A DYNAMIC MIXTURE MODEL FOR UNSUPERVISED TAIL ESTIMATION WITHOUT THRESHOLD SELECTION

ARNOLDO FRIGESSI
NORWEGIAN COMPUTING CENTER
OSLO, NORWAY

OLA HAUG
NORWEGIAN COMPUTING CENTER
OSLO, NORWAY

HÅVARD RUE
DEPARTMENT OF MATHEMATICAL SCIENCES
NTNU, NORWAY

NOVEMBER 28, 2002

SUMMARY

Exceedances over high thresholds are often modeled by fitting a Generalized Pareto distribution (GPD) on $\mathbb{R}^+$. It is difficult to select the threshold, above which the GPD assumption is enough solid and enough data is available for inference. We suggest a new dynamically weighted mixture model, where one term of the mixture is the GPD, and the other is a light-tailed density distribution. The weight function varies on $\mathbb{R}^+$ in such a way that for large values the GPD component is predominant and thus takes the role of threshold selection. The full data set is used for inference on the parameters present in the two component distributions and in the weight function. Maximum likelihood provides estimates with approximate standard deviations. Our approach has been successfully applied to simulated data and to the (previously studied) Danish fire loss data set. We compare the new dynamic mixture method to Dupuis' robust thresholding approach in Peaks-Over-Threshold inference. We discuss robustness with respect to the choice of the light-tailed component and the form of the weight function. We present encouraging simulation results that indicate that the new approach can be useful in unsupervised tail estimation, especially in heavy tailed situations and small percentiles.

KEYWORDS: Peaks Over Threshold; Danish Fire Loss Data; Shape Parameter; Heavy Tailed Distribution; Maximum Likelihood; Mixture Models.

ADDRESSES: A. Frigessi and O. Haug Norwegian Computing Centre, P.O.Box 114 Blindern, N-0314 Oslo, Norway. H. Rue, Department of Mathematical Sciences, The Norwegian University for Science and Technology, N-7491 Trondheim, Norway.

E-MAIL: Arnoldo.Frigessi@nr.no, Ola.Haug@nr.no and Havard.Rue@math.ntnu.no

ACKNOWLEDGMENTS: We thank Peter J. Green (University of Bristol) for inspiring conversations that led to the formulation of the functional mixture model, Holger Rootzen (Chalmers University), Jon Gjerde (Norwegian Computing Centre),

and Jonathan Tawn (University of Lancaster) for discussions, and Debbie Dupuis (Dalhouse University) for providing us with her code. This project was supported by the EU-TMR project on Spatial and Computational Statistics (ERB-FMRX-CT960095) and the Strategic Institute Program of the Norwegian Research Council n. 121144/420.

1 INTRODUCTION

Estimating the tail of a distribution can be important in the understanding of observable natural or man-made systems whose extreme behavior is of interest. Especially when heavy tails are suspected, accurate information about these become crucial for estimation of quantiles which are important in, say, prediction of floods or estimation of financial reserves in insurance. Because extreme data are rare, it is difficult to fit tail models and to support parametric model choices convincingly. An approach that is often taken is based on fitting excesses over a high threshold. Pickands (1975) and Balkema & de Haan (1974) proved that, under certain conditions, the Generalized Pareto Distribution (GPD) is the limit distribution of scaled excesses over high threshold values. More precisely, following Embrechts, Klüppelberg & Mikosch (1997, section 3.4), let $F(x)$ denote the distribution function of a real random variable X and let

$$F_u(x) = P(X \leq u + x | X > u).$$

Let x_∞ denote the right endpoint of the range of X. Then X belongs to the maximum domain of attraction of the generalized extreme value distribution if and only if there exists a positive real function $a(u)$ such that

$$\lim_{u \uparrow x_\infty} |F_u(xa(u)) - G_{\xi,1}(x)| = 0,$$

where $G_{\xi,\sigma}$ is the GPD with parameters ξ and σ. This distribution has the form

$$G_{\xi,\sigma}(x) = 1 - (1 + \frac{\xi x}{\sigma})^{-\frac{1}{\xi}}$$

if $\xi \neq 0$, or if $\xi > 0$ and $x > 0$, or if $\xi < 0$ and $0 < x < -\sigma/\xi$, while $G_{\xi,\sigma}(x) = 1 - \exp(x/\sigma)$ if $\xi = 0$, and $\sigma > 0$. In this paper we consider the case $\xi \geq 0$. A heavy tail is present if $\xi > 0$. For some large threshold u, $F_u(x)$ can be approximated by a GPD $G_{\xi,\sigma}$ with appropriate parameters ξ and σ. Common practice is to fix a threshold u and to fit a GPD to the data exceeding u. In order to reduce model bias, the threshold u should be chosen to be large, but this (often) leaves (very) few data points for the estimation of the parameters ξ and σ, which will have large variances. The selection of an appropriate threshold, above which the GPD assumption is appropriate, is a difficult task in practice, see for instance McNeil (1996), Davison & Smith (1990), Rootzen & Tajvidi (1997), Embrechts et al. (1997, section 6.5). There are a number of methods, beside maximum likelihood, used to estimate ξ and σ once u is fixed, see for instance Resnick (1997), Crovella & Taqqu (1999) and references therein. As is well known, the estimates depend significantly on the choice of the threshold, see for instance Embrechts et al. (1997, Figure 6.2.8.). Citing from Rootzen & Tajvidi (1997, p.79): *"An important practical problem is the choice of the level u for the excesses. ... (We) feel that ... (the) choice of level has to be made from subject matter knowledge."* Research is underway to provide practitioners, who sometimes need to perform many data analysis regularly, with more automatic and robust approaches that do not require an a priori tuning of a threshold. Such an unsupervised approach to tail

estimation is particular relevant in automatic real-time monitoring of financial, industrial and environmental quantities, for instance for warning purposes. Often a supervised analysis will be performed, selectively and off-line as part of a monitoring scheme. In the recent paper Dupuis (1999), a robust model validation mechanism is suggested to guide the threshold selection. The procedure assigns weights between zero and one to each data point, where a high weight means that the point should be retained since a GPD model is fitting it well. The author suggests to start with a low threshold u and increase it, thus reducing the number of data points, until all data left have weights close to 1. This is an interesting idea. Thresholding is still needed at the level of the weights, though this choice is more robust and is more easy to automatise. In this paper an alternative fully unsupervised approach to tail estimation is suggested. There are three key ingredients in our approach: we model all data, not only those belonging to the tail; we use a mixture model, one component of which is a GPD and the other component a density with lighter tail; and, the weight in the mixture is dynamic and determines when the GPD component in the mixture is prevalent.

Let $X_1, X_2, ..., X_n$ be non-negative i.i.d. random variables with common probability density function $l(x)$ given by

$$l(x) = \frac{[1 - p(x;\theta)]f(x;\beta) + p(x;\theta)g(x;\xi,\sigma)}{Z(\theta,\beta,\xi,\sigma)}, \tag{1}$$

where

$$g(x;\xi,\sigma) = \frac{1}{\sigma}(1 + \frac{\xi x}{\sigma})^{-\frac{1}{\xi}-1} \tag{2}$$

is the GPD density, $p(x;\theta)$ is the mixing function taking values in $(0, 1]$, increasing in x and such that, for all θ,

$$\lim_{x \to x_\infty} p(x;\theta) = 1, \tag{3}$$

$f(x,\beta)$ is some other density, and $Z(\theta,\beta,\xi,\sigma)$ is the normalizing constant. The notation indicates that the mixing function p and the f component are parameterized with parameters, or vectors of parameters, θ and β respectively. Model (1) is a mixture of f and g, where the mixing probability $p(x;\theta)$ depends on x. Because of the property (3), the right tail of $l(x)$ is governed for large x by the GPD component, while the left tail is controlled by $f(x;\beta)$.

In the next section we make our choice for f and for p, we discuss the intuition behind our model and compare it to standard mixture models with constant mixing parameter p. In section 3 we present an algorithm for sampling from l. This allows us to perform a simulation study described in section 4. We discuss likelihood based inference for model (1) and describe the numerical procedure we use. In section 4 we present a simulation study to investigate the features of the model and the suggested inferential procedure. We compare our approach to Peaks-Over-Threshold (POT) inference guided by Dupuis' threshold selection (DTS). In section 5 we discuss robustness with respect to the choice of f. We analyse model mismatch situations, when data do not originate from the same model

used for inference, and compare our methods to the POT using DTS. Our study concentrates on the estimation of q-quantiles, sometimes called in the financial context the Value-at-Risk (VaR_q), defined as $\text{P}(X \geq \text{VaR}_q) = q$. Quantiles are often more relevant in applications than the parameter estimates themselves. In section 5 we analyse the Danish fire loss data, which have been previously used in McNeil (1996) and Embrechts et al. (1997). Our method seems to have smaller root mean squared errors for small percentiles ($1/1000$ and $1/10000$) and heavy tails ($\xi = 0.25, 0.5$). We end this paper with some final remarks.

2 The Dynamic Mixture Model

One possible choice for the mixing function in (1) is

$$p(x;\theta) = \frac{1}{2} + \frac{1}{\pi}\arctan(\frac{x-\mu}{\tau}), \qquad \theta = (\mu,\tau), \qquad \mu,\tau > 0 \tag{4}$$

with a location parameter μ and steepness parameter τ^{-1}. We take as the left component in the mixture a Weibull distribution with density

$$f(x;\beta,\lambda) = \beta\lambda^{\beta}x^{\beta-1}e^{-(\lambda x)^{\beta}} \tag{5}$$

for $\beta, \lambda > 0$. As expected, the normalizing constant

$$Z(\theta,\beta,\xi,\sigma) = 1 + \frac{1}{\pi}\int_0^{\infty}\left[\frac{1}{\sigma}(1+\frac{\xi x}{\sigma})^{-\frac{1}{\xi}-1} - \beta\lambda^{\beta}x^{\beta-1}\exp -(\lambda x)^{\beta}\right]\arctan(\frac{x-\mu}{\tau})\,dx$$

has to be computed numerically. We suggest to use the Weibull as the left component since it is flexible and its right tail is not heavy. Other choices are possible and could be needed in particular settings. Notice that the mixture $l(x)$ is a density with tail parameter ξ. In fact, the tail is just as heavy as the heaviest right tail component, if (3) holds. The mixing function (4) also leads to a continuous density $l(x)$. An alternative choice for p is the Heaviside function

$$p(x,\theta) = \begin{cases} 1 & \text{if } x \geq \theta \\ 0 & \text{if } x < \theta. \end{cases} \tag{6}$$

In this case, the parameter θ takes the role of the threshold u after which only the GPD component remains in (1). Using (6) will in general make $l(x)$ discontinuous in $x = \theta$, which is often an unsuitable feature.

In order to better understand how the two components interact, it is useful to give another representation of the model (1) as the pure mixture model

$$l(x) = (1-\pi)g_1(x) + \pi g_2(x), \tag{7}$$

where the mixing parameter π can be computed as the probability

$$\pi = \pi(\xi,\sigma,\theta,\lambda,\beta) = \frac{\int_0^{\infty} g(x,\xi,\sigma)p(x;\theta)dx}{\int_0^{\infty} f(x,\beta,\lambda)(1-p(x;\theta))dx + \int_0^{\infty} g(x,\xi,\sigma)p(x;\theta)dx}$$

and the two components are

$$\begin{aligned} g_1(x) = g_1(x;\beta,\lambda,\theta) &= \frac{f(x,\beta,\lambda)(1-p(x;\theta))}{\int_0^\infty f(x,\beta,\lambda)(1-p(x;\theta))dx}, \\ g_2(x) = g_2(x;\xi,\sigma,\theta) &= \frac{g(x,\xi,\sigma)p(x;\theta)}{\int_0^\infty g(x,\xi,\sigma)p(x;\theta)dx}. \end{aligned}$$

This way of reparametrizing model (1) leads to inference based on the hidden allocation variables $Y_i = 1$ with probability $1-\pi$ when X_i is sampled from g_1 and $Y_i = 0$ when X_i is sampled from g_2. The completed data set $(X_1, X_2, ..., X_n, Y_1, Y_2, ..., Y_n)$ is the basis for maximum likelihood (ML) estimation, using the EM algorithm (Meng & van Dyk, 1997) or MCMC (Gilks, Richardson & Spiegelhalter, 1996; Robert & Casella, 1999), of the parameters π and those of the two components. Conditional on $(Y_1, Y_2, ..., Y_n)$ the data can be assigned to one of the two components and used to estimate the parameters in each of these. However, this approach does not help, because the mixing probability π and the components g_1 and g_2 share $\theta, \beta, \lambda, \xi, \sigma$. Hence the data cannot be split according to what information they carry regarding the different parameters, as usually happens in mixture inference. This is also true when using the the Heaviside function (6). Conditionally on the cut-point θ each data point X_i is assigned to one of the two components, depending on being larger or smaller than θ. But θ is also estimated, and again each data point contributes to the estimation of all parameters.

3 SIMULATION AND ESTIMATION

In order to perform simulation experiments we need to be able to sample from the dynamic mixture model (1) given all parameters. As we can reformulate the model as the pure mixture model (7), we could use the standard simulation algorithm

1. Draw U uniformly on $[0, 1)$.
2. If $U < \pi$ sample x from $g_2(x)$ and return x. Else, sample x from $g_1(x)$ and return x.

This algorithm is not particularly convenient in our case, as we need to compute π, $g_1(x)$, $g_2(x)$, and find algorithms to sample from $g_1(x)$ and $g_2(x)$. To avoid these problems and only use samples from the two original components f and g, we suggest the following alternative sampling scheme:

1. Draw U uniformly on $[0, 1)$.
2. If $U < 1/2$, then sample x from $f(x)$; return x with probability $1 - p(x)$ and stop; or, with probability $p(x)$, return to 1.
3. If $U \geq 1/2$, then sample x from $g(x)$; return x with probability $p(x)$ and stop; or, with probability $1 - p(x)$, return to 1.

The correctness of this algorithm is seen from the following argument. We define the indicator variable $A = 1$ if x is accepted, $A = 0$ otherwise. Let P indicate the probability measure of the output variable X of the above algorithm. Then

$$\begin{aligned} P(X \leq x | A = 1) &= \frac{P(X \leq x, A = 1)}{P(A = 1)} \\ &= \frac{1}{2} \frac{P(X \leq x, A = 1 \mid U < 1/2) + P(X \leq x, A = 1 \mid U \geq 1/2)}{P(A = 1)}, \end{aligned}$$

and the corresponding conditional density is, as required,

$$\frac{(1 - p(x))f(x) + p(x)g(x)}{\int_0^\infty [(1 - p(y))f(y) + p(y)g(y)]dy},$$

dropping for simplicity the parameters in the notation. The algorithm returns a value in geometric time if the mixing function $p(X; \theta)$ is strictly positive with probability one.

To maximise the log-likelihood function, we face two problems: first, the normalization constant has to be computed numerically, and then we need to maximise numerically the log-likelihood function. Unless we compute the normalization constant with high precision, we can introduce false local maxima which would confuse the optimising algorithm. We therefore rewrote the integral from 0 to ∞ as a sum of integrals from 0 to 1, 1 to 2 and so on, and computed each integral with high precision using numerical quadrature (routine `d01ahf` in the NAG-library). (For details about the NAG-algorithms, we refer to the online description of each algorithm at http://www.nag.com.) The sum was truncated when the last contribution did not give any changes in the integral. In order to maximise the log-likelihood function, we experienced with two different methods; a quasi-Newton method (routine `e04jyf` in the NAG-library) where finite differences approximates derivatives, and a Simplex-method (routine `e04ccf` in the NAG-library) which did not required any derivative information about the log-likelihood. The quasi-Newton method was able to handle constrained parameters, like $\tau > 0$, while to use the Simplex-method we had to reparameterise $\tau = \exp(\tau')$.

In our experience, both optimization methods are robust with respect to the choice of initial values and converge for all reasonable initial values we tried. Our scheme is to get initial values for the parameters in f and g by roughly splitting the data into two disjoint sets. The parameters in the weight function are computed accordingly. Both optimisation methods are fast and give the same final result, although the quasi-Newton methods is somewhat faster than the Simplex-method.

Estimates of the variance of the ML-estimators are often obtained by inverting of the empirical Fisher information matrix (i.e. the negative Hessian of the log-likelihood function) evaluated at the maximum. This approach is feasible when the number of observed data points is large. In our simulation study, we will however report variances across simulated datasets.

If data are assumed to originate from the dynamical mixture model (1), it is possible to estimate a threshold $\hat{x}_\epsilon$, beyond which data follow a pure GPD with P-probability $1 - \epsilon$. P is the probability

measure of the output variable of the sampling algorithm described above. Define the threshold x_ϵ as the smallest value x, so that

$$\mathrm{P}(Y = 1 \mid X = x',\ A = 1) \leq \epsilon, \quad \forall x' \geq x, \tag{8}$$

where the left hand side is the probability that an accepted value x' was indeed sampled from the light-tailed component. The binary variable Y equals 1 if the variable actually was sampled from f. The left hand side of (8) can be rewritten as

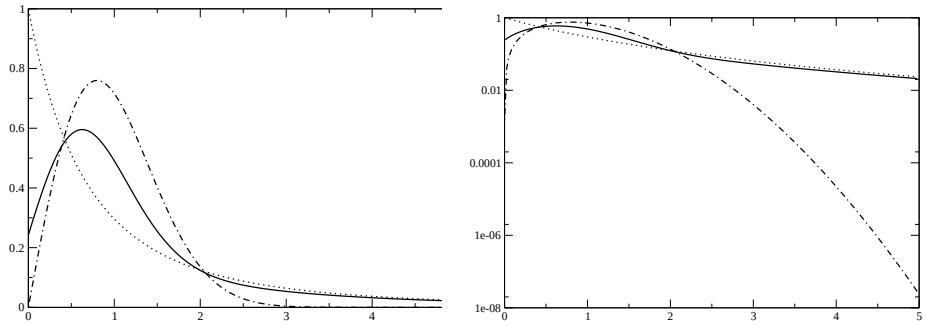
$$x_\epsilon = \inf \left\{ x \ : \ \frac{(1 - p(x', \theta)) f(x')}{(1 - p(x', \theta)) f(x') + p(x', \theta) g(x')} < \epsilon, \ \ \forall\, x' > x \right\}. \tag{9}$$

We use $\epsilon = 0.001$ and plug in the estimated parameters into (9) to obtain a threshold value $\hat{x}_\epsilon$. It is now possible to perform standard POT inference on the parameters of a pure GPD and compute estimated quantiles. Of course, when the data are not a genuine sample from a dynamical mixture model, x_ϵ loses its precise interpretation but can still be used as a rough guide for threshold selection.

4 A SIMULATION STUDY

The probability density $l(x)$ given by (1), (2), (4) and (5) has six parameters: β and λ in the Weibull density component, μ and τ in the weight function p, and ξ and σ in the GPD component. Our first aim is to investigate if inference on these parameters and on quantiles is at all feasible and if so, how the present approach compares with alternative ones. Many data sets have been generated with the algorithm described in Section 3. The parameters have been chosen so that the corresponding density function $l(x)$ is smooth, the data originating from the two components are properly confounding in the overlapping support and a pronounced bi-modal shape is avoided. This puts our simulation study in a realistic and challenging setting. Strong discontinuities or bi-modalities indicate separation of the data and would thus make parameter estimation easier. In addition, only modest (positive) values of ξ have been used, so that $g(x)$ and $l(x)$ are heavy tailed, but data from the two components are sufficiently mixed. In this section we report only on the experiments with two combinations of the six parameters, which were found to be typical. For $\beta < 1$, the Weibull density f has no proper mode and is decreasing for all positive x, while f is unimodal for $\beta > 1$. Here we take $\beta = 2$. The λ parameter is chosen so that the expectation of the Weibull density is 1, that is $\lambda = \Gamma(1 + 1/\beta)$. In the weight-function, we use $\mu = 1$ and $\tau = 1$. Two different values are considered for ξ, namely $1/2$ and $1/4$, and we take $\sigma = 1$. The two parameter combinations are summarized in Table 1. The corresponding density for $\xi = 1/2$ is plotted in Figure 1, on normal and log-scale. The density for $\xi = 1/4$ is similar.

We generated 100 independent data sets for the two parameter combinations, each with $N = 1000$ points. The same is then repeated with $N = 200$, to investigate the effect of sample size. For each experiment, the six parameters are estimated using our approach and by POT maximum likelihood



(a) (b)

FIGURE 1: The density in experiment 1 ($\xi = 1/2$) in normal (a) and log-scale (b). The thick line is the mixture, the dotted line is the GPD and the dashed line is the Weibull density.

	Weibull component		Weight function		GPD component	
	β	λ	μ	τ	ξ	σ
Experiment 1	2.0	$\Gamma(1.5)$	1.0	1.0	0.5	1.0
Experiment 2	2.0	$\Gamma(1.5)$	1.0	1.0	0.25	1.0

TABLE 1: The two parameter combinations used in the simulation experiment.

inference using DTS. Because the data are sampled from the same model our approach uses for inference, it is not surprisingly that our estimates of the tail are clearly better.

DTS is not automatic, but requires specialist intervention to judge the results. In order to apply this method in a large simulation study, we had to make it automatic. The method assigns weights between zero and one to each data point, where a high weight means that the point should be retained since a GPD model is fitting it well. Dupuis suggests to start with a low threshold u and increase it until all data left have weights close to 1, or until the hypothesis that all could have been equal to 1 (at 5% significance level, say) is accepted. We have simply automated this advice, and start with a low threshold and increase it until either of the two conditions apply. For the first of these conditions, at the level of the weights, we used the same thresholding as in Dupuis (1999).

Tables 2 and 3 report estimated parameters, using our approach, when $N = 1000$, and $\xi = 0.5$ or $\xi = 0.25$, respectively. The first column lists the true parameter values and the other columns describe the distributions of the estimators on the basis of the 100 repetitions: smallest estimated value in the second column, then the first quantile, the median of the estimates, their mean, the third quantile and the largest estimate. The last column lists the estimated standard deviations. For all parameters, the estimates are rather spread. Especially τ and β are difficult to estimate and the estimators are clearly dependent. The heavy tail is clearly recognized. Table 3 shows a slightly better picture, and the case $\xi = 0$ (not shown) was in fact rather good, indicating growing difficulties for heavier tails. We see next that errors in the parameter estimates compensate, so that quantile estimation is satisfactory.

	True value	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Stdev
σ	1.0	0.051	0.892	1.020	1.042	1.167	2.306	0.322
ξ	0.5	0.102	0.375	0.430	0.428	0.488	0.681	0.099
μ	1.0	0.016	0.696	0.983	1.154	1.300	3.666	0.762
τ	1.0	0.031	0.638	0.940	1.528	1.309	21.070	2.799
β	2.0	0.440	2.018	2.254	2.423	2.490	9.775	1.026
λ	0.886	0.089	0.813	0.861	0.843	0.916	1.005	0.130

TABLE 2: Summary of the results from 100 simulation from experiment 1.

	True value	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Stdev
σ	1.0	0.298	0.924	1.045	1.052	1.197	1.835	0.262
ξ	0.25	0.002	0.136	0.183	0.182	0.224	0.368	0.072
μ	1.0	0.021	0.752	1.154	1.337	1.635	5.171	0.899
τ	1.0	0.026	0.725	1.022	1.695	1.394	28.380	3.322
β	2.0	0.777	1.984	2.170	2.328	2.387	6.899	0.797
λ	0.886	0.101	0.846	0.883	0.866	0.924	1.032	0.125

TABLE 3: Summary of the results from 100 simulation from experiment 2.

Tables 4 and 5 are concerned with quantile estimation for the two parameter combinations. We estimate the $1/100$, $1/1000$ and $1/10000$ quantiles, which can be computed to be equal to the values reported in the captions by numerical integration (using the true parameter values and NAG-routine `c05adf` to solve the equation). For each of the 100 data sets, we used the estimated parameters obtained with the dynamical mixture approach, to compute the quantiles, again by numerical integration and using NAG-routine `c05adf`. We compare our approach (which we abbreviate "MIX") with two others. In both we use DTS but then estimate the parameters of the GPD beyond the threshold by ML ("POT-ML"). The estimated parameters are used to compute quantiles of the GPD. The numbers reported in Tables 4 and 5 are standardized quantiles, defined as the estimated quantile divided by the true quantile, for all approaches. The target value is hence one. The last column reports the square-root of the squared bias plus the variance (RMSE). We see that our approach is more efficient in particular for larger percentiles: POT-ML is slightly better for the $1/100$ percentile, as good as MIX for $1/1000$ while MIX is best for $1/10000$. As mentioned before the chosen parameter combinations lead to difficult data sets. In view of this fact, estimation of quantiles via the dynamic mixture model is rather good. Both methods tend to underestimate the quantiles, though overestimation also happens. Their behavior is more similar for lower values of ξ. In fact for $\xi = 0$ the methods essentially coincide and are all performing very well.

Next we repeat the same experiments with smaller data sets, each consisting of $N = 200$ points.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Stdev	RMSE
x_{100}^{MIX}	0.622	0.779	0.844	0.852	0.908	1.112	0.104	0.181
$x_{100}^{\text{POT-ML}}$	0.711	0.891	0.978	0.988	1.080	1.450	0.131	0.132
x_{1000}^{MIX}	0.426	0.617	0.725	0.763	0.848	1.306	0.200	0.310
$x_{1000}^{\text{POT-ML}}$	0.366	0.709	0.878	0.944	1.120	1.956	0.339	0.343
x_{10000}^{MIX}	0.275	0.478	0.604	0.697	0.800	1.754	0.315	0.437
$x_{10000}^{\text{POT-ML}}$	0.149	0.510	0.793	1.066	1.302	3.609	0.762	0.765

TABLE 4: Summary of 100 normalized $1/100$, $1/1000$ and $1/10000$-percentiles from experiment 1 with $N = 1000$. The true percentiles are 17.57, 60.17 and 195.19.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Stdev	RMSE
x_{100}^{MIX}	0.718	0.830	0.889	0.879	0.920	1.103	0.073	0.141
$x_{100}^{\text{POT-ML}}$	0.792	0.917	0.996	0.993	1.059	1.225	0.102	0.102
x_{1000}^{MIX}	0.558	0.721	0.777	0.796	0.856	1.179	0.117	0.235
$x_{1000}^{\text{POT-ML}}$	0.543	0.823	0.932	0.966	1.075	1.611	0.220	0.223
x_{10000}^{MIX}	0.404	0.585	0.675	0.724	0.806	1.452	0.197	0.339
$x_{10000}^{\text{POT-ML}}$	0.369	0.685	0.917	1.086	1.302	2.890	0.562	0.569

TABLE 5: Summary of 100 normalized $1/100$, $1/1000$ and $1/10000$-percentiles from experiment 2 with $N = 1000$. The true percentiles are 8.54, 18.39 and 35.92.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Stdev	RMSE
x_{100}^{MIX}	0.404	0.681	0.784	0.826	0.947	1.527	0.209	0.272
$x_{100}^{\text{POT-ML}}$	0.444	0.765	0.936	0.980	1.126	2.099	0.315	0.316
x_{1000}^{MIX}	0.205	0.479	0.633	0.737	0.894	2.378	0.380	0.462
$x_{1000}^{\text{POT-ML}}$	0.217	0.553	0.989	1.231	1.663	3.882	0.848	0.879
x_{10000}^{MIX}	0.092	0.301	0.526	0.712	0.882	3.926	0.630	0.693
$x_{10000}^{\text{POT-ML}}$	0.103	0.382	1.019	1.936	2.812	11.414	2.155	2.349

TABLE 6: Summary of 100 normalized $1/100$, $1/1000$ and $1/10000$-percentiles from experiment 1 with $N = 200$. The true percentiles are 17.57, 60.17 and 195.19.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Stdev	RMSE
x_{100}^{MIX}	0.597	0.781	0.886	0.893	0.988	1.330	0.156	0.189
$x_{100}^{\text{POT-ML}}$	0.621	0.865	0.963	0.996	1.097	1.762	0.214	0.214
x_{1000}^{MIX}	0.452	0.665	0.781	0.838	0.943	2.419	0.281	0.324
$x_{1000}^{\text{POT-ML}}$	0.439	0.825	1.074	1.266	1.546	3.739	0.665	0.717
x_{10000}^{MIX}	0.319	0.524	0.701	0.821	0.935	4.748	0.536	0.565
$x_{10000}^{\text{POT-ML}}$	0.304	0.809	1.212	2.038	2.412	9.078	2.040	2.288

TABLE 7: Summary of 100 normalized $1/100$, $1/1000$ and $1/10000$-percentiles from experiment 2 with $N = 200$. The true percentiles are 8.54, 18.39 and 35.92.

Table 6 and 7 show the results for the two parameter combinations. The situation is more difficult and we see that standard deviations are larger. Our approach is now performing best for all percentiles, although the gain in efficiency is best seen for the largest ones.

We have also tried our alternative threshold (9) with $\epsilon = 0.001$ in the POT approach, estimating a pure GPD beyond $\hat{x}_\varepsilon$. In a few cases this threshold was too large, leaving no data. This happened in correspondence to data sets with particularly few points on the tail. Beside these cases, the threshold $\hat{x}_\varepsilon$ gave estimated quantiles which were comparable to those obtained with DTS and ML.

Our next step is to investigate the performance of the dynamic mixture approach when the data are not generated form model (1), (2), (4) and (5), which is only used to infer on quantiles.

We present two experiments. In the first one we generate samples from a positively truncated Student-t distribution with three degrees of freedom. Results on quantile estimation are reported in Table 8, when the sample size N is 1000, and Table 9, for $N = 200$. In the second experiment we use a standard log-normal distribution, see Tables 10 and 11. Our dynamical mixture model is assumed and parameters estimated as described above. Estimated quantiles are then computed using the estimated parameter values in (1). The POT model using DTS is also applied and paramters are estimated by

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Stdev	RMSE
x_{100}^{MIX}	0.671	0.805	0.860	0.863	0.895	1.249	0.090	0.164
$x_{100}^{\text{POT-ML}}$	0.799	0.931	1.006	0.999	1.070	1.231	0.094	0.094
x_{1000}^{MIX}	0.533	0.644	0.740	0.753	0.826	1.538	0.140	0.284
x_{10000}^{MIX}	0.384	0.488	0.582	0.625	0.737	1.519	0.186	0.419
$x_{10000}^{\text{POT-ML}}$	0.394	0.570	0.820	0.984	1.234	2.869	0.578	0.578

TABLE 8: Summary of the results from 100 simulation from the unsigned student-t distribution with 3 degrees of freedom and $N = 1000$.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Stdev	RMSE
x_{100}^{MIX}	0.624	0.768	0.843	0.883	0.940	1.831	0.189	0.223
$x_{100}^{\text{POT-ML}}$	0.646	0.828	0.914	0.940	1.043	1.367	0.145	0.157
x_{1000}^{MIX}	0.434	0.563	0.654	0.698	0.802	1.285	0.171	0.347
$x_{1000}^{\text{POT-ML}}$	0.381	0.648	0.809	0.916	1.070	1.994	0.368	0.378
x_{10000}^{MIX}	0.229	0.368	0.476	0.543	0.677	1.324	0.224	0.509
$x_{10000}^{\text{POT-ML}}$	0.202	0.420	0.713	1.028	1.293	4.585	0.891	0.892

TABLE 9: Summary of the results from 100 simulation from the unsigned student-t distribution with 3 degrees of freedom and $N = 200$.

ML. Quantiles are estimated using the GPD distribution with such estimated parameters. One hundred independent experiments are performed each time. The standradised estimated quantiles are reported in the tables, i.e. estimated divided true quantiles.

Looking at Tables 8, 9, 10 and 11, we see again that all models tend to underestimate quantiles. For the $1/100$ percentile POT-ML performs slightly better than MIX. One data set leads to a very large estimate of the quantile. For the $1/1000$ percentile POT-ML and MIX are very similar, with MIX more efficient for $N = 200$. Mix is always best for the $1/10000$ percentile.

In conclusion our simulation experiment has shown that inference on quantiles based on the dynamic mixture model is feasible and leads to results that are certainly comparable, sometimes better, than those obtain fitting a pure GPD beyond a threshold chosen using DTS. Maximum likelihood estimation of parameters of the dynamical mixture model and quantile estimation are performed in a truly unsupervised fashion. The estimates only require a few seconds in computing time. Our semi-automatic, C++ implementation of DTS took more time; some of the datasets took a few hours, but with a more clever search algorithm for the threshold it should be possible to reduce the computing time to some minutes. We did not optimise Dupuis' algorithm, hence these results could be improved, at a cost of allowing manual tuning.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Stdev	RMSE
x_{100}^{MIX}	0.695	0.775	0.811	0.818	0.852	1.015	0.064	0.193
$x_{100}^{\text{POT-ML}}$	0.778	0.926	0.995	0.998	1.062	1.282	0.105	0.105
x_{1000}^{MIX}	0.541	0.706	0.767	0.786	0.858	1.163	0.117	0.244
$x_{1000}^{\text{POT-ML}}$	0.582	0.854	0.973	0.985	1.094	1.592	0.191	0.192
x_{10000}^{MIX}	0.432	0.720	0.903	0.939	1.083	1.731	0.288	0.294
$x_{10000}^{\text{POT-ML}}$	0.458	0.794	1.055	1.119	1.309	2.800	0.450	0.466

TABLE 10: Summary of the results from 100 simulation from the standard log-normal distribution and $N = 1000$.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Stdev	RMSE
x_{100}^{MIX}	0.563	0.728	0.810	0.828	0.904	1.294	0.142	0.223
$x_{100}^{\text{POT-ML}}$	0.690	0.831	0.955	0.978	1.076	1.666	0.206	0.207
x_{1000}^{MIX}	0.475	0.612	0.760	0.808	0.909	1.722	0.255	0.319
$x_{1000}^{\text{POT-ML}}$	0.468	0.703	0.864	1.051	1.292	2.572	0.496	0.499
x_{10000}^{MIX}	0.369	0.551	0.775	0.880	1.104	2.357	0.424	0.440
$x_{10000}^{\text{POT-ML}}$	0.328	0.553	0.851	1.366	1.707	5.274	1.162	1.219

TABLE 11: Summary of the results from 100 simulation from the standard log-normal distribution and $N = 200$.

Our new approach seems to work also comparable well on synthetic data sampled from a different distribution which is not a dynamical mixture. It would be interesting to develop a theory to compute bias and variance of the estimated quantiles, but this is beyond the scope of this paper. We conclude this section with a remark, inspired by some arguments in Feuerverger & Hall (1999), that can help to justify the good properties of the dynamical mixture model. We know that as $u \to \infty$ (and $x > u$)

$$F_u(x) \sim 1 - c_1 x^{-1/\xi}.$$

We can then define an error term $\delta(x)$ and write the equality

$$F_u(x) = 1 - c_1 x^{-1/\xi}(1 + \delta(x)),$$

where $\delta(x) \to 0$ as $x \to \infty$. As in Feuerverger & Hall (1999), we can now model $\delta(x)$ and assume

$$\delta(x) = c_2 x^{-1/\alpha} + o(x^{-1/\alpha})$$

for $\alpha > 0$. This leads to

$$F_u(x) = 1 - c_1 x^{-1/\xi} + c_3 x^{-(1/\xi + 1/\alpha)} + o(x^{-(1/\xi + 1/\alpha)})$$

which shows that after the leading term of order $x^{-1/\xi}$, there are in fact infinitely many terms of the order $x^{-\kappa}$ for $\kappa > 1/\xi$. Our choice of a Weibull component is in this respect rather arbitrary, though its flexibility is well known, but it responds to the need to model the error δ. In addition, the functional mixing term p also contributes to the terms in the expansion. Assume for example that data are sampled from a three-term dynamically weighted mixture, with components being a Weibull distribution, a GPD(ξ_1) and a GPD(ξ_2) with $\xi_2 > \xi_1$. (The weight function should be such that the heaviest tail component remains as $x \to \infty$.) If we now fit our original mixture model with only two components (Weibull(θ) and GPD(ξ)), we ignore one component. This will have an effect on the estimates of θ and ξ. We expect $\hat{\xi}$ to be larger than ξ_2, since part of the tail originated by GPD(ξ_1) will be incorporated in the heavy tail component of (1). Also the Weibull component will try to stretch more towards the right tail. While estimates might be unsatisfactory, estimated quantiles and returns might however be reliable, in the sense that q-quantiles, estimated on the basis of weighted mixture model, will be similar to the q-quantile of the correct likelihood with 3 components. The results in this section confirm this intuition.

5 DANISH FIRE LOSS DATA

We will now apply our approach to the Danish fire claim data, which constitute insurance losses over one million Danish kroner (DKK) caused by industrial fires occurring in the period $1980 - 1990$. The losses include damage to buildings, furniture and personal property as well as loss of profit. The unit is 1 million DKK and all data have been adjusted to the 1985 price level leaving a total of 2156

Weibull component parameters		Weight function parameters		GPD component parameters	
$\hat{\beta}$	$\hat{\lambda}$	$\hat{\mu}$	$\hat{\tau}$	$\hat{\xi}$	$\hat{\sigma}$
1.059 (0.041)	1.077 (0.138)	1.039 (0.136)	0.065 (0.101)	0.621 (0.052)	1.044 (0.136)

TABLE 12: The parameter estimates obtained from fitting the weighted mixture model to the Danish fire loss data. The estimated standard deviations are shown in parentheses.

losses in the range 1 to 263.25MDKK. Figure 2 displays a histogram of the data in log-scale. This data set has been analysed in several papers, and is known to exhibit a heavy tail. We translate the data by -1 for simplicity and apply our dynamical mixture model (1), (2), (4) and (5). We used $\beta = 0.5, \lambda = 2, \mu = 1, \tau = 1, \xi = 0.4, \sigma = 2$ as initial values for the maximum likelihood calculation, following the strategy mentioned above. The results are given in Table 12. Standard deviations are estimated by inverting the empirical Fisher information matrix.

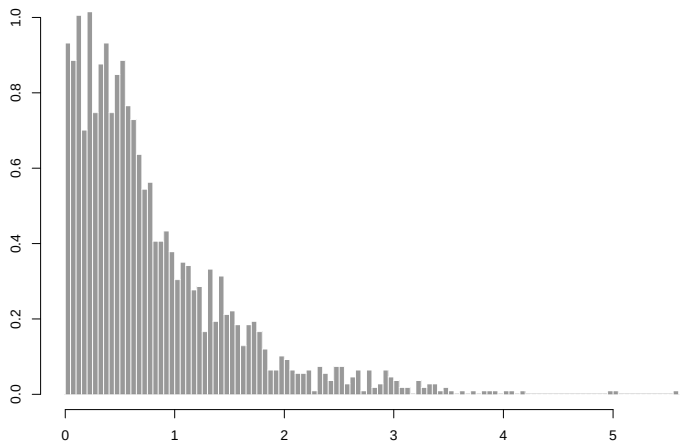


FIGURE 2: Histogram of Danish fire loss data, in log-scale.

How do our estimates compare to those obtained with other approaches? McNeil (1996) uses the POT method and fits a pure GPD to the data above some subjectively chosen threshold u. Table 13 lists the estimates obtained for the various thresholds u that McNeal suggests. (Because data are here translated by one, so are the thresholds with respect to McNeil (1996).) To make up for the offset introduced through the thresholding of the data at $x = u$, the estimates of σ also have been transformed to fit the tail of the distribution of the data themselves and not the exceedances. We observe that the estimate $\hat{\xi}$ obtained with the weighted mixture model are comparable to the POT estimates using McNeil's thresholds, and similarly for $\hat{\sigma}$. The estimated variances of $\hat{\xi}$ and $\hat{\sigma}$ using the mixture model are smaller than for the POT. Alternatively, we used our automated version of

DTS, which gives 6.5 as the threshold, leaving 145 values for inference. The ML estimate of the two GPD parameters beyond this threshold are $\hat{\xi} = 0.45\ (0.114)$ and $\hat{\sigma} = 2.03\ (0.81)$, with estimated standard deviations in parenthesis, as given via the empirical Fisher information matrix. The figures are different, though not significantly.

Threshold	$\hat{\xi}$	$\hat{\sigma}$
2	0.67 (0.07)	0.86 (0.14)
3	0.72 (0.10)	0.72 (0.18)
4	0.63 (0.11)	0.98 (0.32)
9	0.50 (0.14)	1.58 (0.81)
19	0.68 (0.28)	0.58 (0.77)

TABLE 13: Parameter estimates obtained from the POT method, using the thresholds suggested in McNeil (1996) for the Danish fire loss data. The estimated standard deviations, computed using the empirical Fisher information matrix, are shown in parentheses.

We now compare the quantiles of the fitted distributions. McNeil (1996) calculates these based on the POT-fitted GPD above the threshold $u = 9$. The estimated q-quantiles are given in Table 14 together with the estimates we obtain via the fitted dynamic mixture, using the estimated parameter values given in Table 12. The two methods give comparable quantiles for $q = 0.05$ and 10^{-2}. For higher values of q, q-quantiles based on our model are larger. For the 10^{-4} and 10^{-5} quantiles the effect of the deviation in the estimates for ξ, 0.621 compared to 0.5, becomes clear. Still it can be seen that the q-quantiles based on our model lies within the 95%-confidence interval of the POT-ML based quantile when $u = 9$. The last column of Table 14 reports the estimated quantiles obtained using DTS, i.e. $u = 6.5$ and ML fit to the GPD beyond $u = 6.5$. This ML GPD fit is rather similar to that one with $u = 9$. In conclusion, small and medium quantiles are not significantly different, while large ones are estimated to be larger with our unsupervised dynamic mixture model. Given the tendency to underestimate quantiles observed in our simulations, we would tend to report the largest estimated quantiles.

In Table 15 we show the estimated threshold $\hat{x}_\epsilon$ based on (9), which correspond to those tried in McNeil (1996).

6 CONCLUSIONS

In this paper we suggest an unsupervised alternative to the classical POT model, where a GPD is fitted beyond a threshold which is selected in a supervised way. We suggest to model the data with a dynamical mixture: one component is a GPD and becomes predominant, as x moves to infinity, in a continuous and smooth way. We used a lighter Weibull distribution as the other component, though

q	mixture model	POT-ML, $u = 9$	POT-ML, $u = 6.5$
0.05	8.3	9.1	8.6
10^{-2}	25.5	26.3	26.9
10^{-3}	112.0	93.3	90.9
10^{-4}	473.8	303.9	270.2
10^{-5}	1987.0	965.2	772.4

TABLE 14: q-quantile estimates for the Danish fire loss data, using our dynamically weighted mixture model, the POT maximum likelihood approach with threshold $u = 9$, the POT maximum likelihood approach with threshold $u = 6.5$.

ϵ	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
$\hat{x}_\epsilon$	2.60	4.65	6.70	8.65	10.60

TABLE 15: Suggested thresholds $\hat{x}_\epsilon$, using (9) for the Danish fire loss data.

this is just one of the possible choices. Maximum likelihood inference is performed on the heavy tail parameter ξ and empirical quantiles are computed by numerical integration. No data thresholding is performed and all data are used to estimate the parameters present in the two components and in the weight function. We have tried our approach on various simulated data, where it performs well. We implemented Dupuis' robust threshold selection method is an unsupervised way, which might not be optimal for each specific data set. However, our limited experience indicates that a supervised use of Dupuis' threshold selection does not improve results significantly. We compared the unsupervised method with ML and robust estimation of the GPD tail beyond the threshold estimated by DTS. Our results show that our new approach is almost systematically better for $1/10000$ and $1/1000$ percentiles, sometimes doubling efficiency. this was particularly evident for smaller data sets. This is shown in a sensitivity analysis, where data do not follow the mixture model. Our dynamically weighted mixture model applied to the Danish fire loss data gives comparable parameter estimates to those previously reported in the literature, based on guided POT inference. Estimated quantiles are also comparable.

In Feuerverger & Hall (1999) a new estimator for ξ is proposed, based on an asymptotic expansion of $F_u(x)$ that adds a further term to the Pareto fit of relative excesses X/u. The bias and the variance of this new estimator are computed and compared to those of other standard estimators. We have not performed such a theoretical analysis of our approach, which would be interesting indeed.

While we have used a dynamically weighted mixture with two components, it is of course possible to extend our model to more components. We have not performed experiments in this setting, but expect that an even better fit to the data would be achieved.

Tail estimation and corresponding estimation of Value-at-Risk and returns are becoming more and more a routinely performed analysis, for which there is a need for more automatic, less ad hoc and subjective methods, at least as a real time warning device. We hope that this paper makes a contribution in this direction.

REFERENCES

BALKEMA, A. A. & DE HAAN, L. (1974). Residual life time at great age, *Annals of Probability* (2): 792–804.

CROVELLA, M. & TAQQU, M. (1999). Estimating the heavy tail index from scaling properties, *Methodology and Computing in Applied Probability* 1: 55–79.

DAVISON, A. C. & SMITH, R. L. (1990). Models for exceedances over high thresholds (with discussion), *Journal of the Royal Statistical Society, Series B* 5(3): 393–442.

DUPUIS, D. J. (1999). Exceedances over high thresholds: A guide to threshold selection, *Extremes* 1(3): 251–261.

EMBRECHTS, P., KLÜPPELBERG, C. & MIKOSCH, T. (1997). *Modelling Extremal Events*, number 33 in *Applications of Mathematics: Stochastic Modelling and Applied Probability*, Springer Verlag.

FEUERVERGER, A. & HALL, P. (1999). Estimating a tail exponent by modelling departure from a Pareto distribution, *The Annals of Statistics* 27(2): 760–781.

GILKS, W. R., RICHARDSON, S. & SPIEGELHALTER, D. J. (1996). *Markov Chain Monte Carlo in Practice*, London: Chapman & Hall.

MCNEIL, A. J. (1996). Estimating the tails of loss severity distributions using extreme value theory, *Technical report*, Department Mathematik, ETH Zentrum, Zürich.

MENG, X. L. & 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, Series B* 59(3): 511–567.

PICKANDS, J. (1975). Statistical inference using extreme order statistics, *The Annals of Statistics* (3): 119–131.

RESNICK, S. I. (1997). Heavy tail modeling and teletraffic data, *The Annals of Statistics* 25(5): 1805–1869.

ROBERT, C. P. & CASELLA, G. (1999). *Monte Carlo statistical methods*, Springer-Verlag New York.

ROOTZEN, H. & TAJVIDI, N. (1997). Extreme value statistics and wind storm losses: A case study, *Scandinavian Actuarial Journal* (1): 70–94.