Contents lists available at ScienceDirect

NeuroImage

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

# How to compute reliability estimates and display confidence and tolerance intervals for pattern classifiers using the Bootstrap and 3-way multidimensional scaling (DISTATIS)

Hervé Abdi [a,*], Joseph P. Dunlop [a], Lynne J. Williams [b]

[a] *School of Behavioral and Brain Sciences, The University of Texas at Dallas, MS: Gr.4.1., 800 West Campbell Road, Richardson, TX 75080-3021, USA*
[b] *School of Communication Sciences and Disorders, University of Western Ontario, Elborn College, London, ON, Canada N6G 1H1*

ARTICLE INFO

*Article history:*
Received 24 January 2008
Revised 27 October 2008
Accepted 6 November 2008
Available online xxxx

*Keywords:*
Multidimensional scaling (MDS)
DISTATIS
Reliability estimation
Bootstrap resampling
Pattern classifiers
Confidence interval
Tolerance interval
Group analysis
$R_V$ coefficient
Bonferonni correction
Šidàk correction
STATIS

ABSTRACT

When used to analyze brain imaging data, pattern classifiers typically produce results that can be interpreted as a measure of discriminability or as a distance between some experimental categories. These results can be analyzed with techniques such as multidimensional scaling (MDS), which represent the experimental categories as points on a map. While such a map reveals the configuration of the categories, it does not provide a reliability estimate of the position of the experimental categories, and therefore cannot be used for inferential purposes. In this paper, we present a procedure that provides reliability estimates for pattern classifiers. This procedure combines bootstrap estimation (to estimate the variability of the experimental conditions) and a new 3-way extension of MDS, called DISTATIS, that can be used to integrate the distance matrices generated by the bootstrap procedure and to represent the results as MDS-like maps. Reliability estimates are expressed as (1) *tolerance* intervals which reflect the accuracy of the assignment of scans to experimental categories and as (2) *confidence* intervals which generalize standard hypothesis testing. When more than two categories are involved in the application of a pattern classifier, the use of confidence intervals for null hypothesis testing inflates Type I error. We address this problem with a Bonferonni-like correction. Our methodology is illustrated with the results of a pattern classifier described by O'Toole et al. (O'Toole, A., Jiang, F., Abdi, H., Haxby, J., 2005. Partially distributed representations of objects and faces in ventral temporal cortex. J. Cogn. Neurosci. 17, 580–590) who re-analyzed data originally collected by Haxby et al. (Haxby, J., Gobbini, M., Furey, M., Ishai, A., Schouten, J., Pietrini, P., 2001. Distributed and overlapping representation of faces and objects in ventral temporal cortex. Science 293, 2425–2430).

## Introduction

Neuroimaging experiments typically produce large multivariate datasets with many more variables (i.e., voxels) than observations (i.e., scans). In general, multivariate data sets are analyzed with standard multivariate methods (e.g., multiple regression, discriminant analysis, MANOVA), but these methods cannot be used when the variables outnumber the observations because this creates a problem known as multicollinearity. Therefore, standard methods cannot be used directly for the analysis of most brain imaging studies. Multicollinearity can be handled by standard techniques with preprocessing such as principal component analysis, however, some recent techniques, such as machine learning (Duda et al., 2001; Hastie et al., 2001; Bishop, 2006), can cope *directly* with multicollinearity. These techniques, called pattern-based classifiers or sometimes "brain reading" classifiers (Cox and Savoy, 2003), have recently attracted the attention of the brain imaging community (see, e.g., recent reviews by Haynes and Rees, 2006; Norman et al., 2006; O'Toole et al., 2007). Classifiers can be used to assign scans to different a priori categories of interest (e.g., experimental conditions, clinical classifications). In this context, the performance of the classifier is expressed by computing between category distances which reflect the accuracy of the assignments of scans to their a priori categories. With techniques such as multidimensional scaling (MDS), the whole set of the between category distances can, in turn, be represented as a map (see, e.g., Welchew et al., 2002). In such a map, the categories described by the distances are plotted as points and the distances between these points approximate the distances between categories (e.g., Fig. 2). Typically, classifiers are applied separately for each subject, and so, the analysis of an experiment produces, for each subject, one set of between category distances which can displayed as one category map. While this map reveals the configuration of the categories, it does not provide a reliability estimate of the position of the categories, and therefore it cannot be used in lieu of a statistical test.

In this paper, we present a new 3-way generalization of MDS called DISTATIS, which can be applied to sets of distances and which can be used to integrate multiple distance matrices into a single graphical

[*] Corresponding author.
*E-mail address:* herve@utdallas.edu (H. Abdi).

1053-8119/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2008.11.008

representation. Then, we show how to use cross-validation techniques in order to compute reliability estimates for the position of the categories. We also show how to represent these estimates with confidence and tolerance ellipsoids on the representation provided by DISTATIS.

We illustrate our procedure with the results of a pattern classifier described by O'Toole et al. (2005) who, in turn, re-analyzed *f*MRI data originally collected by Haxby et al. (2001). O'Toole et al. (2005) analyzed brain scans collected when subjects were watching the picture of one exemplar from eight categories of objects. In our analysis, we integrate the category data from all six subjects and we estimate the reliability of the category positions. To supplement the analysis of the categories, we also analyze the similarities and differences between the subjects' configurations.

## Methods

### *Analyzing distance matrices with DISTATIS*

The input for MDS is a $K$ by $K$ symmetric matrix whose entries are the distances between $K$ elements of a set of interest. In a brain imaging context, these elements typically represent experimental categories (e.g., experimental conditions, type of stimuli, clinical conditions) and we will refer to them as "categories" in the following presentation. MDS (Togerson, 1958; Abdi, 2007a) transforms the distance matrix into a map where the $K$ categories are points positioned such that the (Euclidean) distances between them give the best approximation of the original distances between categories. For DISTATIS, the input consists of a set of $N$ distance matrices, each of order $K$ by $K$. Each matrix, denoted by $\mathbf{D}_n$, contains the distances provided by the $n$th subject. DISTATIS produces two maps: one for the $K$ categories, and one for the $N$ subjects. The first map is called the *compromise* (or *consensus*) map and it displays the best common position of the $K$ categories for all $N$ subjects. In this map, the categories are represented by $K$ points. Each of the $N$ original distance matrices can also be projected onto the compromise map. Therefore, for each category point, we have $N$ variant points representing the subjects' positions for this category. The second map is called the *subjects'* map or the $R_V$ map (the rationale for this name will be clear later). This map displays the similarity structure of the $N$ subjects represented by $N$ points. This map is also an intermediate step for the computation of the compromise. Below we provide a four step formal sketch of DISTATIS. A summary of the procedure is given in Fig. 1 and a detailed presentation can be found in Abdi et al. (2005) and Abdi et al. (2007).

As in MDS, the *first* step of DISTATIS is to transform each distance matrix, $\mathbf{D}_n$, into a sum of cross-product (SCP) matrix (this matrix is analogous to a variance–covariance matrix). The $K$ by $K$ SCP matrix denoted $\mathbf{S}_n$, is computed by pre- and post-multiplying the distance matrix by a centering matrix $\Xi$ defined as:

$$\Xi = \mathbf{I} - \mathbf{1}\boldsymbol{m}^T \tag{1}$$

where $\mathbf{I}$ is the (conformable, i.e., $K$ by $K$) identity matrix, $\boldsymbol{m}$ a $K$ by 1 vector of masses (i.e., $m_k \geq 0$ and $\Sigma_k = 1$, often we use $m_k = \frac{1}{K}$), and $\mathbf{1}$ a conformable (i.e., $K$ by 1) vector of 1's. Formally, the SCP matrix is obtained as:

$$\mathbf{S}_n = -\frac{1}{2}\Xi\mathbf{D}_n\Xi^T. \tag{2}$$

If the distance matrices are expressed with different scales, the SCP matrices need to be normalized. This is implemented by dividing the entries of each SCP matrix $\mathbf{S}_n$ by its first eigenvalue. The normalized matrices have their first eigenvalue equal to unity (such a normalization is also performed in multiple factor analysis; see, e.g., Escofier and Pagès, 1990; Abdi and Valentin, 2007). Normalization by the first eigenvalue ensures that no matrices dominate the first dimensions of the analysis (when several matrices are concatenated, the matrices

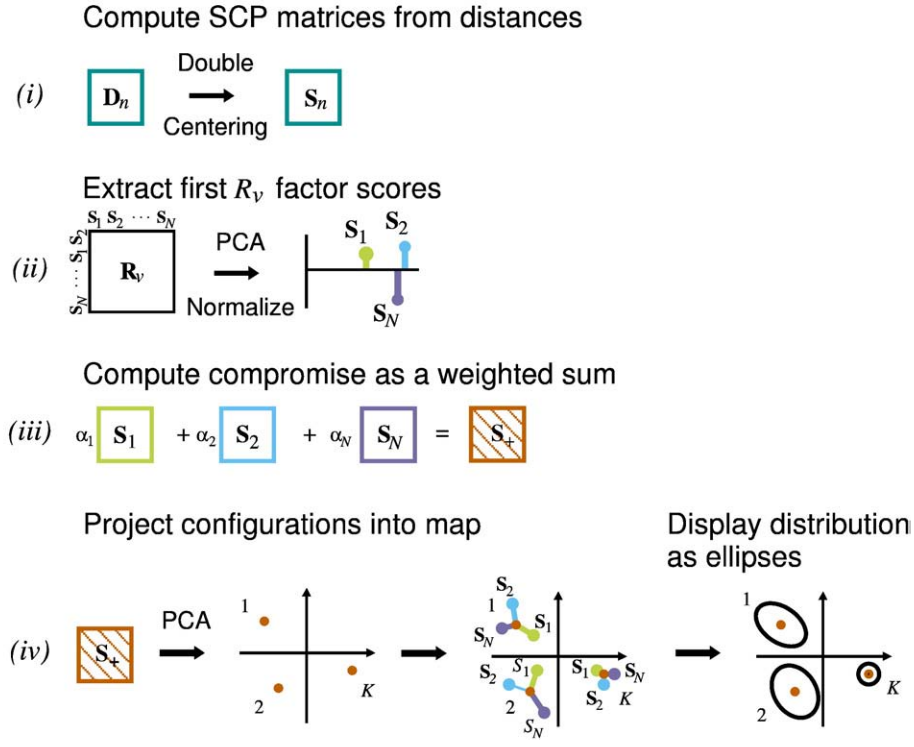
Compute SCP matrices from distances
(i)
D_n
Double
Centering
S_n
Extract first R_V factor scores
(ii)
S_1 S_2 ⋯ S_N
S_1 S_2 ⋯ S_N
R_V
PCA
Normalize
S_1
S_2
S_N
Compute compromise as a weighted sum
(iii)
α_1 S_1 + α_2 S_2 + α_N S_N = S_+
Project configurations into map
Display distribution as ellipses
(iv)
S_+
PCA
1
2
K
S_2
S_1
S_N
S_1
S_2
S_N
S_1
S_2
S_N
1
2
K
1
2
K

**Fig. 1.** The four steps of DISTATIS: (*i*) A set of distance matrices is transformed into a set of SCP matrices; (*ii*) the $R_V$ coefficients between the SCP matrices are stored in the $R_V$ matrix, which is analyzed with a non-centered PCA that provides a set of optimal weights for combining the SCP matrices; (*iii*) a compromise SCP matrix is computed as a weighted sum of the SCP matrices; (*iv*) the compromise is analyzed using standard MDS and the individual SCP matrices are projected onto the compromise map where the common configuration is summarized as a set of ellipses.

Please cite this article as: Abdi, H., et al., How to compute reliability estimates and display confidence and tolerance intervals for pattern classifiers using the Bootstrap and 3-way multidimensional scaling (DISTATIS), NeuroImage (2008), doi:10.1016/j.neuroimage.2008.11.008

with the largest first eigenvalues will dominate the first dimensions of the common matrix; see, e.g., Escofier and Pagès, 1990).

The *second* step of DISTATIS analyzes the similarity between the $N$ SCP matrices. To do so, we create an $N\times N$ matrix, called the $R_V$ (or subjects') matrix, denoted by $\mathbf{C}$ whose $n$, $n'$th element is the $R_V$ coefficient between $\mathbf{S}_n$ and $\mathbf{S}_{n'}$. This $R_V$ coefficient is computed as:

$$R_V(\mathbf{S}_n,\mathbf{S}_{n'}) = \frac{\operatorname{trace}\left(\mathbf{S}_n^T\mathbf{S}_{n'}\right)}{\sqrt{\operatorname{trace}\left(\mathbf{S}_n^T\mathbf{S}_n\right)\times\operatorname{trace}\left(\mathbf{S}_{n'}^T\mathbf{S}_{n'}\right)}}. \tag{3}$$

The $R_V$ coefficient (Escoufier, 1973; Robert and Escoufier, 1976; Abdi, 2007b;Josse et al., 2008) is a squared cosine between (positive semi-definite) matrices and its interpretation is similar to a squared coefficient of correlation (the $R_V$ coefficient can also be used to quantify the similarity between scans; see, e.g., Kherif et al., 2003; Shinkareva et al., 2006, 2008). The $R_V$ coefficient varies between 0 and 1 and indicates how much information is shared between two matrices. The analysis of the subjects' similarity structure is obtained from the eigen-decomposition (Abdi, 2007d) of $\mathbf{C}$:

$$\mathbf{C} = \mathbf{P}\boldsymbol{\Theta}\mathbf{P}^T \quad \text{where} \quad \mathbf{P}^T\mathbf{P} = \mathbf{I} \quad \text{and} \quad \boldsymbol{\Theta}\,\text{is diagonal}. \tag{4}$$

This corresponds to a non-centered principal component analysis (PCA) of $\boldsymbol{C}$. The subjects' map, also called the $R_V$ map, is obtained by plotting the subjects' factor scores that are stored in the matrix $\mathbf{G}$ which is computed as:

$$\boldsymbol{G} = \boldsymbol{P}\boldsymbol{\Theta}^{\frac{1}{2}}. \tag{5}$$

The first column of $\mathbf{P}$, which is the first eigenvector of $\mathbf{C}$, is denoted $\boldsymbol{p}_1$. Because the $R_V$ coefficient is a squared cosine, all the elements of $\boldsymbol{p}_1$ have the same sign, (this is a consequence of the Perron–Frobenius theorem; see, e.g., Lancaster and Tismenestsky, 1985, p. 532*ff.*) and these elements are, by convention, chosen to be positive. The $n$th element of the first eigenvector reflects how much the $n$th matrix has in common with the other matrices: the larger the value of the $n$th element, the more the $n$th matrix shares information with the other matrices. Therefore, the elements of $\boldsymbol{p}_1$ can be used to compute an optimum set of weights for combining the $N$ SCP matrices into a compromise: The weight of a matrix will be proportional to the amount of common information conveyed by this matrix. Specifically, these weights are obtained by rescaling the first eigenvector such that the sum of the weights is equal to one. Formally, if we denote by $\boldsymbol{\alpha}$ the $N\times 1$ weight vector whose elements are denoted $\alpha_n$, then:

$$\boldsymbol{\alpha} = \frac{\boldsymbol{p}_1}{\boldsymbol{p}_1^T\mathbf{1}} \quad \text{where} \quad \mathbf{1}\ \text{is an}\ N\times 1\ \text{vector of}\ 1'\text{s}. \tag{6}$$

In the *third* step of DISTATIS, we combine the $N$ SCP matrices (i.e., the $\mathbf{S}_n$ matrices) into a compromise SCP matrix, denoted $\mathbf{S}_+$, which is computed as the weighted average of the SCP matrices using the elements of $\boldsymbol{\alpha}$ as weights:

$$\boldsymbol{S}_+ = \sum_{n=1}^{N} \alpha_n \boldsymbol{S}_n. \tag{7}$$

In the *fourth* step, the $\mathbf{S}_+$ matrix is decomposed into eigenvalues and eigenvectors:

$$\boldsymbol{S}_+ = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T \quad \text{where} \quad \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}. \tag{8}$$

This step is equivalent to an MDS analysis of the compromise SCP matrix. Like in MDS, the $K$ categories can be represented as $K$ points on a map. The coordinates (i.e., factor scores) of the $K$ categories are obtained as:

$$\boldsymbol{F}_+ = \boldsymbol{V}\boldsymbol{\Lambda}^{\frac{1}{2}} = \boldsymbol{S}_+ \times \boldsymbol{V}\boldsymbol{\Lambda}^{-\frac{1}{2}}. \tag{9}$$

Equivalently, the matrix $\mathbf{V}\boldsymbol{\Lambda}^{-\frac{1}{2}}$ is a projection matrix, which can also be used to project each of the SCP matrices onto the factors of the compromise:

$$\mathbf{F}_n = \mathbf{S}_n \times \mathbf{V}\boldsymbol{\Lambda}^{-\frac{1}{2}}. \tag{10}$$

The compromise configuration is a linear combination of the $N$ subjects' configurations. The observed configurations can simply be projected (by a linear transformation) onto the compromise, and the distribution of observations can be compared to the compromise. We use this general property of DISTATIS both to compare computational estimates of the category configuration and to compare subjects.

### *Bootstrap for reliability estimation*

DISTATIS provides a multivariate *descriptive* analysis of the category structure. In order to complete this description with an *inferential* component, we need to have an estimate of the reliability of the category positions. Such an estimate can be obtained with cross-validation techniques such as bootstrap resampling (Efron and Tibshirani, 1993; Chernick, 2008). From the original data set, the bootstrap generates a large number of samples and it estimates the sampling distribution of a statistic (e.g., mean, variance) from the actual distribution of this statistic for the set of generated samples. The bootstrap procedure provides a nonparametric maximum likelihood estimator of the population distribution as long as the sampling scheme selects *independent* observations from the original data set (see, e.g., Chernick, 2008, p.9).

### *Confidence and tolerance intervals*

Standard statistical approaches often evaluate the reliability of an estimate of a parameter of a population with a *confidence* interval which expresses the variability of this parameter (e.g., a 95% confidence interval of the mean indicates that 95% of the sample means fall into this interval). Confidence intervals generalize standard null hypothesis testing: for example, if the confidence intervals of two means do not overlap, we can conclude that these two means differ for the significance level chosen to compute the interval.

In the context of “brain reading” the problem is to assign elements (e.g., scans) to categories. In this case, a *tolerance* interval (Hahn and Meeker, 1991) is appropriate in order to evaluate the performance of a classifier because a tolerance interval expresses the variability of the elements of the population (e.g., a 95% tolerance interval for a population indicates that 95% of the observations fall into this interval). Tolerance intervals indicate if two categories are separable or “distinguishable.” For example, if the tolerance intervals of two categories do not overlap, we can conclude that the accuracy of the assignment of observations to their respective categories is higher than the level chosen to compute the interval. In addition, the overlap of two tolerance intervals directly reflects the proportion of misclassified observations.

### *General bootstrap procedure*

In order to compute and display confidence or tolerance intervals for a classifier that generates a distance matrix between categories, we follow the steps shown below. First, we generate multiple bootstrap samples from the original data set. Second, we compute (with the classifier) a distance matrix for each bootstrap sample. Third, we compute (with DISTATIS) the compromise of all the distance matrices. This gives the best estimate of the population distance between categories. Fourth, we project (*cf.* Equation 10) the bootstrap samples onto the compromise. This creates, for each category, a set of points which represents the variability of the category position. Fifth, we select the proportion of points corresponding to the chosen level of significance. Finally, for a 2-dimensional display of the interval, we draw an ellipse that includes the proportion of points corresponding to the chosen level of significance (see Appendix for details on the algorithm used to fit the ellipse).

### *Confidence interval: The multiple comparison problem*

As mentioned earlier, a confidence interval generalizes a null hypothesis test. And, just like a standard test, the $\alpha$-level chosen is correct only when there is one single test (i.e., when there are only two categories to be compared). Typically, classifiers are used in experiments with more than two categories. Therefore, the problem of the inflation of Type I error occurs as soon as $K$ (the number of categories) is larger than two. The problem of finding an exact correction for multiple confidence intervals is still open, but a conservative approach can be implemented via a Bonferonni or Šidàk correction (see, e.g., Abdi, 2007c). For $K$ categories, comparing all pairs of categories using confidence intervals creates $\frac{1}{2}K(K-1)$ statistical comparisons. Therefore, if the overall confidence level has been set to the $(1-\alpha)$ level, a Bonferonni corrected confidence level for each category is expressed as:

$$1-\frac{2\alpha}{K(K-1)}. \tag{11}$$

Along the same lines, a Šidàk corrected confidence level for each category is expressed as:

$$(1-\alpha)^{\frac{1}{2}K(K-1)}. \tag{12}$$

For example, to maintain the overall Type I error of $\alpha=.05$ with $K=8$ categories, the Bonferonni and Šidàk corrections will use a value of 99.82% for each confidence interval. This corresponds to a confidence interval of 95% when only two categories are involved.

### *Interpreting overlapping multidimensional confidence intervals*

When a confidence interval involves only one dimension (e.g., when using a confidence interval to compare the means of two groups) the relationship between hypothesis testing and confidence intervals is straightforward. If two confidence intervals do not overlap, then the null hypothesis is rejected. Conversely, if two confidence intervals overlap then the null hypothesis cannot be rejected. The same simple relationship holds with a 2-dimensional display as long as all the variance of the data can be described with only two dimensions. If two confidence ellipses (i.e., the 2-dimensional expression of an interval) do not overlap then the null hypothesis is rejected, whereas if the ellipses overlap then the null hypothesis cannot be rejected.

In most MDS or PCA analyses, however, the 2-dimensional maps used to display the data represent only *part* of the variance of the data (in our application, for example, the first two dimensions explain roughly 50% of the variance). Therefore, the position of a confidence ellipse in a 2-dimensional map gives only an approximation of the real position of the intervals in the complete space. Now, if two confidence ellipses do not overlap in at least one display, then the two corresponding categories do not overlap in the whole space (and the null hypothesis can be rejected). However, when two confidence ellipses overlap in a given display, then the two categories may or may not overlap in the whole space (because the overlap may be due to a projection artifact). In this case, the null hypothesis may or may not be rejected depending upon the relative position of the ellipses in the other dimensions of the space. Therefore, when analyzing data laying in a multidimensional space, the interpretation of confidence intervals are correct only when performed in the *whole* space and the 2-dimensional representations give only a (possibly misleading) approximation. In fact, just like in standard MDS (i.e., without confidence ellipses) using 2-dimensional representations in order to approximate higher dimensional spaces always entails a complex subjective trade-off between the comfort of a visual representation and inevitable data distortion.

### *Confidence and tolerance ellipses for subjects in the $R_V$ map*

The bootstrap approach can also generate confidence and tolerance ellipses for the subjects' map. Looking at intervals on the first dimension will show if subjects differ in their contribution to the compromise. Looking at intervals on dimensions other than the first dimension can reveal the existence of clusters of subjects. If such clusters exist, then one could perform separate analyses for each cluster of subjects (see, e.g., Oliveira and Mexia, 2007a,b; Lazraq et al., 2008 for recent analytical developments on this question).

## An example

As an illustration, we are using a pattern classifier developed by O'Toole et al. (2005) who re-analyzed data originally collected by Haxby et al. (2001). In this experiment, six subjects performed a one-back recognition memory task on visual stimuli from eight different categories: faces, houses, cats, chairs, shoes, scissors, bottles, and scrambled images. The stimuli were presented in 12 runs of 8 blocks comprising 7 stimuli from a given category. This gave a total of 84 scans per category for each subject, 672 scans per subject, and 4,032 scans for all 6 subjects. Information regarding the *f*MRI scanning procedures can be found in Haxby et al. (2001). To increase statistical power, Haxby et al. kept only the few hundred voxels that best discriminated the categories for each subject. These voxels were used in all subsequent analyses.

### *Pattern classifier*

The goal of the analysis performed by O'Toole et al. (2005) was to find out if it was possible to discriminate between scans coming from two different categories. The analysis was performed separately for each subject and for each pair of stimuli. For each subject, the scans were divided into even and odd trial runs for use as the training and testing sets. The activation values of the voxels for a given pair of categories of the training and testing sets were sorted into two “scans by pixel” matrices (one for each set). The first step was to perform a PCA on the “scans by pixels” matrix of the training set. As a preliminary compressing step, the discriminating power of each component of the PCA was evaluated (see O'Toole et al., 2005, for more details) and only the components with the largest discriminating power were kept. A perceptron (see Abdi et al., 1999, for more details on the perceptron) was then trained to combine the factor scores in order to discriminate between the two categories (performance was perfect on the training set). The scans of the testing set were then projected onto the components of the training set. This provided a set of factor scores for the testing set. Finally, these factor scores were used by the perceptron to assign each scan of the testing set to one of the two categories.

The Hit rate was computed as the proportion of scans correctly assigned to their categories. The False Alarm rate (FA) was computed as the proportion of scans incorrectly assigned to the other categories (the proportions of Hit and FA sum to one). Finally these two proportions were integrated into one $d'$ value which quantified the quality of the discrimination between the categories (see, e.g., Abdi, 2007e, for details). Specifically, the value of $d'$ is the difference between the inverse $Z$-transformed proportion of Hits and FA. It is computed as:

$$d' = Z_{\text{Hit}} - Z_{\text{FA}}. \tag{13}$$

Because $d'$ is expressed in $Z$-score units, values obtained from different experiments are directly comparable. Being a difference between two conditions, $d'$ is a distance (it is, in fact, a Mahalanobis distance; see, e.g., Abdi, 2007f). Because all $d'$ distances are expressed in the same unit, there is no need to normalize the SCP matrices in DISTATIS.

Please cite this article as: Abdi, H., et al., How to compute reliability estimates and display confidence and tolerance intervals for pattern classifiers using the Bootstrap and 3-way multidimensional scaling (DISTATIS), NeuroImage (2008), doi:10.1016/j.neuroimage.2008.11.008

This procedure was followed for each of the 28 category pairs and this gave one $d'$ distance matrix per subject. The roles of the training and testing sets were then reversed and the procedure was repeated. This gave two matrices of $d'$ values per subject, and therefore a total of 2 matrices × 6 subjects = 12 distance matrices.

Fig. 2 shows the compromise obtained with DISTATIS of these 12 distances matrices. The first dimension of this map shows a clear opposition between faces, cats and houses, whereas the second dimension separates the small objects (bottles, scissors, and shoes) from houses and faces. The map suggests a separation between some categories. The question is: Are these differences reliable? To answer this question we will derive confidence and tolerance intervals for the position of the categories. In order to keep the example short and focused, we will look only at the category maps and not at the subject $R_V$ map.

### *Bootstrap for fMRI block designs: The temporal correlation problem*

In order to provide correct reliability estimates, the standard bootstrap requires resampling from independent observations. FMRI experiments — especially block designs experiments — violate this requirement because of the low temporal resolution of the hemodynamic response which creates temporal correlation. Therefore, scans close in time are not independent. In particular, scans from the same block are likely to be correlated. This creates a temporal correlation configuration which prevents the use of the standard bootstrap and necessitates a more complex, but adequate, sampling scheme described below. Note that for the current example, preliminary analyses indicated that scans from different blocks were independent.

### *General bootstrap procedure*

In order to create the independent samples required for an accurate bootstrap procedure, we kept the distinction between training and testing sets and drew randomly selected half samples from each block. For each subject, we drew 100 samples of training and testing sets. This entailed computing 200 $d'$ distance matrices (100 × 2 because of the reversal of the role of the training and testing sets). This gave a total of 200 × 6 = 1, 200 $d'$ matrices for our set of 6 subjects. (see Strother et al., 2002, for a similar approach). This sampling plan meets the independence assumption needed by the bootstrap and will thus provide unbiased, though conservative, estimates (because the sample size is half as big as the original sample).

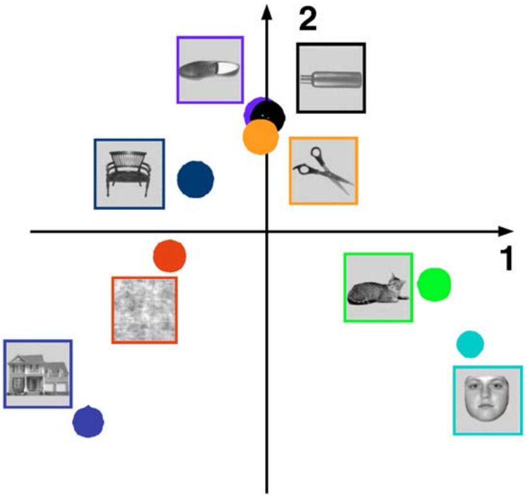

**Fig. 2.** Compromise configuration for 8 object categories computed from 12 $d'$ distances matrices obtained from the procedure described in O'Toole et al. (2005).

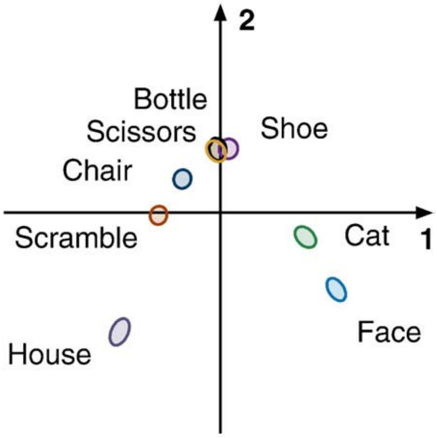

**Fig. 3.** Compromise configuration of categories between subjects, with 99.82% confidence ellipses, corresponding to 95% confidence ellipses after correction for multiple comparisons with a Bonferonni–Šidàk approach. [$\lambda_\ell$ and $\tau_\ell$ denote the eigenvalue and percentage of explained variance of the $\ell$th component ($\lambda_1$=.2517, $\tau_1$=23.84%; $\lambda_2$=.2350, $\tau_2$=22.26%)].

From this original set of 1,200 $d'$ matrices, we created 10,000 standard bootstrap samples (i.e., each sample is obtained by drawing with replacement 1, 200 $d'$ matrices from the original set). For each of these 10,000 bootstrap samples, we computed a compromise map. Finally, we used DISTATIS to integrate all 10,000 compromise maps into a single “grand compromise” map.

### *(Multiple comparison corrected) confidence intervals*

We chose a level of 95% for our confidence intervals. With $K$=8 categories of objects, this translated into a Bonferonni/Šidàk corrected level of 99.82%. Fig. 3 shows the grand compromise map with the 99.82% confidence ellipse for each category. The configuration of the ellipses indicates that the small object categories (i.e., bottle, scissors, and shoe) are not reliably separated, but that all other categories are reliably separated. The tight confidence ellipses signal that the general configuration is very stable because the overall configuration will be preserved for any position of the categories within the confidence ellipses.

### *Tolerance intervals*

The tolerance intervals shown in Fig. 4 include 95% of the 10,000 bootstrap sample projections centered around their respective categories. Recall that tolerance intervals reflect the accuracy of the assignment of scans to categories. Therefore when two categories do not overlap, they can be considered as separable. Reliable differences were found between houses and cats, between faces, shoes, bottles, and scissors; between faces and chairs; and between faces and scrambled images.

The remaining pairwise differences show some overlap, notably for cats and faces. In fact, the cat category has the largest tolerance ellipse of all categories. The size of the cat tolerance ellipse and its position suggest that the scans obtained when subjects were watching pictures of cats were similar either to the scans obtained when the subjects were watching pictures of small objects or of faces. Interestingly, O'Toole et al. (2005) mentioned that, in the analysis of the scans, the cat category was close to the face category but that, in the analysis of the pictures used as stimuli,[1] the cat category was close to the small object category. This suggests that the majority of cat pictures were viewed as faces, but that an important proportion of these pictures

[1] In order to evaluate the influence of the stimuli on the subjects' behavior, O'Toole et al. performed the same analysis on the pictures and on the scans.

Please cite this article as: Abdi, H., et al., How to compute reliability estimates and display confidence and tolerance intervals for pattern classifiers using the Bootstrap and 3-way multidimensional scaling (DISTATIS), NeuroImage (2008), doi:10.1016/j.neuroimage.2008.11.008

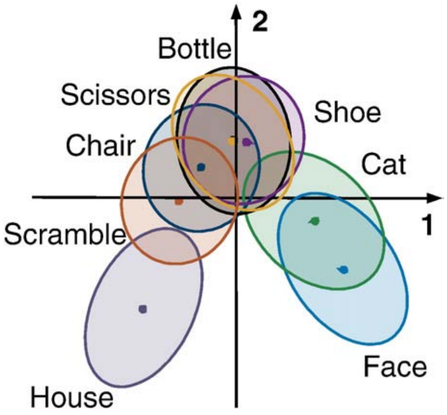

**Fig. 4.** Compromise configuration of categories between subjects, with 95% tolerance ellipses. [$\lambda_\ell$ and $\tau_\ell$ denote the eigenvalue and percentage of explained variance of the $\ell$th component ($\lambda_1$=.2517, $\tau_1$=23.84%; $\lambda_2$=.2350, $\tau_2$=22.26%)].

were instead interpreted as objects. This interpretation is compatible with the existence of two (at least partially) independent systems: one for faces and one for objects with each system being able to process different representations of “pictures of cats” and “faces of cats.”

### *Subject-specific map*

In a multiple subject experiment, such as the one reported here, it is also of interest to look at the specific pattern of a given subject. In order to do so, we consider only the projections of the scans for one subject. For example, Fig. 5a shows the tolerance intervals for subject 1 as well as the centers of gravity of the scans for this subject. The centers of gravity of subject 1 are close to the group center, and their configuration is similar to the group configuration. This indicates that the organization of the categories of subject 1 is similar to the group pattern. Subject 1's tolerance intervals are smaller than the group intervals, this indicates that a pattern classifier trained to identify the scans of a specific subject would perform better for this subject than on data coming from other subjects.

Conversely, it could be of interest to examine how subjects differ for a given category. As an illustration, Fig. 5b displays the tolerance intervals of all 6 subjects for the house category. This figure shows, for example, that the scans of subject 3 are tightly clustered and positioned far away from the center of the graph. This pattern indicates that subject 3's scans of houses are easily identifiable. By contrast, the scans of subject 2 are much more scattered and closer overall to the center of the graph. This pattern indicates that subject 2's scans of houses are less easily identifiable than subject 3's.

## Discussion

In this paper we describe a general procedure to show the reliability of MDS displays in the form of confidence and tolerance intervals. Confidence intervals can be used in lieu of standard null hypothesis testing. However, when the analysis involves more than two categories it is advisable to use a Bonferonni-like correction to adjust the confidence intervals for multiple comparisons. Tolerance intervals, in contrast to confidence intervals, directly express the *accuracy* of pattern classifiers. This makes tolerance intervals of particular interest to neuroimaging researchers.

The procedure described in this paper has been applied to an experiment which involves only one experimental factor. Future work should extend the range of the present technique to multiple factor experiments and adapt it to handle the analysis of interaction(s) between factors.

## Acknowledgments

We want to thank Jim Haxby for providing the original data used in this paper, Alice O'Toole and Fang Jiang for their help in implementing the pattern classifier used here, Nils Pénard for providing help and expertise for this project, Mary K. Reagor, Anjali Krishnan, and two anonymous reviewers for their very helpful comments on a previous version of this paper. MATLAB programs implementing the procedures described in this paper are available from the homepage of the first author (www.utdallas.edu/~herve).

## Appendix

### *Drawing tolerance and confidence ellipses*

In order to draw the tolerance or confidence ellipses, we use a PCA approach to compute the following necessary parameters: (1) the coordinates of the center of the ellipse in the map, (2) the angle between the major axis of the ellipse and the first dimension of the map, and (3) the relative size of the minor axis compared to the major axis of the ellipse. To do this, the center of the ellipse is set at the center of mass of the points. A centered PCA then gives a set of eigenvectors and eigenvalues. The ratio of the minor axis to the major

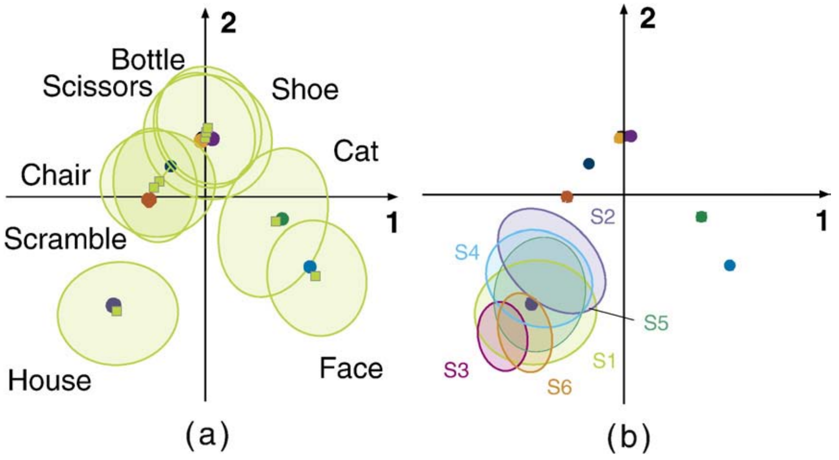

**Fig. 5.** Center of mass for subject 1 in the between subjects DISTATIS compromise, with 95% tolerance ellipses. Note that the large circles represent the category/group centers of mass, while the smaller squares represent the centers of mass for each category for subject 1 (b) The house category compared across subjects, with 95% tolerance ellipses. [$\lambda_\ell$ and $\tau_\ell$ denote the eigenvalue and percentage of explained variance of the $\ell$th component ($\lambda_1$=.2517, $\tau_1$=23.84%; $\lambda_2$=.2350, $\tau_2$=22.26%) for both maps)].

axis is then obtained as the ratio of the second eigenvalue to the first eigenvalue. The angle of rotation is given by the first eigenvector. For a 95% confidence interval, the length of the axes of the ellipse is set to ensure that the ellipse comprises 95% of the points.

## References

Abdi, H., 2007a. Metric multidimensional scaling. In: Salkind, N. (Ed.), Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks, (CA), pp. 598–605.

Abdi, H., 2007b. $R_V$ coefficient and congruence coefficient. In: Salkind, N. (Ed.), Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks, (CA), pp. 849–853.

Abdi, H., 2007c. Bonferroni and Šidàk corrections for multiple comparisons. In: Salkind, N. (Ed.), Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks, (CA), pp. 103–107.

Abdi, H., 2007d. Eigen-decomposition: eigenvalues and eigenvecteurs. In: Salkind, N. (Ed.), Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks, (CA), pp. 304–308.

Abdi, H., 2007e. Signal detection theory. In: Salkind, N. (Ed.), Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks, (CA), pp. 886–889.

Abdi, H., 2007f. Distance. In: Salkind, N. (Ed.), Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks, (CA), pp. 280–284.

Abdi, H., Valentin, D., 2007. Multiple Factor Analysis. In: Salkind, N. (Ed.), Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks, (CA), pp. 657–663.

Abdi, H., Valentin, D., Edelman, B., 1999. Neural Networks. Sage, Newbury Park, (CA, USA).

Abdi, H., Valentin, D., O'Toole, A., Edelman, B., 2005. DISTATIS: the analysis of multiple distance matrices. Proceedings of the IEEE Computer Society: International Conference on Computer Vision and Pattern Recognition. San Diego, CA, USA, pp. 42–47.

Abdi, H., Valentin, D., Chollet, S., Chrea, C., 2007. Analyzing assessors and products in sorting tasks: DISTATIS, theory and applications. Food Qual. Prefer. 18, 627–640.

Bishop, C., 2006. Pattern Recognition and Machine Learning. Springer, New York.

Chernick, M.R., 2008. Bootstrap Methods: A Guide for Practitioners and Researchers. Wiley, New York.

Cox, D.D., Savoy, R.L., 2003. Functional magnetic resonance imaging (fMRI) "brain reading:" detecting and classifying distributed patterns of fMRI activity in human visual cortex. NeuroImage 19, 261–270.

Duda, R.O., Hart, P.E., Stork, D.G., 2001. Pattern Classification. Wiley-Interscience, New York.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman and Hall, New York.

Escofier, B., Pagès, J., 1990. Multiple factor analysis. Comput. Stat. Data Anal. 18, 121–140.

Escoufier, Y., 1973. Le traitement des variables vectorielles. Biometrics 29, 751–760.

Hahn, G.J., Meeker, W.Q., 1991. Statistical Intervals: A Guide for Practitioners. Wiley-Interscience, New York.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning. Springer, New York.

Haxby, J., Gobbini, M., Furey, M., Ishai, A., Schouten, J., Pietrini, P., 2001. Distributed and overlapping representation of faces and objects in ventral temporal cortex. Science 293, 2425–2430.

Haynes, J.D., Rees, G., 2006. Decoding mental states from brain activity in humans. Nat. Rev. Neurosci. 7, 523–534.

Josse, J., Pagès, J., Husson, F., 2008. Testing the significance of the $R_V$ coefficient. Comput. Stat. Data Anal. 53, 82–91.

Kherif, F., Poline, J.-B., Mériaux, S., Benali, H., Flandin, G., Brett, M., 2003. Group analysis in functional neuroimaging: selecting subjects using similarity measures. NeuroImage 20, 2197–2208.

Lancaster, P., Tismenestsky, M., 1985. The Theory of Matrices. Academic Press, Orlando.

Lazraq, A., Hanafi, M., Cléroux, R., Allaire, J., Lepage, Y., 2008. Une approache inférentielle pour la validation du compromis de la méthode STATIS. J. Soc. Fr. Stat. 149, 98–109.

Norman, K., Polyn, S., Detre, G., Haxby, J., 2006. Beyond brain reading: multi-voxel pattern analysis of fMRI data. Trends Cogn. Sci. 10, 424–430.

Oliveira, M.M., Mexia, J., 2007a. ANOVA-like analysis of matched series of studies with a common structure. J. Stat. Plan. Inference 137, 1862–1870.

Oliveira, M.M., Mexia, J., 2007b. Modelling series of studies with a common structure. Comput. Stat. Data Anal. 51, 5876–5885.

O'Toole, A., Jiang, F., Abdi, H., Haxby, J., 2005. Partially distributed representations of objects and faces in ventral temporal cortex. J. Cogn. Neurosci. 17, 580–590.

O'Toole, A., Jiang, F., Abdi, H., Pénard, N., Dunlop, J., Parent, M., 2007. Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data. J. Cogn. Neurosci. 19, 1735–1752.

Robert, P., Escoufier, Y., 1976. A unifying tool for linear multivariate statistical methods: the $R_V$-coefficient. Appl. Stat. 25, 257–265.

Shinkareva, S., Ombao, H., Sutton, B., Mohanty, A., Miller, G., 2006. Classification of functional brain images with a spatio-temporal dissimilarity map. NeuroImage 33, 63–71.

Shinkareva, S.V., Mason, R.A., Malave, V.L., Wang, W., Mitchell, T.M., Just, M.A., 2008. Using fMRI brain activation to identify cognitive states associated with perception of tools and dwellings. PLoS ONE 3, e1394. URL http://dx.doi.org/10.1371/journal.pone.0001394.

Strother, S.C., Anderson, J., Hansen, L.K., Kjems, U., Kustra, R., Sidtis, J., Frutiger, S., Muley, S., LaConte, S., Rottenberg, D., 2002. The quantitative evaluation of functional neuroimaging experiments: the NPAIRS data analysis framework. NeuroImage 15, 747–771.

Togerson, W., 1958. Theory and Methods of Scaling. Wiley, New York.

Welchew, D., Honey, G., Sharma, T., Robbins, T., Bullmore, E., 2002. Multidimensional scaling of integrated neurocognitive function and schizophrenia as a disconnexion disorder. NeuroImage 17, 1227–1239.