unibz
Fakultät für Wirtschaftswissenschaften
Facoltà di Economia
Faculty of Economics and Management

BEMPS –
Bozen Economics & Management
Paper Series

NO 108/2025

A new two-component hybrid model for highly right-skewed data: estimation algorithm and application to finance and rainfall data

Patrick Osatohanmwen

A new two-component hybrid model for highly right-skewed data: estimation algorithm and application to finance and rainfall data

Patrick Osatohanmwen

Faculty of Economics and Management, Free University of Bolzano, Italy

profpato2014@gmail.com; Patrick.Osatohanmwen@unibz.it

Abstract

In many real-life processes, data with high positive skewness are very common. Moreover, these data tend to exhibit heterogeneous characteristics in such a manner that using one parametric univariate probability distribution becomes inadequate to model such data. When the heterogeneity of such data can be appropriately separated into two components: the main innovation component, where the bulk of data is centered, and the tail component which contains some few extreme observations, in such a way, and without a loss in generality, that the data possesses high skewness to the right, the use of hybrid models becomes very viable to model the data. In this paper, we propose a new two-component hybrid model which joins the half-normal distribution for the main innovation of a highly right-skewed data with the generalized Pareto distribution (GPD) for the observations in the data above a certain threshold. To enhance efficiency in the estimation of the parameters of the hybrid model, an unsupervised iterative algorithm (UIA) is adopted. An application of the hybrid model in modeling the absolute log returns of the S&P500 index and the intensity of rainfall which triggered some debris flow events in the South Tyrol region of Italy is carried out.

Keywords: estimation algorithm, generalized Pareto distribution, half-normal distribution, hybrid model, S&P500

JEL Classification: $C02$

MSC Classification: $60G70 \cdot 62E20 \cdot 62F35 \cdot 62P05 \cdot 62P10 \cdot 65D15 \cdot 68W4$

1 Introduction

Skewed distributions are common in real-world processes, particularly when extreme values or outliers significantly influence standard statistical approaches. The analysis of skewed data is critical in various fields, including finance and insurance (Blum and Dacorogna, 2003; Embrechts et al., 1997; Carreau and Bengio, 2009; Dacorogna and Kratz, 2015), communication and signal processing (Rangaswamy et al., 2004; Digham et al., 2007; Broadwater and Chellapa, 2010; Mandava et al., 2011; Sermpezis and Spyropoulos, 2015) and environmental science (Rossi et al., 1984; Davison and Smith, 1990; Furrer and Katz, 2008; Singh et al., 2012; Kollu et al., 2012). In many instances, data characterized by high positive skewness can exhibit heterogeneity, where a single parametric univariate probability distribution falls short in effectively modeling such data. When dealing with highly right-skewed data, it is necessary to understand the asymmetry of observed data, which can be decomposed into two or more distinct components (Osatohanmwen et al., 2024). For the case of two components, the first component is the main innovation, representing where the bulk of the data is centered, while the second is the tail component, which captures the few extreme observations above a certain threshold that contribute to the skewness. Failing to account for this heterogeneity could lead to significant loss in efficiency when modeling the data. Moreover, The modeling of this type of data has received widespread attention in recent years and many models/methods have been put forth including non-parametric models (Guillen et al., 2005) and the Peak over Threshold (PoT) methodology (Pickands, 1975; Davison and Smith, 1990). While the non-parametric models usually offer good fits to the data, they typically fail to account for a few outlying observations in the tail. On the other hand, the PoT methodology only focuses on the extreme observations beyond a certain threshold while ignoring the rest observations in the data and thus does not make use of the entire distribution.

Given these challenges, hybrid models have emerged as effective solutions for modeling complex distributions. A hybrid model combines two or more probability distributions to adequately fit the characteristics of observed data. Several families of two-component hybrid models have been defined and studied in the literature (Cooray and Ananda, 2005; Scollnik, 2007; Carreau and Bengio, 2009; Cooray, 2009; Cooray et al., 2010; Singh et al., 2012; Scollnik and Sun, 2012; Nadarajah and Bakar, 2014; Bakar et al., 2015). In this paper paper, a new two-component hybrid model for data exhibiting high skewness to the right is introduced. The model links a half-normal distribution and a GPD at a certain threshold point determined when imposing a condition of class $\mathcal{C}^1$ (Carreau and Bengio, 2009; Debbabi and Kratz,

2014). Furthermore, the two components of the hybrid distribution are weighted non-uniformly and an unsupervised iterative and convergent estimation scheme based on the Levenberg-Marquardt (L-M) algorithm (Levenberg, 1944; Marquardt, 1963) is adopted to estimate the threshold point and in addition, other free parameters of the two-component model. In standard PoT methodology, this threshold point is usually estimated graphically whereas in this framework it is a parameter to be estimated in the hybrid model. This allows us to determine the point beyond which the extremes are observed algorithmically.

In Section 2, the specification of the two-component non-uniform weights hybrid model framework and the half-normal-GPD model is carried out. A description of the UIA for the estimation of the hybrid model's free parameters is presented in Section 3. Results from numerical studies based on Monte Carlo simulations conducted to assess the UIA's efficiency in estimating the hybrid model's parameters are reported in Section 4. In Section 5, the application of the new hybrid model in fitting the absolute log returns data of the S&P500 index and the data on the intensity of rainfall that triggered some debris flow events in the South Tyrol region of Italy, is performed. A conclusion is used to close the paper in Section 6.

2 Two-component non-uniform weights hybrid model

Suppose we have a data which can be decomposed into two components and the components represent specific behavior of the dichotomized data, and the goal is to use a smooth piecewise probability density function (pdf) to model the data. Assume the data is continuous and follows a non-degenerate distribution. Let f_1 and f_2 be two pdfs each with parameter vector Θ_1 and Θ_2 in such a manner that each of the pdf is suitable for modeling specific component of the data and without a loss in generality, f_1 and f_2 are suitable for modeling the first and second component of the data respectively. Suppose F_1 and F_2 are the respective cumulative distribution functions (cdfs) corresponding to f_1 and f_2, with respective corresponding quantile functions $Q_1(p;\Theta_1)$ and $Q_2(p;\Theta_2)$ where $Q_i(p;\Theta_i) = \inf\{w; F_i(w;\Theta_i) > p\}, 0 < p < 1$. The general two-component non-uniform weights hybrid model for the data can be specified by the pdf of the form

$$f(x;\Theta) = \begin{cases} w_1 f_1(x;\Theta_1), & -\infty < x \leq u, \\ w_2 f_2(x;\Theta_2), & u \leq x < \infty, \end{cases} \tag{1}$$

where Θ is a vector of free parameters in the model, w_1 and w_2 ($w_1 \neq w_2$) are weights associated with the respective component of the pdf in (1) with $(w_1, w_2) \in [0,1]^2$ and u is a junction point or threshold indicating the point of transition from one component or behavior of the data to another.

Suppose that in the pdf in (1) the transition from one component to another is smooth, the following assumptions are made:

(i) The pdf f is positive and satisfies

$$\int_{\mathbb{R}} f(x;\Theta)dx = 1,$$

inferring that

$$w_1 F_1(u;\Theta_1) + w_2\left[1 - F_2(u;\Theta_2)\right] = 1.$$

(ii) The distribution of the data has a heavy right tail belonging to the Fréchet maximum domain of attraction.

(iii) The pdf f is continuous and differentiable at the threshold u and in addition, smooth and $\mathcal{C}^1$-regular, implying that

$$\begin{aligned} w_1 f_1(u;\Theta_1) &= w_2 f_2(u;\Theta_2), \\ w_1 f_1^{'}(u;\Theta_1) &= w_2 f_2^{'}(u;\Theta_2). \end{aligned}$$

Given these assumptions, we obtain

$$\begin{cases} w_1 = w_2 \frac{f_2(u;\Theta_2)}{f_1(u;\Theta_1)}; \\ w_2 = \left\{ \frac{F_1(u;\Theta_1) f_2(u;\Theta_2)}{f_1(u;\Theta_1)} + 1 - F_2(u;\Theta_2) \right\}^{-1}. \end{cases} \tag{2}$$

The pdf in (1) has cdf and corresponding quantile function given respectively by

$$F(x;\Theta) = \begin{cases} w_1 F_1(x;\Theta_1), & -\infty < x \leq u, \\ 1 - w_2\left[1 - F_2(x;\Theta_2)\right], & u \leq x < \infty, \end{cases} \tag{3}$$

$$Q(p;\Theta) = \begin{cases} Q_1\left(\frac{p}{w_1};\Theta_1\right), & \text{if} \quad p \leq w_1, \\ Q_2\left(\frac{p-(1-w_2)}{w_2};\Theta_2\right), & \text{if} \quad p \geq 1-w_2. \end{cases} \tag{4}$$

Remark 1. *One can simulate random samples from the model in (1) using (4) by simply replacing p with the random variable U, where U is uniform on $(0,1)$.*

To define a two-component half-normal-GPD hybrid model, f_1 is taken to be the half-normal distribution with pdf, cdf and quantile function expressed respectively as

$$\begin{aligned} f_1(x;\sigma) &= \frac{\sqrt{2}}{\sigma\sqrt{\pi}}\exp\left(-\frac{x^2}{2\sigma^2}\right) \quad x \in \{0\} \cup \mathbb{R}_+ \quad \sigma \in \mathbb{R}_+, \\ F_1(x;\sigma) &= \operatorname{erf}\left(\frac{x}{\sigma\sqrt{2}}\right) \quad x \in \{0\} \cup \mathbb{R}_+ \quad \sigma \in \mathbb{R}_+, \\ Q_1(p;\sigma) &= \sigma\sqrt{2}\operatorname{erf}^{-1}(p) \quad \sigma \in \mathbb{R}_+ \quad 0 < p < 1, \end{aligned}$$

where σ is a scale parameter, $\operatorname{erf}(\cdot)$ is the error function and $\operatorname{erf}^{-1}(\cdot)$ is its inverse. Furthermore, take f_2 as the GPD with pdf, cdf and quantile function expressed respectively as

$$\begin{aligned} f_2(x-u;\beta,\gamma) &= \frac{1}{\beta}\left(1+\gamma\frac{x-u}{\beta}\right)^{-1-\frac{1}{\gamma}}, \quad \beta \in \mathbb{R}_+ \quad \gamma \in \mathbb{R}, \\ F_2(x-u;\beta,\gamma) &= 1-\left(1+\gamma\frac{x-u}{\beta}\right)^{-\frac{1}{\gamma}}, \quad \beta \in \mathbb{R}_+ \quad \gamma \in \mathbb{R}, \\ Q_2(p;u,\beta,\gamma) &= \frac{\beta}{\gamma}\left[(1-p)^{-\gamma}-1\right]+u, \quad \beta \in \mathbb{R}_+ \quad \gamma \in \mathbb{R} \quad 0 < p < 1, \end{aligned}$$

$$\forall x \geq u \in Z(\beta,\gamma), \quad Z(\beta,\gamma) = \begin{cases} [0,\infty) & \text{if} \quad \gamma \geq 0 \\ [0,-\beta/\gamma) & \text{if} \quad \gamma < 0, \end{cases}$$

where β is a scale parameter and γ is the tail index parameter which controls the shape of the GPD.

Using assumption (i)-(iii) we obtain the following relations for some of the parameters of the distribution:

$$\begin{cases} w_1 = \frac{w_2}{\beta f_1(u;\sigma)}; \\ w_2 = \left\{1 + \frac{F_1(u;\sigma)}{\beta f_1(u;\sigma)}\right\}^{-1}; \\ \beta = -(1+\gamma)\frac{f_1(u;\sigma)}{f_1'(u;\sigma)}. \end{cases} \tag{5}$$

It follows that the vector of parameter Θ will contain only the free parameters including the threshold u. Thus $\Theta = [\sigma, u, \gamma]$. These would be the parameters to be estimated using the proposed UIA. Once Θ has been estimated the estimates of the parameters w_1, w_2 and β can easily be realized from (5).

Remark 2. *Observe that in the half-normal-GPD hybrid model, there were six parameters whose values we needed to estimate. However, with the assumptions (i)-(iii), the number of free parameters to be estimated were reduced to three.*

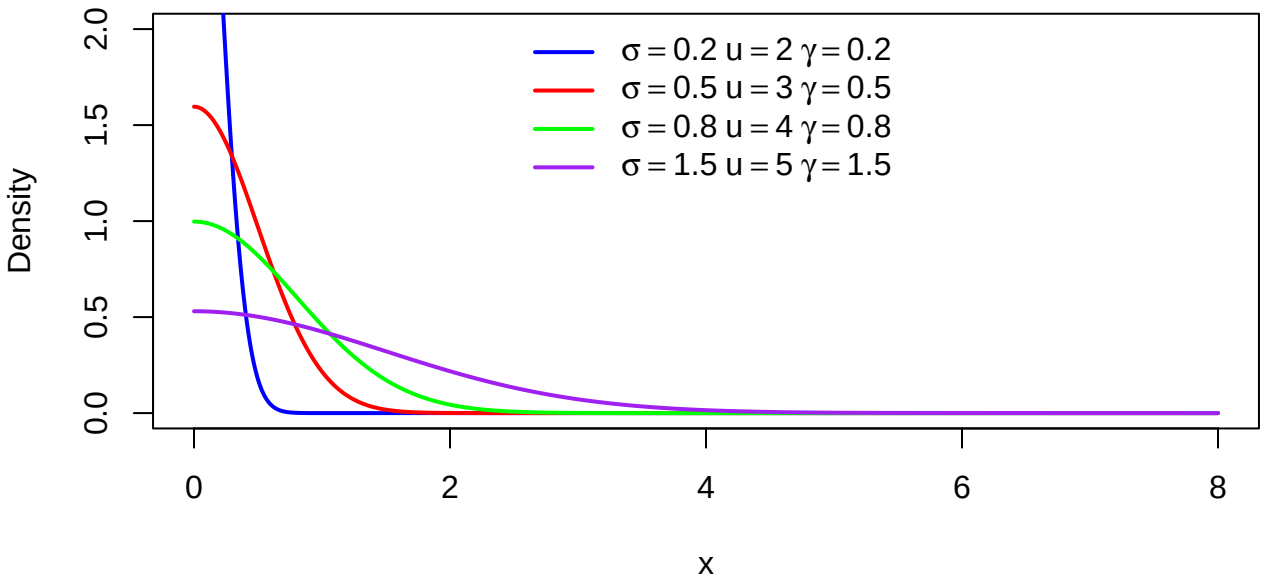

Figure 1: Half-normal-GPD model densities for some selected values of free parameters

Remark 3. *Given that $w_1 \neq w_2$ it follows that the threshold u can be any quantile of the half-normal-GPD model and thus the distribution is free of any constraint which is usually imposed for the case when $w_1 = w_2$ (Carreau and Bengio, 2009; Debbabi and Kratz, 2014). Also, since the half-normal-GPD*

model is positively skewed, the mode of f is equal to zero and less than the median M. This implies that $0 \leq \frac{w_1}{2} \leq F(M;\Theta) = \frac{1}{2}$ *and consequently* $0 \leq w_1 \leq 1$. *Furthermore,* $0 \leq w_2 \leq 1$ *and* $1 - w_1 \leq w_2$. *Lastly, w_2 can be interpreted as the probability of exceeding the threshold u while w_1 is a normalization parameter ensuring that the density f integrates to unity.*

3 Unsupervised iterative estimation algorithm

In this Section, a description of the UIA employed in estimating the vector of free parameter Θ is presented. The model described in Section 2 is taken to belong to the Fréchet maximum domain of attraction (i.e. $\gamma > 0$). For each iteration in the UIA, the UIA breaks down the problem of estimating the vector of free parameters Θ into a double nested sub-problems namely: the parameter $\theta = [\sigma, u]$ and γ. These are estimated successively.

For each iteration of the UIA, the parameter vector θ is first estimated by minimizing the Squared Distance (SD) between the empirical cdf based on some sample and the theoretical one based on the value of γ and then θ is replaced by its estimate obtained in this iteration to estimate γ in the next step of the algorithm using a similar procedure.

Furthermore, θ is estimated again by minimizing the SD between the empirical cdf based on some sample and the theoretical one based on the estimate of γ obtained from the previous iteration and then θ is replaced by its estimate obtained in this iteration to estimate γ in the next step of the algorithm using a similar procedure. Evidently, the algorithm begins with an initial value for γ obtained by minimizing the SD between the empirical cdf based on some sample and the theoretical one based on the initial value chosen for θ and only ends when a stop condition is realized.

Consider a sample $X = (X_i)_{i\in\{1,2,\ldots,n\}}$ from the half-normal-GPD model, and let $\boldsymbol{x} = (x_i)_{i\in\{1,2,\ldots,n\}}$ be a given realization. Suppose $\widetilde{\alpha}^{(0)}$ and $\widetilde{\alpha}^{(k)}$ are the initial value and the estimate of the parameter α at the k^{th} iteration, respectively. To proceed with the UIA for the two-component hybrid model in 1, we start with the initial value $\widetilde{\theta}^{(0)} = \left[\widetilde{\sigma}^{(0)}, \widetilde{u}^{(0)}\right]$, rather than begin by specifying an initial value for γ because the only information we have about γ is that it is positive given that we are dealing with positive data with positive skewness. Morever, $\widetilde{\sigma}^{(0)} = q_{z\%}$ is chosen, where $q_{z\%}$ represents an empirical quantile of order $z\%$ associated to F. Also, $\widetilde{u}^{(0)} = q_{\rho\%}$ is chosen (as we fit a GPD above u). The initial value of $\widetilde{\theta}^{(0)}$ is used to estimate $\widetilde{\gamma}^{(0)}$ by minimizing the SD between the hybrid cdf given $\theta = \widetilde{\theta}^{(0)}$ (fixed), and the empirical cdf F_n associated to the sample $X = (X_i)_{i\in\{1,2,\ldots,n\}}$ of size n, defined, for all $t \in \mathbb{R}$, by

$F_n(t) = \sum_{i=1}^{n} 1_{(X_i \leq t)}/n$. Moreover, the SD is not evaluated only on the realizations x_i (because there might turn out to be just few realizations in the tail), but also on a sequence of generated synthetic data with an increasing property $\boldsymbol{y} = (y_j)_{j\in\{1,2,...,m\}}$, of size m (which may not be the same as n), with a logarithmic step. The synthetic data is added to increase the number of realizations above the tail threshold u. In particular, for any $i \in \{1, 2, ..., n\}$, y_j is defined by:

$$y_j = \min_{i\in\{1,2,...,n\}}(x_i) + (\max_{i\in\{1,2,...,n\}}(x_i) - \min_{i\in\{1,2,...,n\}}(x_i)) \log_{10}\left(1 + \frac{9(j-1)}{m-1}\right). \tag{6}$$

Remark 4. *The introduction of new points between the observations of X only has an impact on F by aiding its evaluation on more points, with no impact on the step function F_n.*

To obtain $\widetilde{\gamma}^{(0)}$ we solve the following minimization problem using the L-M algorithm (Levenberg, 1944; Marquardt, 1963):

$$\widetilde{\gamma}^{(0)} \leftarrow \underset{\gamma>0}{argmin} \left\| F(\boldsymbol{y}; \Theta \mid \widetilde{\theta}^{(0)}) - F_n(\boldsymbol{y}) \right\|_2^2,$$

where $\Theta \mid \widetilde{\theta}^{(0)}$ stands for Θ for $\theta = \widetilde{\theta}^{(0)}$ and $\|.\|_2$ denotes the L_2-norm.

After realizing $\widetilde{\gamma}^{(0)}$, we proceed with the iterations. Now, $\forall k \geq 1$, the k^{th} iteration is divided into two separate minimization problems, which are resolved successively, as described in the following Steps.

Step *1:* Determine $\widetilde{\theta}^{(k)} = \left[\widetilde{\sigma}^{(k)}, \widetilde{u}^{(k)}\right]$ by minimizing the SD between the hybrid cdf given $\widetilde{\gamma}^{(k-1)}$, and the empirical one, as follows:

$$\widetilde{\theta}^{(k)} \leftarrow \underset{(\sigma,u)\in\mathbb{R}_+\times\mathbb{R}_+}{argmin} \left\| F(\boldsymbol{y}; \Theta \mid \widetilde{\gamma}^{(k-1)}) - F_n(\boldsymbol{y}) \right\|_2^2$$

where $\Theta \mid \widetilde{\gamma}^{(k-1)}$ denotes Θ for $\gamma = \widetilde{\gamma}^{(k-1)}$ (fixed). The L-M algorithm is employed to numerically solve this minimization problem.

Step *2:* Determine $\widetilde{\gamma}^{(k)}$ by minimizing the SD between the hybrid cdf given $\widetilde{\theta}^{(k)}$, and the empirical one by solving the following minimization problem using the L-M algorithm:

$$\widetilde{\gamma}^{(k)} \leftarrow \underset{\gamma>0}{argmin} \left\| F(\boldsymbol{y}; \Theta \mid \widetilde{\theta}^{(k)}) - F_n(\boldsymbol{y}) \right\|_2^2,$$

where $\Theta \mid \widetilde{\theta}^{(k)}$, represents Θ for $\theta = \widetilde{\theta}^{(k)}$ (fixed).

Stop condition: The iterations continue till the following stop conditions are satisfied:

$$\left(\underbrace{d\left(F(\boldsymbol{y};\Theta^{(k)}), F_n(\boldsymbol{y})\right) < \varepsilon}_{\text{Condition 1}} \quad \text{and} \quad \underbrace{d\left(F(\boldsymbol{y_{q_\delta}};\Theta^{(k)}), F_n(\boldsymbol{y_{q_\delta}})\right) < \varepsilon}_{\text{Condition 2}} \right) \quad \text{or} \quad \underbrace{k = k_{max}}_{\text{Condition 3}}$$

where ε is a small positive real number, $\boldsymbol{y_{q_\delta}}$ stands for the observations above a fixed quantile q_δ of a given order δ which is associated with the cdf F and $d(x, y)$ denotes the distance between x and y. The distance $d(x, y)$ is chosen in this paper to be the Mean Squared Error (MSE) and it can be further interpreted as the Cramér-von-Mises test of goodness of fit.

To guarantee a good fit of the entire data points and not just the data points lying in the area where the bulk of the distribution lies but also for the tail, the UIA is forced to stop only when the MSE between the hybrid cdf and the empirical one is small enough and this implies the satisfaction of Condition 1 and Condition 2, otherwise, when a fixed maximum number of iterations k_{max} is attained (Condition 3).

Remark 5. *While the hybrid model considered in this paper is assumed to belong to the Fréchet maximum domain of attraction, this algorithm can be extended to the case when the tail index of the GPD is free of any constraints. Also, though the maximum likelihood estimation method appears as a natural estimation method for the parameters of the hybrid model, in practice, this can be very challenging to execute especially when the number of parameters to be estimated is many. Estimating many free parameters at once can be very challenging and the challenge starts when selecting initial values for these parameters as well as obtaining the expression of the gradient function from the likelihood or log-likelihood function. This has informed the choice of the UIA as an alternative method to estimate the parameters of the hybrid model. Nevertheless, when the number of free parameters are relatively small, the maximum likelihood estimation method can be used and is equally robust. Lastly, to understand the convergence properties of the type of iterative algorithm described above see Debbabi et al. (2016).*

Summarily, presented below is the pseudo-code of the algorithm.

Algorithm 1 UIA for the hybrid half-normal-GPD parameters estimation

1: Initialization: Decide start values for $\widetilde{\theta}^{(0)}$ and δ as well as values for ε and k_{max}. Proceed to obtain $\widetilde{\gamma}^{(0)}$ from:

$$\widetilde{\gamma}^{(0)} \leftarrow \underset{\gamma>0}{argmin} \left\| F(\boldsymbol{y}; \Theta \mid \widetilde{\theta}^{(0)}) - F_n(\boldsymbol{y}) \right\|_2^2.$$

2: Iterative process: For

- $k \leftarrow 1$

Step 1 - Estimate $\widetilde{\theta}^{(k)}$ from:

$$\widetilde{\theta}^{(k)} \leftarrow \underset{(\sigma,u)\in\mathbb{R}_+\times\mathbb{R}_+}{argmin} \left\| F(\boldsymbol{y}; \Theta \mid \widetilde{\gamma}^{(k-1)}) - F_n(\boldsymbol{y}) \right\|_2^2.$$

Step 2 - Estimate $\widetilde{\gamma}^{(k)}$ from:

$$\widetilde{\gamma}^{(k)} \leftarrow \underset{\gamma>0}{argmin} \left\| F(\boldsymbol{y}; \Theta \mid \widetilde{\theta}^{(k)}) - F_n(\boldsymbol{y}) \right\|_2^2.$$

- $k \leftarrow k+1$
  until $\Big(d(F(\boldsymbol{y}; \Theta^{(k)}), F_n(\boldsymbol{y})) < \varepsilon \text{ and } d(F(\boldsymbol{y_{q_\delta}}; \Theta^{(k)}), F_n(\boldsymbol{y_{q_\delta}})) < \varepsilon \Big)$ or $\big(k = k_{max}\big)$.

3: Output: Return $\Theta^{(k)} = \big[\widetilde{\sigma}^{(k)}, \widetilde{u}^{(k)}, \widetilde{\gamma}^{(k)}\big]$.

4 Numerical studies

To study the performance of the UIA described in Section 3, we would resort to Monte Carlo simulations. Through simulations we would attempt to test the efficiency of the UIA as applied in the estimating the parameters of the hybrid half-normal-GPD model.

We proceed with the simulations as follows: We consider N (= 100 in this paper) training sets $\{\boldsymbol{x^q} = (x_p^q)_{p\in\{1,2,...,n\}}\}_{q\in\{1,2,...,N\}}$ of size n, and N test sets $\{\boldsymbol{y^q} = (y_p^q)_{p\in\{1,2,...,l\}}\}_{q\in\{1,2,...,N\}}$ of size l, simulated from the hybrid half-normal-GPD model with a fixed parameters vector Θ. Using each training set $\boldsymbol{x^q}$, $q \in \{1, 2, ..., N\}$, Θ is estimated say $\widetilde{\Theta}^q = [\widetilde{\sigma}^q, \widetilde{u}^q, \widetilde{\gamma}^q]$, using the UIA. We denote by $\widetilde{\alpha}^q$ the estimate of the parameter α relative to the q^{th} training set. Furthermore, the empirical mean and variance of $\widetilde{\alpha}^q$ over the N training sets is computed namely $\tilde{\alpha} = \sum_{q=1}^{N} \widetilde{\alpha}^q/N$ and $\tilde{S}_N^{\alpha} = \sum_{q=1}^{N} (\widetilde{\alpha}^q - \tilde{\alpha})^2/(N-1)$, respectively. The significance of $\tilde{\alpha}$ is determined from two criteria: the MSE and test of hypothesis on α. The MSE is expressed for any parameter α as $\text{MSE}_\alpha = \sum_{q=1}^{N} (\widetilde{\alpha}^q - \alpha)^2/N$. A small value of the MSE indicates the efficiency of the UIA in estimating the parameter α. To test $\tilde{\alpha}$ (with unknown variance) we set up the

hypothesis

$$\begin{aligned} H_0 &: \tilde{\alpha} = \alpha \\ H_1 &: \tilde{\alpha} \neq \alpha. \end{aligned}$$

Because N is large, a $z-$test (instead of a t-test) of size κ, with a rejection region of H_0 at risk $\kappa\%$ described by $\left(|T_{\tilde{\alpha}}| > \Phi^{-1}(1-\kappa/2)\right)$ is used, where the statistics $T_{\tilde{\alpha}}$ is given by $T_{\tilde{\alpha}} = (\tilde{\alpha} - \alpha)/\sqrt{\tilde{S}_N^{\alpha}}$, and $\Phi^{-1}(1-\kappa/2)$ denotes the quantile of order $1-\kappa/2$ of the standard normal distribution Φ. Lastly, the hybrid pdf f given Θ is compared with the pdf $\widetilde{f}$ estimated on each test set $\boldsymbol{y^q}$, given $\widetilde{\Theta}^q$. To do so, we compute the average of the log-likelihood ratio $\mathcal{D}$ of $f(\boldsymbol{y^q};\Theta)$ by $\widetilde{f}(\boldsymbol{y^q};\widetilde{\Theta}^q)$, over the N simulations:

$$\mathcal{D} = \frac{1}{Nl}\sum_{q=1}^{N}\sum_{p=1}^{l}\log\left(\frac{f(y_p^q;\Theta)}{\widetilde{f}(y_p^q;\widetilde{\Theta}^q)}\right). \tag{7}$$

A small value of $\mathcal{D}$ indicates an efficient estimation of the parameters of the hybrid half-normal-GPD model using the UIA.

We performed many Monte Carlo simulations by varying Θ and n in order to ascertain the robustness of the UIA for different values of the parameters and sample sizes. We also set $l = n$, $z = 20$, $\kappa = 5\%$, $\delta = 0.3$, $\varepsilon = 10^{-8}$ and $\rho = 0.4$. To conserve space, the results of three of such simulations are reported in Tables 1, 2 and 3, and the other unreported simulations follow a fashion similar to the ones reported here. The efficiency of the UIA, in terms of goodness-of-fit, is shown through the two criteria described above and the average of the log-likelihood ratio $\mathcal{D}$.

We observe from the results in Tables 1, 2 and 3 that as the sample size increases the MSE becomes smaller for all parameters. The variance of σ is also observed to be smaller than the variances of u and γ for all parameters combinations and sample sizes. Observe also that, $\mathcal{D}$ is very small for all parameters combinations and sample sizes. This highlights accuracy and efficiency in the estimation of the parameters using the UIA.

We also made recourse to statistical hypothesis testing as an additional criterion. For the N training sets, we compute the test statistics denoted $T_{\tilde{\alpha},N}$ and the corresponding p-value $p_{T_{\tilde{\alpha},N}} = 2(1-\Phi(|T_{\tilde{\alpha},N}|))$, with respect to the parameter α, that we will compare to κ. Whenever this p-value is higher than κ, we fail to reject H_0. For any $n \in \{10^3, 10^4, 10^5\}$ and for any parameter $\alpha \in \{\sigma, u, \gamma\}$, we obtain $|T_{\tilde{\alpha},N}| < \Phi^{-1}(0.975) = 1.96$, and $p_{T_{\tilde{\alpha},N}} > \kappa = 5\%$, which shows a high acceptance (at the 5% level of significance) of H_0 ($\tilde{\alpha} = \alpha$), that is, a very high level of similarity between the values obtained through

the UIA and the fixed ones.

Lastly, we also observe that as the sample size increases, the average execution time of the UIA increases. It should be noted that the estimation algorithm was implemented using the R programming language. For faster programming languages, the average execution time of the estimation algorithm could significantly reduce.

Table 1: Simulation results for $\Theta = [1, 3, 1.5]$

			$n = 10^3$	$n = 10^4$	$n = 10^5$
Model parameters	$\sigma = 1$	$\tilde{\sigma}$	0.9898	0.9978	0.9999
		$\tilde{S}_N^{\sigma}$	$1.02\ 10^{-3}$	$1.15\ 10^{-4}$	$1.56\ 10^{-5}$
		MSE_{σ}	$1.11\ 10^{-3}$	$1.15\ 10^{-4}$	$1.54\ 10^{-5}$
		$T_{\tilde{\sigma},N}$	-0.3200	-0.2083	-0.0534
		$p_{T_{\tilde{\sigma},N}}$	0.7490	0.8350	0.9574
	$u = 3$	$\tilde{u}$	3.3181	3.1098	3.0008
		$\tilde{S}_N^{u}$	2.0292	$4.49\ 10^{-1}$	$1.52\ 10^{-2}$
		MSE_{u}	2.1101	$4.49\ 10^{-1}$	$1.51\ 10^{-2}$
		$T_{\tilde{u},N}$	0.2233	0.1652	0.0061
		$p_{T_{\tilde{u},N}}$	0.8233	0.8688	0.9951
	$\gamma = 1.5$	$\tilde{\gamma}$	1.7265	1.6312	1.4912
		$\tilde{S}_N^{\gamma}$	$6.89\ 10^{-1}$	$3.12\ 10^{-2}$	$9.37\ 10^{-3}$
		MSE_{γ}	$6.92\ 10^{-1}$	$3.26\ 10^{-2}$	$9.24\ 10^{-3}$
		$T_{\tilde{\gamma},N}$	0.2728	0.1206	-0.0922
		$p_{T_{\tilde{\gamma},N}}$	0.7850	0.8627	0.9265
Execution time in seconds			237.50	1885.94	19884.85
Maximum number of iterations			50	50	50
$\mathcal{D}$			$1.62\ 10^{-3}$	$1.13\ 10^{-4}$	$1.12\ 10^{-6}$

Table 2: Simulation results for $\Theta = [2.5, 5, 1.5]$

			$n = 10^3$	$n = 10^4$	$n = 10^5$
Model parameters	$\sigma = 2.5$	$\tilde{\sigma}$	2.4945	2.4981	2.4997
		$\tilde{S}_N^{\sigma}$	$1.59\ 10^{-2}$	$1.35\ 10^{-3}$	$1.45\ 10^{-4}$
		MSE_{σ}	$1.58\ 10^{-2}$	$1.32\ 10^{-3}$	$1.43\ 10^{-4}$
		$T_{\tilde{\sigma},N}$	-0.0435	-0.0517	-0.0250
		$p_{T_{\tilde{\sigma},N}}$	0.9653	0.9586	0.9801
	$u = 5$	$\tilde{u}$	4.9973	4.9986	5.002
		$\tilde{S}_N^{u}$	$6.45\ 10^{-1}$	$9.71\ 10^{-2}$	$4.49\ 10^{-3}$
		MSE_{u}	$6.39\ 10^{-1}$	$1.00\ 10^{-1}$	$4.41\ 10^{-3}$
		$T_{\tilde{u},N}$	-0.0034	-0.0045	0.0039
		$p_{T_{\tilde{u},N}}$	0.9973	0.9964	0.9976
	$\gamma = 1.5$	$\tilde{\gamma}$	1.5536	1.5217	1.4985
		$\tilde{S}_N^{\gamma}$	$3.60\ 10^{-1}$	$4.23\ 10^{-2}$	$2.24\ 10^{-3}$
		MSE_{γ}	$3.59\ 10^{-1}$	$4.37\ 10^{-2}$	$2.22\ 10^{-3}$
		$T_{\tilde{\gamma},N}$	0.0892	0.1055	-0.0317
		$p_{T_{\tilde{\gamma},N}}$	0.9289	0.9160	0.9747
Execution time in seconds			257.01	1925.24	20184.75
Maximum number of iterations			50	50	50
$\mathcal{D}$			$5.67\ 10^{-5}$	$1.26\ 10^{-5}$	$6.94\ 10^{-8}$

Table 3: Simulation results for $\Theta = [1.5, 3.5, 2]$

			$n = 10^3$	$n = 10^4$	$n = 10^5$
Model parameters	$\sigma = 1.5$	$\tilde{\sigma}$	1.4980	1.4987	1.4991
		$\tilde{S}_N^{\sigma}$	$5.34\ 10^{-3}$	$3.67\ 10^{-4}$	$8.85\ 10^{-5}$
		MSE_{σ}	$5.29\ 10^{-3}$	$3.73\ 10^{-4}$	$9.07\ 10^{-5}$
		$T_{\tilde{\sigma},N}$	-0.0279	-0.0679	-0.0956
		$p_{T_{\tilde{\sigma},N}}$	0.9778	0.9459	0.9238
	$u = 3.5$	$\tilde{u}$	3.5337	3.47898	3.4912
		$\tilde{S}_N^{u}$	$4.65\ 10^{-1}$	$8.60\ 10^{-2}$	$4.38\ 10^{-3}$
		MSE_{u}	$4.61\ 10^{-1}$	$9.01\ 10^{-2}$	$1.42\ 10^{-3}$
		$T_{\tilde{u},N}$	0.0495	-0.0689	-0.1612
		$p_{T_{\tilde{u},N}}$	0.9606	0.9451	0.8943
	$\gamma = 2$	$\tilde{\gamma}$	2.4352	1.9005	1.9601
		$\tilde{S}_N^{\gamma}$	3.1017	$2.19\ 10^{-1}$	$6.13\ 10^{-2}$
		MSE_{γ}	3.2601	$2.27\ 10^{-1}$	$9.42\ 10^{-2}$
		$T_{\tilde{\gamma},N}$	0.2471	-0.2127	-0.1612
		$p_{T_{\tilde{\gamma},N}}$	0.8048	0.8316	0.8719
Execution time in seconds			217.13	1819.07	19163.92
Maximum number of iterations			50	50	50
$\mathcal{D}$			$1.03\ 10^{-3}$	$5.35\ 10^{-5}$	$7.85\ 10^{-6}$

5 Applications

Here the half-normal-GPD hybrid model is used to model two real data sets. The first data set is the Standard & Poor's 500 (S&P 500) index. The S&P 500 index which is reported daily includes Open Prices, High Prices, Low Prices, Close Prices, Adjusted Close Prices and Volume of S&P 500. Our focus is on the indices reported for the period 2nd January, 1987 to 9th May, 2024 with 9411 observations. The goal is to model the absolute log returns of the market rather than the actual returns. The absolute log returns of the market are obtained as the absolute value of the logarithm of the ratio of the current Adjusted Close Price to the previous Adjusted Close Price. The absolute log returns for the afforementioned period has skewness 4.8065 and excess kurtosis 61.4471. The S&P500 index data set is readily available in the Yahoo Finance database.

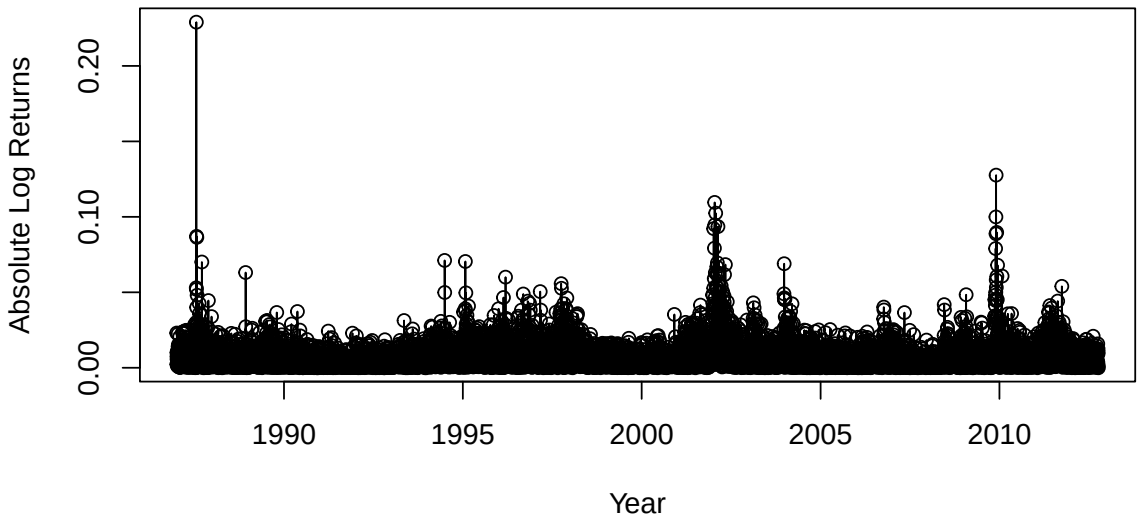

Figure 2: The absolute log returns of the S&P500 from 2nd January, 1987 to 9th May, 2024

Reported in Table 4 are the results obtained from using the half-normal-GPD model in fitting the data set using the UIA. These results include the estimates of the parameters of the distribution and the Kolmogorov-Smirnov (K-S) statistic as well as its corresponding p-value.

Table 4: UIA results for the absolute log returns of S&P500 index

Estimation method	Parameter estimate	$K-S$	p-value
UIA	$\hat{\sigma}=0.0040 \quad \hat{u}=0.0020=q_{25.1\%}$ $\hat{\gamma}=0.0098 \quad \hat{\beta}=0.0073$ $\hat{w}_1=0.5982 \quad \hat{w}_2=0.7492$	0.0067	0.7901

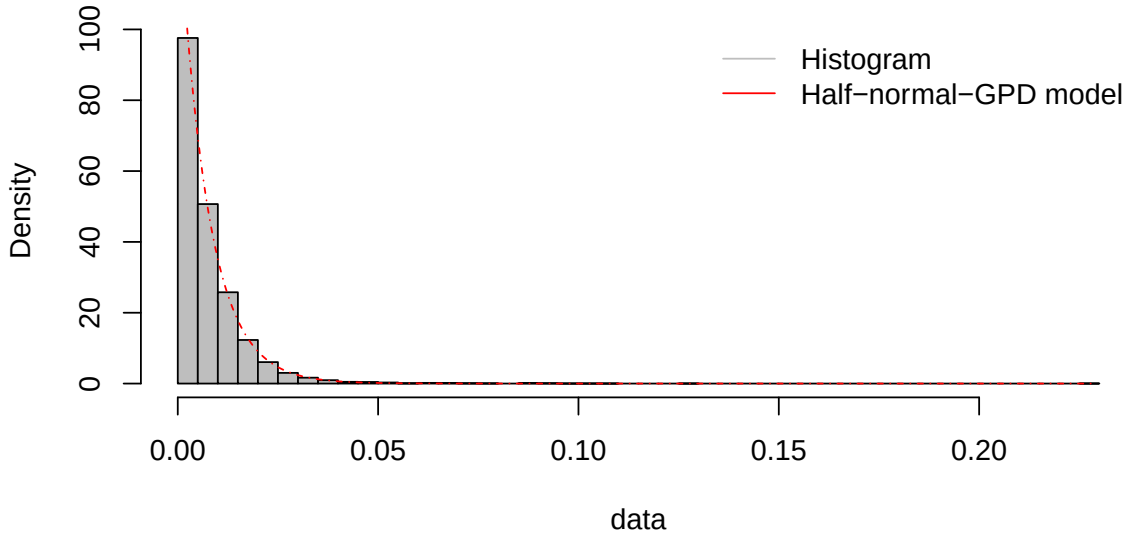

Figure 3: Histogram and fitted density of the absolute log returns of S&P500 index

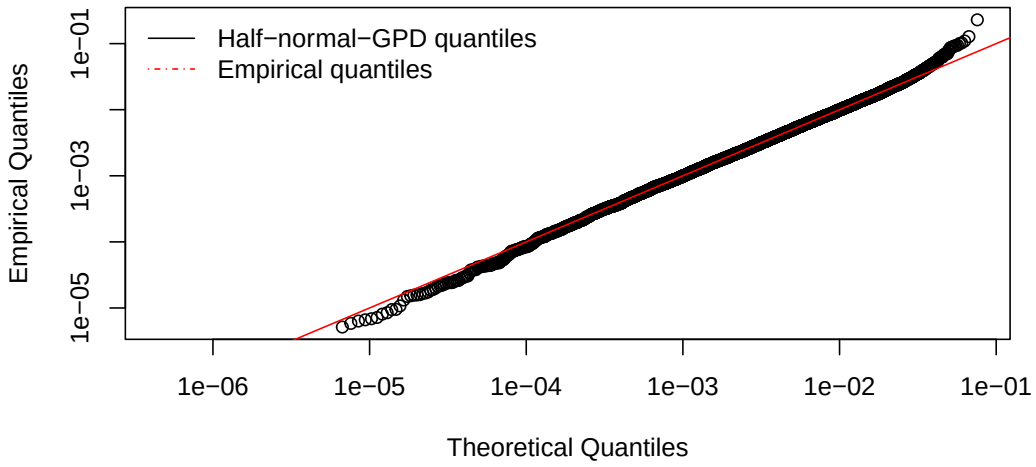

Figure 4: Q-Q plot of the absolute log returns of S&P500 index

The second data set is the rainfall intensity in millimeters per hour which triggered some 785 debris flow events in the South Tyrol region of Italy between 1987 and 2022. The data set was provided by the Agency for Civil Protection of the Autonomous Province of Bozen-Bolzano, Italy. The skewness and excess kurtosis of the absolute log returns are respectively 1.8466 and 3.2528 which clearly shows that the data set is also highly skewed to the right with a heavy tail.

Reported in Table 5 are the results obtained from using the half-normal-GPD model in fitting the data set using the UIA. These results include the estimates of the parameters of the distribution and the Kolmogorov-Smirnov (K-S) statistic as well as its corresponding p-value.

Table 5: UIA results for the rainfall intensity data

Estimation method	Parameter estimate	$K-S$	p-value
UIA	$\hat{\sigma}=3.2855 \quad \hat{u}=4.4253=q_{68.8\%}$ $\hat{\gamma}=0.6329 \quad \hat{\beta}=3.9830$ $\hat{w}_1=0.8248 \quad \hat{w}_2=0.3221$	0.0388	0.1877

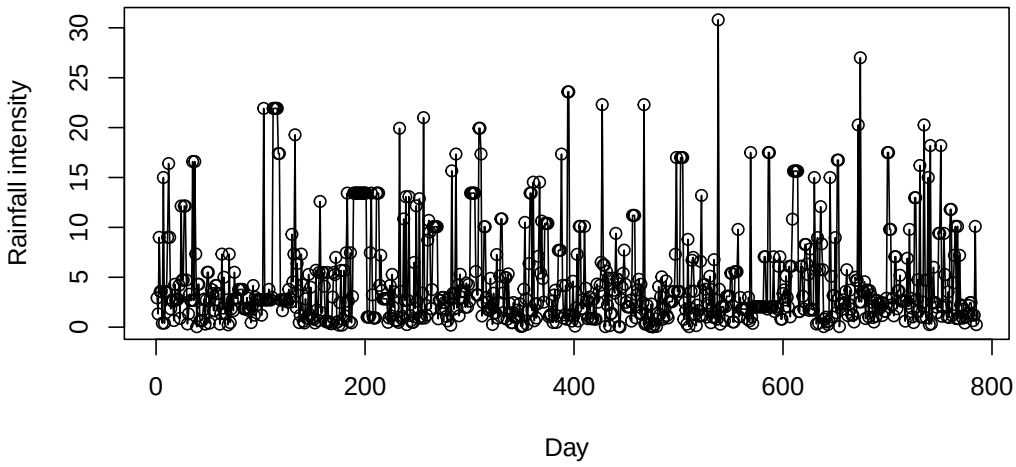

Figure 5: Rainfall intensity in millimeters per hour which triggered some 785 debris flow events in South Tyrol, Italy

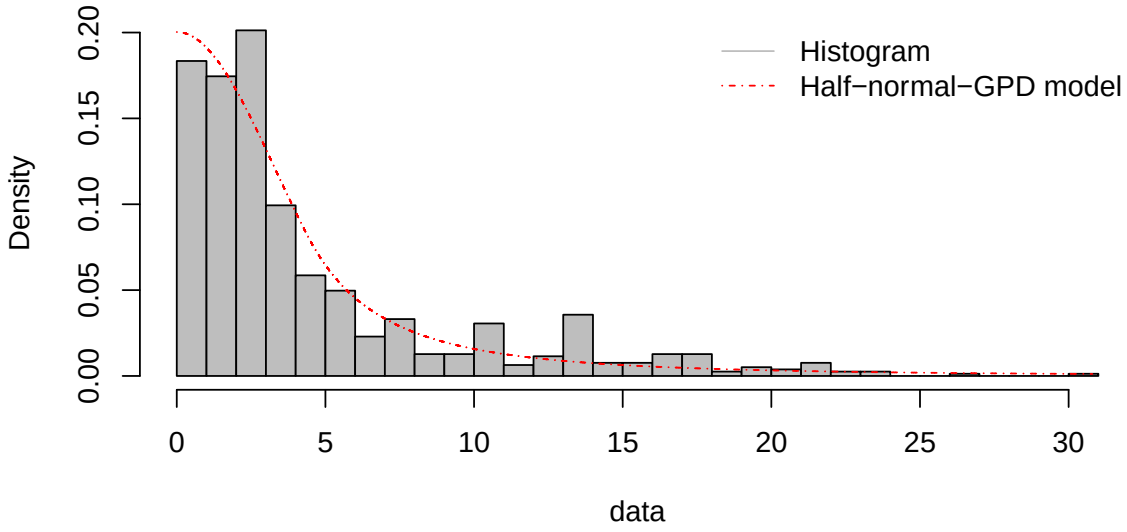

Figure 6: Histogram and fitted density of the rainfall intensity data

The results obtained from using the hybrid half-normal-GPD model to model the two data sets show that the model provided good fits to the data sets.

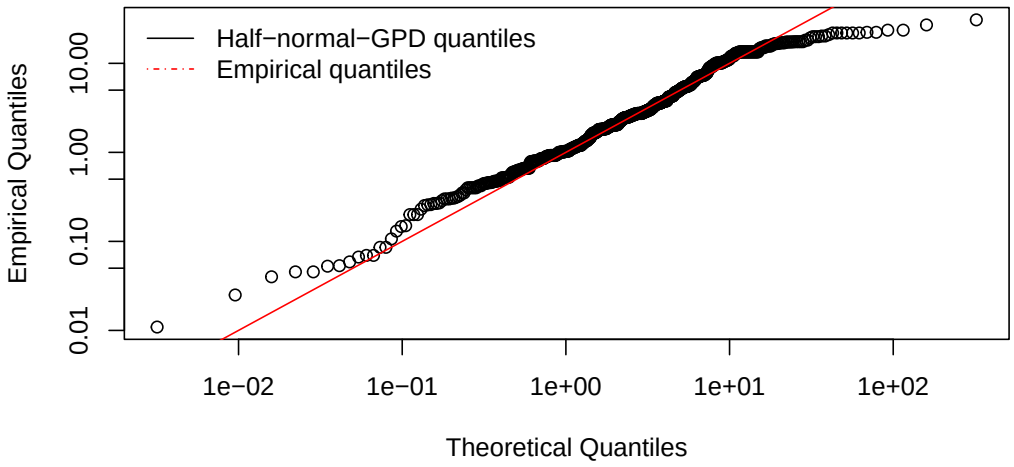

Figure 7: Q-Q plot of the rainfall intensity data

6 Conclusion

A new two-component hybrid model, suitable for modeling data with high right-skewness and estimated by an unsupervised iterative estimation algorithm has been introduced in this paper. We have demonstrated the hybrid model's flexibility and robustness in capturing the unique characteristics of such data through application to real data sets. Moreover, through empirical analyses on synthetic data sets, the unsupervised iterative estimation algorithm for the estimation of the parameters of the hybrid model has shown high accuracy and efficiency, making it a valuable tool for practical applications. This new hybrid model and this estimation technique hold significant promise for a wide range of fields where right-skewed data sets are prevalent, such as finance, environmental studies, signal processing, biomedicine etc. Future research focus could extend this work by exploring multi-component extensions and applying the model and the estimation algorithm to more diverse data sets. Additionally, integrating this hybrid model into more complex statistical frameworks, such as regression models and machine learning algorithms, could further enhance its utility and scope of application. In general, our contribution in this paper provides a substantial advancement in the modeling of data exhibiting high skewness to the right, offering a powerful and versatile tool for statisticians, data scientists and users of statistics.

Data Availability Statement The datasets used in this article can be readily made available upon request from the author.

Declarations

Conflicting Interests The author declares no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Competing Interest The author declares that there are no competing financial or personal interests with any individual or organization that could appear to influence the work reported in this article.

References

Bakar, S. A. A., N. A. Hamzah, and S. Nadarajah (2015). Modeling loss data using composite models. *Insurance: Mathematics and Economics 61*, 146–154.

Blum, P. and M. Dacorogna (2003). Extreme forex moves. *Risk-London-Risk Magazine Limited 16*(2), 63–66.

Broadwater, J. B. and R. Chellapa (2010). Adaptive threshold estimation via extreme value theory. *IEEE Transactions on Signal Processing 58*(2), 490–500.

Carreau, J. and Y. Bengio (2009). hybrid pareto model for asymmetric fat-tailed data: the univariate case. *Extremes 1*, 53–76.

Cooray, K. (2009). The weibull–pareto composite family with applications to the analysis of unimodal failure rate data. *Communications in Statistics: Theory and Methods 38*, 1901–1915.

Cooray, K. and M. M. A. Ananda (2005). Modeling actuarial data with a composite lognormal-pareto model. *Scandinavian Actuarial Journal 2005*(5), 321–334.

Cooray, K., S. Gunasekera, and M. Ananda (2010). Weibull and inverse weibull composite distribution for modeling reliability data. *Model Assisted Statistics and Applications 5*(2), 109–115.

Dacorogna, M. and M. Kratz (2015). Living in a stochastic world and managing complex risks. Available at SSRN 2668468.

Davison, A. C. and R. L. Smith (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society. Series B (Methodological) 52*(3), 393–442.

Debbabi, N. and M. Kratz (2014). A new unsupervised threshold determination for hybrid models. In *in Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, May 2014*, pp. 3440–3444. IEEE.

Debbabi, N., M. Kratz, and M. Mboup (2016). A self-calibrating method for heavy tailed modeling: Application in neuroscience and finance. ESSEC working paper 1619.

Digham, F., M. S. Alouini, and M. K. Simon (2007). On the energy detection of unknown signals over fading channels. *IEEE Transactions on Communications 55*(1), 21–24.

Embrechts, P., C. Kluppelberg, and T. Mikosch (1997). *Modelling Extremal Events for Insurance and Finance*. Springer-Verlag.

Furrer, E. and R. Katz (2008). Improving the simulation of extreme precipitation events by stochastic weather generators. *Water Resources Research 44*(12).

Guillen, M., T. Buch-larsen, J. P. Nielsen, and C. Bolance (2005). Kernel density estimation for heavy-tailed distributions using the champernowne transformation. *Statistics 39*(6), 503–516.

Kollu, R., S. Rayapudi, S. Narasimham, and K. Pakkurthi (2012). Mixture probability distribution functions to model wind speed distributions. *International Journal of Energy and Environmental Engineering 3*(1), 1–27.

Levenberg, K. (1944). A method for the solution of certain nonlonear problems in least squares quart. *Applied Math 2*, 164–168.

Mandava, A., L. Shahram, and E. Regentova (2011). Reliability assessment of microarray data using fuzzy classification methods: A comparative study. *in Advances in Computing and Communications. Communications in Computer and Information Science, Springer, Berlin Heidelberg 190*, 351–360.

Marquardt, D. (1963). An algorithm for least-squares estimation of nonlinear parameters. *SIAM Journal on Applied Mathematics 11*, 431–441.

Nadarajah, S. and S. Bakar (2014). New composite models for the danish fire insurance data. *Scandinavian Actuarial Journal 2014*(2), 180–187.

Osatohanmwen, P., F. Oyegue, S. Ogbonmwan, and W. Muhwava (2024). A general framework for generating three-components heavy-tailed distributions with application. *Journal of Statistical Theory and Applications 23*, 290–314.

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

Rangaswamy, M., J. Michels, and B. Himed (2004). Statistical analysis of the non-homogeneity detector for fastapi applications. *Digital Signal Processing 14* (3), 253–267.

Rossi, F., M. Fiorentino, and P. Versace (1984). Two component extreme value distribution for flood frequency analysis. *Water Resources Research 20* (7), 847–856.

Scollnik, D. P. (2007). On composite lognormal-pareto models. *Scandinavian Actuarial Journal 2007* (1), 20–33.

Scollnik, D. P. and C. Sun (2012). Modeling with weibull-pareto models. *North American Actuarial Journal 16* (2), 260–272.

Sermpezis, P. and T. Spyropoulos (2015). Modelling and analysis of communication traffic heterogeneity in opportunistic networks. *IEEE Transactions on Mobile Computing 14* (11), 2316–2331.

Singh, V., C. Li, and A. Mishra (2012). Simulation of the entire range of daily precipitation using a hybrid probability distribution. *Water Resources Research 48*, 1–17.