution with support on entire real axis, including extension to a mixture of these hybrid Pareto's to capture possible asymmetry, multi-modality and tail heaviness of the underlying density.

The drawback with all the aforementioned approaches is the prior specification of a parametric model for the mode of the distribution (and associated weight function where appropriate), and the complicated inference (and sample properties) for the mixture of hybrid-Pareto's. Tancredi et al. (2006) proposed a mixture model comprising of piecewise uniform distributions from a threshold which is known to be too low, upto the actual threshold above which the PP model is used. Their approach can essentially be seen as a piecewise linear approximation to the distribution function below the actual threshold, with PP model based tail above. Bayesian inference is used with a reversible jump algorithm due to the unknown number of uniforms required. The actual threshold is defined as a parameter of the model, so the inference approach naturally accounts for the threshold uncertainty.

In this paper we propose a flexible model to analyse extremal events which includes a non-parametric smooth kernel density estimator below some threshold accompanied with the PP model for the upper tail above the threshold. This model avoids the need to assume a parametric form for the bulk distribution, and captures the entire distribution function below the threshold using a smooth flexible non-parametric form. This flexible extremal model has one extra (kernel bandwidth) parameter above the usual PP parameters (and threshold), thus potentially simplifying computational aspects of the parameter estimation compared to the uniform mixture based model of Tancredi et al. (2006) and mixture of hybrid-Pareto of Carreau and Bengio (2009). As with the other mixture models, the proposed can automatically be applied to multiple data sets with no prior threshold choice and the threshold uncertainty is fully accounted for as part of the inference process. Section 8 also provides new insights into the complex uncertainties induced in the tail estimates due to the threshold selection.

In the next section we introduce the proposed mixture model, with Sections 4.2 and 5 providing the details of the Bayesian inference using MCMC methods for posterior sampling of the model parameters, including details of the prior distributions. Sections 6.2 and 6.1 assess the performance and features of the approach using simulated data sets. In Section 8, we consider application of the proposed model for physiological measurements of pre-term infants in the neonatal intensive care unit of Christchurch Women's Hospital, New Zealand to aid characterisation of their medical status. We conclude in Section 9 with a discussion of our findings and potential for further research.

4. Proposed Mixture Model

This section details the proposed extreme value mixture model simultaneously describing bulk of the distribution and the tail, encapsulating the threshold as a parameter thus bypassing the issues associated with threshold selection. The observations below the threshold are assumed to follow a non-parametric density $h(\cdot|\eta, \mathbf{X})$, which is dependent on not only the associated parameter η but also the observation vector $\mathbf{X}$. The upper tail (excesses above the threshold) are assumed to follow a GPD(σ_u, ξ) or, equivalently, the PP representation outlined above. The non-parametric and GPD components are assumed to provide a reasonable approximation to the distribution of the data generating process.

Suppose the data comprises of a sequence of n independent observations $\{x_1, ..., x_n\}$ with distribution function F defined by

$$F(x|\eta, \xi, \sigma_u, u, \mathbf{X}) = \begin{cases} (1-\phi_u)\dfrac{H(x|\eta, \mathbf{X})}{\int_{-\infty}^{u} h(x|\eta, \mathbf{X})} & x \leq u \\ \\ (1-\phi_u) + \phi_u G(x|\xi, \sigma_u, u) & x > u \end{cases} \tag{5}$$

where $\phi_u G(\cdot|\xi, \sigma_u, u)$ is the unconditional GPD function given by (2) or equivalently the PP representation with intensity function defined by (3). The probability of being above the threshold ϕ_u, is used to scale the relative contributions represented by the kernel and GPD/PP components, is estimated using the proportion of data points above the threshold. It is possible that there is a discontinuity in the density at the threshold, although the distribution function will be continuous. However, as Bayesian inference with MCMC is utilised below with posterior predictive density estimation, which integrates over the entire posterior, in practice a smooth density estimate (around the threshold) is obtained.

It is possible to express the above model as a pure mixture model

$$f(x) = \pi f_1(x) + (1-\pi) f_2(x)$$

where $\pi = (1 - \phi_u)$ and

$$\begin{aligned} f_1(x) &= \frac{h(x|\eta, \mathbf{X})}{\int_{-\infty}^{u} h(x|\eta, \mathbf{X})\ dx}\ I_{(-\infty,u)}(x) \\ f_2(x) &= g(x|\xi, \sigma_u, u)\ I_{[u,\infty)}(x). \end{aligned}$$

The expectation-maximisation (EM) algorithm is a commonly used likelihood inference approach, using latent variables for component allocation, for pure mixture models due to Meng and van Dyk (1997). However, we cannot make full benefit of the efficiency of the EM algorithm as all the components share common a parameter (u), so the information contained in the data cannot be separated into contributions to each component. Bayesian inference with MCMC will be used instead.

4.1. Kernel Density Model

The univariate Parzen-Rosenblatt kernel estimator for $f(x)$, an unknown true density function, is defined by

$$\hat{f}(x;\lambda) = \frac{1}{n\lambda}\sum_{i=1}^{n} K\left(\frac{x-x_i}{\lambda}\right) \tag{6}$$

where $f(x)$ is defined on $\mathbb{R}$, $\lambda > 0$ is a smoothing parameter and $K(x)$ is a kernel function that usually satisfies the conditions

$$K(x) \geq 0 \text{ and } \int K(x)\ dx = 1$$

The kernel is often defined (Wand and Jones, 1995) using the scale notation $K_\lambda(y) = \lambda^{-1}K(y/\lambda)$ giving:

$$\hat{f}(x;\lambda) = n^{-1}\sum_{i=1}^{n} K_\lambda(x-x_i)$$

The latter notation is used throughout the rest of the article. Typically, K is chosen to be a unimodal probability density function that is symmetric about zero, thus ensuring that $\hat{f}(x;\lambda)$ is a valid density. One can think of the kernel as spreading a "probability mass" of size $1/n$ associated with each data point about its neighbourhood (Wand and Jones, 1995). It is well known that the kernel function used in equation (6) is generally not critical as the tail behaviour associated with the chosen kernel will be diminished by the averaging. Further, in the proposed mixture the GPD (or PP equivalent) is used for the upper tail, so extrapolation of the kernel into the tails is of no concern.

Traditional smooth kernel density estimators are not consistent near the boundary points of a density being estimated. The bias of a kernel is of the order $\mathcal{O}(h)$ at boundary points, compared with bias of the order $\mathcal{O}(h^2)$ at interior points. Jones (1993) and Silverman (1986) note that it is insufficient to simply truncate the density $\hat{f}(x;\lambda)$ at the boundary points and re-normalise. A variety of methods have been developed in the literature to remove these boundary effects, see Jones (1993) for example. For the context of this paper, we have not considered including boundary corrections within the kernel density for situations where there are apparent finite end points to the underlying process being modelled. Extensions of the model to deal with boundaries is considered in follow-up research work. However, in Section 7 we consider a practical alternative to boundary correction which works well in practice, by simply extending the mixture model to have a GPD for both the upper and lower tails.

4.2. Likelihood

The likelihood for the extreme value mixture model in equation (5) with the PP representation for the GPD can be written as:

$$L(\theta|\mathbf{X}) = (1-\phi_u)\prod_{A}\frac{1}{(n-1)}\sum_{\substack{i=1\\ i\neq j}}^{n} K_\lambda(x_j-x_i)\Big/\left[1-\sum_{i=1}^{n}\Phi\left(\frac{u-x_i}{\lambda}\right)\right] \times$$

$$\prod_{B}\exp\left\{-n_b\left[1+\xi\left(\frac{u-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}\prod_{i=1}^{n}\frac{1}{\sigma}\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-1-1/\xi} \qquad \xi \neq 0$$

where $\theta = (\lambda, u, \mu, \sigma, \xi)$, $A = \{j : x_j \leq u\}$ and $B = \{j : x_j > u\}$. This likelihood demonstrates the aforementioned issue as to why the full benefit of the EM algorithm cannot be realised, as conditional on the the latent component variable the contribution of the data to the parameters cannot be seperated into the two components. The PP likelihood is straightforward, however the details of the kernel density component are shown in Section 4.2.1.

4.2.1. Likelihood function for λ

The choice of λ typically plays a more crucial role in the accuracy of the kernel estimate than the choice of kernel $K(\cdot)$. Similar to threshold selection, the choice of λ involves a trade-off between smoothness and bias of the density estimate. Various methods have been proposed for global bandwidth selection in univariate density estimation, from minimising some model fit criterion (Jones et al., 1996), to the use of Bayesian techniques which have also been introduced for both multivariate density estimation, see Zhang et al. (2006) and univariate density estimation (Brewer (1998) and Brewer (2000)).

Likelihood inference for the smoothing parameter λ was first proposed by Habbema et al. (1974) and Duin (1976), who show that the likelihood is unbounded as $\lambda \to 0$, as each sum term in the product of:

$$\prod_{j=1}^{n} n^{-1} \sum_{i=1}^{n} K_\lambda(x_j - x_i)$$

is infinite in the limit $\lambda \to 0$ because the term $(x_j - x_i)$ becomes zero when $i = j$ (Duin, 1976). To avoid this degeneracy they replaced the likelihood function with the cross validation likelihood

$$\prod_{j=1}^{n} \frac{1}{(n-1)} \sum_{\substack{i=1 \\ i \neq j}}^{n} K_\lambda(x_j - x_i), \tag{7}$$

which can be viewed as minimising an estimate of Kullback-Leibler distance, see Bowman (1980) and Bowman (1984) for details.

Habbema et al. (1974) and Duin (1976) showed that the cross-validation likelihood based density estimators work well for short tailed distributions. However, they drastically over smooth heavy-tailed distributions. Schuster and Gregory (1981) showed this problem is due to inconsistency of the ML estimates. For kernels with both finite support i.e. $[-1, 1]$ and left continuous kernels of bounded variation on $(-\infty, \infty)$ that the ML estimates are inconsistent for a wide class of population densities, including the Cauchy. Schuster and Gregory (1981) observed that the smoothing parameter $\lambda = \lambda^*$ which maximises the likelihood in equation (7) for each x_i, must satisfy $|x_i - x_j| \leq \lambda^*$ for some x_j with $j \neq i$. Denoting the order statistics $x_{1n}, x_{2n}, \ldots, x_{nn}$ and when considering the upper tail of the distribution they showed that $|x_{nn} - x_{n-1\,n}| \not\to 0$ as $n \to \infty$, and hence $\lambda^* \not\to 0$. This property leads to inconsistent likelihood estimators for heavy-tailed distributions like the Cauchy (where the distance between the upper order statistics does not decay to zero). Bowman (1984) and Scott and Factor (1981) demonstrated that the cross validation likelihood based inference will tend to give smoothing parameters which are far too large (leading to over-smoothing) for not only heavy tailed distributions, but also in situations where outliers are present.

Within the proposed extreme value mixture model the upper tail is captured by the GPD component, so the inconsistency of the ML estimator using the cross-validation likelihood for heavy upper tails is overcome using our mixture model. If the lower tail is heavy then a simple extension of our approach to allow both the upper and lower tails to be captured using GPD/PP models would also resolve the inconsistency as will be demonstrated in Section 7.

5. Bayesian Inference

Computation for the proposed extreme value model is achieved via MCMC methods. Following the approach illustrated in Behrens et al. (2004), a Metropolis-Hastings sampler is used within a blockwise algorithm. If $\xi < 0$ then there is a finite upper bound on the range of support, so values of the PP parameters which provide a finite upper bound below the maximum of the observed data are invalid. The likelihood function has been defined to explicitly encompass these restrictions. MCMC samplers requires specification of proposal distributions for the parameters. Proposal distributions have been selected to reflect the restrictions to the sample space evident for each of the parameters within the model (which are not covered within the likelihood). The full posterior simulation algorithm is given in Appendix A.

5.1. Prior Specification

One of the benefits of the Bayesian inference approach is that expert prior information can be incorporated, thus allowing fuller account for the uncertainties in the parameters. We will now describe the prior distributions for the parameter set $\theta = (\lambda, u, \mu, \sigma, \xi)$. The joint prior distribution, under the reasonable assumption that the PP parameters are independent of all the other parameters, is expressed as

$$\pi(\lambda, u, \mu, \sigma, \xi) = \pi(\lambda) \cdot \pi(u) \cdot \pi(\mu, \sigma, \xi).$$

The following subsections specify the prior distributions for these three components.

5.1.1. Prior for PP parameters

Coles and Powell (1996) and Coles and Tawn (1996) advocate specification of the priors for extreme value model parameters in terms of extreme quantiles of the underlying process rather than the parameters themselves. They correctly argue that elicitation of expert prior information is easier for quantiles rather than parameters themselves, as the quantiles are a more intuitive quantity for most subject matter experts. Coles and Tawn (1996) construct the prior for the block maxima (generalised extreme value) model. Section 1 showed that by varying n_b and using the transformation given by (4) that it is possible to translate between the parameters for the GPD and block maximum GEV approach. With this in mind we can use the prior elicitation of Coles and Tawn (1996).

The $1-p$ quantile for the GEV distribution can be obtained by inversion of the GEV distribution function (see Coles (2001)) giving:

$$q_p = \mu + \sigma[\{-\log(1-p)\}^{-\xi} - 1]/\xi$$

where q_p is termed the return level associated with a return period of $1/p$ blocks (i.e. the level exceeding once on average every $1/p$ blocks). We can also see that by working with the block maxima representation the parameters are not dependent on the threshold, thus justifying the independence assumption in joint prior distribution.

Coles and Tawn (1996) elicit prior information in terms of the quantiles (q_{p_1}, q_{p_2} and q_{p_3}) for specified upper tail probabilities $p_1 > p_2 > p_3$. As there is a natural ordering to the q_i for $i = 1, 2, 3$, specification of independent priors for the 3 different quantiles would not be valid. Priors are therefore adopted for the quantile differences $(\tilde{q}_1, \tilde{q}_2, \tilde{q}_3)$ such that $\tilde{q}_i = q_{p_i} - q_{p_{i-1}}$ for $i = 1, 2, 3$, where $q_{p_0} = e_1$ is the physical lower end point for the process variable. Naturally in many applications $e_1 = 0$, although we don't make this assumption. Coles and Tawn (1996) suggest marginal priors for these quantities of the form

$$\tilde{q}_i \sim \text{Gamma}(\alpha_i, \beta_i) \quad i = 1, 2, 3$$

The choice of upper tail probabilities is usually not critical, common values for the probabilities are $p_1 = 0.1, p_2 = 0.01$ and $p_3 = 0.001$. The gamma parameters (α_i, β_i) for $i = 1, 2, 3$ are chosen to adhere to an experts belief for specified quantiles for each of the $\tilde{q}_i$. In the case of Coles and Tawn (1996) the median and 90% quantile were used to help determine the variability and location of prior belief.

From this prior specification the differences $(\tilde{q}_2, \tilde{q}_3)$ depend only on the shape and scale parameters (ξ, σ), with prior information on the location μ arising only through $\tilde{q}_1$. The prior is then constructed based on the three independent gamma distributions

$$\begin{aligned} \tilde{q}_1 &= q_{p_1} - e_1 \sim \text{Gamma}(\alpha_1, \beta_1) \\ \tilde{q}_2 &= q_{p_2} - q_{p_1} \sim \text{Gamma}(\alpha_2, \beta_2) \\ \tilde{q}_3 &= q_{p_3} - q_{p_2} \sim \text{Gamma}(\alpha_3, \beta_3) \end{aligned}$$

with the marginal prior distribution for (μ, σ, ξ)

$$\pi(\mu, \sigma, \xi) \propto \mathrm{J} \prod_{i=1}^{3} \tilde{q}_{p_i}^{\alpha_i - 1} \exp\{-\tilde{q}_{p_i}/\beta_i\}$$

with the Jacobian J of the transformation from $(q_{p_1}, q_{p_2}, q_{p_3}) \to (\mu, \sigma, \xi)$ given by

$$J = \left| \frac{\sigma}{\xi^2} \left[-(x_1 x_2)^{-\xi}(\log(x_2) - \log(x_1)) + (x_1 x_3)^{-\xi}(\log(x_3) - \log(x_1)) \right. \right. \\ \left. \left. -(x_2 x_3)^{-\xi}(\log(x_3) - \log(x_2)) \right] \right|$$

where $x_i = -\log(1 - p_i)$ for $i = 1, 2, 3$ and $\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2$ and β_3 are the hyperparameters, potentially based on expert knowledge of the underlying process.

An alternative commonly used prior specification for the PP model parameters is based on a trivariate normal distribution, often with a naive implementation using independent normals. Coles and Powell (1996) discuss this method for spatial modelling of extreme wind speeds, and gave precise values for η (location of MVN) and Σ (covariance structure of MVN). For our simulations, as we wish to provide limited prior information in the following simulation study a trivariate normal distribution, with independent components, is used in Section 6.

5.1.2. Prior for the threshold (u)

The prior for the threshold u, following Behrens et al. (2004), is assumed to follow a truncated normal distribution with parameters (μ_u, ν_u^2), truncated below at e_1 with density

$$\pi(u|\mu_u, \nu_u^2, e_1) = \frac{1}{\sqrt{2\pi\nu_u^2}} \frac{\exp\{-0.5[(u-\mu_u)/\nu_u]^2\}}{\Phi[-(e_1-\mu_u)/\nu_u]}$$

where μ_u is set at some high data percentile. Behrens et al. (2004) show that this prior can be parameterised in many forms, including continuous or discrete uniform prior distributions. We have set ν_u^2 to be sufficiently large to represent a very diffuse prior, to represent lack of knowledge of u.

5.1.3. Prior for Kernel Density parameter (λ)

Brewer (1998), Brewer (2000) and Zhang et al. (2006) consider Bayesian inference for the bandwidth parameter of kernel density estimators. Brewer (2000) consider local varying bandwidths and Zhang et al. (2006) consider bandwidth selection for multivariate kernel density estimation. It is straightforward to extend our kernel density estimator to have a local varying bandwidth estimators, however, as our interest is predominantly on tail quantities we consider only a single global bandwidth parameter.

We follow the prior definitions detailed in Brewer (2000) and Brewer (1998), for the case of a global bandwidth. Instead of specifying a prior for the bandwidth λ we specify the prior for the precision $1/\lambda^2$ as an inverse gamma,

$$\pi(\lambda|d_1, d_2) = \frac{1}{d_2^{d_1}\Gamma(d_1)} \left(\frac{1}{\lambda^2}\right)^{d_1-1} \exp\left(-\frac{1}{\lambda^2 d_2}\right)$$

where d_1 and d_2 are the hyperparameters. Care needs to be given when specifying (d_1, d_2) in cases where the likelihood of $\lambda < 0.50$ is high. This is due to the inverse gamma equalling 0 for most parameter sets when $d_1 \geq 1$ and $\lambda < 0.50$.

5.2. Prediction and Interval Estimation

Prediction of tail quantities is often of primarily concern in extreme value analysis. Prediction can be achieved within a Bayesian framework via the posterior predictive distribution. The predictive distribution allows uncertainty in the parameters to be accounted for (integrated out) as follows:

$$g(y|x) = \int_\Theta f(y|\theta)\pi(\theta|x)d\theta$$

for observations $X_1, ..., X_n$ where θ is the parameter vector of interest. Using Monte Carlo integration,

$$g(y|x) \propto \frac{1}{s}\sum_{i=1}^{s} f(y|\theta_i)$$

where $\theta_1, ..., \theta_s$ are observed realisations of the stationary distribution $\pi(\theta|x)$ from the MCMC chain. Therefore, the predictive distribution is obtained by averaging over the samples generated by the Markov Chain. Results in Sections 7 and 8 use of the predictive distribution.

While posterior uncertainty can be captured within the density described above, we can also summarise posterior uncertainty by using the highest posterior density (HPD) region (a form of credible interval). The HPD region corresponds to the range of parameter values that contain $100(1-\alpha)\%$ of the posterior, with the highest posterior density. All credible intervals given within this paper are HPD intervals unless specified otherwise.

6. Simulation Study

The simulation study to demonstrate the performance of the model and estimation procedure, is broken down into two parts. Firstly, we consider how well the mixture model approximates standard parametric distributions with varying upper and lower tail behaviours, which have easily derivable high quantiles which can be used to assess performance in tail estimation. Secondly, we check how the performance of the estimation procedure when the mixture model is, in some sense, the right model. The second component of the simulation study considers a range of parametric models for the bulk of the distribution, spliced together with three exemplar tail behaviours above some threshold. The principle is that the non-parametric density estimator will approximate the bulk of the distribution, with the PP/GPD approximating the upper tail. We ran our MCMC algorithm on a parallel Linux system, with 64-bit AMD Opteron 1.8GHz processor with and 16GB RAM. The required CPU time is around 90 minutes for a sample size of 1000.

6.1. Application to standard parametric distributions

We have considered three standard parametric population distributions which cover a range of possible tail behaviours and skewness/symmetry of bulk distribution: namely the normal, Student-t (on 3 degrees of freedom) and negative Weibull. The first two are symmetric with the normal distribution having Gumbel type tails ($\xi = 0$) and Student-t has a Fréchet type tails ($\xi > 0$). The negative Weibull is chosen as a skewed example, with Weibull type upper tail ($\xi < 0$). As noted above, the kernel density bandwidth estimator is inconsistent for heavy tailed distributions, so the models in this initial simulation study do not consider these types of model. Instead, the Cauchy distribution is considered as an example in Section 7.

Various parameter sets for the bulk distributions were considered with the results for the Weibull($\lambda = 10, k = 5$), normal($\mu = 0, \sigma = 3$) and Student-t($\nu = 3$) shown for brevity below as they demonstrate the performance of the approach. These parametric forms have a single mode, however the flexible non-parametric density estimator in the mixture model can of course cope with a smooth multi-modal population below the threshold. Note however, that we have deliberately chosen the Weibull parameters so the density is negligible near the lower boundary of the range of support at zero, to avoid the need for boundary corrections for the kernel density estimates, as discussed above.

Performance in the simulations is assessed by considering whether the known asymptotic tail behaviour of these three distributions has been effectively captured by the mixture model, using coverage rates for the HPD credible intervals from each simulated data set. The limiting shape parameter for Student-$t(\nu)$ is $\xi = \frac{1}{\nu}$. For negative-Weibull(l, k) the shape parameter is $\xi = -\frac{1}{k}$, see Beirlant et al. (2004) for details. Note that rate of the convergence of the normal tail to the Gumbel limit ($\xi = 0$) is extremely slow, so in the following results the performance of the estimates uses the sub-asymptotic value for ξ at the estimated threshold.

Table 1 reports the results of 100 replicates of sample size $n = 1000$ from the above population distributions. For every replication an MCMC algorithm as described above is run with 20,000 draws from the posterior distributions for the PP parameter vectors and 0.99 and 0.999 quantiles. We obtained the 95% HPD intervals after a burn-in of 5,000 draws. There is no true bandwidth λ to compare performance and as interest is focussed on tail estimation we consider performance for the shape parameter ξ of the mixture model. The coverage rate for a nominal 95% HPD interval, average length of HPD intervals and average posterior mean for the shape parameter ξ is shown in Table 1. As tail quantities are typically of interest, Table 1 also gives the same performance measure for the 0.99 and 0.999 quantiles. The true parameters/quantiles are also shown.

The coverage rates are well within expectations with 100 replicates, showing the mixture model is providing a reasonable approximation to the tail behaviour of the three population distributions. You will notice that the interval lengths for the shape parameter are very similar for all three population distributions. The average of the posterior means is close to the true values, particularly once the standard errors are taken into account. As we expect the quantiles themselves and the uncertainty associated with them (interval length and it's standard error) increase as the tail probability decrease.

Table 1: Coverage rate for shape parameter and 0.99/0.999 quantiles (for nominal 95% credible intervals) with true values given in [·]. Average posterior means and interval lengths given with standard error in parenthesis.

	Shape Parameter	Quantiles	
	ξ	$\hat{q}_{0.99}$	$\hat{q}_{0.999}$
NEGATIVE-WEIBULL$(l = 10, k = 5)$	[-0.20]	[-3.99]	[-2.51]
Coverage Rate	0.92	0.94	0.96
Interval Length	0.32 (0.044)	0.77 (0.10)	1.73 (0.49)
Average Posterior Mean	-0.22 (0.081)	-3.96 (0.18)	-2.51 (0.36)
STUDENT-t$(\nu = 3)$	[1/3]	[4.54]	[10.21]
Coverage Rate	0.90	0.93	0.92
Interval Length	0.43 (0.054)	1.78 (0.43)	10.55 (4.84)
Average Posterior Mean	0.26 (0.12)	4.72 (0.47)	10.46 (2.41)
NORMAL$(\mu = 0, \sigma = 3)$	[-0.12]	[6.68]	[9.27]
Coverage Rate	0.92	0.89	0.94
Interval Length	0.32 (0.039)	1.08 (0.15)	2.56 (0.70)
Average Posterior Mean	-0.18 (0.076)	7.11 (0.29)	9.24 (0.62)

6.2. Application to Models Spliced with Extremal Tails

The flexibility of the mixture model is now demonstrated by application to the same population distributions in Section 6.1 above spliced together with a GPD/PP upper tail above some threshold. In particular, these spliced distributions can also be used to evaluate the performance in estimating the threshold and the tail model (GPD/PP) parameters.

Denoting the bulk population density by $h^*(x)$) which are the same as considered above: negative Weibull, standard normal and Student-t (on 3 degrees of freedom) distribution. These bulk densities are spliced with examples of three different tail behaviours, with shape parameters $\xi = \{-0.2, 0, 0.4\}$. The threshold u is positioned at the 100×(1-p)% quantile of the bulk distribution and the PP scale parameter σ is chosen to ensure continuity at the threshold, as this is physically sensible in practice. Our sampling algorithm is therefore:

1. For a given p calculate u such that $\int_{-\infty}^{u} h^*(x)\ dx = p$
2. Generate $\mathbf{X} = \{x_1, \ldots, x_n\}$ from $h^*(x)$
3. Replace $\{\mathbf{X} : x_i > u \text{ for } i = 1, \ldots, n\}$ with generated points from the GPD.

As before, various parameter sets for the bulk distributions were considered with the results for the Weibull($\lambda = 10, k = 5$), normal($\mu = 0, \sigma = 3$) and Student-t($\nu = 3$) shown for brevity below as they demonstrate the performance of the approach.

The simulation results are presented in Tables 2 and 3 for 100 replicates of sample size $n = 1,000$ with upper tail probability at the threshold $p = 0.1$ (10% of distribution in the upper tail). Tables 2 and 3 report the coverage level (for a nominal 95% HPD interval), average length of HPD intervals and average posterior mean for the parameters and 0.99 and 0.999 quantiles, respectively. The true parameters and quantiles are also shown. For every replication an MCMC algorithm as described above is run with 20,000 draws from the posterior distributions for the parameter vectors and 0.99 and 0.999 quantiles. We obtained the 95% HPD intervals after a burn-in of 5,000 draws. The PP representation for the upper is used in the mixture model in the simulations, however for brevity the GPD equivalent of the σ_u parameter is shown.

Table 2: Coverage rates (for nominal 95% credible intervals), average credible interval length and posterior means for parameters.

	GPD Parameters								
ξ		$\hat{u}$			$\hat{\xi}$			$\hat{\sigma_u}$	
WEIBULL$(l = 10, k = 5)\mathbb{I}_{(0,u)} + 0.1\times$ ***GPD***$(u = 11.8, \sigma_u = 1.03, \xi)\mathbb{I}_{[u,\infty)}$									
-0.20	0.05	0.29	11.50	0.99	0.33	-0.19	0.98	0.64	1.11
0.00	0.09	0.31	11.51	0.97	0.37	-0.01	0.96	0.56	1.09
0.40	0.08	0.34	11.51	0.96	0.50	0.37	0.88	0.54	0.98
STUDENT-t$(\nu = 3)\mathbb{I}_{(-\infty,u)} + 0.1\times$ ***GPD***$(u = 1.63, \sigma_u = 0.98, \xi)\mathbb{I}_{[u,\infty)}$									
-0.20	0.03	0.29	1.33	0.90	0.33	-0.18	0.93	0.52	1.04
0.00	0.07	0.29	1.02	0.91	0.37	-0.002	0.93	0.52	1.35
0.40	0.10	0.30	1.35	0.99	0.49	0.39	0.87	0.52	0.94
NORMAL$(\mu = 0, \sigma = 3)\mathbb{I}_{(-\infty,u)} + 0.1\times$ ***GPD***$(u = 3.84, \sigma_u = 1.61, \xi)\mathbb{I}_{[u,\infty)}$									
-0.20	0.09	0.60	3.33	0.98	0.33	-0.19	0.97	0.93	1.81
0.00	0.10	0.59	3.34	0.96	0.37	0.01	0.93	0.88	1.73
0.40	0.15	0.61	3.36	0.97	0.49	0.40	0.88	0.88	1.60

In general, ξ is well estimated with coverage rates close to 0.95 (upto sampling variability). The average of the posterior means for the shape parameter are very close to the true value for all three bulk population models spliced with all three combinations

Table 3: Coverage rates (for nominal 95% credible intervals), average credible interval length and posterior means for 0.99/0.999 quantiles. True values for quantiles in square brackets.

	Quantiles					
ξ	$\hat{q}_{0.99}$			$\hat{q}_{0.999}$		
$\boldsymbol{WEIBULL}(l = 10, k = 5)\mathbb{I}_{(0,u)} + 0.1\times \boldsymbol{GPD}(u = 11.8, \sigma_u = 1.03, \xi)\mathbb{I}_{[u,\infty)}$						
-0.20	0.92	0.63	13.75 [12.48]	0.97	1.503	14.94 [14.90]
0.00	0.92	1.02	14.25 [14.18]	0.94	3.63	16.61 [16.54]
0.40	0.93	2.72	15.73 [15.69]	0.94	23.65	25.34 [25.44]
$\boldsymbol{STUDENT\text{-}t}(\nu = 3)\mathbb{I}_{(-\infty,u)} + 0.1\times \boldsymbol{GPD}(u = 1.63, \sigma_u = 0.98, \xi)\mathbb{I}_{[u,\infty)}$						
-0.20	0.95	0.61	3.45 [3.44]	0.94	1.50	4.64 [4.58]
0.00	0.94	0.98	3.96 [3.89]	0.93	3.48	6.28 [6.13]
0.40	0.95	2.59	5.48 [5.33]	0.91	21.24	15.43 [14.59]
$\boldsymbol{NORMAL}(\mu = 0, \sigma = 3)\mathbb{I}_{(-\infty,u)} + 0.1\times \boldsymbol{GPD}(u = 3.84, \sigma_u = 1.61, \xi)\mathbb{I}_{[u,\infty)}$						
-0.20	0.94	1.05	7.00 [7.00]	0.95	2.53	9.04 [8.99]
0.00	0.94	1.70	7.81 [7.78]	0.96	6.08	11.87 [11.72]
0.40	0.95	4.44	10.40 [10.31]	0.93	36.44	27.37 [26.54]

of tail behaviour. As we would expect, the average length of the HPD intervals for the shape parameter are larger for positive values compared to negative values of the shape parameter. The coverage rates for the other tail parameters are also good and well within the bounds due to sampling variability, with the only exception being for the populations with positive shape parameter ($\xi = 0.4$). The reason for the slightly lower than expected coverage is due to higher uncertainty in threshold parameter for positive shape parameter versus those with negative/zero shape, which will influence σ_u due to the dependence mentioned above.

The coverage rates for threshold estimation are very poor, however, this is to be expected. If the GPD (or PP equivalent) is an appropriate models for some threshold u it will be suitable for all higher thresholds $v \geq u$. Further, the standard graphical diagnostics traditionally used for threshold selection generally show a wide range of suitable thresholds, for which the GPD would provide a good fit to the tail. Notice that average posterior mean thresholds for all three bulk populations and tail models are very close to the true value, with consistent standard error (once standard deviation of population is accounted for). However, you will notice that the threshold tends to be biased slightly lower than the true value. It is believed that the threshold is estimated slightly lower than the truth as the kernel density can easily approximate the bulk density, but a slightly lower threshold will provide extra information for estimating the tail model parameters (without substantially impacting on the tail fit), which are intrinsically harder to estimate than the bulk model parameters due to the sparsity of tail data. Therefore, the tendency for a slightly lower estimated threshold is overall a satisfactory property of the proposed mixture model. In fact, when using the aforementioned graphical diagnostics for threshold choice, practitioners generally look for as small a threshold as possible (to maximise the sample tail information) whilst the tail model still provides a sufficiently good fit.

The coverage rates for the quantiles are well within expectations, with small bias in the 100 replications. Notice that the quantiles for distributions spliced with heavier tails (e.g. $\xi = 0.4$) have a higher standard error than those with shorter/lighter tails, which is expected due to the higher uncertainty for quantiles in heavier tailed distributions. While the 0.99 and 0.999 quantile results are given, many other quantiles were considered. Of particular note, are the coverage rates for the 0.9 and 0.95 quantiles which were around 50-60% and 80-90% respectively. We will see new insights in Section 8 that the threshold has a strong influence locally on the distribution function estimate. Hence, the threshold is sensitive to local sample fluctuations, which will reduce the coverage rates for the threshold and those distribution properties close to the threshold. The 90% quantiles are at the threshold, leading to the low coverage rate and the coverage rate quickly increasing as we go further away from threshold.

7. Consistency of Kernel Bandwidth Estimates

Section 4.2.1 outlined issues surrounding consistency of the kernel density bandwidth estimator for distributions exhibiting heavy tails, due to Schuster and Gregory (1981). This problem can be resolved by allowing both the upper and lower tails to be captured using GPD distributions. In particular, the model is defined as

$$F(x|\lambda, \boldsymbol{\xi}, \boldsymbol{\sigma_u}, \boldsymbol{u}, \mathbf{X}) = \begin{cases} \phi_1 G(-x|\xi_1, \sigma_{u1}, u_1) & x < u_1 \\ \phi_1 + (1-(\phi_1+\phi_2)))\dfrac{H(x|\lambda, \mathbf{X})}{\int_{u_1}^{u_2} h(x|\lambda, \mathbf{X})} & u_1 \leq x \leq u_2 \\ (1-\phi_2) + \phi_2 G(x|\xi_2, \sigma_{u2}, u_2) & x > u_2 \end{cases} \quad (8)$$

where $\boldsymbol{\xi} = (\xi_1, \xi_2)$, $\boldsymbol{\sigma_u} = (\sigma_{u1}, \sigma_{u2})$, $\boldsymbol{u} = (u_1, u_2)$, $\phi_1\text{GPD}(- \cdot |\xi_1, \sigma_{u1}, u_1)$ is the unconditional GPD function for x_i's $< u_1$, and $\phi_2\text{GPD}(\cdot|\xi_2, \sigma_{u2}, u_2)$ is the unconditional GPD function for x_i's $> u_2$. Inference for this model, can follow exactly the same methods outlined in Section 5 by using the point process representation for the GPD's outlined above.

Schuster and Gregory (1981) illustrated the consistency problem with the cross-validation maximum likelihood method for kernel density estimation with a pseudo-random sample of size 100 from a standard Cauchy distribution. The above two-tailed model was applied using Bayesian inference for 20,000 iterations with a burn-in of 5000 on random sample of Cauchy(0,1) variables of length 500. Prior distributions for the two sets of point process parameters were set to very diffuse trivariate normal distributions with independent margins given by

$$\pi(\xi_1, \sigma_{u1}, \mu_1) = \pi(\xi_2, \sigma_{u2}, \mu_2) = MVN\left(\mu = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 100 & 0 & 0 \\ 0 & 100 & 0 \\ 0 & 0 & 100 \end{bmatrix}\right)$$

The Cauchy(0,1) distribution is a special case of the Student-t when $\nu = 1$, so the asymptotic tail behaviour has $\xi_1 = \xi_2 = 1$.

Figure 1 and Table 4 give the results from running inference on the generated sample. For comparison, the model with the kernel density estimator only, and with kernel density estimator spliced with the PP/GPD upper tail were also considered. A single sample is shown here, but similar results were achieved across samples considered.

Table 4: Results for Cauchy(0,1) with standard errors given in parenthesis.

Model	Mixture Model Parameters			
	λ	u	ξ	σ_u
Kernel	12.41 (0.42)	-	-	-
GPD + Kernel	13.47 (0.50)	2.45 (0.34)	1.11 (0.30)	2.09 (0.78)
GPD_1 + Kernel $+\text{GPD}_2$	0.41 (0.08)	1: -2.18 (0.23)	0.95 (0.26)	2.30 (0.90)
		2: 2.44 (0.30)	1.11 (0.32)	2.07 (0.84)

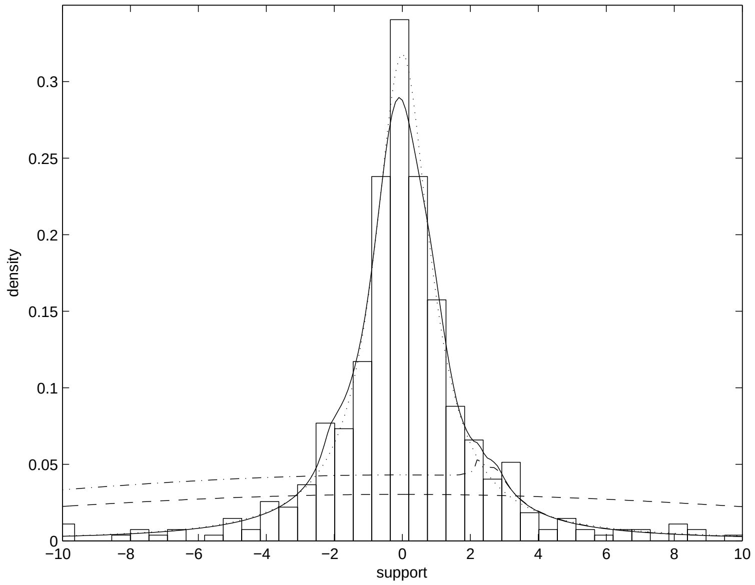


Figure 1: Posterior predictive density estimator for Cauchy(0,1) using various models; kernel density only (- - -); one-tailed mixture model (— · —); two-tailed mixture model (—) and (true) Cauchy(0,1) pdf (· · ·).

Of particular importance in the results, is how modelling either both tails, or the upper tail or neither tail model has effected the resulting kernel bandwidth. As expected when fitting only the kernel density, due to the presence of heavy tails, the bandwidth has been biased too large leading to the over-smoothing of the density estimator as seen in Figure 1. This over-smoothing is also evident in the case where the the upper tail is used. This demonstrates that the heavy lower behaviour can have an strong influence on the bulk distribution estimate, and potentially for low quantiles below/around the threshold. However, you will notice that the upper tail model is still managing to provide a reasonable fit, and provide a very similar upper tail fit to the two-tail mixture model. In particular, the one-tailed upper tail parameters (u, ξ, σ_u) are very similar to those for the upper tail for the two -tailed mixture model $(u_2, \xi_2, \sigma_{u2})$ in Table 4.

It is important to note that the two-tailed mixture model provides a very good fit to both the bulk distribution (main mode), shown by closeness of dotted and solid lines, and a good fit in both tails. Further, the shape parameter estimates for both the upper and lower tails are close to 1, particularly once the standard error has been accounted for. By including both lower and upper tail flexibility we have successfully overcome the inconsistency in the bandwidth estimation for the kernel density estimator.

Many applications often occur in finance where modelling excesses for both tails of a given process are of importance. For example, for simultaneuously modelling the risk associated high returns as well as low returns, and fully accounting for their associated uncertainties. The two-tail model of (8) could be useful in these situations, overcoming the issue of dual threshold estimation (and corresponding) uncertainty estimation in the traditional fixed threshold approach as in McNeil and Frey (2000). It is also common in financial applications to consider asymmetry of the profit/loss profile, evidence for which could be examined by comparing the two-tail model with the same or different tail shape parameters. Thus the two-tail model could also provide a flexible framework for applications where both tails are of interest.

8. Application

Babies born prematurely are vulnerable to tissue and organ injury as a result of immature physiological adaption to extrauterine life. Clinicians take various physiological measurements from premature babies in neonatal intensive care units (NICU's), which are monitored for clinical care. These include oxygenation saturation, pulse rates and respiration rates. The challenged faced by clinicians is the assessment of variation in these measurements, caused by cardio-respiratory instabilities, to determine whether the baby is "premature and stable", "premature and unstable" or "premature and unwell". There are deficiencies in our knowledge of "normal ranges" for these measurements. It is hypothesised that the current normal ranges used in practice need to be refined. The principal goal of this research is to contribute to the refinement of our understanding of "normal ranges" for these high frequency physiological measurements from pre-term babies, which essentially requires reliable estimation of suitably high quantiles (e.g. 95% or 99%).

The proposed one-tailed mixture model is applied to pulse rates from a pre-term baby (gestation age 34 weeks) who was considered stable at the time the study took place and who was not receiving supplementary oxygenation intervention treatment at the NICU at Christchurch Women's Hospital, New Zealand. The data is collected over roughly a 6 hour period at 0.5Hz (once every 2 seconds). Over this time period, the pre-term infant was in various states: including levels of awakeness (awake and quiet, awake and crying, quiet sleep and active sleep), feeding by suckling and through a nasogastric tube feed and exhibited signs of both irregular and regular breathing patterns. Clearly, there will be temporal dependence in these high frequency measurement. We have randomly sub-sampled the data to roughly every 5 measurements, to reduce the dependence and therefore provide a more realistic assessment of the uncertainty associated with our estimates. The pre-term infants commonly exhibit various forms of non-stationary behaviour in both level and variability in time, as can be seen in Figure 2. In this paper will only consider the marginal distribution of the time series, with the non-stationarity to be considered in future work. For this application, we are interested in being able to estimate the lower tail quantiles of the pulse rates.

The MCMC Metropolis-Hastings sampler outlined in Section 5 was initialised at an arbitrary starting parameter vector and run for 25,000 iterations with a burn-in period

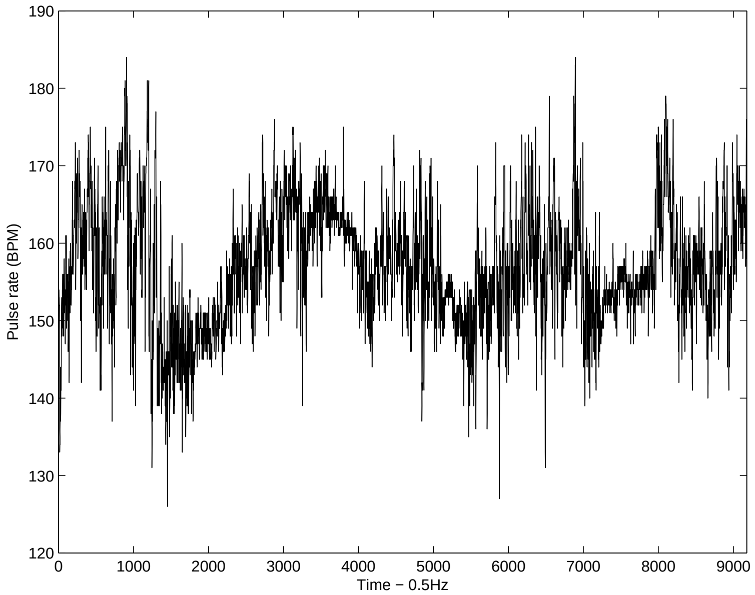


Figure 2: Time series of pulse rates for neonatal patient taken every two seconds, for approximately six hours.

of 5,000, giving 20,000 posterior draws for which subsequent analysis is based on. Convergence of the chains is assessed using the standard diagnostics discussed in Gelman and Rubin (1992). They suggest running multiple chains which are compared to ensure convergence. Starting points for these simulated chains should be dispersed over the sample space. This ensures that all major regions within the target distribution have been considered.

Tables 5 and 6 gives results for both the kernel mixture model and the traditional fixed threshold approach for a range of sensible thresholds. Figure 3 displays the mean residual life (MRL) plot, as discussed in Section 2. As we are interested in whether the GPD/PP model is a good fit to the lower tail, rather than looking for linearity from left to right of the x axis, we look for linearity from right to left. The principle with traditional threshold selection using the MRL is too find as high enough threshold to maximise the sample information in the lower tail, with the lower tail model still providing a good fit which is shown by linearity in the MRL plot if the GPD/PP is an appropriate model to capture the lower tail. A decline in the mean excess plot is seen above around 155 with evidence of a linear trend below this point. The increasing variability for low threshold values is evident due to the limited number of exceedances available out in the lower tail of the data.

Unlike Coles and Tawn (1996), elicitation of the prior structure for $\pi(\mu, \sigma, \xi)$ was not based on an experts knowledge of the process of pulse rates. Very diffuse priors were specified instead, as we desired the data to speak for themselves. The prior for the point process parameters were defined using the 90% quantile, the difference between the 99% and the 90% quantile and the difference between the 99.9% and 99% quantile, giving a prior consisting of three independent gammas with hyper-parameters:

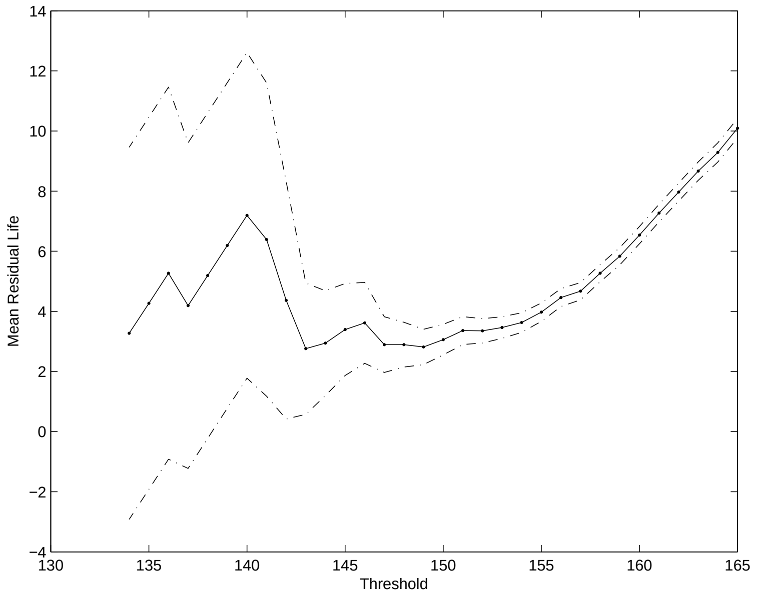


Figure 3: Mean residual life plot for sub-sampled pulse rate data.

- Gamma($\alpha_1 = 1.20, \beta_1 = 28$),
- Gamma($\alpha_2 = 1.20, \beta_2 = 5$) and
- Gamma($\alpha_3 = 1.20, \beta_3 = 10$).

The prior for the threshold was truncated at the minima of the data, centered about the 80% quantile with a standard deviation of 10 and the prior based on the precision of the bandwidth was specified as an inverse Gamma(1,0.25).

Figure 4 gives a comparison of the prior and posterior marginal distributions for each of the parameters within the proposed mixture model. The key thing to notice is that marginal distributions for the mixture model parameters are all very diffuse. It is also evident from Figure 4 that the priors are not carrying any undue influence on the MCMC chain for any of the parameters in the mixture model, shown by the stark differences between the prior and posterior distributions.

Table 5: Posterior means of the mixture model parameters for the pulse rate data.

	Prior			
	Quantile		*Location*	
$\hat{\lambda}$	1.48	(0.90, 2.15)	1.48	(0.87, 2.13)
$\hat{u}$	149.81	(149.07, 150.62)	149.73	(149.03, 150.53)
$\hat{\xi}$	0.049	(-0.106, 0.20)	0.040	(-0.105, 0.213)
$\hat{\sigma}_u$	2.96	(2.21, 3.78)	3.04	(2.25, 3.81)

The MCMC was also run with diffuse priors for the point process parameters based

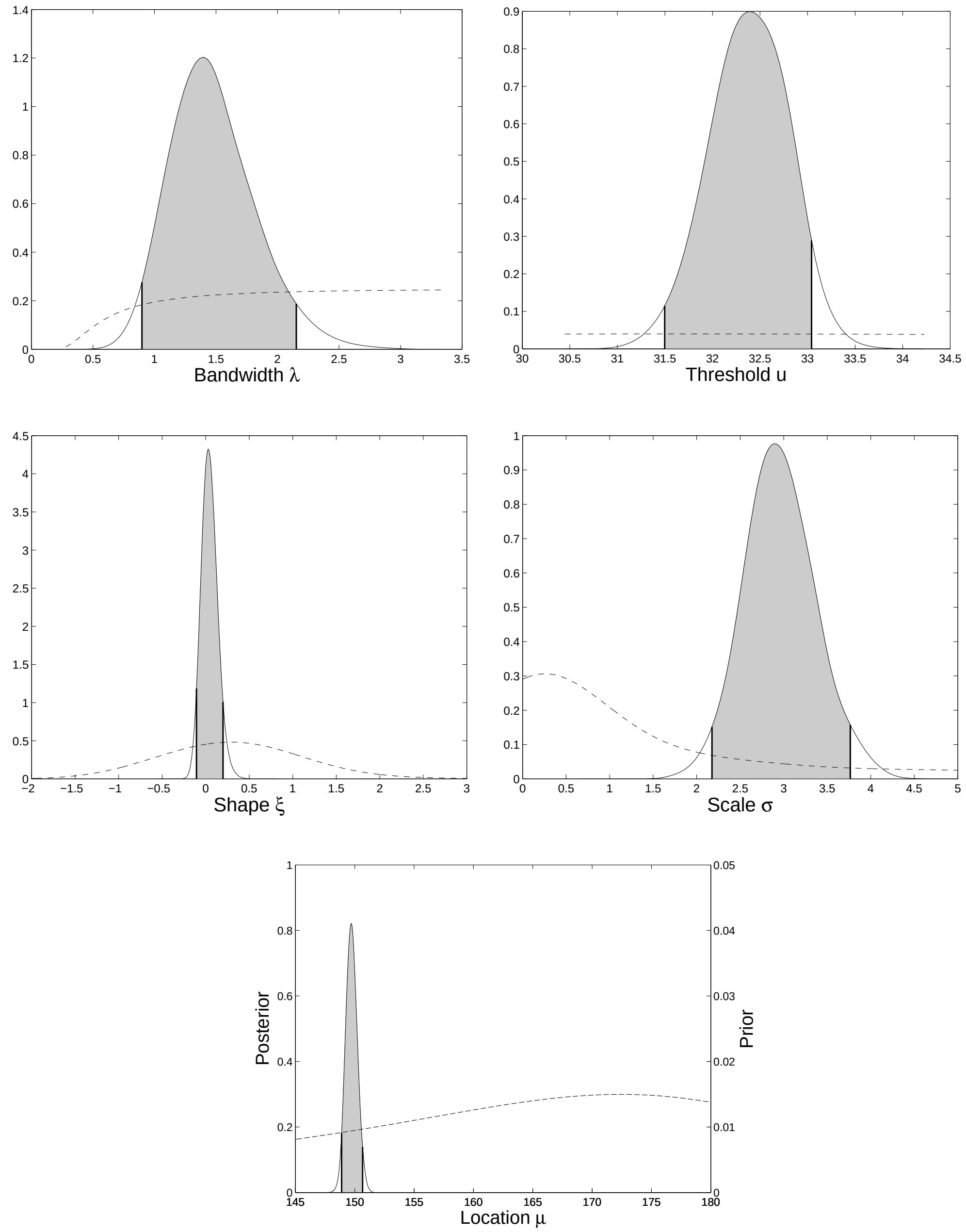


Figure 4: Marginal prior (—) and posterior (- - -) distributions for each parameter within the extreme mixture model. Notice for location parameter that the prior is so diffuse that it has been scaled (see left y-axis) to see the details.

trivariate independent normals as described in Section 5.1.1. The structure of the priors is the same as given in Section 7, with results given in Table 5 alongside those of the quantile difference based priors. The alternative priors were used to ensure the prior specification does not have an undue impact on the posterior distribution and as another diagnostic check for convergence of the MCMC chain. The similarity of the results from the two prior models in Table 5 suggest the Markov chains have successfully converged, and prior structure is not having an adverse effect on the resulting posterior.

Table 6: Posterior means of the GPD parameters for fixed threshold approach.

Fixed threshold	# of exceedances	GPD Parameters			
		Shape (ξ)		*Scale* (σ_u)	
$u = 155$	466	-0.096	(-0.150, -0.039)	4.368	(3.918, 4.824)
$u = 153$	310	-0.036	(-0.120, 0.048)	3.629	(3.131, 4.145)
$u = 149$	110	0.101	(-0.060, 0.253)	2.594	(1.933, 3.266)
$u = 147$	53	0.160	(-0.063, 0.381)	2.557	(1.620, 3.591)
$u = 145$	24	0.132	(-0.155, 0.429)	3.245	(1.611, 5.036)

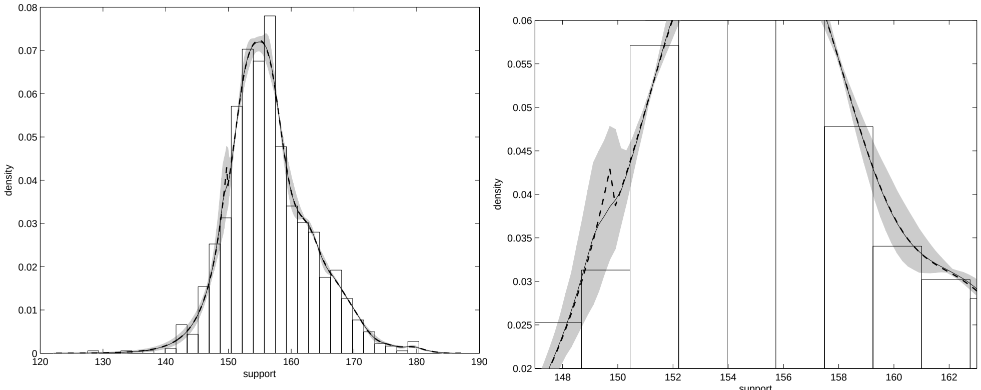
The posterior mean for the shape parameter along with 95% HPD interval for ξ in Table 5 indicates evidence of an exponential type lower tail. The interval length for the threshold u is relatively small in magnitude suggesting the threshold was relatively well defined for the pulse rate data. For comparison Table 6 gives results for running Bayesian inference for the fixed threshold approach, with the same prior specification of the point process parameters as given above. The thresholds considered were chosen based on the MRL plot given in Figure 3. Table 6 indicates one of the issues surrounding threshold selection. For a threshold of 155, inference is suggesting a negative shape parameter ($\xi = -0.10(-0.15, -0.04)$). Based on the MRL plot in Figure 3 a threshold of 155 is a reasonable choice. However, all other possible thresholds give HPD credible intervals include the possibility of zero or positive shape parameter, similar to that suggested by the mixture model estimates in Table 5.

Another useful insight is provided by comparing the interval length for the shape and scale parameters for the mixture model approach in Table 5 and fixed threshold approach in Table 6 for the threshold 149, which is close to that automatically selected for the mixture model. The interval length for the mixture model shape and scale parameters are larger than for the fixed threshold approach, representing the additional uncertainty due to the threshold choice. Thus providing the first insight into the impact of the threshold selection on the tail estimation.

Figure 5(a) shows two density estimates: the solid line is the posterior predictive density and the dashed line is obtained by plugging the point estimates of the posterior means into the density of the mixture model described by (5). The mixture model density using the point estimate is only included to demonstrate that the individual posterior density estimates can exhibit a discontinuity at the threshold, which is easily seen in Figure 5(b). However, the posterior predictive density is continuous at the threshold due to integrating over the whole posterior. The pointwise HPD 95% region for the posterior predictive density is also given in grey. These grey limits provide new insights into the uncertainty about the kernel density component and the tail model (due to threshold choice).

We expect the uncertainty to be highest when the density is at its highest and viceversa. Intuition suggests the uncertainty relative to density will be lowest near the mode (where there is the most data) with increasing relative uncertainty further out into the tails. This intuition is born out in Figure 5, with two key exceptions. Firstly, there is large relative uncertainty where the density is changing the most (i.e. steepest slope), as shown clearly in the width of the intervals on right in Figure 5(b). Secondly, the threshold uncertainty impacts on the tail quantile estimates (seen clearly in Figure 6 below) as we would expect, but it also has a substantial localised effect on the uncertainty on the distribution close to the threshold. The localised threshold uncertainty impacts are shown by the much larger grey intervals on the right in Figure 5(b). The localised effects will therefore have influence on quantiles which are close the threshold, as well as the tail extrapolation. This localised consequence of the threshold choice (as the threshold degree of freedom has predominantly local influence) to the authors knowledge has not been highlighted in previous extremal threshold (mixture) modelling approaches.

In typical extreme value applications, we are interested in describing the behaviour of extremal quantiles. Rather than interpreting these values based on parameter values as above it is often more appropriate to assess the model fit in the tails in terms of the quantiles or so called return levels. Denote z_p the return level associated with the return period $1/p$ (tail probability p), which can be interpreted as the level exceeded on average once every $1/p$ periods (where the time scale of the period is defined by the process being fitted). Estimates of extreme quantiles for threshold exceedances can be obtained by inverting equation (5). Figure 6 shows the return levels (quantiles) for a range of return periods. The return levels are plotted on a negative log-log scale which compresses the tail of the distribution to ensure that the tail extrapolation can be seen in detail. An exponential tail ($\xi = 0$) is shown by straight line under this transformation, with heavier tail than exponential ($\xi > 0$) as convex function and shorter tail ($\xi < 0$) shown by concave function. Return level plots can also be used as a model diagnostic to ensure model-based returns are in reasonable agreement with empirical estimates as seen in Figure 6. Table 7 gives return levels for $p = \{0.1, 0.01, 0.001, 0.0001\}$ for the models



(a) *Histogram with posterior predictive density.*

(b) *Zoom-in of* Figure 5 (a).

Figure 5: Sample density of pulse rates with posterior predictive density estimate (—). The estimated density obtained from plugging-in the posterior mean of the parameters is shown for comparison (- - -).

based on the fixed threshold approach.

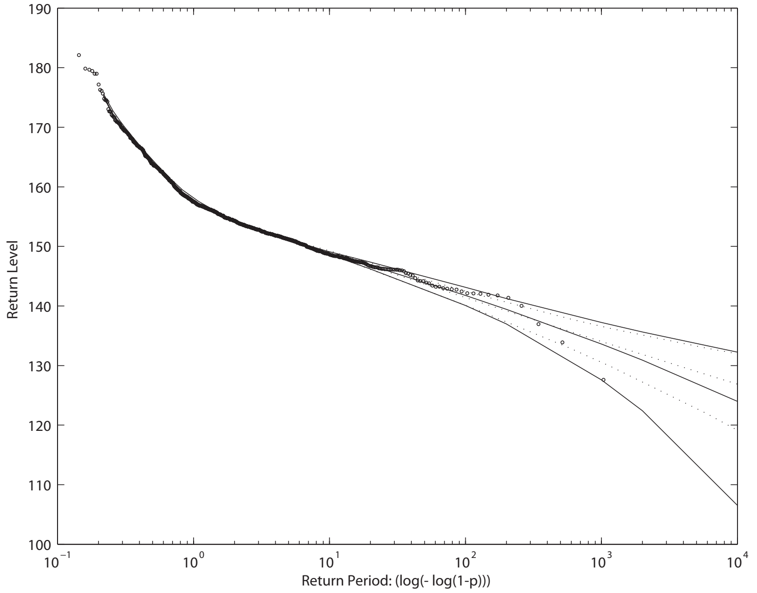


Figure 6: Posterior predictive return level plot for the pulse rate data; mixture model (—); fixed threshold $u = 150$ $(\cdots)$; with 95% HPD region for calculated returns.

Table 7: Return levels for fixed threshold approach fir range of thresholds with 95% HPD intervals given in parenthesis

Fixed Threshold	Return Level			
	10^1	10^2	10^3	10^4
$u = 155$	148.88	141.07	134.77	129.67
	(148.37, 149.39)	(139.81, 142.24)	(132.16, 137.05)	(125.29, 133.13)
$u = 153$	149.10	141.39	134.23	127.49
	(148.64, 149.55)	(140.00, 142.75)	(130.63, 137.30)	(120.39, 133.27)
$u = 149$	148.84	142.09	133.37	121.68
	(148.80, 148.88)	(140.63, 143.41)	(128.20, 137.72)	(106.26, 132.82)
$u = 147$	-	142.27	133.00	118.42
	-	(140.85, 143.62)	(127.26, 137.60)	(97.18, 132.76)
$u = 145$	-	142.14	132.39	117.51
	-	(140.78, 143.37)	(126.61, 137.48)	(94.78, 133.01)

Figure 6 gives the return level plot for the mixture model approach (solid line) and for the fixed threshold approach with $u = 150$ (dotted line) for comparison. The threshold was set to $u = 150$ for the fixed threshold approach as it was generally sensible and is the value chosen by the mixture model in Table 5, so will provide a useful comparison of the return level estimates and uncertainty associated with threshold choice. You will notice that the mixture model and fixed threshold GPD based return level functions are very similar indeed, only showing deviations for quantiles with tail probabilities less than 10^{-3}. The curvature of the returns levels is also suggesting $\xi = 0$, as seen in Table 5, though the HPD intervals include the possibility of positive/zero shape. The sample quantiles are within the pointwise intervals for most return periods, suggesting reasonable model fit after allowance for sampling variability, however there is room for improvement shown by the occasional blocks of sample quantiles outside the HPD intervals, which could clearly be due to possible nonstationary effects which will be considered in future research. In future, improvements to the accuracy of estimates at high return levels could also be achieved by the inclusion of prior knowledge of pulse rates.

Comparing the length of the HPD intervals for the return levels in Figure 6 and Table 7 to those for the fixed threshold approach, shows that the added uncertainty due to threshold selection has been encapsulated in the tail estimates using the mixture model. The extra uncertainty captured by the mixture model approach is particulary noticeable in Figure 6. Further, the extra uncertainty with mixture model estimates has lead to a higher coverage rate for the sample quantiles within the pointwise HPD intervals, this providing further confirmation of the need to account for the uncertainty due to threshold choice.

9. Conclusions

We have proposed a new extreme value mixture model combining a non-parametric density estimator for the bulk of the population distribution below some threshold, with a classical GPD tail model for the excesses above the threshold (or an equivalent point process representation). The mixture model has the benefit of avoiding the subjectivity of the commonly used graphical diagnostic for threshold choice, and permits the complex uncertainties associated with threshold estimation to be fully accounted for. The mixture model can also be automatically applied to multiple data sets, as it avoids user intervention in the threshold choice. Our model has the advantage of a flexible non-parametric component below the threshold avoiding the need to pre-specify a parametric form as in most previous proposed extremal mixture model approaches, and the simple kernel density estimator has just a single extra parameter to be estimated overcoming the computational complexity of other related mixture models.

We have also shown that the addition of upper and lower tail models can be used to overcome the problem in inconsistent kernel density bandwidth estimators for heavy tailed data, e.g. Cauchy distributed populations.

The take home message, clearly demonstrated in Figures 5 and 6, is that the uncertainty associated with threshold choice has a complex structure which not only impacts on the tail extrapolation but also strongly influences distribution estimates close to the threshold due to the inherent local influence of the threshold degree of freedom. It is clear that compared to the traditional fixed threshold approach that the extra uncertainty associated with threshold choice should be accounted for, and the mixture model presented herein appears to have successfully encapsulate this uncertainty.

A key development in ongoing research for the non-parametric component is considering alternative kernel density estimators which can cope with population distributions which have bounded support in the tail captured by the non-parametric component, to overcome the boundary effects experienced using the traditional symmetrical kernels.

10. Acknowledgements

Anna MacDonald received support from both a New Zealand TEC Top Achievers Doctoral scholarship under the Bright Futures Scheme and Canterbury scholarship from the University of Canterbury.

References

Behrens, C. N., Lopes, H. F., Gamerman, D., 2004. Bayesian analysis of extreme events with threshold estimation. Statistical Modelling 4 (3), 227–244.

Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J., 2004. Statistics of Extremes: Theory and Applications. Wiley: London.

Bowman, A. W., 1980. A note on consistency of the kernel method for the analysis of categorical data. Biometrika 67 (3), 682–684.

Bowman, A. W., 1984. An alternative method of cross-validation for the smoothing of density estimates. Biometrika 71 (2), 353–360.

Brewer, M. J., 1998. A modelling approach for bandwidth selection in kernel density estimation. In: Payne, R., Green, P. (Eds.), Proceedings of COMPSTAT 1998. Physica Verlag: Hiedelberg, pp. 203–208.

Brewer, M. J., 2000. A Bayesian model for local smoothing in kernel density estimation. Statistics and Computing 10 (4), 299–309.

Carreau, J., Bengio, Y., 2009. A hybrid Pareto model for asymmetric fat-tailed data: the univariate case. Extremes 12 (1), 53–76.

Castillo, E., Hadi, A., Balakrishnan, N., Sarabia, J., 2004. Extreme Value and Related Models with Applications in Engineering and Science. Wiley.

Coles, S., 2001. An introduction to statistical modeling of extreme values. Springer: London.

Coles, S. G., Powell, E. A., 1996. Bayesian methods in extreme value modelling: A review and new developments. International Statistical Review 64 (1), 119–136.

Coles, S. G., Tawn, J. A., 1996. A Bayesian analysis of extreme rainfall data. Applied Statistics 145 (4), 463–478.

Davison, A. C., Smith, R. L., 1990. Models for exceedances over high thresholds. J. R. Statist. Soc. **B** 52 (3), 393–442.

Duin, R. P. W., 1976. On the choice of smoothing parameters for Parzen estimators of probability density functions. I.E.E.E Transactions on Computers C - 25 (11), 1175–1179.

Dupuis, D. J., 2000. Exceedances over high thresholds: A guide to threshold selection. Extremes 1 (3), 251–261.

Embrechts, P., Klüppelberg, C., Mikosch, T., 2003. Modelling extremal events for insurance and finance. Springer: New York.

Frigessi, A., Haug, O., Rue, H., 2002. A dynamic mixture model for unsupervised tail estimation without threshold selection. Extremes 5 (3), 219–235.

Gelman, A., Rubin, D. B., 1992. Inference from iterative simulation using multiple sequences (with discussion). Stat. Sci. 7 (4), 457–511.

Habbema, J., Hermans, J., van den Broek, K., 1974. A stepwise discriminant analysis program using density estimation. In: Bruckmann, G. (Ed.), Proceedings of COMPSTAT 1974. Physica-Verlag: Vienna, pp. 101–110.

Jones, M., 1993. Simple boundary correction for kernel density estimation. Statistics and Computing 3 (3), 135–146.

Jones, M. C., Marron, J. S., Sheather, S. J., 1996. A brief survey of bandwidth selection for density estimation. J. Am. Statist. Assoc. 91 (433), 401–407.

McNeil, A., Frey, R., 2000. Estimation of tail-related risk measures for heteroscedastic financial time series an extreme value approach. Journal of Empirical Finance 7, 271–300.

Mendes, B., Lopes, H. F., 2004. Data driven estimates for mixtures. Computational Statistics and Data Analysis 47 (3), 583–598.

Meng, X., van Dyk, D., 1997. The EM algorithm - an old folk song sung to a fast new tune (with discussion). J. Roy. Stat. Soc. **B** 59 (3), 511–567.

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

Schuster, E. F., Gregory, C. G., 1981. On the nonconsistency of maximum likelihood nonparametric density estimators. In: Eddy, W. F. (Ed.), Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface. Springer-Verlag: New York, pp. 295–298.

Scott, D. W., Factor, L. E., 1981. Monte Carlo study of three data-based nonparametric probability density estimators. J. Am. Statist. Assoc. 76 (373), 9–15.

Silverman, B., 1986. Density estimation for statistics and data analysis. Chapman and Hall/CRC: London.

Tancredi, A., Anderson, C., O'Hagan, A., 2006. Accounting for threshold uncertainty in extreme value estimation. Extremes 9 (2), 87–106.

Wand, M. P., Jones, M. C., 1995. Kernel Smoothing. Chapman and Hall/CRC: London.

Zhang, X., King, M. L., Hyndman, R. J., 2006. A Bayesian approach to bandwidth selction for multivariate kernel density estimation. Computational Statistics and Data Analysis 50 (11), 3009–3031.

Appendix A.

The sampling algorithm for simulation from the posterior of $\theta = \{h, u, \mu, \sigma, \xi\}$ via a blockwise Metropolis-Hastings algorithm is now presented. The proposal variances $V = \{\mathrm{V}_h, \mathrm{V}_u, \mathrm{V}_\mu, \mathrm{V}_\sigma, \mathrm{V}_\xi\}$, are specified to ensure appropriate acceptance rates result for the marginal posteriors.

`Initialisation:` Choose an arbitrary starting value $\theta^{(0)} = \{h^{(0)}, u^{(0)}, \mu^{(0)}, \sigma^{(0)}, \xi^{(0)}\}$
`Iteration:` $j(j \geq 1)$

- $\xi^{(j)}$

  1. Given $\xi^{(j-1)}$, generate $\xi^* \sim \mathrm{N}(\xi^{(j-1)}, \mathrm{V}_\xi)$
  2. Compute

     $$\alpha_\xi = \min\left\{\frac{\pi(h^{(j-1)}, u^{(j-1)}, \mu^{(j-1)}, \sigma^{(j-1)}, \xi^*|\mathbf{X})}{\pi(h^{(j-1)}, u^{(j-1)}, \mu^{(j-1)}, \sigma^{(j-1)}, \xi^{(j-1)}|\mathbf{X})}, 1\right\}$$

     any constraints placed on ξ are included within the likelihood.
  3. With probability α_ξ, accept ξ^* and set $\xi^{(j)} = \xi^*$; otherwise reject ξ^* and set $\xi^{(j)} = \xi^{(j-1)}$.

- $\sigma^{(j)}$

  1. Given $\sigma^{(j-1)}$, generate $\sigma^* \sim \mathrm{LN}(\log(\sigma^{(j-1)}), \mathrm{V}_\sigma)$
  2. Compute

     $$\alpha_\sigma = \min\left\{\frac{\pi(h^{(j-1)}, u^{(j-1)}, \mu^{(j-1)}, \sigma^*, \xi^{(j)}|\mathbf{X})}{\pi(h^{(j-1)}, u^{(j-1)}, \mu^{(j-1)}, \sigma^{(j-1)}, \xi^{(j)}|\mathbf{X})} \frac{\mathrm{LN}(\sigma^{(j-1)}|\log(\sigma^*), V_\sigma))}{\mathrm{LN}(\sigma^*|\log(\sigma^{(j-1)}), V_\sigma))}, 1\right\}$$

     any constraints placed on σ are included within the likelihood.
  3. With probability α_σ, accept σ^* and set $\sigma^{(j)} = \sigma^*$; otherwise reject σ^* and set $\sigma^{(j)} = \sigma^{(j-1)}$.

- $\mu^{(j)}$

  1. Given $\mu^{(j-1)}$, generate $\mu^* \sim \mathrm{N}(\mu^{(j-1)}, \mathrm{V}_\mu)$

2. Compute

$$\alpha_\mu = \min\left\{\frac{\pi(h^{(j-1)}, u^{(j-1)}, \mu^*, \sigma^{(j)}, \xi^{(j)}|\mathbf{X})}{\pi(h^{(j-1)}, u^{(j-1)}, \mu^{(j-1)}, \sigma^{(j)}, \xi^{(j)}|\mathbf{X})}, 1\right\}$$

any constraints placed on μ are included within the likelihood.

3. With probability α_μ, accept μ^* and set $\mu^{(j)} = \mu^*$; otherwise reject μ^* and set $\mu^{(j)} = \mu^{(j-1)}$.

- $u^{(j)}$

1. Given $u^{(j-1)}$, generate $u^* \sim \mathrm{N}(u^{(j-1)}, \mathrm{V}_u)\mathbb{I}_{(m,M)}$, where $m = \min(x_1, ..., x_n)$ and $M = \max(x_1, ..., x_n)$
2. Compute

$$\alpha_u = \min\left\{\frac{\pi(h^{(j-1)}, u^*, \mu^{(j)}, \sigma^{(j)}, \xi^{(j)}|\mathbf{X})}{\pi(h^{(j-1)}, u^{(j-1)}, \mu^{(j)}, \sigma^{(j)}, \xi^{(j)}|\mathbf{X})} \cdot \right.$$
$$\left.\frac{(\Phi((M-u^*)/\sqrt{(V_u)}) - \Phi((m-u^*)/\sqrt{V_u}))}{(\Phi((M-u^{(j-1)})/\sqrt{V_u}) - \Phi((m-u^{(j-1)})/\sqrt{V_u}))}, 1\right\}$$

all other constraints placed on u are included within the likelihood.

3. With probability α_u, accept u^* and set $u^{(j)} = u^*$; otherwise reject u^* and set $u^{(j)} = u^{(j-1)}$.

- $h^{(j)}$

1. Given $h^{(j-1)}$, generate $h^* \sim \mathrm{LN}(\log(h^{(j-1)}), \mathrm{V}_h)$
2. Compute

$$\alpha_h = \min\left\{\frac{\pi(h^*, u^{(j)}, \mu^{(j-1)}, \sigma^{(j)}, \xi^{(j)}|\mathbf{X})}{\pi(h^{(j-1)}, u^{(j)}, \mu^{(j)}, \sigma^{(j)}, \xi^{(j)}|\mathbf{X})} \frac{\mathrm{LN}(h^{(j-1)}|\log(h^*), V_h))}{\mathrm{LN}(h^*|\log(h^{(j-1)}), V_h))}, 1\right\}$$

3. With probability α_h, accept h^* and set $h^{(j)} = h^*$; otherwise reject h^* and set $h^{(j)} = h^{(j-1)}$.