Hausdorff clustering

Nicolas Basalto,[1] Roberto Bellotti,[2,3,4] Francesco De Carlo,[2,3] Paolo Facchi,[5,3] Ester Pantaleo,[2,3,*] and Saverio Pascazio[2,3]
[1]*Trading Risk Management, Holding, UniCredit SpA, 20121 Milan, Italy*
[2]*Dipartimento di Fisica, Università di Bari, I-70126 Bari, Italy*
[3]*Istituto Nazionale di Fisica Nucleare, Sezione di Bari, I-70126 Bari, Italy*
[4]*TIRES, Center of Innovative Technologies for Signal Detection and Processing, Bari, Italy*
[5]*Dipartimento di Matematica, Università di Bari, I-70125 Bari, Italy*
(Received 18 December 2007; revised manuscript received 31 July 2008; published 28 October 2008)

A clustering algorithm based on the Hausdorff distance is analyzed and compared to the single, complete, and average linkage algorithms. The four clustering procedures are applied to a toy example and to the time series of financial data. The dendrograms are scrutinized and their features compared. The Hausdorff linkage relies on firm mathematical grounds and turns out to be very effective when one has to discriminate among complex structures.

DOI: 10.1103/PhysRevE.78.046112 PACS number(s): 89.65.−s, 89.75.Fb, 02.50.Sk

I. INTRODUCTION

Clustering is the classification of objects into different groups according to their degree of *similarity* [1]. A number of criteria can be used to define this intuitive (and central) concept, leading in general to different partitions. Due to this arbitrariness, clustering is an inherently ill-posed problem, as a given data set can be partitioned in many different ways without any particular reason to prefer one solution to another. It is clear that a clustering technique can be profoundly influenced by the strategy adopted by the observer and his or her own ideas and preconceptions on the problem.

Clustering algorithms can be classified in two main categories [2]: *hierarchical* and *partitive*. Hierarchical methods yield nested partitions, in which any cluster can be further divided in order to observe its underlying structure. Typical examples are the *agglomerative* and *divisive* algorithms that produce *dendrograms* [3]. On the other hand, partitional methods provide one definite partition.

In addition, two different approaches can be adopted: *parametric* and *nonparametric* clustering. Parametric algorithms are adopted when some *a priori* knowledge about the clusters is available and this information is used to make some assumptions on the underlying structure of the data. Vice versa, the nonparametric approach to clustering may represent the optimal strategy when there is no prior knowledge about the data. These and other clustering techniques are reviewed in [4–6].

From the mathematical point of view, given a set of objects $\mathcal{S}\equiv\{s\}$, an allocation function $m:s\in\mathcal{S}\to\{1,2,\ldots,k\}$, must be defined so that $m(s)$ is the class label and k is the total number of clusters (which we assume to be finite for simplicity); k may be chosen *a priori* or computed within the algorithm. The aim of a clustering procedure is to select, among all possible allocation functions, the one performing the best partition of the set $\mathcal{S}$ into subsets $\mathcal{G}_\alpha\equiv\{s\in\mathcal{S}:m(s)=\alpha\}$ $(\alpha=1,\ldots,k)$, relying on some measure of similarity. The space of any clustering solution is the set $\mathcal{M}$ of all possible allocation functions.

In this paper we will focus on a class of clustering techniques called *linkage algorithms*. Linkage algorithms are hierarchical nonparametric methods that merge, at each step, the two clusters with the smallest dissimilarity, starting from clusters made of a single element, ending up in one cluster collecting all data. Notice that, in general, a similarity measure need not be a distance in the mathematical sense; on the other hand, if one aims at clustering in a parameter space, a distance could be the best choice because it does not introduce any arbitrariness. We will analyze the so-called single, average, and complete linkage methods, which are examples of commonly used linkage algorithms that do not rely upon a mathematical definition of distance, and will introduce a linkage method based on a suitable metric in the space of the partitions of the given data set [7]. We shall make use of a distance introduced by Hausdorff and will call the corresponding linkage algorithm *Hausdorff linkage*. It is worth stressing that alternative philosophies are also possible, in which the clustering algorithm is governed by purely topological notions and unveils, e.g., efficient collective dynamics in animal behavior [8]. A comparison among these methods belongs to the realm of statistical mechanics and is beyond the scope of this paper. See [9] for an excellent discussion.

We will focus on finite sets and clusters, although we will keep our analysis on the metric features of the relevant spaces as general as possible. We will start in Sec. II by reviewing and clarifying some mathematical concepts concerning distance and linkage methods, focusing on the single, average, and complete linkage algorithms in Sec. III. The Hausdorff distance and the related clustering procedure will be introduced in Sec. IV. Section V is devoted to the comparison of the different methods on some data sets, including both a toy problem and a case study on financial time series. Some conclusions are drawn in Sec. VI.

II. PRELIMINARES

A. Distances and pseudodistances

We start by recalling the mathematical definition of distance. Given a set $\mathcal{S}$, a distance (or a metric) δ is a nonnegative function

*ester.pantaleo@ba.infn.it

$$\delta: \mathcal{S} \times \mathcal{S} \to \mathbb{R}_+ \tag{1}$$

on $\mathbb{R}_+=[0,\infty)$, endowed with the following properties, valid $\forall\ x,y \in \mathcal{S}$:

$$\delta(x,y) = 0 \quad \Leftrightarrow \quad x = y, \tag{2}$$

$$\delta(x,y) = \delta(y,x), \tag{3}$$

$$\delta(x,y) \leqslant \delta(x,z) + \delta(y,z), \quad \forall\ z \in \mathcal{S}. \tag{4}$$

Incidentally, notice that symmetry (3), as well as non-negativity, are not independent assumptions, but easily follow from Eq. (2) and the triangular inequality (4). If the triangular inequality is written as

$$\delta(x,y) \leqslant \delta(x,z) + \delta(z,y), \quad \forall\ z \in \mathcal{S}, \tag{5}$$

as is often the case, symmetry (3) must be independently postulated. We will henceforth denote a metric space by $(\mathcal{S},\delta)$.

A function (1) is a pseudometric [10] if property (2) is weakened

$$x = y \quad \Rightarrow \quad \delta(x,y) = 0. \tag{6}$$

In such a case, *distinct* elements of the set $\mathcal{S}$ can be at a null distance. A set endowed with a (pseudo)metric is called a (pseudo)metric space.

B. Linkage algorithms

Linkage algorithms are hierarchical methods, yielding a clustering structure that is usually displayed in the form of a tree or dendrogram [3]. We will adopt an agglomerative algorithm, where the clusters are linked through an iterative process, whose successive steps are the following. Given a data set $\mathcal{S}$, made up of n elements, at the first level (leaves of the dendrogram) the number of classes is equal to the number of elements. We assume (without loss of generality) that $\mathcal{S}$ is a metric space [11]. At the first iteration the two closest elements are clustered together, reducing the number of classes to $n-1$ (if more than two elements are at the closest distance, we pick a random couple among them). At the second iteration one has to tackle the subtler problem of defining a distance between the remaining elements of $\mathcal{S}$ and the first cluster formed. When this is done, the distances are recomputed and the two closest objects are joined. At the following iterations one has to tackle the much more subtle problem of defining a distance among classes. Clearly, this can be done in a variety of different ways and entails further elements of arbitrariness. Assume that this procedure can be carried out consistently. After $n-1$ steps, all the points are grouped together in one cluster, corresponding to the whole data set. The agglomerative procedure is reversed in a straightforward way in the so-called divisive approach: starting from one single cluster, this is iteratively divided into smaller and smaller ones, until single elements are obtained.

The most commonly used algorithms of this type are the "single," the "average," and the "complete" linkage, that differ in the definition of "distance" between subsets of points. In the next section we will briefly review these three algorithms.

III. SINGLE, AVERAGE, AND COMPLETE LINKAGE

A. "Distances"

Linkage algorithms differ from each other for the different similarity criteria used to build the clusters. An optimal criterium would rely on a metric d defined on the subsets of the parent space $\mathcal{S}$ as follows:

$$d: \mathcal{K}(\mathcal{S}) \times \mathcal{K}(\mathcal{S}) \to \mathbb{R}_+, \tag{7}$$

where $\mathcal{K}(\mathcal{S})$ is the collection of all the nonempty compact subsets of $\mathcal{S}$. [We restrict the metric to the above class of subsets in order to avoid some pathologies (see later).] Such a metric can be defined in a natural way by using the original metric δ defined on $\mathcal{S}$. If A and B are two nonempty compact subsets of $\mathcal{S}$, the single and complete linkage Ansätze make use of the following "distances"

$$d_s(A,B) = \inf_{a\in A, b\in B} \delta(a,b), \tag{8}$$

$$d_c(A,B) = \sup_{a\in A, b\in B} \delta(a,b). \tag{9}$$

On the other hand, the "distance" upon which the average linkage Ansätz relies needs additional structure on $\mathcal{S}$. Given a regular positive Borel measure μ on $\mathcal{S}$, such that $\mu(X) < \infty$ for all $X \in \mathcal{K}(\mathcal{S})$, one defines

$$d_a(A,B) = \frac{1}{\mu(A)\mu(B)} \int_A \int_B \delta(a,b) d\mu(a) d\mu(b), \tag{10}$$

for $A,B \in \mathcal{K}(\mathcal{S})$ [12].

However, it is easy to check that none of the above functions is a *bona fide* distance in the mathematical sense. Let us discuss the different characteristics of (8)–(10).

The function (8) is non-negative and symmetric, so Eq. (3) is valid. Notice that the pseudometric property is satisfied,

$$A = B \quad \Rightarrow \quad d_s(A,B) = 0, \tag{11}$$

although the converse is not true [so that property (2) is not valid]: consider for instance, two sets A and B such that $A \cap B \neq \emptyset$: in this case $d_s(A,B)=0$, as this is, by definition, the distance δ of a common element from itself. Finally, the triangular inequality (4) is not verified, as can be easily inferred by looking at the counterexample in Fig. 1, for which

$$d_s(A,B) > d_s(A,C) + d_s(B,C). \tag{12}$$

The function d_s is therefore neither a metric nor a pseudometric. As we shall see in Sec. III C, this problem gives rise to the chaining effect.

The function (9) is obviously non-negative and symmetric, so Eq. (3) is valid. Moreover, the triangular inequality (4) is satisfied as follows:

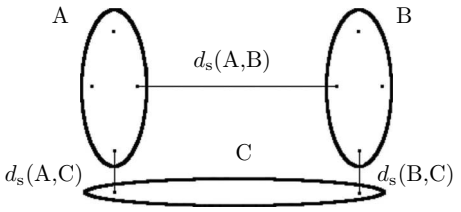

FIG. 1. Three sets A, B, C, containing each two elements, for which d_s does not satisfy the triangular inequality: $d_s(A,B) > d_s(A,C) + d_s(B,C)$.

$$\begin{aligned} d_c(A,B) &= \sup_{a\in A, b\in B} \delta(a,b) \leqslant \sup_{a\in A, b\in B, c\in C} [\delta(a,c) + \delta(b,c)] \\ &\leqslant \sup_{a\in A, b\in B, c\in C} \delta(a,c) + \sup_{a\in A, b\in B, c\in C} \delta(b,c) \\ &= \sup_{a\in A, c\in C} \delta(a,c) + \sup_{b\in B, c\in C} \delta(b,c) \\ &= d_c(A,C) + d_c(B,C). \end{aligned} \tag{13}$$

Yet, property (2) is not valid in general, as for a set A made up of more than one element, the distance of A from itself equals the distance between its farthest objects.

$$d_c(A,A) \neq 0. \tag{14}$$

This is graphically displayed in Fig. 2 and shows that Eq. (9) is not even a pseudodistance [13].

Intuitively, this is not an important issue for "small" sets, but it becomes an increasingly serious problem for "larger" sets. Clearly, the notions of small and large must be properly defined: for a compact metric space of size R, we may say that a subset of size r is small if $r \ll R$ (say by at least one order of magnitude) [14]. This situation will directly concern us in the following sections.

We finally look at the function (10), which is also nonnegative and symmetric, so Eq. (3) is valid. Moreover, the triangular inequality (4) is satisfied: indeed, from Eq. (5), $\forall\ a \in A, c \in C$, integrating over B,

$$\mu(B)\,\delta(a,c) \leqslant \int_B \delta(a,b)\,d\mu(b) + \int_B \delta(b,c)\,d\mu(b), \tag{15}$$

so that

$$\delta(a,c) \leqslant \frac{\int_B \delta(a,b)\,d\mu(b) + \int_B \delta(b,c)\,d\mu(b)}{\mu(B)}, \tag{16}$$

and

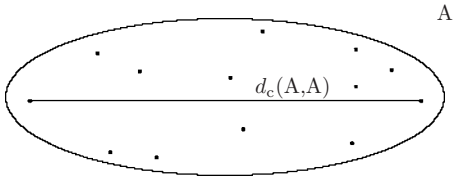

FIG. 2. For a set A containing more than one element, $d_c(A,A) \neq 0$ and neither Eq. (2) nor Eq. (6) are valid.

$$\frac{\int_A \int_C \delta(a,c)}{\mu(A)\mu(C)} \leqslant \frac{\int_A \int_B \delta(a,b)}{\mu(A)\mu(B)} + \frac{\int_B \int_C \delta(b,c)}{\mu(B)\mu(C)}. \tag{17}$$

In conclusion,

$$d_a(A,C) \leqslant d_a(A,B) + d_a(B,C).$$

However, like in the preceding case (complete linkage), property (2) is not valid, in general, as for a set A of positive measure, the distance of A from itself is, in general, strictly positive,

$$d_a(A,A) \neq 0. \tag{18}$$

It is clear that the average linkage is "in between" the single and complete, although it is more similar, in spirit, to the latter (whose features it largely shares).

B. Finite sets

We will explicitly look at the practical case in which Eqs. (8)–(10) are evaluated on *finite* sets. It is therefore convenient to specialize the formulas of the preceding section to such a situation. Let $A = \{a_i\}_{i=1,\ldots,I}$ and $B = \{b_j\}_{j=1,\ldots,J}$ be two finite sets and

$$\delta_{ij} = \delta(a_i, b_j), \tag{19}$$

the distance between any two elements of A and B. Moreover, let μ be the counting measure so that $\mu(A) = |A| = I$. The δ_{ij}'s can be arranged in a $I \times J$ distance matrix. Equations (8), (10), and (9) then read

$$d_s(A,B) = \min_{i\in A} \min_{j\in B} \delta_{ij}, \tag{20}$$

$$d_a(A,B) = \frac{1}{IJ} \sum_{i\in A} \sum_{j\in B} \delta_{ij}, \tag{21}$$

$$d_c(A,B) = \max_{i\in A} \max_{j\in B} \delta_{ij}, \tag{22}$$

for the single, average, and complete linkage algorithms, respectively. In practice, this amounts to determining the smaller and the larger value among the elements of the distance matrix, for the single and complete linkage algorithm, respectively, and the sum of all the elements for the average algorithm, all tasks that can be performed in a polynomial time. These formulas will be applied to the following examples.

C. Comments

It is worth commenting on the features of the three clustering Ansätze introduced, emphasizing their limits and positive aspects. The single linkage algorithm tends to yield elongated clusters, which are sometimes difficult to understand and poorly significant [3]: this is known as the *chaining effect*. On the contrary, the average and complete linkage have the advantage of clustering "compact" clusters and yield well localized classes. In general, the partitions obtained using them are more significant. Their major disadvantage is that they do not set equal to zero the distance of a "compact" set

from itself [see Eqs. (14) and (18), and Fig. 2], performing *de facto* a coarse graining. In a few words, d_a and d_c look at the clustered data points with a "minimal resolution" (that is also, unfortunately, cluster dependent) and are unable to (i) recognize the complexity of a finely structured cluster and (ii) extract "nested" or "concentric" clusters, such as those displayed in Fig. 1 of Ref. [15]. Notice that, by contrast, such nested clusters are very efficiently detected by the single linkage algorithm. For the same reason, this procedure suffers from the so-called "chaining effect," that will be scrutinized in the following sections.

As already emphasized, average linkage is in some sense in between single and complete linkage (although it is closer to the latter). It is often reputed to be the best linkage algorithm; however, it is not a distance, as it does not satisfy property (2), as previously noticed. In the next section we shall introduce a procedure that makes use of an underlying *bona fide* distance. This will have some advantages, also from a conceptual viewpoint, as it enables one to rest on firm mathematical background.

IV. HAUSDORFF DISTANCE AND HAUSDORFF LINKAGE

In the light of the discussion of the preceding section, it appears convenient to approach the clustering problem from a "neutral" perspective, by looking for a linkage algorithm based on a well-defined mathematical similarity criterium. In order to do this, we will use a distance function introduced by Hausdorff [16].

The Hausdorff distance has been used in the literature for image matching [17,18] and partitional clustering algorithms [19]. Application to hierarchical clustering algorithms can be found in [20], through software implementation together with other intercluster and intracluster distances, and in [21], where the hierarchical clustering of nodes in wireless sensors networks is analyzed. We shall focus on hierarchical clustering and analyze this clustering technique from a theoretical point of view, before comparing it to other methods and looking at applications.

A. Hausdorff distance

Given a metric space $(\mathcal{S},\delta)$, the distance between a point $a \in \mathcal{S}$ and a (nonempty and compact) subset $B \in \mathcal{K}(\mathcal{S})$ is naturally given by

$$\tilde{d}(a;B) = \inf_{b\in B} \delta(a,b). \tag{23}$$

Given a subset $A \in \mathcal{K}(\mathcal{S})$, consider the function

$$\tilde{d}(A;B) = \sup_{a\in A} \tilde{d}(a;B) = \sup_{a\in A}\inf_{b\in B} \delta(a,b), \tag{24}$$

which measures the largest distance $\tilde{d}(a;B)$, with $a \in A$. Note that here the strategy is opposite to that used with the single linkage "distance" (8), where one considers instead the *smallest* distance $\tilde{d}(a;B)$, with $a \in A$. The function (24) is not symmetric, $\tilde{d}(A;B) \neq \tilde{d}(B;A)$, and therefore is not a bona fide distance, as it does not satisfy Eq. (3). The Hausdorff distance [16] between two sets $A,B \in \mathcal{K}(\mathcal{S})$ is defined as the largest between the two numbers as follows:

$$d_H(A,B) = \max\{\tilde{d}(A;B),\tilde{d}(B;A)\}, \tag{25}$$

namely,

$$d_H(A,B) = \max\left\{\sup_{a\in A}\inf_{b\in B} \delta(a,b), \sup_{b\in B}\inf_{a\in A} \delta(a,b)\right\}, \tag{26}$$

which is clearly symmetric and satisfies all axioms (2)–(4).

It is worth discussing a bit more the mathematical features of d_H. This will help us grasp its interesting properties, towards physical applications.

Given a set $A \in \mathcal{K}(\mathcal{S})$ and a positive real number $r>0$, define the *open r neighborhood* of A as

$$N_r(A) = \{y:\ \tilde{d}(y;A) < r\}. \tag{27}$$

The Hausdorff distance between two sets $A,B \in \mathcal{K}(\mathcal{S})$ can be reexpressed as

$$d_H(A,B) = \inf\{r:\ A \subseteq N_r(B) \text{ and } B \subseteq N_r(A)\}. \tag{28}$$

Indeed,

$$\begin{aligned} d_H(A,B) &= \inf\{r:\ A \subseteq N_r(B), B \subseteq N_r(A)\} \\ &= \inf(\{r:\ A \subseteq N_r(B)\} \cap \{r:\ B \subseteq N_r(A)\}) \\ &= \max\{\inf\{r:\ A \subseteq N_r(B)\}, \inf\{r: B \subseteq N_r(A)\}\}, \end{aligned} \tag{29}$$

and since

$$\inf\{r:\ A \subseteq N_r(B)\} = \sup_{x\in A}\inf\{r:\ x \in N_r(B)\} = \sup_{x\in A}\inf_{y\in B} \delta(x,y), \tag{30}$$

and analogously for $\inf\{r: B \subseteq N_r(A)\}$, one gets again Eq. (26). Stated differently, the Hausdorff distance can also be defined as the smallest radius r such that $N_r(A)$ contains B and at the same time $N_r(B)$ contains A.

In other words, the Hausdorff distance between A and B is the smallest positive number r, such that every point of A is within the distance r of some point of B, *and* every point of B is within the distance r of some point of A. The geometrical meaning of the Hausdorff distance is best understood by looking at an example, such as that in Fig. 3. We emphasize that the Hausdorff metric on the subsets of $\mathcal{S}$ is defined in terms of the metric δ on the points of $\mathcal{S}$.

The Hausdorff distance enjoys a number of interesting features that are worth discussing. We have defined d_H only on nonempty compact sets for the following reasons. Consider, for example, the real line. Then, by adopting the convention $\inf\{\emptyset\}=\infty$ [22], one gets $\forall\ x, d_H(\emptyset,x)=\infty$, which is not allowed by any definition of metric. This suggests that we should restrict our attention to nonempty sets. Moreover, $d_H(\{0\},[0,\infty))=\infty$, which is again not allowed. We then restrict the use of d_H only to bounded sets. Finally, the Hausdorff distance between two not equal sets could vanish (which would make d_H a pseudometric [see Eq. (6)]): for instance, $d_H((0,1),[0,1])=0$. Therefore we will restrict the application of d_H only to closed sets.

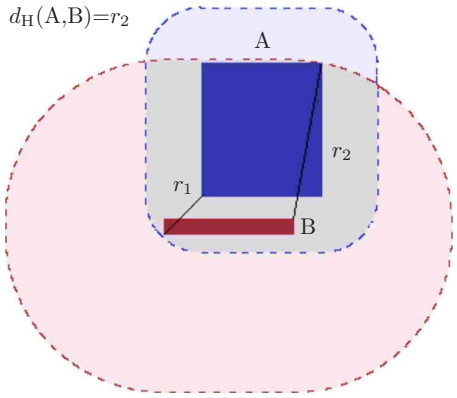

FIG. 3. (Color online) Hausdorff distance between two sets A (a square) and B (a rectangle). The open neighborhoods $N_{r_1}(A)$ and $N_{r_2}(B)$ are shaded, $r_1=\tilde{d}(B;A)$, $r_2=\tilde{d}(A;B)$. The Hausdorff distance is r_2.

More generally, it is easy to prove the following.

Theorem. The Hausdorff function d_{H} is a metric on the set $\mathcal{K}(\mathcal{S})$. Moreover, if $(\mathcal{S},\delta)$ is a complete metric space, then the space $(\mathcal{K}(\mathcal{S}),d_{\mathrm{H}})$ is also complete.

Although of an abstract nature, this is of *physical significance*, as it enables one to be confident about the metric properties of $\mathcal{K}(\mathcal{S})$ even for fine-structured clusters. Notice that the property of completeness could not even be conceived for the "distances" d_a and d_c used for the average and complete linkage in the last section. In conclusion,

$$d_{\mathrm{H}}\colon \mathcal{K}(\mathcal{S})\times\mathcal{K}(\mathcal{S})\to\mathbb{R}_+ \tag{31}$$

is a complete metric. In the cases of interest, $\mathcal{S}$ will be a complete metric space, e.g., a Euclidean space.

We close this section with three remarks. First, if the data set is finite and consists of N elements, all distances can be arranged in a $N\times N$ matrix δ_{ij} and Eq. (26) reads

$$d_{\mathrm{H}}(A,B)=\max\left\{\max_{i\in A}\min_{j\in B}\delta_{ij},\max_{j\in B}\min_{i\in A}\delta_{ij}\right\}, \tag{32}$$

which is a very handy expression, as it amounts to finding the minimum distance in each row (column) of the distance matrix, then the maximum among the minima. The two numbers are finally compared and the largest one is the Hausdorff distance. This sorting algorithm is efficient and can be easily implemented.

Second, $\forall\ A,B\in\mathcal{K}(\mathcal{S})$,

$$d_s(A,B)\leqslant d_{\mathrm{H}}(A,B)\leqslant d_c(A,B), \tag{33}$$

$$d_s(A,B)\leqslant d_a(A,B)\leqslant d_c(A,B). \tag{34}$$

These are simple consequences of Eqs. (26) and (8)–(10) [or Eqs. (32) and (20)–(22) in the discrete case]. In some sense, d_c overestimates the distance between two given sets, essentially because it includes in such a distance the very "size" (14) of the set (see Fig. 2). The same observation is valid for d_a, although the average (10) partially compensates for this drawback. On the other hand, d_s underestimates the distance between two given sets.

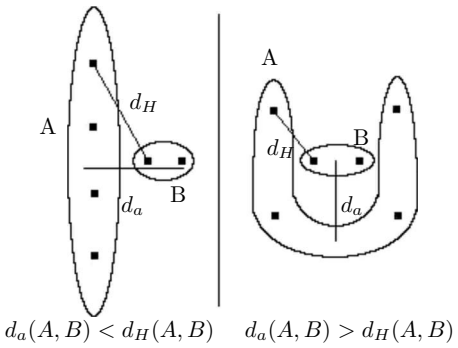

FIG. 4. Comparison between Hausdorff and average. (a) $d_a(A,B)<d_{\mathrm{H}}(A,B)$; (b) $d_a(A,B)>d_{\mathrm{H}}(A,B)$.

Third, by looking at Eqs. (33) and (34) one sees that d_a and d_{H} have a similar behavior. Moreover, there is no order relation between them. Both cases are possible, $d_a<d_{\mathrm{H}}$ and $d_a>d_{\mathrm{H}}$: see Fig. 4. In general, d_a performs better when $d_a\simeq$ distance between the "centers of mass" $=\delta(x_A,x_B)$, where $x_A=\mu(A)^{-1}\int_A a d\mu(a)$. By contrast, for sets that are entangled or intertwined, d_a, being not a distance, starts suffering from the drawbacks that derive from Eq. (18). Clearly, d_{H}, being always a mathematical well-defined distance, can discriminate even sets that are very entangled.

As we shall see, these observations have important consequences when one clusters complex and/or large sets.

B. Hausdorff linkage

We shall take the Hausdorff distance as our dissimilarity measure. This distance naturally translates in a linkage algorithm: at the first level each element is a cluster, the Hausdorff distance between any pair of points reads

$$d_{\mathrm{H}}(\{i\},\{j\})=\delta_{ij}, \tag{35}$$

and coincides with the underlying metric. The two elements of $\mathcal{S}$ at the shortest distance are then joined together in a single cluster. The Hausdorff distance matrix is recomputed, considering the two joined elements as a single set. This iterative process goes on until all points belong to a single final cluster.

Clearly, when evaluating distances among single elements (points), the four procedures d_{H}, d_s, d_a, and d_c yield the same result. The output of the single linkage algorithm will clearly differ very quickly from the other three, due to the drawbacks of the chaining effect. On the other hand, the differences among Hausdorff, average, and complete linkage will become apparent only later in the clustering process. This is a consequence of several factors. First of all, the functions d_{H} and d_c yield the same value when evaluated on a single element $\{a\}$ and a composite set B. Indeed, from Eq. (26),

$$\begin{aligned} d_{\mathrm{H}}(\{a\},B)&=\max\left\{\sup_{x\in\{a\}}\inf_{y\in B}\delta(x,y),\sup_{y\in B}\inf_{x\in\{a\}}\delta(x,y)\right\}\\ &=\max\left\{\inf_{y\in B}\delta(a,y),\sup_{y\in B}\delta(a,y)\right\}\\ &=\sup_{y\in B}\delta(a,y)=d_c(\{a\},B). \end{aligned} \tag{36}$$

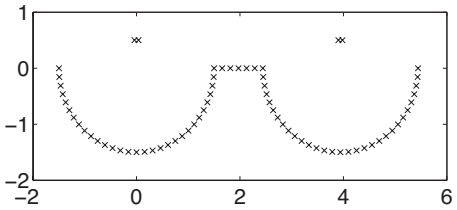

FIG. 5. A two-dimensional toy sample: a pair of glasses (each one made up of 31 points) connected by a short horizontal "bar" (5 points) and two pupils (each one made up of 2 points), for a total of $N=71$ points. In the figure the points are represented by crosses. The axes are in arbitrary units.

As a consequence of this property, at the lowest levels the Hausdorff linkage will yield a partition that is very similar to that obtained by the complete linkage algorithm. As the clustering procedure goes on, the two methods will differ from each other, because of their different criteria in evaluating distances, leading to different aggregations of more complex classes. It is at this point that the output of the complete linkage becomes less reliable, as a consequence of Eqs. (14) and (33). As discussed after Eq. (14), we expect this problem to become serious for large sets, of size comparable to that of the parent space.

By contrast, it is easy to see that given a set composed of a single element $\{a\}$ and a generic finite set B,

$$d_a(\{a\},B) = \frac{1}{|B|}\sum_{b\in B}\delta(a,b) \leqslant d_{\mathrm{H}}(\{a\},B) = d_c(\{a\},B). \tag{37}$$

We expect therefore the partitions obtained by the average linkage algorithm to start forming more "slowly" than in the other two cases (Hausdorff and complete), by virtue of the above property (that entails a greater difficulty in discriminating single points from clusters).

In general, the Hausdorff and average linkage algorithm will yield intermediate results between those obtained by the other two procedures and their performances will tend to be similar (and difficult to discriminate). We shall now compare the four clustering methods, first on an artificial set of points in a two-dimensional Euclidean space, then on set of financial time series.

A final comment is in order. Given a distance matrix, any clustering procedure will yield a tree and an ultrametric, entailing a loss of information on the data set. However, this appears necessary and is inherent in any clustering procedure.

V. APPLICATIONS

A. Two-dimensional data set

Let us analyze the effect of the single, average, complete, and Hausdorff linkage algorithms on the data set shown in Fig. 5. This is a discrete set of points in the plane, resembling a pair of "glasses" (each one made up of 31 points) connected by a short horizontal "bar" (5 points) and two "pupils" (each one made up of 2 points), for a total of $N=71$ points.

The objective of this example is twofold. First, it shows how difficult it can be to discriminate between average, complete, and Hausdorff linkage: while the single linkage will obviously suffer from the chaining effect (as will cluster points at the opposite sides of the figure), the other three procedures will perform in a similar fashion at the beginning, yielding different clusters only when the classes become more complex. Second, the example will clarify that the Hausdorff linkage is the only algorithm, among the four analyzed, that is able to analyze a particular structure in the set of Fig. 5. At the same time, not surprisingly, we shall see that

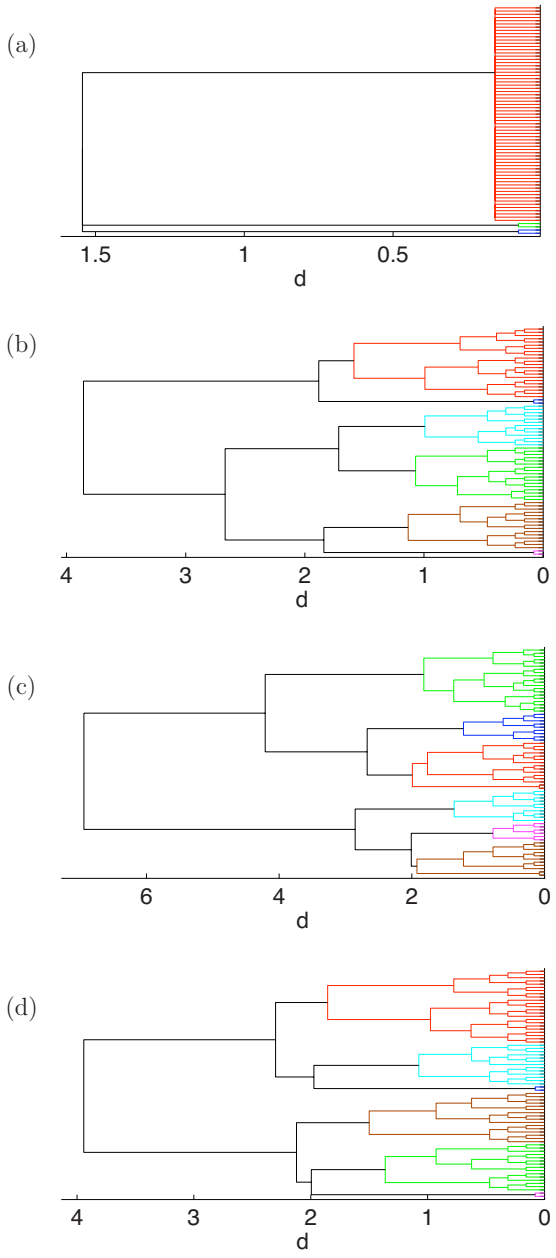

FIG. 6. (Color online) Dendrograms generated by the (a) single, (b) average, (c) complete, and (d) Hausdorff linkage, for the data set of Fig. 5. The pupils are easily recognized because they are always clustered at the smallest distance. Clustering of the remaining part of the "lenses" is more involved and significantly procedure dependent.

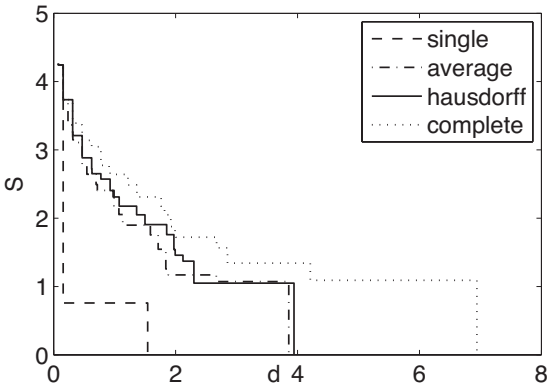

FIG. 7. Cluster entropies of the dendrograms of Fig. 6 vs distance. Dashed line: single; dot-dashed line: average; continuous line: Hausdorff; dotted line: complete.

a discrimination between Hausdorff and average is delicate.

The dendrograms generated by the four algorithms are shown in Fig. 6 and their close scrutiny unearths interesting details. The chaining effect of the single linkage is apparent. This can be an advantage if one wants to bring to light the presence of a "continuous" line of points; it is a drawback in a parameter space because data characterized by opposite values of the parameter on the abscissa in Fig. 5 are clustered together. As anticipated, a discrimination between the Hausdorff, average, and complete algorithms is more difficult. However, as discussed after Eq. (14), the differences should become apparent for "large" sets, of size comparable to that of the parent space: for a parent space made up of $N=71$ (approximately linearly distributed) points, we expect this effect to show up for sets made up of more than 7 points, as one can see in Fig. 6.

It is interesting to understand how different algorithms yield different discrimination of the "pupils." Single, Hausdorff, average, and complete linkage discriminate the pupils with increasing difficulty: starting from the "top" of the dendrogram (where only a single cluster is present), one has to cut the dendrograms at levels 3, 5, 6, and 8, respectively. In this respect, single linkage performs best (as was to be expected), Hausdorff a bit better than average, while complete is the worst. Moreover, as we shall see later, Hausdorff is best at discriminating the central bar.

A hierachical algorithm like the one discussed in this paper generates $n-1$ partitions of the initial set. Clearly, in order to get a significant partition, one must decide where to cut the dendrogram. A proper way to cut the dendrograms could be to search for a stable partition among the whole hierarchy yielded by the algorithms, in correspondence to an approximately constant value of the cluster entropy in a certain range of the dissimilarity measure d [23],

$$S(d) = -\sum_{k=1}^{N_d} P_d(k) \ln P_d(k), \tag{38}$$

where $P_d(k)$ is the fraction of elements belonging to cluster k, and N_d is the number of clusters at level d in the dendrogram. The single, average, complete, and Hausdorff entropies corresponding to the dendrograms in Fig. 6 are shown in Fig. 7 as a function of d. We emphasize that, for the case at hand, the data set was intentionally chosen so that one cannot expect an obvious partition into "sensible" clusters. For this very reason, the entropies in Fig. 7 display no "plateau." The optimal cut is then chosen at a level in which the partition consists of six subsets (except for the single linkage procedure, where the maximum number of subsets in a partition is three), according to a visual optimization of the clustering solution. Figure 8 shows the selected partitions: while the single linkage yields a clear chaining effect, average, complete, and Hausdorff methods share the positive aspect of clustering rather "compact" sets. Moreover, all other clusters being roughly similar, the average and Hausdorff procedure are also able to discriminate the two-points pupils in Fig. 5: in this respect they share the positive spin offs of the single linkage algorithm. On the other hand, the complete linkage algorithm clusters each pupil together with a part of its nearest "glass."

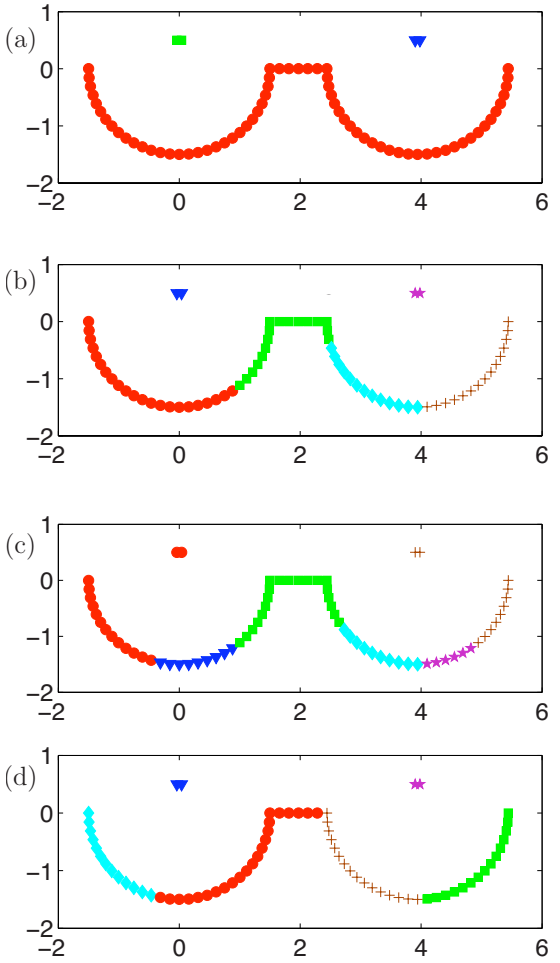

FIG. 8. (Color online) Clustering results for the (a) single, (b) average, (c) complete, and (d) Hausdorff linkage. All dendrograms (except that obtained by the single linkage algorithm) were cut when 6 clusters were present. Objects belonging to the same cluster share the same symbol, which can be a circle, a triangle, a cross, a diamond, a square, or a star: for example, the complete algorithm (c) groups the left pupil with 13 points belonging to its nearest glass (all circles). The axes are in arbitrary units.

Finally, we note that Hausdorff is the only algorithm that is in some sense able to discriminate the central bar: while the single linkage "chains" all points (except the pupils), both complete and average tend to group the central bar with points belonging to both lenses (at the lower levels of the dendrograms), then split the central bar in two parts, that are eventually separately grouped with the left and right "glasses." Although no procedure seems to be able to discriminate the central bar, Hausdorff performs a bit better than the other methods, in that the central bar is associated to the left glass. Since this association is arbitrary and depends on the initial numbering of the points when the algorithm is implemented, it is clear that a different initial labeling would have led to the central bar being clustered with the right glass [24]. The intersection between the clusters obtained using different initial labeling is then able to isolate the bar. By contrast, there is no way of discriminating the central bar if one makes use of the complete or the average clustering Ansätz. Clearly, these differences are subtle and a thorough analysis would require the study of the "robustness" of the different algorithms when "noise" effects are present (e.g., in the coordinates of the data points). This study will not be undertaken here.

B. Financial data

As a further application of the Hausdorff algorithm we consider financial time series. The aspects concerning the financial features of this method were introduced and discussed in [7]. Here we focus on the mathematical features of the dendrograms, on a quantitative basis, by (i) quantifying the difference between dendrograms according to the number of nodes that share the same descendants, and (ii) discussing a typical phenomenon of some linkage procedures, known as "reversal" [25]. In particular, we focus on the $N=30$ shares composing the Dow Jones industrial average (DJIA) index, collecting the daily closure prices of its stocks for a period of five years (1998–2002). The companies of the DJIA stock market are reported in Table I, together with the corresponding industrial areas.

We consider the temporal series of the logarithm of the ratio of two consecutive closure prices

$$X(t) \equiv \ln \frac{P(t)}{P(t-1)}, \tag{39}$$

where $P(t)$ is the closure price of a stock at day t. Both P and X are very irregular functions of time, as one can see in Fig. 9, that displays the typical behavior of a stock value [MSFT (see Table I for stock names corresponding to the ticker symbols)] for the investigated time period. In order to use the linkage algorithm, we quantify the degree of similarity between two time series X and Y by means of the correlation coefficients computed over the investigated time period as follows:

$$\rho(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}, \tag{40}$$

TABLE I. DJIA stocks classified by tickers, names, and industrial categories.

Ticker	Stock name	Industrial category
AA	Alcoa Inc.	Basic materials
AXP	American Express Co.	Financial
BA	Boeing	Capital goods
C	Citigroup	Financial
CAT	Caterpillar	Capital goods
DD	DuPont	Basic materials
DIS	Walt Disney	Services
EK	Eastman Kodak	Consumer cycl.
GE	General Electrics	Conglomerates
GM	General Motors	Consumer cycl.
HD	Home Depot	Services
HON	Honeywell International	Capital goods
HPQ	Hewlett-Packard	Technology
IBM	Int'l Business Machine	Technology
INTC	Intel Corporation	Technology
IP	Internationall Paper	Basic materials
JNJ	Johnson & Johnson	Healthcare
JPM	JP Morgan Chase	Financial
KO	Coca Cola Inc.	Consumer non-cycl.
MCD	McDonalds Corp.	Services
MMM	Minnesota Mining	Conglomerates
MO	Philip Morris	Consumer non-cycl.
MRK	Merck & Co.	Healthcare
MSFT	Microsoft	Technology
PG	Procter & Gamble	Consumer non-cycl.
SBC	SBC Communications	Services
T	AT&T Gamble	Services
UTX	United Technology	Conglomerates
WMT	Wal-Mart Stores	Services
XOM	Exxon Mobil	Energy

where E is the expectation value over the time interval of interest (one year in our case), $\mu_X = E[X]$ and $\sigma_X = \sqrt{E[X^2] - \mu_X^2}$. Figure 10(a) shows the correlation matrix $\rho(X,Y)$ computed for the year 1998: each element is displayed in a gray scale ranging from black (minimum value, zero) to white (maximum value, one). It is worth stressing that, although, in principle, ρ can take negative values, all correlation coefficients in Fig. 10(a) are positive, with values not too close to 1, thus confirming that stocks belonging to the same market do not move independently from each other, but rather share a similar temporal behavior.

The metric function we adopted to quantify the time synchronicity between two stocks is the following [26–29]:

$$d(X,Y) = \sqrt{2[1-\rho(X,Y)]}. \tag{41}$$

The distance (41) is a proper metric in the parent space, ranging from 0 for perfectly correlated series $[\rho(X,Y)=+1]$ to 2 for anticorrelated stocks $[\rho(X,Y)=-1]$. The representa-

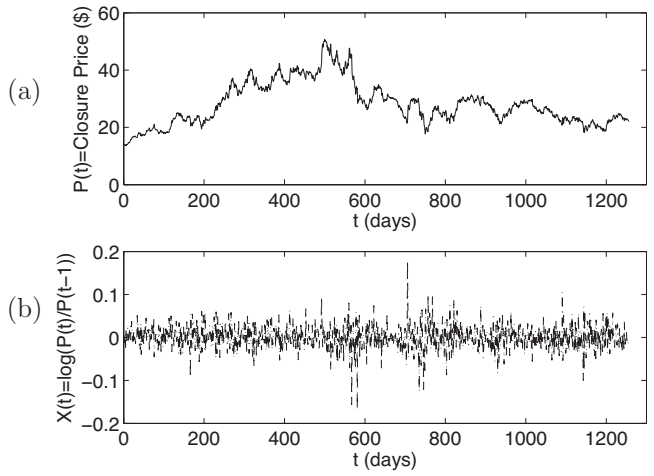

FIG. 9. (a) Time evolution of the closure price $P(t)$ and (b) the logarithm of the ratio of consecutive closure prices $X(t)$ [see Eq. (39)] of a stock value (MSFT), for the period 1998–2002.

tive points lie on a hypersphere and $d(X,Y)$ measures the Euclidean (and not the geodesic) distance between X and Y. Figure 10(b) shows the distance matrix $d(X,Y)$ computed for the year 1998: each element is displayed in a gray scale ranging from black ($d=0$) to white ($d=\sqrt{2}$). The tree structure obtained for this set was already scrutinized and discussed in Ref. [7]. We shall focus here on the features of the dendrograms.

Although our analysis pertains to the period from 1998 to 2002, for the sake of illustration, Fig. 11 only shows the dendrograms obtained by clustering the stocks yearly from 2000 to 2002, with the single, average, complete, and Hausdorff linkage. Some considerations are in order. As expected, the single linkage algorithm suffers from the chaining effect [3], which yields elongated clusters: different points merge

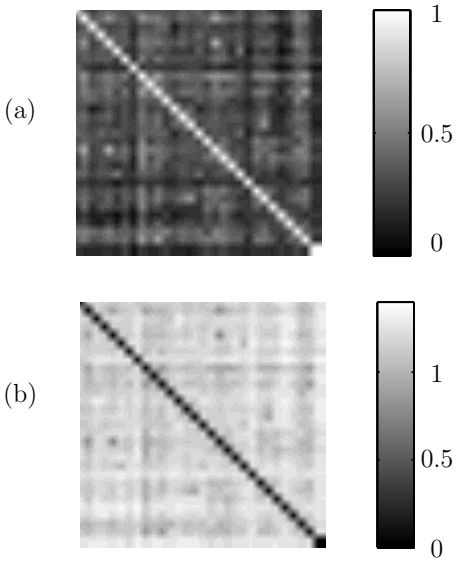

FIG. 10. (a) Correlation matrix $\rho(X,Y)$ computed for the year 1998 and (b) distance matrix $d(X,Y)$ computed for the year 1998. Numerical values are displayed in a gray scale ranging from black (minimum value) to white (maximum value).

into a large cluster almost one at a time during the iterative procedure, with the result of obtaining a poorly defined tree structure, as it can be clearly observed in Fig. 11 [panels (a), (e), and (i)]. Wherever one would choose to cut the dendrogram, no meaningful partition would emerge out of the hierarchical tree. On the other hand, the dendrograms obtained by means of the average, complete, and Hausdorff algorithms show clear inner structures, corresponding to the branches of the hierarchical tree. One recognizes the clusters corresponding to homogeneous (from the industrial viewpoint) groups of companies, belonging to the same industrial area: this is the case of the money center banks {C, JPM AXP}, retail companies {HD, WMT}, companies dealing with basic materials {AA, IP, DD}, and the technological core {IBM, INTC, MSFT}.

The classification of stocks in terms of their economic homogeneity as well as the presence of superclusters and homogeneous subgroups was already discussed in [7] and will not be analyzed here. However, there are characteristic features of the dendrograms that deserve additional attention. An interesting phenomenon, consisting of reversals in the dendrograms [25], sometimes appears in the Hausdorff clustering, as shown in Fig. 11(n), the dendrogram obtained by clustering the financial time series in 2002. This pattern is mathematically spelled out in the Appendix, where its significance is elucidated in terms of an elementary example (see Fig. 12). We take this phenomenon as an indicator of the potentialities of a clustering algorithm based on the Hausdorff distance, that could be exploited in a nonhierarchical algorithm, allowing reversals and hierarchy breaking.

It is interesting to discuss the features of the dendrograms on a more quantitative basis. The difference between dendrograms can be quantified by counting the number of nodes that share the same descendants. Table II shows these numbers for different couples of dendrograms and different years. It is apparent that the performance of the single linkage algorithm is very different from the other three methods. This is clearly reflected in the figures of Table II. By contrast, the other three methods yield results that are similar.

Most of the nodes with common descendants share a set of descendants consisting of a couple of elements. In general, the single linkage dendrogram shares with the other dendrograms only nodes having a low number of descendants. The chaining effect is responsible for this difference: at the inner levels in the dendrogram, all linkage algorithms generate similar clusters because the distance between one-element clusters is the distance of the underlying metric [as one can see, for instance, in Eq. (35)]; at higher levels in the dendrogram, the single linkage tends to link single elements to larger clusters in a chain; therefore it yields less clusters of two elements than other algorithms. As one can see, the single and the Hausdorff dendrogram, as well as the single and the complete dendrogram have almost the same number of nodes sharing common descendants; on the other hand the single and the average dendrogram have a higher number of nodes sharing the same descendants, as a consequence of property (37). Among common subsets of descendants, which mainly consist of two elements sets, there are larger subsets that derive from the union of smaller common subsets.

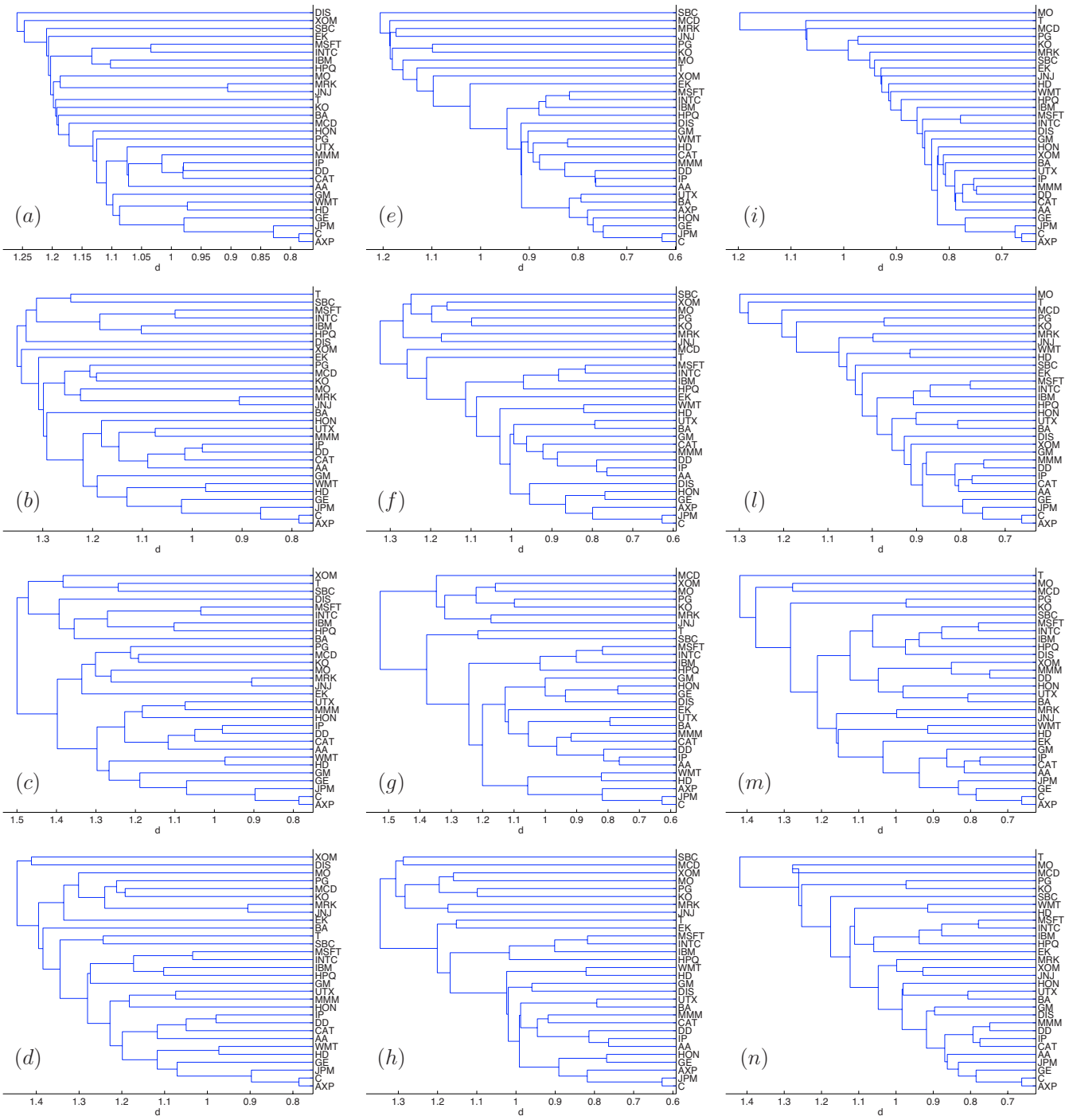

FIG. 11. (Color online) Dendrograms obtained by clustering the stocks for different years: (a) single, year 2000; (b) average, year 2000; (c) complete, year 2000; (d) Hausdorff, year 2000; (e) single, year 2001; (f) average, year 2001; (g) complete, year 2001; (h) Hausdorff, year 2001; (i) single, year 2002; (l) average, year 2002; (m) complete, year 2002; (n) Hausdorff, year 2002. The acronyms are explained in Table I. Some reversals can be clearly observed in one of the dendrograms pertaining to the year 2002, obtained by clustering the stocks with the Hausdorff linkage [panel (n)]. A mathematical explanation of this phenomenon is given in the Appendix.

The complete linkage algorithm has the larger number of common nodes, twice together with the average, three times together with the Hausdorff linkage. It never happens that the average and Hausdorff linkage have the maximum number of common nodes. The significance of this finding is of difficult interpretation. In general, the complete linkage algorithm tends to create more nodes and therefore more common nodes. Moreover, there are probably remarkable differences between the average and Hausdorff linkage algorithms, but these differences are not easy to bring to light, as they involve complicated internal structures of the dendrograms. These open problems require additional investigation.

We also counted the number of nodes, by excluding those nodes that result from joining a set and a single element, in

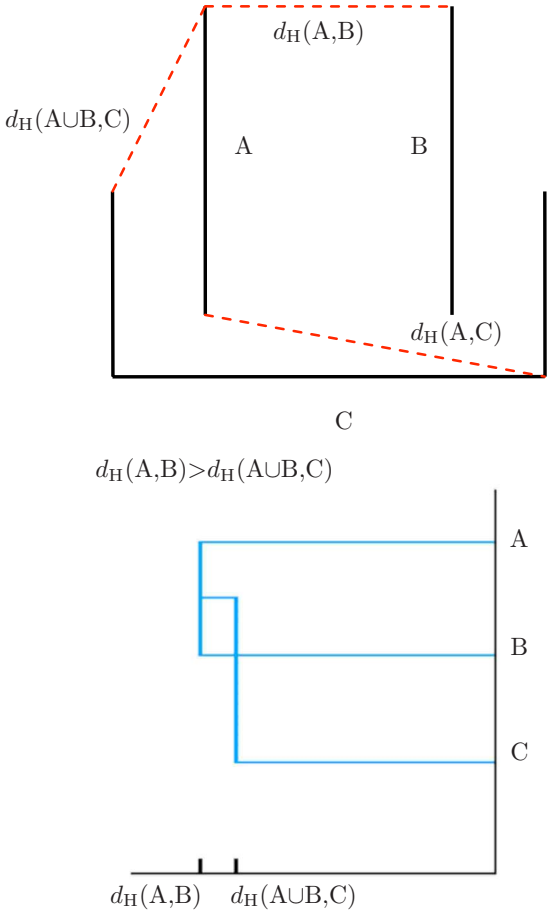

FIG. 12. (Color online) Example of a reversal in the Hausdorff linkage. Given three sets A (a segment), B (another segment), and C (a "U") (upper panel), the Hausdorff linkage algorithm links A and B at a distance $d_H(A,B)$, then links $A \cup B$ and C at a distance $d_H(A \cup B, C) < d_H(A,B)$ (lower panel). Set C is Hausdorff nearer to $A \cup B$ than it is to A and B separately. The corresponding dendrogram is drawn below.

order to avoid the consequences of Eq. (36). The results are displayed in Table III and should be interpreted with care, as the exclusion of nodes at the lowest levels in the dendrograms introduces an arbitrary bias, also as a consequence of Eqs. (36) and (37). One notices again that the single linkage performs differently. On the other hand, the differences among average, Hausdorff, and complete linkage become less significant and year dependent. We also notice that the year 2002 displays no common nodes: this feature was already present, in embryo, in Table II.

TABLE II. Number of nodes sharing the same descendants, in dendrograms pertaining to different couples of linkage algorithms.

	1998	1999	2000	2001	2002
Single—average	8	9	14	13	6
Single—Hausdorff	8	8	11	11	4
Single—complete	8	8	11	10	4
Average—Hausdorff	15	13	16	15	10
Average—complete	17	17	17	14	12
Hausdorff—complete	18	15	16	18	13

TABLE III. Number of nodes sharing the same descendants, which are not the result of the joining of a set and a single element.

	1998	1999	2000	2001	2002
Single—average	1	2	2	0	0
Single—Hausdorff	1	1	2	0	0
Single—complete	1	1	2	0	0
Average—Hausdorff	2	2	2	2	0
Average—complete	3	3	2	1	0
Hausdorff—complete	2	2	1	4	0

VI. CONCLUSIONS

Clustering is a common practice in the analysis of complex data and reflects a human compulsion towards classifying objects or physical phenomena. This can be a difficult task when the phenomena are complicated and the underlying correlations are difficult to bring to light. We have introduced and analyzed a clustering procedure based on a bona fide distance introduced by Hausdorff. The method, that relies on an underlying distance among the elements that make up the "parent" set, has been compared with the single, average, and complete linkage procedures, which only rely on an underlying dissimilarity measure (not a distance). We first looked at a toy problem, in which the Hausdorff method has advantages in comparison with the other ones. We then clustered the financial time series of the DJIA stock market, observing the formation of clusters of "homogeneous" companies and scrutinizing the features of the algorithms. The results are significant from an economical point of view.

An important application of the method introduced here is certainly in portfolio optimization [30–34], where the key issue is to select one (or a few) stocks that are representative of a given cluster, characterized by economic homogeneity, reducing maintenance costs and optimizing risk. Among the possible future developments, one should test the stability of the method against noise effects [35,36] and endeavor to understand the practical consequences of hierarchy breaking due to the reversals discussed in the previous section.

APPENDIX

We explain here the phenomenon of the reversals observed in the Hausdorff dendrogram of Fig. 11(n) and argue that the Hausdorff hierachical clustering does not exploit all the potentialities of the Hausdorff distance.

Let us consider the three compact sets of the Euclidean plane shown in Fig. 12. Set A is a segment, B is another segment, and C is a polygonal "U." They are arranged in such a way that

$$d_H(A,B) < d_H(A,C) \quad \text{and} \quad d_H(A,B) < d_H(B,C). \tag{A1}$$

Therefore, the Hausdorff linkage algorithm starts off by linking A and B at a distance $d_H(A,B)$ into a cluster $D=A \cup B$.

But now it happens that the Hausdorff distance between C and cluster D is smaller than the Hausdorff distance between A and B, namely,

$$d_{\mathrm{H}}(D,C) = d_{\mathrm{H}}(A \cup B, C) < d_{\mathrm{H}}(A,B). \tag{A2}$$

Therefore, set C is nearer to $D=A \cup B$ than it is to A and B separately,

$$d_{\mathrm{H}}(A \cup B, C) < d_{\mathrm{H}}(A,C),\ \ d_{\mathrm{H}}(B,C), \tag{A3}$$

and the corresponding dendrogram exhibits a reversal.

It can therefore happen that two sets, after their aggregation, become Hausdorff closer to a third set than they were separately. This explains (from a mathematical viewpoint) the phenomenon of the reversals observed in Fig. 11(n).

Therefore, reversals are a direct consequence of the very definition of the Hausdorff distance. The existence of reversals implies that d_{H} cannot be used as the Hausdorff hierarchy's aggregation index. Indeed, an aggregation index is a positive function f defined on the hierarchy Y satisfying (i) $f(y)=0$ if and only if y is reduced to a single element of S and (ii) $f(y)<f(y')$ if $y \in y'$. Equation (A3) is at variance with condition (ii). On the other hand, the complete, average, and single hierarchical algorithm generate a hierarchy indexed through d_c, d_a, and d_s, respectively. Nonetheless, the Hausdorff hierarchy can be indexed through a proper choice of the aggregation index f. This will be clarified in a forthcoming paper. From a more intuitive (physical) perspective, condition (A3) can become valid when the sets are rather intertwined, and can be taken as an indication that, although always mathematically consistent, the clustering procedure itself at this level of the dendrogram becomes delicate, in particular, for inherently complex problems, such as that of clustering stock market companies.

[1] K. Fukunaga, *Introduction to Statistical Pattern Recognition* (Academic Press, San Diego, 1990).

[2] A. K. Jain, M. N. Murty, and P. J. Flynn, ACM Comput. Surv. **31**, 264 (1999).

[3] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data* (Prentice Hall, New York, 1988).

[4] A. Gersho and R. M. Gray, *Vector Quantization and Signal Processing* (Kluwer, Academic Publisher, Boston, 1992).

[5] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification* (John Wiley & Sons, New York, 2002).

[6] T. Hofmann and J. M. Buhmann, IEEE Trans. Pattern Anal. Mach. Intell. **19**, 1 (1997).

[7] N. Basalto, R. Bellotti, F. De Carlo, P. Facchi, E. Pantaleo, and S. Pascazio, Physica A **379**, 635 (2007).

[8] M. Ballerini, N. Cabibbo, R. Candelier, A. Cavagna, E. Cisbani, I. Giardina, V. Lecomte, A. Orlandi, G. Parisi, A. Procaccini, M. Viale, and V. Zdravkovic, Proc. Natl. Acad. Sci. U.S.A. **105**, 1232 (2008).

[9] R. Rammal, G. Toulouse, and M. A. Virasoro, Rev. Mod. Phys. **58**, 765 (1986).

[10] J. L. Kelley, *General Topology* (Van Nostrand, New York, 1955); Eric W. Weisstein *et al.*, *Pseudometric.* From MathWorld—A Wolfram Web Resource, http://mathworld.wolfram.com/Pseudometric.html

[11] A metric can always be introduced, in any (finite or infinite) set. The issue here is to understand whether such a metric is "physically" meaningful. A good choice makes the difference between a natural clustering procedure and an artificial one.

[12] Note that, depending on the choice of the measure on the parent space $\mathcal{S}$, Eq. (10) includes weighted averages as well.

[13] We are excluding the very particular case when δ itself is a pseudodistance and all the elements of the cluster are at a vanishing pseudodistance δ. Notice, however, that when one focuses on an iterative clustering algorithm, in all nontrivial cases, some clusters must eventually acquire a nonvanishing distance from themselves at some iteration.

[14] For noncompact sets whose subsets are uniformly distributed with linear density α, a subset is "small" if its size is $r \ll \alpha^{-1}$.

[15] M. Blatt, S. Wiseman, and E. Domany, Neural Comput. **9**, 1805 (1997).

[16] F. Hausdorff, *Grundzüge der Mengenlehre* (von Veit, Leipzig, 1914); republished as *Set Theory*, 5th ed. (Chelsea, New York, 2001).

[17] D. Huttenlocher, D. Klanderman, and A. Rucklige, IEEE Trans. Pattern Anal. Mach. Intell. **15**, 850 (1993).

[18] M. P. Dubuisson and A. K. Jain, in Proceedings of the 12th International Conference on Pattern Recognition (ICPR'94), Jerusalem, Israel (1994), pp. 566–568.

[19] M. Chavent and Y. Lechevallier, in *Classification, Clustering and Data Analysis*, edited by A. Sokolowsky and H. H. Bock (Springer-Verlag, Heidelberg, 2002), p. 53.

[20] N. Bolshakova and F. Azuaje, Signal Process. **83**, 825 (2003).

[21] X. Zhu, L. Shen, and T-S. P. Yum, *The 18th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC'07)* (2007).

[22] This is motivated by thinking of unbounded sets: the Hausdorff distance between a point $A=\{a\}$ and a set B having an accumulation point at infinity (such as a straight line) is indefinitely large, for no open r neighborhood $N_r(A)$ will ever contain B, no matter how large r.

[23] K. Kaneko, Phys. Rev. Lett. **63**, 219 (1989); Physica D **41**, 137 (1990); **75**, 55 (1994).

[24] We emphasize that the symmetries of a data set are, in general, not inherited by the clusters, that always depend on the (arbitrary) initial numbering of the data points. For instance, if two points A and B are equidistant from a third point C (but far from each other), then the linkage algorithm will link either A or B to C, depending on a biased external criterion (e.g., alphabetical order or initial numbering).

[25] M. R. Anderberg, *Cluster Analysis for Applications* (Academic Press, New York, 1973).

[26] J. C. Gower, Biometrika **53**, 325 (1966).

[27] R. N. Mantegna, Eur. Phys. J. B **11**, 193 (1999).

[28] R. N. Mantegna and H. E. Stanley, *Introduction to Econophysics* (Cambridge University Press, Cambridge, England, 2000).

[29] M. Bernaschi, L. Grilli, and D. Vergni, Physica A **308**, 381

(2002); L. Grilli, *ibid.* **332**, 441 (2004).
[30] V. Tola, F. Lillo, M. Gallegati, and R. N. Mantegna, J. Econ. Dyn. Control **32**, 235 (2008).
[31] E. J. Elton and M. J. Gruber, *Modern Portfolio Theory and Investment Analysis* (John Wiley and Sons, New York, 1995).
[32] L. Laloux, P. Cizeau, J.-P. Bouchaud, and M. Potters, Phys. Rev. Lett. **83**, 1467 (1999); V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. Nunes Amaral, and H. E. Stanley, *ibid.* **83**, 1471 (1999).
[33] J.-P. Onnela, A. Chakraborti, K. Kaski, and J. Kertesz, Eur. Phys. J. B **30**, 285 (2002).
[34] J.-P. Onnela, A. Chakraborti, K. Kaski, J. Kertesz, and A. Kanto, Phys. Rev. E **68**, 056110 (2003).
[35] M. Tumminello, T. Aste, T. Di Matteo, and R. N. Mantegna, Proc. Natl. Acad. Sci. U.S.A. **102**, 10421 (2005).
[36] M. Tumminello, F. Lillo, and R. N. Mantegna, Phys. Rev. E **76**, 031123 (2007).