Physica A 593 (2022) 126933

Contents lists available at ScienceDirect

Physica A

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

Statistically validated hierarchical clustering: Nested partitions in hierarchical trees

Christian Bongiorno [a], Salvatore Micciché [b,*], Rosario N. Mantegna [b,c]

[a] Laboratoire de Mathématiques et Informatique pour les Systèmes Complexes, CentraleSupélec, Université Paris Saclay, 3 rue Joliot-Curie, 91192 Gif-sur-Yvette, France

[b] Dipartimento di Fisica e Chimica - Emilio Segrè, Università degli Studi di Palermo, Viale delle Scienze, Ed. 18, 90128, Palermo, Italy

[c] Complexity Science Hub Vienna, Josefstädter Strasse 39, 1080 Vienna, Austria

ARTICLE INFO

Article history:
Received 24 August 2021
Received in revised form 6 January 2022
Available online 20 January 2022

Dataset link: https://github.com/shimo-lab/pvclust

Keywords:
Hierarchical trees
Clusters
Partitions
Multivariate series

ABSTRACT

We develop an algorithm that is fast and scalable in the detection of a nested partition extracted from a dendrogram that is obtained from hierarchical clustering of a multivariate series. Our algorithm provides a *p*-value for each clade observed in the hierarchical tree. The *p*-value is obtained by computing many bootstrap replicas of the dissimilarity matrix and by performing a statistical test on each difference between the dissimilarity associated with a given clade and the dissimilarity of the clade of its parent node. We prove the efficacy of our algorithm with a set of benchmarks generated by a hierarchically nested factor model. We compare results obtained by our algorithm with those of Pvclust. Pvclust is a widely-used algorithm pursuing a global approach originally developed in the context of phylogenetic studies. In our numerical experiments, we focus on the role of multiple hypothesis test correction and the robustness of the algorithms to inaccuracies and errors of datasets. We verify that our algorithm is much faster than Pvclust algorithm and has a better scalability both in the number of elements and in the number of records of the investigated multivariate set. We also apply our algorithm to two empirical datasets, one related to a biological complex system and the other related to financial time-series. We prove that the clusters detected by our methodology are meaningful with respect to some consensus partitioning of the two datasets.

1. Introduction

Hierarchical clustering (HC) is a popular data analysis procedure grouping elements of a set into a hierarchy of clusters [1]. It is widely used in many research fields. Examples include computational biology [2], genomics [3], neuroscience [4–6], psychology [7], finance [8–10] and economics [11], as well as social sciences [12,13]. Once a dissimilarity (or similarity) measure between elements is defined, and a clustering procedure is selected, the hierarchical clustering algorithm is fully defined. The algorithm is deterministic and it is providing a hierarchical tree (also called dendrogram) as an output. However, the generation of a dendrogram does not mean that one also obtains a parsimonious hierarchical nested partition as an output of the HC.

* Corresponding author.
E-mail addresses: christian.bongiorno@centralesupelec.fr (C. Bongiorno), salvatore.micciche@unipa.it (S. Micciché), rosario.mantegna@unipa.it (R.N. Mantegna).

https://doi.org/10.1016/j.physa.2022.126933
0378-4371/

Historically, the simplest and most popular way to obtain a partition from a hierarchical tree was to cut the dendrogram at a fixed dissimilarity (or similarity) value. With this simple approach, such a cut is defining the composition of clusters. They are selected by considering the groups of elements linked in the tree at a dissimilarity value smaller than the threshold value. Several methods have been proposed to select an optimal dissimilarity threshold, as the one discussed in Ref. [14]. Other authors have proposed to determine the most appropriate partition of elements by obtaining its number of clusters with different approaches. Examples are methods based on the gap statistics [15], squared error [16], connectivity [17], Dunn index [18], or silhouette width [19]. The R package `clValid` allows computing hard partitions (i.e. partitions where an element can belong only to a single cluster) with most of the previously cited methods [20]. The dynamical cut tree method provides a different approach, which allows a cut of the dendrogram at different dissimilarity levels [21].

In the context of phylogenetic studies, Felsenstein was one of the first to focus on the problem of obtaining a partition from a hierarchical tree by assessing the statistical significance of clusters obtained by HC [22]. Specifically, he proposed to associate a p-value to each clade of the hierarchical tree. In phylogeny such a p-value provides a piece of information on the evolutionary hypothesis associated with the formation of the clade. The method, that was used to estimate the p-value, was based on a bootstrapping procedure. Since the introduction of the original statistical procedure, a long debate has been ongoing in the statistical literature. Efron proposed a way to refine the test [23]. More recently, Shimodaira implemented the refinement of Efron [23], and developed the so-called approximately unbiased (AU) test based on bootstrap [24,25]. With the AU approach, he achieved a higher accuracy with respect to previously proposed statistical tests. An R-package with the implementation of the AU test, named *Pvclust*, was released in [26] and it is currently widely used in phylogenetic and genomic analyses.

It is worth noting that Felsenstein's approach is a global approach. It is assessing the statistical reliability of the presence of all clades (i.e. groups of elements differentiating above a given value of dissimilarity) in bootstrap replicas of the original data. For this reason, the method is quite slow for large systems and computational time could be so long that its application is unfeasible. Another problem of applicability of Pvclust to a large set of data concerns the aspects of multiple hypothesis test correction [27]. In fact, by repeating many times a statistical test to assess the statistical reliability of the observation of each clade one needs multiple hypothesis test correction. Currently, Pvclust algorithm does not perform multiple hypothesis test correction opening the way to the potential presence of several false positive in large systems. In our numerical investigations (see below), we verify that the presence or absence of a multiple hypothesis test correction may significantly affect Pvclust output.

In addition to the methods motivated by a phylogenetic approach, other methods have been proposed recently to associate a statistical significance to hierarchical partitions obtained by using HC. The primary interest for estimating such p-values originates in microarray expression studies but the methods proposed can be applied to any system investigated by HC. Examples of these approaches are the permutation test that quantifies the significance of each division of a hierarchical tree as proposed in [28] and the comparison of similarity measures with permutation based distribution of similarity between elements obtained under the null hypothesis of no cluster in the data [29].

Close in spirit to the work of [26], we propose an algorithm based on bootstrap resampling of a multivariate series that associates a p-value at each clade of a given hierarchical tree. Our algorithm gives good results when applied to benchmarks mimicking the complexity of hierarchically nested complex systems [8,9]. We call our algorithm statistically validated hierarchical clustering (SVHC). Specifically, for each pair of parent and children nodes in the hierarchical tree, we test the difference between the dissimilarity measure associated with a clade h and the dissimilarity measure associated with the clade of its parent node in the genealogy of the dendrogram. The statistical test we perform assumes as a null hypothesis that the dissimilarity of the parent node is equal to the dissimilarity of the children node, and observed difference might be only due to sample size random fluctuations. Our algorithm deals with dissimilarity values, therefore there is no need to recompute the hierarchical tree for each bootstrap replica of data, as it is required for Pvclust. This makes our approach considerably faster. In fact, our algorithm is primarily developed to investigate large systems. As such, we need to take into account family-wise error rate. For this reason, we perform our statistical tests by including multiple hypothesis test correction. Specifically, we apply the multiple hypothesis test correction based on the control of the false discovery rate (FDR) [30]. By selecting those clades that reject our null hypothesis, we identify a hierarchically nested partition classifying a certain number of elements of the investigated system.

In order to evaluate the performance of our algorithm, we test it with some synthetic benchmarks obtained by using a hierarchical factor model [31]. In our tests, we compare our results with the ones obtained with Pvclust with the AU option. To assess the importance of the multiple hypothesis test correction in large systems, we also apply the FDR multiple hypothesis test correction to the AU Pvclust output obtaining a modified AU Pvclust output corrected for the family-wise error (please note that this modification of the Pvclust output is not currently available in the distributed Pvclust package). Finally, we apply our algorithm and Pvclust (both in the original AU version and in the modified version with FDR correction) to a benchmark and to an empirical dataset. This dataset was originally obtained in [32] and was used as an example in the paper describing Pvclust [26].

Our algorithm has a very good performance when applied to benchmarks obtained from hierarchical factor models and is also highly informative in the analysis of empirical datasets. Furthermore our algorithm is very fast and scalable and therefore it can be used for large datasets that would otherwise need extremely long computational time to provide results. We show that SVHC is much faster than Pvclust, and the difference in computational time increases with the size

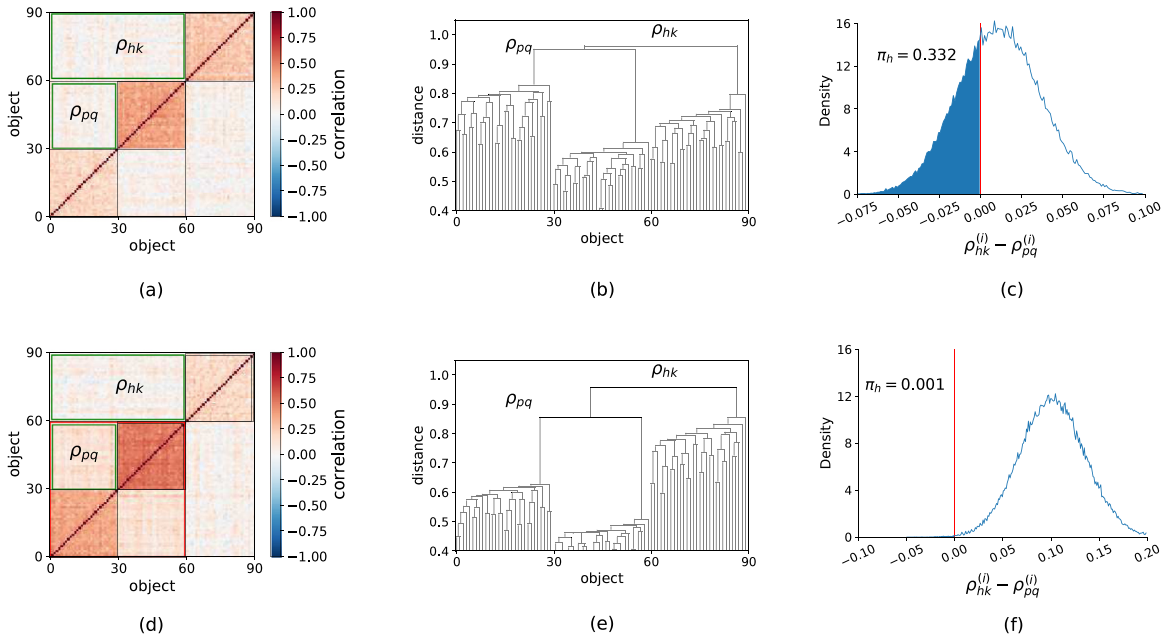

Fig. 1. Panels (a) and (d) show correlation matrices of slightly similar hierarchically nested benchmarks generated with $N = 90$ and $M = 200$. The elements are sorted according to the hierarchical tree of the average linkage HC algorithm. The green boxes highlight the correlation coefficients used to evaluate ρ_{hk} and ρ_{pq}. The red box of panel (d) indicates that cluster of clade h is statistically validated. The elements [0.29] belongs to the set of elements $\mathfrak{C}_p$, the elements [30.59] belongs to the set of elements $\mathfrak{C}_q$, the elements [60.89] belongs to the set of elements $\mathfrak{C}_k$, and the elements [0.59] belongs to the set of elements $\mathfrak{C}_h$. In panels (b) and (e) we show dendrograms of average linkage HC of correlation sample matrices of panels (a) and (d) respectively. In panels (c) and (f) we show the density function of $\rho^{(b)}_{hk} - \rho^{(b)}_{pq}$ to illustrate how the p-value of the null hypothesis $w_h \leq 0$ is estimated (in these examples we perform $N_b = 100{,}000$ bootstrap replicas). In the example of panels (a), (b), and (c) the null hypothesis $w_h \leq 0$ is not rejected whereas in the example of panels (d), (e), and (f) the null hypothesis is rejected. The univariate statistical threshold is set to 0.05 and we use the control of FDR as multiple hypothesis test correction.

N of the system. Therefore better performances of SHVC are even more appreciated when the program is applied to large systems. Similar results apply for the scalability of the computational time with the size M of the records. In particular, we numerically estimate a very low dependance from M in the case of SHVC. We believe that this is due to the fact that, as mentioned above, with our algorithm there is no need to recompute the hierarchical tree for each of the N_b bootstrap replica of data.

We also apply the SHVC algorithm to two real-world large datasets. The first one is a system of biological nature, i.e. a dataset of tissues affected by different types of lung cancer. The second system is a dataset of financial time-series of stocks traded in the US markets in a time period of 5 years from 31/03/2015 to 31/03/2020. We prove that the clusters detected by our methodology are meaningful with respect to some consensus partitioning of the two datasets.

2. Methods

2.1. Statistically validated hierarchical clustering

Let us consider a hierarchical tree obtained from the empirical investigation of a complex system of N elements, where each element has M records, so that a matrix data of order $N \times M$ is available. Let us assume that a clade of the hierarchical tree originating from node h has associated a dissimilarity measure ρ_{pq} (see panels (b) and (e) of Fig. 1). ρ_{pq} is the dissimilarity value where the p and q children clades join in h. In the next step of the agglomerative algorithm, the clade originating at h node joins the clade originating at k node and forms the clade l. The dissimilarity value defining the clade l is ρ_{hk}. The agglomerative procedure of the HC requires that the dissimilarity ρ_{pq} must be lower than ρ_{hk}. However, the empirically estimated ρ_{hk} and ρ_{pq} are affected by an estimation error for finite size data and therefore it is informative to device a methodology aiming at verifying whether their difference is statistically significant. This is the key aspect that we put at the core of our algorithm. In fact in our algorithm, for all pairs of parent–children clades we perform a statistical test of the null hypothesis of the type $\rho_{hk} \leq \rho_{pq}$. For example, when our null hypothesis is rejected for clades h and l, we consider that clade h is statistically distinct from clade l. The p-value associated with each test can

therefore be used to build up a nested partition where elements of statistically validated clades are elements of clusters of the partition. It should be noted that such a partition is in general a hierarchically nested partition where an element can be member of several nested clusters. We will show below that the p-value can be computed analytically under some restrictive assumptions and can be estimated numerically by computing bootstrap replicas of the dissimilarity matrix of the original data.

The worked example illustrated in Fig. 1 puts forward how the simple procedure of cutting dendrograms at a certain arbitrary level can be misleading. In fact, cutting the dendrogram in panel (b) at the clade w_h would give a partition that is not validated by our methodology, given the fact that we reject the null hypothesis $w_h \leq 0$. Hierarchical clustering always detects N-1 clusters that, by construction, can be ranked monotonically by their internal average correlation. With the statistical validation of clusters, we are able to reject such cases where the difference in internal correlation of a pair of nested clusters might be due to random fluctuations.

In our method, we consider a multivariate dataset $\mathbf{X} \in \mathbb{R}^{N\times M}$ with N elements and M records or attributes. We call $\mathbf{R} \in \mathbb{R}^{N\times N}$ Pearson's correlation matrix of $\mathbf{X}$. The matrix $\mathbf{R}$ has entries defined as

$$r_{ij} = \frac{M\sum x_{is}x_{js} - \sum x_{is}\sum x_{js}}{\sqrt{M\sum x_{is}^2 - \left(\sum x_{is}\right)^2}\sqrt{M\sum x_{js}^2 - \left(\sum x_{js}\right)^2}} \tag{1}$$

and we use it as a similarity measure. We perform HC by using a dissimilarity measure. Specifically, in this work we quantify the average linkage dissimilarity measure according to

$$\rho_{hk} = \frac{\sum_{i\in\mathfrak{C}_h}\sum_{j\in\mathfrak{C}_k}\left(1-r_{ij}\right)}{n_h\, n_k} \tag{2}$$

where $\mathfrak{C}_h$ and $\mathfrak{C}_k$ are the sets of nodes of clade h and of clade k in the hierarchical tree, respectively. Each set $\mathfrak{C}_k$ has size n_k. It is worth noting that our choice of Eq. (2) is just one possible choice of similarity measure. In fact our procedure works for a generic definition of similarity matrix and for different linkage algorithms. However, the present version of SVHC algorithm uses Eq. (2) and the average linkage HC.

2.2. Analytical derivation of the p-value

We derive an analytical expression of the p-value π_h associated with the null hypothesis $w_h := \rho_{hk} - \rho_{pq} \leq 0$. It is worth noting that the original sample value of correlation matrix gives $w_h > 0$. Our null hypothesis therefore aims to statistically verify whether the opposite condition to what we observe in the original sample estimation may occur. Our analytical derivation is valid under the hypothesis that $\mathbf{X}$ is a set of multivariate Gaussian random variables, m is a large but finite number, the hierarchical clustering procedure is the average linkage, and the dissimilarity measure is the one of Eq. (2). Our p-value is defined as the value of the cumulative distribution function of the stochastic variable $w_h = \rho_{hk} - \rho_{pq}$, observed when $w_h = 0$ where ρ_{ij} is estimated for each stochastic realization as in Eq. (2) by using the clade composition $\mathfrak{C}_i$ and $\mathfrak{C}_j$ of the original sample dendrogram. To obtain the analytical distribution of w_h we notice that the distribution of a Pearson's correlation coefficient can be well approximated by a normal distribution for large values of m under the assumption of Gaussian variables. Under all the above cited assumptions, w_h is the result of a weighted sum of normal random variables. Due to the central limit theorem, the probability distribution of w_h converges in probability to a normal distribution too. Since the elements of a correlation matrix are not independent variables, such sum is a weighted sum of correlated normal random variables. In particular, according to [33], the covariance between two elements r_{ij} and r_{lu} of a correlation matrix is

$$\begin{aligned}\xi_{(i,j),(lu)} = \frac{1}{2M}\{&\left[\left(r_{il} - r_{ij}r_{lj}\right)\left(r_{ju} - r_{jl}r_{lu}\right)\right] + \\ &\left[\left(r_{iu} - r_{il}r_{lu}\right)\left(r_{jl} - r_{ji}r_{il}\right)\right] + \\ &\left[\left(r_{il} - r_{iu}r_{ul}\right)\left(r_{ju} - r_{ji}r_{iu}\right)\right] + \\ &\left[\left(r_{iu} - r_{ij}r_{ju}\right)\left(r_{jl} - r_{ju}r_{ul}\right)\right]\}\end{aligned} \tag{3}$$

and therefore the variance of element r_{ij} is

$$\xi_{(ij),(ij)} = \frac{1}{M}\left(1 - r_{ij}^2\right)^2. \tag{4}$$

The expected value of the stochastic variable w_h is $\mathbb{E}[w_h] = \rho_{hk} - \rho_{pq}$. To estimate the variance of w_h we must consider the covariance among the elements of the correlation matrix. Let us notice that the elements of the correlation matrix that are used to compute the average dissimilarity ρ_{pq} are identified by the rectangular matrix of n_p and n_q elements of sets $\mathfrak{C}_p$ and $\mathfrak{C}_q$ respectively. Similarly the elements needed to compute ρ_{hk} are identified by a rectangular matrix of elements of sets $\mathfrak{C}_h$ and $\mathfrak{C}_k$ ($n_h\, n_k$ elements). By considering the definition of w_h, its variance is

$$\mathbb{S}[w_h]^2 = \frac{1}{\left(n_p n_q\right)^2}\sum_{i\in\mathfrak{C}_p}\sum_{j\in\mathfrak{C}_q}\sum_{l\in\mathfrak{C}_p}\sum_{u\in\mathfrak{C}_q}\xi_{(ij),(lu)}$$

$$+\frac{1}{(n_h n_k)^2}\sum_{i\in\mathfrak{C}_h}\sum_{j\in\mathfrak{C}_k}\sum_{l\in\mathfrak{C}_h}\sum_{u\in\mathfrak{C}_k}\xi_{(ij),(lu)}$$
$$-\frac{1}{n_p n_q n_h n_k}\sum_{i\in\mathfrak{C}_p}\sum_{j\in\mathfrak{C}_q}\sum_{l\in\mathfrak{C}_h}\sum_{u\in\mathfrak{C}_k}\xi_{(ij),(lu)} \tag{5}$$

Finally the p-value π_h is given by the cumulative distribution of a normal distribution with expected value $\mathbb{E}[w_h]$ and standard deviation $\mathbb{S}[w_h]$

$$\pi_h = \mathcal{P}(w_h \le 0) = \frac{1}{2}\left[1+\mathrm{erf}\left(-\frac{\mathbb{E}[w_h]}{\mathbb{S}[w_h]\sqrt{2}}\right)\right] \tag{6}$$

The analytical estimation of the p-value, that is obtained under the assumption of Gaussian multivariate process is very useful to assess the general validity of our approach but it is of little direct use for empirical investigations because several empirical multivariate processes deviates from the Gaussian statistics In Section 3.1 we will show numerical experiments performed starting from for Gaussian and Student's t multivariate variables in order to compare the p-values obtained starting from bootstrap replicas of data with the analytic result of Eq. (6).

In our algorithm, we indeed rely on numerical investigations of bootstrap replicas of a sample multivariate variable to obtain a p-value without making any assumption about the underlying stochastic multivariate process.

2.3. Numerical estimation of the p-value

Let us call $\mathbf{X}^{(b)} \in \mathbb{R}^{N\times M}$ a bootstrap copy of $\mathbf{X}$ obtained from sampling with replacement M columns of $\mathbf{X}$ matrix. Specifically, we define $\mathbf{X}^{(b)}$ entries as $x_{ij}^{(b)} = x_{i\ell}$, where $\ell = f_j^{(b)}$ is the j record of $\mathbf{f}^{(b)}$ vector of dimension M obtained with random sampling with replacement of the elements of $\{1,2,\ldots,M\}$. Let be $\mathbf{R}^{(b)} \in \mathbb{R}^{N\times N}$ correlation matrix obtained from the bth bootstrap replica $\mathbf{X}^{(b)}$. By using Eq. (2), for each bootstrap replica of the correlation matrix it is possible to compute a dissimilarity for pairs of sets of nodes. For example by considering the set of nodes $\mathfrak{C}_h$ and $\mathfrak{C}_k$ we can compute the set of dissimilarity $\left\{\rho_{hk}^{(1)}, \rho_{hk}^{(2)}, \ldots, \rho_{hk}^{(N_b)}\right\}$ and by considering the set of nodes $\mathfrak{C}_p$ and $\mathfrak{C}_q$ we can compute the dissimilarities $\left\{\rho_{pq}^{(1)}, \rho_{pq}^{(2)}, \ldots, \rho_{pq}^{(N_b)}\right\}$ where $b = 1,2,\ldots,N_b$ is the label of the bth bootstrap replica independently sampled. It is worth noting that such dissimilarities are evaluated by using Eq. (2) according to the composition of sets $\mathfrak{C}_x$ of the original sample dendrogram without computing a hierarchical tree for each bth bootstrap replica. The p-value associated to cluster h is defined as

$$\pi_h = \frac{\sum_{b=1}^{N_b}\delta(\rho_{hk}^{(b)} \le \rho_{pq}^{(b)})}{N_b} \tag{7}$$

where the operator $\delta(\cdot)$ is equal to 1 if the inequality is true, and 0 otherwise. In other words, the p-value is the fraction of times the inequality $\rho_{hk}^{(b)} > \rho_{pq}^{(b)}$ is not satisfied in the bootstrap replicas.

To correct for family-wise error we use the control of the FDR [30] that is implemented as follows: the $N-2$ clades of the system are arranged in increasing order of p-value, labeled as $\pi_1,\ldots,\pi_{N-2}$. We identify the largest integer $k_{\max}$ such that $\pi_k \le k\alpha/(N-2)$ and the clades corresponding to the first $k_{\max}$ p-values are used to define a nested partition of the elements. The statistical threshold α is the expected maximum proportion of false discovery allowed in our statistical test. In this work we set $\alpha = 0.05$.

It is worth noting that in our algorithm the most demanding procedure is the computation of bootstrap replicas of the multivariate dataset $\mathbf{X}$. Since bootstrap replicas are independent the one from the other, the algorithm can be easily parallelized in order to increase its efficiency.

To illustrate our procedure of numerical estimation of the p-value, we show two examples of statistical validation of a clade in Fig. 1. Specifically, we consider two slightly different sample hierarchical trees. They are shown in Fig. 1(b) and Fig. 1(e) respectively. The test aims to evaluate whether the elements from 0 to 59 (i.e. the clade originating at the node of the dendrogram characterized by the ρ_{pq} dissimilarity) are defining a group of elements statistically distinct from the set of all elements. In the top row of Fig. 1 we show three panels referring to the case when the null hypothesis $w_h \le 0$ is not rejected and therefore the clade of elements from 0 to 59 cannot be considered as a group of elements hierarchically distinct from all elements. According to the hierarchical tree, the clade of elements $\mathfrak{C}_p$ (elements [0.29]) and the clade of elements $\mathfrak{C}_q$ [30.59]) join together in the clade $\mathfrak{C}_h$, originating at $\rho_{pq} = 0.95$. Then the clade $\mathfrak{C}_h$ joins with clade $\mathfrak{C}_k$ (composed by the element [60.89]) at the node characterized by the dissimilarity $\rho_{hk} = 0.96$. In the sample tree, the dissimilarity value $\rho_{pq} = 0.95$ is smaller then $\rho_{hk} = 0.96$, as shown in Fig. 1(b). However, in spite of the structure observed in the hierarchical tree of the sample correlation matrix, the bootstrap analysis of $\rho_{pq}^{(b)}$ and $\rho_{hk}^{(b)}$ shows that the null hypothesis $w_h \le 0$ has associated a p-value equal to $\pi_h = 0.332$ and therefore cannot be rejected (see Fig. 1(c)). For this example, we therefore conclude that the set of elements [0.59] cannot be distinguished from the set of elements [0.89].

In the bottom row of Fig. 1 we show a slightly different example. Specifically, in this case the dissimilarity values are $\rho_{pq} = 0.86$ and $\rho_{hk} = 0.96$, as shown in Fig. 1(e). In other words, elements [0.59] are slightly more correlated than in the

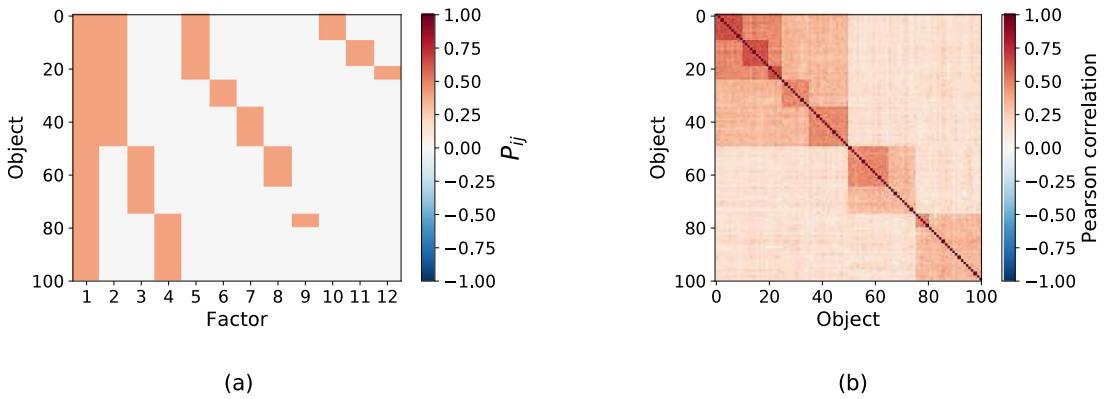

Fig. 2. (*a*) Example of factor loading pattern matrix. (*b*) Person's correlation matrix obtained from a multivariate dataset obtained by using the factor loading matrix shown in panel (a) with $S = 12$ and a factor score matrix $S \times M$ with $S = 12$ and $M = 500$ standardized independent Gaussian variables.

previous case. For this set of data, our approach concludes that the clade [0.59] is statistically distinct from the complete set [0.89] since the inequality $w_h \leq 0$ is observed only for 0.1% of our bootstrap replicas (see Fig. 1(f)). Therefore the null hypothesis $w_h \leq 0$ has associated a p-value $\pi_h = 0.00133$ and after performing the FDR multiple hypothesis test correction we reject it.

Our methodology is implemented in a Python code that is available at: https://github.com/cbongiorno/svhc.

2.4. Hierarchically nested benchmark

In our numerical experiments, we use benchmarks of multivariate datasets obtained with a nested factor model with S common factors [31]. Specifically, we simulate a multivariate dataset $\mathbf{X}$ of N elements with M records by using the equation

$$x_{ij} = \sum_{k=1}^{S} p_{ik}\, a_{kj} + u_i\, \varepsilon_{ij} \tag{8}$$

where $\mathbf{P}$ is the factor loading matrix of dimension $N \times S$ and $\mathbf{A}$ is a factor score matrix of dimension $S \times M$ with entries that are standardized independent Gaussian variables orthogonalized with a Gram–Schmidt algorithm. The vector u_i is called uniqueness and it is given by $u_i = \sqrt{1 - \sum_{j=1}^{S} p_{ij}^2}$. Finally, ε_{ij} is also a standardized Gaussian variable.

A nested factor model is able to generate a multivariate set characterized by a correlation matrix showing hierarchically nested blocks. For example, the multivariate dataset $\mathbf{X}$ obtained from the factor loading matrix $\mathbf{P}$ of Fig. 2(a) with $N = 100$ elements and $S = 12$ factors together with a factor score matrix $\mathbf{A}$ with $S = 12$ factors and $M = 500$ records has associated the correlation matrix shown in Fig. 2(b). With this choice of $\mathbf{P}$ each factor corresponds to a block on the correlation matrix, and an element i is a member of the block associated with the j factor if p_{ij} has a positive value. Specifically, the factor loading matrix of Fig. 2(a) has all positive elements equal to 0.4, and it produces twelve blocks of sizes $\{100,\ 50,\ 25,\ 25,\ 25,\ 10,\ 15,\ 15,\ 5,\ 10.\,10,\ 5\}$.

In numerical experiments discussed in Section 3, we are using the factor loading matrix of Fig. 2(a) and a number of modifications of it. However, we have tested the robustness of our results for many other factor loading matrices.

2.5. Comparing partitions

In this paper, we use the overlapping normalized mutual information (ONMI) [34] as the main comparison metric used to assess the similarity of two hierarchically nested partitions. ONMI is a variant of the normalized mutual information (NMI) [35]. NMI(x, y) measures the amount of information obtained about a partition x through the knowledge of another partition y, or vice-versa. NMI was defined to compare hard partitions. It was generalized to compare overlapping partitions, i.e. coverings, in Ref. [36]. Authors of Ref. [34] later proposed the modification of the ONMI metric that we are adopting in this paper. It is worth stressing that a hierarchical partition is a special case of an overlapping partition, with overlapping groups constrained to be nested.

ONMI measure is widely used but has some limitations [37]. We therefore also quantify the degree of similarity between the partitions obtained with the SVHC and Pvclust algorithms and the generating model by using other measures. These other measures are the Omega index [38] and the recently introduced element-centric (EC) similarity measure [37].

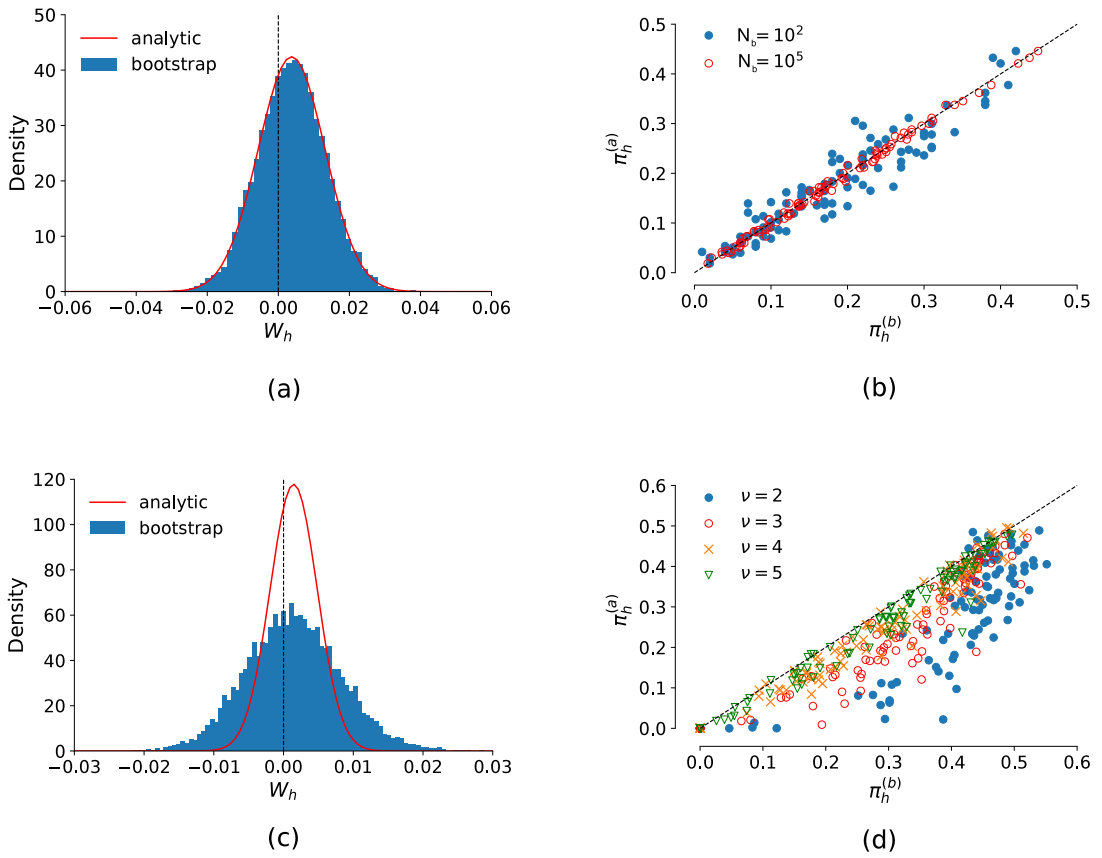

Fig. 3. Numerical experiments performed with uncorrelated multivariate random variables. (a) Histogram of the probability density function of w_h for a selected clade obtained by using $N_b = 10^4$ bootstrap replicas of a Gaussian multivariate random variable **X**. The red line is the analytical probability density function obtained under the hypothesis of Gaussian variables. (b) Scatter plot of p-values of analytic computation $\pi_h^{(a)}$ versus p-values obtained with bootstrap $\pi_h^{(b)}$ for two values of N_b. Each point refers to a clade of the HC of the multivariate series **X**. (c) Histogram of the probability density function of w_h for a selected clade obtained by performing $N_b = 10^4$ bootstrap replicas of a Student's t multivariate random variable **X** with $\nu = 2.00$. The red line is the analytical probability density function expected for Gaussian variable. (d) Scatter plot of $\pi_h^{(a)}$ versus $\pi_h^{(b)}$ for different values of the parameter ν of Student's t random variables. Each point refers to the p-value of a clade of the HC of the multivariate series **X**.

3. Results on synthetic data

3.1. Comparison between analytical and bootstrap based p-value

We first report a numerical experiment comparing the bootstrap based p-value $\pi_h^{(b)}$ of our algorithm, for Gaussian and Student's t multivariate variables, with the analytic p-value $\pi_h^{(a)}$ of Eq. (6) .

3.1.1. The Gaussian case

We numerically generate a set of multivariate uncorrelated Gaussian random variables **X**. The set has $N = 100$ elements with $M = 1000$ records each. Our numerical experiment is done for different values of the number of bootstrap replicas N_b. In Fig. 3(a), we show an example of the probability density function of the stochastic variable w_h obtained starting from bootstrap replicas of the **X** dataset for a selected clade h. We compare the pdf with the result of the analytical computation (red line).

In Fig. 3(b), we also show a scatter plot between $\pi_h^{(a)}$ and $\pi_h^{(b)}$ of **X** for two values of $N_b = (10^2, 10^5)$. It is worth noting that the p-values obtained with the bootstrap procedure converge to their analytical values for large values of N_b. In our numerical experiments, we do not detect any bias in the numerical estimation of bootstrap p-values for Gaussian multivariate data.

3.1.2. Student's t -distribution case

In order to study how crucial is the Gaussian hypothesis for the analytical estimation of the p-value, we compare analytical and numerical p-values in the case of a multivariate dataset $\mathbf{X}$ of uncorrelated t-distributed random variables of $N = 100$ elements with $M = 1000$ records each.

The probability density function of a t-distributed variable is

$$f(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}. \tag{9}$$

The parameter ν controls the finiteness of main moments. Specifically, for $1 < \nu \leq 2$ the variance is not defined, for $2 < \nu \leq 4$ the variance is finite, but the kurtosis is not defined and for $\nu > 4$ both variance and kurtosis are finite.

In Fig. 3(c), we show an example of the probability density function of the stochastic variable w_h obtained with bootstrap for a selected clade h compared with the analytic probability density function expected for Gaussian variables, see Eq. (6). We note that the analytical Gaussian p-value underestimates the variance of the stochastic variable w_h. This conclusion is confirmed by inspecting Fig. 3(d) where we show a scatter plot of $\pi_h^{(a)}$ and $\pi_h^{(b)}$ for different values of ν. It is worth noting that the discrepancy between analytical and numerical p-values is progressively reduced for large values of ν since the t-distribution converges to the Gaussian when $\nu \to \infty$ [39].

We conclude that numerical investigations are therefore essential when the probability density function of the multivariate dataset is not Gaussian.

3.2. Comparison with PVclust: size of the system

In another set of numerical experiments, we investigate the effectiveness of algorithms in retrieving the true hierarchical partition as a function of the number of elements N of the system. In these experiments, we again use a benchmark with 12 nested clusters. Specifically, we use the same benchmark of Fig. 2 modified by increasing the number of elements of each cluster and the total number of elements proportionally. Moreover, the number of records M of the time series is also increased proportionally to N according to $M = 5N$. We perform numerical experiments for systems with a number of elements $N = \{56, 100, 178, 316, 562\}$, where the different values of N present a logarithmic spacing.

Also in this case hierarchical partitions obtained with the SVHC algorithm describes quite well the true hierarchical partition for systems when N is ranging from 56 to 562 (see ONMI values in Fig. 4(a)). For low values of N the AU option of Pvclust ("Pvclust single") performs better than the modification of AU option of Pvclust with multiple hypothesis test correction ("Pvclust FDR"). The reverse is true for high values of N.

An analysis of the number of clusters detected by the algorithms is also highly informative (see Fig. 4(b)) . Also for this indicator the performance of the SVHC algorithm is very good for all values of N. For low values of N, the "Pvclust single" option detects a value that is very close to the true number of clusters. However, as the size N increases the number of detected clusters increases too. This bias of the "Pvclust single" option is probably due to the absence of a multiple hypothesis test correction. In fact, the number of statistical tests performed increases linearly with the size of the system. The results obtained by the "Pvclust FDR" option are different. For low values of N the number of clusters detected is less than the true number. This is probably due to the fact that the control of false positives is obtained at the expenses of not controlling the number of false negatives. For large values of N this limitation is progressively less important and the performance of the hierarchical partition of the "Pvclust FDR" output becomes very good. In summary, the "Pvclust single" output works well for small systems whereas the modified "Pvclust FDR" version is more appropriate for large systems.

3.3. Comparison with PVclust: size of the records

In the last set of numerical experiments, we investigate the performance of the two algorithms as a function of the number of records M of the elements of the multivariate time series. Specifically, we fix $N = 100$, $\lambda = 0$ and $M = \{20, 36, 63, 112, 200, 356, 632, 1125, 2000\}$ (again a set of values with logarithmic spacing). The results summarized in Fig. 4(c) show that SVHC is equivalent or slightly outperforms Pvclust in detecting the true hierarchical partition for high values of M. On the contrary, for low values of M the "Pvclust single" option performs better than SVHC. More details about the ability of the algorithms to detect the true partition can be obtained by inspecting Fig. 4(d). This figure plots the number of clusters detected by the algorithms. For low values of M, the "Pvclust single" option has a low number of false negatives but this performance is obtained at the expenses of a large number of false positives, whereas both SVHC and the "Pvclust FDR" version have a large number of false negatives. Depending on whether the most important aspect is statistical precision or statistical accuracy the most appropriate algorithm turns out to be different.

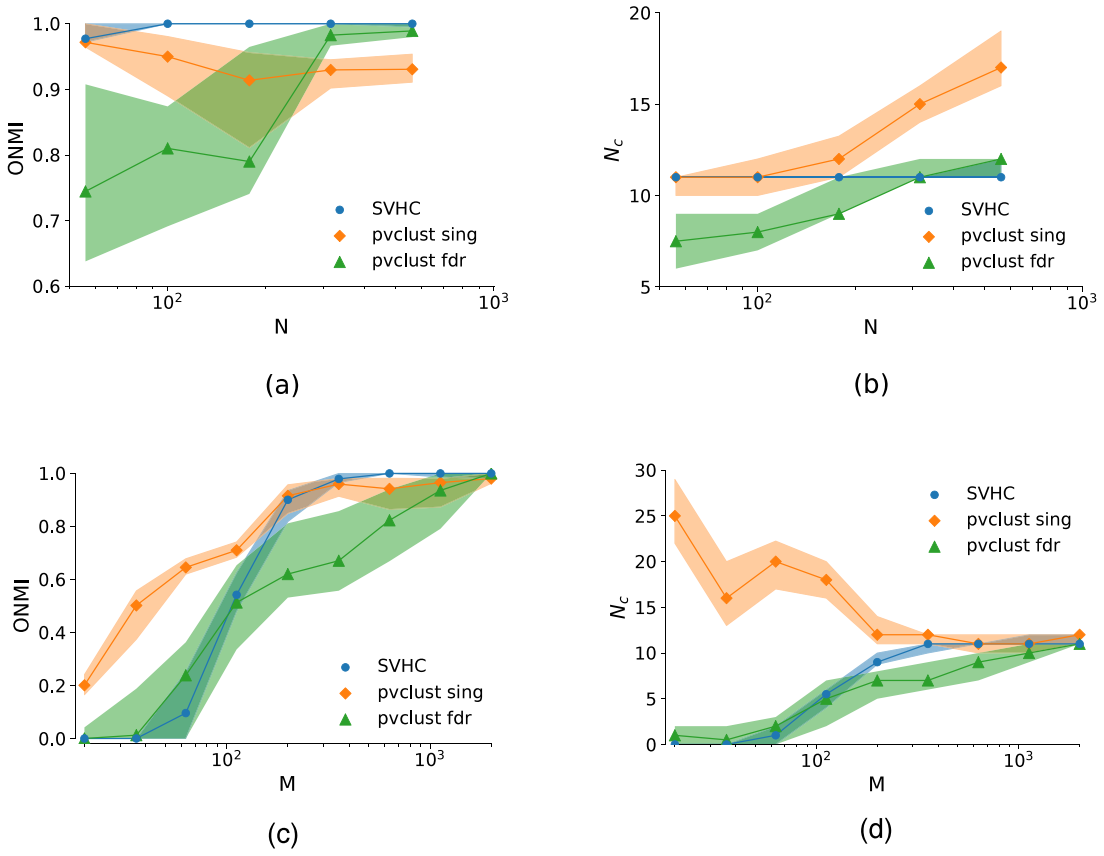

Fig. 4. Numerical experiments with a benchmark of the type shown in Fig. 2 for different values of the size of the system N and for different size of records M. (a) ONMI between the true hierarchical partition of the benchmark and the hierarchical partition obtained with SVHC, the AU option of Pvclust ("Pvclust single") and a modified output of the AU option of Pvclust with the FDR multiple hypothesis test correction ("Pvclust FDR") as a function of the system size N. (b) Number of statistically validated clusters detected by the algorithms as a function of the system size N. In all simulations shown in panels (a) and (b) $M = 5N$ and the univariate statistical threshold is set to 0.05. (c) ONMI between the true hierarchical partition of the benchmark and the hierarchical partition obtained with SVHC, "Pvclust single" or "Pvclust FDR" as a function of M. (d) Number of statistically validated clusters detected by the algorithms as a function of M. In all simulations shown in panels (c) and (d) $N = 100$. Points are the median computed in 100 independent realizations. The color band highlights the interval between the 25 and the 75 percentile. In our numerical experiments, we simulate 1000 bootstrap replicas both for the SVHC and the Pvclust algorithm.

3.4. Comparison with PVclust: computational time and scalabilty

As we mentioned above, although inspired by the work of Ref. [26], our algorithm deals with dissimilarity values, therefore there is no need to recompute the hierarchical tree for each bootstrap replica of data, as it is required for Pvclust. This makes our algorithm considerably faster. In this section we want to investigate this issue by performing numerical experiments with representative benchmarks.

In a first set of numerical experiments, we consider a benchmark with $S = 12$ nested clusters. Specifically, we use the same benchmark of Fig. 2 modified by increasing the number of elements of each cluster and the total number of elements proportionally. Moreover, the number of records M of the time series is also increased proportionally to N according to $M = 5N$. We perform numerical experiments for systems with a number of elements $N = \{56, 100.178, 316, 562\}$, where the different values of N present a logarithmic spacing. In this set of experiments noise is absent ($\lambda = 0$).

In the left panel of Fig. 5 we report the computational time for the SVHC and the Pvclust algorithm. From the left panel of Fig. 5 it is evident that SVHC is much faster than Pvclust, and the difference in computational time increases when the size of the system increases. We numerically estimate the time dependance of computational time τ as a function of N by fitting τ with a power law function $\tau(N) = c_0 + c_1 N^{\gamma}$ in the whole interval of N values. The fitting parameters are summarized in Table 1. The two algorithms present different values of the scaling exponent γ with the value for SVHC ($\gamma = 1.70$) smaller than the value of Pvclust ($\gamma = 1.94$). Moreover, coefficients c_0 and c_1 of τ observed for the SVHC algorithm are much smaller than the same coefficients for Pvclust (see Table 1) .

Computational time and scalability with M values are also very different between the two algorithms. The right panel of Fig. 5 shows computational time for SVHC and Pvclust when $N = 100$ for various values of M. We again fit the

Table 1
Parameters of the power law fit of the computational time τ (i) as a function of N for numerical experiments with a benchmark computed by setting $m = 5n$ (see Fig. 4(c)) and (ii) as a function of m for a benchmark of $n = 100$ objects (see Fig. 4(f)). In both cases the benchmark has 12 nested clusters as described in Fig. 2.

		c_0	c_1	γ
$\tau(n)$ with $M = 5N$	SVHC	2.4	3×10^{-4}	1.70
$\tau(n)$ with $M = 5N$	Pvclust	22	3×10^{-2}	1.94
$\tau(m)$ with $N = 100$	SVHC	2.89	1.07	5.5×10^{-4}
$\tau(m)$ with $N = 100$	Pvclust	123	1.01	0.76

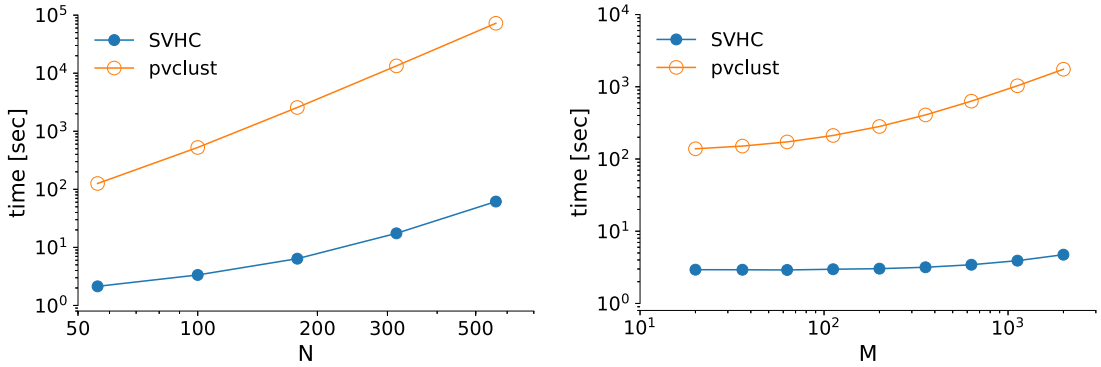

Fig. 5. Numerical experiments with a benchmark of the type shown in Fig. 2 for different values of the size of the system N and for different size of records M. The left panel shows the computational time needed for obtaining the hierarchical partition with SVHC (cyan) and the AU option of Pvclust ("Pvclust single") [and a modified output of the AU option of Pvclust with the FDR multiple hypothesis test correction ("Pvclust FDR")] as a function of the system size n. The right panel shows the same quantities when n is kept fixed at $N = 100$ and M can assume different values. Points are the median computed in 100 independent realizations. In our numerical experiments, we simulate $N_b = 1000$ bootstrap replicas both for the SVHC and the Pvclust algorithm.

computational time with the functional form $\tau(M) = c_0 + c_1 M^{\gamma}$ and the fitting parameters reported in Table 1. From the figure and from the parameters of fitting it is quite evident that the SVHC computational time has a very limited increase as a function of M. In fact, the fitting exponent γ for the SVHC algorithm is very close to zero. On the contrary, Pvclust computational time is characterized both by a sizeable exponent ($\gamma = 0.76$) and by a large minimum constant time ($c_0 = 123$).

These sets of numerical experiments confirm that SVHC algorithm is much faster and presents better scalable characteristics than Pvclust.

4. Results on empirical data

We now apply SVHC to a well known empirical dataset. As in previous numerical experiments, we apply SVHC in parallel with the application of Pvclust.

4.1. Adenocarcinoma in lung data

The first dataset we investigate is a set of microarray data of lung cancer tissues. Specifically, the dataset is the gene expression pattern of $N = 73$ tumor tissues belonging to 56 different patients. The data comprises information on $M = 915$ selected genes.

The dataset was originally collected in Ref. [32], a seminal paper that provided extensive and detailed support for the idea that gene expression-based classification of cancer, with tools such as hierarchical clustering, might become clinically useful for cancer of the lung. In particular, hierarchical clustering provided a classification of tumor tissues that was deemed to be extremely significant from a bio-medical point of view and eventually useful for addressing customized medical therapies.

Moreover, the dataset was used to provide an illustrative example of Pvclust performance in Ref. [26] where authors obtained the p-values of the branching points of the hierarchical tree of tissues.

Here we investigate the hierarchical tree of tissues ($N = 73$). In our investigation, both the SVHC and the Pvclust algorithms perform $N_b = 10{,}000$ bootstrap replicas. In Fig. 6 we show the results of our investigation. A square in the matrix highlights a cluster of elements characterized by a p-value rejecting the statistical null hypothesis.

In Fig. 7 we give a pictorial representation of the hierarchical membership of the 73 tumor tissues to the different clusters by using the two hierarchical clustering procedures of SVHC and the "Pvclust single" algorithm. Each tumor tissue

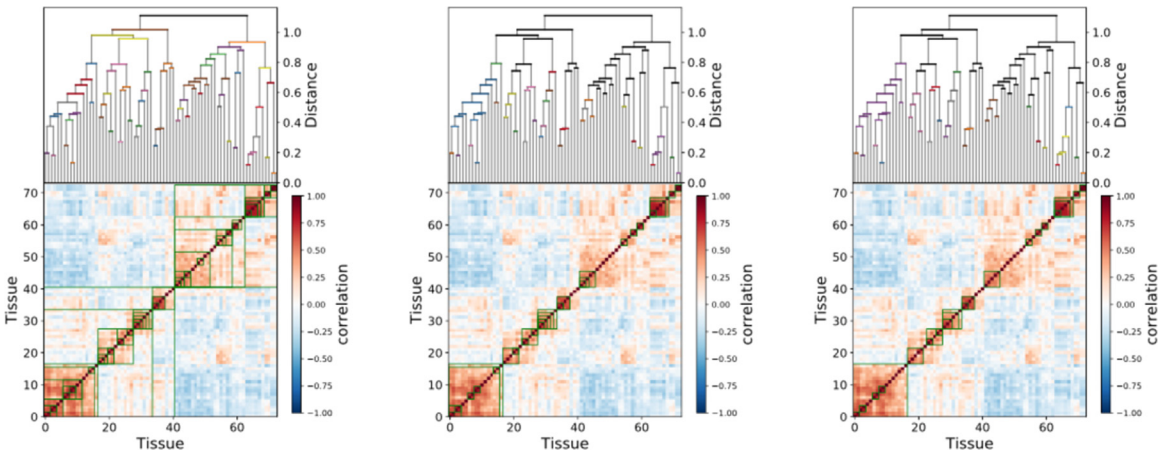

Fig. 6. Hierarchical trees (average linkage HC) and correlation matrices of lung tissues dataset [32]. In the correlation matrices we highlight hierarchically nested clusters detected by different algorithms with boxes. (*a*) SVHC, (*b*) "Pvclust single" (*c*) "Pvclust FDR".

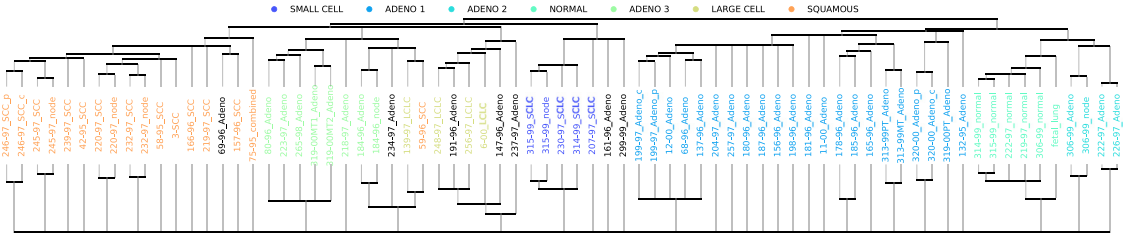

Fig. 7. Pictorial representations of the hierarchical nested partitions obtained with SVHC (upper part) and Pvclust single (lower part). The central part indicates the ids of the cancer tissues represented in a color code according to the reference partition of Ref. [32]. Black elements are non classified. Further explanations are given in the text.

is classified according to an expert partition discussed in Ref. [32]. The partition is summarized in the legend laying at the top of Fig. 7 and we will consider it as a reference partition. The dendrograms shown in Fig. 7 are derived from those of Fig. 6 by only considering the clades that are statistically significant. In fact, rather than showing 72 clades we have 52 validated clades in the upper dendrogram (SHVC) and 22 validated clades in the lower dendrogram ("Pvclust single"). The figure shows how most of the tissues grouped in a group of the reference partition are associated to statistically significant clades obtained with SHVC. When considering "Pvclust single" there are some cases in which tissues do not match any statistically validated clade. A typical example is given by the *ADENO 1* tissues that are mostly linked to the clade that groups all the elements and only in a few cases (5 tissues) to a statistically validated clade at a lower level.

We can put these considerations on a more quantitative basis. Specifically, we can search for each group of the reference partition the most similar cluster of the hierarchical nested partition obtained starting from both methods. As a similarity metric, we use the phi-correlation coefficient [40] defined as

$$\phi_i^{\max} := \max_{j\in\mathcal{D}} \phi_{ij} = \max_{j\in\mathcal{D}} \frac{N n_{ij} - b_i n_j}{\sqrt{b_i n_j (N-n_j)(N-b_i)}}, \tag{10}$$

where $\mathcal{D}$ is the set of statistically significant clusters, b_i is the number of tumor tissues classified in the group i on the reference partition, n_j is the number of tumor tissues that belong to the cluster j in the hierarchical partition, n_{ij} is the number of tumor tissues that are classified in the group i in the reference partition and belong to the group j in the hierarchical nested partition and N is the total number of tumor tissues. This metric is equivalent to a Pearson correlation coefficient when considering binary arrays. Here i labels the clusters of the reference partition, i.e. $i = 1, \ldots, 7$, while j labels the clusters of the SVHC (top) and the "Pvclust single" (bottom) partitions, respectively. In Table 2, for the 7 types of tumor tissues in the legend, we show the highest values of correlation among all clusters detected with the two considered methodologies. One can see that in most cases the SVHC methodology performs better that "Pvclust single". Significantly, the group *ADENO 1* of the reference partition has a phi-correlation value of 0.41 in the case of "Pvclust single" and a phi-correlation value equal to unity in the case of SHVC, in agreement with the outcomes of Fig. 7. A similar situation happened for the *SQUAMOUS* group of the reference partition that has the lowest value of phi-correlation in the case of "Pvclust single" and a phi-correlation value as high as 0.92 for the SHVC partitioning. In fact, when looking at the lower dendrogram of Fig. 7 we see that most tissues in this groups are linked to the clade that groups all the elements.

Table 2
Similarity ($\phi_j^{\max}$) between each reference group and the most similar cluster in the nested partition for both methods.

Classification	SVHC	PVCLUST single
NORMAL	1.00	1.00
ADENO 1	1.00	0.41
ADENO 3	0.83	0.77
ADENO 2	1.00	0.70
LARGE CELL	0.74	0.65
SMALL CELL	0.89	0.89
SQUAMOUS	0.92	0.30

We then aim to quantify an overall degree of similarity of partitions obtained by computing the ONMI between the hierarchical partitions obtained with the SVHC and Pvclust methodologies. We note that the hierarchical partitions obtained with SVHC are quite different from the ones obtained both from Pvclust single (ONMI=0.53) and PVclust FDR (ONMI=0.36). To understand such differences we can notice from Fig. 7 that the SVHC partition contains every cluster of "Pvclust single"; however, the SVHC partition is richer of information since it includes more nested clusters with a biological meaning. The difference with "Pvclust FDR" is even more pronounced being the latter much more parsimonious.

In addition, as already noted in the investigation of the synthetic benchmark, the computational time τ of SVHC (172 s) is significantly lower than Pvclust (4767 s).

4.2. Stocks in the Russel 1000 index

The second dataset we investigate is a set of data relative to the adjusted close prices of $N = 897$ highly capitalized stocks traded in US in a time period of 5 years, i.e. $M = 1260$ days. These stocks are continuously present in the Russel 1000 index from 31/03/2015 to 31/03/2020.

The reason for considering such dataset is threefold: firstly, we want to show that SVHC can provide meaningful insight also on non-biological systems; secondly, we want to show the performances on a quite large system, that cannot be analyzed in reasonable time with Pvclust. Thirdly, with this dataset we aim at better elucidating the types of real-world systems for which our methodology provides better results.

In fact, we have estimated that the computational time needed for performing the same analyses with PVclust would be approximately 105 days, compared to 2.19 h for the SVHC, with $N_b = 100{,}000$ bootstrap replicas of data. Such a large number of bootstrap replicas are necessary on large systems since the number of test increases linearly with the system size. Therefore, we do not present in this section any result obtainable with PVclust.

We performed this analysis in two different ways. In a first investigation we have considered the plain stock price returns defined as

$$r_{ij} = \frac{p_{i,j}}{p_{i,j-1}} - 1, \tag{11}$$

where p_{ij} is the adjusted close price of the stock i in the day j. In a second investigation we consider the excess returns, i.e. the plain return where the daily average return of all considered stocks (market mode) is subtracted [41]:

$$r_{ij}^{(m)} = r_{ij} - \frac{\sum_{i=1}^{N} r_{ij}}{N}. \tag{12}$$

It is known [42,43] that the returns without the market mode show a more pronounced hierarchical organization than the plain ones, and the existence of a genuine hierarchical structure is the underlying assumption for every hierarchical clustering method. We further confirm this statement by showing the distribution of distance among parents clades on both types of returns in the bottom-left panel of Fig. 8. From this analysis ii is clear that plain returns show distances between different levels less separated with respect to the case of excess returns. This implies that for our methodology it is more difficult to partition the system, since the clades all have small distances between each other, and therefore the detection of statistically significant differences would require an enormous computational effort.

When applying SVHC on both types of returns we confirm such an intuition. In fact, the partition of plain returns contains a small number of clusters ($n_c = 232$) of small size. On the contrary, the partition obtained starting from the excess returns have a richer hierarchical structure composed by a larger number of clusters ($n_c = 314$) of heterogeneous sizes, as shown in the top panels of Fig. 8.

In the case of equity stocks there is no unique reference partition. However, it widely known [8,44,45] that the stocks that belong to the same economic sector taxonomy typically have a high level of cross-correlation. In addition, the economic sector classification can be further refined by considering the industrial sector partition, which is naturally nested in the former. For this reason, we computed the ϕ_j^{max} distances between the detected partitions and both reference partitions mentioned above. In Table 3 we summarize $\phi_j^{\max}$ obtained by comparing both SVHC partitions with the

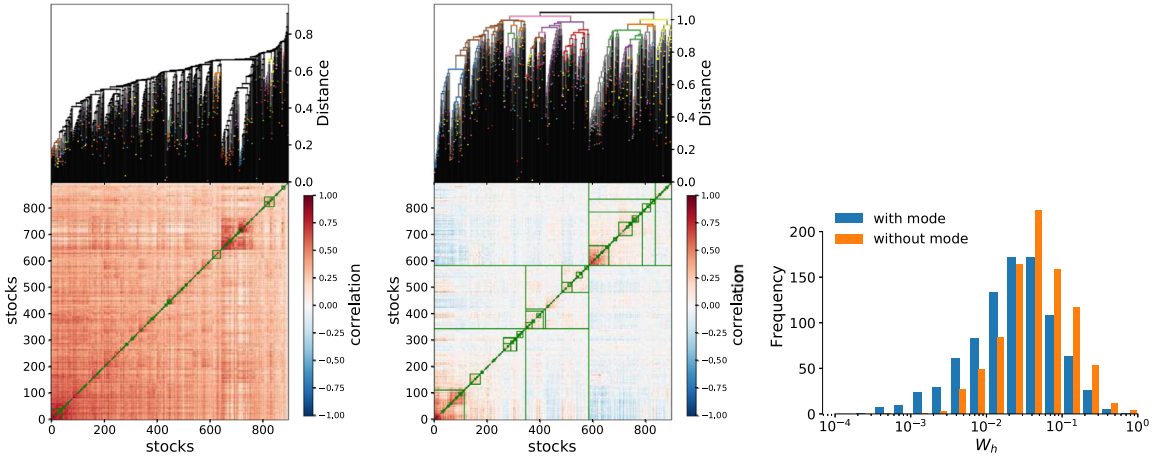

Fig. 8. The left and central panels show the partition obtained with SVHC for the Russell1000 system on the plain return and the excess returns, respectively. The right panel shows the histogram of distances among neighbors clades on the complete dendrogram for both types of returns.

Table 3
Similarity ($\phi_j^{\max}$) between each reference group (sector classification) and the most similar cluster in the nested partition for the SVHC partitioning obtained with and without market mode.

Sector	SVHC without market mode	SVHC with market mode
Basic industries	0.410	0.313
Capital goods	0.313	0.279
Consumer durables	0.311	0.355
Consumer non-durables	0.390	0.451
Consumer services	0.404	0.208
Energy	0.701	0.678
Finance	0.683	0.509
Health care	0.903	0.601
Miscellaneous	0.335	0.274
Public utilities	0.397	0.524
Technology	0.505	0.173
Transportation	0.694	0.573

reference partition given by the economic sectors. Although there are a few cases where SVHC obtained from the plain returns has an higher $\phi_j^{\max}$ value, overall the partition obtained on the excess returns has a higher similarity with the reference partition given by the economic sectors.

Let us now consider the industrial sectors reference partition which is composed of 119 groups. In Fig. 9 we summarize our results by providing an histogram of the $\phi_j^{\max}$ metric. Also in this case, we confirm that the partition obtained removing the market mode has a higher similarity with the reference partition when compared with the partition obtained from the plain returns. Moreover, overall we observed an ONMI between the reference partition and the one obtained from excess returns (ONMI=0.19) higher than the case when we consider the partition obtained from plain returns (ONMI=0.14). However, as a final comment, it worth noticing that both partitions have distributions of $\phi_j^{\max}$ which are significantly different from zero, this indicates that both partitions are able to extract genuine economic information about the equity market.

5. Discussion

Hierarchical clustering is a powerful data analysis tool widely used in many disciplines. Today the widely used Pvclust package is setting the standard for a statistical assessment of a specific hierarchically nested partition obtained from a given hierarchical tree. However, when we compared it to a large benchmark almost never it finds the real underlying partition. This bias seems connected with multiple comparison correction, in fact, when no multiple comparison correction is applied, the errors increases with the problem dimensionality. Differently, when a multiple comparison is applied, it can find the correct partition in a very large sample size limit; however, in a medium or small sample size regime, it lacks on power, missing many nested clusters. The second and most important drawback is the computational time needed to perform the hierarchical clustering estimation for all bootstrap replicas of the multivariate series. Computational time could be quite long for moderately large datasets, and therefore the Pvclust algorithm is of very limited use for big datasets.

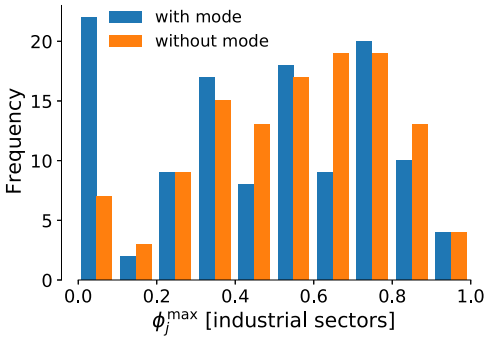

Fig. 9. Histogram of the similarity ($\phi_j^{\max}$) between each group of the reference partition (industrial classification) and the most similar cluster in the nested SHVC partitioning obtained with and without market mode.

In this paper, we introduce an algorithm that is quite effective in the detection of the true hierarchically nested partition of a multivariate series. In particular, it seems more powerful than Pvclust, which allow us to use the multiple comparison correction in almost any condition without a drastical loose of recall. Furthermore, our algorithm is much faster than the Pvclust algorithm and has a better scalability both in the number of elements and in the number of records of the investigated multivariate set.

Indeed, we show that SVHC is much faster than Pvclust, and the difference in computational time increases with the size N of the system. In fact, we numerically estimate the time dependance of computational time τ as a function of N by fitting τ with a power law function $\tau(N) = c_0 + c_1 N^{\gamma}$ in the whole interval of N values. While for SHVC we estimate a scaling exponent $\gamma \approx 1.70$, for Pvclust we find $\gamma \approx 1.94$, thus indicating that the better performances of SHVC are even more appreciated when applied to large systems. Computational time and scalability with the size M of the records are also very different between the two algorithms. As in the previous case we numerically estimate the time dependance of computational time τ as a function of M by fitting τ with a power law function $\tau(N) = c_0 + c_1 M^{\gamma}$. We observe $\gamma \approx 0.76$ for Pvclust, while our estimate of γ for SHVC is close to zero, thus indicating no dependance from the size of records. Indeed, as mentioned above, with our algorithm there is no need to recompute the hierarchical tree for each of the N_b bootstrap replica of data, as it is required for Pvclust. Therefore, after performing the bootstrap, our methodology only deals with the $N-1$ clades of the original dendrogram, which might explain why there is no dependance on M. Furthermore, for Pvclust we have to consider that the computational time spent for computing an average linkage, which has a computational complexity of the order $O(N^3)$, must be multiplied for the N_b bootstrap replica of data. Usually N_b is very large in order to improve the statistical reliability of the p-values, which might explain why Pvclust can be drastically slower than SHVC.

When considering real world data, our algorithm shows a higher ability to extract real biological features from tissue-gene arrays. We applied it also to a large financial dataset, which cannot be explored in a feasible time by Pvclust, and SVHC was able to extract a concise nested partition which strongly overlap with the economic and industrial sector taxonomy. Moreover, as already noted above, the computational time τ of SVHC is of 172 s and 2.19 h for the two considered datasets, respectively. Such numbers should be compared with 4767 s and 105 days respectively estimated for Pvclust. This further confirms the good scalability of our algorithm. We therefore propose to use our algorithm as a valid alternative to Pvclust, and especially in all cases when the Pvclust algorithm is too slow to be used or when it produces outputs that are quite different in the presence or absence of the multiple hypothesis test correction.

Our methodology is implemented in a Python code that is available at: https://github.com/cbongiorno/svhc.

CRediT authorship contribution statement

Christian Bongiorno: Writing – original draft, Python code written. **Salvatore Micciché:** Writing – original draft. **Rosario N. Mantegna:** Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Availability of data

Data dealt with in section 4.1 are publicly available. For example, they can be retrieved from the PVclust software available at the following link: https://github.com/shimo-lab/pvclust. Financial Data dealt with in section 4.2 cannot be shared publicly although they are publicly available online.

Acknowledgments

S.M. and R.N.M. acknowledge financial support by the Italian Ministry of Education, University and Research, Progetto di Ricerca di Interesse Nazionale PRIN 2017WZFTZP, Stochastic forecasting in complex systems. All authors approved the version of the manuscript to be published.

References

[1] J. Han, J. Pei, M. Kamber, Data Mining: Concepts and Techniques, Elsevier, 2011.
[2] J. Felsenstein, Evolutionary trees from DNA sequences: a maximum likelihood approach, J. Mol. Evol. 17 (6) (1981) 368–376.
[3] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, Cluster analysis and display of genome-wide expression patterns, Proc. Natl. Acad. Sci. 95 (25) (1998) 14863–14868.
[4] P. Filzmoser, R. Baumgartner, E. Moser, A hierarchical clustering method for analyzing functional MR images, Magn. Reson. Imaging 17 (6) (1999) 817–826.
[5] C. Goutte, P. Toft, E. Rostrup, F.Å. Nielsen, L.K. Hansen, On clustering fMRI time series, NeuroImage 9 (3) (1999) 298–310.
[6] A. Baune, F.T. Sommer, M. Erb, D. Wildgruber, B. Kardatzki, G. Palm, W. Grodd, Dynamical cluster analysis of cortical fMRI activation, NeuroImage 9 (5) (1999) 477–489.
[7] C. Edelbrock, Mixture model tests of hierarchical clustering algorithms: The problem of classifying everybody, Multivar. Behav. Res. 14 (3) (1979) 367–384.
[8] R.N. Mantegna, Hierarchical structure in financial markets, Eur. Phys. J. B-Conden. Matter Complex Syst. 11 (1) (1999) 193–197.
[9] M. Tumminello, F. Lillo, R.N. Mantegna, Hierarchically nested factor model from multivariate data, Europhys. Lett. 78 (3) (2007) 30006.
[10] F. Musciotto, L. Marotta, S. Micciché, J. Piilo, R.N. Mantegna, Patterns of trading profiles at the nordic stock exchange. a correlation-based approach, Chaos Solitons Fractals 88 (2016) 267–278.
[11] M. Gligor, M. Ausloos, Convergence and cluster structures in EU area according to fluctuations in macroeconomic indices, J. Econ. Integr. (2008) 297–330.
[12] M.E.J. Newman, The structure of scientific collaboration networks., Proc. Natl. Acad. Sci. USA 98 (2001) 404–409.
[13] M. Sales-Pardo, R. Guimerà, A.A. Moreira, L.A. Nunes Amaral, Extracting the hierarchical organization of complex systems, Proc. Natl. Acad. Sci. USA 104 (39) (2007) 15224–15229.
[14] T. Calinski, J. Harabasz, A dendrite method for cluster analysis, Commun. Stat.-Theory Methods 3 (1) (1974) 1–27.
[15] R. Tibshirani, G. Walther, T. Hastie, Estimating the number of clusters in a data set via the gap statistic, J. R. Stat. Soc.: Ser. B Stat. Methodol. 63 (2) (2001) 411–423.
[16] Y. Jung, H. Park, D.-Z. Du, B.L. Drake, A decision criterion for the optimal number of clusters in hierarchical clustering, J. Global Optim. 25 (1) (2003) 91–111.
[17] J. Handl, J. Knowles, D.B. Kell, Computational cluster validation in post-genomic data analysis, Bioinformatics 21 (15) (2005) 3201–3212.
[18] J.C. Dunn, Well-separated clusters and optimal fuzzy partitions, J. Cybern. 4 (1) (1974) 95–104.
[19] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65.
[20] G. Brock, V. Pihur, S. Datta, S. Datta, et al., Clvalid, an R package for cluster validation, J. Stat. Softw. (Brock Et Al., March 2008) (2011).
[21] P. Langfelder, B. Zhang, S. Horvath, Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for R, Bioinformatics 24 (5) (2007) 719–720.
[22] J. Felsenstein, Confidence limits on phylogenies: an approach using the bootstrap, Evolution 39 (4) (1985) 783–791.
[23] B. Efron, E. Halloran, S. Holmes, Bootstrap confidence levels for phylogenetic trees, Proc. Natl. Acad. Sci. 93 (23) (1996) 13429.
[24] H. Shimodaira, Another Calculation of the p-Value for the Problem of Regions Using the Scaled Bootstrap Resamplings, Department of Statistics, Stanford University, 2000.
[25] H. Shimodaira, et al., Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling, Ann. Stat. 32 (6) (2004) 2616–2641.
[26] R. Suzuki, H. Shimodaira, Pvclust: an R package for assessing the uncertainty in hierarchical clustering, Bioinformatics 22 (12) (2006) 1540–1542.
[27] R. Miller Jr., Simultaneous Statistical Inference/RG Jr. Miller, McGraw Hill, New York, 1981.
[28] P.J. Park, J. Manjourides, M. Bonetti, M. Pagano, A permutation test for determining significance of clusters with applications to spatial and gene expression data, Comput. Statist. Data Anal. 53 (12) (2009) 4290–4300.
[29] P. Sebastiani, T.T. Perls, Detection of significant groups in hierarchical clustering by resampling, Front. Genet. 7 (2016) 144.
[30] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. R. Stat. Soc. Ser. B Methodol. (1995) 289–300.
[31] J. Schmid, J.M. Leiman, The development of hierarchical factor solutions, Psychometrika 22 (1) (1957) 53–61.
[32] M.E. Garber, O.G. Troyanskaya, K. Schluens, S. Petersen, Z. Thaesler, M. Pacyna-Gengelbach, M. Van De Rijn, G.D. Rosen, C.M. Perou, R.I. Whyte, et al., Diversity of gene expression in adenocarcinoma of the lung, Proc. Natl. Acad. Sci. 98 (24) (2001) 13784–13789.
[33] J.H. Steiger, Tests for comparing elements of a correlation matrix, Psychol. Bull. 87 (2) (1980) 245.
[34] A.F. McDaid, D. Greene, N. Hurley, Normalized mutual information to evaluate overlapping community finding algorithms, 2011, arXiv preprint arXiv:1110.2515.
[35] L. Danon, A. Diaz-Guilera, J. Duch, A. Arenas, Comparing community structure identification, J. Stat. Mech.: Theory Exp. 2005 (09) (2005) P09008.
[36] A. Lancichinetti, S. Fortunato, J. Kertész, Detecting the overlapping and hierarchical community structure in complex networks, New J. Phys. 11 (3) (2009) 033015.
[37] A.J. Gates, I.B. Wood, W.P. Hetrick, Y.-Y. Ahn, Element-centric clustering comparison unifies overlaps and hierarchy, Sci. Rep. 9 (1) (2019) 8574.
[38] L.M. Collins, C.W. Dent, Omega: A general formulation of the rand index of cluster recovery suitable for non-disjoint solutions, Multivar. Behav. Res. 23 (2) (1988) 231–242.
[39] K.L. Lange, R.J. Little, J.M. Taylor, Robust statistical modeling using the t distribution, J. Am. Stat. Assoc. 84 (408) (1989) 881–896.
[40] R. Fisher, Statistical Methods For Research Workers, Oliver and Boyd, 1950.
[41] It is worth recalling that the procedure of subtracting the average value is also done in the case of microarray data discussed in the previous section; in that case this procedure is necessary if one wants to allow that different microarrays are comparable with each other. In our case the procedure helps in enhancing the local hierarchical organization of the stocks.
[42] C. Bongiorno, D. Challet, Non-parametric sign prediction of high-dimensional correlation matrix coefficients, Europhys. Lett. 133 (4) (2021) 48001.
[43] C. Borghesi, M. Marsili, S. Micciché, Emergence of time-horizon invariant correlation structure in financial returns by subtraction of the market mode., Phys. Rev. E 76 (2007) 026104.
[44] G. Bonanno, G. Caldarelli, F. Lillo, S. Micciché, N. Vandewalle, R.N. Mantegna, Networks of equities in financial markets., Eur. Phys. J. B 38 (2004) 363–371.
[45] C. Coronnello, M. Tumminello, F. Lillo, S. Micciché, R.N. Mantegna, Sector identification in a set of stock return time series traded at the London stock exchange., Acta Phys. Polon. B 36 (9) (2005) 2653–2679.