Statistical Modelling 2004; **4**: 227–244

Bayesian analysis of extreme events with threshold estimation

Cibele N Behrens[1], Hedibert F Lopes[2] and Dani Gamerman[3]
[1]Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
[2]Graduate School of Business, University of Chicago, Chicago, IL, USA
[3]Institute of Mathematics, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

Abstract: The aim of this paper is to analyse extremal events using generalized Pareto distributions (GPD), considering explicitly the uncertainty about the threshold. Current practice empirically determines this quantity and proceeds by estimating the GPD parameters on the basis of data beyond it, discarding all the information available below the threshold. We introduce a mixture model that combines a parametric form for the center and a GPD for the tail of the distributions and uses all observations for inference about the unknown parameters from both distributions, the threshold included. Prior distributions for the parameters are indirectly obtained through experts quantiles elicitation. Posterior inference is available through Markov chain Monte Carlo methods. Simulations are carried out in order to analyse the performance of our proposed model under a wide range of scenarios. Those scenarios approximate realistic situations found in the literature. We also apply the proposed model to a real dataset, Nasdaq 100, an index of the financial market that presents many extreme events. Important issues such as predictive analysis and model selection are considered along with possible modeling extensions.

Key words: Bayesian; extreme value theory; MCMC; mixture model; threshold estimation

Data and software link available from: http://stat.uibk.ac.at/SMIJ
Received May 2003; **revised** December 2003 and April 2004; **accepted** June 2004

1 Introduction

The extreme value theory literature has grown considerably in the last few decades, with applied interest in engineering, oceanography, environment, actuarial sciences and economics, among others. In such areas, the main problem is the scarcity of data or, more specifically, modeling with a fairly small amount of observations. Generally speaking, most of the traditional theory is more concerned with the 'center' of the distributions, the tails being commonly overlooked. Many theoretical developments have been proposed to appropriately study the tail of distributions (Embrechts *et al.*, 1997).

We focus on the class of problems where the behavior of the distributions over (below) a high (small) threshold is of interest, characterizing *extremal events*.

Address for correspondence: Cibele N Behrens, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil. E-mail: cibele@dme.ufrj.br and Dani Gamerman, Instituto de Matematica, Universidade Federal do Rio de Janeiro, Caixa Postal 68530, Rio de Janeiro, RJ, CEP21945-970, Brazil. E-mail: dani@im.ufrj.br

 10.1191/1471082X04st075oa

Pickands (1975) shows that if X is a random quantity with distribution function $F(x)$, then under certain conditions, $F(x|u) = P(X \leq u + x | X > u)$ can be approximated by a generalized Pareto distribution (GPD). A random quantity X follows a GPD if its distribution function is (Embrechts *et al.*, 1997)

$$G(x|\xi, \sigma, u) = \begin{cases} 1 - \left(1 + \frac{\xi(x-u)}{\sigma}\right)^{-1/\xi}, & \text{if } \xi \neq 0 \\ 1 - \exp\{-(x-u)/\sigma\}, & \text{if } \xi = 0 \end{cases} \tag{1.1}$$

where $\sigma > 0$ and ξ are the scale and shape parameters, respectively. Also, Equation (1.1) is valid when $x - u \geq 0$ for $\xi \geq 0$ and for $0 \leq x - u \leq -\sigma/\xi$ for $\xi < 0$. The data exhibit heavy tail behavior when $\xi > 0$.

In general, data analysis with such a model is performed in two steps. In the first one, the threshold, u, is chosen either graphically looking at the mean excess plot (Embrechts *et al.*, 1997) or simply setting it as some high percentile of the data (DuMouchel, 1983). Then, assuming that u is known, the other parameters are estimated, as suggested, for instance, in Smith (1987). The main drawback of this idea is that only the observations above the threshold are used in the second step. Moreover, the threshold selection is by no means an easy task as observed by Davison and Smith (1990) and Coles and Tawn (1994). If, on the one hand, a considerably high threshold is chosen in order to reduce the model bias, on the other hand, this would imply that only a few observations are used for estimating σ and ξ, thus increasing the variances of the estimates.

There is uncertainty in the choice of a threshold, u, even in the traditional theory to select it. As we said before, choosing the threshold through a mean excess plot or choosing a certain percentile does not guarantee that an appropriate selection was made in order to prevent model bias or violation of the independence condition of excess, which is crucial for the use of asymptotic distribution as a model. Most of the literature has shown how the threshold selection influences the parameter estimation (Coles and Powell, 1996; Coles and Tawn, 1996a; Coles and Tawn, 1996b; Frigessi, 2002a, Smith, 1987). We can see some examples where the variation in the estimates of σ and ξ given the selected u is significant and determines the fit of the model. Keeping this in mind we propose a model where we incorporate the uncertainty in the threshold selection by choosing a prior for u, possibly flat.

There have been different approaches proposed in the literature. Beirlant *et al.* (1996), for example, suggest an optimal threshold choice by minimizing bias variance of the model, whereas DuMouchel (1983) suggests the use of the upper 10% of the sample to estimate the parameters. In either of the methods, the estimates of σ and ξ depend significantly on the choice of the threshold. Mendes and Lopes (2004) propose a procedure to fit by maximum likelihood (ML) a mixture model where the tails are GPD and the center of the distribution is a normal. More recently, Frigessi *et al.* (2002b) have proposed a new dynamically weighted mixture model, where one of the terms is the GPD and the other one is a light tailed density function. They use the whole dataset for inference and use maximum likelihood estimation for the parameters in both distributions. However, they do not explicitly consider threshold selection. Bermudez *et al.* (2001) suggest an alternative method for threshold estimation by

choosing the number of upper order statistics. They propose a Bayesian predictive approach to the peaks over threshold (POT) method, extensively studied in the literature (Embrechts *et al.*, 1997). They treat the number of upper order statistics as another parameter in the model, with an appropriate prior distribution, and compute a weighted average over several possible values of the threshold using the predictive distribution avoiding, then, the problem of small sample sizes. They also approach the problem of threshold selection but they do it indirectly, by making inference about the number of order statistics beyond it. However, they do not consider a parametric model for observations below the threshold, only proceeding with simple nonparametric estimates for these data.

In this paper we propose a model to fit data characterized by extremal events where a threshold is directly estimated. The threshold is simply considered as another model parameter. More specifically, we estimate the threshold by proposing a parametric form to fit the observations below it and a GPD for the observations beyond it. It is recommended to have a robust model in order to fit several different situations, usually encountered in practice. It is important to analyse if the chosen form fits data from different distributions and influences the estimates of the threshold and the extreme parameters. All these aspects of robustness, goodness of fit and parameter estimation are treated in this paper.

Therefore, considering $X_1, X_2, \ldots, X_n$ independent and identically distributed observations and u the threshold over which these observations are considered exceedances, then we have $(X_i|X_i \geq u) \sim G(\cdot|\xi, \sigma, u)$. The observations below the threshold are distributed according to H, which can be estimated either parametrically or nonparametrically. In the parametric approach we can model the X_is below u assuming that H is any distribution like Weibull, gamma or normal. The normal distribution is specially used when one is interested in estimating both the lower and upper tails. In the nonparametric approach, mixtures of the parametric forms mentioned earlier provide a convenient basis for H.

Appropriate prior distributions are used for each of the model parameters. This includes the method suggested by Coles and Powell (1996) of eliciting information from experts to build the prior for the GPD parameters. As expected, posterior inference is analytically infeasible and Markov chain Monte Carlo (MCMC) methods are extensively applied, with particular emphasis on the Metropolis–Hastings and Gibbs types.

In the next section we will present the model that considers all the observations, below and above the threshold, in the estimation process. In Section 3 we discuss prior specification and use ideas of Coles and Tawn (1996a) for prior elicitation in the GPD context. A simulation study considering different scenarios is presented in Section 4, also, an analysis of robustness and goodness of fit of our model is included. In Section 5 we apply our approach to real data, the Nasdaq 100 index. The results are analogous to those obtained from the simulation study. We highlight the advantages of our Bayesian method and analyse the sensitivity of the parameter estimates to model selection. General discussion and ideas for future research conclude the paper in Section 6. In the appendix we present the MCMC algorithm for sampling from the posterior distribution along with other computational details.

2 Model

The proposed model assumes that observations under the threshold, u, come from a certain distribution with parameters η, denoted here $H(\cdot|\eta)$, whereas those above the threshold come from a GPD, as introduced in Equation (1.1). Therefore, the distribution function F, of any observation X, can be written as

$$F(x|\eta, \xi, \sigma, u) = \begin{cases} H(x|\eta), & x < u \\ H(u|\eta) + [1 - H(u|\eta)]G(x|\xi, \sigma, u), & x \geq u \end{cases} \tag{2.1}$$

For a sample of size n, $x = (x_1, \ldots, x_n)$ from F, parameter vector $\theta = (\eta, \sigma, \xi, u)$, $A = \{i: x_i < u\}$, and $B = \{i: x_i \geq u\}$, the likelihood function is

$$L(\boldsymbol{\theta}; \mathrm{x}) = \prod_A h(x|\eta) \prod_B (1 - H(u|\eta)) \left(\frac{1}{\sigma} \left[1 + \frac{\xi(x_i - u)}{\sigma} \right]_+^{-(1+\xi)/\xi} \right) \tag{2.2}$$

for $\xi \neq 0$, and $L(\boldsymbol{\theta}; \mathrm{x}) = \prod_A h(x|\eta) \prod_B (1 - H(u|\eta))((1/\sigma) \exp\{(x_i - u)/\sigma\})$, for $\xi = 0$.

Figure 1 represents the model schematically. As it can be seen the threshold u is the point where the density has a discontinuity. Depending on the parameters the density jump can be larger or smaller, and in each case the choice of which observations will be considered as exceedances can be more obvious or less evident. The smaller the jump the more difficult can be the estimation of the threshold. Fitting a nonparametric model to the data below the threshold allows smooth changes in the distribution around u. Strong discontinuities, or large jumps, indicate separation of the data. Consequently, it is expected that the parameter estimation would be easier. On the other hand, density

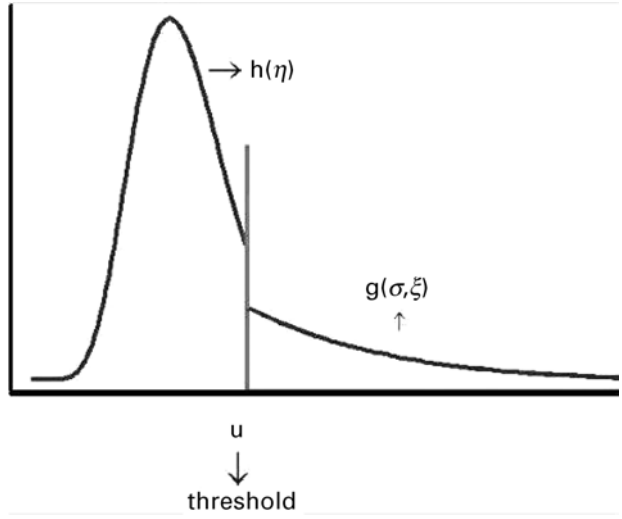

Figure 1 Schematic representation of the model

functions that are relatively smooth might represent an interesting challenge to our modeling structure. The parameters in our simulations were chosen in order to produce both situations.

As stated before, one goal of this work is to analyse whether the choice of the distribution for observations below the threshold influences, and how, the threshold estimation. In addition, we are interested in analysing whether the proposed model exhibits good data fitting when compared with other analyses presented in the literature.

Finally, it is worth mentioning that our mixture model can be extended to, for instance, a mixture of distributions below the threshold. In the next section we combine Equation (2.2) with a prior distribution for the parameters in order to enable one to perform posterior inference.

3 Prior elicitation and posterior inference

Recall that the parameters in the model are $\boldsymbol{\theta} = (\eta, u, \xi, \sigma)$. The prior distribution is now described.

3.1 Prior for parameters above threshold

In extreme value, analysis data are usually sparse, then information from experts can be useful to supplement the information from the data. It is reasonable to hope that experts should provide relevant prior information about extremal behavior, since they have specific knowledge of the characteristics of the data under study. Nonetheless, expressing prior beliefs directly in terms of GPD parameters is not an easy task. The idea we use here is from Coles and Tawn (1996a), Coles and Powell (1996) and Coles and Tawn (1996) and refers to the elicitation of information within a parameterization on which experts are familiar. More precisely, by the inversion of Equation (1.1), we obtain the $1 - p$ quantile of the distribution,

$$q = u + \frac{\sigma}{\xi}(p^{-\xi} - 1) \tag{3.1}$$

where q can be viewed as the return level associated with a return period of $1/p$ time units. The elicitation of the prior information is done in terms of (q_1, q_2, q_3) in the case of location scale parameterization of GPD, for specific values of $p_1 > p_2 > p_3$. Therefore, parameters are ordered and $q_1 < q_2 < q_3$. Therefore, Coles and Tawn suggest to work with the differences $d_i = q_i - q_{i-1}$, $i = 1, 2, 3$ with $q_0 = e_1$, where e_1 is the physical lower bound of the variable. They suggest setting $d_i \sim \text{Ga}(a_i, b_i)$ for $i = 1, 2, 3$. The case of $e_1 = 0$ is used in most applications. Independent prior distributions are assumed for the differences d_is. The prior information is elicited by asking the experts the median and 90% quantile (or any other) estimates for specific values of p that they are comfortable with. Usually, 10, 100 and 1000 time periods are considered, which correspond, respectively, to $p_1 = 0.1$, $p_2 = 0.01$ and $p_3 = 0.001$. After that, the elicited parameters are transformed to obtain the equivalent gamma parameters. For $i > 1$, neither d_i nor q_i depend on u. For $i = 1$, $p(d_1|u)$ was approximated by $(d_1|u*) \sim \text{Ga}(a_1(u^*), b_1(u^*))$ where u^* is the prior mean for u.

In the model proposed here, we are not considering the location parameter of GPD, so only two quantiles are needed in order to specify the GPD parameters, σ and ξ. Therefore, we have the following gamma distributions with known hyperparameters: $d_1 = q_1 \sim \text{Ga}(a_1, b_1)$ and $d_2 = q_2 - q_1 \sim \text{Ga}(a_2, b_2)$ The marginal prior distribution for σ and ξ is

$$\begin{aligned}\pi(\sigma, \xi) \propto & \left[u + \frac{\sigma}{\xi}(p_1^{-\xi} - 1)\right]^{a_1 - 1} \exp\left[-b_1\left\{u + \frac{\sigma}{\xi}(p_1^{-\xi} - 1)\right\}\right] \\ & \times \left[\frac{\sigma}{\xi}(p_2^{-\xi} - p_1^{-\xi})\right]^{a_2 - 1} \exp\left[-b_2\left\{\frac{\sigma}{\xi}(p_2^{-\xi} - p_1^{-\xi})\right\}\right] \\ & \times \left|-\frac{\sigma}{\xi^2}[(p_1 p_2)^{-\xi}(\log p_2 - \log p_1) - p_2^{-\xi}\log p_2 + p_1^{-\xi}\log p_1]\right| \qquad (3.2)\end{aligned}$$

where a_1, b_1, a_2 and b_2 are hyperparameters obtained from the experts information, for example in the form of the median and some percentile, corresponding to return periods of $1/p_1$ and $1/p_2$. The prior for q_1 should in principle depend on u. This would impose unnecessary complications in the prior form. In this paper, this dependence is replaced by dependence on the prior mean of u.

Some authors find that it is interesting to consider the situation where $\xi = 0$. In this case, we can set a positive probability to this point. The prior distribution would consider a probability q if $\xi = 0$ and $1 - q$ if $\xi \neq 0$, spreading the elicited prior shown above to this last case. From the computational point of view this model would not lead to any particular complications.

3.2 Prior for the threshold

There are many ways to set up a prior distribution for u. We can assume that u follows a truncated normal distribution with parameters (μ_u, σ_u^2), truncated from below at e_1 with density

$$\pi(u|\mu_u, \sigma_u^2, e_1) = \frac{1}{\sqrt{2\pi\sigma_u^2}} \frac{\exp\{-0.5(u - \mu_u)^2/\sigma_u^2\}}{\Phi[-(e_1 - \mu_u)/\sigma_u]} \qquad (3.3)$$

with μ_u set at some high data percentile and σ_u^2 large enough to represent a fairly noninformative prior.

This prior is used in the simulation study and the details are shown in the next section. A continuous uniform prior is another alternative. A discrete distribution can also be assumed. In this case, u could take any value between certain high data percentiles, which can be called hyperthresholds, as used in the application in Section 5. As the number of observations is usually high in applications, when the discrete prior is based on the observations the choice between discrete or continuous prior is immaterial for practical purposes.

One approach to the discrete prior for u is presented by Bermudez *et al.* (2001). They suggest threshold estimation by setting a prior distribution for the number of upper order statistics. In this case, the threshold is indirectly chosen and given by the data percentile corresponding to the number of exceedances. We could also have assumed

one more level to set the prior distribution for u, this would require setting a prior distribution for the hyperthresholds.

3.3 Prior for parameters below the threshold

The prior for the parameters η depends on the distribution chosen for data below u, $h(x|\eta)$. It is always better to try and obtain a conjugate prior to simplify the problem analytically. In a general way we assume $\eta \sim P$ with density π.

If the distribution $h(x|\eta)$ chosen is gamma, we have $\eta = (\alpha, \beta)$, α as the shape and β as scale parameter. But, instead of working with α and β, parameters of the gamma distribution, we reparameterize and think, in terms of prior specification, about α and $\mu = \alpha/\beta$. μ has a more natural interpretation; it is the prior expected value for the observational mean below the threshold. Also, it is more natural to assume prior independence between the shape parameter and the mean. We then set, $\alpha \sim \text{Ga}(a, b)$ and $\mu \sim \text{Ga}(c, d)$, where a, b, c and d are known hyperparameters. The joint prior of $\eta = (\alpha, \beta)$ will then be

$$\pi(\eta) = \frac{b^a}{\Gamma(a)}\alpha^{a-1}\,\mathrm{e}^{-b\alpha}\,\frac{d^c}{\Gamma(c)}\left(\frac{\alpha}{\beta}\right)^{c-1}\mathrm{e}^{-d\alpha/\beta}\left(\frac{\alpha}{\beta^2}\right) \tag{3.4}$$

3.4 Posterior inference

From the likelihood [Equation (2.2)] and the prior distributions specified earlier, we can use Bayes theorem to obtain the posterior distribution which has the following: taking a gamma distribution for data below the threshold, functional form, on the logarithm scale

$$\begin{aligned}
\log p(\theta|x) = {} & K + \sum_{i=1}^{n} I(x_i < u)[\alpha\log\beta - \log\Gamma(\alpha) + (\alpha-1)\log x_i - \beta x_i] \\
& + \sum_{i=1}^{n} I(x_i \geq u)\log\left[1 - \int_0^u \frac{\beta^\alpha}{\Gamma(\alpha)} t^{\alpha-1}\,\mathrm{e}^{-\beta t}\,\mathrm{d}t\right] - \sum_{i=1}^{n} I(x_i \geq u)\log\sigma \\
& - \frac{1+\xi}{\xi}\sum_{i=1}^{n} I(x_i \geq u)\log\left[1 + \frac{\xi(x_i - u)}{\sigma}\right] \\
& + (a-1)\log\alpha - b\alpha + (c-1)\log\left(\frac{\alpha}{\beta}\right) - d\left(\frac{\alpha}{\beta}\right) + \log\left(\frac{\alpha}{\beta^2}\right) \\
& - \frac{1}{2}\left(\frac{u-\mu_u}{\sigma_u}\right)^2 - b_1\left\{u + \frac{\sigma}{\xi}(p_1^{-\xi} - 1)\right\} \\
& + (a_2 - 1)\log\left[u + \frac{\sigma}{\xi}\left(p_2^{-\xi} - p_1^{-\xi}\right)\right] - b_2\left\{u + \frac{\sigma}{\xi}(p_2^{-\xi} - p_1^{-\xi})\right\} \\
& + \log\left|-\frac{\sigma}{\xi^2}\left[(p_1p_2)^{-\xi}(\log p_2 - \log p_1) - p_2^{-\xi}\log p_2 + p_1^{-\xi}\log p_1\right]\right|
\end{aligned} \tag{3.5}$$

where K is the normalizing constant. It is clear that this posterior distribution has no known closed form distribution making analytical posterior inference infeasible. Note that the posterior mentioned earlier is shown with a normal prior for the threshold and with the likelihood for the case where $\xi \neq 0$. However, the case where $\xi = 0$ is also considered in the algorithm used in the applications.

The computation is done through the MCMC methods, via Metropolis steps within a blockwise algorithm, which is described in the appendix. We can either sample θ at once, or break it into smaller blocks to be drawn from. It will depend on the convergence rate in each case. Because of the features of the model, we are drawing the shape parameter ξ of GPD first, since the scale parameter σ and the threshold depend on its sign. If ξ is negative, σ and u have restrictions as one can see in the definition of the GPD distribution [Equation (1)]. Following ξ, σ and u are drawn individually and in this order. Lastly, η is jointly drawn. In the case of gamma parameters, $\eta = (\alpha, \beta)$. The use of the Metropolis–Hastings algorithms requires the specification of candidate distributions for the parameters.

4 A simulation study

We entertained a wide range of scenarios, focusing on generating skewed and heavy tailed distributions. For the sake of space, only a small but revealing fraction of them is presented here, with further details directed to Behrens *et al.* (2002) (BLG, hereafter). The parameters used for the simulations presented here are $p = 0.1$, $\alpha = (0.5, 1, 10)$, $\xi = (-0.1, -0.45, 0.2)$ and $n = (1000, 10\,000)$. Also, the scale parameters β and σ were kept fixed ($1/\beta = \sigma = 5$), since their changes do not influence the estimation. The sample size n and p automatically define the value of u. We chose $\xi = -0.45$ for generating lighter tails and for avoiding unstable ML estimation, whereas $\xi = 0.2$ generates heavier tails. Table 1 summarizes our findings based on the 18 datasets, whereas Figure 2 shows the histograms of the marginal distributions of the model parameters when $\alpha = 1.0$ and $\xi = -0.45$. As one would expect, for all entertained datasets, the 95% credible posterior intervals contain the true values. Similar results were found when $p = 0.01$ and $p = 0.001$ (BLG).

It is important to verify if observations from other distributions, different from gamma, are well fitted by our model. Some variations using Weibull data for the center of the distribution were considered and GPD results were not affected. Despite the similarities between the distributions, these results tentatively point to robustness of the models proposed here.

5 Modeling the Nasdaq 100 index

The next step will be the application of the model to real data and to analyse how the methodology performs in different situations in different fields. We now apply our approach to Nasdaq 100, an index of financial market, from January 1985 to May 2002 ($N = 4394$). The dataset was chosen given its importance to financial market and

Table 1 Simulated data – true values, posterior mean (PM), and 95% credibility interval for $p=0.1$ with $n=10.000$ and $n=1.000$

	ALPHA		BETA		U		SIGMA		KSI	
Dataset	True value	PM (CI-95%)	True value	PM (CI-95%)	True value	PM (CI-95%)	True value	PM (CI-95%)	True value	PM (CI-95%)
p=0.1, n=10.000										
1	0.5	0.50 (0.49–0.52)	0.2	0.20 (0.19–0.21)	6.83	6.41 (5.84–6.98)	5	5.41 (4.94–5.87)	−0.1	−0.15 (−0.20–−0.10)
2	1	0.99 (0.96–1.02)	0.2	0.20 (0.19–0.21)	11.48	14.05 (4.16–23.94)	5	4.79 (3.73–5.86)	−0.1	−0.11 (−3.21–2.99)
3	10	10.39 (10.09–10.69)	0.2	0.21 (0.20–0.21)	70.21	70.20 (70.15–70.24)	5	5.04 (4.62–5.46)	−0.1	−0.09 (−0.15–−0.04)
4	0.5	0.50 (0.49–0.52)	0.2	0.20 (0.19–0.21)	6.64	7.78 (5.11–10.44)	5	4.56 (3.11–6.02)	−0.45	−0.47 (−3.47–2.54)
5	1	0.99 (0.97–1.02)	0.2	0.20 (0.19–0.21)	11.40	10.50 (7.62–13.37)	5	5.03 (3.97–6.09)	−0.45	−0.43 (−0.48–−0.37)
6	10	10.18 (9.92–10.45)	0.2	0.20 (0.20–0.21)	70.73	70.72 (70.59–70.84)	5	4.94 (4.57–5.32)	−0.45	−0.45 (−0.50–−0.40)
7	0.5	0.50 (0.49–0.52)	0.2	0.21 (0.19–0.22)	6.64	6.76 (6.56–6.95)	5	5.42 (4.80–6.03)	0.2	0.14 (0.04–0.23)
8	1	0.99 (0.97–1.02)	0.2	0.20 (0.19–0.21)	11.40	14.88 (9.35–20.40)	5	5.60 (3.94–7.25)	0.2	0.16 (0.04–0.28)
9	10	10.13 (9.83–10.43)	0.2	0.20 (0.20–0.21)	70.73	70.71 (70.60–70.83)	5	4.90 (4.41–5.40)	0.2	0.19 (0.11–0.26)
p=0.1, n=1.000										
10	0.5	0.56 (0.52–0.59)	0.2	0.20 (0.18–0.23)	6.61	11.29 (4.40–18.19)	5	4.00 (1.95–6.05)	−0.1	−0.05 (−0.27–0.18)
11	1	1.01 (0.86–1.17)	0.2	0.20 (0.11–0.28)	11.53	13.02 (−1.52–27.56)	5	4.18 (1.59–6.77)	−0.1	−0.08 (−0.32–0.16)
12	10	10.18 (9.08–11.27)	0.2	0.21 (0.18–0.23)	69.13	70.39 (66.86–73.9)	5	5.64 (3.15–8.13)	−0.1	−0.17 (−0.42–0.09)
13	0.5	0.49 (0.45–0.53)	0.2	0.19 (0.17–0.22)	6.85	8.52 (4.46–12.58)	5	4.24 (1.92–6.56)	−0.45	−0.45 (−0.73–−0.17)
14	1	1.03 (0.95–1.11)	0.2	0.21 (0.19–0.24)	11.85	8.67 (4.75–12.58)	5	6.29 (4.46–8.12)	−0.45	−0.37 (−0.52–−0.22)
15	10	9.30 (8.55–10.05)	0.2	0.19 (0.17–0.20)	70.89	70.58 (69.63–71.53)	5	5.18 (3.68–6.67)	−0.45	−0.40 (−0.58–−0.22)
16	0.5	0.50 (0.44–0.56)	0.2	0.22 (0.14–0.29)	6.85	2.77 (−2.79–8.32)	5	3.56 (1.33–5.79)	0.2	0.27 (0.09–0.44)
17	1	1.02 (0.92–1.12)	0.2	0.21 (0.17–0.24)	11.85	9.51 (2.03–17.00)	5	5.33 (2.60–8.06)	0.2	0.31 (0.16–0.46)
18	10	9.31 (8.43–10.20)	0.2	0.19 (0.17–0.21)	70.89	70.64 (69.79–71.48)	5	5.13 (3.44–6.82)	0.2	0.24 (0.07–0.42)

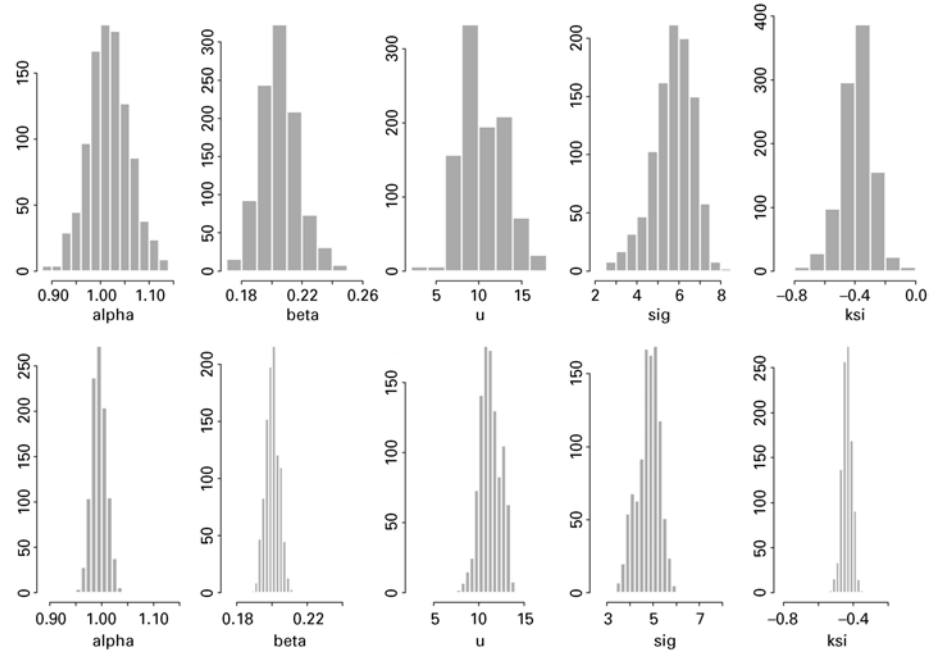

Figure 2 Marginal posterior histograms of model parameters based on data generated from $p = 0.1$ $\alpha = 1$, $\beta = 0.2$, $\sigma = 5$ and $\xi = -0.45$. Top row: $n = 1000$ ($u = 11.85$). Bottom row: $n = 10\,000$ ($u = 11.4$)

the presence of many extreme events and it was taken from Yahoo financial site – http://finance.yahoo.com/q?d=t&s=^IXIC. The original data, daily close index, is converted to daily increments in the following way:

$$y_t = 100\left|\frac{P_t}{P_{t-1}} - 1\right|$$

Absolute values are used since financial datasets usually exhibit clusters of high volatility, caused by either positively or negatively large returns. Both positive and negative large returns are important in most practical volatility evaluations by risk analysts. The usual treatment involves removal of these temporal dependences through time varying volatility formulations. Our interest here, however, is to concentrate on large values of returns and therefore we did not perform any such standardization to the data. Figure 3 displays a histogram of the data. As we can see there is indication of heavy tailed data. Our main goal is to compare the results obtained by our model with those obtained using a ML approach. Also, we want to test the efficiency of our method in the extrapolation issue. The model used in this application uses a gamma distribution to fit the data below the threshold.

A descriptive analysis is presented below in order to get more feeling about the data behavior. For the ease of notation let N be the sample size, $[x]$ is the integer part of the number x, and $y_{(i)}$ is the ith order statistic of data $(y_1, \ldots, y_N)$, such that $y_{([pN])}$ is the $100p$ data percentile, for example, $y_{([0.70N])}$ is the 70% data percentile.

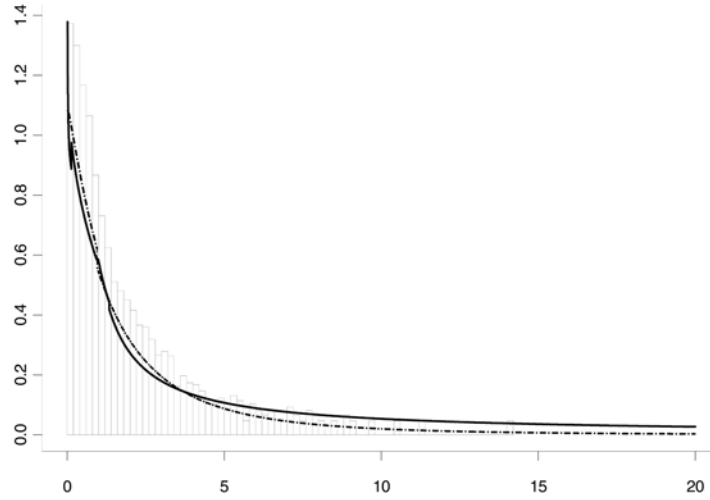

Figure 3 Histogram of the data and predictive distributions: solid line, fully Bayesian and dashed line, approximate Bayesian

Table 2 shows the ML estimator for σ and ξ considering different values of u. There are no important changes in the estimates of σ and ξ as the value of u is changed. We cannot observe any pattern in the ξ estimates with changes in the number of exceedances considered. When the number of exceedances is around 2% of the data, the estimates of σ and ξ become very unstable. We have also calculated the conditional Bayes estimators for the extreme parameters considering the same values of u used to obtain the ML estimators. As we can expect, a small increase in the posterior mean of σ is observed since the greater value of u implies less exceedances or less data points to estimate σ (its variance increases with u). The results are in Table 2 and we can see that the estimate of ξ are consistent if compared with its Bayesian estimate. For the other values we observe an increase in the posterior mean for the scale parameter, since we have more uncertainty incorporated in the model. Also the credibility intervals are larger for both extreme parameters.

Table 2 Summary of extreme parameter estimators: posterior means and 95% credibility intervals, in brackets, of u, σ, and ξ; ML estimators for σ and ξ for different values of $u = y_{([pN])}$

	p	u	σ	ξ
Classical	0.5	0.57	0.6305	0.2396
	0.7	0.93	0.8176	0.1466
	0.9	2.11	0.7435	0.1689
	0.95	2.92	0.7171	0.1786
Bayesian conditional on u	0.5	0.57	0.7480 [0.07; 0.80]	0.2404 [0.18; 0.30]
	0.7	0.93	0.9579 [0.87; 1.04]	0.1609 [0.09; 0.23]
	0.9	2.11	1.0827 [0.92; 1.24]	0.1945 [0.08; 0.32]
	0.95	2.92	1.1869 [0.02; 1.46]	0.2297 [0.03; 0.42]
Bayesian	–	0.9619 [0.79; 1.13]	0.9735 [0.86; 1.08]	0.1567 [0.09; 0.23]

A bivariate analysis of σ and ξ is also performed based on the likelihood and posterior distributions to analyse correlation between blocks of parameters. We have taken the conditional distributions considering different values of u, $y_{([0.5N])}$, $y_{([0.7N])}$, $y_{([0.9N])}$, $y_{([0.95N])}$, and the ML estimators for σ and ξ and the moment estimators for α and β. Conditional on u, the vector (α, β) is independent of (σ, ξ). The values of α and β that maximize the likelihood function are not much affected by changes in the threshold. Only the scale parameter, β, presents a small variation since the number of observations used to estimate it changes with u. The same happens when we look at the conditional likelihood of σ and ξ. Similarly, the values of σ and ξ that maximize the conditional likelihood are close to the ML estimators shown in Table 2.

Flat priors were considered for α, β, σ and ξ, and hyperparameters were calculated as described in Section 3. The chosen values were $a_1 = 0.1$, $b_1 = a_1/19.8$, $a_2 = 0.9$ and $b_2 = a_2/29.8$. A uniform discrete prior was assumed for the threshold u and the values of the hyperparameters are described in the appendix.

Conditional on u, σ and ξ, the values of α and β that maximize the posterior are close to those in the conditional likelihood. The results for σ and ξ are also analogous to those shown in the likelihood analysis. A slight difference can be noticed when the threshold chosen is $y_{([0.5N])}$.

On the basis of the graphs in Figure 3, the initial values to start the chains were chosen and this is described in the appendix. The posterior mean of α and β, 1.0202 and 1.2816, respectively, are very close to the moment estimators. The posterior mean and variance of σ and ξ are shown in Table 2. As we said above, we have chosen two similar discrete prior distributions to perform the analysis, and since we also got similar results with both cases only the second one is shown in Table 2.

Convergence was achieved after few iterations and Figure 4 presents the histograms of the distributions of each parameter. The parameter ξ has a distribution centered in the ML estimator and α and β centered in the moment estimators. The Bayesian approach shows a larger estimate for σ than the classical one, whereas the credibility interval of ξ includes its in estimate. The distribution of the threshold, u, seems to be bimodal, one of the modes being highly concentrated around $y_{([0.58N])}$ and the second mode concentrated around $y_{([0.9027N])}$, so the posterior mean is $y_{([0.7586N])}$. The first mode has probability 0.67 around it and the second mode has probability 0.10 around it.

In order to observe how the model fits the data and to analyse the behavior of the model for future observations, we computed the predictive distribution. Figure 3 shows the predictive distributions superimposing the histogram of the data. The solid line is the Bayesian predictive distribution, $p(y|\text{data}) = \int p(y|\theta)\, p(\theta|\text{data})\, \mathrm{d}\theta$, and the dashed line is an approximate Bayesian (AB) approach, $p(y|\text{data}) \doteq p(y|\tilde{\theta})$ where $\tilde{\theta} = E(\theta|\text{data})$, which corresponds to concentrating all the information in the posterior mean, $\tilde{\theta}$, a reasoning similar to that used for classical prediction. We can see that the difference between the two approaches is not so significant in the center of the distribution, whereas the Bayesian approach gives higher probabilities in the tail. The AB predictive distribution underestimates the probabilities for events considered extremes.

In general terms, the results show that the estimated extreme quantiles obtained from the fully Bayesian (FB) predictive distribution seem to be more conservative than the ones produced by using plug-in estimation such as the AB or the classical approaches.

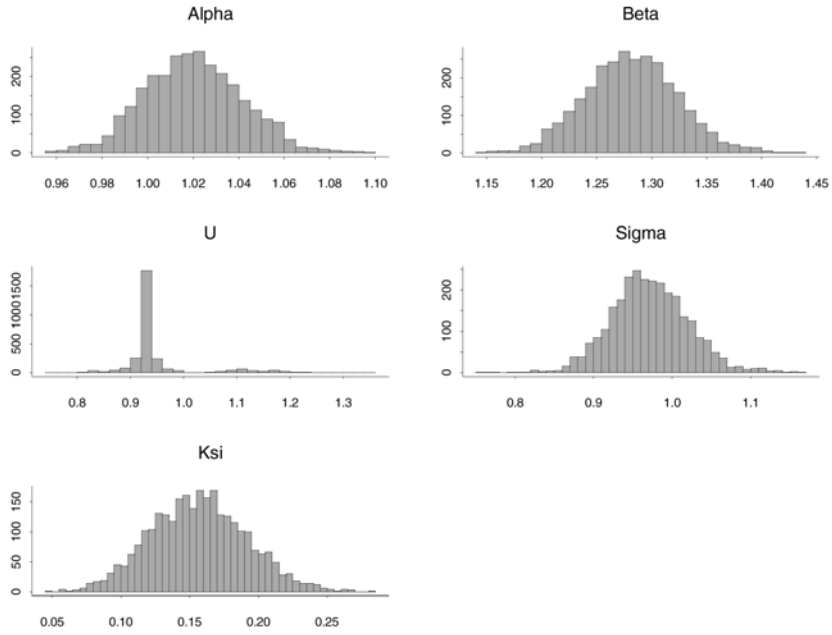

Figure 4 Nasdaq 100 – histograms of the marginal posterior distribution

For instance, in Table 3 we can see that $P(X > 5.35) = 0.01$, which means that an extreme event higher than 5.35, occurs, on the average, once in five months using the AB approach, since our data are taken daily. If we look at the FB estimates, we have $P(X > 5.35) = 0.04$, which means that an extreme event higher than 5.35 only takes 1.25 months to occur on the average. In a decision making setting, the FB approach represents one's risk averse behavior. This is caused by the incorporation of the uncertainty about the parameters of the model. This aspect of the Bayesian approach has already been noted by other authors. Coles and Pericchi (2003) showed that this leads to more sensible solutions to real extremes data problems than plug-in estimation.

Figure 5 shows the return levels associated with return periods from one week to one year. The shorter return periods in the Bayesian approach, associated with any given return level, are a direct consequence of the thicker tail observed in Figure 3. Again, we can see that the FB approach is more conservative than the classical and AB approaches.

Table 3 Extreme tail probability using the empirical data distribution. FB predictive distribution and AB predictive distribution

Quantiles	Empirical	FB	AB
2.11	0.9	0.86	0.897
2.92	0.95	0.91	0.947
5.35	0.99	0.96	0.99
9.00	0.999	0.98	0.998

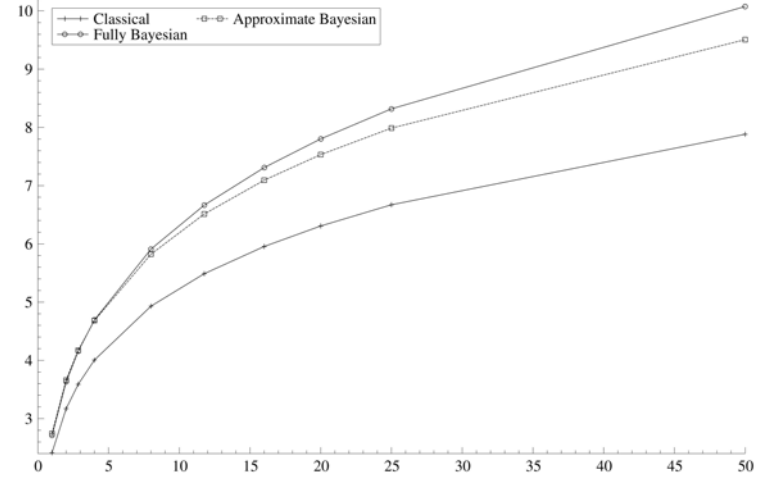

Figure 5 Return level associated with return period from one week to one year. The parameter values used to calculate the return level in the classical approach refer to those where $u = 0.93 = y_{([0.7N])}$.

An extreme return level in the FB approach takes less time to occur on average than the same return level considering the other two approaches.

6 Conclusions

In this paper we suggest an alternative to the usual analysis of extreme events. Inference is based on a mixture model with gamma distribution for observations below a threshold, and a GPD for observations above it. All observations are used to estimate the parameters present in the model, including the threshold.

Different approaches have been tried in the literature, but in none of them is the threshold treated as a parameter in the estimation process. The available methods choose the threshold empirically, even those using Bayesian methodology.

A simulation study was performed and the results have shown that we obtained good estimates of the parameters. In spite of this fact, the threshold was at times hard to estimate, especially when the sample size was not large enough. In general, the gamma parameters converged very fast, whereas the GPD parameters, σ and ξ, and u needed more iterations. These last three parameters sometimes demonstrated a strong correlation between their chains, but this did not affect the convergence.

This wide range of scenarios allowed us to analyse the behavior of posterior densities under different situations. It seems that the shape of the distribution is not a problem in parameter estimation. For sufficiently large samples, parameter estimates were very close to the true value even for data which were strongly skewed and/or did not have a smooth density function. The problems with convergence, and consequently with parameter estimation, arise when the number of observations is small. The proposed model here, although simple, is able to fit different situations. Results obtained applying our gamma GPD model to data simulated from other distributions showed a good performance.

Classical methods for analysing extreme events require a good choice of the threshold to work well. Our proposed method avoids this problem while considering the threshold as a parameter of the model and allows prior information to be incorporated into the analysis.

We could compare the results obtained here with those using a classical approach by applying the proposed model to real data. An important issue here is the threshold being included as a parameter in the model. The results show close estimates for GPD parameters if we compare classical and Bayesian approaches with vague prior distributions. Only the scale parameter, σ, shows a significant difference, which is explained by the fact that we have more uncertainty incorporated into the model in the Bayesian approach. This difference can also be observed when we compare the predictive distributions, where the FB inference seems to fit better to more extreme data than the AB approach.

It is also interesting to look at the return level associated with a return period of $1/p$ units of time. We show the plot of these values for a range of return period from one week to one year and results has shown that the FB approach is, again, more conservative than the classical and AB approaches. This means that a certain return level is associated with a shorter return period in the AB approach.

A similar methodology considering other parametric and nonparametric forms for the distribution of the observations below the threshold can also be considered. Tancredi *et al.* (2002) tackles the threshold problem in a similar but independent way. They model the non-extreme data (below threshold) by a mixture of uniforms. The use of other parametric forms, like Student's t distribution, will allow the estimation of both tails, as needed in many financial and insurance data, where interest lies in the estimation not only on the large claims (or gains) but mainly on large losses. Also, a more exhaustive study about other distributions below the threshold should be performed.

Acknowledgements

We have benefited from invaluable discussions with Professor Richard Smith and Havard Rue. We also thank Professor Bruno Sansó and Professor Josemar Rodrigues for their comments. This research project was partially supported by CAPES Foundation and CNPq.

References

Behrens C, Lopes HF and Gamerman D (2002) Bayesian analysis of extreme events with threshold estimation. Technical Report, Laboratorio de Estatistica, Universidade Federal do Rio de Janeiro.

Beirlant J, Vynckier P and Teugels JL (1996) Excess functions and estimation of the extreme-value index. *Bernoulli* **2**, 293–318.

Bermudez PZ, Turkman MAA and Turkman KF (2001) A predictive approach to tail probability estimation. *Extremes* **4**, 295–314.

Coles SG and Pericchi LR (2003) Anticipating catastrophes through extreme value modelling. *Applied Statistics* **52**, 405–16.

Coles SG and Powell EA (1996) Bayesian methods in extreme value modelling: a review and new developments. *International Statistical Review* **64**, 119–36.

Coles SG and Tawn JA (1991) Modelling extreme multivariate events. *Journal of the Royal Statistical Society B* **53**, 377–92.

Coles SG and Tawn JA (1994) Statistical methods for multivariate extremes: an

application to structural design. *Applied Statistics* **43**, 1–48.
Coles SG and Tawn JA (1996a) A Bayesian analysis of extreme rainfall data. *Applied Statistics* **45**, 463–78.
Coles SG and Tawn JA (1996b) Modelling extremes of the areal rainfall process. *Journal of the Royal Statistical Society B* **58**, 329–47.
Davison AC and Smith RL (1990) Models for exceedances over high thresholds. *Journal of the Royal Statistical Society B* **52**, 393–442 (with discussion.)
Doornik JA (1996) *Ox: Object Oriented Matrix Programming, 3.1 console version*. London: Nuffield College, Oxford University.
DuMouchel WH (1983) Estimating the stable index α in order to measure tail thickness: a critique. *The Annals of Statistics* **11**, 1019–31.
Embrechts P, Klüppelberg C and Mikosch T (1997) *Modelling extremal events for insurance and finance*. New York: Springer.
Embrechts P, Resnick S and Samorodnitsky G (1999) Extreme value theory as a risk management tool. *North American Actuarial Journal* **3**, 30–41.
Frigessi A, Haug O and Rue H (2002a) Tail estimation with generalized Pareto distribution without threshold selection. Unpublished: University of Oslo, Norway.
Frigessi A, Haug O and Rue H (2002b) A dynamic mixture model for unsupervised tail estimation without threshold selection. *Extremes* **5**, 219–35.
Gamerman D (1997) *Markov Chain Monte Carlo: stochastic simulation for Bayesian inference*. London: Chapman & Hall.
Gelman A and Rubin DB (1992) Inference from iterative simulation using multiple sequences. *Statistical Science* **7**, 457–511.
Geweke J (1992) Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian statistics 4*, (Bernardo JM, Berger JO, Dawid AP, Smith AFM). Oxford: Oxford University Press, UK, 169–94.
Heidelberger P and Welch P (1983). Simulation run length control in the presence of an initial transient. *Operations Research* **31**, 1109–44.
Mendes B and Lopes HF (2004) Data driven estimates for mixtures. *Computational Statistics and Data Analysis* (in press).
Pickands J (1975) Statistical inference using extreme order statistics. *Annals of Statistics* **3**, 119–131.
Robert CP and Casella G (1999) *Monte Carlo statistical methods*. New York: Springer.
Smith RL (1987) Estimating tails of probability distributions. *Annals of Statistics* **15**, 1174–1207.
Smith RS (2000) Measuring risk with extreme value theory. In Embrechts P ed. *Extremes and integrated risk management*. London: Risk Book, 19–35.
Smith BJ (2003) *Bayesian output analysis program (BOA) version 1.0*. College of Public Health, University of Iowa, USA.
Tancredi A, Anderson C and O'Hagan A (2002) *Accounting for threshold uncertainty in extreme value estimation*. Technical Report. Italy: Dipartimento di Scienze Statistiche, Università di Padova.

Appendix

In this appendix we describe the MCMC algorithm used to make approximate posterior inference and also the implementation details. Gamerman (1997) and Robert and Casella (1999) are comprehensive references on this subject. We have used the Ox language when developing the MCMC algorithms (Doornik, 1996). It took us about 2 h, on average, to run 10 000 iterations using a Pentium III PC with 833 MHz.

Algorithm

Simulations are done via Metropolis–Hastings steps within blockwise MCMC algorithm. Therefore, candidate distributions for the parameters evolution must be speci-

fied. The candidate distributions used in the algorithm as well as the steps of the algorithm are presented. Suppose that at iteration j, the chain is positioned at $\boldsymbol{\theta}^{(j)} = (\alpha^{(j)}, \beta^{(j)}, u^{(j)}, \sigma^{(j)}, \xi^{(j)})$. Then, at iteration $j+1$ the algorithm cycles through the following steps.

Sampling ξ

ξ^* is sampled from a $N(\xi^{(j)}, V_\xi)I(-\sigma^{(j)}/(M-u^{(j)}), \infty)$ distribution, where V_ξ is an approximation based on the curvature at the conditional posterior mode, and $M = \max(x_1, \ldots, x_n)$. Therefore, $\xi^{(j+1)} = \xi^*$ with probability α_ξ where

$$\alpha_\xi = \min\left\{1, \frac{p(\theta^*|x)}{p(\tilde{\theta}|x)} \frac{\Phi((\xi^{(j)} + \sigma^{(j)}/(M-u^{(j)}))/\sqrt{V_\xi})}{\Phi((\xi^* + \sigma^{(j)}/(M-u^{(j)}))/\sqrt{V_\xi})}\right\}$$

for $\boldsymbol{\theta}^* = (\alpha^{(j)}, \beta^{(j)}, u^{(j)}, \sigma^{(j)}, \xi^*)$, $\tilde{\theta} = \theta^{(j)}$ and $\Phi(\cdot)$ is the standard normal's cumulative distribution function.

Sampling σ

If $\xi^{(j+1)} \geq 0$, σ^* is sampled from a $\mathrm{Ga}(a_j, b_j)$ distribution, where $a_j/b_j = \sigma^{(j)}$ and $a_j/b_j^2 = V_\sigma$. If $\xi^{(j+1)} < 0$, σ^* is sampled from a $N(\sigma^{(j)}, V_\sigma)I(-\xi^{(j+1)}(M-u^{(j)}), \infty)$ distribution, where V_σ is an approximation for the concavity in the conditional posterior mode. Therefore, $\sigma^{(j+1)} = \sigma^*$ with probability α_σ where

$$\alpha_\sigma = \min\left\{1, \frac{p(\theta^*|x)}{p(\tilde{\theta}|x)} \frac{g(\sigma^{(j)}|a_j, b_j)}{g(\sigma^*|a^*, b^*)}\right\}$$

if $\xi^{(j+1)} \geq 0$, and

$$\alpha_\sigma = \min\left\{1, \frac{p(\theta^*|x)}{p(\tilde{\theta}|x)} \frac{\Phi((\sigma^{(j)} + \xi^{(j+1)}(M-u^{(j)}))/\sqrt{V_\sigma})}{\Phi((\sigma^* + \xi^{(j+1)}(M-u^{(j)}))/\sqrt{V_\sigma})}\right\}$$

if $\xi^{(j+1)} < 0$, where $\theta^* = (\alpha^{(j)}, \beta^{(j)}, u^{(j)}, \sigma^*, \xi^{(j+1)})$, $a^*/b^* = \sigma^*$, $a^*/b^{*2} = V_\sigma$, and $\tilde{\theta} = (\alpha^{(j)}, \beta^{(j)}, u^{(j)}, \sigma^{(j)}, \xi^{(j+1)})$.

Sampling u

The threshold parameter u^* is sampled from a $N(u^{(j)}, V_u)I(a^{(j+1)}, M)$ distribution with $a^{(j+1)} = \min(x_1, \ldots, x_n)$, if $\xi^{(j+1)} \geq 0$, and $a^{(j+1)} = M + \sigma^{(j+1)}/\xi^{(j+1)}$, if $\xi^{(j+1)} < 0$. Again,

V_u is a value for the variance which is tuned to allow appropriate chain movements. Therefore, $u^{(j+1)} = u^*$ with probability α_u where

$$\alpha_u = \min\left\{1, \frac{p(\theta^*|x)}{p(\tilde{\theta}|x)} \frac{\Phi((M-u^{(j)})/\sqrt{V_u}) - \Phi((a^{(j+1)}-u^{(j)})/\sqrt{V_u})}{\Phi((M-u^*)/\sqrt{V_u}) - \Phi((a^{(j+1)}-u^*)/\sqrt{V_u})}\right\}$$

for $\boldsymbol{\theta}^* = (\alpha^{(j)}, \beta^{(j)}, u^*, \sigma^{(j+1)}, \xi^{(j+1)})$ and $\tilde{\theta} = (\alpha^{(j)}, \beta^{(j)}, u^{(j)}, \sigma^{(j+1)}, \xi^{(j+1)})$.

In the case where u has a discrete prior we have to follow the same model restrictions as before. Then, the candidate distribution is: If $\xi^{(j+1)} \geq 0$, $u^* \sim U_d(q_1, q_2)$, a discrete uniform distribution on data quantiles from q_1 to q_2, where q_2 can be any high quantile as, for instance, M, whereas q_1 can be any quantile, as long as $q_1 < q_2$. It is important to keep in mind that q_1 must be small enough to reduce model bias and large enough to respect the asymptotic properties of the model. Analogously, if $\xi^{(j+1)} < 0$, $u^* \sim U_d(q_1, q_2)$, with $q_1 \geq M + \sigma^{(j+1)}/\xi^{(j+1)}$. In the discrete case, $\alpha_u = \min\{p(\theta^*|x)/p(\tilde{\theta}|x)\}$.

Sampling α and β

α^* and β^* are sampled, respectively, from a $\log N(\alpha^{(j)}, V_\alpha)$ and a $GI(a_j, b_j)$ distributions, with $a_j/b_j = \beta^{(j)}$ and $a_j/b_j^2 = V_\beta$. V_α and V_β are approximations for the curvatures at the conditional posterior modes. Therefore, $(\alpha^{(j+1)}, \beta^{(j+1)}) = (\alpha^*, \beta^*)$ with probability

$$\min\left\{1, \frac{p(\theta^*|x)}{p(\tilde{\theta}|x)} \frac{h(\alpha^{(j)}|\alpha^*, V_\alpha)}{h(\alpha^*|\alpha^{(j)}, V_\alpha)} \frac{g(\beta^{(j)}|a_j, b_j)}{g(\beta^*|a^*, b^*)}\right\}$$

for $\boldsymbol{\theta}^* = (\alpha^*, \beta^*, u^{(j+1)}, \sigma^{(j+1)}, \xi^{(j+1)})$, $\tilde{\theta} = (\alpha^{(j)}, \beta^{(j)}, u^{(j+1)}, \sigma^{(j+1)}, \xi^{(j+1)})$, $a^*/b^* = \beta^*$, $a^*/b^{*2} = V_\beta$, $h(\cdot|c, d)$ is the lognormal density with mean c and variance d, and $g(\cdot|c, d)$ is the gamma density with parameters c and d.

Implementation

Our implementation involved running a few parallel chains starting from different regions of the parameter space. The first few draws from the chains were used for tuning V_α, V_β, V_σ and V_ξ, the variances of the candidate distributions. Convergence was checked by comparing the marginal distributions of the parameters obtained from the parallel chains and by application of standard tests (Gelman and Rubin, 1992; Geweke, 1992; Heidelberger and Welch, 1983) using the Bayesian Output Analysis Program (Smith, 2003). Values from the chains were merged for posterior inference.