Earthline Journal of Mathematical Sciences
E-ISSN: 2581-8147
Volume 15, Number 3, 2025, Pages 401-417
https://doi.org/10.34198/ejms.15325.401417

A New Hybrid Family of Probability Distributions for Fitting Heavy-tailed Data with Application to Finance

Victor Mazona[1,*], F. B. Adebola[2], A. A. Akomolafe[3] and O. O. Ojo[4]

[1] Department of Statistics, Federal University of Technology, Akure, Nigeria
e-mail: vmmazona@gmail.com

[2] Department of Statistics, Federal University of Technology, Akure, Nigeria

[3] Department of Statistics, Federal University of Technology, Akure, Nigeria

[4] Department of Statistics, Federal University of Technology, Akure, Nigeria

Abstract

The classical continuous univariate probability distributions, which contain one or two parameters, have been observed to break down when complexities exist in the structure of a data set such as when outliers are present, alongside observations centered around the mean. When a data set exhibits heterogeneity or exists in a multi-component form and it becomes impossible to use a single probability distribution to capture the distinct components of the data set, using a composite distribution to model the data set becomes plausible. This situation has led to the formulation of various hybrid or composite models where each component of the hybrid model handles the specific part of the data set that it is well suited for. Furthermore, the approach or method used in the formulation of these hybrid models plays a vital role in determining how meaningful the results obtained from them are. Several approaches or methods for formulating hybrid distributions have appeared in the literature, each with their own pros and cons. We present in this paper a general two-component hybrid model for fitting heterogeneous heavy-tailed data sets with tails to the right. The functional form of the two-component hybrid family is specified by the probability density function (pdf), cumulative distribution function (cdf) and the quantile function. Three members of the family using three different distributions for the right tail are presented. A formal method based on maximum likelihood for the estimation of the parameters of the models belonging to the family is also presented. A Monte Carlo simulation study is carried out to determine the efficiency of the estimation method. An application to a real data set in finance is performed.

1. Introduction

In many situations, particularly in hydrology, finance and insurance, the use of a single probability distributions in fitting observed data can be very inadequate owing to the fact that there exists some reasonable number of outliers in the data and hence the structure of the data can be termed heterogeneous, having both the main part which contains the bulk of the data around the mean, and the tail part which contains observations that can be classified as extremes or outliers. Using a single probability distribution such as the Gaussian, lognormal, Weibull, gamma distributions etc. in fitting such data, can result in either overstating or understating of risk probabilities. Without a loss in generality, the observations which are classified as outliers are assumed to be on the right-tail of the distribution and thus the data can be treated as right heavy-tailed. To adequately

Received: December 30, 2024; Revised: February 27, 2025; Accepted: March 4, 2025; Published: March 7, 2025

2020 Mathematics Subject Classification: 60G70, 62E20, 62F35, 62P05, 62P10, 65D15, 68W40.

Keywords and phrases: heavy-tailed distribution, composite models, maximum likelihood estimation, S&P 500 index.

*Corresponding author

model the data, it is required that a piecewise two-component model be used, where each component of the model is used to handle the specific property of the data for which it is suited. In standard extreme value theory, the tail area usually follows the Generalized Pareto Distribution (GPD) after a certain threshold has been exceeded, and it thus leaves the investigator to find an appropriate distribution for the main part of the data and use some sound mathematical technique to join the two distributions together in such a way that standard conditions for the existence and uniqueness of a probability distribution is satisfied.

Cooray and Ananda (2005) offered a procedure for doing this and since then, several modifications and extensions of that work have appeared in the literature (Scollnik 2007; Carreau and Bengio 2009; Mandava et al. 2011; Li et al. 2012; Kollu et al. 2012; Scollnik and Sun 2012; Debbabi 2015; Baker et al. 2015; Debbabi 2016). The normal, Weibull, lognormal and gamma distributions have been used in several studies in the literature as the distribution of the main part of the data and while useful results have been obtained so far, it has also been observed that these classical distributions can prove inadequate in certain cases (Preda and Ciumara 2006; Teodorescu and Vernice, 2006, 2009; Cooray et al. 2009; Cooray 2010; Teodorescu and Vernice 2013; Benatmane et al. 2020). For the tail area, where the GPD has been extensively used, it remains an open area of research to try out other heavy-tailed distributions in place of the GPD. Some of these heavy-tailed distributions include the Pareto, Burr XII and the Lomax distributions. In this paper, we present a general family of two-components hybrid models which combines distributions for the bulk of the data around the mean and those for the tail where outliers are present.

The rest of the paper is organized as follows. In Section 2, we take a look at the new hybrid family. Section 3 contains the estimation algorithm for the parameters of the new hybrid family. In Section 4, we look at some specific members of the new hybrid family, while in Section 5 a simulation study is conducted to test the performance of the estimation algorithm used in estimating the parameters of the new family. An application to a financial data set is conducted in Section 6. The paper closes in Section 7 with conclusion.

2. The New Hybrid Family

Suppose we have a data set which can be conveniently separated into two components. We assume without a loss in generality that the first component is the main part of the data, while the second part is the part that contains the extreme observations or outliers. Our goal is to use a two-component piecewise pdf to model the data where each component of the data will be modelled by a pdf which is best suited for it, and in general combine the two pdfs together. Let f_1 and f_2 be two pdfs with respective parameter vectors θ_1 and θ_2 where f_1 is the density of the first component of the data and f_2 is the density of the second component. Suppose F_1 and F_2 are cdfs corresponding to f_1 and f_2 respectively, with respective quantile functions $Q_1(p;\theta_1)$ and $Q_2(p;\theta_2)$, $0 < p < 1$. We define the general pdf of the hybrid family of distributions as

$$f(x;\theta) = \begin{cases} u_1 f_1(x;\theta_1), & \text{if } -\infty < x \leq r, \\ u_2 f_2(x;\theta_2), & \text{if } r \leq x < \infty, \end{cases} \tag{1}$$

where θ is a vector which contains all the free parameters, u_1 and u_2 are weights associated with a specific component of the density, and r is a threshold indicating the point of movement from the main innovation component to the tail area or the part that contains the extreme observations. In previous studies, the threshold r is estimated using graphical methods but in this model, we have specified it as a parameter, and it would be estimated algorithmically.

In the model in (1), we assume that the movement from one component of the model is smooth and thus we make the following assumptions for the model:

(a) First, we assume that the pdf in (1) is non-negative and satisfies

$$\int_R f(x;\theta)dx = 1.$$

This implies that

$$u_1F_1(r;\theta_1) + u_2[1 - F_2(r;\theta_2)] = 1. \tag{2}$$

(b) The data has a heavy-tail and the tail is to the right.

(c) The pdf in (1) is smooth and is continuous and differentiable at the threshold r. This implies that

$$\begin{cases} u_1f_1(r;\theta_1) = u_2f_2(r;\theta_2) \\ u_1f_1(r;\theta_1) = u_2f_2(r;\theta_2). \end{cases} \tag{3}$$

Using the result in (3) we have

$$u_1 = \frac{u_2f_2(r;\theta_2)}{f_1(r;\theta_1)}. \tag{4}$$

Substituting (4) into (2) gives

$$u_2 = \left\{\frac{f_2(r;\theta_2)}{f_1(r;\theta_1)}F_1(r;\theta_1) + 1 - F_2(r;\theta_2)\right\}^{-1}. \tag{5}$$

The cdf corresponding to our proposed general family of hybrid probability density function in (1) is given by

$$F(x;\theta) = \begin{cases} u_1F_1(x;\theta_1), & \text{if } -\infty < x \le r, \\ 1 - u_2[1 - F_2(x;\theta_2)], & \text{if } r \le x < \infty. \end{cases} \tag{6}$$

The quantile function corresponding to the cdf in (6) is given by

$$Q(p;\theta) = \begin{cases} Q_1\left(\frac{p}{u_1};\theta_1\right), & \text{if } p \le u_1, \\ Q_2\left(\frac{p - (1 - u_2)}{u_2};\theta_2\right), & \text{if } p \ge 1 - u_2. \end{cases} \tag{7}$$

Random samples X can be simulated from the general family in (1) by replacing p in (7) with U where U is a uniform random variable on (0,1). That is,

$$X = \begin{cases} Q_1\left(\frac{U}{u_1};\theta_1\right), & \text{if } U \le u_1, \\ Q_2\left(\frac{U - (1 - u_2)}{u_2};\theta_2\right), & \text{if } U \ge 1 - u_2. \end{cases} \tag{8}$$

3. Estimation Procedure

Here we present an estimation routine for the vector of free parameters θ in the hybrid density in (1). First, we assumed that the density in (1) depends on the vector of free parameters $\theta = (\delta_1,...,\delta_p,r)^T, p \in \mathbb{N}^+$. Consider a complete independent random sample $x_1,x_2,...,x_n$ of size n from the distribution in (1). Without a loss in generality, assume this sample is ordered such that $x_1 \le x_2 \le ... \le x_n$. To estimate the parameter vector θ using the maximum likelihood approach, one would require some information about a positive integer m such that r lies between the m-th and $(m + 1)$-th observation, so that $x_m \le r \le x_{m+1}$. Assume that m is known. Define the likelihood function of any complete independent random sample $x_1,x_2,...,x_n$ from a distribution g with

parameter vector ϑ as

$$\mathcal{L}(x_1,x_2,...,x_n;\vartheta) = \prod_{i=1}^{n} g(x_i;\vartheta)\,.$$

Consequently, the likelihood function based on the density in (1) is expressed as

$$\begin{aligned}\mathcal{L}\big(x_1,x_2,...,x_n;\delta_1,...,\delta_p,r\big) &= \prod_{j=1}^{n} f(x_i;\theta) = \prod_{k_1=1}^{m} u_1 f_1(x_{k_1};\theta_1) \prod_{k_2=m+1}^{n} u_2 f_2(x_{k_2};\theta_2)\\ &= u_1^m u_2^{n-m} \prod_{k_1=1}^{m} f_1(x_{k_1};\theta_1) \prod_{k_2=m+1}^{n} f_2(x_{k_2};\theta_2)\\ &= u_1^m u_2^{n-m}\mathcal{L}(x_1,x_2,...,x_{m_1};\theta_1)\mathcal{L}(x_{m+1},x_{m+2},...,x_n;\theta_2).\end{aligned}$$

The log-likelihood function corresponding to the likelihood is expressed as

$$\begin{aligned}L = \log\mathcal{L}\big(x_1,x_2,...,x_n;\delta_1,\delta_2,...,\delta_p,r\big) &= m\log u_1 + (n-m)\log u_2\\ &+ \sum_{k_1=1}^{m} \log\big(f_1(x_{k_1};\theta_1)\big) + \sum_{k_2=m+1}^{n} \log\big(f_2(x_{k_2};\theta_2)\big)\,.\end{aligned} \tag{9}$$

In practice, the exact values of m is usually unknown. Observe also that if m changes, the maximum likelihood estimator of θ also changes. A grid search for the optimal value of m can be performed. For an optimal value of m, we obtain $\hat{\delta}_1,...,\hat{\delta}_p, \hat{r}$ as solutions of the systems

$$\begin{cases}\dfrac{\partial L}{\partial \delta_i} = 0, & i = 1,2,...,p\\ \dfrac{\partial L}{\partial r} = 0.\end{cases} \tag{10}$$

If $x_m \le \hat{r} \le x_{m+1}$, then the maximum likelihood estimators of $\delta_1,\delta_2,...,\delta_p$,and r are

$$\hat{\delta}_i^{ML} = \hat{\delta}_i,\ i = 1,2,...,p \quad \hat{r}^{ML} = \hat{r}.$$

4. Specific Members of the General Family

Here, we consider three hybrid distributions which are members of the general family in (1). We let the distribution for the main innovation or main part of the data be the skew-normal distribution. That is, f_1, F_1 and Q_1 are the pdf, cdf and quantile function of the skew-normal distribution respectively, with

$$f_1(x;\alpha,\tau,\varphi) = \frac{2}{\varphi}\phi\Big(\frac{x-\tau}{\varphi}\Big)\Phi\Big(\alpha\Big(\frac{x-\tau}{\varphi}\Big)\Big), \qquad -\infty < x,\alpha < \infty,\ x \ge \tau,\alpha,\tau,\varphi > 0,$$

where $\phi(.) = \Phi_{,}(.)$ and $\Phi(.)$ is the cdf of the normal distribution,

$$F_1(x;\alpha,\tau,\varphi) = \Phi\Big(\frac{x-\tau}{\varphi}\Big) - 2T\Big(\frac{x-\tau}{\varphi},\alpha\Big), \qquad -\infty < x < \infty,\ x \ge \tau,\alpha \ge 0,\tau,\varphi > 0,$$

where $T(h,a)$ is the Owen's T function defined by

$$T(h,a) = \frac{1}{2\pi}\int_0^a \frac{e^{-\frac{1}{2}h^2(1+x^2)}}{1+x^2}\,dx, \qquad -\infty < a,h < \infty,$$

$$Q_1(p;\alpha,\tau,\varphi) = F_1^{-1}(x;\alpha,\tau,\varphi), \qquad 0 < p < 1.$$

In the skew normal distribution, the parameters α,τ and φ are shape, location, and scale parameters respectively. We have decided to make f_1 the skew-normal distribution because, in several literatures, the normal distribution is used and since the skew-normal distribution is more flexible, we consider our choice a good one. For the tail component, we use three heavy-tailed distributions. That is, we shall choose f_2 to be three heavy-tailed distributions to combine with the skew-normal distribution to form the two-component hybrid models. These tail distributions include: the GPD, the Pareto distribution (PD) and the Lomax distribution (LD). Thus, we shall formulate the following hybrid models: the skew-normal GPD (SN-GPD) distribution, the skew-normal Pareto (SN-PD) distribution and the skew-normal Lomax (SN-LD) distribution.

4.1. The skew-normal GPD (SN-GPD) distribution

The GPD has pdf, cdf and quantile function expressed respectively as

$$f_2(x-r;\gamma,c) = \frac{1}{c}\left(1+\gamma\frac{x-r}{c}\right)^{-1-\frac{1}{\gamma}},$$

$$F_2(x-r;\gamma,c) = 1-\left(1+\gamma\frac{x-r}{c}\right)^{-\frac{1}{\gamma}},$$

$$Q_2(p;r,\gamma,c) = \frac{c}{\gamma}[(1-p)^{-\gamma}-1]+r,$$

$$\forall x \geq r \in Z(\gamma,c),\ -\infty < \gamma < \infty, c > 0,$$

$$Z(\gamma,c) = \begin{cases}[0,\infty) & \text{if } \gamma \geq 0\\ [0,-c/\gamma] & \text{if } \gamma < 0,\end{cases} \qquad 0 < p < 1.$$

The parameter c is a scale parameter while the parameter γ is the tail index parameter. Using the result in (1), the pdf of the SN-GPD is given by

$$f(x;\theta) = \begin{cases} u_1 f_1(x;\alpha,\tau,\varphi), & \text{if } -\infty < x \leq r,\\ u_2\frac{1}{c}\left(1+\gamma\frac{x-r}{c}\right)^{-1-\frac{1}{\gamma}}, & \text{if } r \leq x < \infty.\end{cases} \tag{11}$$

Using the assumptions (a-c) in Section 2, we obtain the following equations relating to the weights u_1 and u_2 and the parameter c of the model in (11):

$$u_1 = \frac{u_2}{cf_1(r;\alpha,\tau,\varphi)},$$

$$u_2 = \left\{1+\frac{F_1(r;\alpha,\tau,\varphi)}{cf_1(r;\alpha,\tau,\varphi)}\right\}^{-1},$$

$$c = -(1+\gamma)\frac{f_1(r;\alpha,\tau,\varphi)}{f_1(r;\alpha,\tau,\varphi)}.$$

The parameter vector θ in (11) contains only the free parameters and thus $\theta = [\alpha,\tau,\varphi,r,\gamma]$. After θ has been

estimated, the values of the other constrained parameters which are u_1,u_2 and c can be obtained from the above relations which describe them. Observe that in the model, there are eight parameters which we ought to estimate their values, however, with the imposition of the assumptions or the constraints in Section 2, we were able to reduce the number of parameters to be estimated to just five. This presents one of the benefits of our methods.

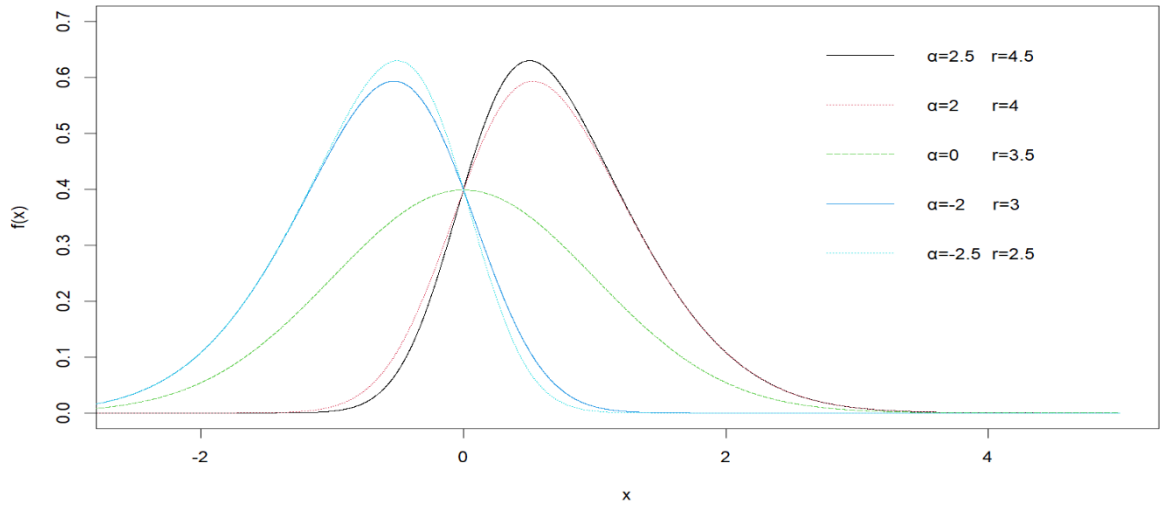

Figure 1: The SN-GPD density $[\tau = 0,\varphi = 1,\gamma = 5]$.

The cdf and quantile function of the SN-GPD are given respectively by

$$F(x;\theta) = \begin{cases} u_1 F_1(x;\alpha,\tau,\varphi), & \text{if } -\infty < x \le r, \\ 1 - u_2\left(1 + \gamma\dfrac{x-r}{c}\right)^{-\frac{1}{\gamma}}, & \text{if } r \le x < \infty, \end{cases} \tag{12}$$

$$Q(p;\theta) = \begin{cases} Q_1\left(\dfrac{p}{u_1};\alpha,\tau,\varphi\right), & \text{if } p \le u_1, \\ \dfrac{c}{\gamma}\left[\left(\dfrac{1-p}{u_2}\right)^{-\gamma} - 1\right] + r, & \text{if } p \ge 1 - u_2. \end{cases} \tag{13}$$

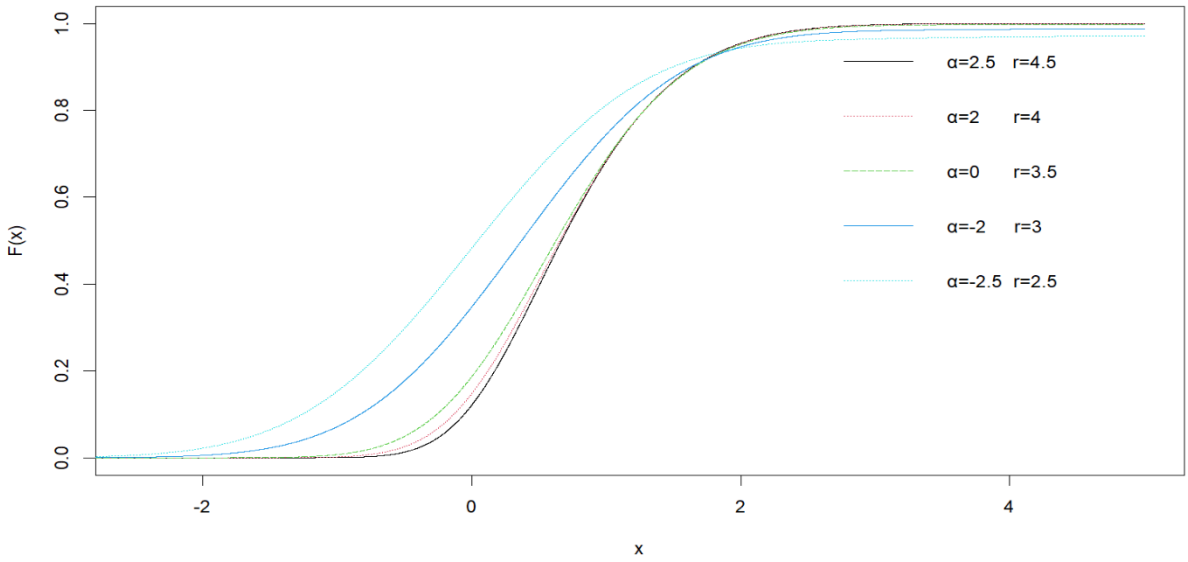

Figure 2: The SN-GPD cdf$[\tau = 0,\varphi = 1,\gamma = 5]$.

Random samples X can be simulated from the SN-GPD by replacing p in (13) with U where U is a uniform random variable on (0,1). That is,

$$X = \begin{cases} Q_1\left(\frac{U}{u_1};\alpha,\tau,\varphi\right), & \text{if } U \le u_1, \\ \frac{c}{\gamma}\left[\left(\frac{1-U}{u_2}\right)^{-\gamma} - 1\right] + r, & \text{if } \quad U \ge 1 - u_2. \end{cases} \tag{14}$$

4.2. The skew-normal Pareto (SN-PD) distribution

The PD has pdf, cdf and quantile function expressed respectively as

$$f_2(x;b,r) = \frac{br^b}{x^{b+1}},$$

$$F_2(x;b,r) = 1 - \left(\frac{r}{x}\right)^b,$$

$$Q_2(p;b,r) = r(1-p)^{-1/b},$$

$$x \ge r, r > 0, b > 0, 0 < p < 1.$$

The parameter b is a shape parameter while the parameter r is a location parameter. Using the result in (1), the pdf of the SN-PD is given by

$$f(x;\theta) = \begin{cases} u_1 f_1(x;\alpha,\tau,\varphi), & \text{if } -\infty < x \le r, \\ u_2 \frac{br^b}{x^{b+1}}, & \text{if } \quad r \le x < \infty. \end{cases} \tag{15}$$

Using the assumptions (a-c) in Section 2, we obtain the following equations relating to the weights u_1 and u_2 and the parameter b of the model in (15):

$$u_1 = \frac{u_2 b}{r f_1(r;\alpha,\tau,\varphi)},$$

$$u_2 = \left\{1 + \frac{b F_1(r;\alpha,\tau,\varphi)}{r f_1(r;\alpha,\tau,\varphi)}\right\}^{-1},$$

$$b = -\frac{r f_1(r;\alpha,\tau,\varphi)}{f_1(r;\alpha,\tau,\varphi)} - 1.$$

The parameter vector θ in (15) contains only the free parameters and thus $\theta = [\alpha,\tau,\varphi,r]$. After θ has been estimated, the values of the other constrained parameters which are u_1,u_2 and b can be obtained from the above relations which describe them. Observe that in the model, there are seven parameters which we ought to estimate their values, however, with the imposition of the assumptions or constraints in Section 2, we were able to reduce the number of parameters to be estimated to just four.

The cdf and quantile function of the SN-PD are given respectively by

$$F(x;\theta) = \begin{cases} u_1 F_1(x;\alpha,\tau,\varphi), & \text{if } -\infty < x \le r, \\ 1 - u_2\left(\frac{r}{x}\right)^b, & \text{if } \quad r \le x < \infty, \end{cases} \tag{16}$$

$$Q(p;\theta) = \begin{cases} Q_1\left(\frac{p}{u_1};\alpha,\tau,\varphi\right), & \text{if } p \le u_1, \\ r\left(\frac{1-p}{u_2}\right)^{-1/b}, & \text{if } p \ge 1-u_2. \end{cases} \quad (17)$$

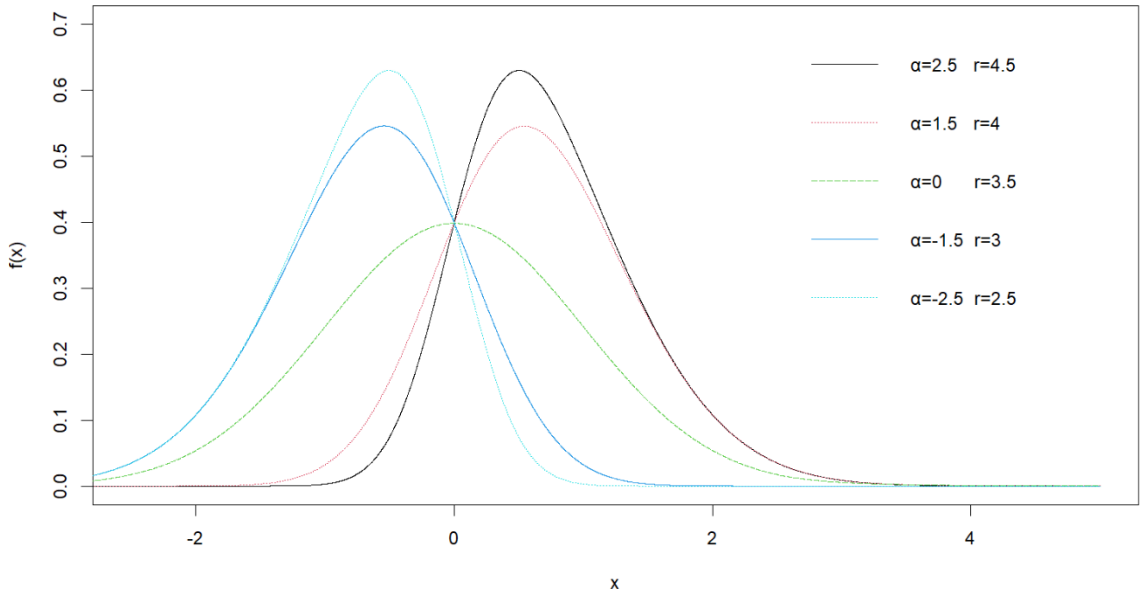

Figure 3: SN-PD density $[\tau = 0, \varphi = 1]$.

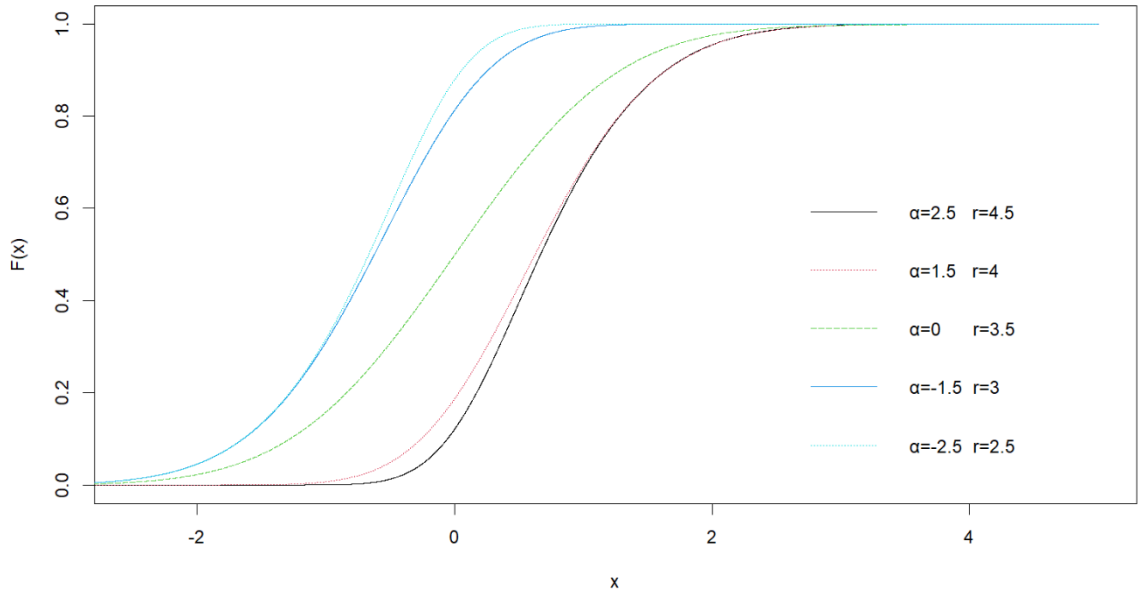

Figure 4: SN-PD cdf$[\tau = 0, \varphi = 1]$.

Random samples X can be simulated from the SN-GPD by replacing p in (17) with U where U is a uniform random variable on (0,1). That is,

$$X = \begin{cases} Q_1\left(\frac{U}{u_1};\alpha,\tau,\varphi\right), & \text{if } U \le u_1, \\ r\left(\frac{1-U}{u_2}\right)^{-1/b}, & \text{if } U \ge 1-u_2. \end{cases} \quad (18)$$

4.3. The skew-normal Lomax (SN-LD) distribution

The LD has pdf, cdf and quantile function expressed respectively as

$$f_2(x-r;c,k) = \frac{k}{c}\left(1+\frac{x-r}{c}\right)^{-k-1},$$

$$F_2(x-r;c,k) = 1-\left(1+\frac{x-r}{c}\right)^{-k},$$

$$Q_2(p;c,k,r) = c\left[(1-p)^{-1/k}-1\right]+r,$$

$$0<p<1, x\geq r, r>0, c>0, k>0.$$

The parameter c,k and r are scale, shape, and location parameters respectively. Using the result in (1), the pdf of the SN-LD is given by

$$f(x;\theta) = \begin{cases} u_1 f_1(x;\alpha,\tau,\varphi), & \text{if } -\infty < x \leq r, \\ u_2 \frac{k}{c}\left(1+\frac{x-r}{c}\right)^{-k-1}, & \text{if } r \leq x < \infty. \end{cases} \tag{19}$$

Using the assumptions (a-c) in Section 2, we obtain the following equations relating to the weights u_1 and u_2 and the parameter c of the model in (19):

$$u_1 = \frac{u_2 k}{c f_1(r;\alpha,\tau,\varphi)},$$

$$u_2 = \left\{1+\frac{k F_1(r;\alpha,\tau,\varphi)}{c f_1(r;\alpha,\tau,\varphi)}\right\}^{-1},$$

$$c = -(1+k)\frac{f_1(r;\alpha,\tau,\varphi)}{f_1(r;\alpha,\tau,\varphi)}.$$

The parameter vector θ in (19) contains only the free parameters and thus $\theta = [\alpha,\tau,\varphi,k,r]$. After θ has been estimated, the values of the other constrained parameters which are u_1,u_2 and c can be obtained from the above relations which describe them. Observe that in the model, there are eight parameters which we ought to estimate their values, however, with the imposition of the assumptions or constraints in Section 2, we were able to reduce the number of parameters to be estimated to just five.

The cdf and quantile function of the SN-PD are given respectively by

$$F(x;\theta) = \begin{cases} u_1 F_1(x;\alpha,\tau,\varphi), & \text{if } -\infty < x \leq r, \\ 1-u_2\left(1+\frac{x-r}{c}\right)^{-k}, & \text{if } r \leq x < \infty, \end{cases} \tag{20}$$

$$Q(p;\theta) = \begin{cases} Q_1\left(\frac{p}{u_1};\alpha,\tau,\varphi\right), & \text{if } p \leq u_1, \\ c\left[\left(\frac{1-p}{u_2}\right)^{-1/k}-1\right]+r, & \text{if } p \geq 1-u_2. \end{cases} \tag{21}$$

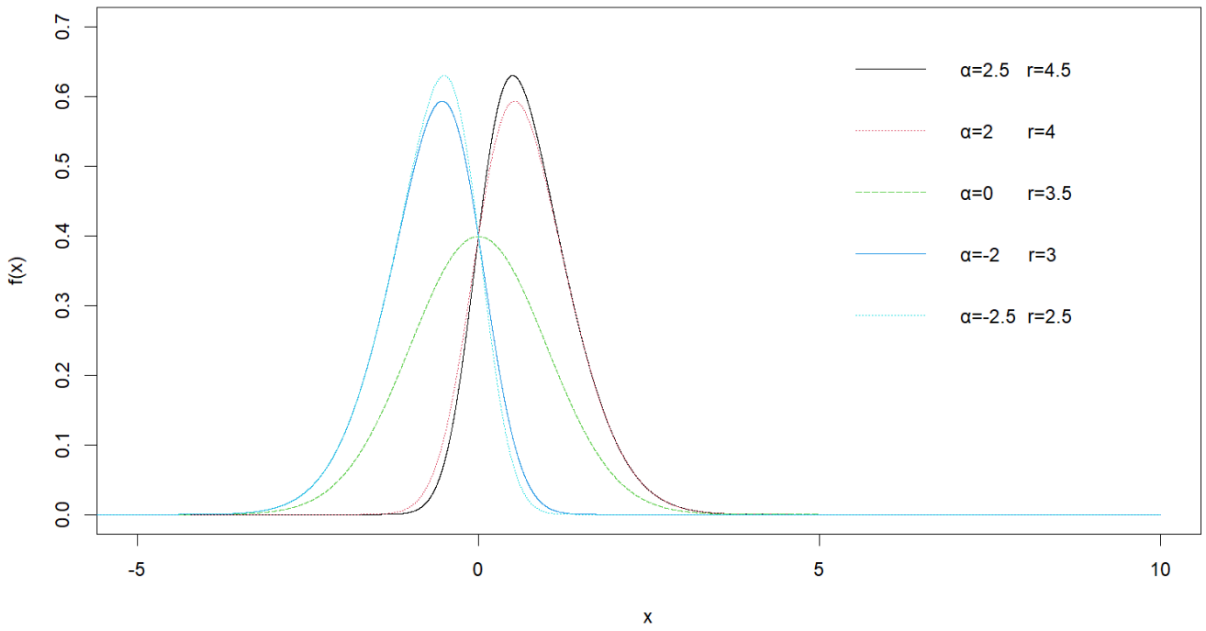

Figure 5: SN-LD density $[\tau = 0, \varphi = 0, k = 1.5]$

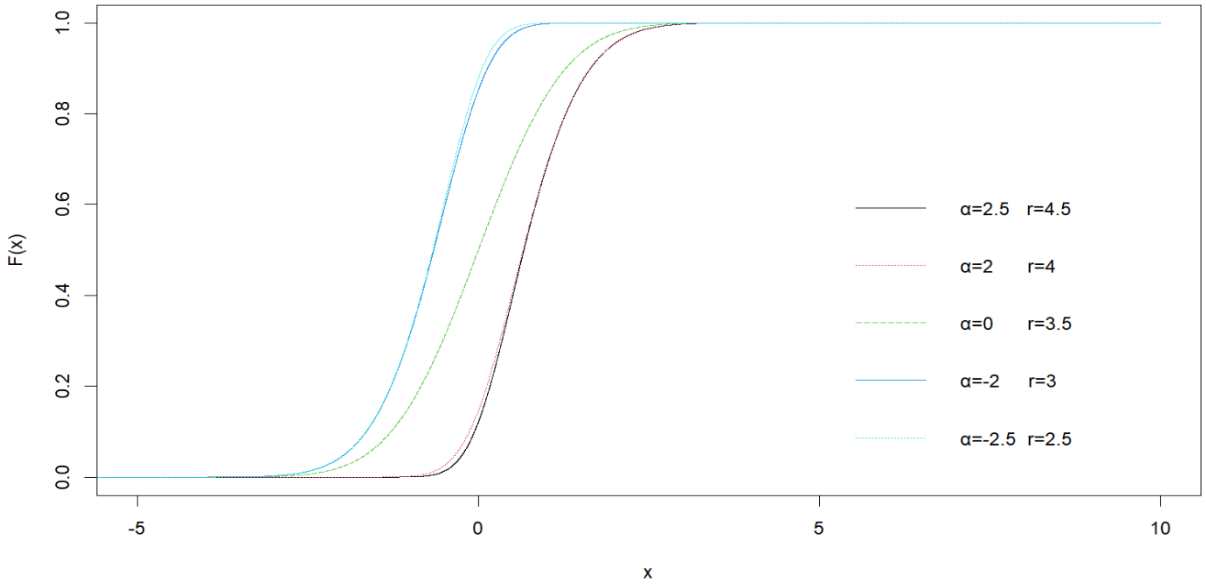

Figure 6: SN-LD cdf$[\tau = 0, \varphi = 0, k = 1.5]$.

Random samples X can be simulated from the SN-GPD by replacing p in (21) with U where U is a uniform random variable on (0,1). That is,

$$X = \begin{cases} Q_1\left(\frac{U}{u_1}; \alpha, \tau, \varphi\right), & \text{if } U \leq u_1, \\ c\left[\left(\frac{1-U}{u_2}\right)^{-1/k} - 1\right] + r, & \text{if } U \geq 1 - u_2. \end{cases} \tag{22}$$

5. Monte Carlo Simulation Study on the Maximum Likelihood Estimator of the Free Parameters of the SN-GPD

A Monte Carlo simulation study is carried out to assess the performance and efficiency of the maximum likelihood-based estimation algorithm described in Section 3 for obtaining estimates of the free parameters of the SN-GPD. The performance of the maximum likelihood estimates is examined for different sample sizes for a given combinations of parameter values. The simulation is repeated for $N = 100$ times using the sample sizes $n = 100, 250,\ 800, 1500$ and 2500 and parameter combination values $r = 4.5, \tau = 0, \varphi = 1, \alpha = 2.5, \gamma = 5$. Random samples are simulated from the SN-GPD using (13) and five quantities are computed in the simulations, and these include:

(a) Mean estimate (ME) of the maximum likelihood estimator of the parameter $\theta = (\alpha\ \tau\varphi\ r\ \gamma)$ where

$$\text{ME} = \frac{1}{N}\sum_{i=1}^{N}\hat{\theta}\,;$$

(b) Average bias (AVB) of the maximum likelihood estimator of the parameter $\theta = (\alpha\ \tau\varphi\ r\ \gamma)$ where

$$\text{AVB} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\theta} - \theta\right);$$

(c) Root mean squared error (RSME) of the maximum likelihood estimator of the parameter $\theta = (\alpha\ \tau\varphi\ r\ \gamma)$ where

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\theta} - \theta\right)^2};$$

(d) Coverage probability (CP) of 95% confidence intervals of the parameters $\theta = (\alpha\ \tau\varphi\ r\ \gamma)$, i.e., the percentage of intervals that contain the true value of parameter Θ;

(e) Average width (AW) of 95% confidence intervals of the parameter $\theta = (\alpha\ \tau\varphi\ r\ \gamma)$.

The results from the simulations are contained in Table 1.

Results from simulations study as presented in Table 1 clearly shows that despite the complexity in terms of the composite model structure, the maximum likelihood-based estimation algorithm designed in this paper offers a very efficient scheme in the estimation of the free parameters of the model. This is evident in the fact that the mean values of each parameter resulting from the given number of simulations are remarkably close to the true parameter values. Also, we observe that as the sample size increases, the root mean square error for each of the parameters decreases. This has helped to sustain the well-established hypothesis that as the sample size increases, the maximum likelihood estimator of a parameter becomes better. We also must mention at this point, that composite models of this nature perform better under conditions of large samples and not small samples.

Our simulation results have also validated this position. Again, we also observe that the average width of 95% confidence intervals of the parameters decreases as the sample size increases. This also emphasizes improvement in precision in terms of the estimation of the parameters. The average biases as well as the coverage probabilities for each of the parameters also emphasize the efficiency of the estimation method.

The same process for the simulations can be conducted for the SN-PD and the SN-LD. Different combinations of parameters values can equally be chosen for the analysis.

Table 1: Results of Monte Carlo simulations $r = 4.5, \tau = 0, \varphi = 1, \alpha = 2.5, \gamma = 5$

Parameter	Sample size	ME	AVB	RMSE	AW	CP
r	$n = 100$	3.0078	-1.4922	1.5558	2.0904	0.4
	$n = 250$	3.1633	-1.3367	1.4456	1.7264	0.3
	$n = 800$	3.4754	-1.0246	1.0829	1.6514	0.5
	$n = 1500$	3.5912	-0.9088	0.9491	1.5561	0.4
	$n = 2500$	3.8228	-0.6772	0.7198	1.4013	0.4
τ	$n = 100$	-0.0792	-0.0042	0.1935	1.2845	1
	$n = 250$	0.0205	0.0302	0.0817	0.3516	0.9
	$n = 800$	-0.0234	-0.0101	0.0502	0.1674	1
	$n = 1500$	0.0025	0.0086	0.0333	0.1259	1
	$n = 2500$	0.0003	0.0029	0.0284	0.0963	0.9
φ	$n = 100$	1.0102	0.0132	0.1867	0.5485	0.9
	$n = 250$	0.9760	-0.0329	0.1006	0.3162	0.9
	$n = 800$	1.0222	0.0089	0.0502	0.1620	0.8
	$n = 1500$	0.9838	-0.0123	0.0299	0.1176	1
	$n = 2500$	1.0005	0.0027	0.0196	0.0908	1
α	$n = 100$	2.7901	0.0132	1.9106	7.419	1
	$n = 250$	2.3493	-0.0329	0.8572	2.9131	0.9
	$n = 800$	2.6526	0.0089	0.3986	1.3723	1
	$n = 1500$	2.4119	-0.0122	0.2349	0.8952	0.9
	$n = 2500$	2.5246	0.0027	0.1555	0.7273	0.9
γ	$n = 100$	3.4259	0.7669	6.6839	23.1340	0.7
	$n = 250$	5.7917	2.8534	5.7116	21.0899	0.6
	$n = 800$	4.8034	-1.3316	3.1263	14.8773	0.7
	$n = 1500$	4.5153	-1.5672	2.8178	13.7926	0.8
	$n = 2500$	4.0469	-0.0853	4.4401	18.1339	0.9

6. Application

In this section, an application of the hybrid distributions framework will be applied to a financial data set: the Standard & Poor's 500 index, often abbreviated as the S&P500 index. The index, which is reported daily includes open prices, high prices, low prices, close prices, adjusted close prices and volume of S&P500. We shall focus on the index reported for the time period 1st January 1995 to 13th September 2023 with 7226 observations. The S&P500 index data set can easily be obtained from the Yahoo Finance database. Our goal is to model the log returns of the market. The log returns are obtained as the logarithm of the ratio of the current adjusted close price to the previous adjusted close price.

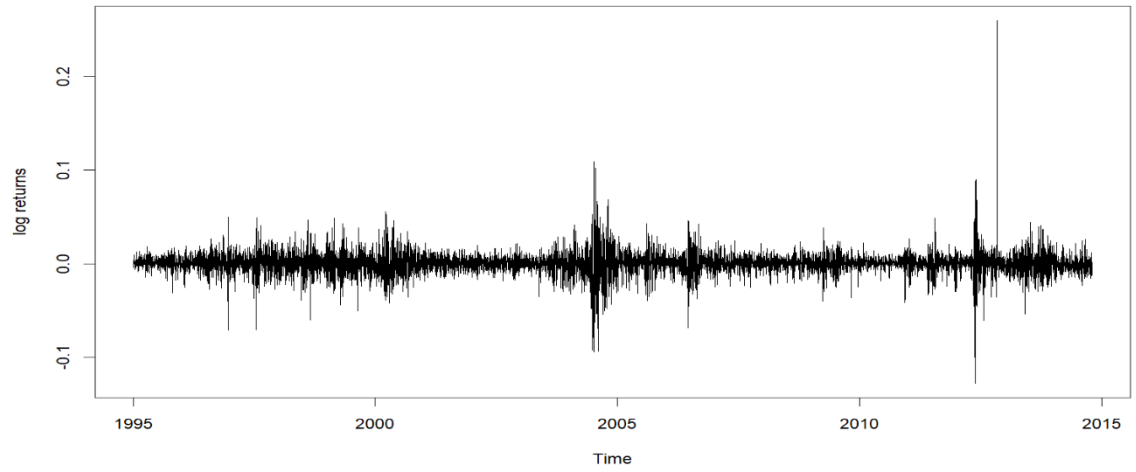

Figure 7: Log returns of the S&P500 from 1st January 1995 to 13th February 2023.

We fit the SNGPD, SNLD and the SNPD distributions to data and also fit the normal-GPD (N-GPD) by Debbabi et al. (2016) to the data and compare the fits of the four distributions. The results are reported in Table 2. These results include the estimates of the free parameters of the distributions (the estimates of the constraint parameters can easily be obtained from the relations specified in the previous sections by substituting the estimated values of the free parameters) and their various standard errors of estimate, the log-likelihood value, and the Akaike Information Criterion (AIC) value for each of the distribution. The Kolmogorov-Smirnov (K-S) statistic is also reported. The density plot (over the histogram of the data), the cdf plot, the Q-Q plots, and the P-P plots of the four fitted distributions are given by Figures 11-14(a-d).

Table 2: Results of estimation of the free parameters of the models

Distribution	SN-GPD	SN-PD	SN-LD	N-GPD
Free Parameter estimates	$\tau = -0.0013$ (0.0006) $\varphi = 0.0111$ (0.0001) $\alpha = 01268$ (0.0684) $\gamma = 0.4200$ (0.0860) $\mathrm{r} = 0.0237$ (0.0010)	$\tau = -0.0014$ (0.0006) $\varphi = 0.0111$ (0.0001) $\alpha = 0135$ (0.0664) $\mathrm{r} = 0.0225$ (0.0005)	$\tau = -0.0036$ (0.0004) $\varphi = 0.0117$ (0.0002) $\alpha = 0.3948$ (0.0511) $\mathrm{k} = 0.7611$ (0.1198) $\mathrm{r} = 0.0289$ (0.0010)	$\mu = -0.0002$ $(1.4e-05)$ $\sigma = 0.0111$ $(9.8e-05)$ $\mathrm{u} = 0.0587$ (0.0155) $\gamma = 0.0186$ $(7.2e-04)$
Loglik	22306.1	22180.57	21992.56	21879.51
AIC	−44602.2	−44353.1	−43975.1	−43751.02
K-S	0.0537	0.0723	0.0894	0.0888

(standard error of estimates in parenthesis)

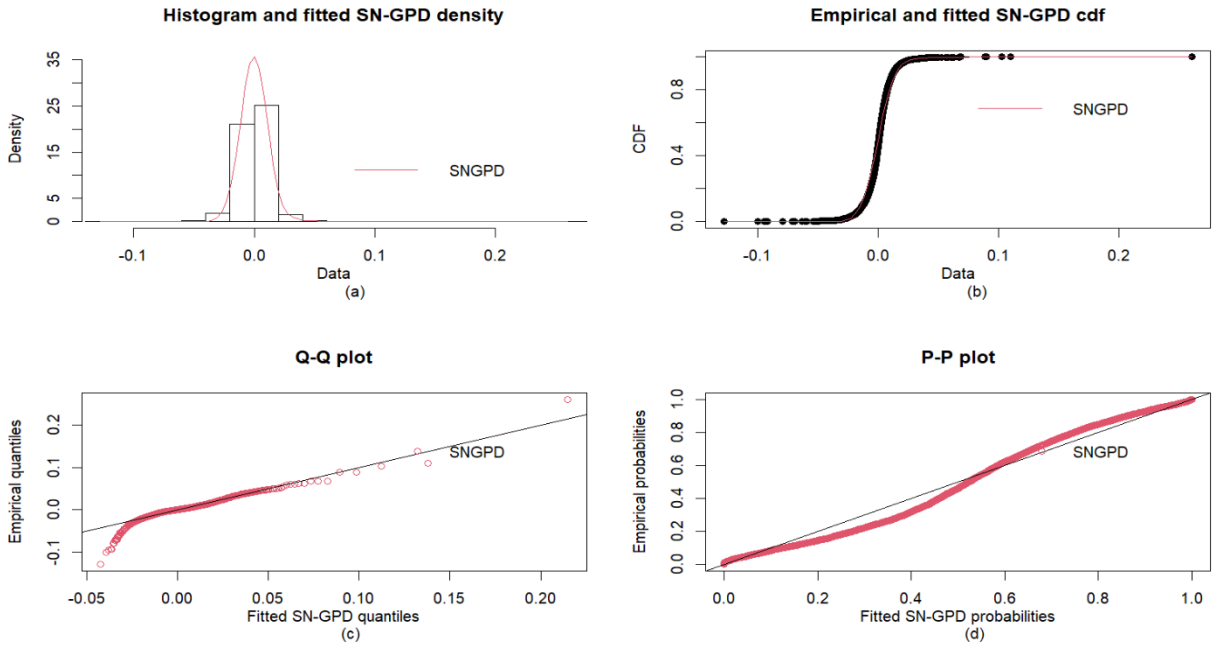

Figure 11 (a-d): Fitted SN-GPD model density, cdf, quantile and probabilities.

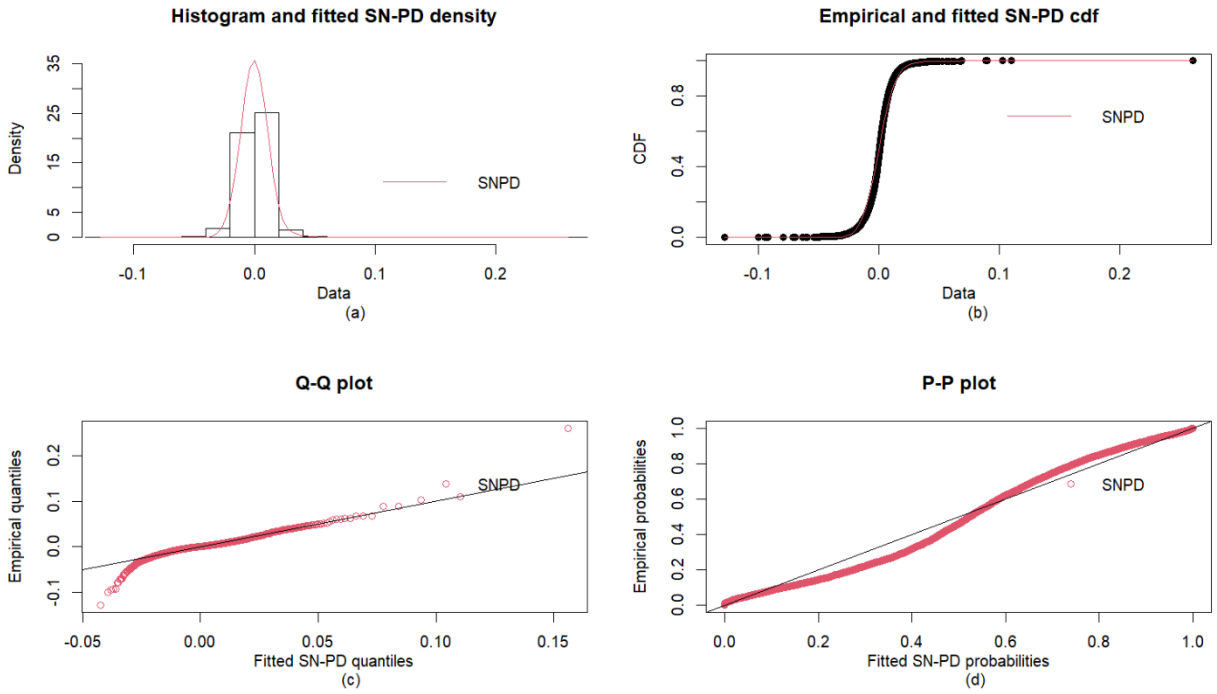

Figure 12 (a-d): Fitted SN-PD model density, cdf, quantile and probabilities.

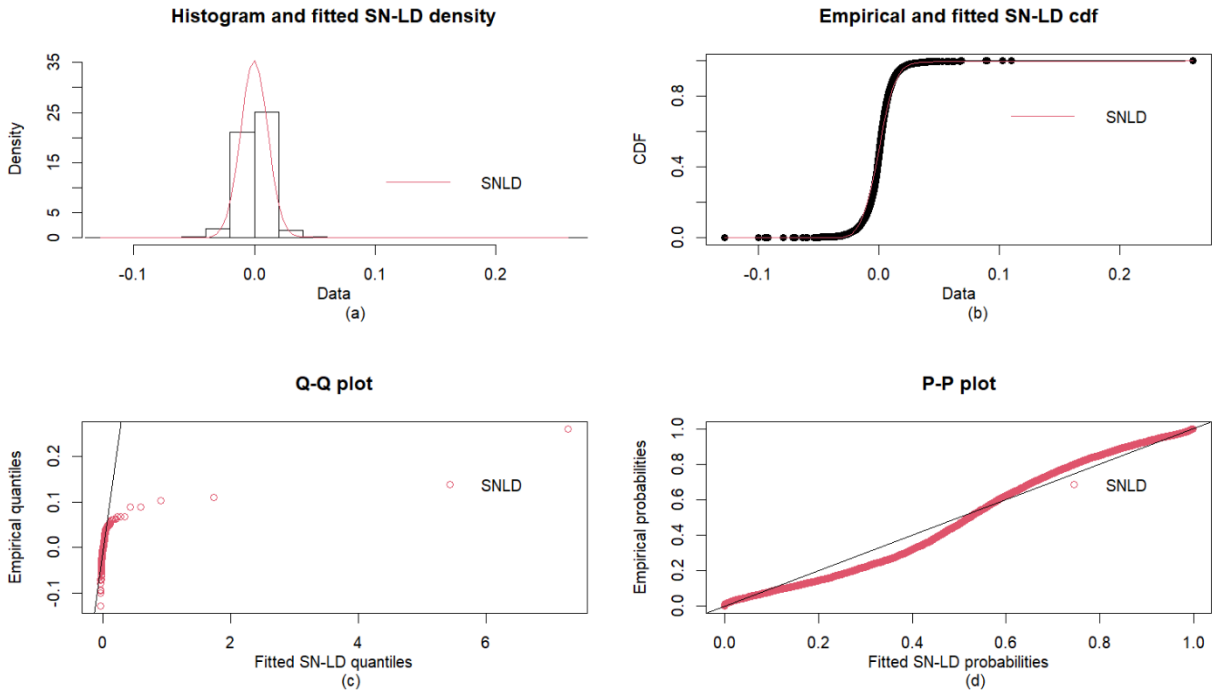

Figure 13 (a-d): Fitted SN-LD model density, cdf, quantile and probabilities.

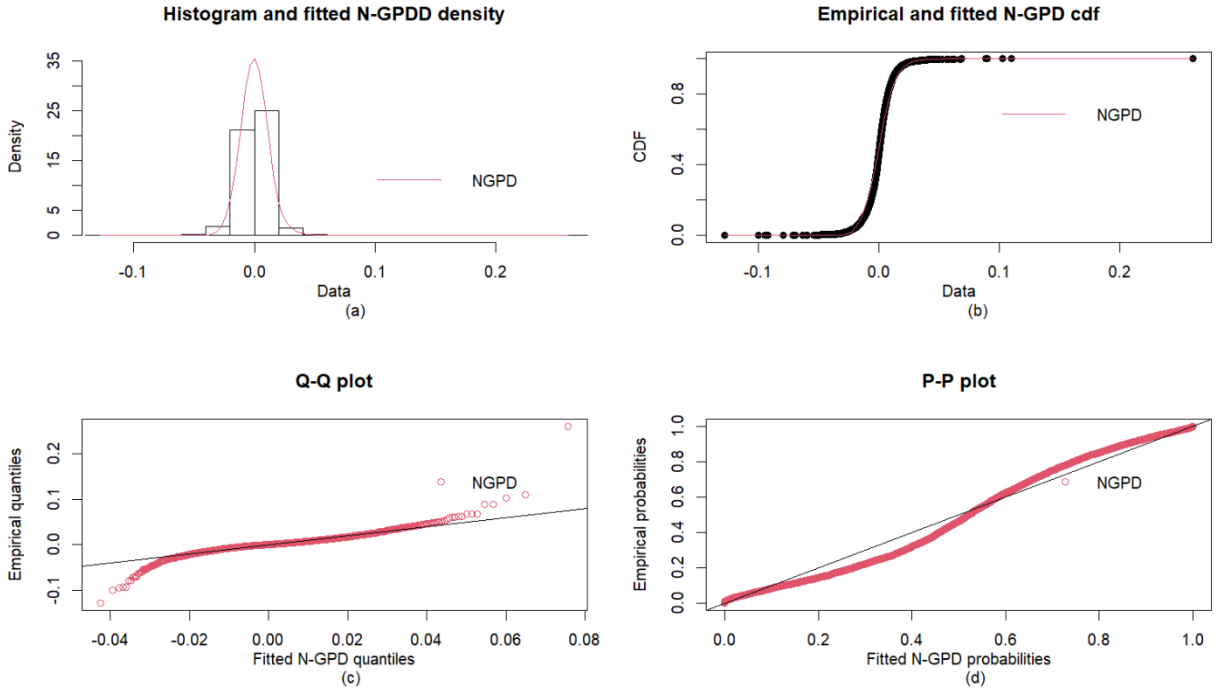

Figure 14 (a-d): Fitted N-GPD model density, cdf, quantile and probabilities.

Results from the application clearly showed that the hybrid models provide very good fit to the data with the SN-GPD reporting the best fit since it has the least AIC value.

7. Conclusion

A general family of two-component hybrid distributions which meets the need of modeling data sets with heavy-tails has been espoused in this study. We have also proposed an estimation procedure for estimating the free parameters of the family. An application of three distributions of the family to a real data set in finance has also been conducted. We hope that the proposed framework will receive further attention in terms of application in other areas of discourse.

References

[1] Bakar, S. A. A., Hamzah, N. A., & Nadarajah, S. (2015). Modeling loss data using composite models. *Insurance: Mathematics and Economics, 61*, 146-154. https://doi.org/10.1016/j.insmatheco.2014.08.008

[2] Benatmane, C., Zeghdoudi, H., Shanker, R., & Lazri, N. (2020). Composite Rayleigh-Pareto distribution : Application to real fire insurance losses data set. *Journal of Statistics and Management Systems, 24*(3), 545-557. https://doi.org/10.1080/09720510.2020.1759253

[3] Carreau, J., & Bengio, Y. (2009). A hybrid Pareto mixture for conditional asymmetric fat-tailed distributions. *IEEE Transactions on Neural Networks, 7*, 1087-1101. https://doi.org/10.1109/TNN.2009.2016339

[4] 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. https://doi.org/10.1080/03610920802484100

[5] Cooray, K., & Ananda, M. M. A. (2005). Modeling actuarial data with a composite lognormal-Pareto model. *Scandinavian Actuarial Journal, 2005*(5), 321-334. https://doi.org/10.1080/03461230510009763

[6] Cooray, K., Gunasekera, S., & Ananda, M. (2010). Weibull and inverse Weibull composite distribution for modeling reliability data. *Model Assisted Statistics and Applications, 5*(2), 109-115. https://doi.org/10.3233/MAS-2010-0149

[7] Debbabi, N., El Asmi, S., & Mboup, M. (2015). Distribution hybride pour la modélisation de données à deux queues lourdes: Application sur les données neuronales. *Groupe d'Études du Traitement du Signal et des Images* (*GRETSI*).

[8] Debbabi, N., Kratz, M., & Mboup, M. (2016). A self-calibrating method for heavy-tailed modeling: Application in neuroscience and finance. *ESSEC Working Paper*, 1619. https://doi.org/10.2139/ssrn.2898731

[9] Li, C., Singh, V. P., & Mishra, A. K. (2012). Simulation of the entire range of daily precipitation using a hybrid probability distribution. *Water Resources Research, 48*, 1-17. https://doi.org/10.1029/2011WR011446

[10] Mandava, A., Latifi, S., & Emma, R. (2011). Reliability assessment of microarray data using fuzzy classification methods: A comparative study. *Communications in Computer and Information Science, 190*, 351-360. https://doi.org/10.1007/978-3-642-22709-7_36

[11] Nadarajah, S., & Bakar, S. (2014). New composite models for the Danish fire insurance data. *Scandinavian Actuarial Journal, 2014*(2), 180-187. https://doi.org/10.1080/03461238.2012.695748

[12] Preda, V., & Ciumara, R. (2006). On composite models: Weibull-Pareto and lognormal-Pareto—A comparative study. *Romanian Journal of Economic Forecasting, 3*, 32-46.

[13] Scollnik, D. P. (2007). On composite lognormal-Pareto models. *Scandinavian Actuarial Journal, 2007*(1), 20-33. https://doi.org/10.1080/03461230601110447

[14] Scollnik, D. P., & Sun, C. (2012). Modeling with Weibull-Pareto models. *North American Actuarial Journal, 16*(2), 260-272. https://doi.org/10.1080/10920277.2012.10590640

[15] Teodorescu, S., & Vernice, R. (2006). A composite exponential-Pareto distribution. *Analele Științifice ale Universității Ovidius Constanța, 14*, 99-108.

[16] Teodorescu, S., & Vernice, R. (2009). Some composite exponential-Pareto models for actuarial prediction. *Romanian Journal of Economic Forecasting, 12*, 82-100.

[17] Teodorescu, S., & Vernice, R. (2013). On some Pareto models. *Mathematical Reports, 1*, 11-29.