Statistically validated hierarchical clustering: Nested partitions in hierarchical trees

Christian Bongiorno, Salvatore Miccichè, Rosario N Mantegna

Nested partitions from hierarchical clustering statistical validation

Christian Bongiorno[(1)],* Salvatore Miccichè[(2)], and Rosario N. Mantegna[(2,3,4)]
[(1)] *Laboratoire de Mathématiques et Informatique pour les Systèmes Complexes, CentraleSupélec, Université Paris Saclay, 3 rue Joliot-Curie, 91192, Gif-sur-Yvette, France*
[(2)] *Dipartimento di Fisica e Chimica, Università di Palermo, Viale delle Scienze, Ed. 18, I-90128, Palermo, Italy*
[(3)] *Complexity Science Hub Vienna, Josefstädter Strasse 39, 1080, Vienna, Austria*
[(4)] *Computer Science Department, University College London, 66 Gower Street, WC1E 6BT, London, UK*
(Dated: June 17, 2019)

We develop a greedy algorithm that is fast and scalable in the detection of a nested partition extracted from a dendrogram 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 a number of 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 using a hierarchical factor model. We compare the results obtained by our algorithm with those of Pvclust. Pvclust is a widely used algorithm developed with a global approach originally motivated by phylogenetic studies. In our numerical experiments we focus on the role of multiple hypothesis test correction and on the robustness of the algorithms to inaccuracy and errors of datasets. We also apply our algorithm to a reference empirical dataset. 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. Our algorithm provides a hierarchically nested partition in much shorter time than currently widely used algorithms allowing to perform a statistically validated cluster analysis detection in very large systems.

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 are computational biology [2], genomics [3], neuroscience [4–6], psychology [7], finance [8, 9] and economics [10]. 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 as an output a hierarchical tree (also called dendrogram). However, the detection of a dendrogram does not mean that one also obtains a hierarchical nested partition as an output of the HC. Historically, the simplest and most popular way to obtain a partition from a hierarchical tree was to cut the dendrogram at a fixed dissimilarity 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. [11]. 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 [12], squared error [13], connectivity [14], Dunn index [15], or silhouette width [16]. The R package clValid allows to compute hard partitions (i.e. partitions where an element can belong only to a single cluster) with most of the previously cited methods [17]. The dynamical cut tree method provides a different approach, which allows a cut of the dendrogram at different distances levels [18].

Felsenstein focused on the problem of assessing the statistical significance of the clusters obtained by HC [19] within phylogenetic studies. Specifically, he proposed to associate a p-value to each clade of hierarchical tree. In phylogeny such a p-value provides a direct information on the evolutionary hypothesis associated with the formation of the clade. The method 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 [20]. More recently, Shimodaira implemented the refinement of Efron [20], developed the so-called approximately unbiased (AU) test based on bootstrap [21, 22], and achieved a higher accuracy with respect to the previous proposed statistical tests. An R-package with the implementation of this test (AU test), named Pvclust, was released in Ref. [23] and it is currently widely used in phylogenetic and genomic analyses.

It is worth noting that Felsenstein's approach is a global approach 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 system. For some large systems the time could be so long that its application is unfeasible. Another problem of applicability of Pvclust to large set of data concerns the aspects of multiple hypothesis

* christian.bongiorno@centralesupelec.fr

test correction [24]. 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. However, the currently available multiple hypothesis test corrections, such as for example the control of the false discovery rate (FDR) [25], guarantee a highly controlled number of false positive at the expenses of a large number of false negative. Due to this limitation, the above cited Pvclust algorithm is often used without multiple hypothesis test correction opening the way to the potential presence of a number of false positive.

In addition to the methods motivated by a phylogenetic approach, other methods have been proposed more recently to associate a statistical significance to hierarchical partitions obtained by using hierarchical clustering. 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 hierarchical clustering. Examples of these approaches are the permutation test that quantifies the significance of each division of a hierarchical tree as proposed in Ref. [26] 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 [27].

In this work, we propose a greedy algorithm based on bootstrap resampling that associates a p-value at each clade of a 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 proximity measure (in our approach a dissimilarity) associated with a clade h and the dissimilarity measure associated with the clade defined by its parent node in the genealogy of the dendrogram. The statistical test we perform consider as a null hypothesis that the dissimilarity of the parent node is larger than the dissimilarity of the children node. Our tests are performed by considering multiple hypothesis test correction. In fact, we always apply the control of FDR [25]. By selecting those clades that reject our null hypothesis, we identify a hierarchically nested partition involving a certain number of elements of the investigated systems. In order to evaluate the performance of our method, we test it with some benchmarks obtained by using a hierarchical factor model [28]. In our tests, we compare our results with the ones obtained with Pvclust with and without a multiple comparison correction. Finally, we apply our algorithm to an empirical dataset. This dataset was originally obtained in Ref. [29] and was used as an example in the paper describing Pvclust [23].

Our algorithm is highly accurate when applied to benchmarks obtained from hierarchical factor models and is also highly informative in the analysis of empirical datasets. Being our approach heuristic and local the algorithm cannot guarantee detection of global optimal solutions. This is of course a limitation of our algorithm. The positive aspect of this limitation is that our algorithm is very fast and highly scalable and therefore can be used for large datasets that would otherwise need extremely long computer time to provide results. With our algorithm one can perform a screening of large data sets, analyze results and then apply most demanding algorithms only to those sets of data that provides interesting results with a greedy approach.

METHODS

Statistically Validated Hierarchical Clustering

Let us assume that a clade originating from node h has associated a dissimilarity measure ρ_{pq}. This 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 form the clade l. The dissimilarity value defining the clade l is ρ_{hk}. The agglomerative procedure of the hierarchical clustering requires that the dissimilarity ρ_{pq} must be lower than ρ_{hk}. This is the basic aspect of the hierarchical clustering procedure that we put at the core of our algorithm. In fact in our algorithm, for each pair of parent children clades, we perform a statistical test of the null hypothesis $\rho_{hk} \leq \rho_{pq}$. When our null hypothesis is rejected, 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 this p-value can be computed analytically for Gaussian multivariate variables and numerically by computing bootstrap replicas of the dissimilarity matrix of the original data.

We consider a multivariate dataset X of dimension $N \times M$ with N elements and M records or attributes. We call R the $N \times N$ Pearson's correlation matrix of X and we use it as a similarity measure. It is worth noting that our choice is just a possible choice of a similarity measure. In fact our procedure works for a generic definition of similarity matrix. We label as $\sigma(x)$ the set of elements of clade defined by node x of the dendrogram and N_x the number of elements that clade x contains. Hierarchical clustering is performed by using a dissimilarity measure. In this work we quantify the dissimilarity measure according to the definition

$$\rho_{hk} = \frac{\sum_{i \in \sigma(h)} \sum_{j \in \sigma(k)} 1 - R_{ij}}{N_h \, N_k} \tag{1}$$

where $\sigma(h)$ and $\sigma(k)$ are the sets of nodes of clade h and of clade k in the hierarchical tree, respectively.

Analytical Derivation of the p-value

We derive an analytical expression of the p-value π_h associated with the null hypothesis $W_h = \rho_{hk} - \rho_{pq} \leq 0$ under the hypothesis that X is a set of multivariate normal distributed random variables and M is a large number. Our analytical results are obtained under the assumption that the hierarchical clustering procedure is the average linkage. Our p-value is defined as the cumulative distribution function in zero of the stochastic variable $W_h^{(s)} = \rho_{hk}^{(s)} - \rho_{pq}^{(s)}$, where, $\rho_{pq}^{(s)}$ is the sample mean of ρ_{pq} defined as in Eq. (1). 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 normal 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 will be a weighted sum of correlated normal random variables. In particular, according to Ref. [30], the covariance between two elements R_{ij} and R_{lm} of a correlation matrix is

$$\begin{aligned}\Xi_{(i,j),(lm)} = \frac{1}{2M}\{&[(R_{il} - R_{ij}R_{lj})(R_{jm} - R_{jl}R_{lm})] + \\ &[(R_{im} - R_{il}R_{lm})(R_{jl} - R_{ji}R_{il})] + \\ &[(R_{il} - R_{im}R_{ml})(R_{jm} - R_{ji}R_{im})] + \\ &[(R_{im} - R_{ij}R_{jm})(R_{jl} - R_{jm}R_{ml})]\}\end{aligned} \tag{2}$$

and therefore the variance of element R_{ij} is

$$\Xi_{(ij),(ij)} = \frac{1}{M}\left(1 - R_{ij}^2\right)^2. \tag{3}$$

The expected value of the stochastic variable W_h is $E[W_h] = \rho_{pq} - \rho_{hk}$. 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 distance ρ_{pq} are identified by the rectangular matrix of N_p and N_q elements of sets $\sigma(p)$ and $\sigma(q)$ respectively. Similarly the elements needed to compute ρ_{hk} are identified by a rectangular matrix of elements of sets $\sigma(h)$ and $\sigma(k)$ ($N_h\, N_k$ elements). By considering the definition of W_h, its variance is

$$\begin{aligned}S[W_h]^2 = &\frac{1}{(N_p N_q)^2}\sum_{i\in\sigma(p)}\sum_{j\in\sigma(q)}\sum_{l\in\sigma(p)}\sum_{m\in\sigma(q)}\Xi_{(ij),(lm)} + \\ &\frac{1}{(N_h N_k)^2}\sum_{i\in\sigma(h)}\sum_{j\in\sigma(k)}\sum_{l\in\sigma(h)}\sum_{m\in\sigma(k)}\Xi_{(ij),(lm)} + \\ &-\frac{1}{N_p N_q N_h N_k}\sum_{i\in\sigma(p)}\sum_{j\in\sigma(q)}\sum_{l\in\sigma(h)}\sum_{m\in\sigma(k)}\Xi_{(ij),(lm)}\end{aligned} \tag{4}$$

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

$$\pi_h = P(W_h < 0) = \frac{1}{2}\left[1 + \mathrm{erf}\left(-\frac{E[W_h]}{S[W_h]\sqrt{2}}\right)\right] \tag{5}$$

Numerical estimation of the p-value

Let us call $X^{(s)}$ a bootstrap copy of X obtained from sampling with replacement of the columns of X matrix. Let be $R^{(s)}$ the correlation matrix obtained from a bootstrap replica. For each bootstrap replica it is possible to compute for each group of elements a dissimilarity. For example by considering the set of nodes $\sigma(h)$ and $\sigma(k)$ we can compute the set of dissimilarity $\left\{\rho_{hk}^{(1)}, \rho_{hk}^{(2)}, \ldots, \rho_{hk}^{(n)}\right\}$ and by considering the set of nodes $\sigma(p)$ and $\sigma(q)$ we can compute the dissimilarities $\left\{\rho_{pq}^{(1)}, \rho_{pq}^{(2)}, \ldots, \rho_{pq}^{(n)}\right\}$ where n is the number of bootstrap replicas. It is worth noting that such dissimilarities are evaluated according to the composition of sets $\sigma(x)$ by using Eq. (1) without computing a hierarchical tree for each bootstrap replica. The p-value associated to cluster h is defined as

$$\pi_h = \frac{\sum_{i=1}^{n}\delta(\rho_{hk}^{(i)} \leq \rho_{pq}^{(i)})}{n} \tag{6}$$

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

Since we are computing a p-value for each node of the hierarchical tree we face family wise error. In this work, to perform a multiple hypothesis test correction we use the procedure of the control of the FDR [25]. The control of the FDR procedure is implemented as follows: the $N-2$ clades 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 $\pi^{(k)} \leq k\alpha/(N-2)$ and the clades corresponding to the first k_{max} p-values are used to build up a nested partition of the elements. The statistical threshold α is the maximum proportion of false discovery allowed in our statistical test. In this work we choose $\alpha = 0.05$.

It should be noticed that in our algorithm the most computational demanding procedure is the computation of bootstrap replicas. Since bootstrap replicas are independent the one from the other, the algorithm can be easily and efficiently parallelized.

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

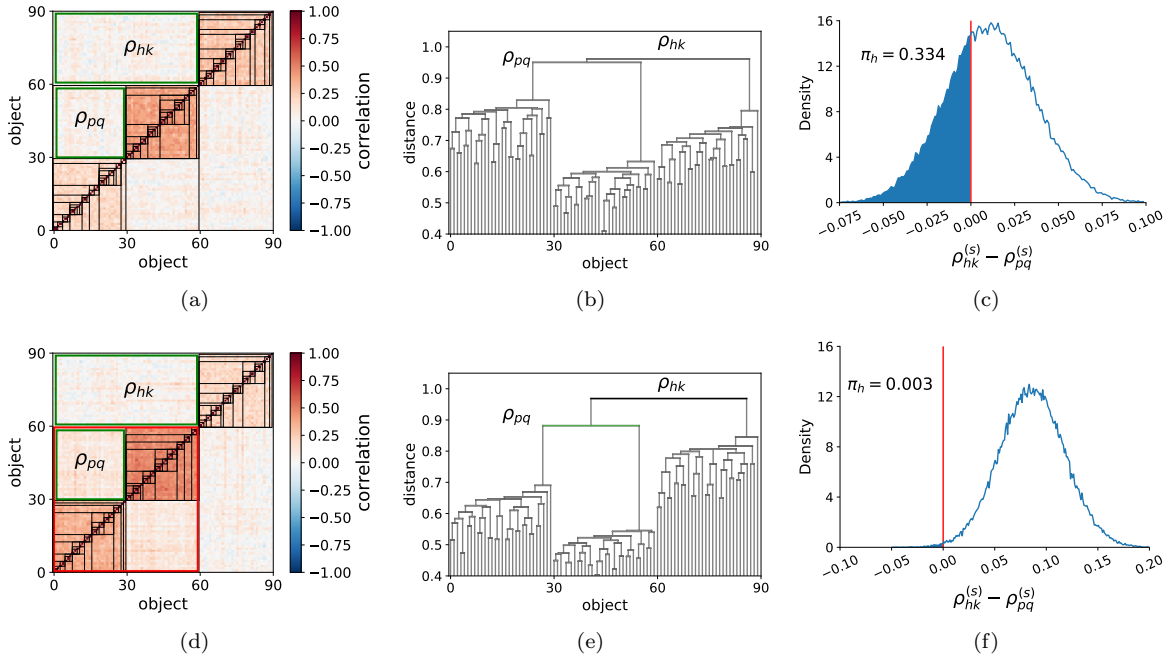


FIG. 1. Panels (a) and (d) show correlation matrices of slightly similar hierachically nested benchmarks generated with $M = 200$. The elements are sorted according to the hierarchical tree of the HC average algorithm. The black boxes indicate the clusters of the all clades of the hierarchical tree. 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 $\sigma(p)$, the elements $[30, 59]$ belongs to the set of elements $\sigma(q)$, the elements $[60, 89]$ belongs to the set of elements $\sigma(k)$, and the elements $[0, 59]$ belongs to the set of elements $\sigma(h)$. In panels (b) and (e) we show dendrograms of average HC of dissimilarity matrices associated with the multivariate datasets with correlation matrices of panels (a) and (d) respectively. In panels (c) and (f) we show the density function of $\rho^{(s)}_{hk} - \rho^{(s)}_{pq}$ to illustrate how the p-value of the null hypothesis $W_h \leq 0$ is estimated (in these examples we perform $n = 100,000$ 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 same null hypothesis is rejected

stocks statistically distinct from the set of all stocks. In the top row of Fig. 1 we show three panels referring to the case when the null hypothesis $W_h \leq 0$ 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 $\sigma(p)$ (elements $[0, 29]$) and the clade of elements $\sigma(q)$ $[30, 59]$) join together in the clade $\sigma(h)$, originating at $\rho_{pq} = 0.95$. Then the clade $\sigma(h)$ joins with clade $\sigma(k)$ (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 this structure observed in the hierarchical tree of the sample correlation matrix, the bootstrap analysis of $\rho^{(s)}_{pq}$ and $\rho^{(s)}_{hk}$ shows that the null hypothesis $W_h \leq 0$ has associated a p-value equal to $\pi_h = 0.333$ 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 [0,99].

In the bottom row of Fig. 1 we show a slightly different example. Specifically, in this case the dissimilarity values are $\rho_{pq} = 0.88$ and $\rho_{hk} = 0.96$, as shown in Fig. 1(e). In other words, elements [0,59] are slightly more correlated than in the previous case. For this set of data, our approach concludes that the clade [0.59] is statistically distinct from the complete set [0.99] since the inequality $W_h \leq 0$ is verified (Fig. 1(f)) only for 0.3% of our bootstrap replicas. Therefore the null hypothesis $W_h \leq 0$ has associated a p-value $\pi_h = 0.003$ and after performing the FDR multiple hypothesis test correction we reject it.

Hierachically nested Benchmark

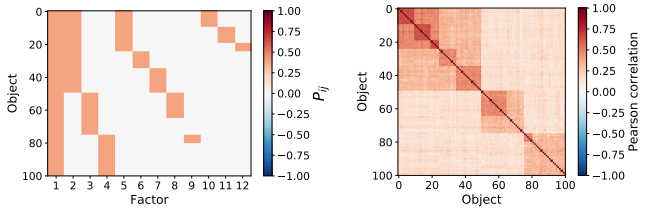
In our numerical experiments, We use benchmarks of multivariate datasets obtained with a nested factor model with r common factors [28]. Specifically, We simulate a multivariate dataset X of N elements with M records by using the equation

$$X_{ij} = \sum_{k=1}^{r} P_{ik} A_{kj} + U_i \varepsilon_{ij} \tag{7}$$

where P is the factor loading matrix of dimension $N \times r$ and A is a factor score matrix of dimension $r \times M$ 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 $U_i = \sqrt{1 - \sum_{j=1}^{r} P_{ij}^2}$. 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 X obtained from the factor loading matrix P of Fig. 2(a) with $N = 100$ elements and 12 factors together with a factor score matrix A with 12 factors and $M = 500$ records has associated the correlation matrix shown in Fig. 2(b). With this choice of P each



(a) (b)

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 of (a) with with $r = 12$ and a factor score matrix $r \times M$ with $r = 12$ and $M = 500$ standardized independent Gaussian variables.

factor corresponds to a block on the correlation matrix, and an element i is a member of the block associated with the r factor if P_{ir} 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 some numerical experiments we add noise to each record to investigate the robustness of the algorithms to inaccuracy and errors of the datasets. This is done by computing

$$X_{ij} = (1 - \lambda)X'_{ij} + \lambda\varepsilon_{ij}$$

where X'_{ij} is the dataset without noise, ε is a standardized Gaussian variable, and $\lambda \in [0, 1]$ is the parameter controlling the amount of noise inserted into the dataset.

In numerical experiments discussed in the Results section, 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.

Comparing partitions

The comparison metric used in this paper to assess the similarity of two hierarchically nested partitions is the overlapping normalized mutual information (ONMI) [31]. ONMI is a variant of the normalized mutual information (NMI) [32]. 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 in Ref. [33]. Later authors of Ref. [31] 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.

RESULTS

Comparison between the analytical and bootstrap based p-value

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

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. In Fig. 3(a), we show an example of the bootstrap probability density function of the stochastic variable W_h for a selected clade h compared with the result of the analytical computation. In Fig. 3(b), we also show a scatter plot between $\pi_h^{(a)}$ and $\pi_h^{(b)}$ of one bootstrap realization for two values of $n = (10^2, 10^5)$. It is worth noting that the bootstrap p-values converge to their analytical values for large values of n. In our numerical experiments we do not detect any bias in the numerical estimation of bootstrap p-values for Gaussian multivariate data.

Student's t-distribution case

In order to study the sensitivity of the analytic estimation to the Gaussian hypothesis we compare analytical and bootstrap p-values in the case of a multivariate dataset 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

$$f(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}. \qquad (8)$$

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. It is worth recalling that a value of

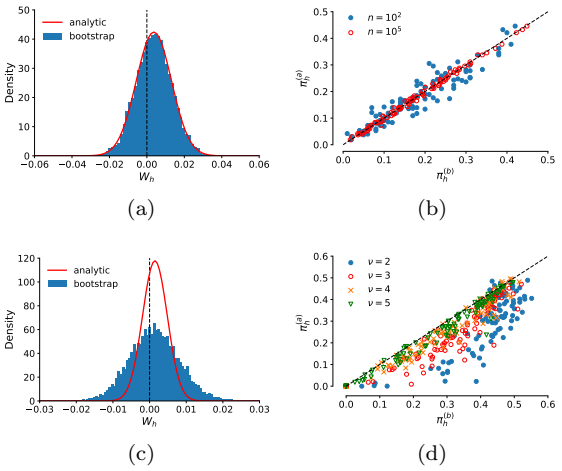


FIG. 3. Numerical experiments performed with uncorrelated multivariate random variables. (a) Histogram of the probability density function of W_h for a selected clade obtained by using $n = 10^4$ bootstrap replicas of Gaussian multivariate random variables. The red line is the analytical probability density function obtained under the hypothesis of Gaussian variables.(b) Scatter plot of e p-values of analytic computation $\pi_h^{(a)}$ versus p-values obtained with bootstrap $\pi_h^{(b)}$ for two values of n. Each point refers to the p-value of a clade (Gaussian random variables). (c) Histogram of the probability density function of W_h for a selected clade obtained by performing $n = 10^4$ bootstrap replicas. The red line is the analytical probability density function expected for Gaussian variables for Student's t multivariate random variables. (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.

$\nu \approx 3$ has been observed in several financial studies [34].

In Fig. 3(c), we show an example of the bootstrap probability density function of the stochastic variable W_h for a selected clade h compared with the analytic probability density function expected for Gaussian variables. 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 between $\pi_h^{(a)}$ and $\pi_h^{(b)}$ for different values of ν. It is worth noting that for large value of ν the discrepancy between analytical and bootstrap p-values is progressively reduced since the t-distribution converges to the Gaussian when $\nu \to \infty$ [35]. We therefore conclude that numerical investigations are therefore essential when the probability density function of the multivariate dataset differs from a Gaussian multivariate.

Experiments on the Benchmark

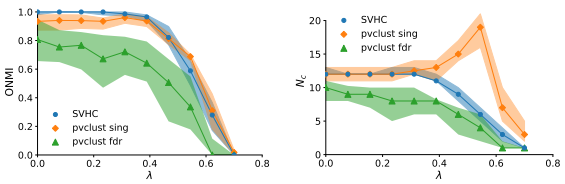
Here we investigate the effectiveness of our algorithm in retrieving the true hierarchical nested partition of a representative benchmark. We also compare our results with the outputs of the algorithm Pvclust [23]. We show that the SVHC algorithm has a good scalability for large systems. For this reason, by considering that our algorithm is preferentially suggested to investigate large systems a multiple hypothesis test correction is part of the algorithm. On the other hand, the multiple hypothesis test correction is an option in the Pvclust algorithm. To take into account this important difference, we are comparing the two algorithms by considering the SVHC output and two outputs of Pvclust, the first obtained without multiple hypothesis test correction (labeled in our figures as "single") and the second obtained with the control of the FDR (labeled as "FDR"). Partitions investigated in this paper are by construction hierarchically nested partitions. We therefore compare hierarchically nested partitions.

Hierarchical partitions

In a second set of numerical experiments (see Fig. 4), we explore the robustness of the two algorithms to different levels of noise in the detection of hierarchically nested partitions of the benchmark. For low levels of noise ($\lambda \leq 0.4$) the SVHC algorithm has a very good performance both in terms of ONMI with the true partition of the benchmark (see Fig. 4(a)) and in terms of the number of clusters detected (see Fig. 4(b)). In the analysis of the figure, it should be noted that the benchmark is characterized by 12 nested clusters. Pvclust "single" has a similar performance for low values of λ but the quality of the detected hierarchical partition is strongly affected by the option about the multiple hypothesis test correction. In fact, in Fig. 4(a) we observe that the hierarchical partition of Pvclust "FDR" has lower performance than the one of Pvclust "single".

For values of the noise parameter $\lambda > 0.4$ both the SVHC and the Pvclust algorithms reduce their ability to retrieve the true hierarchical partition and they have similar performances concerning the ONMI metric. However, the outputs obtained by the two algorithms are characterized by a different types of error. In fact, in Fig. 4(b) we show that starting from noise parameter $\lambda \approx 0.4$ the output of the algorithm Pvclust "single" starts to be characterized by an increasing number of clusters whereas the SVHC presents the opposite case. As a result, in statistical terms the partition retrieved from the SVHC algorithm is more precise than the one obtained from Pvclust "single" , although it lacks in the recall (i.e. it has a large number of false negative).

In a third 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 twelve nested clusters. This is done by using the same type of benchmark of Fig. 2 modified by increasing the number of elements of each cluster and the total number of elements proportionally. Moreover,



(a) (b)

FIG. 4. (*a*) Overlapping normalized mutual information (ONMI) between the true hierarchical partition of the benchmark of Fig. 2 with $N = 100$ and $M = 500$ and the hierarchical partition obtained with SVHC, Pvclust "single" or Pvclust "FDR" as a function of the noise parameter λ. (*b*) Number of statistically validated clusters detected by the algorithms as a function of the noise parameter λ. 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.

the number of records M of the time series is also increased proportional to N according to $M = 5N$. We perform numerical experiments for systems of sizes equal to $N = \{56, 100, 178, 316, 562\}$ where the different values present a logarithmic spacing. In this set of experiments noise is absent ($\lambda = 0$).

Also in this case hierarchical partitions obtained with the SVHC algorithm describes quite well the true hierarchical partition for systems with N ranging from 56 to 562 (see ONMI values in Fig. 5(a)). The algorithm Pvclust has again a performance that is strongly dependent on the multiple hypothesis test correction option. In particular, for low values of N the results obtained by Pvclust "single" perform better then the results obtained with 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. Pvclust hierarchical partitions have different characteristics for low and high values of N. For low values of N, Pvclust "single" 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 Pvclust "single" is probably due to the absence of the multiple hypothesis test correction. In fact, the number of statistical tests performed increases linearly with the size of the system. The profile of the results obtained by Pvclust "FDR" is different. For low values of N the number of clusters detected is less than the true number. This is probably due to the well known limitation of multiple hypothesis test correction. In fact the correction fully controls the amount of false positives but this is done at the expenses of not controlling the number of false negatives. For the large value of N this limitation is progressively less important and the performance of the hierarchical partition of the Pvclust "FDR" algorithm becomes very good for large values of N. In summary, the Pvclust algorithm provides outputs recovering the true partition with one of the two multiple hypothesis test options. Specifically the "single" option works well for small systems whereas the FDR option is more appropriate for large systems.

An important aspect of the two algorithms is computational time. In Fig. 5(c) we report the computational time for the SVHC and the Pvclust algorithms (the two multiple hypothesis test correction options of Pvclust do not significantly affect the computational time of the algorithm). From Fig. 5(c) 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 dependence of computational time T_c as a function of N by fitting T_C with a power law function $T_c = c_0 + c_1\, N^{\gamma}$ in the whole interval of N values and we obtain $\gamma = 1.94$ for the Pvclust algorithm and $\gamma = 1.70$ for the SVHC algorithm. The other fitting parameters are $c_0 = 22$ and $c_1 = 3 \times 10^{-2}$ for Pvclust and $c_0 = 2.4$ and $c_1 = 3 \times 10^{-4}$ for SVHC. In addition to the difference observed in the exponent γ, it should be also noted that the coefficients c_0 and c_1 of T_c for the SVHC algorithm are much smaller than the same coefficients for Pvclust.

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.5(d) show that SVHC outperforms Pvclust in detecting the true hierarchical partition for high values of M. On the contrary, for low values of M Pvclust "single" performs better than SVHC. More details about the ability of the algorithms to detect the true partition can be obtained by inspecting Fig. 5(e). This figure plots the number of clusters detected by the algorithms. For low values of M, the algorithm Pvclust "single" has a low number of false negative but this performance is obtained at the expenses of a large number of false positive, whereas both SVHC and Pvclust FDR have a large number of false negative. Depending whether the most important aspect is statistical precision or statistical accuracy the most appropriate algorithm turns out to be different.

It is again worth noting that computational time is very different for the two algorithms and also the scalability is different. Fig. 5(f) shows that computational time needed for SVHC and Pvclust. We again fit the computational time with the functional form $T_c(M) = c_0 + c_1\, M^{\gamma}$ and obtain $\gamma = 0.76$ for the Pvclust algorithm and $\gamma = 5.5 \times 10^{-4}$ for the SVHC algorithm. The other fitting parameters are $c_0 = 123$ and $c_1 = 1.01$ for Pvclust and $c_0 = 2.89$ and $c_1 = 1.07$ for SVHC. 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

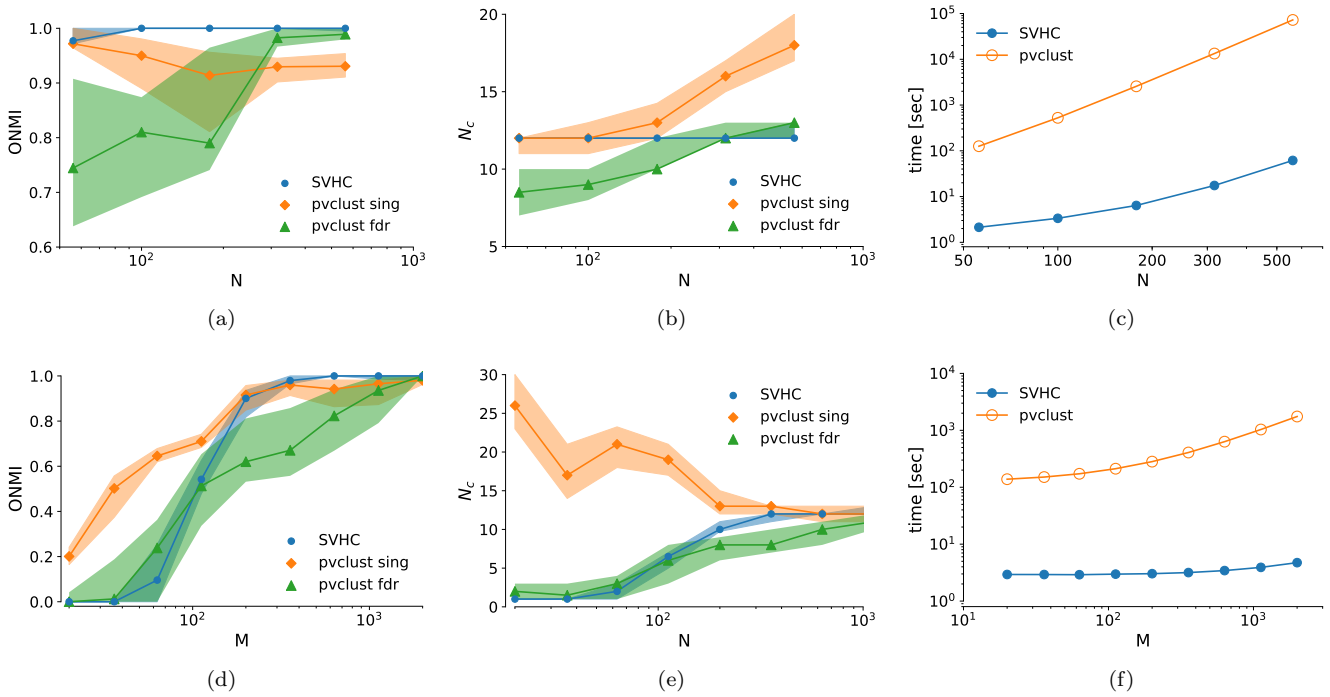


FIG. 5. Numerical experiments with a benchmark of the type shown in Fig. 2 for different values of the size of the system N and for different size of records M. (a) ONMI between the true hierarchical partition of the benchmark and the hierarchical partition obtained with SVHC, Pvclust "single" or Pvclust "FDR" as a function of the system size N. (b) Number of statistically validated clusters detected by the algorithms as a function of the system size N. (b) Computational time T_c of the algorithms as a function of the system size N, In all simulations shown in panels (a), (b), and (c) $M = 5N$. (d) 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. (e) Number of statistically validated clusters detected by the algorithms as a function of M. (f) Computational time T_c of the algorithms as a function of M, In all simulations shown in panels (d), (e), and (f) $N = 100$. 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.

SVHC algorithm is very close to zero. On the contrary, Pvclust computational time is characterized both by a sizeable exponent and also by a large minimum constant time (c_0).

This set of numerical experiments confirms that SVHC algorithm is much faster and presents better scalable characteristics than Pvclust.

Applications to an empirical dataset

We now apply SVHC to a widely investigated empirical dataset. As in previous numerical experiments, the application of SVHC is done in parallel with the application of Pvclust with the two options (i.e. the "single" and the "FDR" option). The 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.

Adenocarcinoma of the lung data

The dataset was originally collected in Ref. [29], and it was used to provide an illustrative example of Pvclust performance in Ref.[23] to obtain p-values of the branching points of hierarchical tree of tissues. Here we investigate both the hierarchical tree of tissues ($N = 73$) and the hierarchical tree of genes ($N = 915$) to provide both a basic example (in the case of tissues) and a more demanding example (in the case of genes) of application of the algorithms to systems of size varying more than one order of magnitude. In our investigation, both the SVHC and the Pvclust algorithms perform $10,000$ 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.

To quantify the degree of similarity of partitions obtained we compute the ONMI between all pairs of hierarchical partitions obtained. For lung tissue the ONMI between the SVHC partition and the Pvclust "single" partition is ONMI(SVHC,Pvclust "S")=0.527 whereas ONMI(SVHC,Pvclust "FDR")=0.368

and ONMI(Pvclust "S",Pvclust "FDR")=0.768. For lung genes we obtain ONMI(SVHC,Pvclust "S")=0.321, ONMI(SVHC,Pvclust "FDR")=0.638 and ONMI(Pvclust "S",Pvclust "FDR")=0.434. By analyzing these values, we note that the hierarchical partitions obtained with SVHC are not too different from the ones obtained with Pvclust both for the small set of elements (tissues) and for the large set of elements (genes). However, in the two cases the highest degree of similarity between partitions is observed for a different option of Pvclust. Specifically, for the hierarchical partition of lung tissues the highest similarity is between the partition of SVHC and the partition of Pvclust "single" whereas for the hierarchical partition of genes the highest similarity is with Pvclust "FDR". This result again suggests that hierarchical partition of Pvclust "single" includes more false positives when the system size increases. A comparison of the hierarchical partition obtained with SVHC with the hierarchical partitions obtained with Pvclust "single" and Pvclust "FDR" can therefore provide indication about the option of Pvclust more appropriate to the size of the system investigated.

As already noted in the investigation of the synthetic benchmark, when the system size increases the computational time grows on. The same beahavior is observed in the analysis of empirical data. In fact, the computational time of the two algorithms for the two hierarchical trees investigated (i.e. for the lung tissues and the lung genes) is as follows: The computational time for lung tissues (i.e. a 73x915 system) is 172 s for SVHC and 4,767 s for Pvclust whereas for lung genes (i.e. a 915x73 system) is 1,434 s for SVHC and 109,660 s for Pvclust. For both investigates sets the computational time of Pvclust is much longer then the computational time of SVHC (27.7 times for lung tissues and 76.5 times for lung genes).

DISCUSSION

Hierarchical clustering is a powerful data analysis tool widely used in many disciplines. The association of hard or hierarchical partitions to hierarchical trees is still an open problem. An approach widely used in genomics has its roots in phylogenetic studies. In such studies bootstrap replicas of each clade are computed and the observation of the number of times the sample composition of the clade is detected in replicas provide a first estimation of the p-value to be associated to a given clade. Over the years this approach has been refined to minimize biases affecting the estimation of the p-value. Today the widely used Pvclust package uses this approach. Pvclust is therefore setting the standard for a statistical assessment of a specific hierarchically nested partition obtained from a given hierarchical tree. Pvclust has a great control of the statistical hypothesis underlying its approach but has also a few drawbacks. One drawback concerns the presence or absence of a procedure of multiple hypothesis test correction. From a statistical point of view, such correction should be present but the results obtained by the algorithm in the presence of a multiple hypothesis test correction are sometimes disappointing due to the fact that such a correction can be too demanding for some systems. The second and most important drawback is the computational time needed to perform the hierarchical clustering estimation for all bootstrap replicas. This time could be quite long for moderately large datasets and therefore the Pvclust algorithm is of limited use for big datasets.

In this paper we introduce a greedy algorithm that is quite effective in the detection of the true hierarchically nested partition of a multivariate series. We prove the efficacy of our algorithm by performing numerical experiments on a set of benchmarks generated by using a hierarchical factor model. The application of Pvclust to the same benchmarks show the efficacy and limitations of Pvclust with the two options of absence (i.e. Pvclust "single") and presence of multiple hypothesis test correction (i.e. Pvclust "FDR"). 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. We therefore propose to use our algorithm 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.

[1] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques* (Elsevier, 2011).

[2] J. Felsenstein, Journal of molecular evolution **17**, 368 (1981).

[3] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, Proceedings of the National Academy of Sciences **95**, 14863 (1998).

[4] P. Filzmoser, R. Baumgartner, and E. Moser, Magnetic resonance imaging **17**, 817 (1999).

[5] C. Goutte, P. Toft, E. Rostrup, F. Å. Nielsen, and L. K. Hansen, NeuroImage **9**, 298 (1999).

[6] A. Baune, F. T. Sommer, M. Erb, D. Wildgruber, B. Kardatzki, G. Palm, and W. Grodd, NeuroImage **9**, 477 (1999).

[7] C. Edelbrock, Multivariate Behavioral Research **14**, 367 (1979).

[8] R. N. Mantegna, The European Physical Journal B-Condensed Matter and Complex Systems **11**, 193 (1999).

[9] M. Tumminello, F. Lillo, and R. N. Mantegna, EPL (Europhysics Letters) **78**, 30006 (2007).

[10] M. Gligor and M. Ausloos, Journal of Economic Integration , 297 (2008).

[11] T. Caliński and J. Harabasz, Communications in Statistics-theory and Methods **3**, 1 (1974).

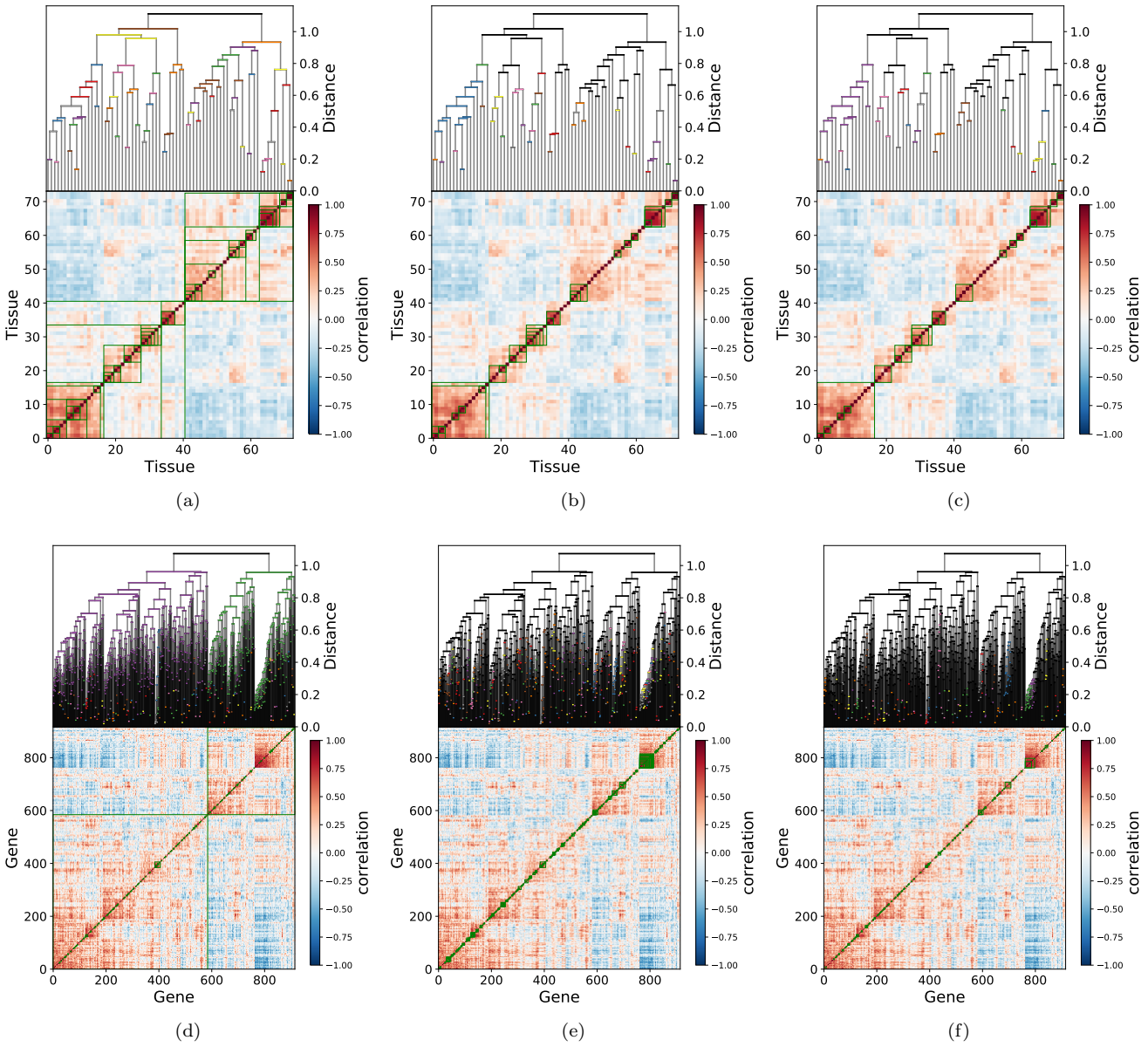


FIG. 6. Hierarchical trees (average HC) and correlation matrices of lung tissues dataset (top panels) and lung genes dataset (bottom panels) [29]. In the correlation matrices we highlight with boxes hierarchically nested clusters detected by different algorithms. (*a*) SVHC on lung tissues, (*b*) Pvclust "single" on lung tissues, (*c*) Pvclust "FDR" on lung tissues. (*d*) SVHC on lung genes, (*e*) Pvclust "single" on lung genes, (*f*) Pvclust "FDR" on lung genes.

[12] R. Tibshirani, G. Walther, and T. Hastie, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **63**, 411 (2001).

[13] Y. Jung, H. Park, D.-Z. Du, and B. L. Drake, Journal of Global Optimization **25**, 91 (2003).

[14] J. Handl, J. Knowles, and D. B. Kell, Bioinformatics **21**, 3201 (2005).

[15] J. C. Dunn, Journal of cybernetics **4**, 95 (1974).

[16] P. J. Rousseeuw, Journal of computational and applied mathematics **20**, 53 (1987).

[17] G. Brock, V. Pihur, S. Datta, S. Datta, *et al.*, Journal of Statistical Software (Brock et al., March 2008) (2011).

[18] P. Langfelder, B. Zhang, and S. Horvath, Bioinformatics **24**, 719 (2007).

[19] J. Felsenstein, Evolution **39**, 783 (1985).

[20] B. Efron, E. Halloran, and S. Holmes, Proceedings of the National Academy of Sciences **93**, 13429 (1996).

[21] H. Shimodaira, *Another calculation of the p-value for the problem of regions using the scaled bootstrap resamplings* (Department of Statistics, Stanford University, 2000).

[22] H. Shimodaira *et al.*, The Annals of Statistics **32**, 2616 (2004).

[23] R. Suzuki and H. Shimodaira, Bioinformatics **22**, 1540 (2006).

[24] R. Miller Jr, "Simultaneous statistical inference/rg jr. miller," (1981).

[25] Y. Benjamini and Y. Hochberg, Journal of the royal statistical society. Series B (Methodological) , 289 (1995).
[26] P. J. Park, J. Manjourides, M. Bonetti, and M. Pagano, Computational statistics & data analysis **53**, 4290 (2009).
[27] P. Sebastiani and T. T. Perls, Frontiers in genetics **7**, 144 (2016).
[28] J. Schmid and J. M. Leiman, Psychometrika **22**, 53 (1957).
[29] 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.*, Proceedings of the National Academy of Sciences **98**, 13784 (2001).
[30] J. H. Steiger, Psychological bulletin **87**, 245 (1980).
[31] A. F. McDaid, D. Greene, and N. Hurley, arXiv preprint arXiv:1110.2515 (2011).
[32] L. Danon, A. Diaz-Guilera, J. Duch, and A. Arenas, Journal of Statistical Mechanics: Theory and Experiment **2005**, P09008 (2005).
[33] A. Lancichinetti, S. Fortunato, and J. Kertész, New Journal of Physics **11**, 033015 (2009).
[34] P. Gopikrishnan, M. Meyer, L. N. Amaral, and H. E. Stanley, The European Physical Journal B-Condensed Matter and Complex Systems **3**, 139 (1998).
[35] K. L. Lange, R. J. Little, and J. M. Taylor, Journal of the American Statistical Association **84**, 881 (1989).

ACKNOWLEDGEMENTS (NOT COMPULSORY)

Acknowledgements should be brief, and should not include thanks to anonymous referees and editors, or effusive comments. Grant or contribution numbers may be acknowledged.

AUTHOR CONTRIBUTIONS STATEMENT

C.B., S.M. and R.N.M. conceived the study. C.B. developed the algorithm and tested it on benchmarks and empirical datasets. C.B., S.M. and R.N.M. analyzed and interpreted the results obtained. C.B. and R.N.M. wrote the manuscript. All authors reviewed the manuscript.

ADDITIONAL INFORMATION

Accession codes A python package of the algorithm is accessible at the github web page `https://github.com/cbongiorno/svhc`; **Competing interests** The authors declare no competing interests.