Unsupervised threshold determination for hybrid models

Nehla Debbabi, Marie Kratz, Mamadou Mboup

UNSUPERVISED THRESHOLD DETERMINATION FOR HYBRID MODELS

Nehla DEBBABI [1,2]*, Marie KRATZ* [3] *and Mamadou MBOUP* [2]

[1] CReSTIC, Faculty of Exact and Natural Sciences, University of Reims Champagne Ardenne
[2] Research Lab. COSIM, Engineering College of Communications, University of Carthage
[3] ESSEC Business School Paris
nehla.debbabi@supcom.rnu.tn, kratz@essec.edu, mamadou.mboup@univ-reims.fr

ABSTRACT

A hybrid Gauss-Pareto model is considered for asymmetric heavy tailed data. The paper presents an unsupervised iterative algorithm to find successively the parameters of the Gaussian density and that of the Generalized Pareto distribution (GPD) with a continuity constrain on the hybrid density and its derivative at the junction point. Simulation results show that the proposed iterative algorithm provides reliable position for the junction point as well as an accurate estimation of the GPD parameters, compared to state of the art methods. Furthermore, a great advantage of the proposed method is that it can be adapted to any hybrid model.

***Index Terms*—** Heavy-tailed data modelling, Hybrid density estimation, Extreme Value Theory (EVT), Generalized Pareto distribution (GPD), Peak Over Threshold (POT) method

1. INTRODUCTION

Modelling non-homogeneous and multi-component data is a challenging problem that interests scientific researchers in several fields [1–4]. In general, it is not possible to find a simple and closed form probabilistic model to describe such data. Therefore, it seems natural to consider non-parametric appraoches, such as *e.g.* Kernel based density estimation [5] or non-parametric Bayesian methods [6, 7] to name few. However, when the multiple components are separable, parametric modelling becomes again tractable. Several hybrid models have been proposed in such context, combining two or more densities. For example in [1], a mixture of two different Gaussian distributions was given to model DNA microarray data. In [2], a hybrid model connecting a Log-Normal distribution with a Pareto one is proposed to model loss ratio in insurance. Another example is given in [3], where asymmetric heavy tailed data were modeled by a hybrid Pareto model linking a Gaussian distribution with a GPD. In this work we are interested on that type of asymmetric heavy tailed data, which are observed in different areas. For instance, in insurance, the distribution of the claims may be asymmetric heavy-tailed [8], as well as the conditional distribution of the profit and loss of a portfolio in Finance [3]. In signal processing, several problems can be resolved by considering the data as an asymmetric heavy-tailed one. Spike detection in neural signals [9] in biomedicine, energy detection for unknown signals over a fading channels [10] in telecommunication and bearing defect early detection in vibratory signals in machine diagnostics [11] are some such examples. For asymmetric heavy tailed data, the hybrid model is essentially introduced to estimate the whole distribution taking into account the tail. Now, the tail can be described via the Extreme Value Theory (EVT). Indeed, according to Pickands's theorem [12], the Peak Over Threshold (POT) method shows that the exceedances above a high threshold follow a Generalized Pareto Distribution (GPD) regardless the underlying distribution. This observation motivates the idea of separating the mean behaviour from the tail, by introducing a hybrid model, with a normal distribution around the mean and a GPD for the tail. Note that we chose the normal law around the mean, since it can be encountered in many situations by virtue of the Central Limit Theorem (CLT). In EVT, much work has been done on how to select and estimate a threshold above which the observations can be modeled by a GPD [9, 13, 14]. This is usually achieved using the method of the Hill plot [15], or more often, the mean excess plot method [13, 16]. The difficulty faced when applying these methods is that they are graphical ad hoc approaches. Moreover, it may take some time to determine the threshold since it requires hand-tuning. The offline solution of those methods represents an important disadvantage especially when complexity burden and/or delay processing are critical. A recent analytical study gives, however, an optimal way to separate mean and tail behaviours [17], providing thus an interesting alternative to the ad-hoc hand-tuning approaches. A combined statistical and numerical approach has been proposed [3, 18]. In [3], a hybrid model linking a normal distribution with a GPD in the right tail has been considered. The estimation of the model parameters was done iteratively using the maximum likelihood method. The initial parameters of the algorithm were based on the Hill plot estimation of the GPD parameters

from data above an arbitrary fixed threshold. In [18], the data were modeled by a Gaussian distribution with a GPD at the left and the right tails with given thresholds chosen as a small and a high quantile, respectively, and the GPD parameters estimated using the maximum likelihood method. Unlike existing methods, in this work we propose an unsupervised numerical approach that we illustrate when considering the hybrid normal-GPD model

$$h(x) = \begin{cases} f_{\mathcal{N}}(x), & \text{if } x \leq u \\ f_{GPD}(x), & \text{if } x > u \end{cases}$$

where $f_{\mathcal{N}}$ and f_{GPD} are the probability density functions (pdf) of the normal distribution and the GPD respectively and u is the junction point. Note that we chose to look only at the right tail, since it occurs in many applications, as for instance when looking at losses insurance [3], or at spike detection in neural signals [9]. Our approach is numerical only and consists in

- Fitting the normal distribution on the trimmed sample taking away the observations above the 68% quantile. Then evaluating the two parameters, mean and variance, of this truncated normal distribution, via the Levenberg Marquart algorithm [19, 20]. Those parameters will be kept to estimate the first threshold.
- Using an iterative procedure with a moving threshold representing the junction point between the normal and GPD distributions and fitting the hybrid distribution on the empirical one for each threshold. Thereafter, keeping the threshold that minimizes well the relative distance between the hybrid pdf and the empirical one. The first selected threshold will then be used for estimating the new Gaussian parameters below which, as in the first step, where those later will be used to estimate the next threshold.
- Continuing the iterative algorithm until the convergence to a fixed threshold with respect to some error.

It is important to note that the determination of the threshold is unsupervised, which is an advantage compared with the standard POT methods of threshold determination [13–15]. The proposed numerical algorithm provides a reliable evaluation of the threshold as well as the GPD parameters, when compared with statistical approaches, with a fast rate of convergence (only a few iterations are needed). Another great benefit of our method is that it can be adapted to any other hybrid model, hence it can resolve some non-homogenous data modeling problems.

The remainder of this paper is organized as follows. In section 2 we describe the proposed iterative algorithm for the unsupervised threshold determination. Simulation results are discussed in section 3. Conclusions follow in the last section.

2. DESCRIPTION OF THE PROPOSED NUMERICAL APPROACH FOR UNSUPERVISED THRESHOLD SELECTION

In this paragraph, we introduce the hybrid model that we will consider, linking a Gaussian distribution and a GPD at a junction point u, as follows.

$$h^{\mu,\sigma}_{\xi,\beta}(x,u) = \begin{cases} \gamma f_{\mu,\sigma}(x), & \text{if } x \leq u \\ \gamma g_{\xi,\beta}(x-u), & \text{if } x > u \end{cases}$$

where $\mu \in \mathbf{R}$ and $\sigma \in \mathbf{R}_+$ are the mean and the variance respectively of the Gaussian pdf $f_{\mu,\sigma}$ expressed as

$$f_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x-\mu)^2}{2\sigma^2}), \quad \forall x \in \mathbf{R}$$

while ξ and β represent the tail index and the scale parameter respectively of the GPD pdf $g_{\xi,\beta}$ defined by

$$g_{\xi,\beta}(x) = \begin{cases} \frac{1}{\beta}(1+\frac{\xi}{\beta}x)^{-1-\frac{1}{\xi}}, & \text{if } \xi \neq 0 \\ \frac{1}{\beta} exp(-\frac{x}{\beta}), & \text{if } \xi = 0 \end{cases} \quad x \in \mathcal{D}(\xi,\sigma)$$

on the following domain

$$\mathcal{D}(\xi,\sigma) = \begin{cases} [0,\infty), & \text{if } \xi \geq 0 \\ [0,-\frac{\beta}{\xi}], & \text{if } \xi < 0 \end{cases}$$

Here, γ is a regulator factor ensuring $\int_{\mathbf{R}} h(x)dx = 1$. To determinate this parameter, we express the hybrid pdf using the following formula

$$h^{\mu,\sigma}_{\xi,\beta}(x,u) = \gamma(1-H(x-u))f_{\mu,\sigma}(x) + \gamma H(x-u) g_{\xi,\beta}(x-u) \tag{1}$$

where H is the heaviside function defined as

$$H(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ 0, & \text{else.} \end{cases}$$

Hence, by integrating (1) we obtain

$$\int_{\mathbf{R}} h^{\mu,\sigma}_{\xi,\beta}(x,u)dx = \gamma F_{\mu,\sigma}(u) + \gamma \int_0^{+\infty} g_{\xi,\beta}(x)dx = 1$$

where $F_{\mu,\sigma}$ represents the cumulative distribution function (cdf) of the Gaussian distribution. Since $g_{\xi,\beta}$ is a pdf defined on $\mathbf{R}_+$, it integrates to one on that domain. Consequently, γ can be expressed as

$$\gamma = \frac{1}{F_{\mu,\sigma}(u) + 1}$$

An example of the hybrid pdf is illustrated in Fig. 1, where the difference between this later and the Gaussian distribution is clearly noticeable

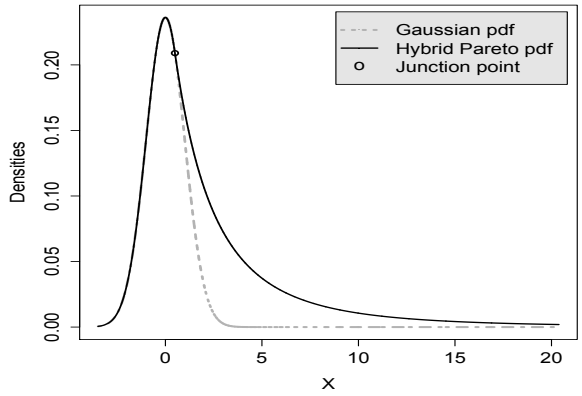

Fig. 1. Gaussian pdf (dotted curve) with parameters $\mu = 0$ and $\sigma = 1$ and hybrid pdf (continuous curve) with parameters $\mu = 0$, $\sigma = 1$ and $\xi = 0.2$

2.1. Fitting the normal distribution

Our numerical algorithm begins with estimating the parameters of the Gaussian distribution. To do so, we consider a truncated normal distribution taking away the extreme observations. We keep only observations under a fixed quantile that we denote by q. Thereafter we evaluate the mean and the variance of the Gaussian distribution from this trimmed sample using the Levenberg-Marquart method [19, 20] which is a robust method for nonlinear system resolution based on a minimization in the least squares sense. However, in our case, the likelihood estimators or other empirical estimators of the Gaussian parameters become non efficient since we manipulate a truncated part of the Gaussian distribution. Hence, the parameters μ and σ of the Gaussian distribution are obtained as a result of the following minimization problem

$$[\mu, \sigma] \leftarrow \underset{\substack{(\mu,\sigma)\in\mathbf{R}\times\mathbf{R}_+ \\ x\leq u=q}}{argmin} \parallel f_{\mu,\sigma}(x) - \widetilde{h}(x) \parallel_2^2 \qquad (2)$$

Here, $\widetilde{h}$ represents the empirical hybrid pdf using the kernel density estimation method [5] and is defined by

$$\widetilde{h}(x) = \frac{1}{nb_w} \sum_{i=1}^{n} K(\frac{x - x_i}{b_w})$$

Where K denotes the Gaussian kernel $K(x) = \frac{1}{\sqrt{2\pi}} \exp(\frac{-x^2}{2})$, n the number of the total samples including extremes, and b_w a smoothing parameter.

2.2. New iterative algorithm for fitting a GPD

In order to obtain a smooth pdf, the continuity and the derivability of the hybrid pdf at the junction point is enforced. We obtain then the following system of constraints

$$(C_1) \begin{cases} f_{\mu,\sigma}(u) = g_{\xi,\beta}(0) \\ f'_{\mu,\sigma}(u) = g'_{\xi,\beta}(0) \end{cases}$$

where $f'_{\mu,\sigma}$ and $g'_{\xi,\beta}$ are the derivatives of $f_{\mu,\sigma}$ and $g_{\xi,\beta}$ respectively. Knowing that

$$\begin{cases} f'_{\mu,\sigma}(x) = \frac{-(x-\mu)}{\sqrt{2\pi}\sigma^3} \exp(-\frac{(x-\mu)^2}{2\sigma^2}) = -\frac{(x-\mu)}{\sigma^2} f_{\mu,\sigma}(x) \\ g'_{\xi,\beta}(x) = \frac{-(\xi+1)}{\beta^2}(1 + \frac{\xi}{\beta}x)^{-2-\frac{1}{\xi}} \end{cases}$$

we transform the system (C_1) into

$$(C_2) \begin{cases} f_{\mu,\sigma}(u) = \frac{1}{\beta} \\ f'_{\mu,\sigma}(u) = -\frac{(u-\mu)}{\sigma^2} f_{\mu,\sigma}(u) = \frac{-(\xi+1)}{\beta^2} \end{cases}$$

By replacing $f_{\mu,\sigma}(u)$ in the second equation of (C_2) by $\frac{1}{\beta}$ we conclude the following relations between the threshold u, the parameters of the GPD and those of the Gaussian distribution

$$(C_3) \begin{cases} \beta = \frac{1}{f_{\mu,\sigma}(u)} \\ \xi = \beta \frac{(u-\mu)}{\sigma^2} - 1 \end{cases}$$

Those relations will be helpful for evaluating the hybrid model parameters. To achieve that, we propose the following new iterative procedure. Assuming that the Gaussian parameters are estimated (see section 2.1), we propose a moving threshold in the positive part of the distribution. For each fixed threshold, we compute the GPD parameters given in (C_3). Once ξ and β are estimated for this specific threshold, we introduce them in the hybrid model. Thereafter, using this model, we select the threshold value, denoted by u^*, minimizing the relative distance, according to a certain norm, between the hybrid pdf and the empirical one. In other words, u^* is the solution of the following minimization problem

$$u^* = \underset{u\geq 0}{argmin} \; N(h^{\mu_0,\sigma_0}_{\xi_u,\beta_u}(x,u) - \widetilde{h}(x))$$

where ξ_u and β_u are the GPD parameters relatives to u, and N is the norm to be minimized. The choice of the norm will be discussed in the section 3.4.

We chose the initial threshold arbitary at 0, but it could be chosen larger than the mean because of the nature of our hybrid model.

2.3. Summary of the proposed iterative algorithm

In this section we give the principal steps of the proposed iterative algorithm for unsupervised threshold selection

Step 1: Estimation of the empirical hybrid pdf $\widetilde{h}$ using the kernel density estimation method

Step 2: Choice of an arbitrary initial threshold $u_0 = q_{68}$ equal to the 68% quantile and estimation of the Gaussian parameters μ_0 and σ_0 from the data below u_0 using the method described in section 2.1

Step 3: Determination of the threshold u_1, when considering our hybrid model with μ_0 and σ_0

$$u_1 = \underset{u \geq 0}{argmin} \; N(h_{\xi_u,\beta_u}^{\mu_0,\sigma_0}(x,u) - \widetilde{h}(x))$$

Step 4: Iterative procedure: $\forall k \geq 1$

- μ_k and σ_k are estimated as in Step 2
- $u_{k+1} = \underset{u \geq 0}{argmin} \; N(h_{\xi_u,\beta_u}^{\mu_k,\sigma_k}(x,u) - \widetilde{h}(x))$
- The algorithm stops when the following condition is satisfied

$$\| \, u_{k+1} - u_k \, \| < \epsilon$$

Pseudocode of the algorithm

Algorithm 1 Iterative algorithm for the determination of the unsupervised threshold

1: Initialization of u_0, ϵ

2: Determination of the empirical pdf $\widetilde{h}$

$$\widetilde{h}(x) \leftarrow \frac{1}{nb_w} \sum_{i=1}^{n} K(\frac{x - x_i}{b_w})$$

3: Estimation of μ_0 and σ_0

$$[\mu_0, \sigma_0] \leftarrow \underset{\substack{(\mu,\sigma) \in \mathbf{R} \times \mathbf{R}_+ \\ x \leq u_0}}{argmin} \| \, f_{\mu,\sigma}(x) - \widetilde{h}(x) \, \|_2^2$$

4: Determination of the first threshold u_1

$$u_1 \leftarrow \underset{u \in \mathbf{R}_+}{argmin} \; N(h_{\xi_u,\beta_u}^{\mu_0,\sigma_0}(x,u) - \widetilde{h}(x))$$

5: Iterative process
$k \leftarrow 0$
while $\| \, u_{k+1} - u_k \, \| < \epsilon$

$$[\mu_{k+1}, \sigma_{k+1}] \leftarrow \underset{\substack{(\mu,\sigma) \in \mathbf{R} \times \mathbf{R}_+ \\ x \leq u_k}}{argmin} \| \, f_{\mu,\sigma}(x) - \widetilde{h}(x) \, \|_2^2$$

$$u_{k+1} \leftarrow \underset{u \in \mathbf{R}_+}{argmin} \; N(h_{\xi_u,\beta_u}^{\mu_k,\sigma_k}(x,u)) - \widetilde{h}(x))$$

$k \leftarrow k + 1$
end while
return $[\mathbf{u_{k+1}}, \; \mu_{\mathbf{k+1}}, \; \sigma_{\mathbf{k+1}}]$

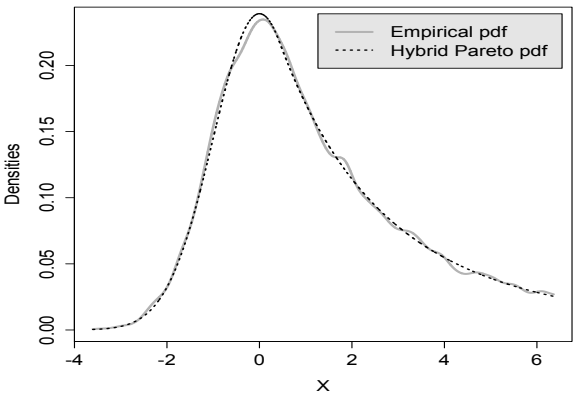

Fig. 2. Empirical density estimation using the kernel density estimation method

3. SIMULATION RESULTS AND DISCUSSION

To carry out the performance of our method, we apply it to a simulated data set following the hybrid model for a given μ, σ and ξ.

3.1. Hybrid pdf estimation

The Fig. 2 depicts the use of the kernel density estimtion method [5], for the estimation of the empirical hybrid pdf. Indeed, the black dotted curve represents a zoomed part of the hybrid pdf with parameters $\mu = 0$, $\sigma = 1$, and $\xi = 0.2$, while the gray continuous curve represents the empirical one. The empirical pdf appears to fit pretty well the hybrid one, on Fig. 2.

3.2. Evolution of the algorithm convergence

For $\mu = 0$, $\sigma = 1$ and $\xi = 0.2$, the convergence of the algorithm is drawn in Fig. 3. In order to detect clearly the evolution of the the threshold determination as well as the reliable estimation of the hybrid pdf, we represent in Fig. 3 only the fixed initial threshold u_0 (dashed vertical line), the first threshold u_1 (dotted vertical line), and the final one u_k (dash-dotted vertical line). Indeed, as u_0 is fixed arbitrary, it is obvious that the hybrid distribution using the determined parameters according to it (dashed curve) do not fit well the empirical one (black continuous curve). The hybrid pdf according to u_1 (dotted curve) fits better the empirical pdf since the threshold u_1 is closer to the exact value of the threshold (black vertical continuous line) than u_0 is. Accordingly, the closest the iterative threshold is from the optimal one, the best is the estimation of the hybrid pdf. Finally, we represent the final value of the threshold. As shown in the figure, u_k superposes with the threshold fixed value and the according hybrid

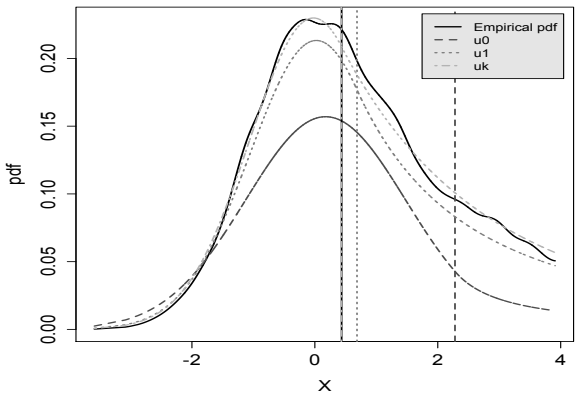

Fig. 3. Evolution of the convergence of the proposed threshold determination algorithm. Example with $\mu = 0$, $\sigma = 1$ and $\xi = 0.2$

pdf to u_k (dash-plotted curve) fit well the empirical one. This results ensures the reliable convergence of the proposed algorithm to the true value of the junction point.

3.3. Results

As already mentioned in the introduction, besides the reliable determination of the threshold, the proposed algorithm shows a good performance in terms of the GPD parameters estimation. To highlight this point, we compare the obtained results using the Hybrid Model (HM) with standard methods. Indeed, several methods in literature have been developed for the estimation of the tail index and the scale parameter. Most of them are statistical methods such as the MOments Method (MOM). This method is based on the use of the first and the second moment of the GPD. Knowing that the GPD distribtion admits a finite moment of order k if and only if $\xi < \frac{1}{k}$, the performances of the MOM decreases when $\xi \geq \frac{1}{2}$. To go further, a Probability Weighted Moments (PWM) method was proposed [21] extending the MOM when $\xi < 1$, but the estimation of the GPD parameters still unreliable for $\xi \geq 1$. Another statistical method well used in literature is the Maximum Likelihood Method (MLM) [21] based on the maximization of a likelihood function. The limitation of this method is that it can returns a local maximum instead of the global one. To overcome the encountered problem using the moments of the GPD, Zhang has proposed a specific method [22] for any value of ξ based on a bayesian analysis. Nevertheless, it remains an ad-hoc method. We show in the Table 1 that the hybrid model goes over the encountered problems with standard methods. The Table 1 illustrates two cases, the first case when $\xi = 0.5$ and the second one for $\xi = 1$. For both cases, the Gaussian parameters are $\mu = 0$ and $\sigma = 1$. Using our algorithm, we can identify the junction point. Thereafter, we compare the estimated GPD parameters using our method to those estimated via the standard methods from the observations above the determined junction point. Through the absolute error between the exact and the estimated parameters using different methods, we can easily infer the reliable estimation of the GPD parameters using the hybrid model.

Table 1. Comparison of the GPD parameters estimation

Param	HM	MOM	PWM	MLM	Zhang
Generated data with $\mu = 0$, $\sigma = 1$ and $\xi = 0.5$					
$\xi = 0.5$	0.4858	0.4589	0.3978	0.3390	0.3387
Error(ξ)	0.0141	0.0410	0.1021	0.1609	0.1612
$\beta = 2.8727$	2.902	3.4965	3.8915	3.9267	3.9275
Error(β)	0.0297	0.6238	1.0188	1.0540	1.0547
Generated data with $\mu = 0$, $\sigma = 1$ and $\xi = 1$					
$\xi = 1$	0.9974	0.4986	0.8858	0.7414	0.7415
Error(ξ)	0.0025	0.5013	0.1141	0.2586	0.2585
$\beta = 3.0905$	3.0798	17.5238	3.9910	4.5232	4.5239
Error(β)	0.0107	14.4333	0.9004	1.4327	1.4334

Table 2. Distance choice

Name	Definition
Relative Manhatten	$d_1(x,y) = \sum_{i=1}^{n} \mid \frac{x_i - y_i}{y_i} \mid$
Relative Euclidean	$d_2(x,y) = \sqrt{\sum_{i=1}^{n} \mid \frac{x_i - y_i}{y_i} \mid^2}$
Relative least squares	$d_3(x,y) = \frac{1}{n} \sum_{i=1}^{n} \mid \frac{x_i - y_i}{y_i} \mid^2$
Kullback-Leibler	$d_4(x,y) = \sum_{i=1}^{n} x_i \log \frac{x_i}{y_i}$
Bhattacharyya	$d_5(x,y) = -\log \sum_{i=1}^{n} \sqrt{x_i y_i}$
log	$d_6(x,y) = -\sum_{i=1}^{n} \log \frac{x_i}{y_i}$
Adapted Kullback-Leibler	$d_7(x,y) = \{\frac{d_4(x,y)}{\max(y)}\}_{x \leq u} + \{\frac{d_4(x,y)}{\max(y)}\}_{x > u}$

Table 3. Distance selection

Param	μ	σ	u	ξ	β	$t(s)$
Exact	0	1	0.4354	0.2	2.7559	x
d_1	0.0038	0.9953	0.4443	0.2234	2.7514	509.71
E_{d_1}	0.0038	0.0047	0.0089	0.0234	0.0045	x
d_2	0.0038	0.9952	0.4387	0.2053	2.7446	487.15
E_{d_2}	0.0038	0.0048	0.0033	0.0053	0.0113	x
d_3	0.0038	0.9952	0.4387	0.2053	2.7446	500.22
E_{d_3}	0.0038	0.0047	0.0033	0.0053	0.0113	x
d_4	0.0216	1.0079	0.1437	-0.6942	2.5451	516.69
E_{d_4}	0.0216	0.0079	0.2917	0.8942	0.2108	x
d_5	0.2105	1.3231	2.4499	16.7668	13.8898	1062.6
E_{d_5}	0.2105	0.3231	2.0145	16.5668	11.1339	x
d_6	0.0039	0.9953	0.4487	0.2379	2.7569	541.59
E_{d_6}	0.0039	0.0047	0.0133	0.0379	0.001	x
d_7	0.0048	0.9961	0.3621	-0.0412	2.6626	453.22
E_{d_7}	0.0048	0.0039	0.0733	-0.2412	0.0933	x

3.4. Choice of the norm

To enhance the performance of the proposed algorithm, we will determine carefully the distance that minimizes the error between the empirical pdf and the hybrid one. To achieve that, we tested different distances, described in Table 2, taking into account of the difference of the amplitude order between small and large order statistics. The reliability of the estimation and the execution time of the algorithm are given in Table 3 for the various distances described in the Table 1. As shown in this table, through the absolute error between the exact parameters of the simulated data and the estimated ones, denoted by $E_{d_i}\ \forall i \in \{1, ..., 7\}$, the best estimation was found using the relative Euclidean distance and that of the relative least squares distance. Both relative distances return the same estimation since they are related to each other. But, in terms of execution time, the relative Euclidean distance takes less time than the relative squares one to converge to the true threshold.

4. CONCLUSION

This paper proposes a new numerical method, with an iterative algorithm for unsupervised threshold detection for a hybrid model that links a Gaussian distribution to a GPD. Thanks to our simulation results, we see that we can obtain a reliable evaluation of the hybrid junction point with our method. Moreover, we have a fast convergence to this point. Besides the accurate determination of the threshold (or junction point), the GPD parameters appear better estimated than with standard methods. As a follow up, it would be interesting to compare our method with the ones detailed in [3] and [18]. Likewise, we are exploring a new hybrid model with two junction points introducing an intermediate behaviour between the Gaussian distribution and the GPD.

References

[1] A. Mandava, S. Latifi, and R. EmmaE, "Reliability assessment of microarray data using fuzzy classification methods: A comparative study," in *Advances in Computing and Communications*. 2011, vol. 190 of *Communications in Computer and Information Science*, pp. 351–360, Springer Berlin Heidelberg.

[2] M. Knecht and S. Küttel, "The czeledin distribution function," in *XXXIV ASTIN Colloquium*, August.

[3] J. Carreau and Y. Bengio, "A hybrid pareto model for asymmetric fat-tailed data: the univariate case," *Extremes*, vol. 1, pp. 53–76, 2009.

[4] M.E. Kuhl and P.S. Bhairgond, "Nonparametric estimation of nonhomogeneous poisson processes using wavelets," in *Simulation Conference, 2000. Proceedings. Winter*, 2000, vol. 1, pp. 562–571 vol.1.

[5] K. JooSeuk and D.S Clayton, "Robust kernel density estimation," *Journal of Machine Learning Research*, vol. 13, pp. 2529–2565, 2012.

[6] Stephen G. Walker, Paul Damien, Purushottam W. Laud, and Adrian F. M. Smith, "Bayesian nonparametric inference for random distributions and related functions," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 61, no. 3, pp. 485–527, 1999.

[7] P. Orbanz and Y.-W. Teh, "Bayesian nonparametric models," in *Encyclopedia of Machine Learning*. Springer, 2010.

[8] J. Carreau and Y. Bengio, "A hybrid pareto mixture for conditional asymmetric fat-tailed distributions," *Neural Networks, IEEE Transactions on*, vol. 20, no. 7, pp. 1087–1101, 2009.

[9] N. Debbabi, M. Kratz, M. Mboup, and S. El Asmi, "Combining algebraic approach with extreme value theory for spike detection," in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, 2012, pp. 1836–1840.

[10] F. F. Digham, M. S. Alouini, and M. K. Simon, "On the energy detection of unknown signals over fading channels," *Communications, IEEE Transactions on*, vol. 55, no. 1, pp. 21–24, 2007.

[11] Y. Yang, D. Yu, and J. Cheng, "A roller bearing fault diagnosis method based on {EMD} energy entropy and {ANN}," *Journal of Sound and Vibration*, vol. 294, no. 12, pp. 269 – 277, 2006.

[12] J. Pickands, "Statistical inference using extreme order statistics," *Annals of Statistics*, vol. 3, pp. 119–131, 1975.

[13] P. Embrechts, C. Klppelberg, and T. Mikosch, *Modelling Extremal Events for Insurance and Finance*, Springer-Verlag, 1997.

[14] S. Tanaka, K. Takara, A. Snorrason, H. Finnsdottir, and E. M. Moss, "A study on threshold selection in pot analysis of extreme floods," *IAHS-AISH publication*, , no. 271, pp. 299–304, 2002, eng.

[15] B. M. Hill, "A simple general approach to inference about the tail of a distribution," *The Annals of Statistics*, vol. 3, pp. 1163–1174, 1975.

[16] Y. Demichel, A. Estrade, M. Kratz, and G. Samorodnitsky, "How fast can the chord length distribution decay?," *Advances in Applied Probability*, vol. 43, pp. 504–523, 2011.

[17] M. Kratz, ," in *There is a VaR beyond usual approximations*. Preprint 2013.

[18] W. Gehin, "Modlisation des queues de distribution des rendements des actif financiers," M.S. thesis, EURIA, Brest, 2011.

[19] K. Levenberg, "A method for the solution of certain nonlonear problems in least squares quart," *Applied Math*, vol. 2, pp. 164–168, 1944.

[20] D. W. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *SIAM Journal on Applied Mathematics*, vol. 11, pp. 431–441, 1963.

[21] J. R. M. Hosking and J. R. Wallis, "Parameter and quantile estimation for the generalized pareto distribution," *Technometrics*, vol. 29, no. 3, pp. 339–349, 1987.

[22] J. Zhang and M. A. Stephens, "A new and efficient estimation method for the generalized pareto distribution," *Technometrics*, vol. 51, no. 3, pp. 316–325, 2009.