Computational Statistics and Data Analysis 55 (2011) 2137–2157

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

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 Healthcare, Queen Charlotte's and Chelsea Hospital, London, UK

ARTICLE INFO

Article history:
Received 22 April 2010
Received in revised form 6 December 2010
Accepted 5 January 2011
Available online 14 January 2011

Keywords:
Extreme values
Mixture model
Kernel density
Threshold selection

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 non-parametric 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. The potential of the proposed model for overcoming the lack of consistency of likelihood based kernel bandwidth estimators when faced with heavy tailed distributions is also demonstrated.

1. Introduction

Extreme value theory is unlike most traditional statistical theory, which typically examines the "usual" or "average" behaviour of processes, in that it is used to motivate models for describing unusual behaviour or rare events. Practical applications are seen in many fields of endeavour including finance (Embrechts et al., 2003), engineering (Castillo et al., 2004) and environmental science (Reiss and Thomas, 2007), where the risk of rare events is of interest. At the heart of extreme value techniques is reliable extrapolation of risk estimates beyond 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 it 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.

1.1. The 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 $x - u$ over some suitably high threshold u can be well approximated

[*] Corresponding author. Tel.: +64 3 3642987x8338; fax: +64 3 3642587.
E-mail address: carl.scarrott@canterbury.ac.nz (C.J. Scarrott).
URL: http://www.math.canterbury.ac.nz/~c.scarrott (C.J. Scarrott).

0167-9473/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2011.01.005

by a generalised Pareto distribution, denoted by GPD(σ_u, ξ), given by

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

where $x > u$, $y_+ = \max(y, 0)$ and ξ and $\sigma_u > 0$ are the shape and scale parameters respectively. The GPD is defined conditionally on being above the threshold u. Although often not explicitly defined in the presentation of the GPD in the literature, a third implicit parameter is required in most applications: that of the probability of being above the threshold $\phi_u = P(X > u)$. The unconditional survival probability is then given by

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

This implicit parameter ϕ_u is typically estimated using the maximum likelihood estimator, which is simply the sample proportion above the threshold.

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 an exponential (e.g. power law decay); and
- $\xi < 0$: short tail with finite upper end point $u - \sigma_u/\xi$.

Sample estimates of the shape parameter and associated uncertainty intervals are used to determine the population tail behaviour. Typically, interest focuses on whether the population tail behaviour has an exponential decay ($\xi = 0$), which is the case for the normal distribution, or whether the tail is heavier or lighter (and therefore will be a short tail) than an exponential tail. However, due to the inherent sparsity of observations in the tails of the distributions, this determination is often not straightforward.

In applications, the asymptotically motivated GPD is typically assumed to approximate the upper (or lower) tail of the population distribution above some chosen threshold, and this is followed by various checks on the model fit (e.g. PP and QQ plots). Once the threshold is fixed, standard likelihood and Bayesian inference approaches are applicable; see Coles (2001) and Coles and Tawn (1996) respectively. The GPD tail model can then be used for extrapolation purposes, e.g. to estimate the likelihood of unobserved events occurring or the level that is exceeded with a certain probability (e.g. a 1 in 100 year rainfall event).

We return to the challenging problem of threshold choice in the next section, as partial motivation for the development of extreme value mixture models in Sections 1.3 and 2. In the meantime, we consider a more general tail model which includes the GPD as a special case. The main benefit of this more general model is that it overcomes the dependence of the parameter σ_u on the threshold which can be problematic in inferences. In particular, if undertaking Bayesian inference, this dependence will require a suitable joint prior for both the scale σ_u and the threshold u, rather than potentially two independent priors, and will also lead to a need to jointly consider the posterior information on these parameters. In general, dependence between the parameters can also lead to poor mixing and, depending on the algorithm used, inefficiency in Markov chain Monte Carlo based posterior sampling.

Pickands (1971) showed that the point process defined by

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

is well approximated, in the limit as $n \to \infty$, by an inhomogeneous Poisson process on the region $A = [0, 1] \times (u, \infty)$, for a sufficiently high threshold u, with the intensity function on the subregion $B = (t_1, t_2) \times (x, \infty)$ given by

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

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 the threshold exceedances using the point process (PP) framework follows a process similar to that of threshold modelling via the GPD, with the PP conditional probability model parametrised by (μ, σ, ξ). Application of the GPD and PP representations relies on the choice of a suitably high threshold u, above which the asymptotically motivated models provide a reliable approximation.

Another key benefit of the PP tail model over the GPD is that the former can more easily be extended to the non-stationary case by allowing a change in the intensity/number of threshold exceedances over time or space, which has a natural physical interpretation. Further, as the PP model also describes the threshold exceedance process over time (or space) it is more straightforward to consider representations of dependence in the extremes, e.g. clustering of exceedances, which are frequently observed in applications ranging from finance (volatility clustering) to hydrology (for example due to climatic variation).

The value of n_b can be seen as completely arbitrary, as for any particular choice the PP parameters change deterministically. In particular, n_b can be defined for deriving the classical extreme value models as special cases. The GPD is a special case where n_b is the number of threshold exceedances, with the major benefit of the PP representation that the parameters (μ, σ, ξ) are independent of the threshold. The shape parameters for the PP and GPD models are the same and the GPD scale parameter is related to the PP parameters by $\sigma_u = \sigma + \xi(u - \mu)$. The value of n_b will change the properties of the likelihood function, and so can be chosen to improve the likelihood properties and therefore computational aspects of the inference process; see Wadsworth et al. (2010) for details. The relationship between the PP parameters for n_b blocks (μ_b, σ_b, ξ_b) and for n_y blocks (μ_y, σ_y, ξ_y) is given by

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

We will return to the selection of n_b for improving the likelihood properties in Section 2.2.1, which also substantially improves the computational aspects of the Bayesian inference approach undertaken in Section 3.

1.2. Threshold choice

It is common practice to use properties of the GPD/PP models to aid threshold selection 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/PP model) the mean excesses above any higher threshold $v \geq u$ will be linear, as follows:

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

for $\xi < 1$; 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 the tail extrapolation. General principles to follow Coles (2001) are choosing as low a threshold as possible to maximise the sample data for efficient inference, but not selecting so low a threshold that the asymptotic theory underlying the tail models is invalidated (as indicated by these graphical diagnostics).

Further, traditional inference approaches for the GPD/PP models assume that the threshold, once chosen, is a fixed quantity, so the estimation uncertainty is not accounted for in further inferences. The goal of automating the threshold choice for efficient application to many data sets has also proven elusive. Dupuis (2000) has recently developed a more robust technique for aiding 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. The following section will review pertinent existing approaches to overcoming these problems and detail our proposed methodology.

1.3. 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/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 problems of threshold choice and uncertainty estimation discussed in the previous section.

Mendes and Lopes (2004) propose a simple mixture model where the main mode is assumed to be normal and two separate GPD models are used for the tails, with threshold estimation carried out by either a 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 distribution) for the main mode, to the GPD which will dominate the upper tail. There is no explicit threshold in this approach, as they have essentially replaced the threshold estimation problem with that of estimating the transition function parameters; however a threshold could be determined as the point at which the weighted contribution from the Weibull distribution 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) up to 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 (a combination of normal and GPD tails, with the resultant density constrained to be continuous up to the first derivative) to approximate the distribution with support on the entire real axis, including extension to a mixture of these hybrid Pareto distributions 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 bulk of the distribution (and the associated weight function where appropriate), and the complicated inference (and sample properties) for the mixture of hybrid Pareto distributions. Tancredi et al. (2006) proposed a semi-parametric mixture model comprising piecewise uniform distributions from a threshold which is known to be too low, up to 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 threshold, with a PP model based tail above. Bayesian inference is used with a reversible jump algorithm due to the unknown number of uniform distributions. The threshold is defined as a parameter, so the inferences naturally account for the threshold uncertainty.

In this paper we propose a flexible model for analysing 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 just one extra (kernel bandwidth) parameter above the usual PP parameters (and the 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 distributions of Carreau and Bengio (2009). As with the other mixture models, the proposed one 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 6 also provides new insights into the complex uncertainties induced in the tail estimates due to the threshold selection. The proposed model is also shown to overcome a long-standing problem with standard likelihood based kernel density bandwidth estimators when applied to data with heavy tails (e.g. Cauchy distributed) in Section 5.

In the next section we outline the proposed mixture model, with Sections 2.2 and 3 providing the details of the Bayesian inference using MCMC methods for posterior sampling of the model parameters, including details of the prior distributions. Sections 4.1 and 4.2 assesses the performance and features of the approach using simulated data sets. In Section 6, we consider application of the proposed model in studying 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 7 with a discussion of our findings and the potential for further research.

2. The proposed mixture model

This section details the proposed extreme value mixture model, which simultaneously describes the bulk of the distribution and the tail, encapsulating the threshold as a parameter and thus bypassing the issues associated with threshold selection. The observations below the threshold are assumed to follow a non-parametric density $h(\cdot|\boldsymbol{\lambda}, \mathbf{X})$, which is dependent on a parameter λ and the observation vector $\mathbf{X}$. The upper tail (excesses above the threshold) is 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 comprise a sequence of n independent and identically distributed observations $\mathbf{X} = \{x_i : i = 1, \dots, n\}$ with distribution function F defined by

$$F(x|\lambda, u, \sigma_u, \xi, \mathbf{X}) = \begin{cases} (1-\phi_u)\dfrac{H(x|\lambda, \mathbf{X})}{H(u|\lambda, \mathbf{X})} & x \le u \\ (1-\phi_u) + \phi_u G(x|u, \sigma_u, \xi) & x > u \end{cases} \tag{5}$$

where $\phi_u G(\cdot|u, \sigma_u, \xi)$ 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, 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 sampling is utilised below with posterior predictive density estimation, which integrates over the entire posterior, in practice a smooth density estimate (at 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

$$f_1(x) = \frac{h(x|\lambda, \mathbf{X})}{H(u|\lambda, \mathbf{X})} I_{(-\infty,u]}(x)$$

$$f_2(x) = g(x|u, \sigma_u, \xi)\, I_{(u,\infty)}(x).$$

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 take full benefit from the efficiency of the EM algorithm, as all the components share a common parameter (u), so the information contained in the data cannot be separated into contributions to each component. Bayesian inference with MCMC sampling is undertaken instead, which is a common alternative.

2.1. The 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

$$K(x) \geq 0 \quad \text{and} \quad \int K(x)\mathrm{d}x = 1.$$

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 Eq. (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. A mean zero gaussian pdf is used as the kernel in this article, so the bandwidth λ is the standard deviation of the kernel.

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 $O(\lambda)$ at boundary points, compared with bias of the order $O(\lambda^2)$ 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 renormalise. A variety of methods have been developed in the literature for removing 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 (lower) end points to the underlying process being modelled. Extensions of the model for dealing with boundaries are considered in follow-up research. However, in Section 5 we consider a practical alternative to boundary correction which works well in practice: simply extending the mixture model to have a GPD for both the upper and lower tails.

2.2. The likelihood function

The likelihood for the extreme value mixture model in Eq. (5) can be separated out into the contributions from the observations below the threshold (the kernel density component) and those above the threshold (the GPD or PP tail model), with the PP representation shown here:

$$L(\theta|\mathbf{X}) = L_K(\lambda, u|\mathbf{X})L_{PP}(u, \mu, \sigma, \xi|\mathbf{X}) \tag{7}$$

where the parameter vector is $\theta = (\lambda, u, \mu, \sigma, \xi)$. The likelihood functions for each component will be detailed in the following sections. However, it worth noting at this stage that this likelihood function cannot take full benefit from the EM algorithm, which is commonly used for mixture model parameter estimation in a likelihood framework. Conditional on the latent component variable used in the EM algorithm, the contribution of the data to the parameter vector cannot be separated into the two components due to the common threshold parameter u in the two likelihood components.

2.2.1. The likelihood function for the PP model

The PP likelihood is obtained by straightforward application of the Poisson likelihood with inhomogeneous intensity function given by Eq. (3):

$$L_{PP}(u, \mu, \sigma, \xi|\mathbf{X}) = \begin{cases} \exp\left\{-n_b\left[1+\xi\left(\frac{u-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}\prod_B \frac{1}{\sigma}\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-1-1/\xi} & \text{for } \xi \neq 0, \\ \exp\left[-n_b\exp\left(\frac{u-\mu}{\sigma}\right)\right]\prod_B \frac{1}{\sigma}\exp\left(\frac{x_i-\mu}{\sigma}\right) & \text{for } \xi = 0, \end{cases}$$

where $B = \{i : x_i > u\}$. The GPD likelihood function (if required) is easily derived using the products of the density defined in Eq. (1). Regularity conditions for PP/GPD likelihood functions, including conditions for existence, for extreme value models are discussed by Smith (1985), but are of no practical concern here. Wadsworth et al. (2010) investigated the shape of the likelihood for the PP model and, in particular, how this varies with the choice of n_b. They suggest that n_b should be set at the number of exceedances of the threshold u, to ensure good properties for optimisation purposes (including good mixing and efficiency for Bayesian MCMC based methods).

2.2.2. The likelihood function for kernel bandwidth estimation

The choice of kernel bandwidth 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 kernel bandwidth 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, ranging from minimising various model fit criteria (Jones et al., 1996) to the use of Bayesian techniques. The latter approach has also been developed for both multivariate density estimation (Zhang et al., 2006) and univariate density estimation, by Brewer (1998, 2000).

Likelihood inference for the global kernel bandwidth λ was first proposed by Habbema et al. (1974) and Duin (1976). They treat the bandwidth as a parameter to be estimated, and showed that the likelihood is unbounded as $\lambda \to 0$, as each sum term in the product of

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

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

$$L(\lambda|\mathbf{X}) = \prod_{i=1}^{n} \frac{1}{(n-1)} \sum_{\substack{j=1 \\ j \neq i}}^{n} K_\lambda(x_i - x_j),$$

which can be viewed as minimising an estimate of Kullback–Leibler distance; see Bowman (1980, 1984) for details. This cross-validation likelihood for the kernel bandwidth (now considered a parameter) is used in the likelihood for our proposed model defined in Eq. (7). In the proposed model, all observations are used as kernel centres; however, only those below the threshold in the set $A = \{i : x_i \leq u\}$ contribute to the likelihood in Eq. (7). Hence, the kernel density is renormalised to get the appropriate contribution to the likelihood giving

$$L_K(\lambda, u|\mathbf{X}) = \left\{ \frac{(1-\phi_u)}{\frac{1}{n}\sum\limits_{k=1}^{n} \Phi\left(\frac{u-x_i}{\lambda}\right)} \right\}^{|A|} \prod_{A} \frac{1}{(n-1)} \sum_{\substack{j=1 \\ j \neq i}}^{n} K_\lambda(x_i - x_j). \tag{8}$$

Brewer (2000) consider local varying bandwidth estimators based upon adapting the cross-validation likelihood based approach for a global kernel bandwidth discussed in the previous section. They also implement a Bayesian inference approach, including both a locally adaptive cross-validation likelihood and prior information. A straightforward extension to the global kernel bandwidth utilised in our approach is possible, using the ideas presented in Brewer (2000). However, as our interest is predominantly in tail quantities, the increased performance derived from the locally adaptive kernel density estimates is not considered of immediate relevance, so we consider only a global bandwidth parameter.

2.2.3. *Consistency of the likelihood based kernel bandwidth estimator*

Habbema et al. (1974) and Duin (1976) showed that the cross-validation likelihood based kernel bandwidth estimators work well for short tailed distributions (i.e. with the tail lighter than an exponential). However, they drastically oversmooth heavy tailed distributions. Schuster and Gregory (1981) showed that this problem is due to inconsistency of the likelihood based bandwidth estimate. For kernels with both bounded support, i.e. $[-1, 1]$, and left continuous kernels of bounded variation on $(-\infty, \infty)$, they showed that the likelihood based estimates are inconsistent for a wide class of population densities, including the Cauchy distribution. Schuster and Gregory (1981) observed that the smoothing parameter $\lambda = \lambda^*$ which maximises the likelihood in Eq. (8) for each x_i must satisfy $|x_i - x_j| \leq \lambda^*$ for some x_j with $j \neq i$. Denoting the order statistics as $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ and considering the upper tail of the distribution, they showed that for heavy tailed data the difference in the two largest order statistics $|x_{(n)} - x_{(n-1)}| \not\to 0$ as $n \to \infty$, and hence $\lambda^* \not\to 0$. This property leads to an inconsistent likelihood estimator for the bandwidth, for heavy tailed distributions like the Cauchy distribution. 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 oversmoothing) not only for heavy tailed distributions, but also in situations where outliers are present.

Section 5 demonstrates that another potential application of our proposed mixture is in overcoming the lack of likelihood based bandwidth estimation consistency for heavy tailed data. Within the proposed extreme value mixture model the upper tail is captured by the GPD component, so the inconsistency of the likelihood based 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 proposed approach to allow both the upper and lower tails to be captured using GPD/PP models would also resolve the inconsistency.

3. Bayesian inference

Bayesian inference is a commonly used inference approach for mixture models; see for example Robert (2007). In our model, Bayesian inference overcomes the issue of the common threshold parameter u across the mixture components mentioned in Section 2.2, which means that the full benefit of the EM algorithm cannot be obtained. Further, Bayesian inference is also potentially of great value in extreme value applications due to the possibility of supplementing the inherent lack of sample information in the tails, for expert prior information; see Coles and Powell (1996) for discussion. However, in the simulation study and application shown in this article, diffuse prior information has been provided to show what is, in some sense, the worse case scenario in terms of performance where there is essentially no prior expert information.

3.1. Prior specification

The joint prior distribution of $\theta = (\lambda, u, \mu, \sigma, \xi)$, under the reasonable assumption that the PP parameters are independent of all of the threshold and kernel bandwidth (a parameter), is expressed as

$$\pi(\lambda, u, \mu, \sigma, \xi) = \pi(\lambda) \cdot \pi(u) \cdot \pi(\mu, \sigma, \xi). \tag{9}$$

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

A similar prior can be specified for the simpler GPD tail model; however, this is somewhat complicated by the dependence on the threshold u and scale σ_u parameters in this case and so needs more careful consideration. In practice, we have utilised diffuse independent priors for these two GPD related parameters (similar to those presented for the PP model below), which lead to inefficient posterior sampling and occasionally poor chain mixing, but no important changes to the resultant inferences. As discussed in Section 1.1, the GPD is a special case of the PP tail model, so the details of the alternative GPD posterior sampling are not given, for brevity.

3.1.1. Priors 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 than for the 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, GEV) model. Section 1 showed that by varying n_b and using the transformation given by (4) it is possible to translate between the parameters for the GPD and block maximum GEV approach ($n_b = 1$), thus permitting the prior elicitation approach 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. level exceeded on average once every $1/p$ blocks). We can also see that on working with the block maxima representation, the parameters are not dependent on the threshold, thus justifying the independence assumption in the 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 $\{q_i : i = 1, 2, 3\}$, specification of independent priors for the three different quantiles would not be inappropriate. Priors are therefore specified 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. Coles and Tawn (1996) suggest marginal priors for the differences of

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

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 expert 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 the 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 on the basis of 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

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

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 margins. Coles and Powell (1996) discuss this method for spatial modelling of extreme wind speeds, and gave precise values for $\boldsymbol{\eta}$ (the location vector) and Σ (the covariance matrix). For the simulation study in Section 4, limited prior information is desirable to allow the data to "speak for themselves", so a simple trivariate normal distribution with independent components and high variances is used.

3.1.2. The prior for the threshold (u)

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

3.1.3. The prior for the kernel density bandwidth parameter (λ)

Brewer (1998, 2000) and Zhang et al. (2006) consider Bayesian inference for the bandwidth parameter of kernel density estimators. Brewer (2000) considers local varying bandwidths and Zhang et al. (2006) consider bandwidth selection for multivariate kernel density estimation.

We follow the prior definitions detailed in Brewer (2000, 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 taken 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 $\lambda < 0.50$.

3.2. The posterior sampling scheme

The posterior is defined in the usual way as proportional to the likelihood defined in Eq. (7) multiplied by the prior defined in Eq. (9), with the latter made up of prior components specified in the preceding sections, giving

$$\pi(\lambda, u, \mu, \sigma, \xi|\mathbf{X}) \propto L(\theta|\mathbf{X}) \cdot \pi(\lambda|d_1, d_2) \cdot \pi(u|\mu_u, \nu_u^2, e_1) \cdot \pi(\mu, \sigma, \xi).$$

Posterior sampling for the proposed extreme value model is achieved via Markov chain Monte Carlo (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. Further, the threshold cannot be below the minimum data point or above the maximum data point. The likelihood function has been defined to explicitly encompass these restrictions.

MCMC samplers require specification of proposal distributions for the parameters. Proposal distributions have been selected to reflect the other restrictions on the model parameters (i.e. log-normal to ensure positive scale σ and bandwidth λ parameters). For simplicity a random walk sampler is used, with variances for the normal and log-normal proposals used as tuning parameters to ensure that there are suitable rates of acceptance of the sampling chain, following the guidance provided by Gelman (1996). The full posterior simulation algorithm is given in Appendix A.

3.3. Prediction and interval estimation

Prediction of tail quantities is often of primary concern in extreme value analysis, which can be achieved within a Bayesian framework via the posterior predictive distribution. The predictive distribution is given by

$$g(y|\mathbf{X}) = \int_{\Theta} f(y|\theta)\pi(\theta|\mathbf{X})\mathrm{d}\theta,$$

where θ is the parameter vector, $\pi(\theta|\mathbf{X})$ captures the parameter uncertainty and $f(y|\theta)$ gives the uncertainty associated with future observations, allowing the posterior predictive estimates to fully capture all uncertainties about tail estimates. Defining Z as the exceedance level for a given return period $1/p$, the predictive distribution for Z is defined as

$$\Pr\{Z \leq z|\mathbf{X}\} = \int_{\Theta} \Pr\{Z \leq z|\theta\}f(\theta|\mathbf{X})\mathrm{d}\theta = 1 - p$$

and solving for z will give the return level of the process. If the posterior chain $\theta_1, \ldots, \theta_m$ is regarded as estimates from the stationary distribution $f(\theta|\mathbf{X})$, the Monte Carlo integration approximates this integral:

$$\Pr\{Z \leq z|\mathbf{X}\} \approx \frac{1}{m}\sum_{i=1}^{m} \Pr\{Z \leq z|\theta_i\} = 1 - p \tag{10}$$

where the solution for z can be found using a standard numerical solver. The results in Section 5 and application in Section 6 make use of the predictive distribution.

While posterior uncertainty can be captured within the predictive density, it can also be summarised using the highest posterior density (HPD) region (a form of credible interval). The $(1-\alpha)$ HPD region comprises those parameter values with the highest posterior densities and containing $100(1-\alpha)\%$ of the posterior probability. All credible intervals presented are HPD intervals unless otherwise specified.

4. A simulation study

The simulation study used to demonstrate the performance of the model and estimation procedure is split into two parts. Firstly, we study the performance of the mixture model for approximating standard parametric distributions with various upper and lower tail behaviours. Secondly, we study 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. The MCMC algorithm was run on a parallel Linux system, with a 64-bit AMD Opteron 1.8 GHz processor with 16 GB RAM. The required CPU time is around 90 min for a sample size of 1000.

The MCMC Metropolis–Hastings sampler outlined in Section 3 was initialised at an arbitrary starting parameter vector and run for 20,000 iterations with a burn-in period of 5000, giving 15,000 posterior draws for each simulation. Convergence of the chains is assessed using the standard diagnostics discussed in Gelman and Rubin (1992). The Gelman and Rubin (1992) and Gelman (1996) approach is based on the idea that the variance within a single chain will be less that the variance in combined sequences. In particular the Gelman–Rubin approach monitors scalar quantities of interest in the analysis (i.e. θ) where for example the scalar summary can be the mean elements of a given parameter chain of interest. The Gelman–Rubin approach is based on running multiple chains where the starting points of the chains are widely dispersed over the target distribution ensuring that all major regions are considered. Convergence is assessed on the basis of the potential scale reduction which is essentially a measure of how much the multiple chains are overlapping. Gelman and Rubin (1992) suggest that convergence occurs once the potential scale reduction is below 1.2 or 1.1.

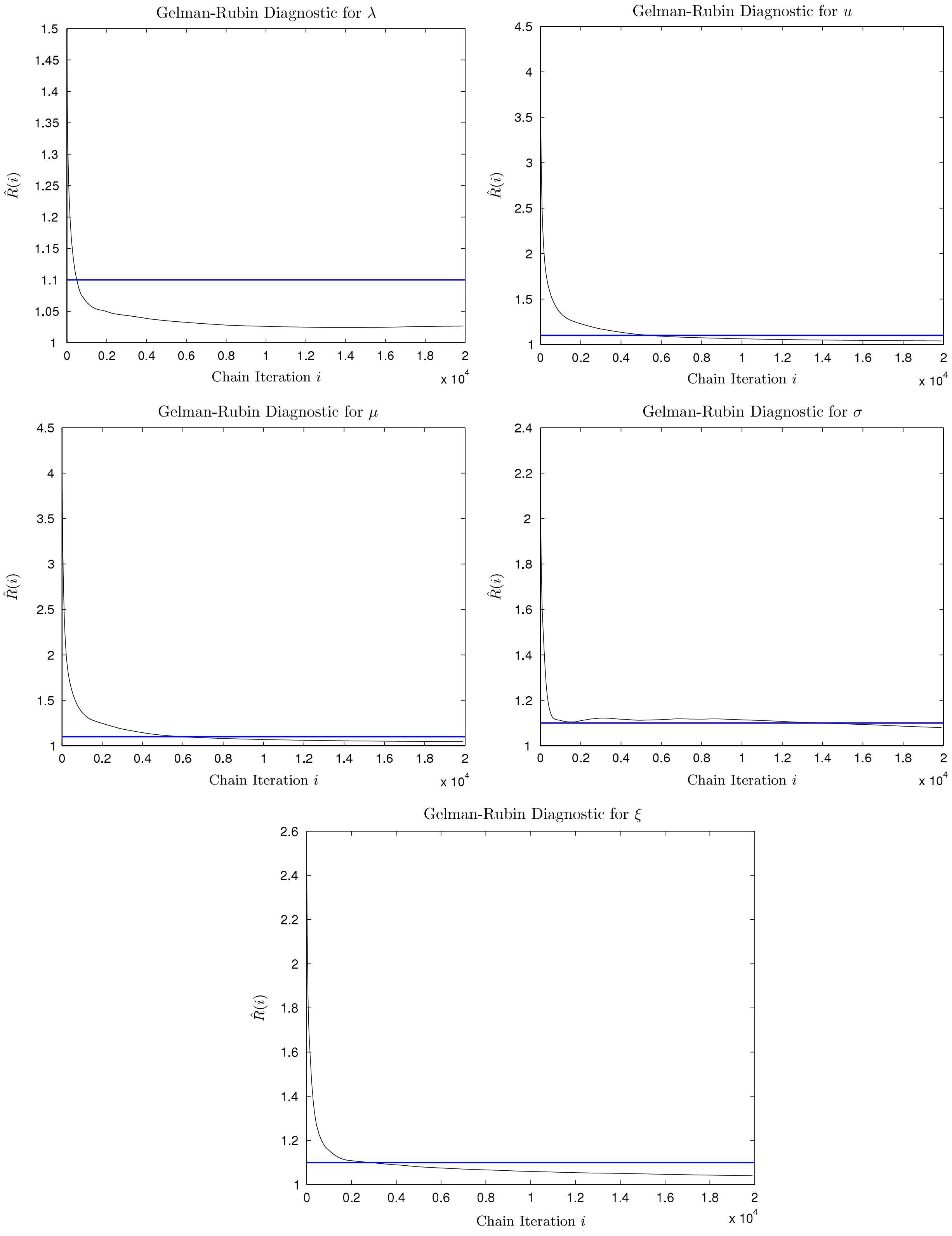
Fig. 1 shows the results for the Gelman–Rubin convergence test based on a simulated $N(0, 3)$ data set as an example, with six chains run with disperse starting points. The results for chains from each starting point are not shown for brevity. However, it is clear that the chains converge quickly, as evidenced in Fig. 1, where all scalar summaries (means of parameters) show signs of convergence (dropping below the reference line of 1.1) after 5000 iterations or earlier. The only slight exception is for σ, where it is very close to 1.1 from around 1500 iterations and is still well below 1.2. Hence, we are confident in using a burn-in of 5000 with the remaining 15,000 draws from the posterior being used for inference purposes.

4.1. Application to standard parametric distributions

Three standard parametric population distributions which cover a range of possible tail behaviours and skewness/symmetry of bulk distribution are considered: namely the normal, Student-t (with three degrees of freedom) and negative Weibull. The first two are symmetric with the normal distribution having exponential type tails ($\xi = 0$) and Student-t has 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, heavy tailed distributions (namely the Cauchy distribution) are considered separately in Section 5.

Various parameter sets for the bulk distributions were considered with the results for the Weibull($\lambda = 10, k = 5$), $N(0, 3)$ and Student-$t(\nu = 3)$ shown for brevity 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 such that 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 proportion of simulations in which the HPD interval includes the true parameter/quantile). 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 exponential limit ($\xi = 0$) is extremely



Gelman-Rubin Diagnostic for λ

Gelman-Rubin Diagnostic for u

Gelman-Rubin Diagnostic for μ

Gelman-Rubin Diagnostic for σ

Gelman-Rubin Diagnostic for ξ

Fig. 1. Results from applying the Gelman–Rubin method for checking the convergence of Markov chains using running potential scale reduction for a posterior sampling chain for $\theta = (\lambda, u, \mu, \sigma, \xi)$.

Table 1
Summary of the performance of the mixture model using Bayesian inference for estimating the shape parameter ξ and 0.99/0.999 quantiles for three population distributions across 100 simulations. True values for the shape and quantiles are shown in $[\cdot]$. Coverage rates for nominal 95% credible intervals, average posterior means and interval lengths are given with the standard error in parentheses.

	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)
$N(\mu = 0, \sigma^2 = 3)$	[−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)

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. We obtained the 95% HPD intervals after a burn-in of 5000 draws. There is no true bandwidth λ to compare the performance with, and as interest is focused on tail estimation so we consider the 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 are 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 sample expectations with 100 replicates, showing that 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 uncertainties associated with them (the interval length and its standard error) increase as the tail probability decreases.

4.2. Application to models spliced with extremal tails

The flexibility of the mixture model is now demonstrated by application to the same population distributions as in Section 4.1 above spliced together with a GPD/PP upper tail above some threshold. In particular, these spliced distributions are used to evaluate the performance in estimating the threshold and all the tail model (GPD/PP) parameters. These bulk densities are spliced with representative examples of three different tail behaviours, with shape parameters $\xi = \{-0.2, 0, 0.4\}$. The threshold u is positioned at the $100 \times (1 - p)\%$ 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 $\int_{-\infty}^{u} h^*(x)\mathrm{d}x = p$.
2. Generate $\mathbf{X} = \{x_1, \dots, x_n\}$ from $h^*(x)$.
3. Replace $\{\mathbf{X} : x_i > u \text{ for } i = 1, \dots, n\}$ with points generated from the GPD.

As before, various parameter sets for the bulk distributions were considered with the results for the Weibull$(\lambda = 10, k = 5)$, $N(\mu = 0, \sigma^2 = 3)$ 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 $n = 1000$ with upper tail probability at the threshold $p = 0.1$ (10% of the 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. The true parameters and quantiles are also shown. While the PP representation for the upper tail is used in the mixture model in the simulations, for brevity the GPD equivalent of the σ_u parameter is shown.

In general, ξ is well estimated with coverage rates close to 0.95 (up to sampling variability). The average of the posterior means for the shape parameter is very close to the true value for all three bulk population models spliced with all three combinations of tail behaviour. As we would expect, the average lengths 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 threshold uncertainty for positive shape parameter versus those with negative/zero shape, which will influence σ_u due to the aforementioned dependence.

Table 2
Summary of the performance of the mixture model using Bayesian inference for estimating the threshold, GPD scale σ_u and shape parameter ξ for three population distributions (Weibull, Student-t and normal) spliced with GPD tail for three tail behaviours ($\xi = -0.2$, 0 and 0.4) across 100 simulations. The true values for the threshold and GPD scale parameters are shown in the population distribution definition and the true shape parameters are shown in the first column. Coverage rates for nominal 95% credible intervals appear in the first column for each parameter, followed by the average posterior means and interval lengths in the third and second columns respectively.

ξ	GPD parameters								
	$\hat{u}$			$\hat{\xi}$			$\hat{\sigma_u}$		
Weibull$(l = 10, k = 5)\mathbb{I}_{(0,u)} + 0.1 \times \text{GPD}(u = 11.8, \sigma_u = 1.03, \xi)\mathbb{I}_{[u,\infty)}$									
−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-$t(\nu = 3)\mathbb{I}_{(-\infty,u)} + 0.1 \times \text{GPD}(u = 1.63, \sigma_u = 0.98, \xi)\mathbb{I}_{[u,\infty)}$									
−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
$N(\mu = 0, \sigma^2 = 3)\mathbb{I}_{(-\infty,u)} + 0.1 \times \text{GPD}(u = 3.84, \sigma_u = 1.71, \xi)\mathbb{I}_{[u,\infty)}$									
−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

Table 3
Summary of the performance of the mixture model using Bayesian inference for the 0.99/0.999 quantiles for three population distributions (Weibull, Student-t and normal) spliced with the GPD tails of three tail behaviours ($\xi = -0.2$, 0 and 0.4) across 100 simulations. The true value for the quantiles is shown in [·]. Coverage rates for nominal 95% credible intervals appear in the first column for each quantile, followed by average posterior means and interval lengths in the third and second columns respectively.

ξ	Quantiles					
	$\hat{q}_{0.99}$			$\hat{q}_{0.999}$		
Weibull$(l = 10, k = 5)\mathbb{I}_{(0,u)} + 0.1 \times \text{GPD}(u = 11.8, \sigma_u = 1.03, \xi)\mathbb{I}_{[u,\infty)}$						
−0.20	0.92	0.63	13.75 [12.48]	0.97	1.50	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]
Student-$t(\nu = 3)\mathbb{I}_{(-\infty,u)} + 0.1 \times \text{GPD}(u = 1.63, \sigma_u = 0.98, \xi)\mathbb{I}_{[u,\infty)}$						
−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]
$N(\mu = 0, \sigma^2 = 3)\mathbb{I}_{(-\infty,u)} + 0.1 \times \text{GPD}(u = 3.84, \sigma_u = 1.71, \xi)\mathbb{I}_{[u,\infty)}$						
−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]

The coverage rates for threshold estimates are very low; however, this is to be expected. If the GPD (or PP equivalent) is an appropriate model for some threshold u it will be suitable for all higher thresholds $\upsilon \geq u$. 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 the standard deviation of the 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 true value 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 traditional graphical diagnostics for threshold choice, practitioners 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 little bias in the 100 replicates. 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 insight in Section 6: that the threshold has a strong local influence (close to the threshold) on the distribution function. 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. However, the coverage rates quickly increased further away from threshold, as seen in Table 3.

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

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 + 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)

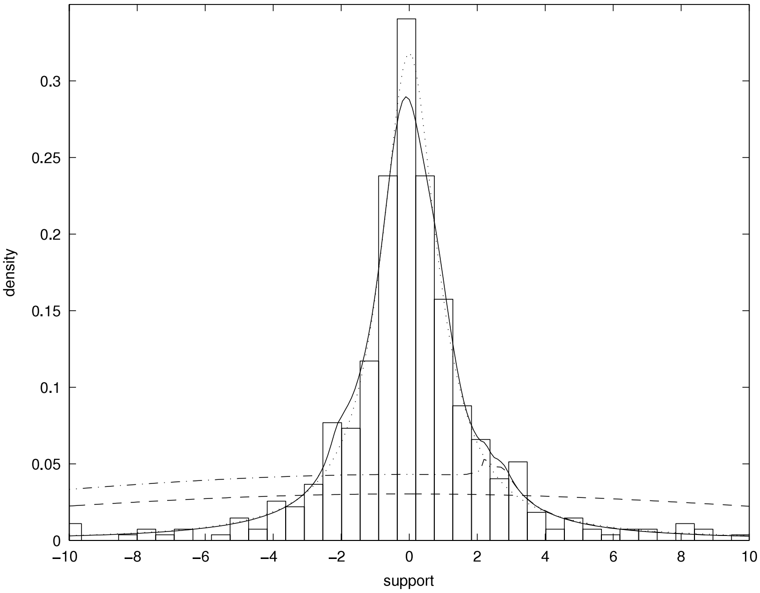


Fig. 2. Posterior predictive density estimator for Cauchy(0, 1) using various models: kernel density only (- - -); one-tailed mixture model (— · —); two-tailed mixture model (—) and (true) Cauchy(0, 1) pdf (· · ·).

5. Consistency of bandwidth estimates

Section 2.2.3 outlined issues surrounding the 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 the lower tails to be captured using GPD distributions. In particular, the model is defined as

$$F(x|\lambda, \boldsymbol{u}, \boldsymbol{\sigma_u}, \boldsymbol{\xi}, \mathbf{X}) = \begin{cases} \phi_1 G(-x|u_1, \sigma_{u1}, \xi_1) & x < u_1 \\ \phi_1 + (1 - (\phi_1 + \phi_2)) \dfrac{H(x|\lambda, \mathbf{X})}{\int_{u_1}^{u_2} h(x|\lambda, \mathbf{X}) \mathrm{d}x} & u_1 \le x \le u_2 \\ (1 - \phi_2) + \phi_2 G(x|u_2, \sigma_{u2}, \xi_2) & x > u_2 \end{cases} \tag{11}$$

where $\boldsymbol{\xi} = (\xi_1, \xi_2)$, $\boldsymbol{\sigma_u} = (\sigma_{u1}, \sigma_{u2})$, $\boldsymbol{u} = (u_1, u_2)$, $\phi_1 \text{GPD}(- \cdot |u_1, \sigma_{u1}, \xi_1)$ is the unconditional GPD function for the x_i that are $< u_1$, and $\phi_2 \text{GPD}(\cdot|u_2, \sigma_{u2}, \xi_2)$ is the unconditional GPD function for the x_i that are $> u_2$. Inference for this model can follow exactly the same methods as are outlined in Section 3 using the PP representation for the GPDs 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 to a sample of 500 Cauchy random variables, using Bayesian inference for 20,000 MCMC iterations with burn-in of 5000. Prior distributions for both sets of PP parameters were set to diffuse trivariate normal distributions with independent margins:

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

The Cauchy(0, 1) distribution is a special case of the Student-t when $\nu = 1$, so the asymptotic tail behaviour has $\xi_1 = \xi_2 = 1$. Fig. 2 and Table 4 give the key results for a single sample representative of all the samples considered. For comparison, the model with the kernel density estimator only and that with the kernel density estimator spliced with the PP/GPD upper tail were also considered.

Note that the two-tailed mixture model provides a very good fit to the bulk distribution, shown by the closeness of the 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. However, when no tail model is used, or just a single tail model, the kernel bandwidth is substantially biased upwards due the heavy tails, providing drastic oversmoothing, as shown in Fig. 2. This demonstrates that the heavy lower tail behaviour can have a strong influence on the bulk distribution estimate, and potentially for low quantiles below/around the threshold. However, another good feature of the proposed mixture model is that even given the drastic oversmoothing in the bulk model it will be noticeable that the upper tail model is still managing to provide a reasonable fit similar to that of the two-tailed mixture model. In particular, the one-tailed upper tail parameters (u, σ_u, ξ) are very similar to those for the upper tail for the two-tailed mixture model $(u_2, \sigma_{u2}, \xi_2)$ in Table 4.

These results show that the proposed two-tailed mixture model can overcome the long-standing inconsistency of the likelihood based kernel bandwidth estimator for heavy tailed distributions. Effectively the positive bias in the traditional likelihood based bandwidth estimates due to lack of decay of the separation between the uppermost (and lowermost) order statistics is irrelevant in the mixture model, as the tails which are the source of the uncertainty are approximated by the GPD, which is flexible enough to allow for both short tailed, exponential and heavier tailed distributions. Only a small number of extra degrees of freedom are required for the two tails.

Many applications in finance require modelling excesses for both tails, for example, simultaneously modelling the risk associated gains as well as losses, and fully accounting for their associated uncertainties. The two-tailed model of (11) could also be useful in these situations, overcoming the issue of dual-threshold estimation (and the 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-tailed model with the equal (symmetric) or unequal (asymmetric) tail shape parameters. Thus the two-tailed model could also provide a flexible framework for applications where both tails are of interest.

6. An application

Babies born prematurely are vulnerable to tissue and organ injury as a result of immature physiological adaptation to extrauterine life. Clinicians take various physiological measurements from premature babies in neonatal intensive care units (NICUs), which are monitored for clinical care, including oxygenation saturation, pulse rates and respiration rates. The challenge faced by clinicians is the assessment of variations 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 could be refined. A 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) considered well and stable and not receiving supplementary oxygenation intervention treatment at the NICU at Christchurch Women's Hospital, New Zealand. The data are collected over roughly a 6 h period at 0.5 Hz (once every 2 s). During monitoring, the infant was in various states, including showing various levels of awakeness (awake and quiet, awake and crying, quiet sleep and active sleep), and feeding by suckling or 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 measurements. We have randomly subsampled the data roughly every five 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 Fig. 3. In this paper we will only consider the marginal distribution of the time series, with the non-stationarity to be considered in future work. For this application, interest is in estimating the lower tail quantiles of the pulse rates.

The MCMC Metropolis–Hastings sampler outlined in Section 3 was initialised at an arbitrary starting parameter vector and run for 25,000 iterations with a burn-in period of 5000, giving 20,000 posterior draws. Convergence of the chains was assessed using the standard diagnostics discussed in Section 4, and similar results were found using the Gelman–Rubin convergence diagnostic as shown in Fig. 1 for an example simulated data set in Section 4. The resultant convergence diagnostics are not shown for brevity.

Tables 5 and 6 give results for both the kernel mixture model and the traditional fixed threshold approach for a range of sensible thresholds. Fig. 4 displays the mean residual life (MRL) plot, as discussed in Section 1.1. As interest is 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 to find a 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 for capturing the lower tail. A decline in the mean excess plot is seen above approximately 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 in the lower tail.

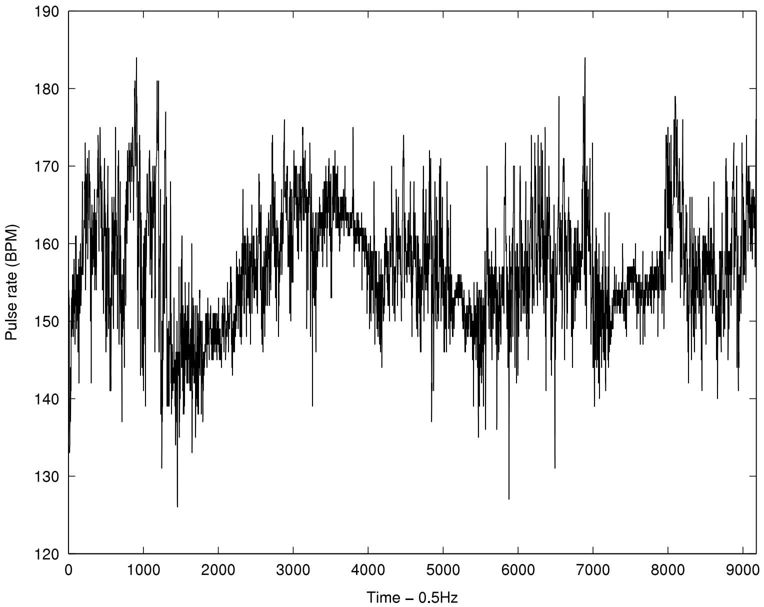


Fig. 3. Time series of pulse rates for a single neonatal patient taken every 2 s, for approximately 6 h.

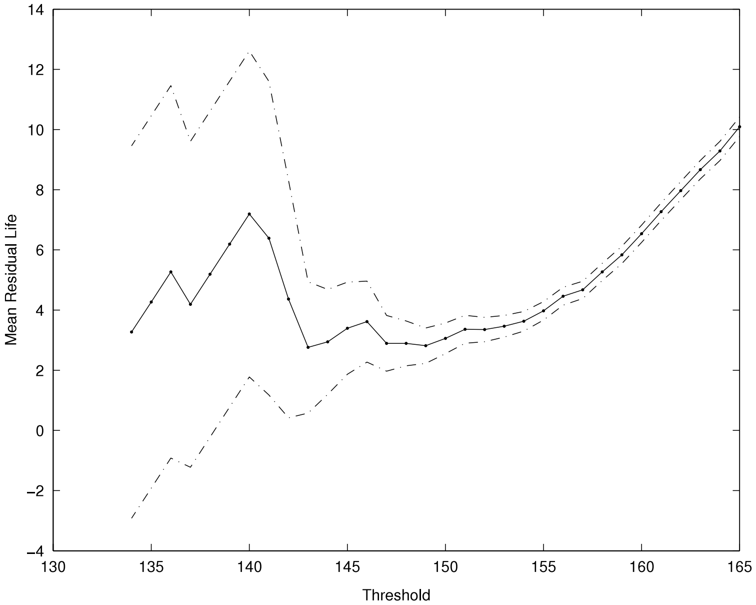


Fig. 4. Mean residual life plot for subsampled pulse rate data.

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

- 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, centred 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).

Fig. 5 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.

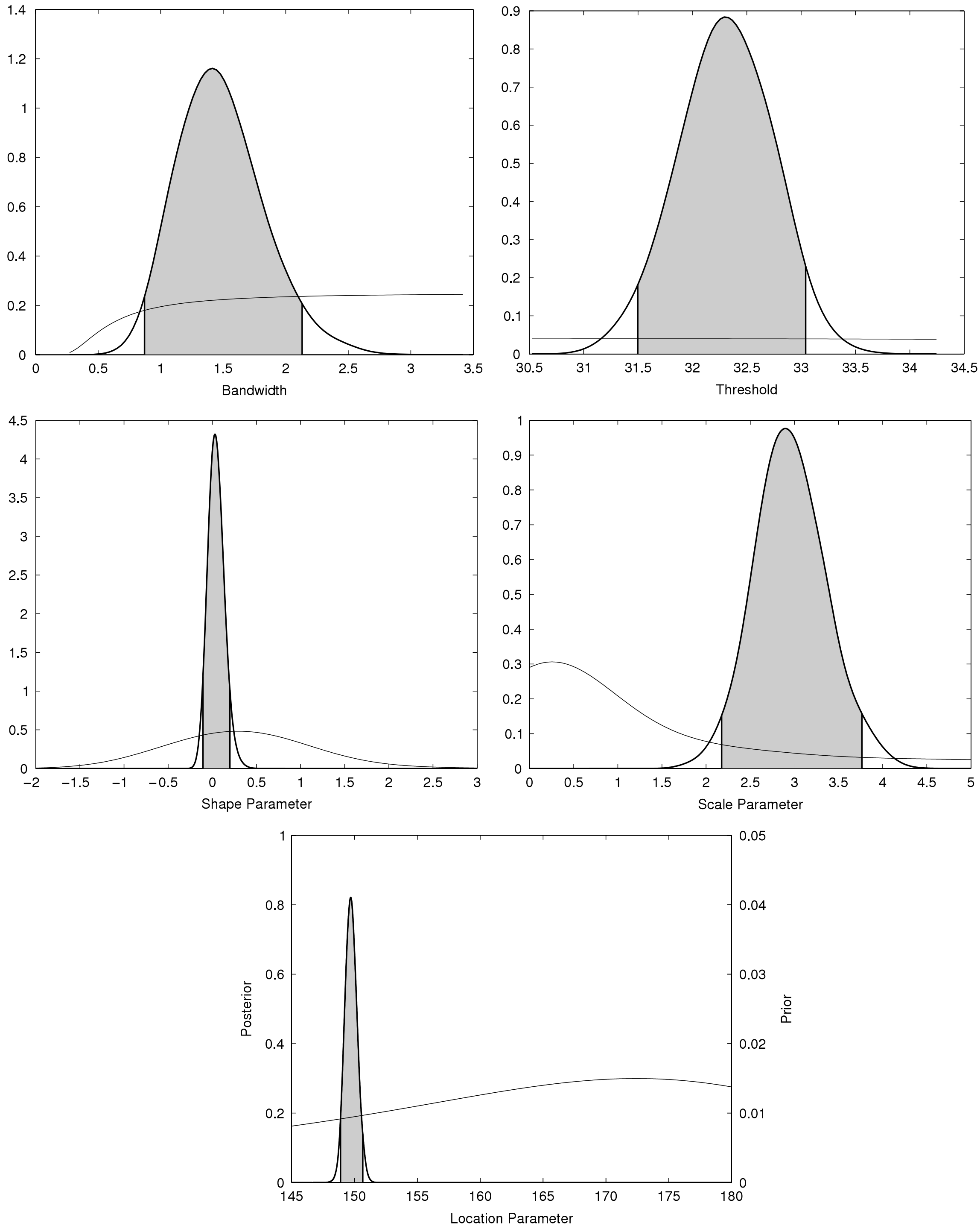


Fig. 5. Marginal prior (—) and posterior (**—**) distributions for each parameter within the extreme mixture model. Notice that for the location parameter the prior is so diffuse that it has been scaled (see the left y axis) to show the details.

It is also evident from Fig. 5 that the priors are not having 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.

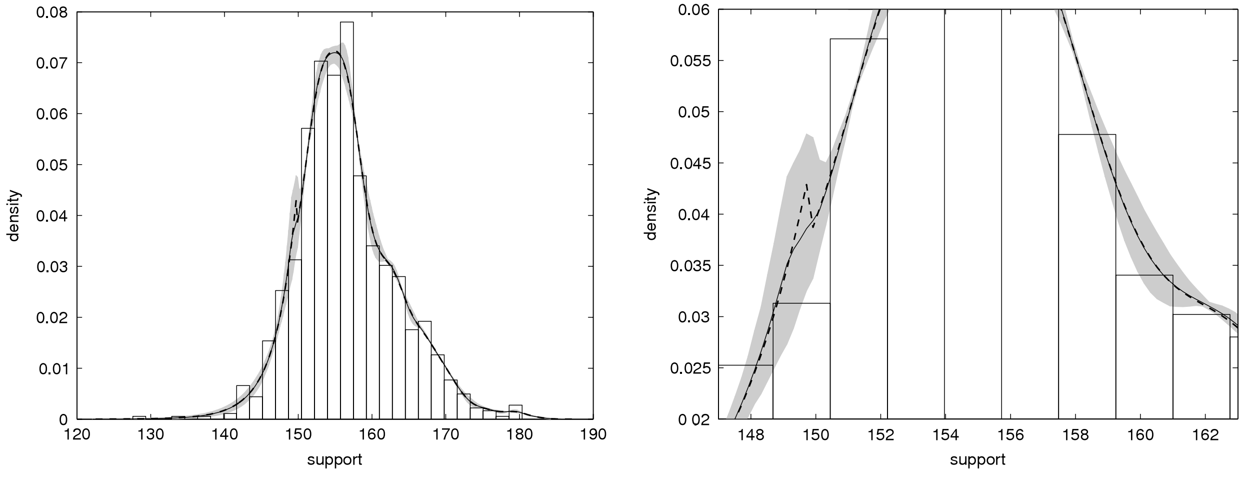
The MCMC algorithm was also run with diffuse priors for the point process parameters based on the trivariate independent normals as described in Section 3.1.1, with the results given in Table 5 alongside those for the quantile difference based priors. The alternative priors were used to ensure that the 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

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)

Table 6
Posterior means of the GPD parameters for the 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)



(a) Histogram with posterior predictive density.

(b) Zoom-in of (a).

Fig. 6. Sample density of pulse rates with a posterior predictive density estimate (—). The estimated density obtained from plugging in the posterior mean of the parameters is shown for comparison (- - -).

prior models in Table 5 suggests that the Markov chains have successfully converged, and prior structure is not having an adverse effect on the resulting posterior.

The posterior mean for the shape parameter, along with the associated 95% HPD interval, in Table 5 indicates an exponential type lower tail. The interval length for the threshold u is relatively small in magnitude, suggesting that 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 on the basis of the MRL plot given in Fig. 4. Table 6 indicates one of the issues surrounding threshold selection. For a threshold of 155, inference suggests a negative shape parameter $\xi = -0.10(-0.15, -0.04)$. On the basis of the MRL plot in Fig. 4 a threshold of 155 is a reasonable choice. However, all other possible thresholds give credible HPD intervals including the possibility of a 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 the fixed threshold approach in Table 6 for the threshold 149, which is close to that automatically selected for the mixture model. The interval lengths 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.

Fig. 6(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

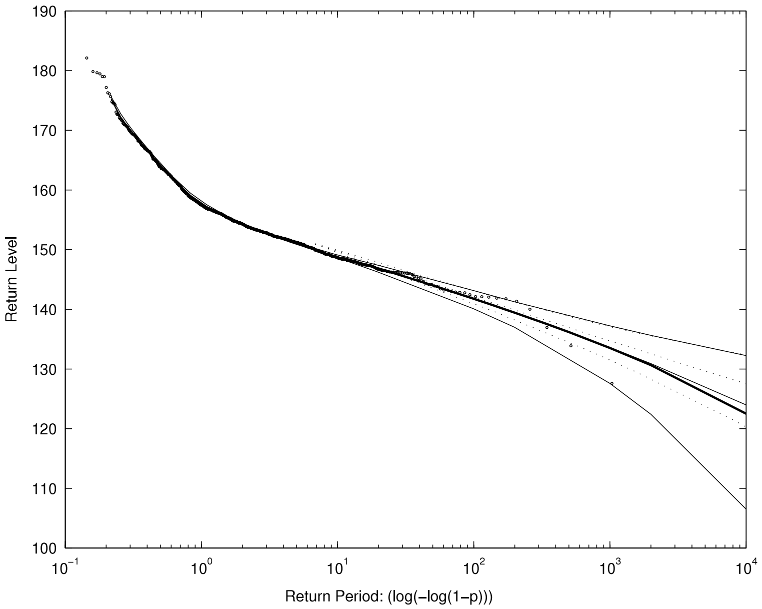


Fig. 7. Return level plot for the pulse rate data. The posterior predictive return level for the mixture model is shown by the bold solid line, with the corresponding posterior mean return level with a 95% HPD region shown by thin solid lines. The posterior mean return level for the fixed threshold model ($u = 150$) with a 95% HPD region is shown by dotted lines.

model density using the point estimates is only included to demonstrate that the individual posterior density estimates can exhibit a discontinuity at the threshold, as seen in Fig. 6(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 high when the density is at its highest and vice versa. Intuition suggests that the uncertainty relative to the density will be lowest near the mode (where we have the most data) with increasing relative uncertainty further out into the tails. This intuition is borne out in Fig. 6, with two key exceptions. Firstly, there is large relative uncertainty where the density is changing the most (i.e. the steepest slope), as shown clearly in the width of the intervals on the right in Fig. 6(b). Secondly, the threshold uncertainty impacts on the tail quantile estimates (seen clearly in Fig. 7 below) as expected, but it also has a substantial localised effect on the uncertainty of the distribution close to the threshold. The localised threshold uncertainty impacts are shown by the much larger grey intervals on the right in Fig. 6(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) has not to the authors' knowledge been highlighted in previous extremal threshold (mixture) modelling approaches.

In typical extreme value applications, interest is in describing the behaviour of extremal quantiles. Rather than interpreting these values on the basis of 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 as 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). Posterior estimates of extreme quantiles for threshold exceedances can be obtained by inverting Eq. (5), for given values of the parameter vector θ sampled from the posterior. These posterior sample return levels can be used to estimate posterior mean return levels and suitable HPD intervals. Posterior predictive estimates of the return levels are obtained by numerical solution of Eq. (10).

Fig. 7 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 a straight line under this transformation, with heavier tail than an exponential ($\xi > 0$) shown as a convex function and a shorter tail ($\xi < 0$) shown by a concave function. Return level plots can also be used as a model diagnostic to ensure that model based returns are in reasonable agreement with empirical estimates, as seen in Fig. 7. Table 7 gives return levels for $p = \{0.1, 0.01, 0.001, 0.0001\}$ for the models based on the fixed threshold approach.

Fig. 7 gives the return level plot for the mixture model approach (solid lines) 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 (validated using traditional graphical diagnostics) and is close to the posterior mean for the mixture model in Table 5, and so will provide a useful comparison of the return level estimates and uncertainty associated with the threshold choice. You will notice that the mixture model and fixed threshold GPD based return level functions are very similar indeed,

Table 7
Return levels for the fixed threshold approach for a range of thresholds and return periods, with 95% HPD intervals given in parentheses.

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

only showing deviations for quantiles with tail probabilities less than 10^{-3}. Further, it is also interesting to note that the posterior mean and posterior predictive return levels are extremely similar, with relevant deviations only observed above a return period of 10^3, which is around the largest observed return level (sample data point).

The curvature of the return levels suggests $\xi \geq 0$, 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 a 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 non-stationary 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 Fig. 7 and Table 7 to those for the fixed threshold approach reveals 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 particularly noticeable in Fig. 7. Further, the extra uncertainty with mixture model estimates has led to a higher coverage rate for the sample quantiles within the pointwise HPD intervals, thus providing further confirmation of the need to account for the uncertainty due to threshold choice.

7. 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 problem of 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 of inconsistency in the traditional likelihood based estimators of the kernel density bandwidth for heavy tailed data, e.g. Cauchy distributed populations.

The take home message, clearly demonstrated in Figs. 6 and 7, 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 the extra uncertainty, compared to that in the traditional fixed threshold approach, associated with the threshold choice should be accounted for, and the mixture model presented herein successfully encapsulates 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 considered in this paper.

Acknowledgements

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.

Appendix

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

```
Initialisation: Choose an arbitrary starting value $\theta^{(0)} = \{\lambda^{(0)}, u^{(0)}, \mu^{(0)}, \sigma^{(0)}, \xi^{(0)}\}$
Iteration: $j$ $(j \geq 1)$
• $\xi^{(j)}$
  1. Given $\xi^{(j-1)}$, generate $\xi^* \sim N(\xi^{(j-1)}, \mathrm{V}_\xi)$.
  2. Compute
        $\alpha_\xi = \min\left\{\frac{\pi(\lambda^{(j-1)}, u^{(j-1)}, \mu^{(j-1)}, \sigma^{(j-1)}, \xi^*|\mathbf{X})}{\pi(\lambda^{(j-1)}, u^{(j-1)}, \mu^{(j-1)}, \sigma^{(j-1)}, \xi^{(j-1)}|\mathbf{X})}, 1\right\}$
     where any constraints placed on $\xi$ are included within the likelihood.
  3. With probability $\alpha_\xi$, accept $\xi^*$ and set $\xi^{(j)} = \xi^*$; otherwise reject $\xi^*$ 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
        $\alpha_\sigma = \min\left\{\frac{\pi(\lambda^{(j-1)}, u^{(j-1)}, \mu^{(j-1)}, \sigma^*, \xi^{(j)}|\mathbf{X})}{\pi(\lambda^{(j-1)}, u^{(j-1)}, \mu^{(j-1)}, \sigma^{(j-1)}, \xi^{(j)}|\mathbf{X})} \frac{\mathrm{LN}(\sigma^{(j-1)}|\log(\sigma^*), V_\sigma)}{\mathrm{LN}(\sigma^*|\log(\sigma^{(j-1)}), V_\sigma)}, 1\right\}$
     where any constraints placed on $\sigma$ are included within the likelihood.
  3. With probability $\alpha_\sigma$, accept $\sigma^*$ and set $\sigma^{(j)} = \sigma^*$; otherwise reject $\sigma^*$ and set $\sigma^{(j)} = \sigma^{(j-1)}$.
• $\mu^{(j)}$
  1. Given $\mu^{(j-1)}$, generate $\mu^* \sim N(\mu^{(j-1)}, \mathrm{V}_\mu)$.
  2. Compute
        $\alpha_\mu = \min\left\{\frac{\pi(\lambda^{(j-1)}, u^{(j-1)}, \mu^*, \sigma^{(j)}, \xi^{(j)}|\mathbf{X})}{\pi(\lambda^{(j-1)}, u^{(j-1)}, \mu^{(j-1)}, \sigma^{(j)}, \xi^{(j)}|\mathbf{X})}, 1\right\}$
     where any constraints placed on $\mu$ are included within the likelihood.
  3. With probability $\alpha_\mu$, accept $\mu^*$ and set $\mu^{(j)} = \mu^*$; otherwise reject $\mu^*$ and set $\mu^{(j)} = \mu^{(j-1)}$.
• $u^{(j)}$
  1. Given $u^{(j-1)}$, generate $u^* \sim N(u^{(j-1)}, \mathrm{V}_u)\mathbb{I}_{(m,M)}$, where $m = \min(x_1, \ldots, x_n)$ and $M = \max(x_1, \ldots, x_n)$.
  2. Compute
        $\alpha_u = \min\left\{\frac{\pi(\lambda^{(j-1)}, u^*, \mu^{(j)}, \sigma^{(j)}, \xi^{(j)}|\mathbf{X})}{\pi(\lambda^{(j-1)}, u^{(j-1)}, \mu^{(j)}, \sigma^{(j)}, \xi^{(j)}|\mathbf{X})} \cdot \frac{(\Phi((M-u^*)/\sqrt{(V_u)}) - \Phi((m-u^*)/\sqrt{V_u}))}{(\Phi((M-u^{(j-1)})/\sqrt{V_u}) - \Phi((m-u^{(j-1)})/\sqrt{V_u}))}, 1\right\}$
     where all other constraints placed on $u$ are included within the likelihood.
  3. With probability $\alpha_u$, accept $u^*$ and set $u^{(j)} = u^*$; otherwise reject $u^*$ and set $u^{(j)} = u^{(j-1)}$.
• $\lambda^{(j)}$
  1. Given $\lambda^{(j-1)}$, generate $\lambda^* \sim \mathrm{LN}(\log(\lambda^{(j-1)}), \mathrm{V}_\lambda)$.
  2. Compute
        $\alpha_\lambda = \min\left\{\frac{\pi(\lambda^*, u^{(j)}, \mu^{(j-1)}, \sigma^{(j)}, \xi^{(j)}|\mathbf{X})}{\pi(\lambda^{(j-1)}, u^{(j)}, \mu^{(j)}, \sigma^{(j)}, \xi^{(j)}|\mathbf{X})} \frac{\mathrm{LN}(\lambda^{(j-1)}|\log(\lambda^*), V_\lambda)}{\mathrm{LN}(\lambda^*|\log(\lambda^{(j-1)}), V_\lambda)}, 1\right\}.$
  3. With probability $\alpha_\lambda$, accept $\lambda^*$ and set $\lambda^{(j)} = \lambda^*$; otherwise reject $\lambda^*$ and set $\lambda^{(j)} = \lambda^{(j-1)}$.
```

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

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

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. Journal of the Royal Statistical Society B 52 (3), 393–442.

Duin, R.P.W., 1976. On the choice of smoothing parameters for Parzen estimators of probability density functions. IEEE 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., 1996. Inference and monitoring convergence. In: Gilks, W., Richardson, S., Spiegelhalter, D. (Eds.), Markov Chain Monte Carlo in Practice. Chapman and Hall, London, pp. 131–143.

Gelman, A., Rubin, D.B., 1992. Inference from iterative simulation using multiple sequences (with discussion). Statistical Science 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. 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. Journal of the American Statistical Association 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). Journal of the Royal Statistical Society B. 59 (3), 511–567.

Pickands, J., 1971. The two dimensional Poisson process and extremal processes. Journal of Applied Probability 8, 745–756.

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

Robert, C., 2007. Bayesian Core: A Practical Approach to Computational Bayesian Statistics. Springer-Verlag.

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. Journal of the American Statistical Association 76 (373), 9–15.

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

Smith, R., 1985. Maximum likelihood estimation in a class of non-regular cases. Biometrika 72, 67–90.

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

Wadsworth, J., Tawn, J., Jonathan, P., 2010. Accounting for choice of measurement scale in extreme value modeling. The Annals of Applied Statistics 4 (3), 1558–1578.

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 selection for multivariate kernel density estimation. Computational Statistics and Data Analysis 50 (11), 3009–3031.