Journal of Computational and Applied Mathematics 344 (2018) 521–531

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

journal homepage: www.elsevier.com/locate/cam

A novel scale-invariant, dynamic method for hierarchical clustering of data affected by measurement uncertainty

Federica Vignati *, Damiano Fustinoni, Alfonso Niro

Politecnico of Milano, Energy Department, via Lambruschini, 4 –20156 Milan, Italy

ARTICLE INFO

Article history:
Received 31 May 2017
Received in revised form 30 May 2018

MSC:
62H30
68T99

Keywords:
Hierarchical clustering
Non-uniqueness
Proximity measure
Computational cost
Uncertainty

ABSTRACT

An enhanced technique for hierarchical agglomerative clustering is presented. Classical clusterings suffer from non-uniqueness, resulting from the adopted scaling of data and from the arbitrary choice of the function to measure the proximity between elements. Moreover, most classical methods cannot account for the effect of measurement uncertainty on initial data, when present.

To overcome these limitations, the definition of a weighted, asymmetric function is introduced to quantify the proximity between any two elements. The data weighting depends dynamically on the degree of advancement of the clustering procedure. The novel proximity measure is derived from a geometric approach to the clustering, and it allows to both disengage the result from the data scaling, and to indicate the robustness of a clustering against the measurement uncertainty of initial data.

The method applies to both flat and hierarchical clustering, maintaining the computational cost of the classical methods.

1. Introduction

With the advancements in computer science, the study of clustering has gained increasing interest in the last decades, due to its several technical applications, ranging from machine learning [1], to data mining [2], including a number of tools for scientific research [3], logistics [4], and everyday life [5].

Indeed, in the study of many engineering and physical phenomena, data appear to be conditioned by several parameters. When analytical models are not available, the correlation between multivariate data and the corresponding parameter configurations can be usually obtained only by means of numerical simulations or experiments. Unfortunately, the nature of this correlation may remain unknown. Moreover, it is sometimes observed that experiments performed with different operational conditions may produce similar results. The first step to understand the data underlying structure, therefore, may consist in determining the families of configurations which lead to comparable results.

Since the first works, e.g., [6], statistical clustering has been used to provide a classification of a set of multivariate data, based on some similarities among the members of the same set (termed *cluster*). Flat clustering methods, in particular, identify a given number of sets and assign each configuration to one of them. Algorithms for flat clustering are usually very fast. On the contrary, the quality of the results depends in general on a-priori choices, e.g., the number of clusters: the research of the optimal number of clusters requires additional computations. Moreover, flat clustering methods cannot distinguish between "close" and "closer" configurations. These limitations are overcome by hierarchical clustering: the clusters are recursively identified and partitioned into sub-cluster, each characterized by an increasing degree of similarity among the

* Corresponding author.
E-mail addresses: federica.vignati@polimi.it (F. Vignati), damiano.fustinoni@polimi.it (D. Fustinoni), alfonso.niro@polimi.it (A. Niro).

https://doi.org/10.1016/j.cam.2018.05.062

belonging elements. Despite the slightly larger computational cost, therefore, hierarchical clustering methods retrieve a larger number of informations. Classical hierarchical clustering methods can be either agglomerative (or *bottom-up*) or divisive (or *top-down*): the first ones start out with as many clusters as the number of data – i.e., all the clusters are singleton sets – and, at each iteration, they gather them together until a single cluster is built, containing all data. On the contrary, the second ones start out with a whole set containing all data, and recursively partition it into smaller subsets until the selected level of detail. For complete introductions and descriptions of statistic clustering, the reader is addressed to [7,8].

Despite the better insight in the data structure provided, also hierarchical clustering methods suffer from some limitations. The resulting clustering, indeed, is non-unique, as it depends on the method used to quantify the similarity between its elements [9,10], on the treatment adopted to deal with special cases [11] and, mostly, on the data scaling [12–14]. To overcome the latter limitation, several solutions have been proposed so far. A first attempt to produce scale-invariant clusters [15] was based on the so-called Mahalanobis distance [16]. The latter is derived from the Euclidean distance, but it is made scale-invariant by introducing a normalization with respect to the data covariance. However, as pointed out by later works, the use of the variance to normalize data may result in meaningless clusters [17]. One of the most effective technique to find robust, scale-invariant clustering is based on a geometric approach, the so-called "minimum volume ellipsoids method", which allows to both disengage the results from the data scaling and to control the significance of the results [18–20]. Unfortunately, the computational efficiency of these geometric methods is usually lower than non-geometric ones.

To fill this gap, in this paper we introduce a novel, enhanced technique for hierarchical agglomerative clustering. The method is based on a geometric approach and, at the same time, on the definition of a weighted asymmetric distance function. The advantages of the adopted function are threefold: on the one hand, it allows to disengage the resulting clustering from the data scaling. On the other hand, thanks to the definition of the distance function, it requires the same computational effort as non-geometric methods. Eventually, the proposed technique intrinsically manages the measurement uncertainty, if any, even if the statistical distribution of the error is unknown. Therefore, this method can be successfully applied to both fully deterministic data and to experimental results affected by errors. When experimental results are analyzed, indeed, the choice of a suitable, robust clustering procedure is particularly relevant. On the one hand, the numerical values of the data, indeed, depend on arbitrary choices, e.g the units of measurement, the use of dimensional or dimensionless variables, the normalization of the results with respect to a reference value. An effective clustering method, therefore, must be scale-invariant. On the second hand, experimental data are usually affected by the measurement uncertainty: for this reason, a robust procedure should be recommended. Previous works on the clustering of non-deterministic data often require to know in advance the probability density function of the data [21–23], which may be unknown.

This work stems from the results of a long-term experimental campaign carried out at the ThermALab of Energy Department of Politecnico di Milano on the enhancement of heat transfer in forced convection of air flows through rectangular channels by means of square ribs in large variety of geometrical configurations. The Nusselt number and the friction factor – indicators of thermal and hydraulic performances, respectively – were experimentally measured. The ribs enhance heat transfer by periodically deflecting streamlines, interrupting boundary layer growth and destabilizing the flow. All these effects bring about an early transition to turbulent regime or promote turbulence. Unfortunately, to force a flow through a ribbed channel at a given Reynolds number, a larger pumping power is required, resulting in lower hydraulic performances with respect to a smooth channel in the same conditions. For each flow Reynolds number, a large number of rib geometries and configurations were tested, since the program was aimed at investigating the optimal rib configuration, i.e., producing the best compromise between the heat transfer enhancement and the induced hydraulic losses. The results showed a large dispersion of the data, and the apparent lack of an underlying criterion. For this reason, a cluster analysis was first performed on experimental data, in order to determine a possible structure, by means of classical clustering methods. Both plain data – i.e., on the Nusselt number and on the friction factor resulting from the experiments – and the same data normalized with respect to the Nusselt number and on the friction factor of the reference configuration – the smooth channel – were analyzed, in order to highlight both the absolute performances of each configurations and the difference with respect to the reference case. Unfortunately, the results of the clustering analysis of non-normalized and of normalized data were completely different. This fact represented an unexpected additional difficulty, since it is commonly enough to use non-dimensional number – i.e., Reynolds and Nusselt numbers and the friction factor – in order to provide unique correlations in most thermo-fluid-dynamical problems. On the contrary, the obtained clustering was not unique, highlighting a sensitivity problem. As observed also in other scientific researches dealing with different metrics [24], it can become a hard task to understand a priori whether non-normalized or normalized data should be used for the clustering. Moreover, classical clusterings proved to be not helpful in the perspective of determining the channel performances for two additional reasons. On the one hand, the effect of the measurement uncertainty was not known. On the other hand, classical methods forced to attribute the same relevance to both the performance indicators, whereas it is known that, depending on the technical applications, either the thermal or the hydraulic aspect must be privileged. To overcome these limitations, the novel method has been devised.

The paper is structured as follows: Section 2 reports the formulation of the method. In particular, the novel function adopted to measure the proximity between elements is defined, and its physical and geometrical interpretation is provided. An example of clustering obtained by means of the proposed method is provided in Section 3, including a comparison with the results of a classical clustering procedure. Conclusions and final remarks are reported in Section 4.

2. Description of the method

2.1. Overview and geometrical formulation

For both agglomerative and divisive hierarchical clustering, it is mandatory to define some criteria to quantify the similarity among the elements of each cluster. In general, these criteria concern the following possibilities.

- A measure of the proximity between two elements (two configuration, two clusters or a configuration and a cluster), which depends only on a geometrical property, i.e., the metrics of the space which represent the data: in general, p-norms are used.
- The so-called *linkage*, i.e., a criterion to select a representative of each cluster, that is a method to associate to a set of points the coordinates of a single point only. This choice is non-trivial, as it is based on both statistical and circumstantial considerations, e.g. the physics of the problem under scrutiny.
- A scaling of the data. This point strongly influences the resulting clustering, and may depend significantly on the field of application.

To compute the hierarchical clustering of a dataset by means of agglomerative methods, only iterative procedures have been proposed so far [7]. Classical agglomerative methods will be hereafter defined "static" as they prescribe that, at each iteration, a matrix of distances between each couple of elements is computed, in accordance with the selected proximity measure. The so-called *inclusion criterion* required to build the clusters is that the couple of elements resulting in the least distance is merged together to form a new cluster. The process then continues until all the data are merged into a unique cluster. At each iteration, the level of dissimilarity between the two merged elements is given by the least value of the distances matrix. Of course, once the above parameters have been set, the calculation of the distances matrix is straightforward. The general schematic of a static clustering method is represented in the flowchart plotted in Fig. 1(a). The main drawback of static methods is that the clustering is very sensitive to the scaling, and quite sensitive with respect to both the metric and the linkage. If on one hand static methods result unsuitable for clustering experimental data due to this lack of robustness, on the other hand they represent a good starting point to develop enhanced, more robust and physically-oriented clustering methods.

The proposed clustering method is called "dynamic", and it is still an agglomerative, bottom-up technique, but adopts a similarity criterion which disengages the results from the data scaling. With reference to Fig. 1(b), the structure of the dynamic method partially follows the static methods'. In the geometric interpretation of the method, multivariate data are arranged in a Cartesian hyper-plane, whose axes represent each of the parameters. The dynamic method prescribes, at each iteration, to build a bounding box around each element (configuration or cluster). The bounding box is a hyper-rectangle centered on the element. Each side of the bounding box is parallel to one of the Cartesian axes, and its length is twice a given percentage of the value of the coordinate of the point along the considered direction. The inclusion criterion states that if an element is included in the bounding box of another element, then the two elements are merged in a cluster. In the proposed dynamic method, the level of dissimilarity between the two merged elements is the selected percentage. The value of the percentage is step-by-step increased at each iteration, starting from a generic "low" initial value. If, in a given range of percentages, more than one clustering are found, the percentage increment is reduced until only one new clustering is built.

The method is termed dynamic because the proximity is independent from the elements coordinates, and therefore the ordinate of the resulting dendrogram represents the real increasing relative similarity among the elements of the same cluster. Therefore, the increasing size of the bounding box drives the continuation of the agglomerative procedure.

A special treatment must be introduced to deal with the so-called *ties in proximity problem* [25,26]. This consists in the inclusion of multiple elements in one bounding box for the same level of dissimilarity, that is, one element can be merged with more than one other element at the same time. Unlike most static methods, the proposed dynamic method does not require to merge exactly two elements per iteration, but it allows to merge multiple points. This solution allows to avoid the use of arbitrary solutions to the ties in proximity problem or complex representations of the resulting clustering.

2.2. Application to experimental data

Considering the possible applications of the proposed method, i.e., the clustering of experimental data, the dynamical clustering presents two additional advantages. The first one is that it allows to distinguish between meaningful and spurious aggregations, the latter resulting from experimental errors. The measurement uncertainty on each point represents indeed the half-side of a bounding box itself. Therefore, all the clusters with an internal level of dissimilarity larger than the measurement uncertainty represent genuine families of configurations; on the contrary, a lower internal similarity level indicates only a generic "proximity" among some elements, which cannot be further quantified.

The second advantage concerns the nature of multivariate data, i.e., observation on a number of variables. Depending on the problem, it is sometimes useful to make the cluster more inclusive with respect to one or more variables, and more selective with respect to the other ones. The dynamic method allows to account for the effect of weighting of the diverse variables which make up the multivariate data. This results from the disengagement of the inclusion criterion from the data value, which introduces a new degree of freedom. The aspect ratio of the bounding boxes can be therefore modified in order to make the clustering more inclusive along a specific direction. The practical consequence is that the clustering can

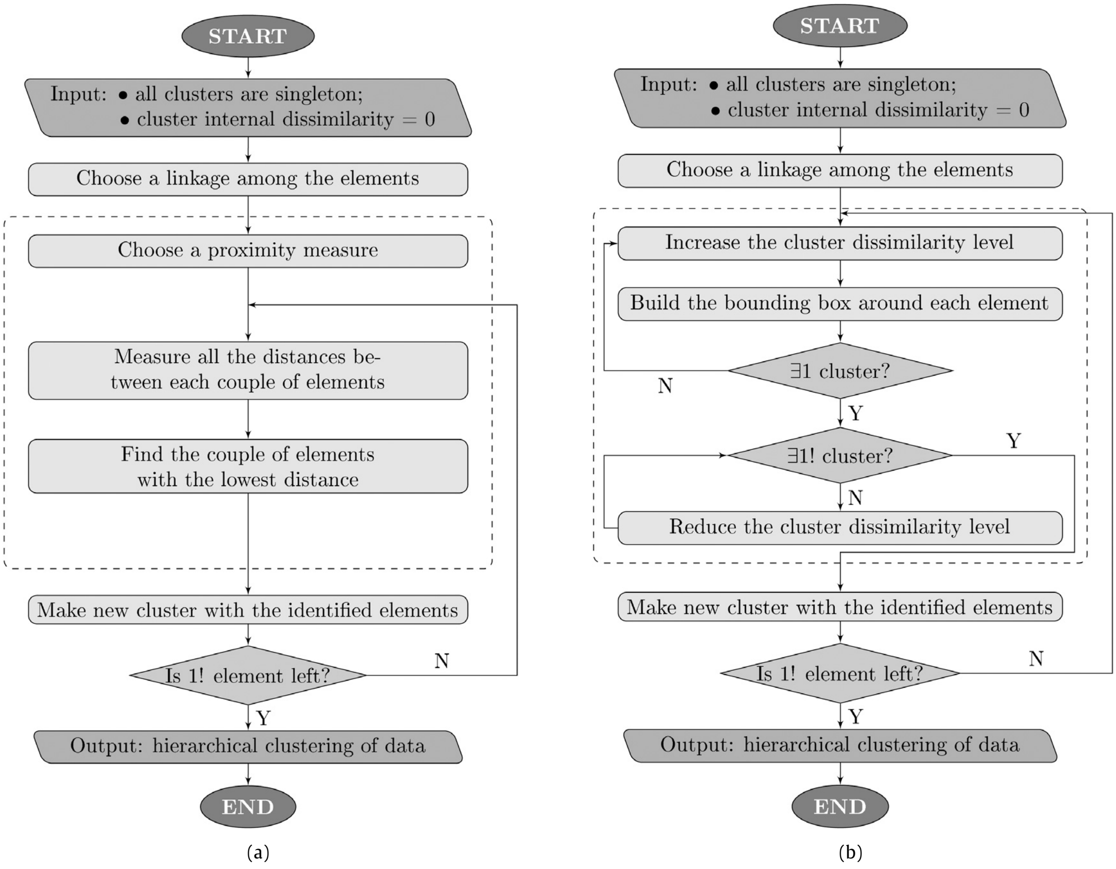

START
Input: • all clusters are singleton;
• cluster internal dissimilarity = 0
Choose a linkage among the elements
Choose a proximity measure
Measure all the distances between each couple of elements
Find the couple of elements with the lowest distance
Make new cluster with the identified elements
Is 1! element left?
N
Y
Output: hierarchical clustering of data
END
(a)
START
Input: • all clusters are singleton;
• cluster internal dissimilarity = 0
Choose a linkage among the elements
Increase the cluster dissimilarity level
Build the bounding box around each element
∃1 cluster?
N
Y
∃1! cluster?
Y
N
Reduce the cluster dissimilarity level
Make new cluster with the identified elements
Is 1! element left?
N
Y
Output: hierarchical clustering of data
END
(b)

Fig. 1. Flowchart of (a) a generic static and (b) the dynamic methods, respectively. The dashed boxes in the background highlight the difference between the two clustering approaches.

be tailored to different applications, depending on the relevance assigned to each variable. The choice of the aspect ratio numerical value – equal to 1, by default – is therefore left to the user, who can either rely on previous experience and design specifications, or, otherwise, perform sensitivity analyses.

2.3. Definition and considerations on the adopted proximity measure

The geometric approach is direct and robust; unfortunately, its computational cost is much larger than that required by standard clustering techniques. Therefore, it is strongly recommended to connect the dynamic method to classical ones, in order to take advantage of their computational efficiency. To link the hierarchical-clustering dynamic method to static ones, therefore, a new proximity measure is introduced, that will be hereafter termed "weighted asymmetric ∞-pseudometric" and indicated with the symbol d^*_∞. Let the multivariate data be represented by the elements P, Q depending on N variables in V, where $V = \mathbb{R}^N$. The new measure of proximity is defined as a function $d^*_\infty : V \times V \to [0, \infty)$ such that

$$\begin{aligned} d^*_\infty(P,Q) &= \min_{R=P,Q}\left(\max_i\left(\lim_{w_i\to R_i}\left|\frac{P_i-Q_i}{w_i}\right|\right)\right) \\ &= \min\left(\max_i\left(\lim_{w_i\to P_i}\left|\frac{P_i-Q_i}{w_i}\right|\right),\max_i\left(\lim_{w_i\to Q_i}\left|\frac{P_i-Q_i}{w_i}\right|\right)\right) \end{aligned} \tag{1}$$

where P_i, Q_i are the values of the ith variable making up the multivariate data, and w_i the corresponding value of a weight. The fractions are computed by means of a limit operator to account for the possibility of zero-value variables.

The function d^*_∞ is called "∞-pseudometric" since it derives from the Chebyshev distance L_∞, as they would be coincident if the difference between P_i and Q_i were not weighted with w_i. However, d^*_∞ does not represent a metric, as it satisfies only three out of the four conditions required for its definition [27]:

1. non-negativity: $d^*_\infty(P,Q) \geq 0$. It directly stems from the definition, being d^*_∞ the minimum between two non-negative numbers, i.e., the maxima over a set of absolute values.

2. identity of indiscernibles: $d^*_\infty(P, Q) = 0 \iff P = Q$. The proof is provided by biconditional introduction. If $P \equiv Q$, then $P_i = Q_i\ \forall i$, and therefore $|P_i - Q_i| = 0\ \forall i$. Since the denominator is always non-zero by definition, also $|P_i - Q_i|/w_i = 0\ \forall i$, including the maximum over i. Vice versa, to prove the claim in the opposite direction ad absurdum, let $P \neq Q$. There is at least one variable i such that $P_i \neq Q_i$, i.e., P_i and Q_i are separated. Therefore, $\exists i \;\big|\; |P_i - Q_i| > 0$ and $\exists i \;\big|\; |P_i - Q_i|/w_i > 0$, since $w_i > 0$ by definition. Therefore, $\max\limits_i |(P_i - Q_i)/w_i| > 0$ for both $w_i \to P_i$ and $w_i \to Q_i$ and, eventually, $d^*_\infty(P, Q) > 0$ is computed as the minimum between two non-zero elements, which is in contradiction with the claim that $d^*_\infty(P, Q) = 0$.
3. symmetry: $d^*_\infty(P, Q) = d^*_\infty(Q, P)$. For each i, $|Q_i - P_i| = |-1\,(P_i - Q_i)| = |-1|\,|P_i - Q_i| = |P_i - Q_i|$. Therefore,

$$\begin{aligned}
d^*_\infty(Q, P) &= \min\left(\max_i\left(\lim_{w_i\to Q_i}\left|\frac{Q_i - P_i}{w_i}\right|\right), \max_i\left(\lim_{w_i\to P_i}\left|\frac{Q_i - P_i}{w_i}\right|\right)\right) = \\
&= \min\left(\max_i\left(\lim_{w_i\to Q_i}\left|\frac{P_i - Q_i}{w_i}\right|\right), \max_i\left(\lim_{w_i\to P_i}\left|\frac{P_i - Q_i}{w_i}\right|\right)\right) = \\
&= \min\left(\max_i\left(\lim_{w_i\to P_i}\left|\frac{P_i - Q_i}{w_i}\right|\right), \max_i\left(\lim_{w_i\to Q_i}\left|\frac{P_i - Q_i}{w_i}\right|\right)\right) = \\
&= d^*_\infty(P, Q)
\end{aligned} \tag{2}$$

where the last equality is made possible by the commutativity of the minimum operator.

By contrast, the fourth condition, i.e., triangle inequality ($\forall R \in V$, $d^*_\infty(P, Q) \leq d^*_\infty(P, R) + d^*_\infty(R, Q)$), does not hold for d^*_∞. This can be verified by counterexample, e.g. by computing the relative proximity between the points $P(10, 1)$, $Q(1, 4)$ and $R(2, 2)$:

$$\begin{aligned}
d^*_\infty(P, Q) &= \min\left(\max\left(\left|\frac{10-1}{10}\right|, \left|\frac{1-4}{1}\right|\right), \max\left(\left|\frac{10-1}{1}\right|, \left|\frac{1-4}{4}\right|\right)\right) = \\
&= \min(3, 9) = 3
\end{aligned} \tag{3a}$$

$$\begin{aligned}
d^*_\infty(P, R) &= \min\left(\max\left(\left|\frac{10-2}{10}\right|, \left|\frac{1-2}{1}\right|\right), \max\left(\left|\frac{10-2}{2}\right|, \left|\frac{1-2}{2}\right|\right)\right) = \\
&= \min(1, 4) = 1
\end{aligned} \tag{3b}$$

$$\begin{aligned}
d^*_\infty(R, Q) &= \min\left(\max\left(\left|\frac{2-1}{2}\right|, \left|\frac{2-4}{2}\right|\right), \max\left(\left|\frac{2-1}{1}\right|, \left|\frac{2-4}{4}\right|\right)\right) = \\
&= \min(1, 1) = 1
\end{aligned} \tag{3c}$$

which give $d^*_\infty(P, Q) > d^*_\infty(P, R) + d^*_\infty(R, Q)$. Therefore, the fourth condition required by the definition of a metric is not satisfied, and d^*_∞ cannot be considered a metric: for this reason, it is defined "pseudometric".

Moreover, d^*_∞ is not a norm, as well. Indeed, d^*_∞ satisfies the first two conditions required by the definition of a norm (it is a positive-semidefinite function and it is sub-additive, as previously proven), but it is not homogeneous of degree 1 [27]. For all scalar $\lambda \in \mathbb{R}$, indeed $d^*_\infty(\lambda P, \lambda Q) \neq |\lambda|\, d^*_\infty(P, Q)$. From Eq. (1),

$$d^*_\infty(\lambda P, \lambda Q) = \min\left(\max_i\left(\lim_{w_i\to \lambda P_i}\left|\frac{\lambda P_i - \lambda Q_i}{w_i}\right|\right), \max_i\left(\lim_{w_i\to \lambda Q_i}\left|\frac{\lambda P_i - \lambda Q_i}{w_i}\right|\right)\right) \tag{4}$$

For the linearity of the limit operator, one has

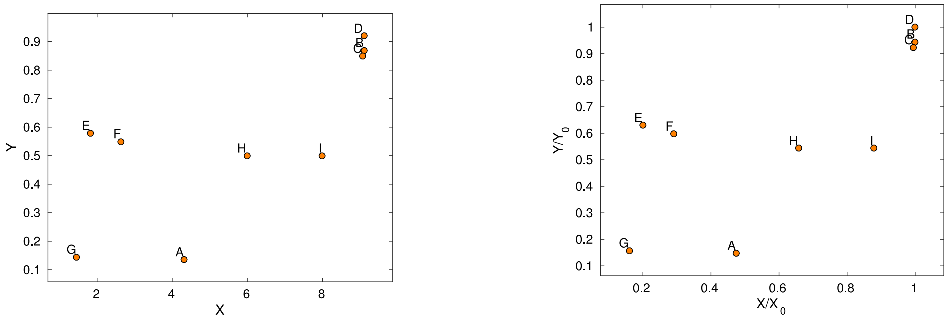
$$d^*_\infty(\lambda P, \lambda Q) = \min\left(\max_i\left(\lim_{w_i\to P_i}\left|\frac{\lambda P_i - \lambda Q_i}{\lambda w_i}\right|\right), \max_i\left(\lim_{w_i\to Q_i}\left|\frac{\lambda P_i - \lambda Q_i}{\lambda w_i}\right|\right)\right) \tag{5}$$

From Eqs. (4)–(5),

$$\begin{aligned}
d^*_\infty(\lambda P, \lambda Q) &= \min\left(\max_i\left(\lim_{w_i\to P_i}\frac{|\lambda|\cdot|P_i - Q_i|}{|\lambda|\cdot|w_i|}\right), \max_i\left(\lim_{w_i\to Q_i}\frac{|\lambda|\cdot|P_i - Q_i|}{|\lambda|\cdot|w_i|}\right)\right) \\
&= \min\left(\max_i\left(\lim_{w_i\to P_i}\left|\frac{P_i - Q_i}{w_i}\right|\right), \max_i\left(\lim_{w_i\to Q_i}\left|\frac{P_i - Q_i}{w_i}\right|\right)\right) = \\
&= d^*_\infty(P, Q)
\end{aligned} \tag{6}$$

which does not satisfy homogeneity of degree 1.

d^*_∞ is defined an "asymmetric" pseudometric despite it satisfies property 3.. The non-symmetry, indeed, is observed only in the geometrical interpretation of the distance, i.e., the one based on the construction of bounding boxes around the element. Indeed, for a given proximity $d^*_\infty(P, Q)$ between two elements P and Q, in general, only one element is included in the bounding box of the other one. The bounding box of an element, therefore, can be interpreted as a circle centered in the element, and whose radius, measured by means of the novel pseudometric, has value equal to d^*_∞. As a consequence, if



(a) Non-normalized data.

(b) Normalized data.

Fig. 2. Two-dimensional data for the toy-problem, arranged in the X-Y plane. Data are (a) non-normalized and (b) normalized with the largest coordinates, to fit in the unit square.

the boxes aspect ratio is varied by the user to make the clustering more inclusive in a specific direction, the bounding boxes become ellipsoids.

Due to the definition of the d^*_∞-pseudometric, the dynamic clustering can be computed by means of standard algorithm for agglomerative hierarchical clustering, which are based on the definition of a proximity measure. Therefore, no additional computational cost is required by the dynamic method.

2.4. *Application to flat clustering*

By analogy with other hierarchical methods, a flat clustering can be easily extracted from the hierarchical one resulting from the proposed procedure. To obtain the flat clustering, if the hierarchical clustering is represented by means of a dendrogram, the latter can be simply cut in correspondence of either the selected level of dissimilarity or the chosen number of clusters.

Obtaining a flat clustering from hierarchical ones is in general less efficient than directly computing it. However, the optimization of the flat clustering resulting from dedicated algorithms can require a non-negligible increment of the computational time. Moreover, flat clustering methods are based on the divisive approach, which is in general more sensitive to the data distribution. For these reasons, computing a flat clustering from hierarchical may turn out to be convenient in terms of both computational time and accuracy.

3. Results of the clustering method

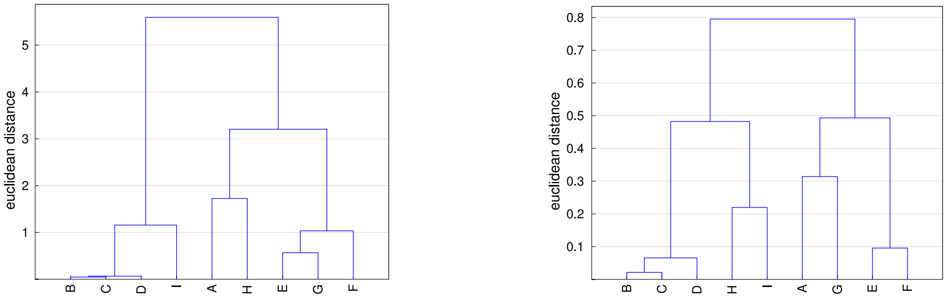
To represent the proposed evolution of the agglomerative hierarchical clustering method, a toy-problem is adopted. The configurations are represented by a generic set of points. In the example, each point is identified by two coordinates, i.e., X and Y, but the method can be generalized to spaces of any dimension. Both non-normalized and normalized data are analyzed. Fig. 2 shows the initial, unclustered points: Fig. 2(a) represents the non-normalized data, whereas Fig. 2(b) concerns the same data after they have been normalized to fit in the unit-square.

A static method for hierarchical agglomerative clustering is applied, using the euclidean norm to measure the distance and selecting the so-called *centroid linkage*, i.e., the representation of a cluster by means of its centroid. The resulting hierarchical clustering is represented in Fig. 3 by means of dendrograms, where each node represents a cluster containing all the points (reported in abscissa) that branch from it. The ordinate of every node indicates the dissimilarity among the cluster elements, i.e., their euclidean distance. Fig. 3(a) concerns non-normalized data, whereas 3(b) reports the dendrogram resulting from the clustering of the normalized data. As expected, the two dendrograms are totally different, that is, the static procedure is too sensitive with respect to the scaling of the data. The occurrence of this problem is systematically observed in all the static clustering methods, regardless of the adopted metric or linkage.

Conversely, the dynamic method produces stable results with respect to the scaling. Figs. 4 and 5 depict the construction of the bounding boxes around each element, for non-normalized and normalized data, respectively. It is evident that, when the "weighted asymmetric ∞-pseudometric" d^*_∞ is adopted, the data clustering is not sensitive with respect to the data scaling.

Fig. 6 shows the dendrograms resulting from the clustering of the set of points represented in Fig. 2 by means of the dynamic building of bounding boxes. The dendrogram of non-normalized data is shown in Fig. 6(a). The dendrogram resulting from the clustering of the same data scaled to fit in the unit square is reported in Fig. 6(b): it can be observed that the provided clustering is identical, indicating the stability of the dynamic method.

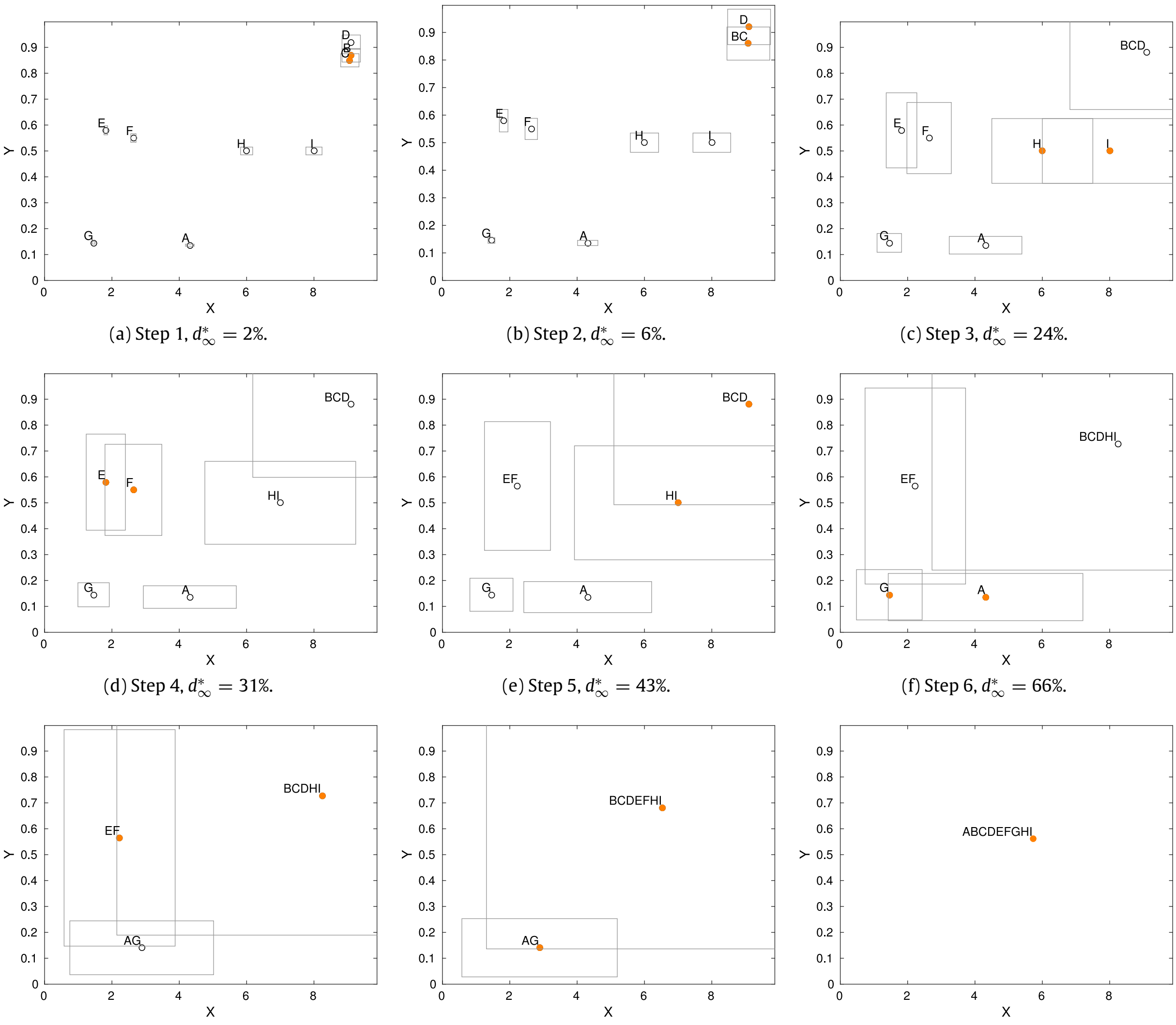
Since the measurement error is usually expressed in terms of a fraction of the reference numerical value of each measured variable, the dynamic method provides an immediate indication of which clusters are meaningful and which spurious. The



(a) Non-normalized data.

(b) Normalized data.

Fig. 3. Dendrograms resulting from the static clustering procedure applied to the toy-problem: (a) non-normalized and (b) normalized data, both coming from the same dataset. The relevant difference in the resulting dendrograms highlights the influence of the data scaling on a static clustering.



(a) Step 1, $d^*_\infty = 2\%$.

(b) Step 2, $d^*_\infty = 6\%$.

(c) Step 3, $d^*_\infty = 24\%$.

(d) Step 4, $d^*_\infty = 31\%$.

(e) Step 5, $d^*_\infty = 43\%$.

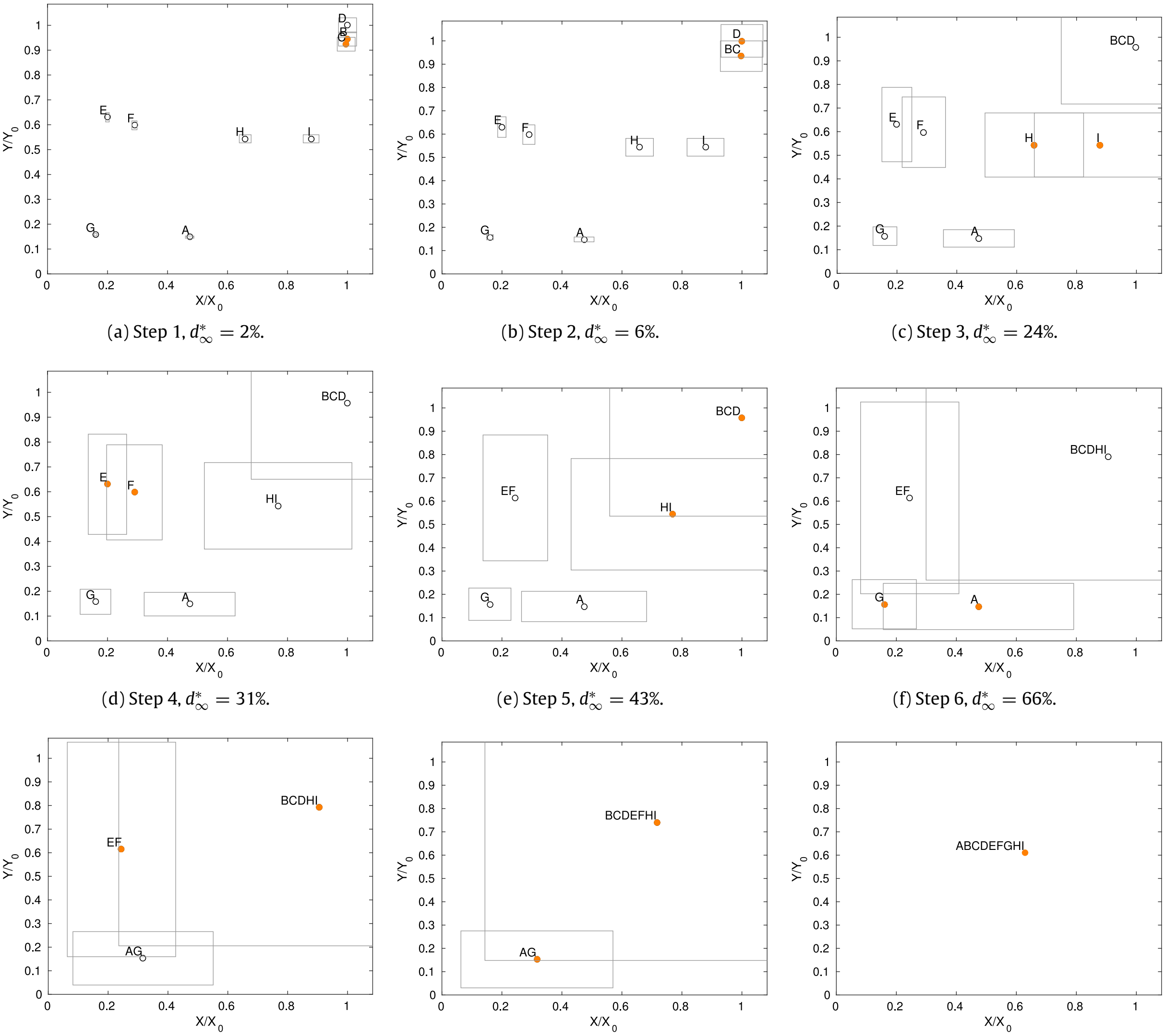
(f) Step 6, $d^*_\infty = 66\%$.

(g) Step 7, $d^*_\infty = 73\%$.

(h) Step 8, $d^*_\infty = 79\%$.

(i) Final step, $d^*_\infty = 79\%$.

Fig. 4. Step-by-step computation of the agglomerative hierarchical clustering by means of the dynamic method. The adopted dataset is the one depicted in Fig. 2, without normalization. The empty circles represent the generic elements. In each snapshot, the pair of merged elements is represented by a bullet. The rectangles with increasing size are the bounding boxes of each elements. When two elements cluster at a given iteration, they are substituted by the centroid of the cluster in next iterations.



(a) Step 1, $d^*_\infty = 2\%$.

(b) Step 2, $d^*_\infty = 6\%$.

(c) Step 3, $d^*_\infty = 24\%$.

(d) Step 4, $d^*_\infty = 31\%$.

(e) Step 5, $d^*_\infty = 43\%$.

(f) Step 6, $d^*_\infty = 66\%$.

(g) Step 7, $d^*_\infty = 73\%$.

(h) Step 8, $d^*_\infty = 79\%$.

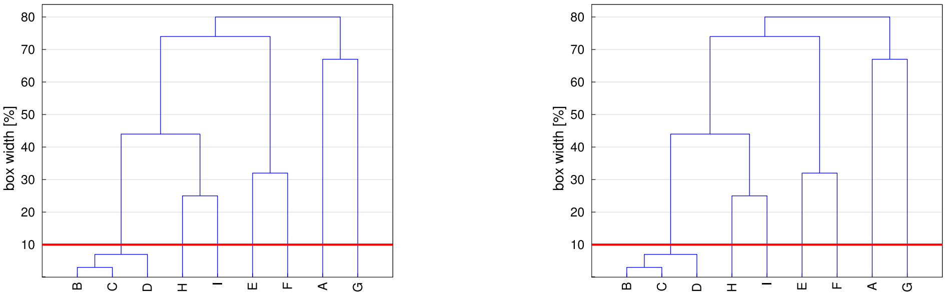
(i) Final step, $d^*_\infty = 79\%$.

Fig. 5. Step-by-step computation of the agglomerative hierarchical clustering by means of the dynamic method. The adopted dataset is the one depicted in Fig. 2, with a normalization to fit in the unit square. The empty circles represent the generic elements. In each snapshot, the pair of merged elements is represented by a bullet. The rectangles with increasing size are the bounding boxes of each elements. When two elements cluster at a given iteration, they are substituted by the centroid of the cluster in next iterations.

latter ones, indeed, are associated to an internal level of dissimilarity lower than the measurement error. Therefore, to provide a robust clustering with respect to the measurement uncertainty, the cluster whose internal dissimilarity (measured with d^*_∞) is lower than the uncertainty are treated as flat ones, since any further insight in their internal structure lies in the uncertainty interval of data.

In the described toy-problem, the measurement uncertainty value is set equal to 10%. In the corresponding dendrogram obtained by means of the dynamic method (see Fig. 6), the threshold corresponding to the measurement uncertainty is represented by a horizontal, thick line. With reference to Fig. 5(b), the level of internal dissimilarity of the cluster containing the configuration B, C and D is of 6%, lower than 10%: therefore, B, C and D are assumed to cluster together, but without any internal hierarchy.

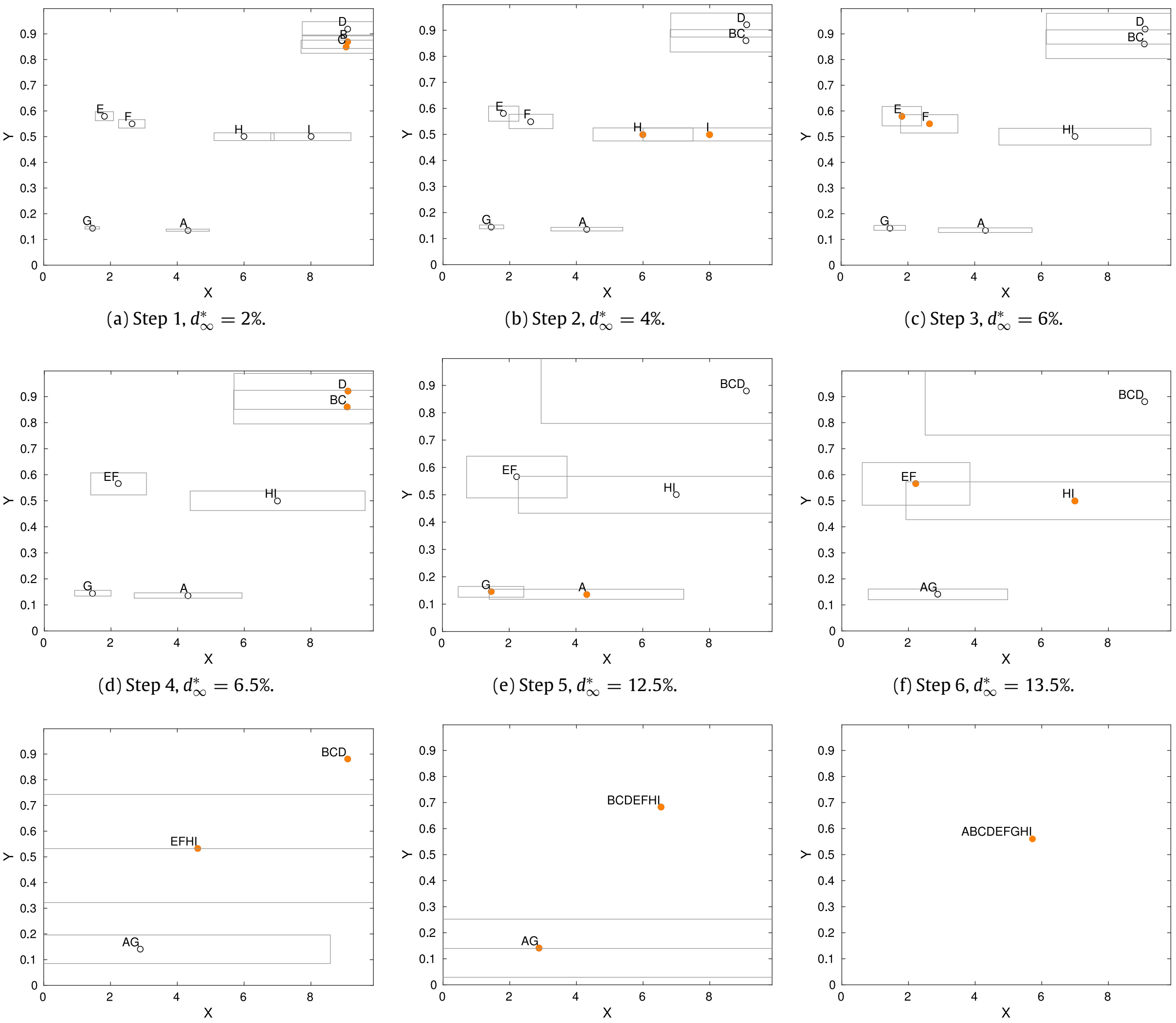
Figs. 7 and 8 show the procedure and the dendrogram resulting from the dynamic clustering of the same initial data as in Fig. 2, when the boxes aspect ratio is set equal to 5 in the horizontal direction. As it is visible, the bounding boxes appear to be stretched, and therefore less selective, in the X-direction. The adoption of non-unit aspect ratio for the bounding box can be useful in a number of cases, e.g., when the goal of the clustering is the optimization of an objective function. In this case, indeed, the diverse variables which make up the multivariate data have, in general, a different relevance. Therefore,



(a) Non-normalized data.

(b) Normalized data.

Fig. 6. Dendrograms resulting from the dynamic clustering procedure applied to the toy-problem: 6(a) non-normalized and 6(b) normalized data, both coming from the same dataset. The absolute coincidence of the resulting dendrograms highlights the capability of the novel dynamic method of disengaging the result from the data scaling. The horizontal line separates clustering with an internal dissimilarity level larger or lower than the measurement uncertainty.



(a) Step 1, $d^*_\infty = 2\%$.

(b) Step 2, $d^*_\infty = 4\%$.

(c) Step 3, $d^*_\infty = 6\%$.

(d) Step 4, $d^*_\infty = 6.5\%$.

(e) Step 5, $d^*_\infty = 12.5\%$.

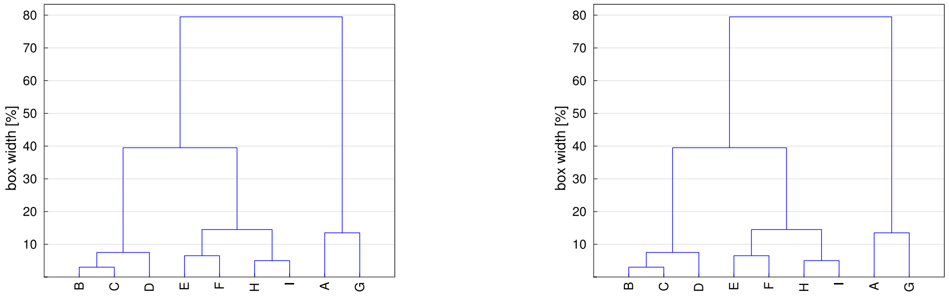
(f) Step 6, $d^*_\infty = 13.5\%$.

(g) Step 7, $d^*_\infty = 38.5\%$.

(h) Step 8, $d^*_\infty = 78.5\%$.

(i) Final step, $d^*_\infty = 78.5\%$.

Fig. 7. Step-by-step computation of the agglomerative hierarchical clustering by means of the dynamic method. The adopted dataset is the one depicted in Fig. 2, without normalization. The clustering is made more inclusive in the X-direction, with an aspect ratio of the bounding boxes set equal to 5.



(a) Non-normalized data.

(b) Normalized data.

Fig. 8. Dendrograms resulting from the dynamic clustering procedure applied to the toy-problem: 8(a) non-normalized and 8(b) normalized data, both coming from the same dataset. The boxes aspect ratio is set equal to 5. The absolute coincidence of the dendrograms resulting from the dynamic method is preserved also when the aspect ratio of the bounding box is different from one.

the lower the aspect ratio in a given direction, the more selective is the clustering with respect to the specific variable and, for this reason, the more sensitive is the optimum with respect to it. By tuning the aspect ratio, the user can directly tailor the method to fit to the problem under scrutiny.

Unfortunately, when different bounding box aspect ratios are devised from the unity, the comparison between the cluster internal dissimilarity level and the measurement uncertainty is less immediate than in the case of unity-aspect ratio. Therefore, the dendrograms do not report the horizontal line which represents the minimum meaningful internal dissimilarity of the clusters.

For brevity, the clustering procedure is illustrated only for non-normalized data, because the adoption of a different box aspect ratio does not affect the disengagement of the dynamic method from the data scaling. This can be observed by comparing the resulting dendrograms depicted in Fig. 8(a) and 8(b), concerning the non-normalized and the normalized data, respectively.

4. Conclusions and final remarks

A novel method for clustering of experimental data was presented and duly discussed. The new formulation stems from the observation that the clustering resulting from classical methods are not unique, since the computation of the proximity between the elements is not robust with respect to the data scaling and to the measurement error, if any.

The proposed clustering method is an agglomerative technique, based on the detection of the reciprocal inclusion of each configuration in another one's bounding box. The bounding box are built around the elements, and their size is dynamically associated to the level of advancement of the clustering procedure.

A new distance function was defined, i.e. the "weighted asymmetric ∞-pseudometric". The latter allowed to compute the dynamic clustering by means of standard algorithms, which are in general more efficient than the one based on the direct building of bounding boxes. The dynamic method was tested against static ones by means of a toy-problem. The example showed that the advantages are threefold. On the one hand, the dynamic method provides identical clustering for initial data with the same distribution, regardless of their scaling. On the other hand, information on the measurement error are not lost during the clustering procedure, but are considered a-posteriori, by means of a direct comparison between the uncertainty on initial data and the internal dissimilarity of the clusters. The a-posteriori retrieval of these information does not increase the required computational cost. Eventually, the aspect ratio of the boxes can be modified, in order to make the clustering more selective with respect to the most relevant variables. This possibility results from the disengagement of the clustering from the data scaling, and it allows to tailor the procedure to the nature of each specific dataset.

It is the authors' belief that further developments may include different aspects. The two most immediate improvements involve, on the first hand, the derivation of a technique for the evaluation of the most suitable bounding box aspect ratio, in the perspective of providing a complete method to any user. On the other hand, the integration of the proposed clustering technique with pattern identification should be worked out, in particular for making data analysis more automatic and robust.

References

[1] A.K. Jain, M.N. Murty, P. Flynn, Data clustering: A review, ACM Comput. Surv. 31 (3) (1999) 264–323.

[2] S.K. Murthy, Automatic construction of decision trees from data: a multi-disciplinary survey, Data Min. Knowl. Discov. 2 (1998) 345–389.

[3] G. Hormiga, Cladistics and the comparative morphology of linyphiid spiders and their relatives (Arneae Araneoidea, Linyphiidae), Zoological J. Linnean Soc. 111 (1) (1994) 1–71.

[4] J.F. Campbell, Hub location and the p-hub median problem, Oper. Res. 44 (6) (1996) 923–935.

[5] E. Pampalk, S. Dixon, G. Widmer, On the evaluation of perceptual similarity measures for music, in: Proc. Sixth Internat. Conf. on Digital Audio Effects (DAFx-03), 2003, pp. 7–12.
[6] R.R. Sokal, P.H.A. Sneath, Principles of Numerical Taxonomy, W.H. Freeman and Company, San Francisco, 1963.
[7] R.A. Johnson, D.W. Wichern, Applied Multivariate Statistical Analysis, Pearson Education, New York, 1990.
[8] T.W. Anderson, An Introduction to Multivariate Statistical Analysis, John Wiley & Sons, New York, 1984.
[9] F. James Rohlf, Robert R. Sokal, The description of taxonomic relationships by factor analysis, Syst. Zoology 11 (1) (1962) 1–16.
[10] W.H.E. Day, H. Edlesbrunner, Investigation of proportional link linkage clustering methods, J. Classification 2 (1985) 239–254.
[11] A. Fernández, S. Gómez, Solving non-uniqueness in agglomerative hierarchical clustering using multidendrograms, J. Classification 25 (2008) 43–65.
[12] A.K. Jain, Data clustering: 50 years beyond K-means, Pattern Recognit. Lett. 31 (2010) 651–666.
[13] J. Kleinberg, An impossibility theorem for clustering, in: Advances in Neural Information Processing Systems, vol. 15, MIT Press, Boston, 2002, pp. 446–453.
[14] S. Fortunato, Community detection in graphs, Phys. Rep. 486 (2010) 75–174.
[15] H.P. Friedman, J. Rubin, On some invariant criteria for grouping data, Amer. Statist. Assoc. J. (1967) 1159–1178.
[16] P.C. Mahalanobis, On the generalized distance in statistics, Proc. Natl. Inst. Sci. India 2 (1936) 49–55.
[17] E.M. Knorr, R.T. Ng, R.H. Zamar, Robust space transformations for distance-based operations, in: Proceedings of the 7th International Conference on Knowledge Discovery and Data Mining, 2001, pp. 126–35.
[18] J. Jolion, P. Meer, S. Bataouche, Robust clustering with applications in computer vision, IEEE Trans. Pattern Anal. Mach. Intell. 13 (8) (1991) 791–802.
[19] M. Kumar, J.B. Orlin, Scale-invariant clustering with minimum volume ellipsoids, Comput. Oper. Res. 35 (2008) 1017–1029.
[20] Rajesh N. Davé, Raghu Krishnapuram, Raghu krishnapuram robust clustering methods: a unified view, IEEE Trans. Fuzzy Syst. 5 (2) (1997) 270–293.
[21] B. Jiang, J. Pei, Y. Tao, X. Lin, Clustering uncertain data based on probability distribution similarity, IEEE Trans. Knowl. Data Eng. 25 (4) (2013) 751–763.
[22] H.P. Kriegel, M. Pfeifle, Density-based clustering of uncertain data, in: Proceedings of the 11th ACM KDD Conference pn Knowledge Discovery in Data Mining, 2005, pp. 672–677.
[23] C.C. Aggarwal, P.S. Yu, A survey of uncertain data algorithms and applications, IEEE Trans. Knowl. Data Eng. 21 (5) (2009) 609–623.
[24] J.L. Margot, A quantitative criterion for defining planets, Astron. J. 150 (6) (2015) 185–191.
[25] J. MacCuish, C. Nicolaou, N.E. MacCuish, Ties in proximity and clustering compounds, J. Chem. Inf. Comput. Sci. 41 (2001) 134–146.
[26] A.K. Jain, R.C. Dubes, Algorithms for Clustering Data, in: Prentice Hall Advanced Reference Series, Englewood Cliffs, NJ, 1998.
[27] W. Rudin, Functional Analysis, McGraw-Hill Science/Engineering/Math, 1991.