Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Computational Statistics & Data Analysis 47 (2004) 583–598
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Data driven estimates for mixtures

Beatriz Vaz de Melo Mendes*, Hedibert Freitas Lopes
Statistics Department, Federal University at Rio de Janeiro, RJ, Brazil

Received 1 September 2001; received in revised form 1 November 2003

Abstract

Data with asymmetric heavy tails can arise from mixture of data from multiple populations or processes. We propose a computer intensive procedure to fit by quasi-maximum likelihood a mixture model to a robustly standardized data set. The robust standardization of the data set results in well-defined tails which are modeled using extreme value theory. The data are assumed to be a mixture of a normal distribution contaminated by a distribution with heavy tails. This procedure provides an analytical expression for the mixture distribution of the data, which may be used in simulations and construction of scenarios, while providing an accurate estimation of quantiles associated with probabilities close to zero or one. The performance of the proposed data driven procedure is assessed by simulation experiments and also by its application to real data.

Keywords: Extreme value theory; Mixtures; Robustness; Simulations

1. Introduction

In several situations the primary concern is the protection against the occurrence of some catastrophic scenario or extreme event associated with some (monitored) variable. Dikes must be constructed to withstand extreme sea levels. Environmental officers need to take actions when monitors show that air pollution is high. Risk managers compute risk measures to assess probabilities of adverse events. An actuary fits statistical models to claim data to establish the amount of capital in reserve to cover extreme losses. Common to all these situations is the need to obtain a reliable estimate for the

* Corresponding author. Rua Marquesa de Santos 22, apt. 1204, Rio de Janeiro, 22221-080, RJ, Brazil. Tel.: +55-021-22652914.
E-mail addresses: beatriz@im.ufrj.br (B.V.M. Mendes), hedibert@im.ufrj.br (H.F. Lopes).

0167-9473/$ - see front matter
doi:10.1016/j.csda.2003.12.006

probability density function of the variable of interest, which should be especially accurate on regions related to extreme events where, usually, the data are sparse.

The main role of a robust procedure is typically to identify different data structures or data patterns (Hampel et al., 1986; Huber, 1981). For example, the daily returns of a portfolio or a stock market index should not be treated as if every data point had come from the same probability distribution. Instead, they may be thought as generated from different processes related to different states of the economy.

Mixture models are the subject of vast amounts of research under classical and Bayesian approaches. Modeling complex data via mixtures has proven to be a powerful data analytic technique. Escobar and West (1995) and Roeder and Wasserman (1997) used mixture models in density estimation. West (1991, 1993) and Richardson and Green (1997) study fitting and applications of mixture of normals. Escobar (1994) used mixture of Dirichlet processes for inferences on finite normal mixture models. A comprehensive account of theory on mixture distributions and their applications may be found in McLachlan and Peel (2000) or Titterington et al. (1985).

When the objective is the computation of probabilities associated with extreme events, procedures based on models from the extreme value theory (EVT) work very well (Embrechts et al., 1997; Leadbetter et al., 1983). There are two main approaches when using EVT. The so-called blocks method makes use of the generalized extreme value (GEV) distribution to model maxima and minima collected in blocks of fixed size (see, for example, McNeil (1996, 1998) and Smith (1999), for applications in finance and insurance). In this paper we are not extending methods for GEV. The peaks over threshold method uses the generalized Pareto distribution (GPD) to model the tails beyond a high threshold (see McNeil and Frey, 1998; Mendes, 2000, among others). These two methods will give precise probability estimates of extreme events. However, there are a number of situations—for example, when simulating the evolution of a process in finance, in epidemiology, etc.—where one would like to have the analytical expression for the whole distribution and not just for the extreme tails.

When modeling tails using the GPD, a difficult problem is the estimation of the thresholds, i.e., the points where the tails begin. Several authors, including Smith (1987) and Pickands (1975), have addressed this problem. Danielsson and de Vries (1998) proposed a computer intensive bootstrap technique. Other authors proposed to obtain the thresholds using graphical techniques (see Embrechts et al., 1997 and references therein).

We propose to first robustly standardize the data in order to make the extreme points appear more obvious (Hoaglin et al., 1983), thus distinguishing the bulk of the data from the extreme tails. Then we fit a mixture model to the robustly standardized data set. We assume that the center and majority of the standardized data is normal and use the appropriate EVT models, two separate GPD models, on the well-defined tails, therefore combining three distributions to represent the data. We use the maximum likelihood principle combined with L-moments estimation to estimate the best proportion of data in each tail. As a sub product we obtain the thresholds. In summary, we let the data speak for themselves.

Related work can be found in Danielsson and de Vries (1997), where they propose a semiparametric approach combining GPDs and non-parametric empirical distribution.

Our procedure has the advantage to provide an analytical expression for the (mixture) distribution of the data, which may be used to simulate data and create scenarios, while providing an accurate estimation of quantiles associated with probabilities close to zero or one.

Simulations indicate that the proposed procedure is able to accurately estimate the entire distribution and to provide close estimates for the true proportions of data in the tails. We found that for extreme quantiles, especially those in regions where there are no observed data, the proposed procedure is clearly superior to non-parametric fits (when they exist).

In Section 2 we explain the proposed data driven procedure and show its usefulness when modeling the distribution of real financial data and computing risk measures. Another illustration is provided using measurements of oceanographic variables at a location off the southwest cost of England. In Section 3 we carry out Monte Carlo experiments to investigate the performance of the proposed procedure. In Section 4 we summarize our results and give our conclusions.

2. The data driven procedure

2.1. Data

The data type motivating this investigation, as described in the Introduction, are typically skew and fat tailed, containing a few extreme points, such as, for example, data from the area of finance. Fig. 1 shows on the left-hand side the histogram of data simulated from a mixture of 2% GPD on the left tail, 4% GPD on the right tail and 94% standard normal in the center.[1] On the right-hand side this figure shows the histogram of daily returns of a robustly standardized index representing the Mexican market.[2] We note some similarities between these distributions: high kurtosis, presence of extreme observations and skewness. The Mexico minimum and maximum standardized values are respectively -14.60 and 9.42, whereas for the simulated data these extreme values are -14.42 and 9.98.

2.2. Alternative methods

Empirical and non-parametric techniques (Silverman, 1986) for density estimation typically do a good job at the center of the data but are poor estimates for the tails. Moreover, being non-parametric they are not suitable for reproducing the data through simulations. Simulations play an important role in finance, for example, when investigating the likelihood of adverse complex scenarios.

Several authors have emphasized the poor quality of estimates of risk from normal-based measures (for example, Danielsson and de Vries, 1997; Susmel, 1998; Mendes, 2000). The t-Student distribution is the most used alternative. However, this distribution

[1] The GPD parameters used to simulate this data set in SPLUS are: shape (left) = 0.55, scale (left) = 1, shape (right) = 0.3, scale (right) = 1.

[2] Data obtained from the site of the Morgan Stanley Capital International, MSCI. Series length is 1350, from March/1995 to March/2000. From now on this series will be referred to as Mexico.

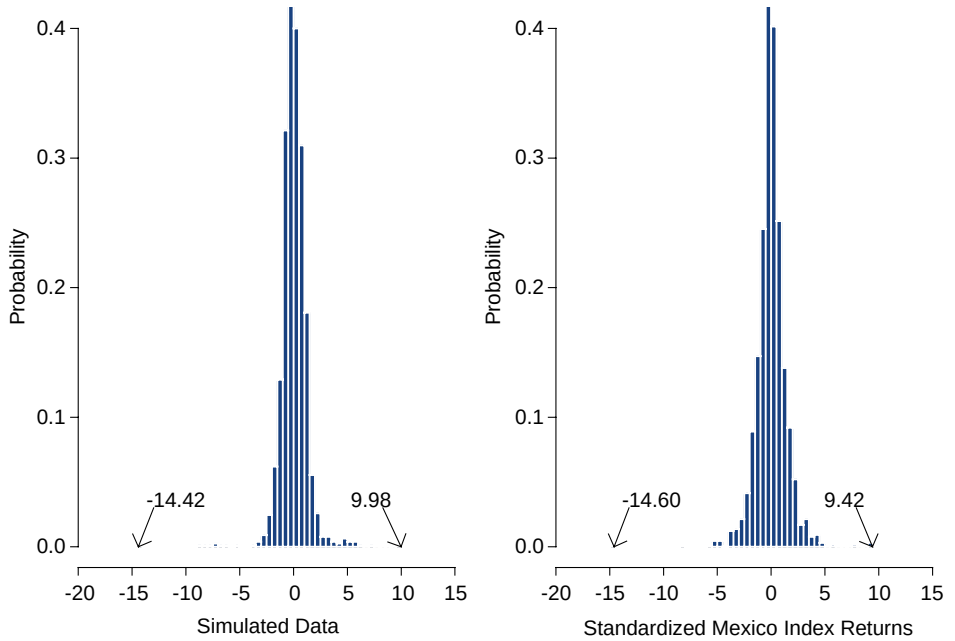

Fig. 1. Left: Histogram of data simulated from a mixture of 2% GPD on the left tail, 4% GPD on the right tail and 94% standard normal in the center. Right: Histogram of robustly standardized daily returns of Mexico index.

does not also handle another important characteristic of these data sets, its asymmetry (see Harvey and Siddique, 1999).

A main result from the EVT states that the generalized Pareto distribution appears as the limit distribution of scaled excesses over high thresholds. This means that the extreme tail (excesses beyond a high threshold) may be well modeled using the GPD, which has cumulative distribution function (cdf) $G_{\xi,\psi}(\cdot)$ given by

$$G_{\xi,\psi}(y) = \begin{cases} 1 - \left(1 + \dfrac{\xi y}{\psi}\right)^{-1/\xi} & \text{if } \xi \neq 0, \\ 1 - \exp\left(\dfrac{-y}{\psi}\right) & \text{if } \xi = 0, \end{cases} \tag{1}$$

where ψ is the scale parameter, ξ is the shape parameter, $y \geqslant 0$ when $\xi \geqslant 0$, and $0 \leqslant y \leqslant -\psi/\xi$ when $\xi < 0$. In expression (1) the location parameter is zero. For non-zero location just introduce a location parameter μ and write $(y-\mu)/\psi$ instead of y/ψ.

2.3. Data driven procedure

We assume that the data generating process may be a mixture of different structures, or made up from percentages of different distributions. For the bulk of the data we assume a left and right truncated normal. As a strong support for this assumption we

recall the famous Winsor's principle, quoted by Tukey (1960, p. 457): "all distributions are normal in the middle". The proportion of data in each tail, as well as the parameters of the GPDs, are not necessarily the same. Therefore, the left and the right truncation points (thresholds) are not necessarily symmetric about the center. By estimating the proportion of data in each tail providing the best overall fit we obtain the thresholds.

More formally, to accurately estimate the density of a unimodal possibly asymmetric random variable X with cdf F we propose a computer intensive maximum likelihood procedure to fit the model

$$F(x) = p_\ell^* G^\ell_{\xi_\ell,\psi_\ell}(x - t_\ell) + (1 - p_\ell^* - p_\mathrm{r}^*) H^{t_\mathrm{r}}_{t_\ell}(x - t_\mathrm{r}) + p_\mathrm{r}^* G_{\xi_\mathrm{r},\psi_\mathrm{r}}(x), \tag{2}$$

where p_ℓ^* and p_r^* are the proportions of data in the left and right tails, $G^\ell_{\xi_\ell,\psi_\ell}$ and $G_{\xi_\mathrm{r},\psi_\mathrm{r}}$ are the GPD models corresponding to the left and right tails, $G^\ell_{\xi_\ell,\psi_\ell}(x) = 1 - G_{\xi_\mathrm{r},\psi_\mathrm{r}}(x)$, and $H^{t_\mathrm{r}}_{t_\ell}$ may be a standard normal distribution or a t-Student with ν degrees of freedom truncated at t_ℓ and t_r, $t_\ell < 0$ and $t_\mathrm{r} > 0$. The data should be firstly robustly located at zero and scaled to one.

Let $y_1, y_2, \ldots, y_n$ be a set of independent and identically distributed observations and $\boldsymbol{y}$ its vector representation.[3] The procedure starts by robustly standardizing the data using the most robust pair of location and scale estimates (Huber, 1981; Hampel et al., 1986), the median (*med*) and the median absolute deviation (*mad*). We recall that the robust scale estimate *mad* is defined as $mad(\boldsymbol{y}) = med(\boldsymbol{e})$, where $\boldsymbol{e} = (e_1, e_2, \ldots, e_n)$, $e_i = |y_i - med(\boldsymbol{y})|$, and $|a|$ represents the absolute value of a. Thus, for the data $x_1, x_2, \ldots, x_n$, where $x_i = (y_i - med(\boldsymbol{y}))/mad(\boldsymbol{y})$, we fit model (2) by applying the data driven procedure presented in a frame in the next page.

At step 2 we set up two grids G_ℓ and G_r on $[0.0, 0.5)$ for the proportions of data p_ℓ and p_r on the left and right tails. In practice, the two grids can be defined on smaller intervals, such as $[0.0, 0.2]$.

At step 3 instead of setting up probability grids one could directly define the thresholds t_ℓ and t_r on the range of empirical quantiles.

At step 4 the GPD parameters are estimated using the L-moments (Hosking and Wallis, 1987) estimation procedure, which is more robust than the maximum likelihood method. They are convenient computationally efficient estimators, and for several distributions they yield closed-form expressions. According to Hosking et al. (1985) and Hosking and Wallis (1987), for small and moderate sample sizes, the L-moments estimators are more efficient than maximum likelihood.

The distribution H at step 5 represents either the cdf of the standard normal (Φ) or the standard t-Student with ν degrees of freedom (t_ν).

At steps 6 and 7 we form the mixture density $\hat{f}$ based on model (2) using the set of estimates $\{p_\ell^*, p_\mathrm{r}^*, t_\ell, t_\mathrm{r}, \xi_\ell, \psi_\ell, \xi_\mathrm{r}, \psi_\mathrm{r}\}$, and compute the corresponding estimated cdf $\hat{F}$. Two optimality criteria were used. The first criterion chooses the pair $(p_\ell^*, p_\mathrm{r}^*)$ that maximizes, over all possible pairs of proportions, the log-likelihood $\hat{l} = \log(\hat{f})$ of the data. This may be seen as a quasi-maximum likelihood criterion (similarly to Engle

[3] In the general case one may assume that the data present time dependence in the mean and in the variance. In this case we may fit a suitable time series model and apply the proposed procedure on the residuals.

and González-Rivera (1991) in the case of GARCH models), since it was assumed a normal (or t-Student) distribution for the center of the data even knowing that the data may not be normal (or t-Student).[4] We denote this data driven procedure as DDP.

Maximum likelihood estimation of mixture models has been investigated at length in the past few decades (Titterington et al., 1985, Chapter 4). Its limited usage may be due to identifiability problems, unbounded solutions, complexity. Moreover, obtaining asymptotic properties and solving computational issues may not be straightforward. It is worth mentioning that mixture of models have been extensively studied within the Bayesian framework. For example, the number of components of the mixture is exhaustively discussed in Richardson and Green (1997), and in the seminal work of Diebolt and Robert (1994).

The second optimality criterion chooses the pair (p_ℓ^*, p_r^*) that minimizes, over all possible pairs of proportions, the mean squared distance (MSD) between $\hat{F}$ computed at each observed data point and the empirical cumulative distribution function (EMP, in our tables below).

Data driven procedure (DDP)

1. Order the data. For the sake of simplicity we continue to denote by $(x_i, i = 1,\ldots,n)$ the ordered data. Thus, $x_1 \leqslant x_2 \leqslant \cdots \leqslant x_n$.
2. Set up two grids on $[0.0, 0.5)$, G_ℓ and G_r, for the proportions of data p_ℓ and p_r on the left and right tails.
3. For all pairs (p_ℓ, p_r), $p_\ell \in G_\ell$, and $p_r \in G_r$, we compute the left and right empirical thresholds as

$$t_\ell = x_j, \quad \text{where } j = [n * p_\ell],$$
$$t_r = x_j, \quad \text{where } j =]n * (1 - p_r)[,$$

   where $[a]$ represents the largest integer smaller or equal than a and $]a[$ represents the smallest integer greater or equal than a.
4. Fit the $G^{\ell}_{\xi_\ell,\psi_\ell}$ to the left tail data $\{x_j,$ such that $x_j \leqslant t_\ell\}$, and fit the G_{ξ_r,ψ_r} to the right tail data $\{x_j,$ such that $x_j \geqslant t_r\}$. The shape and scale parameters are obtained using the L-moments procedure. Note that the location parameters are the thresholds t_ℓ and t_r.
5. Compute $p_\ell^* = H(t_\ell)$ and $p_r^* = 1 - H(t_r)$.
6. Form the mixture density $\hat{f}$ based on model (2) using the set of estimates $\{p_\ell^*, p_r^*, t_\ell, t_r, \xi_\ell, \psi_\ell, \xi_r, \psi_r\}$. Compute estimated mixture cumulative distribution function $\hat{F}$, and evaluate the log-likelihood $\hat{l} = \log(\hat{f})$ of the data.
7. Final solution is the pair (p_ℓ^*, p_r^*) (and corresponding parameters $t_\ell, t_r, \xi_\ell, \psi_\ell, \xi_r, \psi_r$, and distribution estimates $\hat{F}$ and $\hat{f}$), satisfying the optimality criterion: Choose the pair (p_ℓ^*, p_r^*) that maximizes, over all possible pairs of proportions, the log-likelihood $\hat{l}$ of the data.

[4] A more rational justification for the name might be just that it is a mixture of different procedures, as opposed to true maximum likelihood which would estimate all parameters (including the degrees of freedom v in the case of a t_v) in the same way.

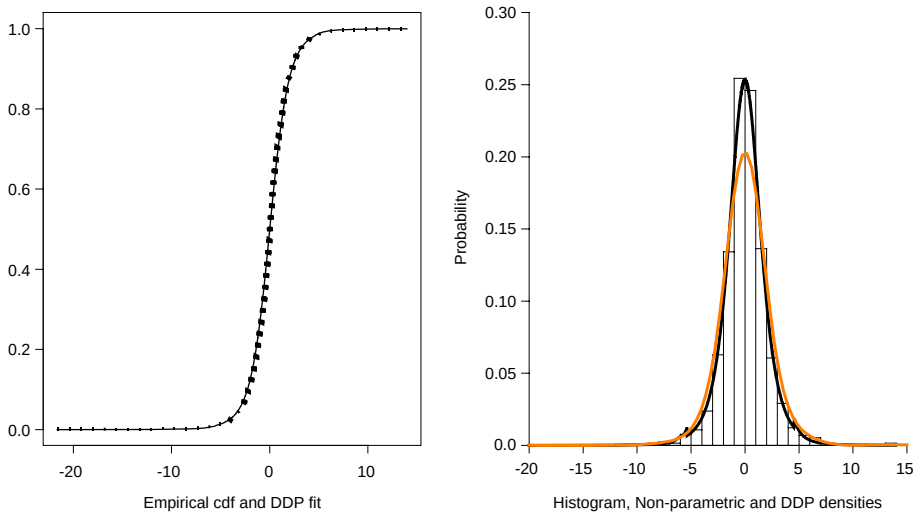

Fig. 2. DDP, empirical, and non-parametric fits for the robustly standardized Mexico index. The plot on the left-hand side shows the empirical (dotted) and DDP (solid) cdf. The plot on the right-hand side shows the DDP density (black) and the non-parametric density (lighter) superposed to the histogram.

Fig. 2 shows, for the robustly standardized Mexico index, the good quality of the DDP fit based on a t-Student with 4 degrees of freedom for the center of the data. The point estimates are $p_\ell^* = 0.01102$, $p_r^* = 0.01923$, $t_\ell = -5.39069$, $t_r = 4.50460$, $\xi_\ell = 0.61129$, $\psi_\ell = 0.93769$, $\xi_r = 0.44911$, $\psi_r = 0.93769$, $\hat{l} = -2654.84$). On the left-hand side we show the empirical cdf and the DDP distribution function $\hat{F}$. On the right-hand side we show the histogram of the data superposed by the DDP estimated density $\hat{f}$ (orange), and by a non-parametric fit (green). The non-parametric fit was obtained using the SPLUS function "density". This is essentially a smoothing operation using kernel estimates. We experimented with its arguments and found a good fit by setting the "cosine" window and the number of points used in the computations, "n", equal to 100.

Table 1 shows the quantiles obtained from the DDP fit and the empirical ones for the Mexico index. The minimum and maximum of this data set (before standardization) are, respectively, -21.76 and 14.10. In finance, an important risk measure is the unconditional value-at-risk (VaR), which is simply a quantile on the left tail of the returns distribution of a portfolio or an index. For example, the 1% quantile is the 1%-VaR, an event with return period of approximately 100 business days. In order to set aside capital in reserve to cover extreme losses, reliable estimates of extreme events associated with very low probabilities are needed. For this data set we cannot empirically estimate the 0.01%-VaR, but we can offer a DDP estimate of -31.24.

It would be interesting to contrast the DDP and empirical quantile estimates to classical procedures and EVT estimates. Table 1 also gives the normal and the t_4

Table 1
Quantile estimates for Mexico index

	0.01%	0.1%	1.0%	5.0%	25.0%	75.0%	95.0%	99.0%	99.9%	99.99%
EMP	—	−16.87	−5.48	−3.00	−0.98	1.03	3.16	5.56	12.77	—
DDP	−31.24	−10.58	−5.52	−3.18	−1.10	1.10	3.18	5.26	10.42	24.95
Normal	−7.82	−6.49	−4.87	−3.43	−1.37	1.48	3.54	4.98	6.60	7.93
Classical t_4	−27.55	−15.14	−7.88	−4.46	−1.51	1.62	4.57	7.99	15.25	27.66
Robust t_4	−19.41	−10.67	−5.56	−3.15	−1.07	1.14	3.22	5.63	10.74	19.48
EVT	−48.79	−22.05	−8.55	—	—	—	—	7.19	15.55	39.43

Negative quantiles are the unconditional value-at-risk.

based quantiles with location and scale estimated using the classical sample mean and sample standard deviation. More robust quantiles estimates based on the t_4 with the median and mad (constant = 1.4826) estimates are also given. The EVT notation refers to the estimates proposed in Danielsson and de Vries (1997). In this case, the quantiles are obtained by combining the observed proportion of data in tails and GPD estimates for the tails. These estimates are very sensitive to the choice of number of observations in the tails. Estimates given in Table 1 are based on 14 and 21 observations on the left and right tail, respectively. This choice makes the thresholds equal to −5.42 and 5.00, and the proportions of data in tails equal to 1.07% and 1.61%, similar to those used by the DDP estimates. Of course, it is possible to EVT estimate only quantiles associated to probabilities smaller than the chosen proportions in the tail.

In order to compare the estimates given in Table 1 we observe that the empirical and the EVT estimates are unable to estimate certain quantiles. The normal assumption with maximum likelihood estimates underestimate quantiles on the extreme tails. The t-Student assumption combined with the robust mad scale estimate may underestimate extreme quantiles since scale is not inflated. Classical (and robust) t_4 estimates, as well as the normal quantiles do not allow for asymmetric tails. DDP estimates differentiate the left and right tails and seem to agree with others for probabilities ranging between 5% and 95%.

Following a referee's suggestion we provide another illustration using a data set from the recent book by Stuart Coles (Coles, 2001, Example 1.10, p. 13). As commented already, an interesting research topic is the selection of thresholds in GPD analyzes. Even though this is not the main objective of our procedure, we do obtain the thresholds as a by-product. We compare our results to those obtained by the procedure suggested in Coles (2001), where the main interest is the selection of thresholds.

The original data set is a 2894 by 2 matrix giving simultaneous measurements of oceanographic variables at a location off the southwest cost of England. We use just the second column which gives surge height measured in meters and taken as typical over a 15 h time window. We first robustly standardize the data using the median and mad with constant equal to one. The minimum, first quartile, third quartile and maximum of the standardized data are, respectively, −4.333, −0.963, 1.057, and 8.816, indicating right asymmetry.

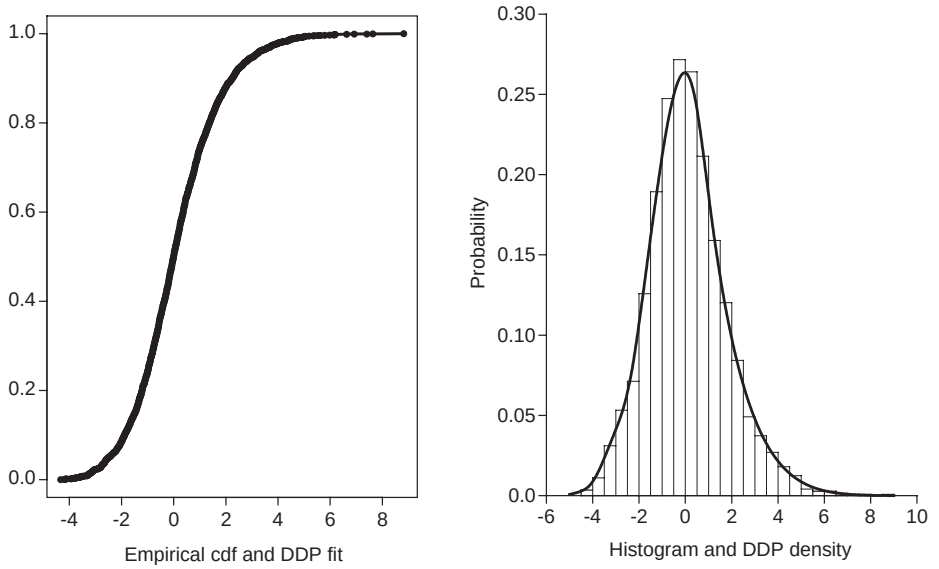

Fig. 3. DDP fit for the robustly standardized surge data. The plot on the left-hand side shows the empirical (points) and DDP cdf (black line). The plot on the right-hand side shows the DDP density superposed to the histogram.

The DDP estimates of left and right thresholds are, respectively, -2.2184 and 0.3678. The corresponding proportions of data in the tails are, respectively, 0.0673 and 0.3990. The good quality of the DDP fit may be seen in Fig. 3 where we show the fitted cdf and estimated density.

We now apply the graphical procedure given in Coles (2001) to help selecting the thresholds. This tool plots estimates for the scale and shape GPD parameters together with their sampling errors for varying threshold values. We should look for a change on the pattern of the estimates. Applying this procedure to both tails we found thresholds (and corresponding GPD estimates) very close to the ones found by the DDP procedure. For the right tail, the graphical tool indicates a threshold close to 0.50, and for the left tail it seems to indicate a value around -2.00. We show in Fig. 4 this plot for the left tail. For example, the plot indicates a value of approximately -0.40 for ξ, and the DDP estimate for ξ was -0.439.

We should recall that the two thresholds chosen by the DDP procedure were found jointly with the estimates for the center of the data. So, we should not expect exactly the same results from a method focusing only on the tails.

3. Simulations

To assess the performance of the proposed procedure we carried out Monte Carlo experiments. The programs were written using the publicly available R language and R

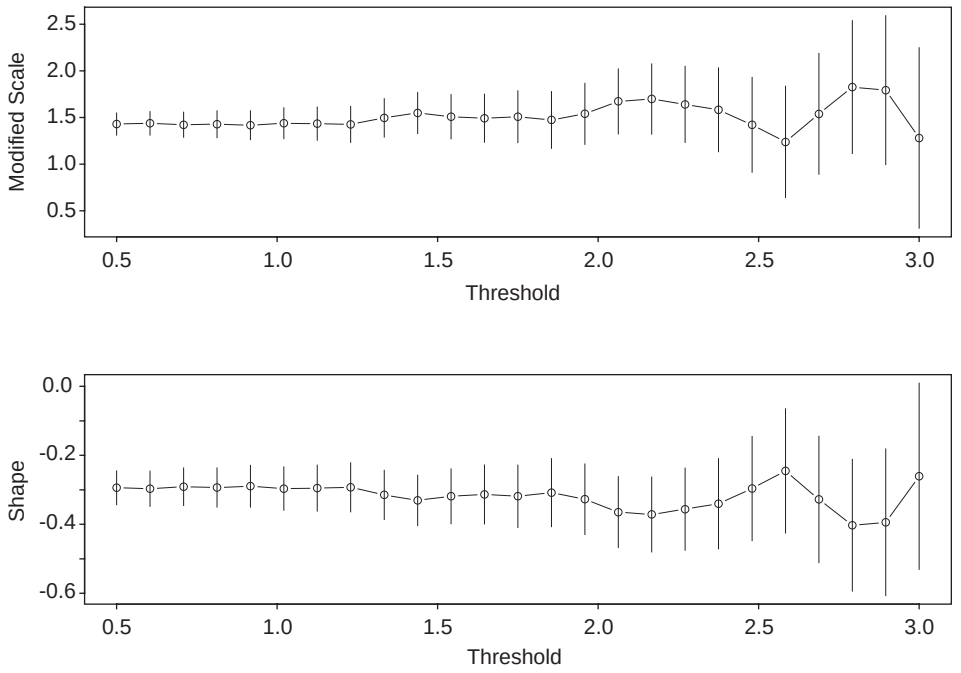

Fig. 4. Figure shows the absolute value of the threshold in the horizontal axis and, in the vertical axis, the corresponding estimates and sampling errors for the scale parameter (at top), and for the shape parameter (at bottom).

functions,[5] which do not require any dynamic loading and can be run on any personal computer. All data sets were generated from the mixture distribution

$$G = p_\ell G_{\xi_\ell,\psi_\ell} + (1 - p_\ell - p_\mathrm{r})H_{t_\ell}^{t_\mathrm{r}} + p_\mathrm{r} G_{\xi_\mathrm{r},\psi_\mathrm{r}}. \tag{3}$$

The bulk of the data were generated from a distribution H. The choices for H were the normal and the t-Student distributions. A proportion p_ℓ in the left extreme tail and a proportion p_r in the right extreme tail were generated from GPDs. We found that the scale parameter of the GPD did not have too much influence on the form of the mixture, so, we fixed it at 1. We experimented with different choices for the shape parameter and for the proportions. The proportions $p_\ell = p_\mathrm{r} = 0.0$ were also included. For each mixture type we simulated 100 series of lengths 1350 and 2500.[6] The series lengths are justified by the currently available sizes of financial daily data, which usually are ten or five years.

The output from the simulations included all parameters and proportions in (3) and a selection of quantiles. In the tables below we report the means, the standard deviations

[5] The R software and manuals can be freely downloaded from www.r-project.org. Our routines are available upon request.

[6] We do not report here the results for the larger data sets since they are pretty much the same as those obtained with the ones of size 1350.

Table 2
Mean, (standard deviation), and RMSE of proposed estimates from the first simulation experiment

		ξ_ℓ	ψ_ℓ	ξ_r	ψ_r	p^*_ℓ	p^*_r
DDP	Mean	0.5180	0.7583	0.3070	1.0110	0.0251	0.0238
	Std. deviation	(0.139)	(0.320)	(0.148)	(0.340)	(0.006)	(0.011)
	RMSE	0.1363	0.3950	0.1443	0.3313	0.0079	0.0106
MSD	Mean	0.5493	0.6116	0.4807	0.5538	0.0331	0.0337
	Std. deviation	(0.121)	(0.326)	(0.154)	(0.255)	(0.010)	(0.010)
	RMSE	0.1282	0.5019	0.2350	0.5108	0.0161	0.0167
TRUE		0.5000	1.0000	0.3000	1.0000	0.0200	0.0200

DDP uses first criterion. MSD uses second criterion.

and the root mean squared error (RMSE) of the results obtained. In these tables the notation DDP refers to the estimates obtained under the first criterion. The notation MSD refers to the estimates obtained under the second criterion.

The simulation results showed that the proposed procedure was able to accurately reproduce the entire distribution and to provide close estimates for the true proportions of data in the tails and for the GPD's parameters when the bulk of the data was either normal or t-Student.

3.1. First and second experiments

Data were generated from model $0.02G_{0.5,1} + 0.96\Phi^{t_r}_{t_\ell} + 0.02G_{0.3,1}$. So, in the first run of simulations H was the standard normal distribution, mixed with equal proportions of GPD data on the tails with same scale but different shape parameters. This choice of shape parameters will result in a longer left tail. We note that proportions smaller than 0.02 would make the GPD estimates very inefficient.

In Table 2 we give summaries of the parameters estimates from the simulations. For most parameters, the DDP criterion provided estimates closer to the true values, smaller standard errors and smaller RMSE when compared to the MSD. In Table 3 we observe that the DDP quantiles estimates were closer to the true ones and to the empirical quantiles estimates. As expected, the normal based quantiles using either the classical (mean and standard deviation) or robust (median and mad with tuning constant 1.4826) location-scale estimates under-estimated the extreme quantiles.

Data for the second experiment were generated from model $0.03G_{0.3,1} + 0.92\Phi^{t_r}_{t_\ell} + 0.05G_{0.3,1}$, that is, a normal center combined to different proportions of observations from GPDs with same scale and shape parameters in each tail. Results are omitted since they are basically the same as those in Tables 2 and 3. In summary, in the second experiment the DDP estimates showed smaller standard deviation and RMSE, and provided a good fit for the whole distribution.

Table 3
Results from the first simulation experiment: true and mean estimated quantiles

	Probabilities								
	0.0001	0.0010	0.0100	0.0200	0.0300	0.0400	0.0500	0.1000	0.1500
TRUE	−28.34	− 8.998	−2.882	−2.054	−1.881	−1.751	−1.645	−1.281	−1.036
DDP	−26.30	− 7.940	−2.758	−2.086	−1.875	−1.746	−1.640	−1.278	−1.033
MSD	−26.77	− 7.776	−2.719	−2.091	−1.866	−1.741	−1.640	−1.277	−1.033
EMP	NA	− 10.49	−2.910	−2.177	−1.917	−1.774	−1.681	−1.301	−1.040
N-cl	−5.453	− 4.533	−3.416	−3.017	−2.764	−2.574	−2.419	−1.888	−1.530
N-ro	−3.707	− 3.080	−2.318	−2.046	−1.874	−1.744	−1.639	−1.277	−1.032

	0.8500	0.9000	0.9500	0.9600	0.9700	0.9800	0.9900	0.9990	0.9999
TRUE	1.036	1.281	1.645	1.751	1.881	2.054	2.824	6.908	15.06
DDP	1.033	1.278	1.640	1.746	1.885	2.110	2.837	7.208	17.06
MSD	1.033	1.277	1.640	1.742	1.872	2.088	2.641	6.627	19.81
EMP	1.038	1.277	1.644	1.744	1.878	2.067	2.763	6.948	NA
N-cl	1.502	1.860	2.392	2.546	2.737	2.990	3.388	4.505	5.425
N-ro	1.034	1.278	1.641	1.746	1.876	2.048	2.320	3.081	3.708

Notation in table. TRUE: true model quantiles; DDP and MSD: mean estimates obtained under the first and second criterion, respectively; EMP: mean empirical estimate; N-cl and N-ro: mean estimates based on the Normal distribution using classical and robust estimates, respectively.

3.2. Third and fourth experiments

In the third run of simulations the model was $0.03G_{0.3,1} + 0.94H_{t_\ell}^{t_r} + 0.03G_{0.3,1}$, with H being the t-Student distribution with 5 degrees of freedom centered at zero and having scale equal to 1. In each tail we used the same proportion of observations from GPDs with same scale and shape parameters.

In steps 5 and 6 of the estimation procedure we used the t-Student distribution and considered the options of 4–7 degrees of freedom when optimizing both criteria. To assess the performance of the DDP and MSD estimates under the wrong assumption of a normal center, we also used this distribution. We found that, for this model, the normal assumption provided poor results. The procedure typically was not able to find the thresholds, in the sense that most of the runs the proportion of data in the tails would increase to close to 50%.

In Table 4 we give the results under the t-Student fit. The mean estimated number of degrees of freedom was 6. We note that most of the DDP estimates provided smaller RMSE and standard deviations. The proportions of data in the tails were not so close to the true ones and this might had reflected on the extreme quantile estimates (see Table 5). The quantiles estimates based on the other methods and given in Table 5 show the same deficiencies already noted in Table 3.

In the fourth run of simulations the model was $0.02G_{0.3,1} + (0.94)H_{t_\ell}^{t_r} + 0.04G_{0.3,1}$, with H being the t-Student distribution with 3 degrees of freedom with location zero and scale 1. Thus, in each tail, we had different proportions of observations from

Table 4
Mean, (standard deviation), and RMSE of proposed estimates from the third simulation experiment

		ξ_ℓ	ψ_ℓ	ξ_r	ψ_r	p^*_ℓ	p^*_r
DDP	Mean	0.3256	1.1858	0.3188	1.3646	0.0493	0.0477
	Std. deviation	(0.123)	(0.316)	(0.097)	(0.380)	(0.011)	(0.017)
	RMSE	0.1419	0.3052	0.1039	0.4899	0.0240	0.0247
MSD	Mean	0.3779	0.9959	0.3676	1.1598	0.0584	0.0553
	Std. deviation	(0.199)	(0.378)	(0.082)	(0.316)	(0.013)	(0.018)
	RMSE	0.1944	0.3981	0.1041	0.4327	0.0312	0.0312
TRUE		0.3000	1.0000	0.3000	1.0000	0.0300	0.0300

DDP uses first criterion. MSD uses second criterion.

Table 5
Results from the third simulation experiment: true and mean estimated quantiles

	Probabilities								
	0.0001	0.0010	0.0100	0.0200	0.0300	0.0400	0.0500	0.1000	0.1500
TRUE	−17.54	− 8.336	−3.723	−2.853	−2.422	−2.191	−2.015	−1.476	−1.156
DDP	−13.68	− 5.584	−2.394	−1.920	−1.708	−1.580	−1.429	−1.057	−0.832
MSD	−13.18	− 5.581	−2.378	−1.917	−1.698	−1.544	−1.426	−1.057	−0.832
MP	NA	−11.61	−3.776	−2.866	−2.442	−2.210	−2.027	−1.482	−1.152
N-cl	−5.526	− 4.589	−3.451	−3.045	−2.787	−2.593	−2.435	−1.894	−1.529
T-cl	−11.96	− 7.760	−4.683	−3.893	−3.447	−3.136	−2.896	−2.146	−1.690
T-ro	−5.890	− 3.822	−2.307	−1.917	−1.698	−1.544	−1.426	−1.057	−0.832

	0.8500	0.9000	0.9500	0.9600	0.9700	0.9800	0.9900	0.9990	0.9999
TRUE	1.156	1.476	2.015	2.191	2.422	2.853	3.723	8.336	17.54
DDP	0.832	1.057	1.429	1.580	1.708	1.926	2.405	5.810	14.12
MSD	0.832	1.057	1.426	1.544	1.698	1.920	2.422	5.935	13.79
EMP	1.168	1.475	2.031	2.194	2.466	2.910	3.814	8.286	NA
N-cl	1.560	1.925	2.467	2.624	2.818	3.076	3.482	4.621	5.558
T-cl	1.690	2.146	2.896	3.136	3.447	3.893	4.683	7.760	11.96
T-ro	0.832	1.057	1.426	1.544	1.698	1.917	2.307	3.822	5.890

Fit for the center of the data used the t-Student distribution. Notation in table. TRUE: true model quantiles; DDP and MSD: mean estimates obtained under the first and second criterion, respectively; EMP: mean empirical estimate; N-cl, T-cl, and T-ro: mean estimates based on the normal and t-Student(5) distributions together with classical and robust estimates, respectively.

GPDs with same scale and shape parameters. We do not show the results since they are basically the same obtained in the third simulation experiment.

As suggested by a referee, we carried out some simulations in which the assumed model is not the correct one. We generated data from the model $0.00G^{\ell}(\cdot,\cdot)+(1.00)H^{t_r}_{t_\ell}+$

$0.00G(\cdot,\cdot)$, with H being the normal and the t_4 distributions. We considered samples of size 1350 and also smaller samples of size 300.

We observed that in almost all simulations the proposed procedure was not able to find the thresholds in the sense that, the smallest the proportion of data in the tails, the greater the likelihood. In summary, the DDP estimates were robust enough to identify "no mixture", and kept the assumed central distribution for the whole data set. We do not report summaries of these simulations.

4. Conclusions

To estimate the probability distribution of a unimodal fat tailed random variable we proposed to fit a mixture model to a robustly standardized data set. The robust standardization distinguishes the bulk of the data from the tails by amplifying extreme points. We assumed that the bulk of the data is normal and used the appropriate EVT model, the GPD, on the well-defined tails. We provided very reasonable theoretical support for these assumptions and the simulations showed that they work well even when the bulk of the data is non-normal.

We used the maximum likelihood principle to estimate the best proportion of data in each tail. As a sub product we obtained the thresholds, the points where the tails begin. In summary, we let the data speak for themselves. The GPD parameters were estimated using the L-moments estimation procedure. The output is the analytical expression for the (mixture) distribution of the data, which may be used to simulate data and analyze extreme scenarios associated with probabilities close to zero or one.

We showed the DDP usefulness when modeling the distribution of real financial data and when computing risk measures. We observed a nicer DDP fit, especially in the tails, when compared to a poor non-parametric one.

The performance of the proposed data driven procedure was assessed by four Monte Carlo experiments. Our goal was to find the best overall possible fit, and not just a good fit for the bulk of the data, nor just a good fit for the tails. To this end we compared the estimated quantiles to the true ones. Our simulations indicated that the proposed data driven procedure works well for data containing observations from different structures. We found that for extreme quantiles, especially those in regions where there are no observed data, the proposed procedure is clearly superior to non-parametric and empirical fits.

The simulations studies assumed an specific model. Even though the model was general enough to generate symmetric and asymmetric data with different proportions of observations in the tails, it is limited to unimodal and bell-shaped distributions.

The procedure seems to be appropriate in situations when clearly the data are generated from different structures. An indication of this situation is the presence of a proportion of outliers in the data, or when the robust and classical estimates of location and scale are quite different. Its main usefulness is to provide, using a very flexible model, an analytical expression for the data distribution.

References

Coles, S., 2001. An Introduction to Statistical Modeling of Extreme Values, Springer Series in Statistics. Springer, Berlin.

Danielsson, J., de Vries, C.G., 1997. Value-at-risk and extreme returns. Working paper, Department of Economics, University of Iceland.

Danielsson, J., de Vries, C.G., 1998. Beyond the sample: extreme quantile and probability estimation. Paper provided by Financial Markets Group and ESRC in its Series FMG Discussion Papers, Number dp0298.

Diebolt, J., Robert, C.P., 1994. Estimation of finite mixture distributions by Bayesian sampling. J. Roy. Statist. Soc. B 56, 363–375.

Embrechts, P., Klüppelberg, C., Mikosch, T., 1997. Modelling Extremal Events for Insurance and Finance. Springer, Berlin.

Engle, R.F., González-Rivera, G., 1991. Semi-parametric estimation of ARCH model. J. Business Econom. Statist. 9 (4), 345–358.

Escobar, M.D., 1994. Estimating normal means with a Dirichlet process prior. J. Amer. Statist. Assoc. 89, 268–277.

Escobar, M.D., West, M., 1995. Bayesian density estimation and inference using mixtures. J. Amer. Statist. Assoc. 90, 577–588.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A., 1986. Robust Statistics: The Approach Based on Influence Functions. Wiley, New York.

Harvey, C., Siddique, A., 1999. Autoregressive conditional skewness. Preprint, Duke University. J. Financial Quantitative Anal. 34(4), 465–488

Hoaglin, D.C., Mosteller, F., Tukey, J.W., 1983. Understanding Robust and Exploratory Data Analysis. Wiley, New York.

Hosking, J., Wallis, J., 1987. Parameter and quantile estimation for the generalized Pareto distribution. Technometrics 29, 339–349.

Hosking, J., Wallis, J., Wood, E., 1985. Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. Technometrics 27, 251–261.

Huber, P.J., 1981. Robust Statistics. Wiley, New York.

Leadbetter, M., Lindgren, G., Rootzén, H., 1983. Extremes and Related Properties of Random Sequences and Processes. Springer, Berlin.

McLachlan, G.J., Peel, D., 2000. Finite Mixture Models. Wiley, New York.

McNeil, A.J., 1996. Estimating the Tails of Loss Severity Distributions using Extreme Value Theory, Department Mathematik, ETH Zentrum, Zürich.

McNeil, A.J., 1998. Calculating Quantile Risk Measures for Financial Return Series Using Extreme Value Theory, Department Mathematik, ETH Zentrum, Zürich.

McNeil, A., Frey, R., 1998. Estimation of Tail-Related Risk Measures for Heteroscedastic Financial Time Series: an Extreme Value Approach, Department Mathematik, ETH Zentrum, Zürich.

Mendes, B.V.M., 2000. Computing robust risk measures in emerging equity markets using extreme value theory. Emerging Markets Quart. 4 (2), 25–41.

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

Richardson, S., Green, P., 1997. On Bayesian analysis of mixtures with an unknown number of components. J. Roy. Statist. Soc. Ser. B 59, 731–792.

Roeder, K., Wasserman, L., 1997. Practical Bayesian density estimation using mixtures of normals. J. Amer. Statist. Assoc. 92, 894–902.

Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. Chapman & Hall, London.

Smith, R.L., 1987. Estimating tails of probability distributions. Ann. Statist. 15, 1174–1207.

Smith, R.L., 1999. Measuring risk with extreme value theory. Working paper, Department of Statistics, University of North Carolina, Chapel Hill.

Susmel, R., 1998. Switching volatility in Latin American emerging equity markets. Emerging Markets Quart. 2 (1), 44–56.

Titterington, D.M., Smith, A.F.M., Makov, U.E., 1985. Statistical Analysis of Finite Mixture Distributions, Wiley Series in Probability and Mathematical Statistics. Wiley, New York.

Tukey, J.W., 1960. A survey of sampling from contaminated distributions. In: Olkin, I., Ghurye, S.G., Hoeffding, W., Madow, W.G., Mann, H.B. (Eds.), Contributions to Probability and Statistics, Essays in Honor of Harold Hotelling. Stanford University Press, Stanford, CA, pp. 448–485.

West, M., 1991. Kernel density estimation and marginalization consistency. Biometrika 78, 421–425.

West, M., 1993. Approximating posterior distributions by mixtures. J. Roy. Statist. Soc. 55, 409–422.