Research Article

Statistics in Medicine

Received 3 December 2013, Accepted 12 August 2014 Published online 3 September 2014 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6292

A non-parametric statistical test to compare clusters with applications in functional magnetic resonance imaging data

André Fujita,[a*†] Daniel Y. Takahashi,[b,c] Alexandre G. Patriota[d] and João R. Sato[e]

Statistical inference of functional magnetic resonance imaging (fMRI) data is an important tool in neuroscience investigation. One major hypothesis in neuroscience is that the presence or not of a psychiatric disorder can be explained by the differences in how neurons cluster in the brain. Therefore, it is of interest to verify whether the properties of the clusters change between groups of patients and controls. The usual method to show group differences in brain imaging is to carry out a voxel-wise univariate analysis for a difference between the mean group responses using an appropriate test and to assemble the resulting 'significantly different voxels' into clusters, testing again at cluster level. In this approach, of course, the primary voxel-level test is blind to any cluster structure. Direct assessments of differences between groups at the cluster level seem to be missing in brain imaging. For this reason, we introduce a novel non-parametric statistical test called analysis of cluster structure variability (ANOCVA), which statistically tests whether two or more populations are equally clustered. The proposed method allows us to compare the clustering structure of multiple groups simultaneously and also to identify features that contribute to the differential clustering. We illustrate the performance of ANOCVA through simulations and an application to an fMRI dataset composed of children with attention deficit hyperactivity disorder (ADHD) and controls. Results show that there are several differences in the clustering structure of the brain between them. Furthermore, we identify some brain regions previously not described to be involved in the ADHD pathophysiology, generating new hypotheses to be tested. The proposed method is general enough to be applied to other types of datasets, not limited to fMRI, where comparison of clustering structures is of interest. Copyright © 2014 John Wiley & Sons, Ltd.

Keywords: clustering; silhouette method; statistical test; fMRI

1. Introduction

Biological datasets are growing enormously in size, leading to an information-driven science [1] and allowing previously impossible breakthroughs. More than ever, there is now an increasing need for statistical methods that can identify relevant characteristics among these large datasets. For example, in medicine, the identification of features that characterize control and disease subjects is key for the development of diagnostic procedures, prognosis, and therapy [2]. Among several exploratory methods, the study of clustering structures is a very appealing candidate method, mainly because several biological questions can be formalized in the form: Are the features of populations A and B equally clustered? One typical example occurs in neuroscience. It is thought that the brain is organized in clusters of neurons with different major functionalities, and deviations from the typical clustering pattern can lead to

[a]*Department of Computer Science, Institute of Mathematics and Statistics, University of São Paulo, São Paulo, Brazil*
[b]*Department of Psychology, Princeton University, Princeton, NJ, U.S.A.*
[c]*Neuroscience Institute, Princeton University, Princeton, NJ, U.S.A.*
[d]*Department of Statistics, Institute of Mathematics and Statistics, University of São Paulo, São Paulo, Brazil*
[e]*Center of Mathematics, Computation, and Cognition, Universidade Federal do ABC, Santo André, Brazil*
**Correspondence to: André Fujita, Department of Computer Science, Institute of Mathematics and Statistics, University of São Paulo, Rua do Mato, 1010-Building C, Cidade Universitária São Paulo, São Paulo, SP, CEP 05508-090, Brazil.*
[†]*E-mail: andrefujita@gmail.com*

a pathological condition [3]. Another example is in molecular biology, where the gene expression clustering structures depend on the analyzed population (control or tumor, for instance) [4, 5]. Therefore, in order to better understand diseases, it is necessary to differentiate the clustering structures among different populations. This leads to the problem of how to statistically test the equality of clustering structures of two or more populations followed by the identification of features that are not equally clustered. The traditional approach is to compare some descriptive statistics of the clustering structure (number of clusters, common elements in the clusters, etc.) [6–8], but to the best of our knowledge, little have been carried out regarding formal statistical methods to test the equality of clustering structures among populations. With this motivation, we introduce a new statistical test called analysis of cluster structure variability (ANOCVA) in order to statistically compare whether the items of two or more populations are equally clustered.

Our method is an extension of two well-established ideas: the silhouette statistic [9] and analysis of variance. Essentially, we use the silhouette statistic to measure the 'variability' of the clustering structure in each population. Next, we compare the silhouette among populations. The intuitive idea behind this approach is that we assume that populations with the same clustering structures also have the same 'variability'. This simple idea allows us to obtain a powerful statistic test for equality of clustering structures, which (i) can be applied to a variety of clustering algorithms; (ii) allows us to compare the clustering structure of multiple groups simultaneously; (iii) is fast and easy to implement; and (iv) identifies features that significantly contribute to the differential clustering.

We illustrate the performance of ANOCVA through simulation studies and demonstrate the power of the test in identifying small differences in clustering among populations. We also applied our method to study the whole brain functional magnetic resonance imaging (fMRI) recordings of 759 children with typical development (TD), attention deficit hyperactivity disorder (ADHD) with hyperactivity/impulsivity and inattentiveness, and ADHD with hyperactivity/impulsivity without inattentiveness. ADHD is a psychiatric disorder that usually begins in childhood and often persists into adulthood, affecting at least 5–10% of children in the US and non-US populations [10]. Given its prevalence, impacts on the children's social life, and the difficult diagnosis, a better understanding of its pathophysiology is fundamental. The statistical analysis using ANOCVA on this large fMRI dataset composed of ADHD and subjects with TD identified brain regions that are consistent with already known literature of this physiopathology. Moreover, we have also identified some brain regions previously not described as associated with this disorder, generating new hypotheses to be tested empirically.

2. Methods

We can describe our problem in the following way. Given k populations $T_1, T_2, \ldots, T_k$ where each population T_j ($j = 1, \ldots, k$), is composed of n_j subjects, and each subject has N items that are clustered in some manner, we would like to verify whether the cluster structures of the k populations are equal and, if not, which items are differently clustered. To further formalize our method, we must define what we mean by cluster structure. The silhouette statistic is used in our proposal to identify the cluster structure. We briefly describe it in the next section.

2.1. The silhouette statistic

The silhouette method was proposed in 1987 by [9] with the purpose of verifying whether a specific item was assigned to an appropriate cluster. In other words, the silhouette statistic is a measure of goodness of fit of the clustering procedure. Let $\mathcal{X} = \{x_1, \ldots, x_N\}$ be the items of one subject that are clustered into $\mathcal{C} = \{C_1, \ldots, C_r\}$ clusters by a clustering algorithm according to an optimal criterion. Note that $\mathcal{X} = \bigcup_{q=1}^{r} C_q$. Denote by $d(x, y)$ the dissimilarity (e.g., Euclidian and Manhattan) between items x and y and define

$$d(x, C) = \frac{1}{\#C} \sum_{y \in C} d(x, y) \tag{1}$$

as the average dissimilarity of x to all items of cluster $C \subset \mathcal{X}$ (or $C \in \mathcal{C}$), where $\#C$ is the number of items of C. Denote by $D_q \in \mathcal{C}$ the cluster to which x_q has been assigned by the clustering algorithm and

by $E_q \in C$ any other cluster different of D_q, for all $q = 1, \ldots, N$. All quantities involved in the silhouette statistic are given by

$$a_q = d\left(x_q, D_q\right) \quad \text{and} \quad b_q = \min_{E_q \neq D_q} d\left(x_q, E_q\right), \quad \text{for } q = 1, \ldots, N,$$

where a_q is the 'within' dissimilarity and b_q is the smallest 'between' dissimilarity for the sample unit x_q. Then a proposal to measure how well item x_q has been clustered is given by the silhouette statistic [9]

$$s_q = \begin{cases} \frac{b_q - a_q}{\max\{b_q, a_q\}}, & \text{if } \#D_q > 1, \\ 0, & \text{if } \#D_q = 1. \end{cases} \tag{2}$$

The choice of the silhouette statistic is interesting owing to its interpretations. Notice that, if $s_q \approx 1$, this implies that the 'within' dissimilarity is much smaller than the smallest 'between' dissimilarity ($a_q \ll b_q$). In other words, item x_q has been assigned to an appropriate cluster because the second-best choice cluster is not nearly as close as the actual cluster. If $s_q \approx 0$, then $a_q \approx b_q$; hence, it is not clear whether x_q should have been assigned to the actual cluster or to the second-best choice cluster because it lies equally far away from both. If $s_q \approx -1$, then $a_q \gg b_q$, so item x_q lies much closer to the second-best choice cluster than to the actual cluster. Therefore, it is more natural to assign item x_q to the second-best choice cluster instead of the actual cluster because this item x_q has been 'misclassified'. To conclude, s_q measures how well item x_q has been labeled.

Let $\mathbf{Q} = \{d(x_l, x_q)\}$ be the $(N \times N)$-matrix of dissimilarities, and then it is symmetric and has zero diagonal elements. Let $\mathbf{l} = (l_1, l_2, \ldots, l_N)$ be the labels obtained by a clustering algorithm applied to the dissimilarity matrix $\mathbf{Q}$, that is, the labels represent the cluster each item belongs to. It can be easily verified that the dissimilarity matrix $\mathbf{Q}$ and the vector of labels $\mathbf{l}$ are sufficient to compute the quantities $s_1, \ldots, s_N$. In order to avoid notational confusions, we will write $s_q^{(\mathbf{Q},\mathbf{l})}$ rather than s_q for all $q = 1, \ldots, N$, because we deal with many datasets in the next section.

2.2. *Extension of the silhouette approach*

In the previous section, we introduced notations when we have N items in one subject. In the present section, we extend the approach to many populations and many subjects in each population. Let $T_1, T_2, \ldots, T_k$ be k types of populations. For the jth population, n_j subjects are collected, for $j = 1, \ldots, k$. In order to establish notations, the items of the ith subject taken from the jth population are represented by the matrix $\mathbf{X}_{i,j} = (\mathbf{x}_{i,j,1}, \ldots, \mathbf{x}_{i,j,N})$, where each item $\mathbf{x}_{i,j,q} (q = 1, \ldots, N)$ is a vector (i.e., the item can be of any dimension. For example, each $\mathbf{x}_{i,j,q}$ can be a vector containing a time series or also the spatial position in the cartesian space).

First, we define the $(N \times N)$ matrix of dissimilarities among items of each matrix $\mathbf{X}_{i,j}$, by

$$\mathbf{A}_{i,j} = \left\{d\left(\mathbf{x}_{i,j,q}, \mathbf{x}_{i,j,q'}\right)\right\}, \quad \text{for } i = 1, \ldots, n_j, \quad j = 1, \ldots, k.$$

Notice that each $\mathbf{A}_{i,j}$ is symmetric with diagonal elements equal zero. Also, we define the following average matrices of dissimilarities:

$$\bar{\mathbf{A}}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} \mathbf{A}_{i,j} = \frac{1}{n_j} \sum_{i=1}^{n_j} \left\{d\left(\mathbf{x}_{i,j,q}, \mathbf{x}_{i,j,q'}\right)\right\} \quad \text{and} \quad \bar{\bar{\mathbf{A}}} = \frac{1}{n} \sum_{j=1}^{k} n_j \bar{\mathbf{A}}_j,$$

where $n = \sum_{j=1}^{k} n_j$, $q, q' = 1, \ldots, N$. The $(N \times N)$-matrices $\bar{\mathbf{A}}_1, \ldots, \bar{\mathbf{A}}_k$ and $\bar{\bar{\mathbf{A}}}$ are the only quantities required for proceeding with our proposal.

Now, based on the matrix of dissimilarities $\bar{\bar{\mathbf{A}}}$, we can use a clustering algorithm to find the clustering labels $\mathbf{l}_{\bar{\bar{\mathbf{A}}}}$. Then, we compute the following silhouette statistics:

$$s_q^{\left(\bar{\bar{\mathbf{A}}}, \mathbf{l}_{\bar{\bar{\mathbf{A}}}}\right)} \quad \text{and} \quad s_q^{\left(\bar{\mathbf{A}}_j, \mathbf{l}_{\bar{\bar{\mathbf{A}}}}\right)}, \quad \text{for } q = 1, \ldots, N.$$

The former is the silhouette statistic based on the matrix of dissimilarities $\bar{\bar{\mathbf{A}}}$, and the latter is the silhouette statistic based on the dissimilarity matrix $\bar{\mathbf{A}}_j$, both obtained by using the clustering labels computed

via the matrix $\bar{\bar{\mathbf{A}}}$. We expect that if the items from all populations $T_1, \ldots, T_k$ are equally clustered, the quantities $s_q^{\left(\bar{\bar{\mathbf{A}}}, \mathbf{l}_{\bar{\bar{\mathbf{A}}}}\right)}$ and $s_q^{\left(\bar{\mathbf{A}}_j, \mathbf{l}_{\bar{\mathbf{A}}}\right)}$ must be close for all $j = 1, \ldots k$ and $q = 1, \ldots, N$.

2.3. *Statistical tests*

Given a clustering algorithm and a distance metric, define the following vectors:

$$\mathbf{S} = \left(s_1^{\left(\bar{\bar{\mathbf{A}}}, \mathbf{l}_{\bar{\bar{\mathbf{A}}}}\right)}, \ldots, s_N^{\left(\bar{\bar{\mathbf{A}}}, l_{\bar{\bar{\mathbf{A}}}}\right)} \right)^\top \quad \text{and} \quad \mathbf{S}_j = \left(s_1^{\left(\bar{\mathbf{A}}, \mathbf{l}_{\bar{\mathbf{A}}}\right)}, \ldots, s_N^{\left(\bar{\mathbf{A}}, \mathbf{l}_{\bar{\mathbf{A}}}\right)} \right)^\top ,$$

and let $\mathbf{l}_j$ ($j = 1, \ldots, k$) be the vector of labels for the jth population with $\mathbf{l}_j = \mathbf{l}_{\mathbf{A}_j^+}$, where $\mathbf{A}_j^+$ is the average matrix of dissimilarities for the jth population, which is defined as follows. Here, we consider that $x_{1,j,q}, \ldots, x_{n_j,j,q}$ are independent and identically distributed for each combination of $j = 1, \ldots, k$ and $q = 1, \ldots, N$. Assume that $\mathbb{E}\left(d\left(\mathbf{x}_{1,j,q}, \mathbf{x}_{1,j,q'}\right)\right) < \infty$, for all j, q, and q', and then we define $\left\{\mathbf{A}_j^+\right\}_{q,q'} = \mathbb{E}\left(d\left(x_{1,j,q}, x_{1,j,q'}\right)\right)$. Notice that, by the strong law of large numbers, the sample average matrix of dissimilarities $\bar{\mathbf{A}}_j$ converges almost surely to $\mathbf{A}_j^+$.

Let us define $\delta S_j = \mathbf{S} - \mathbf{S}_j$. We want to test if all k populations are equally clustered (present the same clustering structure), that is,

H_0 : $\mathbb{E}(\delta S_1) = \cdots = \mathbb{E}(\delta S_k) = 0$, i.e., 'Given the clustering algorithm, the data from $T_1, T_2, \ldots, T_k$ are equally clustered'.
versus
H_1 : 'At least one is clustered in a different manner'.

We will use the statistic $\Delta S = \sum_{j=1}^k \delta S_j^\top \delta S_j$ to build our test statistic. Notice that under the null hypothesis, all N items are equally clustered along the k populations. Therefore, $s_q^{\left(\bar{\bar{\mathbf{A}}}, \mathbf{l}_{\bar{\bar{\mathbf{A}}}}\right)} \approx s_q^{\left(\bar{\bar{\mathbf{A}}}, \mathbf{l}_{\bar{\bar{\mathbf{A}}}}\right)}$ for all $q = 1, \ldots, N$, and consequently, it is expected to obtain small ΔS while large ΔS suggests a rejection of the null hypothesis.

Now, suppose that the null hypothesis is rejected by the previous test. Thus, a natural next step is to identify which item is not equally clustered among populations.

Let us define $\delta s_q = s_q^{\left(\bar{\bar{\mathbf{A}}}, \mathbf{l}_{\bar{\bar{\mathbf{A}}}}\right)} - \frac{1}{k}\sum_{j=1}^k s_q^{\left(\bar{\mathbf{A}}_j, \mathbf{l}_{\bar{\bar{\mathbf{A}}}}\right)}$, for $q = 1, \ldots, N$. Then, the test can be defined as

H_0 : $\mathbb{E}(\delta s_q) = 0$, i.e., 'Given the clustering algorithm, the qth item ($q = 1, \ldots, N$) is equally clustered among populations'.
versus
H_1 : 'The qth item is not equally clustered among populations'.

We will use the statistic $\Delta s_q = \delta s_q^2$, for $q = 1, \ldots, N$ to build the test statistic. Again, under the null hypothesis, we expect small Δs_q while large Δs_q suggests a rejection of the null hypothesis.

The exact or asymptotic distributions of both ΔS and Δs_q are not trivial; therefore, we use a computational procedure based on bootstrap [11] to construct the empirical null distributions. The bootstrap implementation of both tests is as follows:

(1) Resample with replacement n_j subjects from the entire dataset $\{T_1, T_2, \ldots, T_k\}$ in order to construct bootstrap samples T_j^*, for $j = 1, \ldots, k$.
(2) Calculate $\bar{\mathbf{A}}_j^*$, $\bar{\bar{\mathbf{A}}}^*$, $s_q^{\left(\bar{\bar{\mathbf{A}}}, \mathbf{l}_{\bar{\bar{\mathbf{A}}}}\right)^*}$ and $s_q^{\left(\bar{\mathbf{A}}, \mathbf{l}_{\bar{\bar{\mathbf{A}}}}\right)^*}$, for $q = 1, \ldots, N$, using the bootstrap samples T_j^*.
(3) Calculate $\hat{\Delta S}^*$ and $\hat{\Delta s}_q^*$.
(4) Repeat steps 1 to 3 until the desired number of bootstrap replications is obtained.
(5) The p-values from the bootstrap tests based on the observed statistics ΔS and Δs_q are the fraction of replicates of $\hat{\Delta S}^*$ and $\hat{\Delta s}_q^*$ on the bootstrap dataset T_j^*, respectively, that are at least as large as the observed statistics on the original dataset.

The data analysis can be described as shown in Figure 1.

2.4. *Simulation description*

Here, we study both the control of the rate of false positives and the sensitivity of ANOCVA in verifying whether items are equally clustered between two populations.

In this scenario (Figure 2), there are two populations. Each subject of each population is composed of $N = 20$ items clustered into two clusters of equal sizes (i.e., each cluster has 10 items). The items of clusters 1 and 2 are generated by bivariate normal distributions with mean $(0, 0)$ and $(2, 0)$, respectively, and with an identity covariance matrix. All items are well clustered except by the 10th item of cluster 1

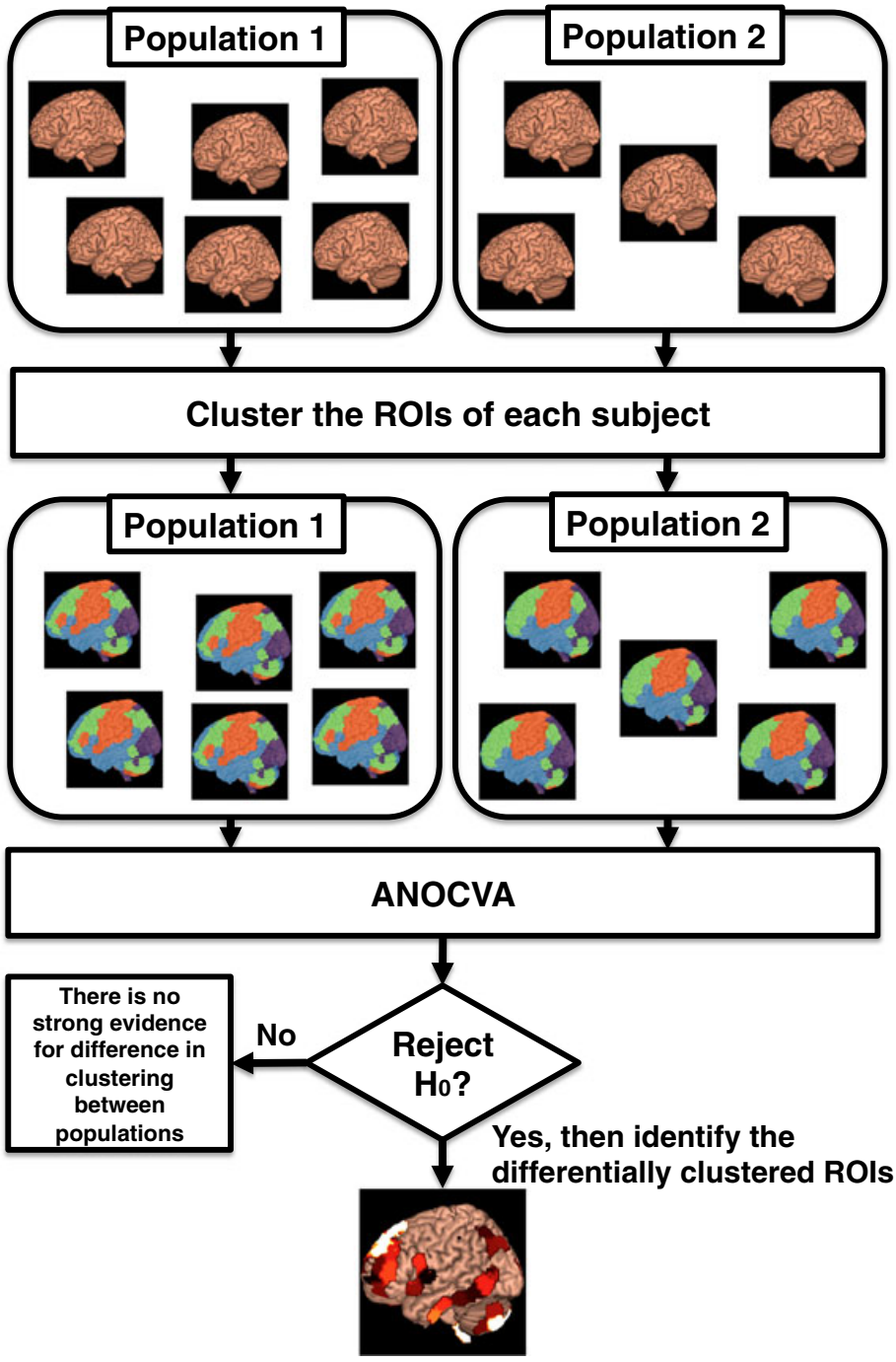

Figure 1. Data analysis pipeline. The analysis consists in clustering the regions of interest (ROIs) of each subject and then testing by ANOCVA (analyzing the test statistic ΔS) whether the ROIs are equally clustered between populations. If they are not equally clustered, that is, the null hypothesis (H_0) is rejected, the ROIs that most contribute to this differential clustering can be identified by using the test statistic Δs_q.

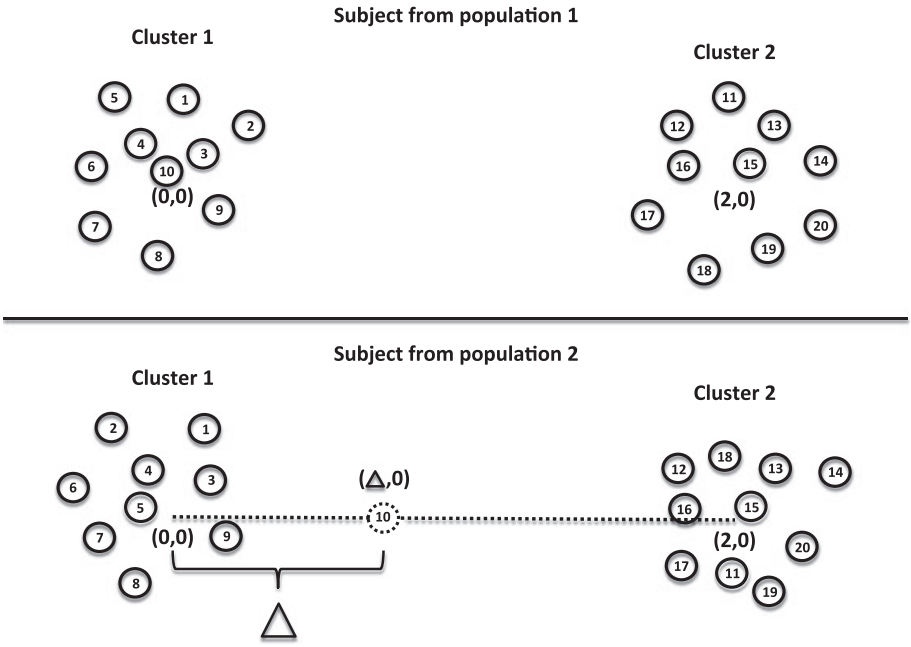

Figure 2. Illustration of the simulation. This figure illustrates the items of one subject from population 1 and one subject from population 2. There are two clusters, one centered at $(0, 0)$ and another at $(2, 0)$. Each cluster is composed of 10 items. The 10th item 'moves' from cluster 1 to cluster 2. Δ is the distance of the 10th item from the center of its original cluster.

that 'moves' from position $(0, 0)$ (center of cluster 1) in direction to position $(2, 0)$ (center of cluster 2). The distance between this item and the center of its original cluster $(0, 0)$ is given by Δ. We performed 100 Monte Carlo realizations of this scenario for all combinations of parameters and number of subjects ($n_1 = n_2 = 10, 20, 30, 40$) and $\Delta = 0, 0.25, 0.50, \ldots, 2$.

The clustering algorithm and the dissimilarity measure used to this simulation are the complete linkage hierarchical clustering procedure and the Euclidian distance, respectively.

2.5. ADHD data description

ANOCVA was applied to an fMRI dataset composed of children with TD and with ADHD, both under a resting state protocol, totaling 759 subjects. This dataset is available at the ADHD-200 Consortium [12] website (http://fcon_1000.projects.nitrc.org/indi/adhd200/). fMRI data were collected in eight sites that compose the ADHD-200 consortium, and data collection was conducted with local internal review board approval, and also in accordance with local internal review board protocols. Further details about this dataset can be obtained from the ADHD-200 consortium website.

This dataset is composed of 479 controls—children with TD (253 boys, mean age $\pm$ standard deviation of 12.23 ± 3.26 years) and three sub-groups of ADHD patients: (i) combined—hyperactive/impulsive and inattentive—(159 children, 130 boys, 11.24 ± 3.05 years); (ii) hyperactive/impulsive (11 subjects, 9 boys, 13.40 ± 4.51 years); and (iii) inattentive (110 subjects, 85 boys, 12.06 ± 2.55 years).

2.6. ADHD data pre-processing

The pre-processing of the fMRI data was performed by applying the Athena pipeline (http://www.nitrc.org/plugins/mwiki/index.php/neurobureau:AthenaPipeline). The pre-processed data are publicly available at the Neurobureau website (http://neurobureau.projects.nitrc.org/ADHD200). Briefly, the steps of the pipeline are as follows: exclusion of the first four scans; slice timing correction; deoblique dataset; correction for head motion; masking the volumes to discard non-brain voxels; co-registration of mean image to the respective anatomic image of the children; spatial normalization to MNI space (resampling to 4 mm $\times$ 4 mm $\times$ 4 mm resolution); removing effects of white matter, cerebrospinal fluid, head motion (six parameters) and linear trend using linear multiple regression; temporal band-pass filter ($0.009 < f < 0.08$ Hz); spatial smoothing using a Gaussian filter (full width at half maximum = 6 mm). The CC400 atlas (based on the approach described in [13]) provided by the Neurobureau was used to

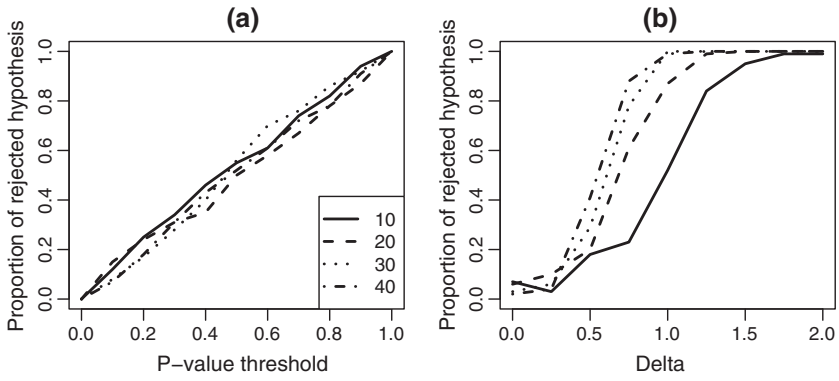

Figure 3. (a) The x-axis represents the p-value threshold. (b) The x-axis represents the distance (Δ) of one of the items from the center of the original cluster. The y-axis represents the proportion of rejected null hypotheses in 100 repetitions carried out for each number of subjects ($n_1 = n_2 = 10, \ldots, 40$). Different lines (three dashed lines and one solid line) represent different numbers of subjects in each population. Notice that under the null hypothesis, the rate of false positives is as expected by the p-value threshold (a). Moreover, the greater the number of subjects in each population (or the distance (Δ)), the greater the power to reject the null hypothesis (b).

define the 351 regions of interest (ROIs) used in this study. The average signal of each ROI was calculated and used as representative of the region.

For each child, a correlation matrix was constructed by calculating Spearman's correlation coefficient (which is robust against outliers and suitable to identify monotonic non-linear relationships) among the 351 ROIs (items) in order to identify monotonically dependent ROIs. Then, the correlations' matrices were corrected for site effects by using a general linear model (GLM). Site effects were modeled as a GLM (site as a categorical variable), and the effect was removed by considering the residuals of this model. p-values corresponding to Spearman's correlation for each pair of ROIs were calculated. Then, the obtained p-values were corrected by false discovery rate (FDR) [14]. Thus, the dissimilarity matrices $\mathbf{A}_j$ for $j = 1, \ldots, 759$ are symmetric with diagonal elements equal zero and non-diagonal elements ranging from zero to one. The higher the correlation, the lower the p-value, and consequently, the lower the dissimilarity between two ROIs. Notice that the p-value associated with each Spearman's correlation is not used as a statistical test, but only as a measure of dissimilarity normalized by the variance of the ROIs. The choice of the proposed dissimilarity measure instead of the standard one minus the correlation coefficient is because we are interested in ROIs that are highly correlated, independent whether they present positive or negative correlation. Here, we are interested in calculating how much ROIs are dependent (the dissimilarity among them) and not how they are correlated.

3. Results

3.1. Simulation analysis

For each $\Delta = 0, 0.25, 0.50, \ldots, 2$, one hundred Monte Carlo realizations were constructed and tested by our approach. The results obtained by simulations are illustrated in Figure 3. Under the null hypothesis ($\Delta = 0$), Figure 3(a) shows that the test actually controls the rate of type I error. In other words, the proportion of falsely rejected null hypotheses as expected by the p-value threshold. Figure 3(b) illustrates the power of the test under different Δ. The higher the Δ, the higher the power. Moreover, it is also possible to see that as the number of subjects ($n_1 = n_2 = 10, \ldots, 40$) in each population increases, the power also increases.

Figure 4(a, b) depicts illustrative examples of one realization of the scenario described in Section 2.4 with $\Delta = 0$ and one with $\Delta = 2$, respectively. The x-axis represents the items from 1 to 20. The y-axis represents the z-scores of the p-values (computed for each Δs_q for $q = 1, \ldots, N$) corrected for multiple comparisons by the FDR method [14]. Items with z-score higher than 1.96 represent the statistically significant ones at a p-value threshold of 0.05. Notice that, as expected, Figure 4(a) does not present any item as statistically significant because all the items are equally clustered between the two populations ($\Delta = 0$). Figure 4(b) highlights the 10th item as statistically significant, which is exactly the item that

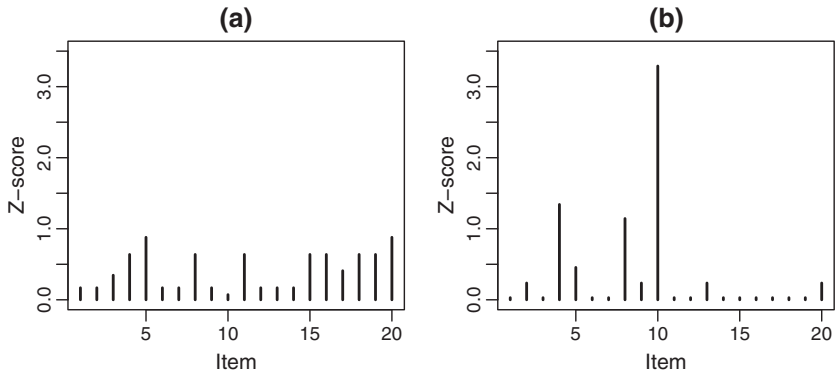

Figure 4. Statistically significant items. (a) and (b) represent how much the items 1 to 20 contribute to the differential clustering with $\Delta = 0$ (null hypothesis) and $\Delta = 2$ (alternative hypothesis), respectively. Items with z-score higher than 1.96 represent the statistically significant items at a p-value threshold of 0.05 after FDR correction for multiple comparisons. Notice that, as expected, (a) does not present any items as statistically significant. (b) presents the 10th item as statistically significant.

Table I. ANOCVA applied to the ADHD dataset.

Comparison	p-value
TD vs Combined ADHD vs Inattentive ADHD	0.020
TD vs Combined ADHD and Inattentive ADHD	<0.001
TD vs Combined ADHD	<0.001
TD vs Inattentive ADHD	0.700
Combined ADHD vs Inattentive ADHD	0.615

The number of bootstrap samples is set to 1000. p-values are corrected by Bonferroni method for multiple comparisons.

moved from one cluster to another. Therefore, by analyzing the statistic Δs_q for $q = 1, \ldots, N$, it is possible to identify which item is not equally clustered among populations.

For a better comprehension of the properties of ANOCVA, additional scenarios (e.g., when one item 'jumps' from one cluster in one population to another cluster in another population, or when the number of clusters change between two populations) were simulated and studied. Results are illustrated in the Supporting Information. For the analyzed scenarios, ANOCVA indeed discriminates populations that present different clustering structures and also identifies items that are clustered in a different manner.

3.2. ADHD data analysis

ANOCVA was applied to the ADHD dataset, in order to identify ROIs associated with the disease. As we are interested in identifying ROIs that are not equally clustered in terms of their connectivity, the clustering algorithm used to determine the labels based on the dissimilarity matrix $\bar{\bar{\mathbf{A}}}$ was the unnormalized spectral clustering algorithm [15]. The numbers of clusters for each group were estimated by the slope method [16]. The number of bootstrap samples was set to 1 000. The group of children with hyperactive/impulsive ADHD was excluded from our analysis owing to the low number of subjects (11 children). The performed tests and the respective p-values corrected for multiple comparisons by the Bonferroni method are listed in Table I. First, the test was applied to the entire dataset (excluded the group of hyperactive/impulsive ADHD owing to the low number of subjects) in order to verify if there is at least one population that differs from the others. The test indicated a significant difference (p-value = 0.020), suggesting that there is at least one population that is not equally clustered. In order to identify which populations are not equally clustered, pairwise comparisons among the groups were carried out. By observing Table I, it is not possible to verify significant differences between TD versus inattentive ADHD (p-value = 0.700), and combined ADHD versus inattentive ADHD (p-value = 0.615), but there are significant differences between TD versus combined and inattentive ADHD (p-value < 0.001), and TD versus combined ADHD (p-value < 0.001). These results indicate that the significant differences obtained for TD, combined ADHD, and inattentive ADHD were probably due to the differences between TD and combined ADHD.

Thus, the analysis of the fMRI dataset was focused on identifying the differences between children with TD and children with combined ADHD. The results of the clustering procedure are visualized in Figure 5, where (a) and (b) represent children with TD and children with combined ADHD, respectively. Interestingly, the clusters were composed of anatomically contiguous and almost symmetric areas of the brain, although these constraints were not included *a priori* in our analyses. This is consistent with the hypothesis that the spectral clustering method groups areas with similar brain activities in the same cluster.

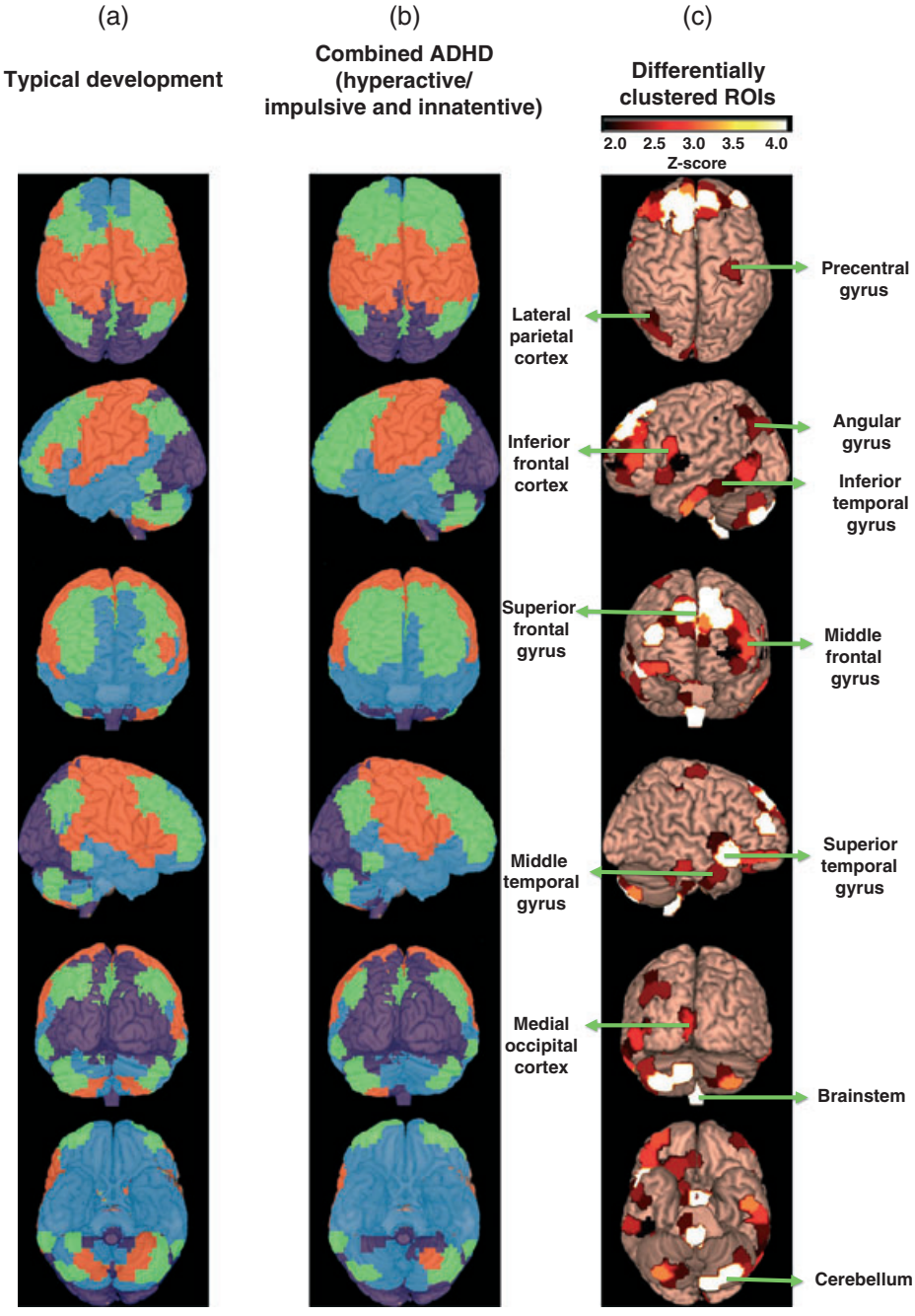

Figure 5. Clustering of ROIs. (a) and (b) represent the ROIs clustered by the spectral clustering algorithm applied to the dissimilarity matrices $\bar{\mathbf{A}}_{\text{TD}}$ and $\bar{\mathbf{A}}_{\text{combined ADHD}}$, respectively. The number of clusters was estimated by the silhouette method. (c) highlights the ROIs that are not equally clustered between children with TD and ADHD. Regions highlighted in white represent high z-scores, while regions in red represent lower z-scores. The z-scores were calculated by using the p-values obtained by ANOCVA after FDR correction for multiple comparisons.

Then, each ROI was tested in order to identify the one that significantly contribute to the difference in clustering between children with TD and with combined ADHD. p-values were corrected for multiple comparisons by the FDR method and then, converted to z-scores. Figure 5(c) illustrates the statistically significant ROIs at a p-value threshold of 0.05 after FDR correction. The regions highlighted in white are the ROIs with the highest z-scores, while the regions highlighted in red represent ROIs with lower z-scores, but still statistically significant.

By comparing Figure 5(a) with (b), it is possible to verify that the highlighted regions in Figure 5(c) correspond to ROIs that are not equally clustered between children with TD and with combined ADHD. Cluster analysis has suggested a very similar network organization between children with TD and combined ADHD patients. Apparently, sensory-motor systems, frontoparietal control networks, visual processing, and fronto-temporal systems are similarly distributed between the two groups. However, the application of ANOCVA unveiled that anterior portion of inferior, middle and superior frontal gyri, inferior temporal gyrus, angular gyrus, and some regions of cerebellum, lateral parietal, medial occipital, and somato-motor cortices have a distinct clustering organization between the two populations.

Motor system (precentral and postcentral giri and cerebellum) alterations in ADHD are associated with hyperactivity symptoms, and this finding has already been extensively described in the literature (for a detailed review, see [17]). In addition, Dickstein *et al*. [18] carried out a meta-analysis of fMRI studies comparing controls and ADHD and identified some portions of parietal cortex, inferior prefrontal cortex, and primary motor cortex as regions with activation differences between the groups.

The inferior frontal cortex highlighted by ANOCVA is described in literature as a key region for inhibition of responses [19]. In this sense, the impulsivity symptoms present in combined ADHD can be related to an abnormal participation of this region in the context of global brain network organization, when compared with healthy controls. This finding is reinforced by the findings of recent studies. Schulz *et al*. [20] investigated the role of this area in therapeutic mechanisms of treatments for ADHD. In addition, Vasic *et al*. [21] and Whelan *et al*. [22] explored the neural error signaling in these regions (in adults) and the impulsivity of adolescents with ADHD. The inferior frontal cortex is also implied to participate in language production, comprehension, and learning, and therefore, our finding is consistent with the language impairment reported in ADHD subjects [23].

An interesting finding from the application of the proposed method to the resting state fMRI dataset was the identification of angular gyrus as a region with functional abnormalities in ADHD in the context of brain networks. Although the angular gyrus contributes to the integration of information, playing an important role in many cognitive processes [24], to the best of our knowledge, there are very few studies in literature suggesting activation differences in this region when comparing ADHD with healthy controls [24,25]. The angular gyrus contributes to the integration of information and plays a role in many cognitive processes [24]. Tamm *et al*. [25] have found that this region exhibited less activation in adolescents with ADHD during a target detection task. Moreover, Simos *et al*. [26] have shown that angular and supramargial gyri play a role in brain mechanisms for reading and its correlates in ADHD. We note that despite that the angular gyrus has a crucial role in the integration of information, both studies have not explored the relevance of this region from a network connectivity perspective. Using the proposed method, we properly carried out this analysis, and the findings indicate that the role of this region in brain spontaneous activity is different, when comparing the two groups.

One point to be noticed is the fact that several regions such as precentral gyrus, inferior frontal cortex, and brain stem are assigned by the clustering algorithm to be in the same cluster between TD and ADHD populations, although they were identified by ANOCVA to be clustered in a different manner. This fact can be explained as follows: ANOCVA analyzes how much the silhouette statistic differs among different populations. Thus, suppose that we have one item that is located in the center of one cluster in population one. Therefore, its silhouette statistic is close to one because it is well clustered. In population two, suppose that the same item is located in the border of two clusters. In this case, its silhouette statistic is close to zero, but it still belongs to the same cluster as in population 1. In this situation, the clustering label of this item is the same in both populations, but their silhouette statistics differ significantly, and consequently, this item is identified by ANOCVA as clustered in a different manner.

The existence of temporal and spatial correlation is inherent in fMRI data, and ignoring intrinsic correlation may lead to misleading or erroneous conclusions. This dependence structure makes clustering analysis more challenging and should be accounted for [27, 28]. Notice that the proposed bootstrap incorporates the spatial correlations in the clustering process and also preserves the temporal structure.

In order to certify that the bootstrap based statistical test is correctly working in actual data and that the results obtained in this study are not due to numerical fluctuations or another source of error that was

not taken into account, we verified the control of the rate of false positives in biological data. The set of 479 children with TD was split randomly into two subsets, and the clustering test was applied between them. This procedure was repeated 700 times. The proportion of falsely rejected null hypothesis for p-values lower than 1%, 5%, and 10% were 2.14%, 5.70%, and 9.83%, respectively, confirming that the type I error is effectively controlled in this biological dataset. Moreover, the Kolmogorov–Smirnov test was applied to compare the p-values' distribution obtained in the 700 repetitions with the uniform distribution. The Kolmogorov–Smirnov test indicated a p-value of 0.664, meaning that there is no statistical evidence to reject the null hypothesis that the p-values' distribution obtained in 700 repetitions follows a uniform distribution. Furthermore, we also verified in the same manner whether the highlighted ROIs are indeed statistically significant. The proposed method was applied to each ROI, that is, 351 p-values were calculated in each repetition. Thus, 351 p-values' distributions, one for each ROI, were constructed. Each of the 351 p-values' distribution was compared with the uniform distribution by the Kolmogorov–Smirnov test. After correcting the obtained 351 p-values by the Kolmogorov–Smirnov test by FDR [14], only two null hypotheses were rejected at a p-value threshold of 0.05, confirming that the type I error is also controlled for ROIs. These results suggest that the differences in clustering between children with TD and with combined ADHD are indeed statistically significant.

4. Final remarks

By both simulations and an application in an actual dataset, we showed that the method proposed here statistically identifies differences in the clustering structure of two or more populations of subjects simultaneously. However, it is important to discuss some limitations of ANOCVA. First, the method is only defined if the estimated number of clusters for the average of the matrix of dissimilarities $\bar{\bar{\mathbf{A}}}$ is greater than one. It is because the silhouette statistic s is only defined when the number of clusters is greater than one. In practice, one may test whether each dissimilarity matrix $\bar{\mathbf{A}}_j$ and the average of dissimilarity matrix $\bar{\bar{\mathbf{A}}}$ are composed of one or more clusters by using the gap statistic proposed by [29]. If $\bar{\bar{\mathbf{A}}}$ is composed of one cluster while one of the matrices of dissimilarities $\bar{\mathbf{A}}_j$ is composed of more than one cluster, clearly, the clustering structures among populations are different. Another limitation is the fact that ANOCVA does not identify changes in both rotated and translated data. In other words, the test does not identify alterations that maintain the relative dissimilarities among items. If one is interested in identifying this kind of difference, one simple solution is to test the joint mean by Hotteling's T-squared test [30]. However, it is important to point out that ANOCVA is sensitive to identify different clustering assignments that have the same goodness of fit (silhouette). Notice that [9] proposed the use of the average of s_q in order to obtain a goodness of fit. Here, we do not use the average value but the distance between the entire vectors $s_q^{\left(\bar{\bar{\mathbf{A}}},\mathbf{l}_{\bar{\bar{\mathbf{A}}}}\right)}$ and $s_q^{(\bar{\mathbf{A}}_j,\mathbf{l}_{\bar{\mathbf{A}}})}$. In other words, we take into account the label of the items that are clustered. Therefore, if one or more items are not equally clustered among populations, our statistic ΔS is able to capture this difference. However, the use of the average value of $s_q^{\left(\bar{\bar{\mathbf{A}}},\mathbf{l}_{\bar{\bar{\mathbf{A}}}}\right)}$ and $s_q^{(\bar{\mathbf{A}}_j,\mathbf{l}_{\bar{\mathbf{A}}})}$ is not. Moreover, ANOCVA requires a considerable number of subjects in each group to be able to reject the null hypothesis (when the clustering structures are in fact different). It is very difficult to define a minimum number of subjects because it depends on the variance, but we suppose that an order of dozens (based on our simulations) is necessary.

The use of appropriate clustering algorithm and distance measure is essential to obtain reasonable results with ANOCVA. As the definition of a cluster depends on the problem of interest—clusters are usually determined as the result of the application of a clustering algorithm (e.g., k-means, hierarchical clustering, and spectral clustering)—the results obtained by ANOCVA may change in function of the clustering procedure and the chosen metric (e.g., Euclidean and Manhattan). The selection of both clustering algorithm and metric depends essentially on the type of data, what is clustered, and the hypothesis to be tested. ANOCVA assumes that the clustering algorithm is 'correctly' clustering the items. Therefore, if the output of the clustering algorithm is not correct (reasonable), the statistical test performed by ANOCVA may fail or provide unexpected results. For example, suppose that the dataset is composed of two populations generated by non-Gaussian distributions. If the items of this dataset are clustered by an algorithm that assumes mixture of Gaussian distributions and one considers that a cluster is a probability distribution, the obtained clustering labels will be wrong, not because of the clustering algorithm but because of the incorrect choice of the clustering algorithm. In this case, it is clear that ANOCVA will not work correctly because the clustering labels provided by the clustering algorithm are not 'correct'.

There are other measures for similarity between clustering structures [31, 32] that might be used to develop a statistical test. However, these similarity measures cannot be extended in a straightforward manner to simultaneously test more than two populations. The work proposed by Alexander-Block *et al*. [33] was successfully applied in neuroscience to statistically test differences in network community structures. However, again, this method cannot test simultaneously more than two populations. Notice that in our case, we are interested in comparing controls and several sub-groups of ADHD simultaneously. Therefore, this method is not applicable. It is necessary to point out that the characteristic of ANOCVA that allows to statistically test whether the structure of the clusters of several populations—not limited to pairwise comparisons—is all equal avoids the increase in type I error due to multiple tests. Another advantage is the use of a bootstrap approach to construct the empirical null hypothesis distribution from the original data. It allows the application of the test to datasets that the underlying probability distribution is unknown. Moreover, it is also known that for actual datasets, the bootstrap procedure provides better control of the rate of false positives than asymptotic tests [34].

One question that remains is what is the difference between ANOCVA and a test for equality of dissimilarities, for example, a *t*-test for each distance. The main difference is that, for a *t*-test, only the mean and variance of the measure to be tested are taken into account to determine whether two items are 'far' or 'close', while ANOCVA is a data-based metric that uses the clustering structure to determine how 'far' the items are from others. Furthermore, we also remark that testing whether the data are equally distributed is not the same of testing whether the data are equally clustered, as items may come from very different distributions and be clustered in a quite similar way depending on the clustering algorithm. One may also ask the difference between an *F*-test of the clustering coefficient and ANOCVA. Notice that the clustering coefficient is a measure of degree to which nodes in a graph tend to cluster together, while ANOCVA tests whether the structure of the clusters of several populations is all equal.

Here, the purpose of the test was to identify ROIs that are associated with ADHD, but the same analysis can be extended to other large datasets. Specifically in neuroscience, the recent generation of huge amounts of data in collaborative projects, such as the Autism Brain Image Data Exchange project (http://fcon_1000.projects.nitrc.org/indi/abide/index.html), which generated fMRI data of more than 1000 individuals with autism, the ADHD-200 project [12] previously described here that provides fMRI data of $\approx$ 700 children with ADHD, the fMRI Data Center, which is a public repository for fMRI (http://www.fmridc.org/f/fmridc) [35], the Alzheimer's Disease Neuroimaging Initiative, which collected magnetic resonance imaging of $\approx$800 subjects [36], and many others that will certainly be produced owing to the decreasing costs in data acquisition, makes cluster analysis techniques indispensable to mine information useful to diagnosis, prognosis, and therapy.

The flexibility of our approach that allows the application of the test on several populations simultaneously (instead of limiting to pairwise comparisons), along with its performance demonstrated in both simulations and actual biological data, will make it applicable to many areas where clustering structure is of interest.

Acknowledgements

A. F. was partially supported by São Paulo Research Foundation—FAPESP (2011/50761-2, 2013/01715-3, and 2014/09576-5), CNPq (306319/2010-1 and 473063/2013-1), and NAP eScience—PRP—USP. D. Y. T. was partially supported by Pew Latin American Fellowship and Ciência sem Fronteiras Fellowship (CNPq 246778/2012-1). A. G. P. was partially supported by FAPESP (2012/21788-2). J. R. S. was supported by FAPESP (2013/10498-6).

References

1. Stein LD. Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges. *Nature Reviews Genetics* 2008; **9**:678–688.
2. Rubinov M, Sporn O. Complex network measures of brain connectivity: uses and interpretations. *NeuroImage* 2010; **52**:1059–1069.
3. Grossberg S. The complementary brain: unifying brain dynamics and modularity. *Trends in Cognitive Sciences* 2000; **4**:233–246.
4. Furlan D, Carnevali IW, Bernasconi B, Sahnane N, Milani K, Cerutti R, Bertolini V, Chiaravalli AM, Bertoni F, Kwee I, Pastorino R, Carlo C. Hierarchical clustering analysis of pathologic and molecular data identifies prognostically and biologically distinct groups of colorectal carcinomas. *Modern Pathology* 2011; **24**:126–37.
5. Wang YK, Print CG, Crampin EJ. Biclustering reveals breast cancer tumour subgroups with common clinical features and improves prediction of disease recurrence. *BMC Genomics* 2013; **14**:102.

6. Meila M. Comparing clusterings—an information based distance. *Journal of Multivariate Analysis* 2007; **98**:873–895.
7. Cecchi GA, Garg R, Rao AR. A cluster overlap measure for comparison of activations in fMRI studies. *Lecture Notes in Computer Science* 2009; **5761**:1018–1025.
8. Kluger Y, Basri R, Chang JT, Gerstein M. Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Research* 2003; **13**:703–716.
9. Rousseeuw P. Sihouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 1987; **20**:53–65.
10. Fair DA, Dosenbach NU, Church JA, Cohen AL, Brahmbhatt S. Development of distinct control networks through segregation and integration. *Proceedings of the National Academy of Sciences of the United States of America* 2007; **104**: 13507–13512.
11. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Chapman & Hall/CRC: Boca Raton, 1994.
12. The ADHD-200 Consortium. The ADHD-200 Consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. *Frontiers in Systems Neuroscience* 2012; **6**:62.
13. Craddock RC, James GA, Holtzheimer PE, Hu XP, Mayberg HS. A whole brain fMRI Atlas generated via spatial constrained spectral clustering. *Human Brain Mapping* 2012; **33**:1914–1928.
14. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B* 1995; **57**:289–300.
15. Ng A, Jordan M, Weiss Y. On spectral clustering: analysis and an algorithm. In *Advances in Neural Information Processing Systems*, Dietterich T, Becker S, Ghahramani Z (eds), Vol. 14. MIT Press, 2002; 849–856.
16. Fujita A, Takahashi DY, Patriota AG. A non-parametric method to estimate the number of clusters. *Computational Statistics & Data Analysis* 2014; **73**:27–39.
17. Castellanos FX, Proal E. Large-scale brain systems in ADHD: beyond the prefrontal-striatal model. *Trends in Cognitive Sciences* 2012; **16**:17–26.
18. Dickstein SG, Bannon K, Castellanos FX, Milham MP. The neural correlates of attention deficit hyperactivity disorder: an ALE meta-analysis. *Journal of Child Psychology and Psychiatry* 2006; **47**:1051–1062.
19. Aron AR, Robbins TW, Poldrack RA. Inhibition and the right inferior frontal cortex. *Trends in Cognitive Sciences* 2004; **8**:170–177.
20. Schulz KP, Fan J, Bédard AC, Clerkin SM, Ivanov I, Tang CY, Halperin JM, Newcorn JH. Common and unique therapeutic mechanisms of stimulant and nonstimulant treatments for attention-deficit/hyperactivity disorder. *Archives of General Psychiatry* 2012; **1**:952–961.
21. Vasic N, Plichta MM, Wolf RC, Fallgatter AJ, Sosic-Vasic Z, Gron G. Reduced neural error signaling in left inferior prefrontal cortex in young adults with ADHD. *Journal of Attention Disorders* 2014; **18**(8):659–670.
22. Whelan R, Conrod PJ, Poline JB, Lourdusamy A, Banaschewski T, Barker GJ, Bellgrove MA, Bchel C, Byrne M, Cummins TD, Fauth-Bhler M, Flor H, Gallinat J, Heinz A, Ittermann B, Mann K, Martinot JL, Lalor EC, Lathrop M, Loth E, Nees F, Paus T, Rietschel M, Smolka MN, Spanagel R, Stephens DN, Struve M, Thyreau B, Vollstaedt-Klein S, Robbins TW, Schumann G, Garavan H, IMAGEN Consortium. Adolescent impulsivity phenotypes characterized by distinct brain networks. *Nature Neuroscience* 2012; **15**:920–925.
23. Tirosh E, Cohen A. Language deficit with attention-deficit disorder: a prevalent comorbidity. *Journal of Child Neurology* 1998; **13**:493–497.
24. Seghier ML. The angular gyrus: multiple functions and multiple subdivisions. *Neuroscientist* 2013; **19**:43–61.
25. Tamm L, Menon V, Reiss AL. Parietal attentional system aberrations during target detection in adolescents with attention deficithyperactivity disorder: event-related fMRI evidence. *American Journal of Psychiatry* 2006; **163**:1033–1043.
26. Simos PG, Rezaie R, Fletcher JM, Juranek J, Passaro AD, Li Z, Cirino PT, Papanicolaou AC. Functional disruption of the brain mechanism for reading: effects of comorbidity and task difficulty among children with developmental learning problems. *Neuropsychology* 2011; **25**:520–534.
27. Kang H, Ombao H, Linkletter C, Long N, Badre D. Spatio-spectral mixed-effects model for functional magnetic resonance imaging data. *Journal of the American Statistical Association* 2012; **107**:568–577.
28. Liao W, Chen H, Yang Q, Lei X. Analysis of fMRI data using improved self-organizing mapping and spatio-temporal metric hierarchical clustering. *IEEE Transactions on Medical Imaging* 2008; **10**:1472–1483.
29. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2002; **63**:411–423.
30. Hotteling H. The generalization of Student's ratio. *Annals of Mathematical Statistics* 1931; **2**:360–378.
31. Bae E, Bailey J, Dong G. Clustering similarity comparison using density profiles. *Proceedings of the 19th Australian Joint Conference on Artificial Intelligence: Advances in Artificial Intelligence*, Hobart, 2006, 342–351.
32. Torres GJ, Basnet RB, Sung AH, Mukkamala S, Ribeiro BM. A similarity measure for clustering and its applications. *Proceedings of World Academy of Science, Engineering and Technology* 2008; **31**:490–496.
33. Alexander-Block A, Lambiotte R, Roberts B, Giedd J, Gogtay N, Bullmore E. The discovery of population differences in network community structure: a new methods and application to brain functional networks in schizophrenia. *NeuroImage* 2012; **59**:3889–3900.
34. Bullmore ET, Suckling J, Overmeyer S, Rabe-Hesketh S, Taylor E, Brammer MJ. Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain. *IEEE Transactions on Medical Imaging* 1999; **18**:32–42.
35. Van Horn JD, Grethe JS, Kostelec P, Woodward JB, Aslam JA, Rus D, Rockmore D, Gazzaniga MS. The functional Magnetic Resonance Imaging Data Center (fMRIDC): the challenges and rewards of large-scale databasing of neuroimaging studies. *Philosophical Transactions of the Royal Society B—Biological Sciences* 2001; **356**:1323–1339.
36. Jack CR, Jr, Berstein MA, Fox NC, Thompson P, Alexander G, Harvey D, Borowski B, Britson PJ, Whitwell JL, Ward C, Dale AM, Felmlee JP, Gunter JL, Hill DLG, Killiany R, Schuff N, Fox-Bosetti S, Lin C, Studholme C, DeCarli CS,

Krueger G, Ward HA, Metzger GJ, Scott KT, Mallozzi R, Blezek D, Levy J, Debbins JP, Fleisher AS, Albert M, Green R, Bartzokis G, Glover G, Mugler J, Weiner MW. The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging* 2008; **27**:685–691.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.