doi:10.20944/preprints202104.0083.v1

Article

Distribution-based entropy weighting clustering of skewed time series

Abstract: The goal of clustering is to identify common structures in a data set by forming groups of homogeneous objects. The observed characteristics of many economic time series have motivated the development of classes of distributions that can accommodate properties such as heavy tails and skewness. Thanks to its flexibility, the Skew Exponential Power Distribution (also called Skew Generalized Error Distribution) ensures a unified and general framework for clustering possibly skewed time series. This paper develop a clustering procedure of model-based type, assuming that the time series are generated by the same underlying probability distribution but with different parameters. Moreover, we propose to optimally combine all the parameter estimates to form the clusters with an entropy weighing *k*-means approach. The usefulness of the proposal is showed by means of an application to financial time series, showing also how the obtained clusters can be used to form portfolio of stocks.

Keywords: Classification, Generalized Error Distribution, Skewness, Skewed Exponential Power Distribution, Portfolio selection

1. Introduction

The goal of clustering is to identify common structures in a data set by forming groups of homogeneous data. This objective can be achieved by minimizing the within-group similarity and by maximizing the between-group dissimilarity.

Time series clustering has been shown effective in providing useful information in various domains from science and engineering to finance and economics. For example, trough clustering methods it is possible to build portfolios of similar stocks for financial applications (e.g. [1–3]). From a methodological point of view, time series clustering methods can be divided into three main classes [4]: observation-based, feature-based and model-based.

In the observation-based clustering the raw are clustered according to a specified distance measure. Several authors proposed fuzzy extensions of common clustering algorithms for raw data (e.g. [5–9]). The time series involved could have either the same length or not. In the second case, the observation-based clustering methods are usually built upon the so-called Dynamic Time Warping (DTW), a technique that allows finding an optimal alignment between two given sequences of different length. (e.g. [9,10])

In the feature-based clustering the objects are clustered according to some data's features. The main advantage of this class of clustering approaches lies on the fact that the time series length is not an issue because also the objects with different length can be clustered together. Common time series features considered for clustering are the autocorrelation function (ACF) [11,12], the partial autocorrelation function (PACF) [13], the features of wavelet decomposition of the time series (e.g. [14,15]) or the ceptral (e.g. [16,17]).

The model-based clustering approaches assume, instead, that the time series are generated by the same statistical model (e.g. [18–21]) or that they have the same probability distribution (e.g. [22,23]). The spirit of most of the model-based clustering procedures is to group objects according to the estimated parameters. Important examples are the clustering methods based on ARMA processes distances (e.g. [18,19,24]), GARCH-based distances for heteroskedastic time series [19,20,25], estimates of the probability distribu-

tions' parameters (e.g. [22,23]) or, more recently, conditional higher moments (e.g. see [26]).
This paper develop a clustering procedure of model-based type, assuming that the time series are generated by the same underlying probability distribution but with different parameters. Clearly, at this aim the specification of a very general distribution is required in order to account for a wide range of possible special cases.
Observed characteristics of many economic time series have motivated exploration of classes of distributions that can accommodate properties such as fat-tailedness and skewness while nesting distributions typically used in estimation such as the normal (and skew-normal).
An important desired property of any such class is that it permits maximum likelihood estimation of all parameters. Classes of asymmetric distributions that nest the skew normal were firstly constructed by [27]. Other classes of distributions with the desired properties of accommodating heavy tails and skewness, the Skewed Exponential Power Distribution (SEPD) classes, were proposed in [28–31]. They all generalize the generalized error distribution (GED) class. Many financial applications of the GED as well as its skew extensions have been considered (e.g. [30–38]).
For example, [31] explored moments (also see [30]) as well as measures such as value at risk and expected shortfall useful in financial applications. Moreover, [39] studied the maximum likelihood properties of the [27] version of the SEPD, while [40], introduced a new different parametrization where the shape of the tails are not symmetric.
In general, the Exponential Power Distribution (either symmetric or not) encompasses a very wide amount of special cases. Examples are the Gaussian, the skew Normal, the Laplace, the Asymmetric Laplace distribution and much others [37,40–42].
Therefore, in what follows we consider the Skew Exponential Power Distribution family as the underlying assumption for all the considered time series. Thanks to its flexibility it ensures a unified and general framework for clustering possibly skewed time series.
The paper is structured as follows. In the next section, the Skew Exponential Power Distribution is presented. Then, in Section 3, we discuss the clustering algorithm in details. To show the usefulness of the proposed approach we provide an application to financial time series with the aim of building a portfolio of stocks. At the end some conclusions are offered.

2. Skewed Exponential Power Distribution

A very general and flexible family of distribution is represented by the Exponential Power Distribution (also called Generalized Error Distribution or Exponential Power Function).
The EPD random variable Z has the following probability density function [42,43]:

$$f(z) = \frac{p\exp(\frac{1}{2}|\frac{z-\mu}{\sigma}|^p)}{2p^{(1+\frac{1}{p})}\sigma\Gamma(\frac{1}{p})} \tag{1}$$

where $z \in \mathbf{R}$, $\mu \in (-\infty, +\infty)$ is called location parameter, $\sigma > 0$ is called scale parameter, $p > 0$ is a measure of fatness of tails and is called shape parameter (see [40]) and:

$$\Gamma(a) = \int_0^\infty x^{a-1}\exp(-x)dx. \tag{2}$$

is the Gamma function. Since the distribution is symmetric and unimodal, the location parameter is also the mode, median and mean of the distribution (Fig. 1).

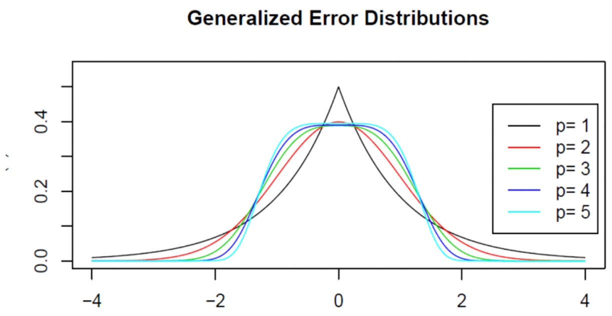

Figure 1. Generalized Error Distribution for different values of skewness.

It is possible to write the EPD probability density (1) in more compact form by means of [40]:

$$f(z) = \frac{1}{\sigma} C_{EPD}(p) exp\left(-\frac{1}{p} \left| \frac{z-\mu}{\sigma} \right|^{p} \right) \tag{3}$$

where $C_{EPD}(p)$ is a normalizing constant, $C_{EPD}(p) = 1/[2p^{1/p}\Gamma(1+1/p)]$.
The shape parameter p controls the tails and the peak of the distribution; a small value of p means that the tails of the distribution become flat, with the center becoming largely peaked.
A very important feature of this family of distributions, that has been proved to be useful in modeling stock market volatility (e.g. [37,38,44]), is that they include also other common distributions, for different values of shape parameter p.
In particular, the Gaussian distribution is a special case of the GED when $p = 2$, and when $p < 2$ the distribution has fatter tails than a Gaussian distribution [37]. Moreover, when $p = 1$ we have a Laplace distribution, and for $p = +\infty$ we have the Uniform distribution [42].
So far, there are two different methods to extend the EPD for skewness (see Fig. 2).

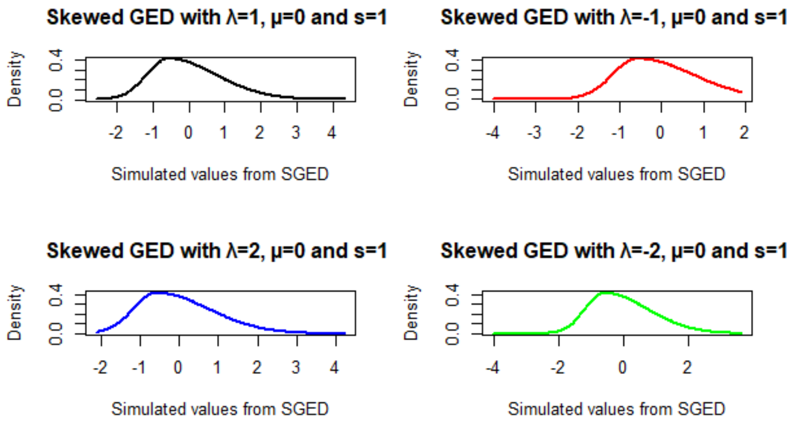

Figure 2. Skewed Generalized Error Distribution for different values of skewness.

A first approach is represented by the one of [27], that defined the first family of SEPD. Later, [28,29] extended the EPD class to another family of SEPD by using a two-piece method, in which an additional skew parameter γ (that henceforth we define as λ) is introduced. By a method similar to that of [28,29] [30] [31], respectively, constructed seemingly different classes of SEPD, which are actually reparametrizations of the one developed by [28,29]. In what follows we consider the SPED family of [28,29,35].
A random variable Z is said to have an Skew Exponential Power Distribution if there

exist parameters $p > 0$, $\mu \in \mathbb{R}$, $\sigma > 0$, and $\lambda > 0$ such that the density function has the form:

$$f(z) = \frac{p}{\sigma\Gamma\left(\frac{1}{p}\right)} \frac{\lambda}{1+\lambda^2} \exp\left(-\frac{\lambda^p}{\sigma^p}[(z-\mu)^+]^p - \frac{1}{\sigma^p\lambda^p}[(z-\mu)^-]^p\right) \tag{4}$$

where:

$$(z-\mu)^+ = \begin{cases} (z-\mu) & \text{if} \quad (z-\mu) \geq 0, \\ 0 \quad \text{if} & (z-\mu) < 0, \end{cases} \quad \text{and} \quad (z-\mu)^- = \begin{cases} (\mu - z) & \text{if} \quad (z-\mu) \leq 0, \\ 0 \quad \text{if} & (z-\mu) > 0 \end{cases}$$

The parameters μ and σ correspond to location and scale, respectively, while λ controls skewness, and p is the shape parameter. For $\lambda = 1$, the distribution is symmetric about μ so we obtain the symmetric exponential power distribution. In case $\lambda \neq 1$, letting $p = 1$ leads to the skew Laplace distribution with density:

$$f(z) = \frac{1}{\sigma} \frac{\lambda}{1+\lambda^2} \begin{cases} \exp(-\frac{\lambda}{\sigma}|z-\mu|) & \text{for} \quad z \geq \mu, \\ \exp(-\frac{1}{\sigma\lambda}|z-\mu|) & \text{for} \quad z < \mu \end{cases} \tag{5}$$

For $p = 2$, we obtain the skew normal distribution. The moments of the Skew Exponential Power Distribution are the following [35].
The mean is equal to:

$$E(Z) = \mu + \sigma\left(\frac{1}{\lambda} - \lambda\right)\frac{\Gamma(\frac{2}{p})}{\Gamma(\frac{1}{p})}, \tag{6}$$

while the variance:

$$V(Z) = \sigma^2 \frac{\Gamma(\frac{3}{p})}{\Gamma(\frac{1}{p})} \frac{1+\lambda^6}{\lambda^2(1+\lambda^2)} - \sigma^2 \frac{\Gamma^2(\frac{2}{p})}{\Gamma^2(\frac{1}{p})} \frac{(1-\lambda^2)^2}{\lambda^2}. \tag{7}$$

The skewness and the excess of kurtosis, instead, can be retrieved by means of:

$$S(Z) = E\left[\left(\frac{Z-\mu}{\sigma}\right)^3\right]$$

$$K(Z) = E\left[\left(\frac{Z-\mu}{\sigma}\right)^4\right] - 3$$

Hence, the skewness is equal to:

$$S(Z) = \frac{(1-\lambda^8)\Gamma^2(\frac{1}{p})\Gamma(\frac{4}{p}) - 3(1-\lambda^2)(1+\lambda^6)\Gamma(\frac{1}{p})\Gamma(\frac{2}{\lambda})\Gamma(\frac{3}{\lambda}) + 2(1-\lambda^2)^3(1+\lambda^2)\Gamma^3(\frac{2}{p})}{(1+\lambda^2)\left(\sqrt{\Gamma(\frac{1}{\alpha})\Gamma(\frac{3}{p})\frac{1+\lambda^6}{1+\lambda^2} - \Gamma^2(\frac{2}{p})(1-\lambda^2)^2}\right)^3} \tag{8}$$

while the (excess) kurtosis is:

$$
\begin{aligned}
K(Z) = \frac{1+\lambda^2}{\left[\Gamma(\frac{1}{p})\Gamma(\frac{3}{p})(1+\lambda^6)-\Gamma^2(\frac{2}{p})(1-\lambda^2)^2(1+\lambda^2)\right]^2}\times \\
\Bigg\{\Gamma^3\left(\frac{1}{p}\right)\Gamma\left(\frac{5}{p}\right)(1+\lambda^{10})-4\Gamma^2\left(\frac{1}{p}\right)\Gamma\left(\frac{2}{p}\right)\Gamma\left(\frac{4}{p}\right)(1-\lambda^2)(1-\lambda^8) \\
+6\Gamma\left(\frac{1}{p}\right)\Gamma^2\left(\frac{2}{p}\right)\Gamma\left(\frac{3}{p}\right)(1-\lambda^2)^2(1+\lambda^6)-3\Gamma^4\left(\frac{2}{p}\right)(1-\lambda^2)^4(1+\lambda^2)\Bigg\}-3
\end{aligned}
\tag{9}
$$

Note that in the special case $p = 1$ (Laplace distribution) we have:

$$S(Z) = 2\frac{\frac{1}{\lambda^3}-\lambda^3}{(\frac{1}{\lambda^2}+\lambda^2)^{\frac{3}{2}}}$$

and:

$$K(Z) = 6 - \frac{12}{(1/\lambda^2+\lambda^2)^2}$$

which coincide with the parameters of the Asymmetric Laplace Distribution [35,41].
An interesting alternative parametrization is the Asymmetric Exponential Power Distribution developed by [40] with the following density:

$$
\begin{aligned}
f(z) &= \left(\frac{\lambda}{\lambda*}\right)\frac{1}{\sigma}C_{EPD}(p_1)exp\left(-\frac{1}{p_1}\left|\frac{z-\mu}{2-(\lambda^*\sigma)}\right|^{p_1}\right) \quad \text{if} \quad (z \geq \mu) \\
&= \left(\frac{1-\lambda}{1-\lambda^*}\right)\frac{1}{\sigma}C_{EPD}(p_2)exp\left(-\frac{1}{p_2}\left|\frac{z-\mu}{2(1-\lambda^*)\sigma}\right|^{p_2}\right) \quad \text{if} \quad (z < \mu)
\end{aligned}
$$

where $\mu \in R$ and $\sigma > 0$ still represent location and scale, respectively, $\lambda \in (0,1)$ is the skewness parameterization, $p_1 > 0$ and $p_2 > 0$ are the left and right tail parameters, respectively, and $C_{EPD}(p)$ is the constant defined before. This representation suppose that the two tails have different shapes p_1 and p_2. If $p_1 = p_2 = p$, implying $\lambda^* = \lambda$, the AEPD (SPED) reduces to:

$$
\begin{aligned}
f(z) &= \frac{1}{\sigma}C_{EPD}(p)exp\left(-\frac{1}{p}\left|\frac{z-\mu}{2\lambda\sigma}\right|^{p}\right) \quad \text{if} \quad (z \leq \mu) \\
&= \frac{1}{\sigma}C_{EPD}(p)exp\left(-\frac{1}{p}\left|\frac{z-\mu}{2(1-\lambda)\sigma}\right|^{p}\right) \quad \text{if} \quad (z > \mu)
\end{aligned}
$$

which is equivalent, but with a different parametrization, to those developed by [28–31,35].

3. The entropy weighting clustering approach

The proposed clustering algorithm uses the estimated moments from the Skew Exponential Power Distribution introduced in previous section to form clusters. In other words, time series with similar moments' estimates will be placed in the same group. Therefore, with a k-means clustering algorithm, the dissimilarity measure is computed on the basis of these estimates.
Therefore, assuming to have $N(n = 1,\dots,N)$ time series all generated by the Skew Exponential Power Distribution of parameters μ_n, σ_n, p_n and λ_n, we can store the moments' estimates in the following matrix:

$$\mathbf{X} = \begin{bmatrix} \mu_1 & \sigma_1 & p_1 & \lambda_1 \\ \mu_2 & \sigma_2 & p_2 & \lambda_2 \\ \vdots & \vdots & \vdots & \vdots \\ \mu_n & \sigma_n & p_n & \lambda_n \\ \vdots & \vdots & \vdots & \vdots \\ \mu_N & \sigma_N & p_N & \lambda_N \end{bmatrix} \tag{10}$$

that we can use to compute the time series' dissimilarities.

However, since the specified distribution has more than one parameter, a natural question is how would we use all this information. Indeed, one can cluster the time series only according to the location estimates or with respect to the time series distribution's scale. Similarly, one can be interested in clustering time series with similar skewness or shape.

In this paper, following in the spirit [22,23], we propose to optimally combine all the information to form the clusters.

A useful approach for optimally weighting different data features is represented by the weighted k-means algorithm of [45]. The algorithm of [45] is based on the k-means approach, where the weights are incorporated into the distance function. The main idea is that the weights are a measure of the relative importance of each variable with regard to the membership of the observations to that cluster.

Formally, the Weighted k-Means algorithm (WKM) can be formalized as follows:

$$\min : \sum_{c=1}^{C} \sum_{n=1}^{N} \sum_{m=1}^{M} u_{n,c} w_{m,c}^{\beta} D_{m,c}^{2} \tag{11}$$

under the constraints:

$$\sum_{c=1}^{C} u_{n,c} = 1, \quad u_{n,c} \geq 0, \tag{12}$$

$$\sum_{m=1}^{M} w_{m,c} = 1, \quad 0 \leq w_{m,c} \leq 1 \tag{13}$$

where $D_{m,c} = d(x_{n,m}, x_{c,m})$, represents the (Euclidean) distance between the m-th feature in the matrix $\mathbf{X}$ shown in (10) of the n-th time series and the one of the c-th centroid.

Applied to the context of the distribution-based clustering, the weights are suitable values associated to each parameter m of the specified distribution within the c-th cluster. Note that, the weight $w_{m,c}$ are intrinsically associated with the squared distance $D_{n,c}^2$ for the specified distribution parameters, while the overall dissimilarity is just a sum of the squares of these weighted distances. This allows us to appropriately tune the influence of the different distribution features when computing the dissimilarity between time series. Moreover, another appealing feature is that each c-th group has its own optimal weighting.

In the end, the exponent β has to be analyzed. With $\beta = 0$ we obtain the usual k-means clustering algorithm, while with a value of $\beta = 1$, we have that the estimated distribution parameter with the smallest value of the weighted dissimilarity is equal to 1 and all the others $w_{m,c}$ are equal to zero.

When $\beta > 1$, the larger the D_m, the smaller the weight w_m. Therefore, the effect of a moment with a large D_m is reduced. When $\beta < 0$, the larger D_m, the larger the weight w_m. However, w_m becomes smaller and has less weighting to the moment in the distance calculation because of negative β.

In the end, if $0 < \beta < 1$ the larger the parameters' dissimilarity, the larger is the weight w_m and this is against the variable weighting principal [45].

Therefore we cannot choose $0 < \beta < 1$ but we can choose $\beta < 0$ or $\beta > 1$ in the weighted k-means algorithm.

However the exponent β is an artificial device, lacking a strong theoretical justification [7]. Indeed, the value of β in the formula (11) is similar to the fuzzines parameter in the fuzzy c-means algorithm. Consequently, the adoption of regularization terms to be juxtaposed to the maximum internal homogeneity criterion has been seen as a valid alternative (e.g. [7,46]). In this case, the burden of represented by β is shifted to the regularization term, in the form of a weighting factor multiplying the contribution of the regularization function to the clustering criterion.

With this respect, [46] proposed a clustering algorithm where the weight of a given dimension in a cluster represents the probability of contribution of that dimension in forming the cluster. The *entropy* of the dimension weights represents the certainty of dimensions in the identification of a cluster.

Therefore, [46] modified the objective function (11) by adding the weight entropy term to it so that we can simultaneously minimize the within cluster dispersion and maximize the negative weight entropy to stimulate more dimensions to contribute to the identification of clusters.

The new objective function can be written as follows:

$$\min : \sum_{c=1}^{C} \left[\sum_{n=1}^{N} \sum_{m=1}^{M} u_{n,c} w_{n,m} D^2_{m,c} + \gamma \sum_{m=1}^{M} w_{n,m} \ \log(w_{n,m}) \right] \tag{14}$$

subject to the constraints:

$$\sum_{c=1}^{C} u_{n,c} = 1, \quad u_{n,c} \geq 0, \tag{15}$$

$$\sum_{m=1}^{M} w_{m,c} = 1, \quad 0 \leq w_{m,c} \leq 1 \tag{16}$$

The first term is the sum of the within cluster dispersion, and the second term the negative weight entropy. The positive parameter γ controls the strength of the incentive for clustering on more dimensions [46].

The algorithm works as follows. An initial set of k means are identified as the starting centroids. An initial cluster is defined considering that the observations are clustered to the nearest centroid according to the Euclidean distance measure among distribution parameter estimates. Then, the centroids are identified based on these clusters, while the weights are computed for each time series within each cluster. New centroids are then calculated, and using the weighted distance measure each observation is once again clustered to its nearest centroid until convergence of the algorithm.

In the case of Skew Exponential Power Distribution, the optimal weights of the SEPD-DWEKM model, obtained by the solution of the optimization problem (14), are equal to:

$$w_{m',c} = \frac{\exp(\frac{-D_{m',c}}{\gamma})}{\sum_{m=1}^{4} \exp(\frac{-D_{m,c}}{\gamma})} \tag{17}$$

Proof of 17. By using the Lagrangian multiplier technique we ontain the following unconstrained minimization problem:

$$\min : \sum_{c=1}^{C} \left[\sum_{n=1}^{N} \sum_{m=1}^{4} u_{n,c} w_{m,c} D^2_{m,c} + \gamma \sum_{m=1}^{4} w_{m,c} (\log w_{m,c}) \right] - \sum_{c=1}^{C} \lambda_c \left(\sum_{m=1}^{4} w_{m,c} - 1 \right) \tag{18}$$

where $[\lambda_1, \ldots, \lambda_C]$ is a vector containing the Lagrange multipliers corresponding to the constraints. The optimization problem can be decomposed into C independent minimization problems:

$$\min : \sum_{n=1}^{N} \sum_{m=1}^{4} u_{n,c} w_{m,c} D_{m,c}^2 + \gamma \sum_{m=1}^{4} w_{m,c} \log(w_{m,c}) - \lambda_c \left(\sum_{m=1}^{4} w_{m,c} - 1 \right)$$

for $c = 1, \ldots, C$. By setting the gradient with respect to $w_{m,c}$ and λ_c to zero, we obtain:

$$\left(\sum_{m=1}^{4} w_{m,c} - 1 \right) = 0$$

and:

$$\sum_{n=1}^{N} u_{n,c} D_{m',c}^2 + \gamma \left[1 + \log(w_{m',c}) \right] - \lambda_c = 0.$$

From the last equality we get:

$$w_{m',c} = \exp\left(\frac{-D_{m',c}^2 - \gamma + \lambda}{\gamma} \right) = \exp\left(\frac{\lambda_c - \gamma}{\gamma} \right) \exp\left(\frac{-D_{m',c}^2}{\gamma} \right)$$

where $D_{m',c}^2$ can be interpreted as a measure of the data dispersion of the m' dimension for the objects placed within the c-th cluster. By substitution of the above equations we get:

$$\begin{aligned} \sum_{m=1}^{4} w_{m,c} = \sum_{m=1}^{4} \exp\left(\frac{\lambda_c - \gamma}{\gamma} \right) \exp\left(\frac{-D_{m,c}^2}{\gamma} \right) = \\ = \exp\left(\frac{\lambda_c - \gamma}{\gamma} \right) \sum_{m=1}^{4} \exp\left(\frac{-D_{m,c}^2}{\gamma} \right) = 1. \end{aligned}$$

it follows that:

$$\exp\left(\frac{\lambda_c - \gamma}{\gamma} \right) = \frac{1}{\sum_{m=1}^{4} \exp\left(\frac{-D_{m,c}^2}{\gamma} \right)}$$

Substituting back we obtain the (17):

$$w_{m',c} = \frac{\exp(\frac{-D_{m',c}^2}{\gamma})}{\sum_{m=1}^{4} \exp(\frac{-D_{m,c}^2}{\gamma})}$$

Similarly to the standard k-means algorithm $u_{n,c}$ is updated:

$$\begin{cases} u_{n,c} = 1 & \text{if} \quad \sum_{m=1}^{4} w_{m,c} D_{m,c}^2 \leq \sum_{m=1}^{4} w_{m,c'} D_{m,c'}^2 \\ u_{n,c} = 0 & \text{otherwise} \end{cases}$$

where $u_{n,c} = 1$ means that the n-th object is assigned to the c-th cluster, so we have an hard, not fuzzy, final assignment. If the distances between an object and two cluster centers are equal, the object is arbitrarily assigned to the cluster with the smaller cluster index number.

As clearly appears from the clustering model, the parameter γ plays an important role since it is used to control for the size of the weights.

Indeed, if $\gamma > 0$, the weights $w_{m,c}$ are inversely proportional to squared distance $D^2_{m,c}$. Therefore, the smaller $D^2_{m,c}$, the larger the weights $w_{m,c}$ and, hence, the more important the corresponding dimension m.

Instead, if $\gamma < 0$, the weights $w_{m,c}$ is proportional to the distance $D^2_{m,c}$. Therefore, the larger is the distance the larger is the associated weight. This is a contradictory result and, hence, γ cannot be smaller than zero.

In the end, γ can be set equal to zero. In this case, the dimension m' with the smallest distance has a weight equal to 1, $w_{m',c} = 1$, while all the others are zero $w_{m,c} = 0$. Therefore, each cluster contains only one important dimension.

A final crucial aspect of the any clustering procedure is the selection of the number of clusters C. With this respect we compute the Silhouette Width Criterion (SWC).

In order to define this criterion, let us consider an n-th time series that belongs to a given cluster $c \in \{1, ..., C\}$. Then, let the distance of this object to all other time series in the cluster c be denoted by $a_{n,c}$ and the average distance of this n-th time series from all the others in another cluster c', $c' \neq c$, be called $d_{n,c'}$. Finally, let $b_{n,c'}$ be the minimum $d_{n,c'}$ computed over $c' = 1, ..., C$, which represents the average dissimilarity of object n to its closest neighboring cluster. Then, the silhouette of the individual n-th time series is defined as:

$$s_n = \frac{b_{n,c'} - a_{n,c}}{\max\{a_{n,c}, b_{n,c'}\}} \tag{19}$$

where the denominator is a normalization term. The higher s_n the better is the assignment of the n-th time series to the c-th cluster. In the case c is constituted by a single time series, then by convention we assume $s_n = 0$. This prevents the SWC, defined as the average of s_n over $n = 1, 2, ..., N$:

$$\mathcal{S} = \frac{1}{N} \sum_{n=1}^{N} s_n \tag{20}$$

to elect the trivial solution, where each time series form a single cluster, as the best one. Clearly, the best partition is expected to be pointed out when the (20) is maximized, which implies the minimization of the intra-group distance the maximization of the intergroup distance.

4. Application to financial time series

To show the effectiveness of the proposed clustering approach, in what follows we provide an application to stock market data. Financial market data represent a clear example of the possible application since the empirical densities of the financial time series are proven to be non-Gaussian, asymmetric and heavy tailed [47].

Many are financial applications that take advantage from modeling stock returns by means of the Exponential Power Distribution as well as its skew extensions.

For example, it is found especially useful for modeling the volatility of portfolios such as SP500 and NASDAQ, where the ex post innovations from estimated GARCH models (even with a leverage effect) are not normally distributed with asymmetry and heavy tails (e.g. [38,48,49]), while other authors (e.g. [37,40]) showed its usefulness for modeling market risk.

For the experiment with real data, in this paper we select the stock prices included in the FTSE100 Index, that is a stock market index that contains the most important companies quoted at the London Stock Exchange (LSE) in terms of market capitalization. More in details, we downloaded the last 10 years of daily observations only for the 25 stocks quoted from the 1^{th} January 2010 and, therefore, that do not include any missing value (Fig. 3).

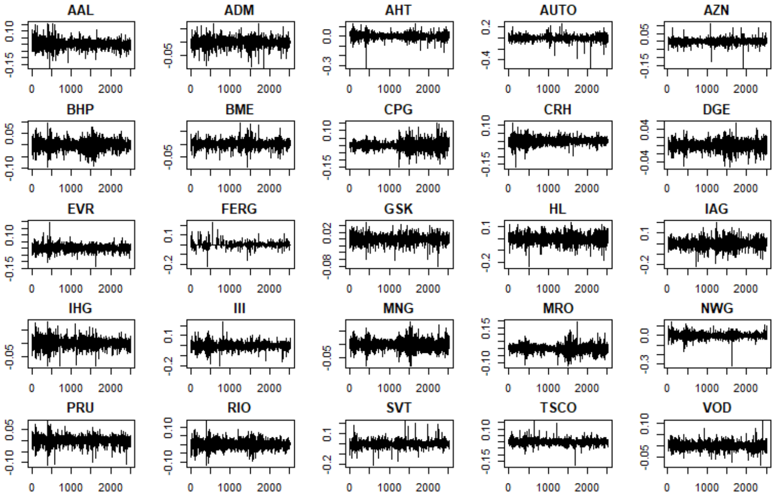

Figure 3. Dataset containing the time series of stock returns

In order to empirical show the aforementioned stock returns characteristics (i.e. heavy tails and skewness) in the Fig. 4 are showed the empirical densities for all the considered stock returns.

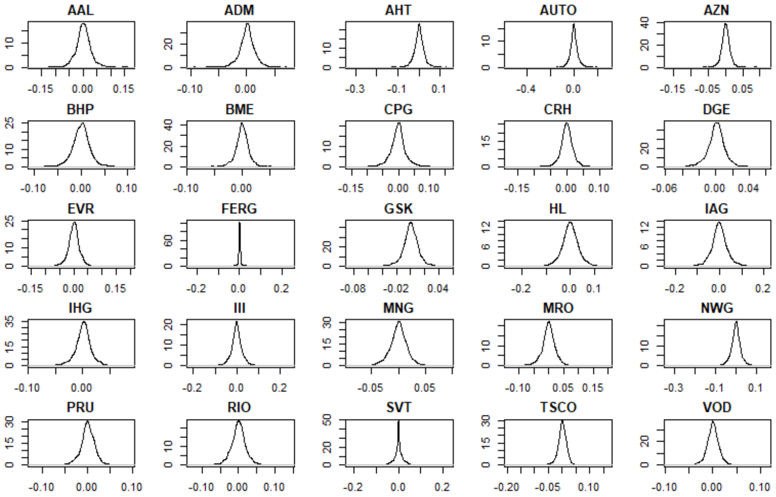

Figure 4. Dataset containing the empirical densities of stock returns

From Fig. 4 it is possible to note that the time series show very different distributions. For example, the stock SVT exhibits a very peaked distribution, similar to a Laplace one, with a low degree of skewness, while, on the other side, the stock AAL shows a peaked distribution with much higher degree of skewness.
Therefore, from these simple considerations appear clearly the need for the specification of a very flexible distribution able to accurately capture these diversities.
As previously described, the first step of the proposed clustering procedure involves the estimation of the Skew Exponential Power Distribution parameters (i.e. location, scale, skewness and shape) by means of maximum likelihood method (see Tab. 1).

	Location	Scale	Skewness	Shape
AAL	0.000738	0.027954	1.139503	1.018270
ADM	0.000249	0.014571	1.068612	0.936824
AHT	0.000189	0.025201	1.008579	0.955910
AUTO	-0.000388	0.038164	0.836474	0.992476
AZN	0.000480	0.013382	1.037042	0.995797
BHP	0.000067	0.019585	1.218680	0.940325
BME	0.000547	0.012799	1.067420	0.989826
CPG	-0.000750	0.026240	1.015652	0.980069
CRH	0.000231	0.019413	1.140867	0.992628
DGE	0.000288	0.010269	1.999935	0.999934
EVR	0.000448	0.020609	1.158119	0.967239
FERG	0.002564	0.006946	0.817407	1.347677
GSK	0.000250	0.011136	1.999924	0.999944
HL	-0.000359	0.034408	1.312343	0.981018
IAG	-0.000616	0.036465	1.199736	1.036676
IHG	0.000652	0.015839	1.041454	0.938918
III	0.000218	0.028553	0.917216	1.005544
MNG	-0.000188	0.016659	1.210081	0.962476
MRO	-0.000050	0.024692	1.055687	0.971691
NWG	-0.000145	0.025434	1.033912	0.953965
PRU	0.000318	0.017418	1.111372	0.951754
RIO	0.000213	0.021568	1.170911	0.987042
SVT	0.002635	0.021796	0.422242	1.131838
TSCO	0.000820	0.017106	1.108148	0.985365
VOD	0.000143	0.013780	1.151410	1.007114

Table 1: Maximum likelihood parameter estimates of a Skew Exponential Power Distribution for the considered stock returns

The estimated parameters clearly highlight the differences among the stock returns empirical densities. None of the considered stock returns is normally distributed, since all the estimated shape parameters are $p \neq 2$, showing a skew distribution with very heavy tails.

As usual, the second step of the clustering procedure involves the decision of the number of clusters C. As specified in the section 3 of the paper, we take advantage of the Silhouette Width Cirterion (SWC), whose results are shown in the Fig. 5.

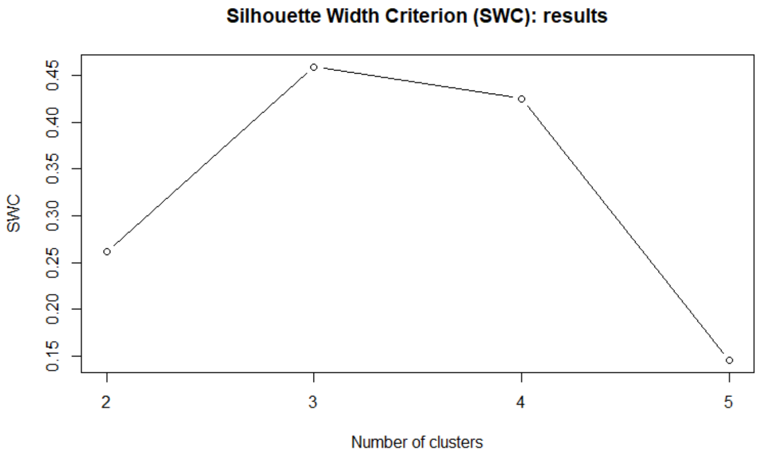

Figure 5. Silhouette Width Criterion for different number of clusters C

The higher value of $\mathcal{S}$ are obtained with $C = 3$ clusters and, then, its value dramatically decreases with an increasing number of clusters. Therefore we choose $C = 3$.
Then, from the Distribution-based Entropy Weighting k-Means (DEWKM) algorithm we obtain the hard partition showed in the Tab. 2.

Stock	Group
AAL	2
ADM	1
AHT	1
AUTO	3
AZN	1
BHP	2
BME	1
CPG	1
CRH	2
DGE	2
EVR	2
FERG	3
GSK	2
HL	2
IAG	2
IHG	1
III	3
MNG	2
MRO	1
NWG	1
PRU	2
RIO	2
SVT	3
TSCO	2
VOD	2

Table 2: Entropy weighting clustering: results

The third cluster $c = 3$ is the less numerous with only the 16% of the assets (AUTO, FERG, III and SVT) placed within. By looking at the parameter estimates in Tab. 1, it appears clear that the stock within the group $c = 3$ are those showing an heavy tailed distribution with a low degree of skewness.
The second clusters $c = 2$, on the other side, is the most numerous since a proportion of 52% of stock is included within. In this case, looking at Tab. 1, we can conclude the in the second group are placed the stock with the highest degree of skeweness. Indeed, for example, the stock DGE, GSK and HL are those with the three highest value of skeweness (close to 2 for both DGE and GSK and 1.31 for HL).
In the end, the residual cluster $c = 1$ contains all the other 32% of stocks.
An important feature of the proposed clustering approach is that, in any of the C clusters, we obtain the relative importance of any parameter in determining the cluster. The results for the proposed financial application are shown in the Tab. 3.

	Location (w_1)	Scale (w_2)	Skewness (w_3)	Shape (w_4)
Group 1	0.250450	0.250385	0.249591	0.249572
Group 2	0.303423	0.303187	0.092584	0.300805
Group 3	0.264247	0.264112	0.228034	0.243606

Table 3: Estimated weights from the entropy weighting clustering model

At the end, once can be interested in the possible usage of these groups in the real world. An immediate example for any clustering approach is, once it is applied to financial data, represented by the portfolio selection. In what follows we provide the results about the financial performance of the portfolios build by means of the proposed clustering model.

4.1. Portfolio analysis

The clusters obtained in the previous Section by the proposed approach can be seen as possible portfolios from an asset allocation perspective.

Financial literature provided various approaches to portfolio selection. Nevertheless, [50] showed that empirically the naive or *Talmudic*[1] diversification rule returns the highest performances in out-of-sample analysis with respect most of alternatives.

This result highlights the relevance of the estimation error in portfolio selection, coming from the fact that the investors estimates unknown quantities. Indeed, the equally weighted strategy ($1/N$) is the only diversification strategy with zero estimation error, since nothing is estimated.

Because estimation of expected returns is much more challenging than covariance matrix estimation [51], for reducing estimation error several scholars (e.g. [52–54]) focused on the following portfolio problem: find an asset allocation that minimize portfolio variance. This is called Global Minimum Variance (GMV) strategy. Assuming to have N time series, the portfolio problem can we written as:

$$\min_{w} w'\Sigma w \tag{21}$$

under the constraint:

$$\sum_{n=1}^{N} w_m = 1 \tag{22}$$

The optimal global minimum variance weights w, as solution of the minimization problem (21), are:

$$w = \frac{\Sigma^{-1} 1_N}{1'_N \Sigma^{-1} 1_N} \tag{23}$$

where $1_N = (1, 1, \ldots, 1)$ and $1'_N \Sigma^{-1} 1_N$ is the sum of all elements within the vector $\Sigma^{-1} 1_N$. By replacing Σ^{-1} with $\hat{\Sigma}^{-1}$ we get the optimal *estimated* GMV portfolio weights that we call $\hat{w}$.

In what follows we consider each cluster as a possible set of stock and we use the Global Minimum Variance approach to build C different portfolios.

In order to evaluate the out-of-sample performances we follow the empirical procedure of [50], based on a "rolling-sample" approach. Specifically, given a T daily observation of the securities returns, we choose a rebalancing window of length $M = 180, 360$. Then, in each day t, starting from $t = M + 1$, we use the M observations to form the portfolio according to the naive diversification strategy.

This process is repeated $T - M$ times, by adding the return for the next period in the dataset and dropping the earliest one until the end of the dataset is reached. The outcome is, for each portfolio, a time series of $T - M$ daily out-of-sample portfolio returns.

Given the time series of monthly out-of-sample returns, we compute the out-of-sample Sharpe ratio of the portfolio c, $\hat{\theta}_c$, defined as the sample mean of out-of-sample portfolio returns divided by their standard deviation:

[1] The *Talmud* is the central text of Rabbinic Judaism that provides the following investment advice: "let every man divide his money into three parts, and invest a third in land, a third in business and a third let him keep by him in reserve".

$$\hat{\theta}_c = \frac{\hat{\mu}_c}{\hat{\sigma}_c} \tag{24}$$

where $\hat{\mu}_c$ is the average of the $T - M$ out of sample returns for the c portfolio and $\hat{\sigma}_c$ its standard deviation.
The Fig. 6 shows the over the time returns of the $C = 3$ different portfolios formed with the proposed clustering approach, assuming a rolling-window $M = 180$.

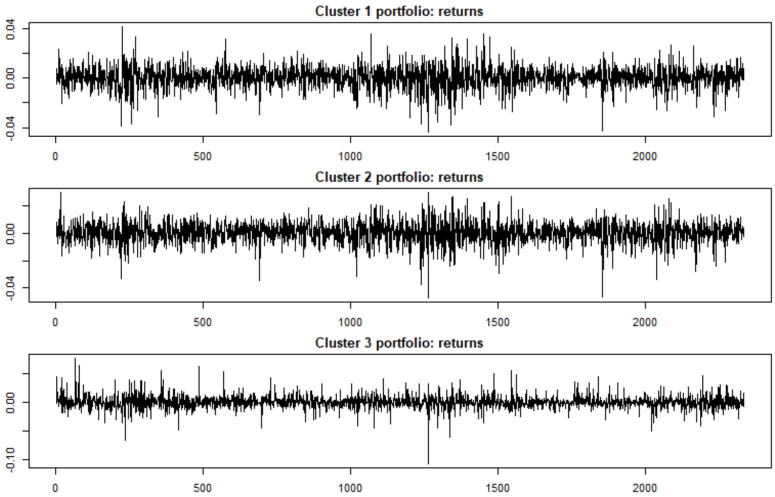

Figure 6. Returns of the $C = 3$ clustered portfolios (with $M = 180$)

Assuming a window of $M = 180$ means that the portfolio is re-balanced every six months. The performance comparisons among the different clustered portfolios are reported in the Tab. 4.

	Cluster 1	Cluster 2	Cluster 3
Sharpe Ratio	0.054163	0.022751	0.038398
Mean	0.000452	0.000185	0.000491
St. Dev.	0.008353	0.008149	0.012774
Max	0.035995	0.031026	0.111629
Min	-0.042813	-0.047531	-0.124435

Table 4: Sharpe ratios of the $C = 3$ clustered portfolios (with $M = 180$)

The highest Sharpe ratio is achieved with the first cluster $c = 1$ of stocks, that have been used to build a portfolio diversified according to the GMV rule. More in details, the $c = 1$ portfolio achieves 314 basis point higher trade-off- between risk and return with respect the $c = 2$ portfolio, that contains the greatest number of assets.
On the other side, the difference in terms of Sharpe Ratio between the portfolios $c = 1$ and $c = 3$ is of 157 basis points. It is interesting to note that the portfolios containing the lower amount of assets are those with the highest performances.
Then the analysis has been repeated assuming an annual portfolio re-balancing, where $M = 360$ (see Fig. 7).

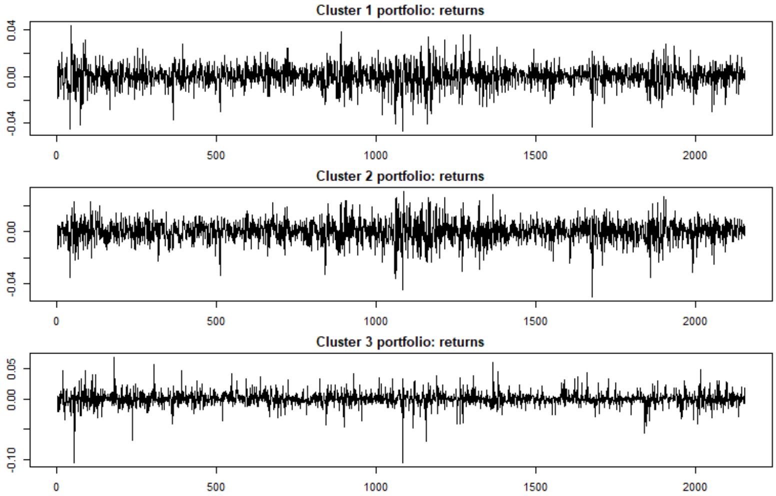

Figure 7. Returns of the $C = 3$ clustered portfolios (with $M = 360$)

In this second scenario, the Sharpe ratios dramatically reduces: the first clustered portfolio decreases its performance of 156 basis points, the portfolio $c = 2$ of 54 while the last $c = 3$ of 115 basis points (see Tab. 5).

	Cluster 1	Cluster 2	Cluster 3
Sharpe Ratio	0.038523	0.017291	0.026889
Mean	0.000342	0.000148	0.000358
St. Dev.	0.008874	0.008586	0.013309
Max	0.038319	0.031303	0.097478
Min	-0.044596	-0.049410	-0.150932

Table 5: Sharpe ratios of the $C = 3$ clustered portfolios (with $M = 360$)

However these differences are given by the different re-balancing timing. Nevertheless, the global ranking of the build clustered portfolios remain the same meaning that, despite the out-of-sample performances change with a different re-balancing, their overall performances are not affected by this choice.

In conclusion, we showed how this clustering algorithm can be used to form portfolios of stocks. Obviously, we only supposed the Global Minimum Variance (GMV) strategy here. However, it is possible also to assume different diversification rules (e.g. naive, mean-variance, etc.) either for all the clusters or assuming a different diversification strategy for each cluster.

5. Conclusions

In this paper, we propose a new model-based clustering approach for classifying skewed time series, by means of an entropy weighting clustering algorithm.

Clustering, by grouping objects that have maximum similarity with other objects within the group, and minimum similarity with objects in other groups, is a useful approach for exploratory data analysis as it identifies structure(s) in an unlabeled dataset by objectively organizing data into similar groups.

Indeed, a possible application of financial time series clustering concerns the asset allocation, where groups of similar stocks could be seen as portfolios of asset that shares similar characteristics.

There are many on-going research projects aimed to improve the existing techniques. This article proposes a model clustering model that refers to data based on a very important family of Asymmetric functions: the Asymmetric Exponential Power Distribution (SEPD) also called in literature as the Skewed Generalized Error Distribution (SGED) .

This distribution is very useful for classifying time series in presence of fat-tailed and asymmetric time series.
Three important particular cases of SEPD are analyzed in the paper and exactly they are the Gaussian, the Laplace and the Asymmetric Laplace.
The clustering algorithm, which represents the innovative aspect of this paper, uses the moments estimated by the introduced exponential power distribution Skew to form the clusters.
The criterion is that time series with similar moment estimates are placed in the same group. Therefore, with a k-means clustering algorithm, the measure of dissimilarity is determined on the basis of these estimates. In this paper we therefore propose to combine all the information in an optimal way to form clusters.
The approach we devised to optimally weight the different data characteristics is represented by the entropy weighting k-mean algorithm. The algorithm is based on the k-means approach, where the weights are inserted and positioned in the distance function. The idea behind the new method proposed in this paper is that weights are a measure of the relative importance of each variable with respect to the belonging of the observations to that particular cluster
. In details, we formalized the weighted K-Means algorithm (WKM) considering the family of Asymmetric Distributions: SEPD.
Finally, to demonstrate the effectiveness of the proposed clustering approach, the work ends with an application to stock market data. Financial market data lend themselves well to adhering to our methodological proposal. In fact, the empirical densities of the financial time series proved to be non-Gaussian, asymmetric and heavy.
Ours wants to be a fairly innovative research address and certainly many can be there financial applications that benefit from modeling equity returns via exponential power distribution and its skew extensions.
Indeed a final important result allows us to conclude that the new clustering algorithm we described in the paper can be used to form equity portfolios. However, we have developed the Global Minimum Variance (GMV) strategy but as a starting point for future research it is also possible to hypothesize different diversification rules (e.g. mean-variance, etc.) both for all clusters and considering a different diversification strategy for each cluster. analyzed.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mantegna, R.N. Hierarchical structure in financial markets. *The European Physical Journal B-Condensed Matter and Complex Systems* **1999**, *11*, 193–197.
2. Tola, V.; Lillo, F.; Gallegati, M.; Mantegna, R.N. Cluster analysis for portfolio optimization. *Journal of Economic Dynamics and Control* **2008**, *32*, 235–258.
3. Iorio, C.; Frasso, G.; D'Ambrosio, A.; Siciliano, R. A P-spline based clustering approach for portfolio selection. *Expert Systems with Applications* **2018**, *95*, 88–103.
4. Liao, T.W. Clustering of time series data—a survey. *Pattern recognition* **2005**, *38*, 1857–1874.
5. D'Urso, P. Dissimilarity measures for time trajectories. *Statistical Methods & Applications* **2000**, *1*, 53–83.
6. D'urso, P. Fuzzy C-means clustering models for multivariate time-varying data: different approaches. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **2004**, *12*, 287–326.
7. Coppi, R.; D'Urso, P. Fuzzy unsupervised classification of multivariate time trajectories with the Shannon entropy regularization. *Computational statistics & data analysis* **2006**, *50*, 1452–1477.
8. Coppi, R.; D'Urso, P.; Giordani, P. A fuzzy clustering model for multivariate spatial time series. *Journal of Classification* **2010**, *27*, 54–88.
9. D'Urso, P.; De Giovanni, L.; Massari, R. Robust fuzzy clustering of multivariate time trajectories. *International Journal of Approximate Reasoning* **2018**, *99*, 12–38.

doi:10.20944/preprints202104.0083.v1

10. Caiado, J.; Crato, N.; Peña, D. Comparison of times series with unequal length in the frequency domain. *Communications in Statistics—Simulation and Computation®* **2009**, *38*, 527–540.
11. Alonso, A.M.; Maharaj, E.A. Comparison of time series using subsampling. *Computational statistics & data analysis* **2006**, *50*, 2589–2599.
12. D'Urso, P.; Maharaj, E.A. Autocorrelation-based fuzzy clustering of time series. *Fuzzy Sets and Systems* **2009**, *160*, 3565–3589.
13. Caiado, J.; Crato, N.; Peña, D. A periodogram-based metric for time series classification. *Computational Statistics & Data Analysis* **2006**, *50*, 2668–2684.
14. D'Urso, P.; Maharaj, E.A. Wavelets-based clustering of multivariate time series. *Fuzzy Sets and Systems* **2012**, *193*, 33–61.
15. Maharaj, E.A.; D'Urso, P.; Galagedera, D.U. Wavelet-based fuzzy clustering of time series. *Journal of classification* **2010**, *27*, 231–275.
16. Maharaj, E.A.; D'Urso, P. Fuzzy clustering of time series in the frequency domain. *Information Sciences* **2011**, *181*, 1187–1211.
17. D'Urso, P.; De Giovanni, L.; Massari, R.; D'Ecclesia, R.L.; Maharaj, E.A. Cepstral-based clustering of financial time series. *Expert Systems with Applications* **2020**, *161*, 113705.
18. Piccolo, D. A distance measure for classifying ARIMA models. *Journal of Time Series Analysis* **1990**, *11*, 153–164.
19. Otranto, E. Clustering heteroskedastic time series by model-based procedures. *Computational Statistics & Data Analysis* **2008**, *52*, 4685–4698.
20. D'Urso, P.; De Giovanni, L.; Massari, R. GARCH-based robust clustering of time series. *Fuzzy Sets and Systems* **2016**, *305*, 1–28.
21. Iorio, C.; Frasso, G.; D'Ambrosio, A.; Siciliano, R. Parsimonious time series clustering using p-splines. *Expert Systems with Applications* **2016**, *52*, 26–38.
22. Maharaj, E.A.; Alonso, A.M.; D'Urso, P. Clustering seasonal time series using extreme value analysis: An application to Spanish temperature time series. *Communications in Statistics: Case Studies, Data Analysis and Applications* **2015**, *1*, 175–191.
23. D'Urso, P.; Maharaj, E.A.; Alonso, A.M. Fuzzy clustering of time series using extremes. *Fuzzy Sets and Systems* **2017**, *318*, 56–79.
24. Corduas, M.; Piccolo, D. Time series clustering and classification by the autoregressive metric. *Computational statistics & data analysis* **2008**, *52*, 1860–1872.
25. D'Urso, P.; Cappelli, C.; Di Lallo, D.; Massari, R. Clustering of financial time series. *Physica A: Statistical Mechanics and its Applications* **2013**, *392*, 2114–2129.
26. Cerqueti, Roy, G.M.; Mattera, R. Model-based fuzzy time series clustering of conditional higher moments. *International Journal of Approximate Reasoning* **2021**. Forthcoming.
27. Azzalini, A. Further results on a class of distributions which includes the normal ones. *Statistica* **1986**, *46*, 199–208.
28. Fernandez, C.; Osiewalski, J.; Steel, M.F. Modeling and inference with v-spherical distributions. *Journal of the American Statistical Association* **1995**, *90*, 1331–1340.
29. Fernández, C.; Steel, M.F. On Bayesian modeling of fat tails and skewness. *Journal of the american statistical association* **1998**, *93*, 359–371.
30. Theodossiou, P. Skewed generalized error distribution of financial assets and option pricing. *Multinational Finance Journal* **2015**, *19*, 223–266.
31. Komunjer, I. Asymmetric power distribution: Theory and applications to risk measurement. *Journal of applied econometrics* **2007**, *22*, 891–921.
32. Hsieh, D.A. Modeling heteroscedasticity in daily foreign-exchange rates. *Journal of Business & Economic Statistics* **1989**, *7*, 307–317.
33. Nelson, D.B. Conditional heteroskedasticity in asset returns: A new approach. *Econometrica: Journal of the Econometric Society* **1991**, pp. 347–370.
34. Duan, J.C. Conditionally fat-tailed distributions and the volatility smile in options. *Rotman School of Management, University of Toronto, Working Paper* **1999**.
35. Ayebo, A.; Kozubowski, T.J. An asymmetric generalization of Gaussian and Laplace laws. *Journal of Probability and Statistical Science* **2003**, *1*, 187–210.
36. Christoffersen, P.; Dorion, C.; Jacobs, K.; Wang, Y. Volatility components, affine restrictions, and nonnormal innovations. *Journal of Business & Economic Statistics* **2010**, *28*, 483–502.
37. Cerqueti, R.; Giacalone, M.; Panarello, D. A Generalized Error Distribution Copula-based method for portfolios risk assessment. *Physica A: Statistical Mechanics and its Applications* **2019**, *524*, 687–695.

38. Cerqueti, R.; Giacalone, M.; Mattera, R. Skewed non-Gaussian GARCH models for cryptocurrencies volatility modelling. *Information Sciences* **2020**.
39. DiCiccio, T.J.; Monti, A.C. Inferential aspects of the skew exponential power distribution. *Journal of the American Statistical Association* **2004**, *99*, 439–450.
40. Zhu, D.; Zinde-Walsh, V. Properties and estimation of asymmetric exponential power distribution. *Journal of econometrics* **2009**, *148*, 86–99.
41. Kotz, S.; Kozubowski, T.; Podgorski, K. *The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance*; Springer Science & Business Media, 2012.
42. Giacalone, M.; Panarello, D.; Mattera, R. Multicollinearity in regression: an efficiency comparison between L p-norm and least squares estimators. *Quality & Quantity* **2018**, *52*, 1831–1859.
43. Giacalone, M. A combined method based on kurtosis indexes for estimating p in non-linear Lp-norm regression. *Sustainable Futures* **2020**, *2*, 100008.
44. Trucíos, C. Forecasting Bitcoin risk measures: A robust approach. *International Journal of Forecasting* **2019**, *35*, 836–847.
45. Huang, J.Z.; Ng, M.K.; Rong, H.; Li, Z. Automated variable weighting in k-means type clustering. *IEEE transactions on pattern analysis and machine intelligence* **2005**, *27*, 657–668.
46. Jing, L.; Ng, M.K.; Huang, J.Z. An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data. *IEEE Transactions on knowledge and data engineering* **2007**, *19*, 1026–1041.
47. Cont, R. Empirical properties of asset returns: stylized facts and statistical issues **2001**.
48. Nelson, D.B. Conditional heteroskedasticity in asset returns: A new approach. *Econometrica: Journal of the Econometric Society* **1991**, pp. 347–370.
49. Wilhelmsson, A. GARCH forecasting performance under different distribution assumptions. *Journal of Forecasting* **2006**, *25*, 561–578.
50. De Miguel, V.; Garlappi, L.; Uppal, R. Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy? *The Review of Financial Studies* **2007**, *22*, 1915–1953.
51. Merton, R.C. On estimating the expected return on the market: An exploratory investigation. *Journal of financial economics* **1980**, *8*, 323–361.
52. Ledoit, O.; Wolf, M. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance* **2003**, *10*, 603–621.
53. Ledoit, O.; Wolf, M. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis* **2004**, *88*, 365–411.
54. Kourtis, A.; Dotsis, G.; Markellos, R.N. Parameter uncertainty in portfolio selection: Shrinking the inverse covariance matrix. *Journal of Banking & Finance* **2012**, *36*, 2522–2531.

doi:10.20944/preprints202104.0083.v1