NIH Public Access
Author Manuscript
Stat Med. Author manuscript; available in PMC 2014 September 10.

Published in final edited form as:
Stat Med. 2013 September 10; 32(20): 3436–3448. doi:10.1002/sim.5788.

A method for processing multivariate data in medical studies

Olivier A. Coubard[1] **for the Alzheimer's Disease Neuroimaging Initiative***
[1]The Neuropsychological Laboratory, CNS-Fed; 39 rue Meaux; 75019 Paris; France

Abstract

Traditional displays of principal component analyses lack readability to discriminate between putative clusters of variables or cases. Here the author proposes a method that clusterizes and visualizes variables or cases through principal component analyses thus facilitating their analysis. The method displays pre-determined clusters of variables or cases as urchins that each has a soma (the average point) and spines (the individual variables or cases). Through three examples in the field of neuropsychology, the author illustrates how urchins help examine the modularity of cognitive tasks on the one hand, and identify groups of healthy vs. brain-damaged participants on the other hand. Some of the data used in this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. The urchin method was implemented in MATLAB, and the source code is available in the online Supporting information. Urchins can be useful in biomedical studies to identify distinct phenomena at first glance, each having several measures (clusters of variables) or distinct groups of participants (clusters of cases).

Keywords

Statistics; Multivariate analysis; Principal component analysis; Neuropsychology

1. INTRODUCTION

Principal component analysis (PCA) is a mathematical procedure that transforms a number of possibly correlated variables into a small number of uncorrelated variables – the principal components. PCA is useful to study the variation of numerous variables simultaneously and has been applied to anatomy [1], biocatalysis [2], biology [3, 4], cancer research [5], chemistry [6], genetics [7, 8], morphometry [9, 10], proteomics [11-13], metabonomics [14], physiology [15], and psychology [16].

When performing PCA, several practical issues to which theory may lack clear responses need to be addressed: How many factors should one keep? Which ones? Are cloud configurations satisfactory? Is such a contribution strong enough? Etc. Let us illustrate how to decide on a number of factors empirically, using the values of Table 1 in the first study described below. First, one keeps the factors whose eigenvalue is above 1, the average eigenvalue: factors 1 to 3 meet this criterion and cumulate a variance of 73.1%. Second, one also considers the factors whose eigenvalue approximates 1 and whose variance is not negligible, while keeping an eye on cumulative variance: factor 4 meets this criterion (eigenvalue of .91 and variance of 10.1%), resulting in a cumulative variance of 83.2%.

Corresponding author: Olivier A. Coubard The Neuropsychological Laboratory, CNS-Fed; Paris; France Tel: 331-42-080943 olivier.coubard@cns-fed.com.

*__Note__: Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Third, searching for a break in the scree plot of variances may help to distinguish between the first useful factors and the others. Finally, one should have a look at variable or case contributions for the different factors as appreciated by their communalities ($cosine^2$) or contributions ($cosine^2$/eigenvalue) to keep the factors for which the variable or case contributions may be high. There have been several suggestions to determine the number of principal components including the recommendation that they should represent at least 97% of the variation of the data [17, 18].

The purpose of this article is to propose a method to test through the visualization of PCA the existence of clusters that have been determined in advance on the basis of a theoretical question. The researcher first selects the clusters of variables or of cases that are of interest, and then plots them using urchins in 2D displays. For all factors, plots show variables or cases regardless of their contribution on the one hand, and for their significant contribution on the other, as it will be described in Section 3. A quick overview of the entire dataset enables the researcher to select three factors relevant to his/her issue, and test the relevance of the expected pre-determined clusters through their visualization in 2D or 3D displays.

Two points need to be clarified before the author gets to the core of the method. First, the urchin method is *not* a clustering method. Clustering is defined as the fractionation of objects into clusters in which the elements share common properties. A clustering algorithm has to determine the appropriate number of clusters, as does for instance the Davies-Bouldin index [19], and to determine the measure that will permit to assign the elements into the said clusters [20]. PCA has been used as a clustering method among other techniques like functional enrichment and expectation-maximization [21], kernel PCA [22], kernel PCA compared to the Davies-Bouldin index [20, 23], Mercer kernel-based clustering [24], or spectral clustering [25]. Clustering has been applied to data mining [26], image processing [25], pattern recognition [27], self organizing systems [28, 29], etc. Importantly, the urchin method is distinct from clustering methods in the sense that it is the reality of clusters known in advance that is sought through the visualization of PCA. This approach requires that the investigator must first determine clusters on the basis of theoretical issues relevant to his/her field of research.

Second, the urchin method is *not only* a visualization method of PCA. Because of the difficulty to visualize more than three dimensions in Euclidean space, new techniques have been developed as substitutes to traditional PCA biplots or triplots: Independent Component Analysis combined with PCA (IPCA) in bioinformatics [30], PCA combined with model-based individual ancestry analysis in genetics [31], multidimensional visualizations using parallel coordinates for examining historical data [18, 32, 33], porcupine plots for studying protein motions [34], PCA combined with VARIMAX rotation in spectroscopy [35], etc. In the present method, clusters are built a priori and visualized using urchins that have a soma and spines and can be displayed in 2D or 3D space for representing the average point and individuals, respectively. Importantly, though this visualization may be reminiscent of existing plots of PCA in 2D or 3D space [36, 37], the urchin method proposes to visualize clusters that have been pre-determined in compliance with a theoretical question, which is not the case of the studies that use PCA to determine a posteriori putative clusters within their multidimensional data.

2. URCHINS FOR CLUSTERIZING VARIABLES OR CASES

The urchin is a marine invertebrate belonging to the echinodermata family. This extraordinary animal with no brain or eyes is made of a ball to which spines are attached all around. On the ocean ground, the urchin slowly moves using its spines like stilts. The spines also enable it to scrape and slash vegetables it then masticates, thanks to its Aristotle’s

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

lantern organizing buccal and digestive apparata in a pentaradiate symmetry. For the author' statistical purpose the configuration of a spherical soma with spines of variable length and orientation is retained, as illustrated in Figure 1.

Here the author proposes to use urchins for clusterizing and visualizing PCAs. Contrary to traditional plots of factor coordinates that link individual points to the center of gravity, the present method displays clusters of variables or of cases as urchins, each having a soma (the average point) and spines (the individual variables or cases). The key point is to decide a priori which clusters of variables or cases one wants to study, so that if those clusters actually correspond to uncorrelated principal components, they will result in distinct urchins popping out in the display. Such a method has several advantages which the author illustrates through three studies in the field of neuropsychology. The two first studies will present on purpose a small number of variables and cases to introduce and describe the urchin method for respectively variables and cases. The third study will use data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database to illustrate the interest of the urchin method in a large dataset, and will apply statistics on the urchins.

Data of the two first studies were extracted from a neuropsychological report in which the authors explored the effects of normal aging and of Alzheimer disease (AD) on the frontal attentional system [38]. To specify the attentional controller, they referred to the Supervisory System model, which recruits at least six cognitive processes to perform up to seven basic attentional tasks: sustaining, concentrating, sharing, suppressing, switching, preparing, and setting attention [39, 40]. The authors used neuropsychological tests to assess four of these attentional tasks: (i) the Tower of London test (TOL) and the Zoo map test for setting attention; (ii) the Stroop test and letter verbal fluency for suppressing attention; (iii) the Rule shift card sorting test (RSCT), the Trail making test (TMT) and the Plus-minus test for switching attention; and (iv) a reaction time (RT) task for preparing attention. Technical details on the tests are provided in Section A of the Supporting information.

Before performing PCAs, the author selected one measure per test – except for the RT task for which there were two measures – which resulted in nine measures: a ratio ranging 0-1 for the TOL, a score ranging 0-4 for the Zoo and the RSCT, an interference rate and a switch error rate in percentage for respectively the Stroop and the Plus-minus, the number of words for the Fluency, time difference in seconds between the two parts of the TMT, and the mean (Pme) and standard deviation (Psd) of the preparing RT task in milliseconds. For clarity of display, five of the nine measures were multiplied by −1 so that a better performance corresponds to a higher value for all tests: the interference rate of the Stroop, the switch error rate of the Plus-minus, time difference of the TMT, Pme and Psd of the RT task.

As mentioned above, deciding which clusters of variables or cases one wants to study is the preliminary key step to the present method. Here, there were four clusters of variables, namely those corresponding to attentional tasks (setting, suppressing, switching, and preparing attention), and three clusters of cases, namely the three groups of participants (healthy young adults, healthy older adults, and AD patients).

3. TESTING THE MODULARITY OF COGNITIVE TASKS USING URCHINS

3.1. Introduction

The first issue concerned the modularity of attentional tasks as postulated by the Supervisory System model [39, 40]. To achieve this goal, PCA was performed using the four tasks (setting, suppressing, switching, and preparing) as clusters of variables, with 2, 2, 3 and 2 measures, respectively. Cases were restricted to healthy young adults, considering that normal aging or AD may alter the putative modularity.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

3.2. Methods

Using STATISTICA 7.0 (StatSoft, U.S.), PCA based on correlations for the nine measures in young adults (N=18, age=20-30 years) provided the eigenvalues and related statistics shown in Table 1. Factor coordinates and contributions of variables are provided in Section B of the Supporting information. Using a traditional approach (see Section 1), the author would have empirically decided that the first four factors were sufficient for describing the PCA (Table 1, lines 1-4). But the author wished to explore the entire set of data using the urchin program.

What does the urchin program do? In the present example, it displays the four clusters of variables as four urchins that each has a soma (the average point) and spines (the individual variables). For example in the case of the 'setting' cluster, the urchin has a soma which is the average factor coordinates of the cluster's two variables, and two spines which link the soma to the factor coordinates of each of the two variables. The urchin program displays the urchins for all factors: factor 1 against factor 2 in the first plot, factor 2 against factor 3 in the second plot, and so on until factor N-1 against factor N. Finally, the urchin program produces two types of subplots: first, a subplot of all variables regardless of their contribution; second, two subplots (one for each factor) of variables that have a significant contribution, which is defined as a contribution (cosine^2/eigenvalue) above the average contribution (1 divided by the number of variables). Inputs and outputs of the urchin program are detailed in Section 6. The content of the urchin program and related files is provided in Section C of the Supporting information.

3.3. Results and Discussion

Figure 2 shows the plots for the first four factors. Figure 2A shows the projection of the variables for factor 1 on the *x* axis (variance=33.6%) and factor 2 on the *y* axis (25.8%). Figure 2B shows the projection of the variables for factor 3 on the *x* axis (13.7%) and for factor 4 on the *y* axis (10.1%). Left panels show all variables regardless of their contribution. Middle and right panels show the same information, except that only variables with a significant contribution appear according to the *x* axis in the former (i.e., factor 1 in Figure 2A and factor 3 in Figure 2B) or to the *y* axis in the latter (i.e., factor 2 in Figure 2A and factor 4 in Figure 2B), as outlined by the black double arrows. Such a visualization of PCA using urchins brings a clear answer to the first issue: do the clusters of attentional tasks hypothesized by Stuss et al. [39, 40] correspond to any reality in participants' cognitive performance? If 'attentional clusters' actually correspond to uncorrelated components, this should result in distinct urchins popping out in the display.

In Figure 2A, the fact that setting/switching, suppressing, and preparing attention correspond to distinct modules immediately stands out. Factor 1 clearly distinguishes between setting/switching on the one hand and suppressing on the other, as outlined in the left panel and confirmed by significant variables in the middle panel. Factor 2 separates setting/switching from preparing as sketched out in the left panel and corroborated in the right panel (see Figure 2A). What about the modularity of setting vs. switching attention? This issue is addressed in Figure 2B where factor 3 suggested some segregation between the two tasks (see left panel), which is confirmed by significant variables in the middle panel. Finally, factor 4 produces an effect size as can be seen in the left panel and more clearly in the right panel: all attentional tasks appeared in the positive field whereas nothing occurs in the negative field (see Figure 2B). Therefore, it can be concluded that the four attentional tasks are modular, based on the performance of young adults, although setting and switching might share some common properties.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Traditional displays of the projection of variables on planes 1-2 and 3-4 are illustrated in Figure S1 of the Supporting information. They show that linking all variables to the center of gravity without any discrimination between them would have been less readable.

It is worth noting that the urchin program also plots all other factors, here from 5 to 9. Although factors of decreasing variance are thus of decreasing interest, the researcher may wish to look at those remaining plots, as illustrated in Figure S2 and commented upon in Section D of the Supporting information. This exhaustive examination of the entire set of data also satisfies the recommendation according to which almost all variation of the data should be taken into account [17, 18].

Once factors are examined, it can be useful to project all variables according to three factors of interest. The urchin3 program fills this function. The inputs are the same as for urchin except that the urchin3 program also requires specifying the three selected factors. In output, the 2D plots are restricted to these three factors, which results in two plots each made of three subplots. The three factors can also be viewed in a 3D plot for better visualization. Inputs and outputs of the urchin3 program are detailed in the Section 6. The content of the urchin3 program and related files is provided in Section E of the Supporting information.

4. TESTING GROUP DISPARITIES USING URCHINS

4.1. Introduction

The second issue is about the disparities between groups of participants in their performance in attentional tasks depending on their age and neurological health. PCA was performed on the full set of data, using the clusters of variables of the first study (setting, suppressing, switching, and preparing) and the three groups as cases (healthy young adults, healthy older adults, and AD patients).

4.2. Methods

PCA based on correlations for the nine measures in younger (N=18, age=20-30 years), older adults (N=17, age=62-89 years), and AD patients (N=17, age=63-85 years) performed using STATISTICA 7.0 provided the eigenvalues and related statistics shown in Table 2. The first observation is that the first factor explained almost half the variance, while other factors had a variance of less than 15%. Here, minimum and maximum factor coordinates were −3.05 and +5.12, respectively, and significant contributions were defined as a contribution ($cosine^2$/eigenvalue) above the average contribution corresponding to 1 divided by the number of cases.

4.3. Results and Discussion

The results for variables are illustrated in Figure 3A for factors 1-2, and in Figure S3 for other factors (see Supporting information). As shown in Figure 3A, factor 1 produced an effect size with all attentional tasks gathered in the negative field vs. nothing in the positive field (left and middle panels), while factor 2 discriminated suppressing from setting attention (left and right panels). Factors 3, 4 and 6 failed to overcome the effect size showing no clear discrimination between clusters. Nevertheless, some contrasts were found between setting vs. preparing for factor 5 (Figure S3B-right panel), and between setting vs. switching for factor 8 (Figure S3D-middle panel). Otherwise, some dissociation was observed between tests within a task: for factor 7 between the Trail making and the Rule shift cards tests within switching attention (Figure S3C-right panel), and for factor 9 between the average vs. variability of reaction time within preparing attention (Figure S3D-right panel). In all, the inclusion of older adults and AD patients modified the organization of variables as if normal

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

aging and AD had altered the modularity nature of attentional tasks based only on the results of young adults.

The results for cases are shown in Figure 3B for factors 1-2, and in Figure S4 for other factors. Based on variable profiles, it was expected a shift from the negative field to the positive one with age and neurological disorder for factor 1. As shown in Figure 3B (left-middle panels), this is exactly what was found as evidenced by urchins of young adults, older adults, and AD patients whose somas clearly shifted from left to right. Interestingly, the increasing heterogeneity with age and disease also appeared at first glance through longer and more distributed spines for older adults and AD patients as compared to young adults. For other factors, there was no expectation based on variable profiles, and plots show how intricate the groups were according to factors of decreasing variance (see Figure S4 of the Supporting information).

Again, traditional displays would have failed to show the disparities between the groups of participants as clearly as the urchin method (see Figure S5 of the Supporting information).

Finally, the author illustrates how useful the urchin3 program may be after the entire set of data has been examined. For that purpose, the two extreme factors were plotted one against the other: factor 1 with variance of 46.0% against factor 9 (variance=0.54%). For variables, recall that factor 1 had shown an effect size with all tasks in the negative (left) field, and factor 9 some dissociation within preparing reaction time between the mean in the negative field (bottom) and the variability in the positive one (up) (see Section D and Figure S3D of the Supporting information). For cases, the projection on the plane 1-9 is shown in Figure 3C. Interestingly, beside the decline in attentional control shown horizontally by a left-to-right shift in urchins of younger, older adults and AD patients (see middle panel), there was a slight vertical bottom-up shift compatible with an increasing variability of reaction time with age and neurological disorder (see right panel).

5. APPLYING THE URCHIN METHOD TO THE ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE (ADNI) DATABASE

5.1. Introduction

The two first studies intentionally presented a small number of variables and cases for easiness of description and clarity of display in the author's pedagogic purpose of introducing the urchin method. But the urchin method may gain interest with huge sets of data. To test this possibility in the field of neuropsychology, the author has downloaded data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu) as ADNI investigator. A simple question was addressed: can the Mini-Mental State Examination (MMSE), known for differentiating between cognitive normal (NL) older adults and Alzheimer disease (AD) patients [41, 42], also discriminate NL vs. mild cognitive impairment (MCI) adults? To achieve this goal, PCA was performed using as a cluster of variables the MMSE and as clusters of cases NL, MCI and AD participants.

5.2. Methods

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers,

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California – San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 adults, ages 55 to 90, to participate in the research, approximately 200 cognitively normal older individuals to be followed for 3 years, 400 people with MCI to be followed for 3 years and 200 people with early AD to be followed for 2 years. For up-to-date information, see www.adni-info.org.

Diagnosis and neuropsychological data were downloaded on August 6th 2012. Pre-processing was done as follows using home-made scripts under MATLAB 7.0 (The MathWorks, U.S.). When participants had undergone more than one examination, only the first one was taken into account. For diagnosis, only stable NL, stable MCI, and stable AD were kept resulting in 814 cases distributed as follows: 271 NL, 419 MCI, and 124 AD. For neuropsychological data, only scores to the 30 items of the MMSE were kept, for which raw values were 1 or 2 for respectively correct and incorrect responses. For our purpose, the score 2 was replaced by 0 for all items so that a better performance always corresponds to a higher value. In all, PCA was performed using as a cluster of variables the MMSE and its 30 measures, and as clusters of cases NL, MCI and AD with respectively 271, 419, and 124 measures.

5.3. Results

Using STATISTICA 7.0 (StatSoft, U.S.), PCA based on correlations for the 30 measures in NL (N=271), MCI (N=419), and AD (N=124) provided the eigenvalues and related statistics shown in Table 3. As first observation, the first factor explained 21.3% of the variance, whereas other factors had a variance of less than 10%. To answer our question, the author used the urchin3 program, restricted the analysis to the first three factors, and plotted in 3D space all variables and cases regardless of their contribution. For cases, minimum and maximum factor coordinates were −17.7 and +27.0 but the plot was limited to −10 and +10 for clarity of display.

The results for variables are illustrated in Figure 4A for factors 1-2-3. A corresponding video is provided in Figure 4bis of the Supporting information, in which the 3D plot is rotated 360-degree in azimuth for full visualization of the urchin in 3D space. As evidenced in Figures 4A and 4bis, factor 1 produced an effect size with all measures of the MMSE gathered in the negative field vs. nothing occurred in the positive field. From this configuration and since the MMSE is a battery of general cognition [41, 42], the negative field could be interpreted as the field of cognitive functions while the positive one as the absence of cognitive functions. With respect to our issue, factors 2 and 3 were of little interest and no further subjected to comments.

The results for cases are shown in Figure 4B for factors 1-2-3, and a corresponding video is available in Figure 4ter of the Supporting information, showing as for variables a rotation of the 3D plot 360-degree in azimuth. Based on variable profile and according to the first factor (factor X in Figures 4B and 4ter), it was expected a shift from the negative field to the positive one with cognitive decline, that is to say from NL to MCI on the one hand and from MCI to AD on the other. This is exactly what was found as demonstrated by urchins of NL,

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

MCI, and AD whose somas shifted from left to right according to the first factor. However, though NL and MCI urchins were clearly dissociated from the AD urchin, it is striking how intricate the urchins of NL and MCI are, as shown in Figures 4B and 4ter. Thus there is a need for statistics to disentangle between NL and MCI so as to answer the author's question.

Using STATISTICA 7.0 (StatSoft, U.S.), univariate statistics were performed on case factor coordinates and individual case distances to the soma, i.e. spine lengths. Descriptive statistics are given in Table 4. To test the distance between urchin somas, one-way ANOVA on case factor coordinates showed a group effect ($F(2,811)=408.2$, $P<.001$). For post-hoc comparisons, the Newman-Keuls (NK) method was used to prevent type I errors, and all two-by-two comparisons were statistically significant (NK, $P<.001$). To test individual case distances to the average point, one-way ANOVA on the absolute distance of cases to the soma showed a group effect (($F(2,811)=125.4$, $P<.001$). The difference in spine lengths was statistically different between NL and AD and between MCI and AD (NK, $P<.0001$), but not between NL and MCI (NK, $P=.872$).

Using STATISTICAL PACKAGE FOR THE SOCIAL SCIENCES 12.0 (SPSS Inc., U.S.), multivariate statistics were also performed on case factor coordinates and individual case distances to the average point. When factor coordinates and spine lengths were entered as dependent variables and diagnosis as fixed factor, Hotelling's t test showed a group effect ($F(4,1618)=209.6$, $P<.001$). Tukey test was chosen for post-hoc comparisons as it also prevents type I errors. For case factor coordinates, all two-by-two comparisons were statistically significant (Tukey, $P<.001$). For spine lengths, the difference was statistically different between NL and AD and between MCI and AD (Tukey, $P<.001$), but there was no difference between NL and MCI (Tukey, $P=.978$).

To end, traditional displays would have been less efficient to show between-group disparities as clearly as the urchin method (see Figure S6 of the Supporting information).

5.4. Discussion

Using a large set of neuropsychological data from the ADNI, the urchin method enabled the author to efficiently address the issue according to which the MMSE battery may be useful to dissociate MCI from NL older adults. Indeed, pre-determining a cluster of variable for the MMSE and clusters of cases for NL vs. MCI vs. AD resulted in clearly distinct urchins in the plot, though NL and MCI were intricate, suggesting their relative independency. The somas of NL, MCI and AD urchins made visible at first glance a higher distance between NL and AD (5.27) or MCI and AD (4.69) than between NL and MCI (0.58). The distribution in space of individual case spines immediately suggested higher between-participant variability for AD than for NL and MCI. Statistics, either univariate or multivariate, confirmed the significant difference between the three average points given by urchin somas, demonstrating that the MMSE shows some difference between not only NL/MCI and AD but also NL and MCI. Only between-participant variability in MMSE performance was not different between NL and MCI groups of participants, both of which were significantly lower than that of AD patients.

6. INPUTS AND OUTPUTS OF THE URCHIN AND URCHIN3 PROGRAMS

6.1. Inputs

The urchin and urchin3 programs were implemented by the author in MATLAB 7.0 (The MathWorks, U.S.). Before running the program, the user has to build three matrices for the eigenvalues, the factor coordinates, and the contributions of variables in ASCII format under the file names eigen, coord, and contr, respectively. These matrices can easily be built by a

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

copy-paste of numbers from the statistical software used for performing the PCA (see Tables 1, S1, and S2 provided by STATISTICA 7.0).

Let us define the inputs of the urchin program and illustrate them with the first study (see Section 3). The user has to specify the following information. “Variables(v) or cases(c)” defines the type of data: in our example, “v” was entered. “First number of clusters [1...N]” is the key information where the user specifies his/her clusters: here, [1 3 5 8] were set as first lines of setting, suppressing, switching, and preparing clusters, respectively. Then, the “Lower limit?”, “Upper limit?”, and “Step?” refer to the limits of the plot and grid step: for variables as here, lower and upper limits and the step can be set to −1, +1 and 0.5 as their projection in the factor plane never exceed those values. In contrast for cases, the user should first calculate the minimum and maximum of the case factor coordinates in order to fix these limits (e.g., lower limit, upper limit and grid step were respectively set to −3.5, +5.5 and 1.5 in the second study). “Tags on soma(o) or spine(p) or both(b)?” defines the location of tags identifying the somas or the spines or both of them. The input depends on the information the user wants to examine in his/her data: for a small number of variables (as here), tags were set on both the soma and spines. The content of tags is defined apart in the tagvar and tagcas subprograms. To end, the user decides the “Black-and-white(b) or color(c)?” format.

In addition to the urchin program, the urchin3 program requires as inputs selecting three factors of interest: in response to “Factors [f1 f2 f3]?”, the user enters the factor values, for instance [1 5 9].

6.2. Outputs

The outputs of both the urchin and urchin3 programs are N-1 plots, where N is the total number of factors: factor 1 against factor 2 in the first plot, factor 2 against factor 3 in the second plot, and so on until factor N-1 against factor N. Each plot is made of three subplots: a subplot of all variables regardless of their contribution, and two subplots (one for each factor) of variables with a significant contribution. Plots are saved under *.emf format to quickly overview all plots, and *.fig format to retouch them if necessary. The output of the urchin3 program is restricted to two 2D plots of the first selected factor against the second one, and of the second one against the third one, each made of three subplots similar to those of the urchin program. The urchin3 program also produces a 3D plot of these three factors.

7. CONCLUSION

This article presents a method to test pre-determined clusters through the visualization of PCA. Contrary to traditional displays showing variables or cases individually (e.g., snowflakes), the method proposes to decide clusters of variables or of cases and to plot them using urchins having a soma and several spines for representing the average point and the individual variables or cases, respectively. Though previous reports may have occasionally used average points or urchin-like plots, the present article systematizes the visualization of PCAs using urchins for testing pre-determined clusters, with the option of seeing separately all variables or cases and/or only the significant ones. The present article also provides a code that any researcher can apply to his/her data and adapt to his/her purpose.

The advantages of the urchin method are fivefold: (i) the pre-determination of clusters of variables or cases results in popping-out segregate urchins in the plot as soon as they correspond to any reality of distinct components; (ii) the average point of clusters of variables or cases given by the soma of the urchin enables to catch at a glance the average location of every cluster and the distances (or not) between them; (iii) the distance and distribution in space of individual variables or cases with respect to the average point given

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

by the spines of the urchin allows to directly appreciate the homogeneity (or heterogeneity) of the clusters; (iv) the display in separate plots of variables or cases with above average contribution further facilitates the discrimination between clusters and the identification of those variables or cases which significantly contribute to the contrasts of every factor; (v) finally, statistics can be performed on factor coordinates and on absolute distances of individuals to the average point to test, respectively, the spatial shift in urchin somas indicating clusters' distances in average, and the varying length of spines indicating variations in clusters' homogeneity. Urchins in both their plots and the statistics they allow on their somas and spines enable researchers to immediately and efficiently answer their initial theoretical question, while a standard visualization of PCA offers exploratory directions for which researchers have to build a sense a posteriori.

In the field of neuropsychology, the author presented basic issues but the urchin method can also be useful to study subtle phenomena such as the effects of motor or cognitive training on neuropsychological health. It is likely that the urchin method will also gain interest with huge sets of data in anatomy [1], biology [3, 4], cancer research [5], genetics [7, 8], proteomics [11-13], metabonomics [14], or physiology [15]. Finally, the urchin method may also be a useful tool to pool heterogeneous data sets together. For example, let us suppose the different data sets for different groups of participants: historical, medical, biological, genetic, imaging, psychometric data, etc. The urchin method will simply set as a cluster of variables each of these data sets and as clusters of cases the different populations (healthy, unhealthy, intermediate, etc.) to provide investigators 2D and 3D plots of their corresponding urchins as presented in this article and to enable them univariate or multivariate statistics on urchin somas and/or spines.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott; Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Amorfix Life Sciences Ltd.; AstraZeneca; Bayer HealthCare; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles. This research was also supported by NIH grants P30 AG010129 and K01 AG030514.

REFERENCES

1. Cox PG. A quantitative analysis of the Eutherian orbit: correlations with masticatory apparatus. Biological Reviews of the Cambridge Philosophical Society. 2008; 83:35–69. [PubMed: 18211281]
2. Braiuca P, Ebert C, Basso A, Linda P, Gardossi L. Computational methods to rationalize experimental strategies in biocatalysis. Trends in Biotechnology. 2006; 24:419–425. [PubMed: 16870286]
3. Higgs BW, Weller J, Solka JL. Spectral embedding finds meaningful (relevant) structure in image and microarray data. BMC Bioinformatics. 2006; 7:74. [PubMed: 16483359]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

4. Bhattarai D, Karki BB. Atomistic visualization: space-time multiresolution integration of data analysis and rendering. Journal of Molecular Graphics and Modelling. 2009; 27:951–968. [PubMed: 19278879]
5. Skerman HM, Yates PM, Battistutta D. Multivariate methods to identify cancer-related symptom clusters. Research in Nursing and Health. 2009; 32:345–360. [PubMed: 19274688]
6. Kaliszan R. QSRR: quantitative structure-(chromatographic) retention relationships. Chemical Reviews. 2007; 107:3212–3246. [PubMed: 17595149]
7. Rudi K, Zimonja M, Trosvik P, Naes T. Use of multivariate statistics for 16S rRNA gene analysis of microbial communities. International Journal of Food Microbiology. 2007; 120:95–99. [PubMed: 17602772]
8. Jombart T, Pontier D, Dufour AB. Genetic markers in the playground of multivariate analysis. Heredity. 2009; 102:330–341. [PubMed: 19156164]
9. Krey KF, Dannhauer KH. Morphometric analysis of facial profile in adults. Journal of Orofacial Orthopedics. 2008; 69:424–436. [PubMed: 19169639]
10. Vezzetti E, Calignano F, Moos S. Computer-aided morphological analysis for maxillo-facial diagnostic: a preliminary study. Journal of Plastic and Reconstructive Aesthetic Surgery. 2010; 63:218–226.
11. Stenberg P, Pettersson F, Saura AO, Berglund A, Larsson J. Sequence signature analysis of chromosome identity in three Drosophila species. BMC Bioinformatics. 2005; 6:158. [PubMed: 15975141]
12. Carpentier SC, Panis B, Swennen R, Lammertyn J. Finding the significant markers: statistical analysis of proteomic data. Methods in Molecular Biology. 2008; 428:327–347. [PubMed: 18287781]
13. Marengo E, Robotti E, Bobba M. 2D-PAGE maps analysis. Methods in Molecular Biology. 2008; 428:291–325. [PubMed: 18287780]
14. Izquierdo-Garcia JL, Rodriguez I, Kyriazis A, Villa P, Barreiro P, Desco M, Ruiz-Cabello J. A novel R-package graphic user interface for the analysis of metabonomic profiles. BMC Bioinformatics. 2009; 10:363. [PubMed: 19874601]
15. Knock SA, McIntosh AR, Sporns O, Kotter R, Hagmann P, Jirsa VK. The effects of physiologically plausible connectivity structure on local and global dynamics in large scale brain models. Journal of Neuroscience Methods. 2009; 183:86–94. [PubMed: 19607860]
16. Hutchison KA. Attentional control and the relatedness proportion effect in semantic priming. Journal of Experimental Psychology. Learning, Memory, and Cognition. 2007; 33:645–662.
17. Valle S, Li W, Qin SJ. Selection of the number of principal components: the variance of the reconstruction error criterion with a comparison to other methods. Industrial & Engineering Chemistry Research. 1999; 38:4389–4401.
18. Wang XZ, Medasani S, Marhoon F, Albazzaz H. Multidimensional visualization of principal component scores for process historical data analysis. Industrial & Engineering Chemistry Research. 2004; 43:7036–7048.
19. Jain, AK.; Dubes, RC. Algorithms for clustering data. Prentice Hall; Englewood Cliffs, NJ: 1988.
20. Nasser A, Hamad D, Nasr C. Kernel PCA as a visualization tools for clusters identifications. Lecture Notes in Computer Science. 2006; 4132:321–329.
21. Boyle J. Gene-Expression Omnibus integration and clustering tools in SeqExpress. Bioinformatics. 2005; 21:2550–2551. [PubMed: 15746290]
22. Cristianini, N.; Shawe-Taylor, J.; Kandola, J. Spectral kernel methods for clustering. In: Dietterich, T.; Becker, S.; Ghahramani, Z., editors. Advances in Neural Information Processing Systems. MIT Press; Cambridge, MA: 2002. p. 649-655.
23. Nasser A, Hébert PA, Hamad D. Clustering evaluation in feature space. Lecture Notes in Computer Science. 2007; 4669:321–330.
24. Girolami M. Mercer kernel-based clustering in feature space. IEEE Transactions on Neural Networks. 2002; 13:780–784. [PubMed: 18244475]
25. Ng, AY.; Jordan, MI.; Weiss, Y. On spectral clustering: analysis and an algorithm. In: Dietterich, T.; Becker, S.; Ghahramani, Z., editors. Advances in Neural Information Processing Systems. MIT Press; Cambridge, MA: 2002. p. 849-856.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

26. Twining CJ, Taylor CJ. The use of kernel principal component analysis to model data distributions. Pattern Recognition. 2003; 36:217–227.
27. Roberts SJ, Everson R, Rezek I. Maximum certainty data partitioning. Pattern Recognition. 2000; 33:833–839.
28. Mu-Chun S, Hsiao-Te C. A new model of self-organizing neural networks and its application in data projection. IEEE Transactions on Neural Networks. 2001; 12:153–158. [PubMed: 18244371]
29. Lebart L. Assessing self organizing maps via contiguity analysis. Neural Networks. 2006; 19:847–854. [PubMed: 16777380]
30. Yao F, Coquery J, Le Cao KA. Independent Principal Component Analysis for biologically meaningful dimension reduction of large biological data sets. BMC Bioinformatics. 2012; 13:24. [PubMed: 22305354]
31. Di Gaetano C, Voglino F, Guarrera S, Fiorito G, Rosa F, Di Blasio AM, Manzini P, Dianzani I, Betti M, Cusi D, Frau F, Barlassina C, Mirabelli D, Magnani C, Glorioso N, Bonassi S, Piazza A, Matullo G. An Overview of the Genetic Structure within the Italian Population from Genome-Wide Data. PLoS One. 2012; 7:e43759. [PubMed: 22984441]
32. Albazzaz H, Wang XZ, Marhoon F. Multidimensional visualisation for process historical data analysis: a comparative study with multivariate statistical process control. Journal of Process Control. 2005; 15:285–294.
33. Albazzaz H, Wang XZ. Historical data analysis based on plots of independent and parallel coordinates and statistical control limits. Journal of Process Control. 2006; 16:103–114.
34. Barrett CP, Hall BA, Noble ME. Dynamite: a simple way to gain insight into protein motions. Acta Crystallographica. Section D: Biological Crystallography. 2004; 60:2280–2287.
35. Broersen A, van Liere R, Heeren RMA. Comparing three PCA-based methods for the 3D visualization of imaging spectroscopy data. Visualization, Imaging, and Image Processing. 2005; 480:540–545.
36. Hibbs MA, Dirksen NC, Li K, Troyanskaya OG. Visualization methods for statistical analysis of microarray clusters. BMC Bioinformatics. 2005; 6:115. [PubMed: 15890080]
37. Sharov AA, Dudekula DB, Ko MS. A web-based tool for principal component and significance analysis of microarray data. Bioinformatics. 2005; 21:2548–2549. [PubMed: 15734774]
38. Coubard OA, Ferrufino L, Boura M, Gripon A, Renaud M, Bherer L. Attentional control in normal aging and Alzheimer's disease. Neuropsychology. 2011; 25:353–367. [PubMed: 21417533]
39. Stuss DT, Shallice T, Alexander MP, Picton TW. A multidisciplinary approach to anterior attentional functions. Annals of the New York Academy of Sciences. 1995; 769:191–211. [PubMed: 8595026]
40. Stuss DT, Alexander MP. Is there a dysexecutive syndrome? Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences. 2007; 362:901–915.
41. Folstein MF, Folstein SE, McHugh PR. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. Journal of Psychiatric Research. 1975; 12:189–198. [PubMed: 1202204]
42. Kalafat M, Hugonot-Diener L, Poitrenaud J. Standardisation et étalonnage français du "Mini Mental State" (MMS) version GRECO. Revue de Neuropsychologie. 2003; 13:209–236.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

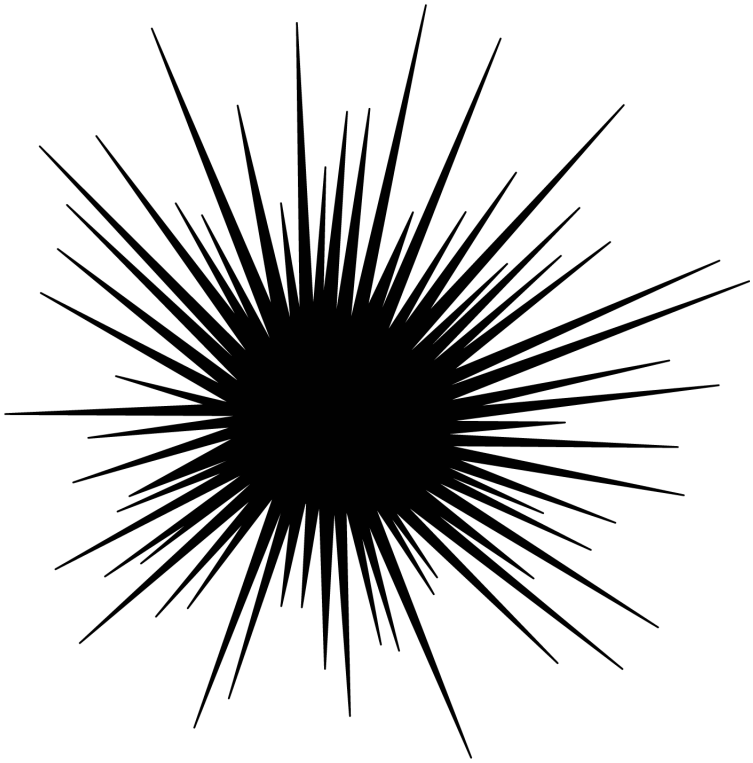

Figure 1.
Schema of an urchin with a soma in the center and spines all around.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

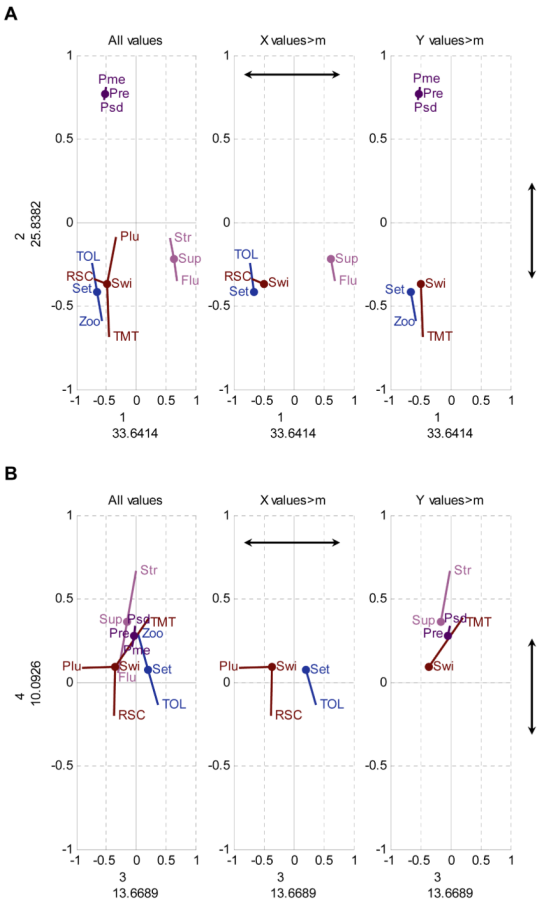

Figure 2.

2D plots of variable factor coordinates on planes 1-2 (**A**), and 3-4 (**B**) in the first study using the urchin program. Factors are identified by their value number and variance in percentage. In body graphs, urchins are shown for setting (Set in blue), suppressing (Sup in light purple), switching (Swi in burgundy), and preparing (Pre in deep purple) attention, either for all variables (All values in left panels), or for variables of significant contribution according to the *x* axis (X values>m in middle panels) or the *y* axis (Y values>m in right panels). Tags on spines: TOL, Tower of London; Str, Stroop; Flu, Fluency; RSC, Rule shift cards; TMT, Trail making test; Plus, Plus-minus; Pme, mean Preparing reaction time; Psd, standard deviation of Preparing reaction time.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

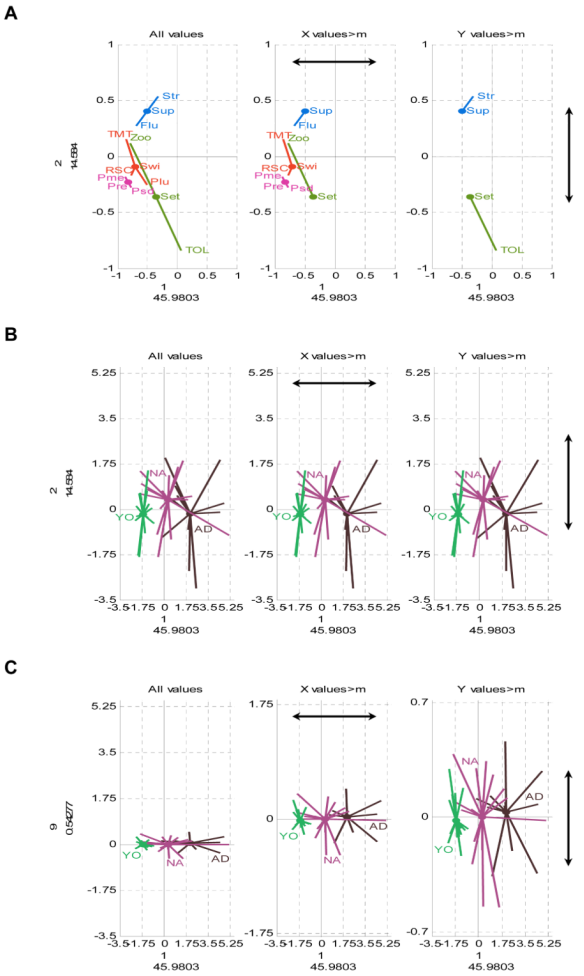

Figure 3.

2D plots of variable (**A**) and case (**B**) factor coordinates on the plane 1-2 in the second study using the urchin program. (**C**) 2D plots of case factor coordinates on the plane 1-9 using the urchin3 program. Factors are identified by their value number and variance in percentage. In body graphs, urchins are shown (**A**) for setting (Set in green), suppressing (Sup in blue), switching (Swi in red), and preparing (Pre in purple) attention; (**B-C**) for healthy young adults (YO in green), normal aged adults (NA in purple), and Alzheimer disease patients (AD in dark brown), (**A-B-C**) either for all variables (All values in left panels), or for variables of significant contribution according to the *x* axis (X values>m in middle panels) or the *y* axis (Y values>m in right panels). (**A**) Tags on spines as in Figure 2; (**B-C**) tags were set only on somas.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

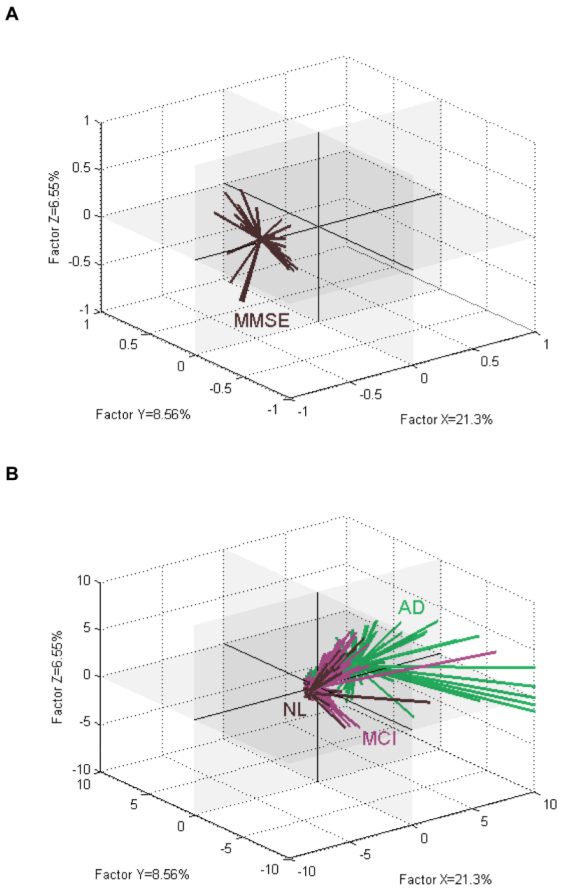

Figure 4.

3D plots of variable (**A**) and case (**B**) factor coordinates on the planes 1-2-3 in the third study using the urchin3 program. Factors 1-3 are respectively identified by the letters X, Y, and Z, and their variance in percentage. In body graphs, urchins are shown (**A**) for the Mini-Mental State Examination (MMSE); (**B**) for normal older adults (NL in dark brown), mild cognitive impairment adults (MCI in purple), and Alzheimer disease patients (AD in green), for all variables (**A**) or cases (**B**). (**A-B**) Tags were set only on somas. Notice that corresponding videos of Figures **4A** and **4B** are provided in respectively Figures **4bis** and **4ter** of the Supporting information, in which 3D plots are rotated 360-degree in azimuth to appreciate the visualization of the urchins in 3D space.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 1

Eigenvalues of correlation matrix and related statistics in the first study.

Value number	Eigenvalue	% Total variance	Cumulative Eigenvalue	Cumulative %
1	3.03	33.64	3.03	33.64
2	2.33	25.84	5.35	59.48
3	1.23	13.67	6.58	73.15
4	0.91	10.09	7.49	83.24
5	0.57	6.37	8.06	89.61
6	0.46	5.10	8.52	94.71
7	0.27	2.97	8.79	97.67
8	0.18	2.03	8.97	99.70
9	0.03	0.30	9.00	100.00

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 2

Eigenvalues of correlation matrix and related statistics in the second study.

Value number	Eigenvalue	% Total variance	Cumulative Eigenvalue	Cumulative %
1	4.14	45.98	4.14	45.98
2	1.31	14.58	5.45	60.56
3	0.93	10.36	6.38	70.93
4	0.80	8.94	7.19	79.87
5	0.77	8.57	7.96	88.44
6	0.45	5.00	8.41	93.44
7	0.30	3.35	8.71	96.79
8	0.24	2.67	8.95	99.46
9	0.05	0.54	9.00	100.00

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 3

Eigenvalues of correlation matrix and related statistics (restricted to the ten first factors) in the third study.

Value number	Eigenvalue	% Total variance	Cumulative Eigenvalue	Cumulative %
1	6.38	21.26	6.38	21.26
2	2.57	8.56	8.95	29.82
3	1.96	6.55	10.91	36.37
4	1.65	5.51	12.56	41.88
5	1.22	4.07	13.79	45.96
6	1.13	3.77	14.92	49.73
7	1.06	3.54	15.98	53.27
8	1.03	3.43	17.01	56.70
9	0.98	3.26	17.99	59.96
10	0.92	3.05	18.90	63.02

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 4

Mean±Standard deviation of urchins according to factor 1 in the third study.

Factor 1	NL	MCI	AD
Factor coordinates	−1.102±0.733	−0.521±1.099	4.171±3.965
Spine lengths	0.728±0.533	0.749±0.842	2.836±3.016