Noname manuscript No.
(will be inserted by the editor)

A tail dependence-based dissimilarity measure for financial time series clustering

Giovanni De Luca · Paola Zuccolotto

Received:

Abstract In this paper we propose a clustering procedure aimed at grouping time series with an association between extremely low values, measured by the lower tail dependence coefficient. Firstly, we estimate the coefficient using an Archimedean copula function. Then, we propose a dissimilarity measure based on tail dependence coefficients and a two-step procedure to be used with clustering algorithms which require that the objects we want to cluster have a geometric interpretation. We show how the results of the clustering applied to financial returns could be used to construct defensive portfolios reducing the effect of a simultaneous financial crisis.

Keywords Time series · Clustering · Tail dependence · Copula function

Mathematics Subject Classification (2000) 62H30 · 62M10

1 Introduction

The discrimination and clustering literature has a very long history. However, the issues related to the time series clustering procedures have arisen more recently.

There are various approaches for computing a distance or a dissimilarity between two time series. Historically, the first approach has been based on the simple comparison between observations or some statistics computed on the data. For instance, Bohte et al. (1980) used the cross-correlation to cluster time series into homogeneous groups.

G. De Luca (corresponding author)
University of Napoli Parthenope, via Medina, 40, 80133 Napoli, Italy
E-mail: giovanni.deluca@uniparthenope.it

P. Zuccolotto
University of Brescia, C.da S. Chiara, 50, 25122 Brescia, Italy
E-mail: zuk@eco.unibs.it

A second approach is based on the comparison between the underlying generating processes which have generated the time series. Piccolo (1990) proposed a distance measure for ARIMA models, based on the comparison between the parameters of the corresponding AR(∞) representations. Corduas and Piccolo (2008) derived the asymptotic distribution of the squared AR distance and implemented a hypothesis test to capture the independence between two time series revealed by a very small distance. The same approach has been followed by Otranto (2008) for financial time series clustering. Considering GARCH models as main statistical description of such time series, the author exploits the well known result that if a time series follows a GARCH model, then its squares can be represented by an ARMA model. The distance measure between the parameters of autoregressive models can then be utilized. The approach has been extended in various directions. Galeano and Peña (2000) considered the generalized distance between the autocorrelation functions of two time series while Caiado et al. (2006) introduced a metric based on the normalized periodogram. Recently, a fuzzy approach has been developed by D'Urso and Maharaj (2009).

An alternative method, based on the forecast densities, was discussed in Alonso et al. (2006). The motivation was that a clustering method should take into account the dynamic properties of a time series. A classification provided by statistical models built on the past observations might change in time. In an out-of-sample period, the evolution of the time series could imply a different clustering. For this reason, the authors proposed an algorithm based on forecast densities. Vilar et al. (2010) extended the procedure with nonparametric forecast densities for non-linear time series.

Finally, Kakizawa et al. (1998) and Taniguchi and Kakizawa (2000) considered dissimilarity measures in the frequency domain, Weng and Shen (2008) proposed a new approach for classifying multivariate time series using two-dimensional singular value decomposition, while Pattarin et al. (2004) implemented a clustering procedure using a robust evolutionary algorithm.

Actually, the existence of so many proposals to face the problem of defining a distance function points out that the main problem in clustering consists in identifying the most appropriate distance measure. So, the classification of time series into homogeneous groups involves the definition of a criterion of homogeneity.

In this paper, we propose to cluster time series of returns of financial assets in groups which are homogeneous in the sense that there is an association between extreme low values. The idea behind this way of defining homogeneity within groups is to cluster financial assets according to their extreme comovements. To realize this idea we resort to the concept of tail dependence and propose a dissimilarity measure based on tail dependence coefficients. For the estimation of tail dependence coefficients we suggest to employ copula functions. The results of the corresponding clustering method could then be used for a portfolio selection purpose, when the goal is to protect investments from the effects of a financial crisis.

The article is organized as follows. In Section 2 we define the concept of tail dependence, while a dissimilarity measure between time series based on tail dependence coefficients is introduced in Section 3. In Section 4 we propose a two-step clustering procedure to be used with clustering algorithms which require that the objects we want to cluster have a geometric interpretation and thus cannot be applied directly to arbitrary dissimilarities. Subsequently, we describe the results of some simulation studies in Section 5. An application to real data is presented and discussed in Section 6. Finally, Section 7 concludes.

2 Tail dependence

The analysis of association between extreme values of two variables has always interested researchers in many areas, e.g. epidemiology, hydrology, finance. In fact, an extreme event is usually of great interest, either for its positive effects or for its negative consequences. An association between extreme values strengthens this interest. The problem of quantifying this association is crucial. In the literature, a great deal of attention has been directed toward the bivariate extreme value theory, based on particular distributions aimed at capturing the tail behavior.

An alternative way of describing the association of extreme values consists in the consideration of a conditional probability, that is the probability that one variable assumes an extreme value, given that an extreme value has occurred to the other variable (see Cherubini et al. 2004). This probability is known as tail dependence where we must distinguish between upper and lower tail dependence, according to whether the extreme values are very high or very low, respectively.

Let Y_1 and Y_2 be two random variables and let $U_1 = F_1(Y_1)$ and $U_2 = F_2(Y_2)$ be their distribution functions, respectively. The lower tail dependence coefficient is defined as

$$\lambda_L = \lim_{v \to 0^+} P\left[U_1 \leq v | U_2 \leq v\right]$$

while the upper tail dependence coefficient is

$$\lambda_U = \lim_{v \to 1^-} P\left[U_1 > v | U_2 > v\right].$$

More clearly: if $\lambda_L > 0$, then the random variables Y_1 and Y_2 are called asymptotically dependent in the lower tail. Conversely, if $\lambda_L = 0$, then we speak of asymptotic independence. An analogous interpretation can be extended to λ_U.

The quantification of this probability only depends on the assumed bivariate distribution. In other words, the existence of tail dependence is a property of the bivariate distribution. The bivariate Gaussian distribution has been for a long time the benchmark in the multivariate statistical analysis. However, it does not admit tail dependence, that is for each couple of variables represented by the normal distribution there exists no association between extreme values.

In practice, the tail dependence coefficients have to be estimated from observed data. In general, a distributional assumption is needed, but this is usually hard, especially with with financial data, that are the focus of this paper. A very effective way is the use of copula functions, thanks to which tail dependence estimation is both simple and flexible. A two-dimensional copula function for two random variables Y_1 and Y_2 is defined as a function $C : [0,1]^2 \to [0,1]$ such that

$$F(y_1, y_2; \theta) = C(F_1(y_1; \vartheta_1), F_2(y_2; \vartheta_2); \tau)$$

for all y_1, y_2, where $F(y_1, y_2; \theta)$ is the joint distribution function of Y_1 and Y_2 (see Nelsen 2006) and $\theta = (\vartheta_1, \vartheta_2, \tau)$. The function C allows to model the joint density separating the marginal behavior from the dependence structure. As a result, given a random sample $(y_{11}, y_{21}), \ldots, (y_{1n}, y_{2n})$ drawn from (Y_1, Y_2), the parameters of the copula functions are usually estimated by maximum likelihood in a two-step procedure. First, the estimates $\hat{\vartheta}_j$ are obtained from $(y_{j1}, \ldots, y_{jn})$, $j = 1, 2$. Then, the estimated distribution functions $\hat{F}_1$ and $\hat{F}_2$ are plugged into the likelihood function of the copula, which is then maximized to estimate τ.

When the observed data are time series, the marginal models estimated in the first step have to take into account the possible autocorrelated and heteroskedastic nature of the data. In other words, after modelling the time series the estimated distribution functions to be used in the second step are computed on the *i.i.d.* residuals time series rather than on the original time series.

It is straightforward to show that the tail dependence coefficients can be expressed in terms of the copula function. In particular, the lower tail dependence coefficient, which will be hereafter the focus of the paper, is given by

$$\lambda_L = \lim_{v \to 0^+} \frac{C(v, v)}{v}. \tag{1}$$

The value of the limit depends on the specific copula adopted. There are some copula functions for which the above limit is zero. One of the most popular copulas, the Gaussian copula, does not admit any tail dependence, i.e. the tail events occur independently. On the other hand, the class of Archimedean copulas is quite attractive because it presents a rich variety of cases (a comprehensive review is in Joe 1997) modelling the dependence either in one of the tails or in both.

In the analysis of the relationship between financial returns, the tail dependence offers a very important point of view, because it allows to go beyond the association of the entire distribution which is usually described by the correlation coefficient. The upper tail dependence shows a concordance between very high returns which could be exploited in an aggressive investment strategy. On the other hand, the lower tail dependence indicates the risk of investing on assets for which very negative returns could occur simultaneously. A defensive investment strategy would suggest to avoid such a situation. Undoubtedly, the lower tail dependence is strictly linked to the diversification in a financial crisis period when the assets composing the portfolio should ensure a counterbalance in order to avoid that negative returns of some assets push down the value of the portfolio, thus generating a very relevant loss. To avoid such an event, the choice of the assets in the portfolio should take into account the mutual relationship between extreme negative returns. Assets with a low value of the lower tail dependence coefficient should be selected. With this purpose, De Luca et al. (2010) proposed an algorithmic asset selection procedure based on the concept of multivariate tail dependence.

In this paper we face this problem by means of a different approach, based on time series clustering according to a dissimilarity measure defined as a function of the lower tail dependence. The idea is to obtain clusters of assets with high tail dependence in the lower tail, thus characterized by a high probability of simultaneous extremely negative events. From a portfolio selection perspective, we should then avoid the construction of a portfolio containing assets belonging to the same cluster.

3 A dissimilarity measure between time series based on the tail dependence coefficient

Let $\{y_{it}\}_{t=1,\ldots,T}$ be a time series. We define a dissimilarity measure between two time series $\{y_{1t}\}$ and $\{y_{2t}\}$ based on their estimated tail dependence coefficient $\hat{\lambda}_L$, as the inverse of its natural logarithm:

$$\delta(\{y_{1t}\},\{y_{2t}\}) = -\log\left(\hat{\lambda}_L\right). \tag{2}$$

The reason for this choice is that (2) ranges from 0 to infinity, is small when the two time series are "near" to each other (i.e. when their tail dependence is high), and monotonically increases as the two time series become more "different" (i.e. when their tail dependence decreases). Of course, one can decide to use a different function in place of the logarithm. As a matter of fact, as we will see later on, the first step of clustering procedure proposed in this paper (Section 4) involves an optimally defined monotonic transformation of the dissimilarities. In other words, more than to their absolute value, we are interested to the rank information they contain, hence the specific function used in (2) to transform the estimated tail dependence coefficients into dissimilarities should not be a crucial choice if we desire to use the clustering procedure of Section 4.

It is well known that the definition of a dissimilarity measure δ is less restrictive than that of a metric or a distance function. We usually assume that a dissimilarity function satisfies the following properties:

1. non-negativity: $\delta(\{y_{1t}\},\{y_{2t}\}) \geq 0$;
2. identity: $\delta(\{y_{1t}\},\{y_{1t}\}) = 0$;
3. symmetry: $\delta(\{y_{1t}\},\{y_{2t}\}) = \delta(\{y_{2t}\},\{y_{1t}\})$.

Note that the triangle inequality is not required for dissimilarities, in contrast to the case of distances. It is easy to see that properties 1 and 2 hold for expression (2). The symmetry (property 3) is ensured thanks to an analogous property characterizing tail dependence coefficients (e.g. Cherubini et al. 2004): $\lim_{v \to 0^+} P\left[U_1 \leq v | U_2 \leq v\right] = \lim_{v \to 0^+} P\left[U_2 \leq v | U_1 \leq v\right]$. In addition to properties 1-3, we require that the dissimilarity measure $\delta(\{y_{1t}\}, \{y_{2t}\})$ decreases in a monotone way as $\{y_{1t}\}$ and $\{y_{2t}\}$ are more and more similar, according to the idea of similarity one has adopted. Here, this requirement derives from the choice of the logarithmic function in (2), provided that the idea of similarity we are describing is that of lower tail dependence, that is, we consider that the two series $\{y_{it}\}$ and $\{y_{jt}\}$ are more similar to each other than the two series $\{y_{ht}\}$ and $\{y_{lt}\}$ if $\hat{\lambda}_L(\{y_{it}\}, \{y_{jt}\}) > \hat{\lambda}_L(\{y_{ht}\}, \{y_{lt}\})$.

If the dissimilarity measure has to be used for clustering purposes, we should remark that actually none of above mentioned properties is really essential for clustering and there are clustering methods that do not require any of them (Kaufman and Rousseeuw, 1990, p.16). On the other hand, in absence of the triangle inequality, the objects we want to cluster may have no geometric interpretation, for example as points in a high-dimensional Euclidean space. This could prevent us from using some kinds of clustering algorithms, such as for instance some partitioning methods which require as input some quantitative measurement vectors. Hence, since expression (2) defines a dissimilarity function only, it will require either to be handled with a proper clustering method or to be somehow transformed into a distance function, if we desire to use procedures that need the triangular inequality. In Section 4 we will propose a two-step clustering procedure for the latter case. Then, in section 5, we will present the results of a simulation study comparing the two strategies: (a) to apply a hierarchical clustering procedure directly to the dissimilarity measures defined in (2), and (b) to use the proposed two-step clustering procedure in order to apply a k-means algorithm requiring that the objects are represented as points in a high-dimensional Euclidean space.

4 The two-step clustering procedure

As pointed out before, there are many clustering methods which do not really require that the dissimilarities between objects are measured by a distance function. For example, hierarchical clustering algorithms can be applied also to a dissimilarity matrix. On the other hand, the absence of the triangle inequality could prevent us from using some kinds of clustering algorithms, such as, for example, the k-means algorithm, which requires as its input the observed measurement vectors. In this Section we propose a clustering procedure in this context and we focus attention on the case when the objects we want to cluster are p time series of financial returns. The starting point is the tail dependence dissimilarity matrix $\mathit{\Delta} = (\delta_{ij})_{i,j=1,\ldots,p}$ with $\delta_{ij} = \delta(\{y_{it}\}, \{y_{jt}\})$ as defined in (2).

The idea is to conduct clustering in two steps: firstly, starting from the dissimilarity matrix Δ, we find an optimal representation of the p time series $\{y_{1t}\},\ldots,\{y_{pt}\}$ as p points $\mathbf{y}_1,\ldots\mathbf{y}_p$ in R^q; secondly, we perform a clustering algorithm using either the obtained geometric representation or the distance matrix D of the p points, according to the requirements of the clustering algorithm.

First step: In the first step we carry out a non-metric Multidimensional Scaling (MDS) in order to find an optimal representation $\mathbf{y}_1,\ldots\mathbf{y}_p$ of $\{y_{1t}\},\ldots,\{y_{pt}\}$ as points in R^q. The term optimal refers to the fact that the Euclidean distance matrix $D=(d_{ij})_{i,j=1,\ldots,p}$, with $d_{ij}=\|\mathbf{y}_i-\mathbf{y}_j\|$, of the obtained points has to fit as closely as possible the dissimilarity matrix Δ, in a sense that will be specified below. In other words, by means of MDS, each time series $\{y_{it}\}$ will be represented by a q-dimensional vector $\mathbf{y}_i$. For a given dimension q, the representation of the time series $\{y_{1t}\},\ldots,\{y_{pt}\}$ by the vectors $\mathbf{y}_1,\ldots\mathbf{y}_p$, is such that the interpoint distances d_{ij} in some sense "match" the dissimilarities δ_{ij}. Several approaches have been proposed for MDS (see Coombs, 1964 for a general discussion). In the present work, we use the algorithm developed by Shepard (Shepard 1962a, 1962b) and further refined by Kruskal (Kruskal 1964a, 1964b). Thereby we must consider the following monotonicity constraint:

$$\delta_{ij}<\delta_{hl}\Longleftrightarrow d_{ij}\leq d_{hl}\quad\forall i,j,h,l. \tag{3}$$

Of course this property may possibly not hold for any Euclidean point configuration, since in general the scatterplot of the couples (δ_{ij},d_{ij}) is not monotonic. Through a non-decreasing monotonic unknown function f we obtain a set of new values $\hat{d}_{ij}=f(\delta_{ij})$ such that the couples $(\delta_{ij},\hat{d}_{ij})$ exhibit a monotonic scatterplot and the so-called *stress function*

$$s=\sqrt{\frac{\sum_{i<j}\sum_{j=1}^{p}(f(\delta_{ij})-d_{ij})^2}{\sum_{i<j}\sum_{j=1}^{p}d_{ij}^2}},\quad 0\leq s\leq 1, \tag{4}$$

is minimized. The function f is usually a non-linear function obtained in a non-parametric way, but it can also be restricted to be linear, piece-wise linear, logarithmic, etc. For a given dimension q, the Shepard-Kruskal algorithm is iterative: once started from an initial point configuration $\mathbf{y}_1,\ldots\mathbf{y}_p$, the values $\hat{d}_{ij}$ are determined so as to minimize s subject to the monotonicity constraint (3), thus solving a problem equivalent to the so-called *monotone regression* (see for example Bartholomew, 1959). Then, an improved points configuration is obtained by considering s as a function, with fixed $\hat{d}_{ij}$, of the $q\times p$ coordinates of the p points, and so on. Thus, the function f corresponding to the optimal Euclidean configuration is itself an outcome of the algorithm. The expression of the stress function s which has to be minimized clearly highlights that with the Shepard-Kruskal algorithm we define an Euclidean configuration such that the Euclidean interpoint distances d_{ij} are as close as possible to a (possibly non-linear and non-parametric) weakly monotonic optimal transformation of the dissimilarities δ_{ij}.

The algorithm works for a given value of the dimension q, which has to be given in input. In the procedure presented in this paper we propose to start with the dimension $q = 2$ and then to repeat the analysis by increasing q until the minimum stress of the corresponding optimal configuration is lower than a given threshold $\bar{s}$. In other words, we choose the lowest dimension q for which we obtain $\min(s) \leq \bar{s}$. In fact, the purpose of MDS in this procedure is not, as usual, the graphical representation of the final configuration (which of course requires small values of q), because we are interested only in obtaining the optimal Euclidean configuration (or, equivalently, in computing its distance matrix D), to be used in the clustering step. Hence we are allowed to fix a very small threshold $\bar{s}$ (e.g., $\bar{s} = 0.005$), neglecting the fact that this usually necessitates a high-dimensional configuration.

Second step: In the second step the optimal Euclidean configuration $\mathbf{y}_1, \ldots \mathbf{y}_p$ or its distance matrix D can be used as an input for whatever clustering algorithm. In the following we will use the k-means algorithm.

5 Simulation studies on the clustering procedure

A set of simulation studies has been carried out in order to explore the performance of the proposed clustering procedure, with the comparison of two alternative strategies: (a) to apply a hierarchical clustering directly to the dissimilarity matrix Δ, and (b) to use the proposed two-step clustering procedure, i.e., to consider the points $\mathbf{y}_1, \ldots \mathbf{y}_p$ in R^q obtained from MDS and to cluster them with the classical k-means algorithm. The simulation scheme has been arranged in order to focus only on the clustering aspect, without considering the additional matters connected with the estimation of the tail dependence coefficients. To do that, in practice, at each iteration we generated a random similarity matrix that describes a clustering structure with k clusters of sizes $n_1, \ldots, n_k$. The latter values was obtained by sampling independently the number of clusters k and the number n_h of time series belonging to each cluster, $h = 1, \cdots, k$ from a discrete uniform distribution on $\{2, 3, \ldots, 10\}$. In other words we simulated a random symmetric matrix

$$\mathbf{TD} = \begin{bmatrix} \mathbf{TD}_1^W & \mathbf{TD}_{12}^B & \ldots & \mathbf{TD}_{1k}^B \\ \mathbf{TD}_{21}^B & \mathbf{TD}_2^W & \ldots & \mathbf{TD}_{2k}^B \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{TD}_{k1}^B & \mathbf{TD}_{k2}^B & \ldots & \mathbf{TD}_k^W \end{bmatrix},$$

where

$$\mathbf{TD}_h^W = \begin{bmatrix} 1 & \dot{\lambda}_{L,12}^W & \ldots & \dot{\lambda}_{L,1n_h}^W \\ \dot{\lambda}_{L,21}^W & 1 & \ldots & \dot{\lambda}_{L,2n_h}^W \\ \vdots & \vdots & \ddots & \vdots \\ \dot{\lambda}_{L,n_h1}^W & \dot{\lambda}_{L,n_h2}^W & \ldots & 1 \end{bmatrix} \quad \mathbf{TD}_{hl}^B = \begin{bmatrix} \dot{\lambda}_{L,11}^B & \dot{\lambda}_{L,12}^B & \cdots & \dot{\lambda}_{L,1n_l}^B \\ \dot{\lambda}_{L,21}^B & \dot{\lambda}_{L,22}^B & \cdots & \dot{\lambda}_{L,2n_l}^B \\ \vdots & \vdots & \ddots & \vdots \\ \dot{\lambda}_{L,n_h1}^B & \dot{\lambda}_{L,n_h2}^B & \cdots & \dot{\lambda}_{L,n_hn_l}^B \end{bmatrix}$$

denote, respectively, the matrix containing the similarities of the series belonging to the same cluster h and the matrix containing the similarities of the series belonging to cluster h and l. Each value in the matrix $\mathbf{TD}$ has been sampled from a normal distribution. More specifically, we defined two different distributions to model the similarity within and between clusters, $\dot{\lambda}^W_{L,ij} \sim N(\lambda^W_L, \sigma^2_W)$ and $\dot{\lambda}^B_{L,ij} \sim N(\lambda^B_L, \sigma^2_B)$. We decided to start our simulation experiment by setting $\lambda^W_L = 0.4$ and $\lambda^B_L = 0.3$. The reason for these choices about the distributions from which $\dot{\lambda}^W_{L,ij}$ and $\dot{\lambda}^B_{L,ij}$ are sampled is the following: we desire that the generated similarities resemble the behavior of tail dependence coefficients which are likely to occur in practice with financial time series (see for example Fortin and Kuzmics 2002; Patton 2006). We denote these "generated tail dependence coefficients" with $\dot{\lambda}_L$ in order to clearly distinguish them from the tail dependence coefficients estimated from data, $\hat{\lambda}_L$, which are the basis of the dissimilarity measure (2) and will be used in the case study of section 6. In addition, the generated values are sufficiently challenging for a clustering task, as the mean similarities within and between clusters are close each other.

We examined nine different parameters configurations, obtained by crossing three different values for the standard deviations of the distributions: $\sigma_W = 0.02, 0.05, 0.08$, $\sigma_B = 0.02, 0.05, 0.08$. Each experimental configuration affects the extent to which the distributions of the similarities within and between the clusters tend to overlap. With this choice of parameters the sampled values only occasionally lied outside the interval $[0, 1]$. When it was the case, a new value was drawn. For each of the nine combinations we built the dissimilarity matrix Δ as follows: the similarities $\dot{\lambda}^W_{L,ij}$ and $\dot{\lambda}^B_{L,ij}$, which, as stated before, are generated in a way that mimic the tail dependence coefficients, was transformed by using (2), thus obtaining the dissimilarity measures $\dot{\delta}^W_{L,ij} = -\log(\dot{\lambda}^W_{L,ij})$ and $\dot{\delta}^B_{L,ij} = -\log(\dot{\lambda}^B_{L,ij})$, respectively, to be put into Δ. Then we performed the clustering according to the two strategies described above. More specifically, strategy (a) consists of applying three hierarchical clustering algorithms (single, average, complete linkage) directly to the dissimilarity matrix Δ, while strategy (b) consists of employing the proposed two-step clustering procedure, setting $\bar{s} = 0.025$ in the first step and then using a k-means algorithm in the second step. We fixed a relatively high size of the threshold $\bar{s}$ in order to reduce computational times.

As a measure of goodness of clustering, we used the misclassification error. We denote the misclassification errors of the single, average, complete linkage and that of the k-means algorithm with $ME(s)$, $ME(a)$, $ME(c)$, $ME(k)$, respectively. Figures 1 through 3 display the results obtained with $R = 500$ iterations for each parameters configuration. Since our main focus is on the two-step clustering procedure, we report the boxplot of $ME(k)$ and, in order to compare the performance of the two strategies, those of the differences $ME(k) - ME(s)$, $ME(k) - ME(a)$, $ME(k) - ME(c)$. Negative values of these differences denote a better performance of the two-step clustering procedure applied with the k-means algorithm.

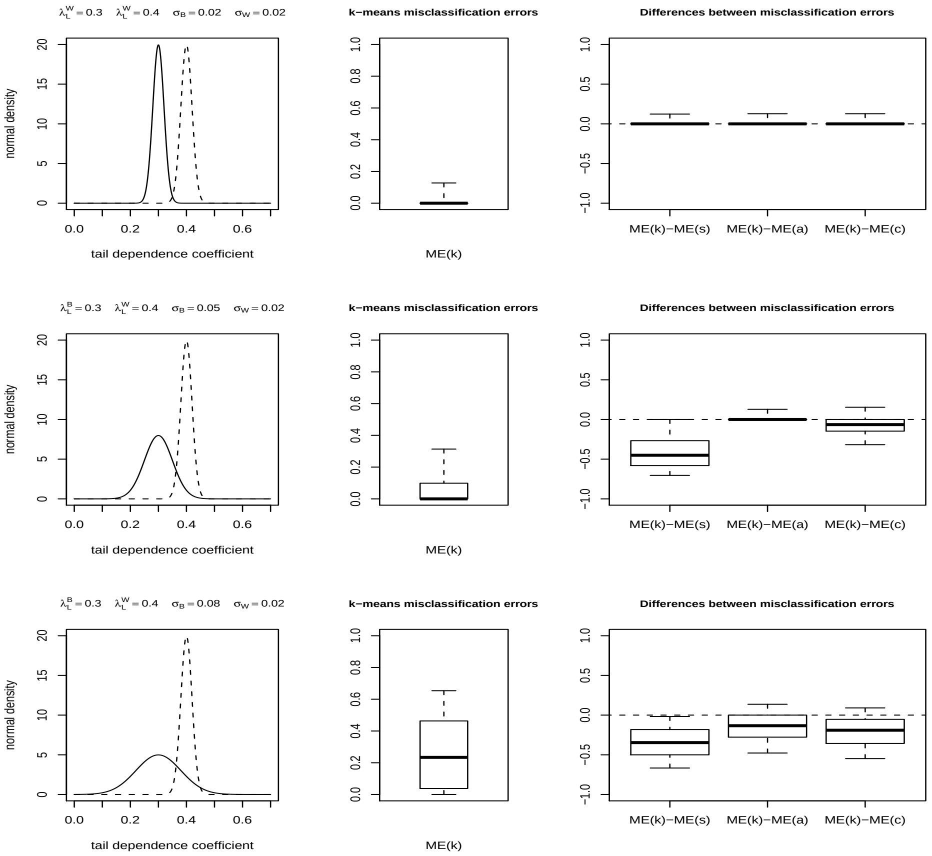

Fig. 1 Simulation results for the case $\sigma_B = 0.02, \sigma_W = 0.02$ (top), $\sigma_B = 0.05, \sigma_W = 0.02$ (middle), $\sigma_B = 0.08, \sigma_W = 0.02$ (bottom). Left: pdf of the distributions from which the similarities between and within the clusters have been sampled; Middle: boxplot of the misclassification errors obtained with the k-means algorithm; Right: boxplot of the differences between the misclassification errors obtained with the k-mean algorithm and with hierarchical clustering. The ends of the whiskers are the 5th and 95th percentile.

Looking at the boxplots of $ME(k)$, we notice that the misclassification error is low when the distributions of the similarities are well separated (for example, when $\sigma_W = \sigma_B = 0.02$). As expected, it tends to be large if overlap increases. Further simulation studies, performed with different values of λ_L^W and λ_L^B (not reported here) confirm these results. More specifically, we always obtain perfect clusterings with non-overlapping distributions. On the other hand, from the point of view of comparing the two strategies, the two-step clustering procedure applied with the k-means algorithm tends to outperform the others in many configurations, especially if compared to the single and complete linkage.

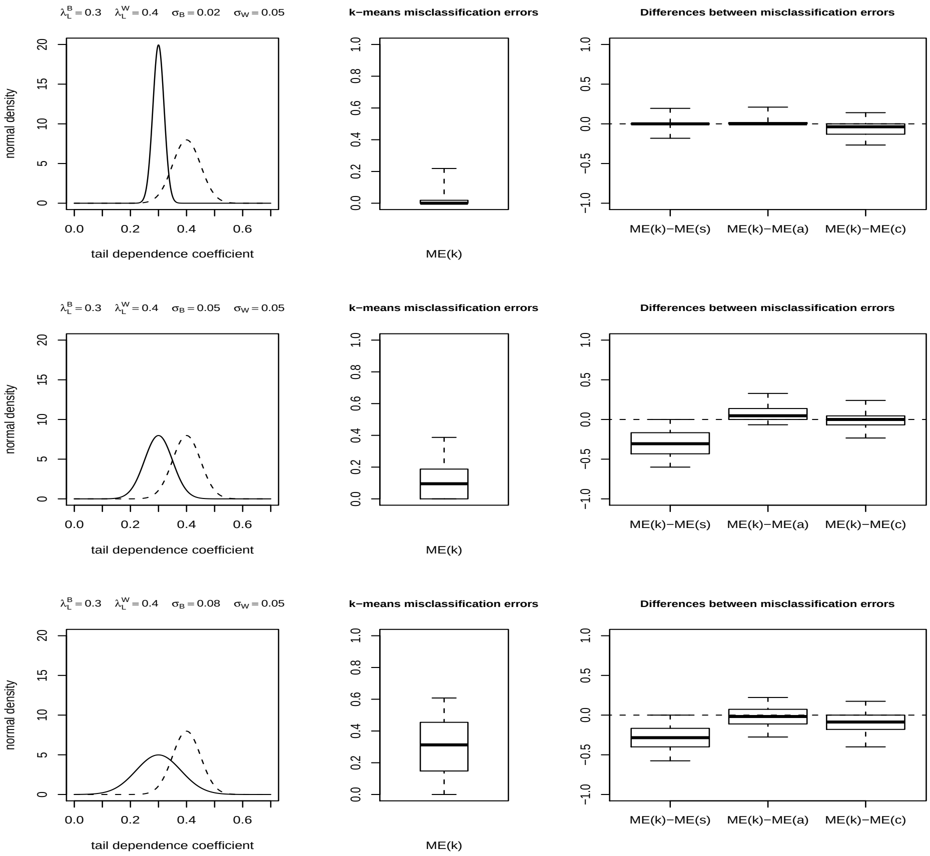

Fig. 2 Simulation results for the case $\sigma_B = 0.02, \sigma_W = 0.05$ (top), $\sigma_B = 0.05, \sigma_W = 0.05$ (middle), $\sigma_B = 0.08, \sigma_W = 0.05$ (bottom). Left: pdf of the distributions from which the similarities between and within the clusters have been sampled; Middle: boxplot of the misclassification errors obtained with the k-means algorithm; Right: boxplot of the differences between the misclassification errors obtained with the k-mean algorithm and with hierarchical clustering. The ends of the whiskers are the 5th and 95th percentile.

6 Application to real-case data

6.1 An application to geographic MSCI indices

In this section we analyze a dataset that consists of time series of the returns of 23 Morgan Stanley Capital International (MSCI) Developed Markets indices, $\{y_{1t}\}, \ldots, \{y_{pt}\}$, $p = 23$, $t = 1, \ldots, T$, designed to measure the equity market performance of developed markets[1], recorded daily from June 4, 2002 to June 10, 2010 (T=2093 observations; Source: MSCI Barra).

[1] The MSCI World Index consists of the following developed market country indices: Australia, Austria, Belgium, Canada, Denmark, Finland, France, Germany, Greece, Hong

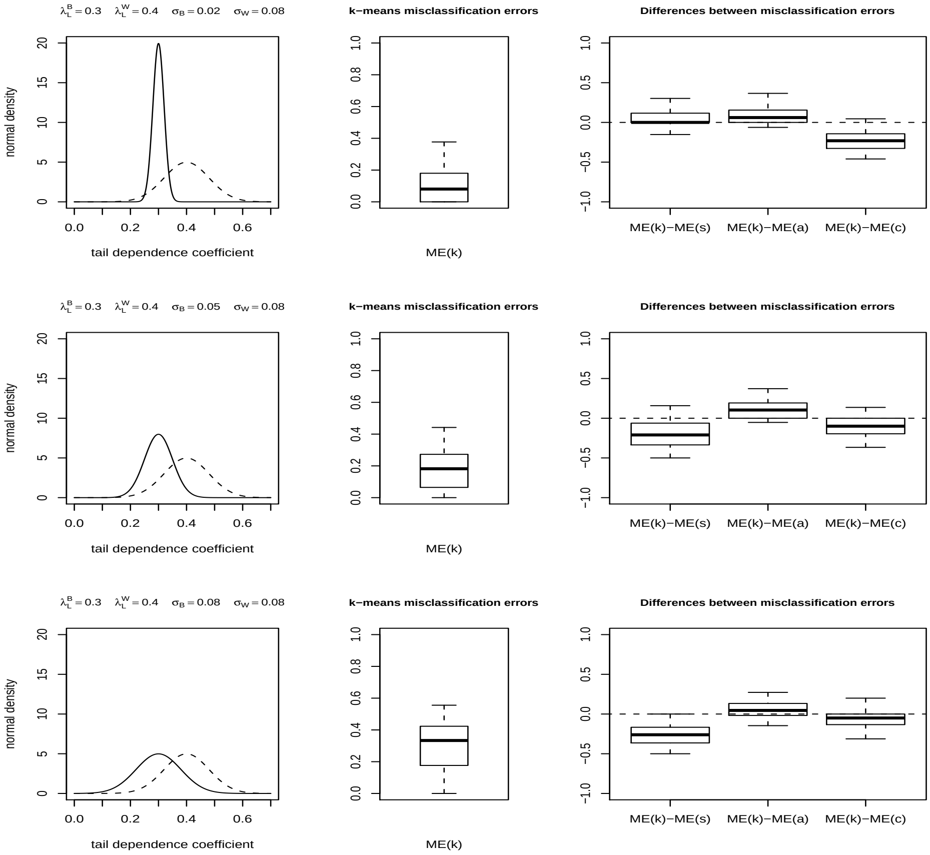

Fig. 3 Simulation results for the case $\sigma_B = 0.02, \sigma_W = 0.08$ (top), $\sigma_B = 0.05, \sigma_W = 0.08$ (middle), $\sigma_B = 0.08, \sigma_W = 0.08$ (bottom). Left: pdf of the distributions from which the similarities between and within the clusters have been sampled; Middle: boxplot of the misclassification errors obtained with the k-means algorithm; Right: boxplot of the differences between the misclassification errors obtained with the k-mean algorithm and with hierarchical clustering. The ends of the whiskers are the 5th and 95th percentile.

We preliminary applied a univariate Student-t AR-GARCH model to each time series of returns (see Bollerslev et al. (1992) for a comprehensive review on GARCH-type models) to remove autocorrelation and heteroskedasticity from the data and we computed the standardized residuals. For each couple of returns we estimated a bivariate Joe-Clayton copula function,

$$C(u_1, u_2) = 1 - \left\{1 - \left[(1-(1-u_1)^\kappa)^{-\theta} + (1-(1-u_2)^\kappa)^{-\theta} - 1\right]^{-1/\theta}\right\}^{1/\kappa} \tag{5}$$

Kong, Ireland, Italy, Japan, Netherlands, New Zealand, Norway, Portugal, Singapore, Spain, Sweden, Switzerland, the United Kingdom, and the United States.

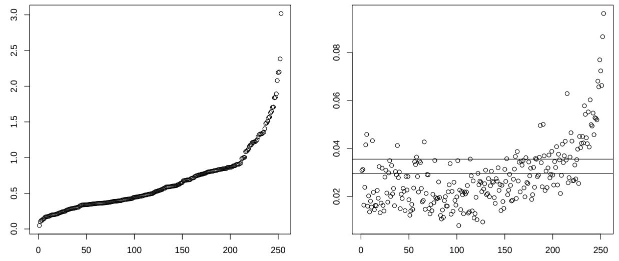

Fig. 4 Ranked values of the estimated θ's (left) and, accordingly, estimated Kolmogorov statistics (right). Horizontal lines denotes 5% and 1% critical values.

using the estimated distribution functions of the standardized residuals. The Joe-Clayton is a quite general copula (Joe 1997) belonging to the so-called Archimedean copulas, allowing for lower and upper tail dependence.[2] In particular, the lower tail dependence coefficient of the Joe-Clayton copula function is given by

$$\lambda_L = 2^{-1/\theta}, \tag{6}$$

and depends only on one of the two parameters κ, θ in (5).

Given the $p = 23$ time series, the total number of bivariate joint distribution to be estimated through the copula (5) is $p(p-1)/2 = 253$. The maximum likelihood method has been used to estimate all the bivariate copula functions. The left part of Figure 4 reports the 253 ranked estimates of θ. The middle 90% of the 253 values ranges from 0.2005 to 1.4600 and in all cases the parameter is significant ($\alpha = 0.05$). Moreover, in order to provide a measure of goodness-of-fit of the estimated copula functions (5), we considered a goodness-of-fit procedure based on the Rosenblatt transformation (see Breymann et al. 2003) of two random variables Y_1 and Y_2,

$$S(Y_1, Y_2) = \left[\Phi^{-1}(F_1(Y_1))\right]^2 + \left[\Phi^{-1}(C(F_2(Y_2)|F_1(Y_1)))\right]^2 \tag{7}$$

where Φ denotes the cdf of a normal distribution and $C(F_2(Y_2)|F_1(Y_1))$ is the conditional copula. If C is the true copula function of the variables Y_1 and Y_2, then the statistics (7) has a χ^2_2 distribution. This hypothesis has been tested with the Kolmogorov test. In our application we applied the transformation to the standardized residuals. The right part of Figure 4 reporting the corresponding Kolmogorov statistics with two horizontal lines at the 5% and 1% critical values ensures that only a few part of the estimated copula functions is not fully satisfactory.

After estimating the 253 lower tail dependence coefficients $\hat{\lambda}_L = 2^{-1/\hat{\theta}}$, we carried out MDS using the dissimilarity matrix $\Delta = (\delta_{ij})_{i,j=1,\ldots,p}$, where δ_{ij} is is the dissimilarity between time series $\{y_{it}\}$ and $\{y_{jt}\}$ computed through (2) as

[2] As pointed out by a referee, extreme events represented by an increase of the conditional volatility are not encountered.

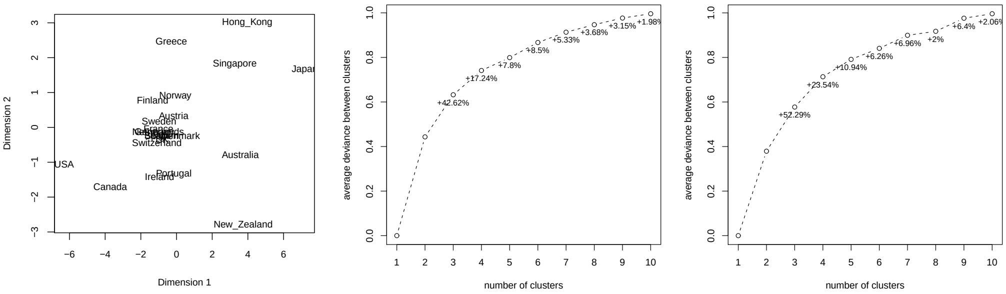

Fig. 5 Left: Two-dimensional MDS configuration. Middle: Average deviance between clusters vs number of clusters (k-means algorithm performed on data from June 4, 2002 to June 10, 2010). Right: Average deviance between clusters vs number of clusters (k-means algorithm on data from June 4, 2002 to May 31, 2008).

a function of the estimated tail dependence coefficient of the two series. We set $\bar{s} = 0.005$. The dimension allowing a final configuration with minimum stress $\min(s) \leq \bar{s}$ resulted in $q = 11$, with $\min(s) = 0.0032$. The left part of Figure 5 displays the graphical representation of the two-dimensional configuration, characterized by $\min(s) = 0.1020$.

In the second step, we performed a k-means clustering algorithm using the 11-dimensional MDS point configuration in the Euclidean space, $\mathbf{y}_1, \ldots \mathbf{y}_p$. The graph in the middle of Figure 5 shows the pattern of the average deviance between clusters as a function of the number of clusters k. The average deviance between clusters is given by

$$Dev_b = 1/q \sum_{h=1}^{k} \sum_{d=1}^{q} \left(m_{dh} - m_d\right)^2 n_h$$

with

$$m_{dh} = 1/n_h \sum_{i_h=1}^{n_h} y_{i_h d}$$

$$m_d = 1/p \sum_{i=1}^{p} y_{id}$$

where y_{id} denotes the d-th dimension measurement of a generic ith point $\mathbf{y}_i = (y_{i1}, \ldots, y_{id}, \ldots, y_{iq})$, the subscripts i_h and i refer to the objects within cluster h and in the whole set of points, respectively, and n_h is the number of points within cluster h. The same graph reports the increments in the average deviance between clusters obtained with k clusters, with respect to $k-1$ clusters. Looking at these increments, we find out that $k = 4$ or $k = 5$ can be considered good solutions. We chose the more parsimonious one in terms of number of clusters, that is $k = 4$ (Table 1).

Table 1 Cluster composition.

Cluster 1		**Cluster 2**	**Cluster 3**	**Cluster 4**
Austria	Italy	Hong Kong	Australia	Canada
Belgium	Netherland	Japan	New Zealand	USA
Denmark	Norway	Singapore		
Finland	Portugal			
France	Spain			
Germany	Sweden			
Greece	Switzerland			
Ireland	UK			

It results that the MSCI indices are clustered according to a geographic pattern, which means that the lower tail dependence tends to be higher within financial markets characterized by geographic proximity. More specifically, we clearly distinguish one European and one North-American cluster. Pacific countries tend to be divided in two separate clusters where New Zealand and Australia are joined together as well as Hong Kong, Japan and Singapore.

6.2 An example of using the clustering for portfolio selection purposes

Clustering for tail dependence coefficients has an interesting practical use when employed for a portfolio selection purpose. For example, we could decide to construct a portfolio composed by as many assets as given by the number k of clusters. The assets should be selected by imposing the restriction that each asset belongs to a different cluster. This strategy should protect the investments from parallel extreme losses during crisis periods, because the clustering solution is characterized by a moderate lower tail dependence between clusters. Hence we investigated the performance of this method during the period from June 2, 2008 to June 30, 2008, that is in the month following the beginning of the crisis. In order to draw the analysis from an out-of-sample perspective, we carried out again the above described clustering procedure, using the $p = 23$ time series $\{y_{1t}\}, \ldots, \{y_{pt}\}$, observed from June 4, 2002 to May 31, 2008. After choosing the solution with $k = 4$ clusters, we obtained exactly the same clusters of the analysis performed when using the complete sample (average deviances in the right part of Figure 5), which could mean that the financial crash of 2008 has not affected the lower tail dependence structure of the analyzed indices.

With the $k = 4$ clusters mentioned above, 192 different selections can be made according to the criterion of selecting one index from each cluster. Using the popular Markowitz portfolio selection procedure, the minimum variance portfolio is built for each one of these 192 selections, again using only data from June 4, 2002 to May 31, 2008. Figure 6 shows the returns of these 192 portfolios next month, from June 2, 2008 to June 30, 2008, compared to the returns of 500 minimum variance portfolios built with 4 randomly selected indices and with the returns of the naive minimum variance portfolio built

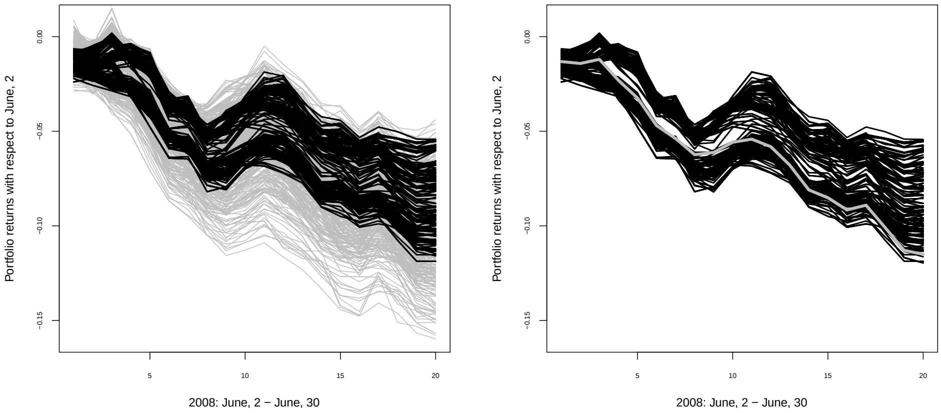

Fig. 6 Returns (black lines) of the 192 minimum variance portfolios composed by 4 indices, selected by imposing the restriction that each index belongs to a different cluster, compared to (left) returns of 500 minimum variance portfolios composed by 4 randomly selected indices (thin gray lines) and to (right) returns of the naive minimum variance portfolio composed by all the indices (bold gray line), June 2008.

using all the indices. The 192 portfolios outperform a part of the competitors during the whole considered period, while the returns of the naive portfolio are clearly below the 192 portfolios returns.

In order to select one portfolio out of the 192 possible choices, we plot their variance against their expected return (Figure 7). According to the investor's propensity to risk, 10 different portfolios could be selected (Table 2).

Table 2 Indices composing the 10 portfolios marked in Figure 7.

Portfolio	Index 1	Index 2	Index 3	Index 4
1	New Zealand	Portugal	USA	Hong Kong
2	New Zealand	Portugal	USA	Singapore
3	New Zealand	Austria	USA	Hong Kong
4	New Zealand	Portugal	USA	Singapore
5	New Zealand	Portugal	Canada	Hong Kong
6	New Zealand	Portugal	Canada	Singapore
7	New Zealand	Austria	Canada	Hong Kong
8	New Zealand	Austria	Canada	Singapore
9	Australia	Austria	Canada	Hong Kong
10	Australia	Austria	Canada	Singapore

The returns of these 10 portfolios during the crisis period (from June 2, 2008 to June 30, 2008) are compared to the returns of 500 minimum variance portfolios built with 4 randomly selected indices, (Figure 8, left panel) and to the naive minimum variance portfolio (Figure 8, right panel).

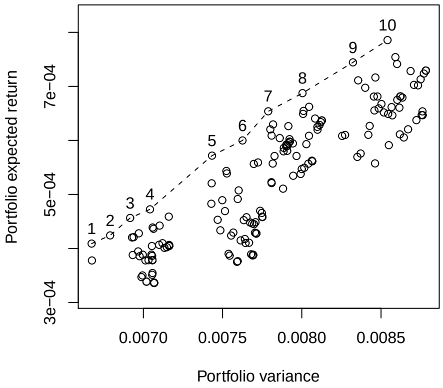

Fig. 7 Variance against expected returns for the 192 minimum variance portfolios composed by 4 indices, selected by imposing the restriction that each index belongs to a different cluster.

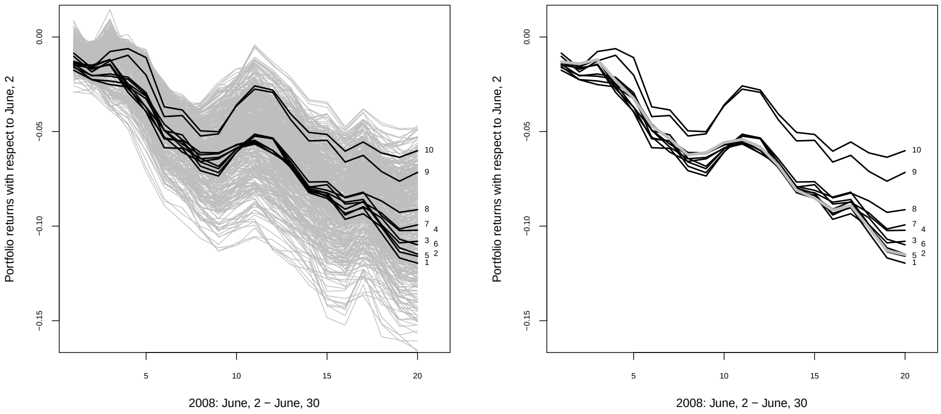

Fig. 8 Returns of the 10 selected minimum variance portfolios (black lines), compared (left) to the returns of 500 minimum variance portfolios composed by 4 randomly selected indices (thin gray lines) and (right) to the returns of the naive minimum variance portfolio composed by all the indices (bold gray line), June 2008.

Higher-risk portfolios (labelled as 9 and 10) outperform the main part of the randomly selected portfolios. Moreover, they exhibit a much better result than the naive minimum variance portfolio.

7 Concluding remarks

In this paper we propose a clustering procedure for financial time series, focusing attention on extreme values. Based on the concept of tail dependence, we define a dissimilarity measure able to quantify the extent to which extremely negative events tend to occur simultaneously for pairs of financial assets. This results in a clustering structure grouping financial time series moving similarly during crisis periods and this can be useful for a portfolio selection purpose. The proposed tail dependence dissimilarity measure does not fulfil the triangle inequality and can therefore not be used for clustering methods that are based on a distance measure. For this reason we have proposed a two-step clustering procedure to be used with algorithms needing in input either a distance matrix or the actual measurements. After checking the performance of the proposed clustering procedure by simulation studies, we present a real data case study based on MSCI geographical indices.

Future research will be aimed at adapting the time series clustering procedure introduced in this paper to the analysis of financial contagion, that is the transmission of financial market crises. Masson (1999) highlights the various concepts of contagion and the different underlying economic and financial mechanisms. From a statistical point of view, we point out that co-movements of returns do not have to be necessarily considered as an evidence of contagion, as they could be simply the effect of markets reactions to each other, as driven by their traditional relationships of cross-correlation. We can consider contagion only a situation when there is a significant increase in cross-correlation during the crisis periods (Baig and Goldfajn 1999 and Forbes and Rigobon 2002). The temporal lag of the analysis depends on the rapidity of propagation of the crisis, which is usually very fast (frequently almost simultaneous) in financial markets, and somehow slower when it is driven by macroeconomic fundamentals. In any case, contagion is a phenomenon requiring a deep analysis, both for its potentially dangerous effects and for the fact that economic and financial globalization makes it more and more frequent. Our idea is to develop a measure of the excess of correlation between assets during crisis periods using tail dependence coefficients and then to define a dissimilarity measure for clustering time series on this concept. The resulting clustering structure could then be used to draw a map of contagion within financial markets.

References

1. Alonso AM, Berrendero JR, Hernández A, Justel A 2006 Time series clustering based on forecast densities. Comput Stat Data Anal 51:762-776
2. Baig T, Goldfajn I 1999 Financial market contagion in the Asian crisis. IMF Staff Papers 46:167-195
3. Bartholomew DJ 1959 A test of homogeneity for ordered alternatives. Biometrika 38:337-344
4. Bohte Z, Cepar D, Kosmelj K 1980 Clustering of time series. In: Barritt MM, Wishart D (eds) COMPSTAT 1980, Proceedings in Computational statistics, Physica-Verlag, Wien, pp. 587-593

5. Bollerslev T, Chou R, Kroner K 1992 ARCH modeling in finance: a review of the theory and empirical evidence. Journal of Econometrics, 52:5-59
6. Breymann W, Dias A, Embrechts P 2003 Dependence structures for multivariate high-frequency data in finance. Quant Fin 3:1-14
7. Caiado J, Crato N, Peña D 2006 A periodogram-based metric for time series classification. Comput Stat Data Anal 50:2668-2684
8. Cherubini U, Luciano E, Vecchiato W 2004 Copula methods in finance. Wiley, New York
9. Coombs CH 1964 A theory of data. Wiley, New York
10. Corduas M, Piccolo D 2008 Time series clustering and classification by the autoregressive metrics. Comput Stat Data Anal 52:1860-1872
11. De Luca G, Rivieccio G, Zuccolotto P 2010 Combining random forest and copula functions: a heuristic approach for selecting assets from a financial crisis perspective. Int Syst Acc Fin Man 17:91-109
12. D'Urso P, Maharaj EA 2009 Autocorrelation-based fuzzy clustering of time series. Fuzzy Sets and Systems 160:3565-3589
13. Forbes K, Rigobon R 2002 No contagion, only interdependence: measuring stock market co-movements. The Journal of Finance 57:2223-2261
14. Fortin I, Kuzmics C 2002 Tail-dependence in stock-return pairs. Int J Intell Sys Acc Fin Mgmt 11:89-107
15. Galeano P, Peña D 2006 Multivariate analysis in vector time series. Resenhas 4:383-404
16. Joe H 1997 Multivariate models and dependence concepts. Chapman & Hall/CRC, New York
17. Kakizawa Y, Shumway RH, Taniguchi M 1998 Discrimination and clustering for multivariate time series. J Am Stat Assoc 93:328-340
18. Kaufman L, Rousseaw PJ 1990 Finding groups in data. Wiley, New York
19. Kruskal JB 1964a Multidimensional scaling by optimizing a goodness of fit to a nonmetric hypothesis. Psychometrica 29:1-27
20. Kruskal JB 1964b Nonmetric multidimensional scaling: a numerical method. Psychometrica 29:115-129
21. Masson P 1999 Contagion. J Int M Fin 18:587-602
22. Nelsen R 2006 An introduction to copulas. New York, Springer
23. Otranto E 2008 Clustering heteroskedastic time series by model-based procedures. Comput Stat Data Anal 52:4685-4698
24. Pattarin F, Paterlini S, Minerva T 2004 Clustering financial time series: an application to mutual funds style analysis. Comput Stat Data Anal 47:353-372
25. Patton AJ 2006 Modelling asymmetric exchange rate dependence. Int Ec Rev 47:527-556
26. Piccolo D 1990 A distance measure for classifying ARMA models. J Time Ser Anal 11:153-164
27. Shepard RN 1962a The analysis of proximities: multidimensional scaling with an unknown distance function-I. Psychometrica 27:125-140
28. Shepard RN 1962b The analysis of proximities: multidimensional scaling with an unknown distance function-II. Psychometrica 27:219-246
29. Taniguchi M, Kakizawa Y 2000 Asymptotic theory of statistical inference for time series. Springer, New York
30. Vilar JA, Alonso AM, Vilar JM 2010 Non-linear time series clustering based on nonparametric forecast densities. Comput Stat Data Anal 54:2850-2865
31. Weng X, Shen J 2008 Classification of multivariate time series using two-dimensional singular value decomposition. Knowledge-Based Systems 21:535-539