SIViP (2014) 8:913–922
DOI 10.1007/s11760-012-0326-0

ORIGINAL PAPER

A spatial discrepancy measure between voxel sets in brain imaging

Ludmila I. Kuncheva · David Martínez-Rego · Kenneth S. L. Yuen · David E. J. Linden · Stephen J. Johnston

Received: 1 September 2011 / Revised: 21 April 2012 / Accepted: 24 April 2012 / Published online: 15 May 2012

Abstract Functional Magnetic Resonance Imaging serves to identify networks and regions in the brain engaged in various mental activities, represented as a set of voxels in the 3D image. It is important to be able to measure how similar two selected voxel sets are. The major flaw of the currently used correlation-based and overlap-based measures is that they disregard the spatial proximity of the selected voxel sets. Here, we propose a measure for comparing two voxel sets, called Spatial Discrepancy, based upon the average Hausdorff distance. We demonstrate that Spatial Discrepancy can detect genuine similarities and differences where other commonly used measures fail to do so. A simulation experiment was carried out where distorted copies of the same voxel sets were compared, varying the level of distortion. The experiment revealed that the proposed measure correlates better with the level of distortion than any of the other measures. Data from a 10-subject experiment were used to demonstrate the advantages of the Spatial Discrepancy measure in multi-subject studies.

Keywords Medical imaging · fMRI · Voxel selection · Pattern recognition · Discrepancy measure

1 Introduction

Functional magnetic resonance imaging (fMRI) is currently the most advanced non-invasive technology for identifying regions in the brain as they increase output in response to task demand. Traditionally, voxel activity maps are created by relating the individual voxel responses to some expected response. A voxel is deemed important if its evaluated relationship with the stimulus exceeds a pre-defined threshold. Pattern recognition takes voxel selection a step further by treating voxels as the features, the stimuli as the class labels, and the brain responses to the stimuli as the instances to be labelled. There are two groups of voxel selection methods. Univariate methods evaluate the importance of each voxel in the image separately. They are fast, reliable, reproducible and statistically sound, but they cannot capture any joint behaviour of the voxels [9]. Multivariate methods, on the other hand, based largely on pattern recognition, evaluate a subset of voxels at a time, where the voxels do not necessarily form a spatial neighbourhood. Multivariate methods are more time-consuming but offer higher accuracy and deeper insight into distributed patterns of brain functionality [5,12,30,31]. Feature selection is now a mature, rich and well-structured sub-domain of pattern recognition and machine learning [6,11,13,21]. Only recently has attention been turned to the problem of very large-scale feature selection, addressing

L. I. Kuncheva (✉)
School of Computer Science, Bangor University,
Dean Street, Bangor, Gwynedd LL57 1UT, UK
e-mail: l.i.kuncheva@bangor.ac.uk

D. Martínez-Rego
LIDIA Group, University of A Coruña, Campus de Elviña, s/n,
15407 A Coruña, Spain
e-mail: dmartinez@udc.es

K. S. L. Yuen
Institute of Systems Neuroscience, University Medical Centre
Hamburg-Eppendorf, Martinistr. 52, Hamburg, 22466, Germany
e-mail: k.yuen@uke.de

D. E. J. Linden
School of Psychology, Cardiff University, Tower Building,
70 Park Place, Cardiff CF10 3AT, UK
e-mail: LindenD@cardiff.ac.uk

S. J. Johnston
Psychology Department, Brunel University, Uxbridge UB8 3PH, UK
e-mail: stephen.johnston@brunel.ac.uk

classification tasks such as micro-array data analysis [19,42]. Given the abundance of voxel selection methods, it is surprising that comparison of their outputs has received so little attention. The aim of this article is to present a measure of discrepancy between two subsets of voxels which takes into account the spatial positioning of the voxel sets within the brain.

Similarity between sets of voxels has been quantified for studying variability or reliability of fMRI results across multiple subjects as well as multiple runs for a single subject (test–retest experiments) [25,26,40]. In addition to multi-subject variability, and variability due to the differences in the experimental protocol from one run to the next, variability in fMRI data can be due to random fluctuations or artefacts. These fluctuation can be of environmental, physiological and psychological nature (random errors), technical imperfections of the data collecting equipment, faults in the data pre-processing, etc. [23]. Thus, different voxel sets may be identified as important from two identically planned runs carried out at different times. Reliability studies are primarily focused on cleaning the noise and producing a more certain brain map showing the voxels deemed to be important (often termed active voxels) [25,40]. Reproducibility or activation maps do not give a value that measures the discrepancy between two fMRI outcomes.

Intra-class correlation coefficients (ICC) [37] have been used to capture the difference between within- and between-subject variability. They can serve as reliability measures for multi-subject and multi-session experiments [2,27,33,39]. A similar measure of dependence, the RV-coefficients [34], has been adapted to fMRI data analysis [16]. By design, neither ICC nor the RV-coefficients take into account the spatial relationship between the voxels in the brain.

fMRI reliability studies often base voxel importance on the value of a continuous-valued measure or statistic that can be thresholded to identify a set of significant voxels [25,36,39]. Feature selection methods, on the other hand, produce a *subset* of voxels. While correlation between two voxel subsets can still be calculated, the insightful scatterplots proposed by Specht et al. [39] and the certainty maps developed by Maitra [25] will not be applicable.

Discrepancy between two subsets of voxels can be evaluated by the overlap measure [24,35]. It is calculated as twice the ratio between the cardinality of the intersection and the sum of the cardinalities of the two voxels sets. Despite being well accepted for the purpose [33,39,44], the overlap measure has been criticised for its adverse sensitivity to the cardinalities of the compared sets [25]. The overlap measure is simple and intuitive but it ignores the spatial relationship between selected voxels.

A step closer to taking spatial relationships into account are the methods based upon labelling the selected voxels into clusters and comparing the clustering results. After the clusters have been identified, a matching procedure is applied, e.g., the Rand index. A problem with the clustering approach is that the results will depend on the clustering algorithms used (k-means, fuzzy clustering [3], genetic algorithm [1], neural networks [41], SVM [43], etc.). A straightforward clustering method coming from image analysis is to keep only voxels forming connected components larger than a certain size (η). For example, Thirion et al. [40] experiment with $\eta = 10$ and $\eta = 30$.

Having identified a gap in the fMRI comparison toolbox, here we propose a discrepancy measure that bypasses the clustering step and compares the voxels subsets directly, taking into account their spatial positioning within the brain. The rest of the paper is organised as follows. Section 2 details six commonly used discrepancy measures from the fMRI literature and introduces the proposed *Set Discrepancy* measure, D_S. Section 3 presents a simulation study to demonstrate the advantage of D_S over the other discrepancy measures. An experiment with a 10-subject data set is given in Sect. 4.

2 Discrepancy between voxel subsets

Consider a set of features (voxels) $V = \{v_1, v_2, \ldots, v_n\}$. Let $A \subseteq V$ and $B \subseteq V$ be two non-empty subsets of V with respective cardinalities $|A| = N_A$ and $|B| = N_B$. Let $r = |A \cap B|$ be the cardinality of the intersection of the two subsets. The discrepancy between A and B can be quantified in different ways.

2.1 Existing measures

Taking the overlap and the correlation measures from the fMRI reliability literature, we convert them into discrepancy indices in the following way

- Overlap index

$$D_O = 1 - \frac{2r}{N_A + N_B}. \tag{1}$$

The overlap discrepancy index is effectively the Rombouts et al.'s [35] overlap measure, negated and shifted by a constant to scale the discrepancy index between 0 (complete match) and 1 (complete mismatch).

- Correlation Index

$$D_\rho = \frac{1}{2} - \frac{rn - N_A N_B}{2\sqrt{N_A N_B (n - N_A)(n - N_B)}}. \tag{2}$$

The correlation index is the scaled negated correlation coefficient between the two binary variables corresponding to subsets A and B [37]. Both variables have length n and each bit corresponds to a voxel. The bits for the selected voxels are set to 1 and the bits for the non-selected voxels are set to 0. The discrepancy measure D_ρ takes value 0 if A and B are identical, and value 1 if they are complementary subsets

($A \cup B = V$, $A \cap B = \emptyset$). Statistical literature offers a myriad of measures of dependency between binary variables that can be used to the same effect [8,38].

Two simple discrepancy measures can be added to the list. These have been used before to study consistency of feature selection methods [7,15,17].

- Intersection-union cardinality ratio

$$D_{\mathrm{IU}} = 1 - \frac{|A \cap B|}{|A \cup B|} = 1 - \frac{r}{N_A + N_B - r}. \tag{3}$$

- Relative hamming distance

$$D_{\mathrm{RH}} = \frac{N_A + N_B - 2r}{n}. \tag{4}$$

Both D_{IU} and D_{RH} take value 0 if the two sets are identical and have maximum value 1. D_{IU} takes its maximum value 1 when the two sets are not intersecting, as does the overlap index D_O, regardless of the cardinalities of A and B. On the other hand, D_{RH}, like D_ρ, takes its maximum value of 1 for complementary A and B.

None of the four measures takes into account the spatial relationship between the voxels. Suppose that the two sets contain one voxel each. If the voxels are different, the intersection will be empty, hence $D_{\mathrm{IU}} = D_O = 1$, labelling the two sets as completely different. The Hamming distance will be $D_{\mathrm{RH}} = \frac{2}{n}$, which is approximately 0 given that n in a typical fMRI data set is of the order of tens of thousands. Finally, the correlation measure will be $D_\rho \approx 0.5$, signifying independence between the two sets. Now assume that the voxel in A is situated in the brain as a neighbour to the voxel in B. To account for this case, we need a measure that will recognise the proximity between the two selections. Therefore, we add a discrepancy measure based upon the Hausdorff distance between sets

- Hausdorff distance

$$D_H = \frac{1}{d_{\max}} \times \max\left\{\sup_{a\in A}\inf_{b\in B} d(a,b),\ \sup_{b\in B}\inf_{a\in A} d(a,b)\right\} \tag{5}$$

where $d(a, b)$ is the Euclidean distance between voxel a from set A and voxel b from set B within the three-dimensional brain array. The normalising constant $d_{\max}$ is the maximum possible value of the distance. For a 3D box containing the fMRI data, the largest distance between any pair of voxels is the length of the diagonal.[1] The maximum value $D_H = 1$ is attained when A and B contain a single voxel each situated at the two furthest corners in the fMRI volume. For identical sets, $D_H = 0$. Note that the empirical values measured on fMRI maps are likely to be orders of magnitude smaller than the maximum.

- Cluster-based distance. This measure was proposed by Thirion et al. [40] for evaluating reproducibility of fMRI activation maps across multiple subjects or multiple runs. Here, we use it for comparing two maps, A and B. First, clusters are identified within A and B as connected components consisting of η or more selected voxels. The centres of the clusters are calculated next. Let $\{x_1, \ldots, x_{C_A}\}$ be the C_A centres of clusters derived from voxel set A, and $\{y_1, \ldots, y_{C_B}\}$ be the C_B centres of clusters derived from voxel set B. The Cluster-Based Distance measure is

$$D_C = \frac{1}{2}\left\{\frac{1}{C_A}\sum_{i=1}^{C_A}\min_{j=1}^{C_B}\phi\left(||x_i - y_j||\right) + \frac{1}{C_B}\sum_{j=1}^{C_B}\min_{i=1}^{C_A}\phi\left(||x_i - y_j||\right)\right\} \tag{6}$$

where

$$\phi(\zeta) = 1 - \exp\left\{-\frac{\zeta^2}{2\sigma^2}\right\} \tag{7}$$

Given the two selected voxel sets, A and B, this distance requires the values of two parameters: the minimum cluster size η and the spread σ. The values used by Thirion et al. [40] are $\eta \in \{10, 30\}$ and $\sigma = 6$ mm.

2.2 The spatial discrepancy measure

- The spatial discrepancy measure is defined as follows

$$D_S = \frac{1}{d_{\max}(N_A + N_B)} \times \left\{\sum_{a\in A}\min_{b\in B} d(a,b) + \sum_{b\in B}\min_{a\in A} d(a,b)\right\} \tag{8}$$

The measure is the average spatial distance between the voxels selected in the two sets, scaled to span the interval [0, 1]. To calculate D_S, we sum up the distances from each selected voxel in A to the nearest voxel from B. Then, we add the sum of the distances from each selected voxel in B to the nearest voxel from A and finally divide by the number of distances, $N_A + N_B$. The advantage of the proposed measure over the other six measures is demonstrated through the examples shown in Fig. 1. Three cases of non-intersecting sets of selected voxels are displayed.

The values of the seven measures for the three cases are given in Table 1. The four measures that do not consider spatial context, D_O, D_ρ, D_{IU} and D_{RH}, fail to recognise the differences between the three cases, having constant values across (a), (b) and (c). On the other hand, D_H, D_C and D_S

[1] If $X_{\max}$, $Y_{\max}$ and $Z_{\max}$ are the maximum dimensions of the three axes, $d_{\max} = \sqrt{X_{\max}^2 + Y_{\max}^2 + Z_{\max}^2}$.

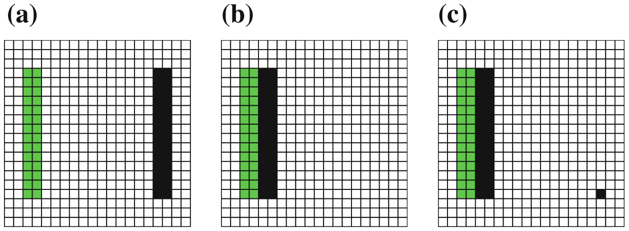

Fig. 1 Examples of non-intersecting A (*light*) and B (*dark*): **a** distant, **b** close, **c** close with an outlier

Table 1 Values of the seven measures for the example in Fig. 1

Measure	(a)	(b)	(c)
D_O	1.00	1.00	1.00
D_ρ	0.56	0.56	0.56
D_{IO}	1.00	1.00	1.00
D_{RH}	0.14	0.14	0.14
D_H	0.49	0.07	0.49
D_C	0.16	0.05	0.05
D_S	0.48	0.05	0.06

have lower values for the close selections (subplot (b)), indicating low discrepancy compared with the value for distant selections (subplot (a)). Comparing subplots (b) and (c), the Hausdorff measure, D_H, leaps to 0.49, wrongly indicating high discrepancy due to the outlier in set B. For computing the cluster-based discrepancy, D_C, we set the limit $\eta = 10$. The clustering, achieved by filtering out smaller connected components, is a form of smoothing of the image and will therefore wipe out the outlier in B. Thus, D_C will not be able to distinguish between cases (b) and (c). The proposed Spatial Discrepancy measure, D_S, is stable and consistent with the perceived discrepancy across the three subplots. Its value drops from (a) to (b) and increases only marginally from (b) to indicate the outlier in (c). To the best the authors' knowledge, a discrepancy measure with these properties has not been applied before for the analysis of fMRI data.

3 Simulations

3.1 Data 1. Single-subject, single-run fMRI experiment

The fMRI data were collected on a 3 Tesla Philips Achieva system (TR = 2 s, TE = 30 ms, 30 slices, 3 mm slice thickness, inplane resolution 2 mm × 2 mm). Pre-processing was performed using the Brainvoyager software QX (Braininnovation, Maastricht, The Netherlands). The data were corrected for intra-subject angular and translational motion and filtered to remove long-term drift.

The participant was a right handed male with corrected to normal vision, with no history of neurological or psychiatric illness. Prior to the start of the experiment, informed consent was obtained. The experimental protocol was approved by the ethics committees of the School of Psychology, Bangor University, and the North West Wales NHS Trust. The data set was a part of a neurofeedback experiment [14]. In a pilot run, the participant views images with negative and positive content. Subsequently, the specific brain regions corresponding to positive and negative emotion were localised for this participant. He was then instructed to upregulate his target brain region corresponding to negative emotion for periods of 20 s "up", alternating with baseline periods of 14 s "rest". Twelve up-rest cycles were recorded in the run, spanning a total of 204 scans (408 s).

To construct the data set, we averaged the 3D brain images recorded in the five scans around the peak of the predicted haemodynamic response function (HRF)[2] during the up-regulation phase and also the five scans around the trough during the rest phase. By doing this we apply 'temporal compression', found to be a useful pre-processing heuristic in single-subject experiments [29]. A grey-matter mask estimated from the anatomical MR image was applied, leaving a data set of 24 instances (12 negative emotion and 12 rest) × 33,274 voxels (features). The classification task is to recognise the state (negative or rest) from the brain image alone.

3.2 Comparison of distorted voxel subsets

Five hundred voxels were selected by sorting and cutting the p value of the two-sample t test, using all 24 instances. This set will be called "the original set", O. One hundred distorted versions of O were created and the seven discrepancy measures were calculated between O and each distorted version. The distortion was implemented in the following way. A magnitude of the distortion, Δ, in voxel positions, was a chosen randomly from the set $\{-5, -4, \ldots, 4, 5\}$. Then, $K\%$ of the *important* voxels were randomly chosen for shifting. Here, we experimented with $K \in \{10, 25, 50\}$. Each such voxel was shifted in a random direction: Left, Right, Anterior, Posterior, Dorsal and Ventral, by Δ positions. If the new location was already occupied by a selected voxel, the move was cancelled, otherwise the new voxel was marked as important and the voxel at the starting location was marked as unimportant. The processing of the chosen voxels was done in random order. Since some of the voxels might not move as a result, the distortion is said to be of *up to* $K\%$ of the labels. If exactly $K\%$ of the labels of important voxels were flipped, and the important voxels were moved to new locations, D_O, D_ρ, D_{IU} and D_{RH} would have constant values regardless of

[2] The haemodynamic response function (HRF) models the expected changes in bloodflow that follows a neural event.

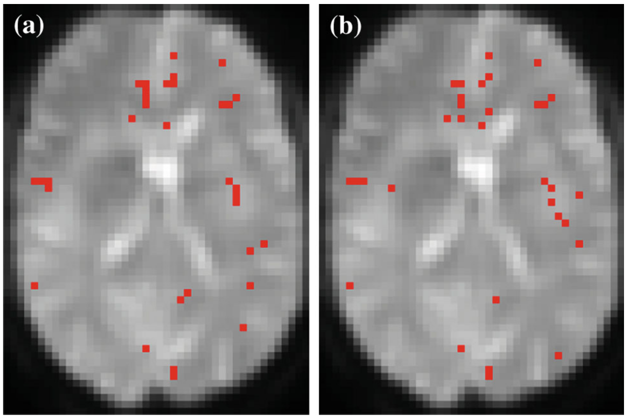

Fig. 2 **a** A slice of the original image O; **b** The same slice after distortion of 25 % of the labels, jump distance 4

Δ because none of r, N_A, N_B and n changes from one distorted set to the next. We added two outliers in each distorted image. Two random non-important voxels were re-labelled as important. Figure 2a shows O, and Fig. 2b shows an example of a distorted set.

Figure 3 shows the plots of the measure values against Δ. Each plot contains 3 curves corresponding to $K \in \{10, 25, 50\}$. The grey points show the individual (Δ, measure) values from which the averages are calculated. The estimates of the correlation $\hat{\rho}$ and the Spearman's rank correlation coefficient $\hat{\rho}_R$ between the measure and Δ are shown in Table 2. The proposed spatial discrepancy measure correlates very well with the magnitude of the distortion. The farther the jump, the larger the D_S value. All other measures show much lower correlations with Δ. The two outliers completely fool the Hausdorff measure which would have had a high correlation with Δ in the absence of outliers. D_S manages to ignore the outliers while being finely responsive to the Δ.

4 A multi-subject experiment

The purpose of this experiment is to evaluate the how well the seven measures capture the similarities and difference between the voxel sets selected through three different methods.

4.1 Data 2. Multi-subject, 10 fMRI runs

The experimental protocol was approved by the ethics committees of the School of Psychology, Bangor University, and the North West Wales NHS Trust. The participants' task was to passively view a set of 'emotionally charged' images in a block type design, while fMRI data were collected and preprocessed as explained in Sect. 3.1. Each block of images consisted of pictures of a single emotional valence type, either positive, negative or neutral. The images were selected from the international affective picture system (IAPS) [20], which have been pre-tested in normative samples for their valence (emotion evoked in participants with a scale of 1 to 9, ranging from "unhappy" to "happy") and arousal (scale from 1 to 9, ranging from "calm" to "excited").

The data set for each subject consisted of 329 instances where each TR in the fMRI sequence was taken to be one instance. The instances were labelled in four classes: positive stimuli, negative stimuli, neutral stimuli and fixation. The labels corresponded to the stimuli at the time of taking the TR. A common grey-matter mask consisting of 59,707 voxels was calculated.

4.2 Comparison of three voxel selection methods

We ran voxel selection separately on each subject data. Three voxel selection methods were applied:

- t-selection. The voxels were ranked according to the t test scoring function and the top 1,000 were retained.
- W-selection. The voxels were ranked according to the scoring function for the Wilcoxon rank sum test and the top 1,000 were retained.
- SVM-selection. The support vector machine classifier SVM has proven its worth for high-dimensional data [10]. SVM is particularly suited to wide data because it scales linearly along the feature dimension while tolerating the small sample size by ensuring large classification margins. The linear-kernel SVM can be used as a feature ranking algorithm. A feature's relevance is measured by the absolute value of the weight for this feature in the linear discriminant function of the trained SVM [18,28].

Since there are four classes and all three voxel selection methods operate for two classes, we calculated the measures for all pairs of classes and averaged the scores. The top voxels were those which scored the best on the average scores. By applying the three methods to the 10 subjects, we created 30 voxel subsets. A "random subject" was created for comparison. Three independent random subsets of 1,000 voxels, one for each method, were sampled from the voxel set. Thus, there were 33 voxel subsets altogether. A good discrepancy measure is expected to demonstrate the following:

- The measure should find the T and W methods more similar to one another than T and SVM.
- It is expected that a good measure should detect larger variability of the selected sets *between* subjects compared with the variability *within* subjects.

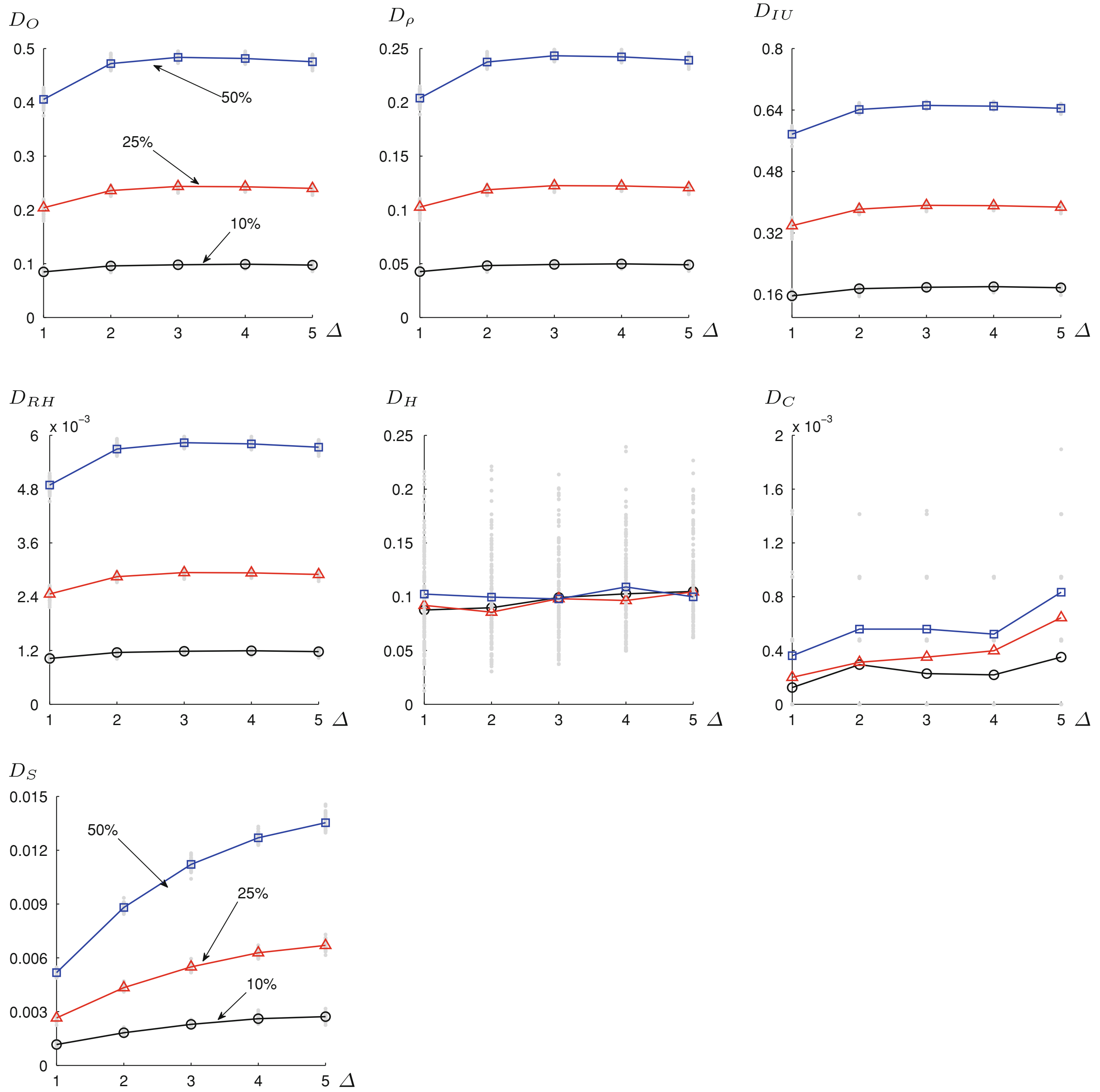

Fig. 3 The values of the seven discrepancy measures versus the magnitude of the distortion Δ calculated from 100 distorted versions of the original set O

– The measure should be able to indicate that the voxel sets for the random subject (randomly and independently sampled subsets) are notably different from the sets for all real subjects.

We calculated the $11 \times 11 = 121$ discrepancies between the pairs of voxel sets selected through T and W (denoted T/W) and also the 121 discrepancies between T and SVM (denoted T/SVM). Figure 4 shows the behaviour of the seven discrepancy measures in 2d. Each plot contains 121 points. Consider subjects i and j ($i, j \in \{1, 2, \ldots, 11\}$). The ordered pair of subjects (i, j) generated a point on each graph. The x-coordinate is the discrepancy between the T subset of i and the W subset of j. The y-coordinate is the discrepancy between the T subset of i and the SVM subset of j. If the three voxel sets were identical, the point would lie at the origin of the coordinate

Table 2 Correlation ρ and rank correlation ρ_R (both in %) between the seven measures and the distortion value Δ for $K =$ 10, 25 and 50 %, as plotted in Fig. 3

Measure	10 %		25 %		50 %	
	ρ	ρ_R	ρ	ρ_R	ρ	ρ_R
D_O	64.4	58.1	70.2	60.5	69.7	53.5
D_ρ	64.4	58.1	70.2	60.5	69.7	53.5
D_{IO}	64.4	58.1	70.1	60.5	69.7	53.5
D_{RH}	64.4	58.1	70.2	60.5	69.7	53.5
D_H	16.3	14.5	12.0	11.5	1.5	3.5
D_C	15.6	14.7	34.3	32.1	32.3	32.0
D_S	94.3	94.2	96.5	97.3	96.3	97.8

system. If the sets were very different, the point would be further up the diagonal indicating high discrepancy between the T and W subsets as well as high discrepancy between T and SVM subsets. If, however, T and W subsets were close whereas T and SVM subsets were different, the point is expected to appear towards the top left corner of the plot. The scatterplots were annotated by colour to facilitate the interpretation. Blue circles indicate the pairs where one or both are the random subject. Green squares indicate the same subject. Thus, in each plot, there are 21 blue circles: 10 for pairs (i = random, j = other), 10 for (i = other, j = random) and one for (i = random, j = random). There will be 11 green points for the pairs (i, i). The point (i = random, j = random) will be in both colours.

As indicated by green colour, the cluster on the left corresponds to within-subject discrepancies. Its position shows that the discrepancy between T and W is smaller than the discrepancy between T and SVM and also that the within-subject discrepancies are smaller than between-subject discrepancies (the main cluster in the middle). Apart from the plots for D_H and D_C, the cluster with the individual discrepancies is quite pronounced. The random subject is clearly identifiable only in the plots for D_H and D_S as a cluster far along the diagonal. The position of this cluster should be above and to the right of the black dot cluster. The reason is that random voxel subset should have high discrepancies with a selected subset compared with the discrepancy between any pair of selected subsets, whatever the selection method. The figure shows that only the proposed spatial discrepancy measure D_S behaves as desired.

As with any pairwise measure, D_S can be used to evaluate the discrepancy between a collection of M voxel subsets by averaging the $M(M-1)/2$ pairwise values. Alternatively, D_S could be extended to carry out the calculation in one pass, as is done with other non-pairwise measures [4,22]. One possible extension could be by creating M sums in (8), where the distance will be to the nearest selected voxel in any of the other $M-1$ sets. Such an extension, however, merits a separate study in the context of similar multi-set measures.

5 Conclusions

The identification of the degree of spatial discrepancy of voxel sets in fMRI has a wide range of potential applications. Here, we propose a novel discrepancy measure, D_S, calculated as the average distance between each selected voxel and the nearest selected voxel from the alternative set. We argue that set-theoretic measures, often applied for comparing two voxel sets, are not well suited for the task because they are not equipped to take into consideration any spatial relationship between the elements of the two sets. We have also shown that the Hausdorff metric is adversely affected by outliers in the selected voxel set. We demonstrated the use of D_S on fMRI data for getting insights about the results of six voxel selection methods. The measure could equally be applied to maps obtained with univariate and multivariate analyses. It can be used to assess similarity across trials, runs, or sessions of the same experiment or across participants for assessing group homogeneity [16].

It would be interesting to include the brain surface map in the calculation of the distance function. Voxels in close proximity to one another may be at a considerable distance on the brain surface. Such account of the brain surface is only applicable if we are analysing the results from different methods *on the same* brain. It has been argued that spatial discrepancies from one subject's brain to another may vary as far as 1 cm [40]. In this case, a correction of the distance to account for the brain surface will be of little use and will rather contribute to the noise, unless the multi-subject alignment procedure takes care of inter-individual variability of cortical folding patterns [32].

While some possible uses of D_S were demonstrated through illustrations and simulations, we should also declare what the proposed measure *is not* designed/ suitable for. None of the pairwise measures discussed here is a straightforward measure of reliability. However, they can be taken as a component in the process of assessing reliability of fMRI, as demonstrated through the real-data experiment. Second, the analyses in this paper have not been designed with activation maps in mind, nor are they based upon thresholding some statistic, calculated for the individual voxels. We

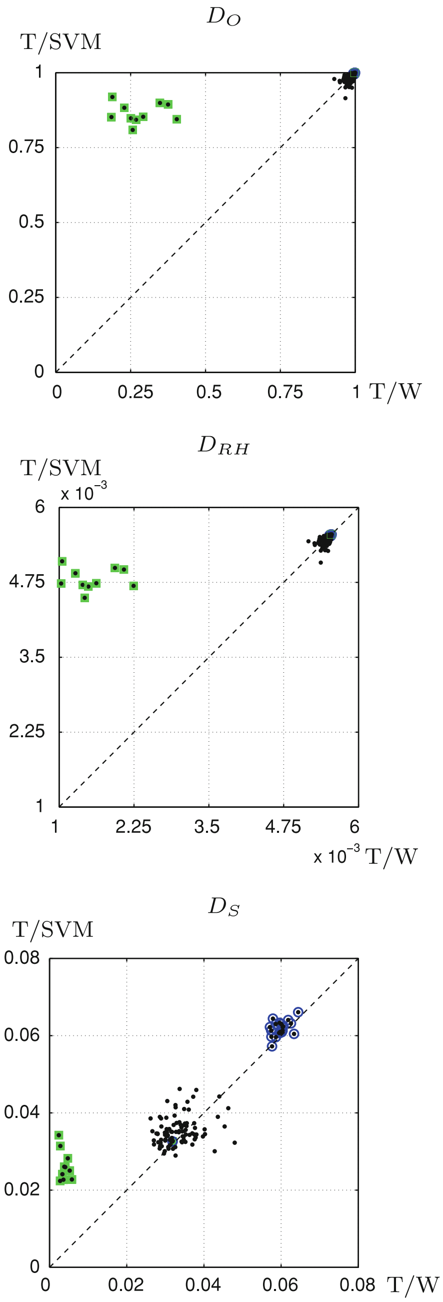

Fig. 4 The seven discrepancy measures for the real data. *Blue circles* indicate the random subject. *Green squares* indicate the same subject. Only D_S manages to separate clearly the random subject from the rest, while recognising at the same time the within-subject similarities (colour figure online)

designed D_S to quantify the outcomes of *feature selection* methods, where the features are selected as a *group*, and the cardinality of the selected sets is similar. Third, the paper does not advocate a particular voxel selection method nor does it look for an optimal number of selected voxels. The fMRI data are used as an illustration; hence, we have not carried out cross-validation experiments to evaluate classification accuracies or regions of interest in the brain related to discrimination between positive and negative emotions.

An exciting future use of D_S is for creating a landscape of the existing voxel selection methods. Through gauging their classification success, a gap may be found for new, even more successful voxel selection methods. The landscape may suggest what characteristics these new methods should possess, e.g., multivariate versus univariate, correlation-rewarding versus correlation-penalising, and so on.

Variations of D_S can be tried too. For example, the contributions from the two sets of voxels are scaled differently. Instead of dividing by $N_A + N_B$, the sums in Eq. (8) can be weighted, respectively, by $1/N_A$ and $1/N_b$. The Euclidean distance in (8) can be replaced by the squared Euclidean distance, the Minkovski distance or any other distance that is deemed suitable.

Acknowledgments We would like to thank Dr Federico De Martino for the useful comments on the draft of this study. David Martínez Rego was supported by the grant TIN2006-02402 and the FPU program, both financed by Ministerio de Ciencia e Innovacion of Spain.

References

1. Aberg, M.B., Wessberg, J.: An evolutionary approach to the identification of informative voxel clusters for brain state discrimination. In: IEEE J. Sel. Top. Signal Process. **2**(6), 919–928 (2008)
2. Aron, A.R., Gluck, M.A., Poldrack, R.A.: Long-term test–retest reliability of functional mri in a classification learning task. NeuroImage **29**(3), 1000–1006 (2006)
3. Baumgartner, R., Somorjai, R., Summers, R., Richter, W., Ryner, L., Jarmasz, M.: Resampling as a cluster validation technique in fMRI. Magnetic Resonance Imaging **11**, 228–231 (2000)
4. Cabral, C., Silveira, M., Figueiredo, P.: Decoding visual brain states from fmri using an ensemble of classifiers. Pattern Recogn. **45**(6), 2064–2074 (2012)
5. Clithero, J.A., Carter, R.M., Huettel, S.A.: Local pattern classification differentiates processes of economic valuation. NeuroImage **45**(4), 1329–1338 (2009)
6. Dash, M., Liu, H.: Feature selection for classification. Intell. Data Anal. **1**, 131–156 (1997)
7. Dunne, K., Cunningham, P., Azuaje, F.: Solution to instability problems with sequential wrapper-based approaches to feature selection. Technical Report TCD-CS-2002-28. Department of Computer Science. Trinity College. Dublin, Ireland (2002)
8. Fleiss, J.L.: Statistical Methods for Rates and Proportions. Wiley, New York (1981)
9. Friston, K.J., Ashburner, J., Kiebel, S.J., Nichols, T.E., Penny, W.D.: editors. Statistical Parametric Mapping: The Analysis of Functional Brain Images. Academic Press, New York (2007)
10. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3**, 1157–1182 (2003)
11. Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L.A. (eds.): Feature Extraction, Foundations and Applications. Studies in Fuzziness and Soft Computing. Physica-Verlag, Springer (2006)
12. Haxby, J.V., Gobbini, M., Furey, M.L., Ishal, A., Schouten, J.L., Pietrini, P.: Distributed and overlapping representation of faces and objects in ventral termporal cortex **293**, 2425–2430 (2001)
13. Jain, A.K., Mao, J.: Guest editorial: special issue on artificial neural networks and statistical pattern recognition. In: IEEE Trans. Neural Netw. **8**(1), 1–3 (1997)
14. Johnston, S.J., Boehm, S.G., Healy, D., Goebel, R., Linden, D.E.J.: Neurofeedback: a promising tool for the self-regulation of emotion networks. Neuroimage **49**(1), 1066–1072 (2010)
15. Kalousis, A., Prados, J., Hilario, M.: Stability of feature selection algorithms. In Proceedings of 5th In: IEEE International Conference on Data Mining (ICDM'05), pp. 218–225 (2005)
16. Kherif, F., Poline, J.-B., Mériaux, S., Benali, H., Flandin, G., Brett, M.: Group analysis in functional neuroimaging: selecting subjects using similarity measures. NeuroImage **20**(4), 2197–2208 (2003)
17. Kuncheva, L.I.: A stability index for feature selection. In Proceedings of IASTED. Artificial Intelligence and Applications, pp. 390–395. Innsbruck, Austria (2007)
18. LaConte, S., Strother, S., Cherkassky, V., Anderson, J., Hu, X.: Support vector machines for temporal classification of block design fmri data. NeuroImage **26**(2), 317–329 (2005)
19. Lai, C., Reinders, M.J.T., Wessels, L.: Random subspace method for multivariate feature selection. Pattern Recogn. Lett. **27**(10), 1067–1076 (2006)
20. Lang, P.J., Bradley, M.M., Cuthbert, B.N.: International Affective Picture System (IAPS): Technical Manual and Affective Ratings. NIMH Center for the Study of Emotion and Attention. University of Florida (1997). http://csea.phhp.ufl.edu/media/iapsmessage.html
21. Langley, P.: Selection of relevant features in machine learning. In: Proceedings of AAAI Fall Symposium on Relevance, pp. 140–144 (1994)
22. Langs, G., Menze, B.H., Lashkari, D., Golland, P.: Detecting stable distributed patterns of brain activation using gini contrast. NeuroImage **56**(2), 497–507 (2011)
23. Liou, M., Su, H.R., Lee, J.D., Cheng, P.E., Huang, C.C., Tsai, C.H.: Bridging functional MR images and scientific inference: reproducibility maps. J. Cogn. Neurosci. **15**(7), 935–945 (2003)
24. Machielsen, W.C.M., Rombouts, S.A.R.B., Barkhof, F., Scheltens, P., Witter, M.P.: fMRI of visual encoding: reproducibility of activation. Hum. Brain Mapp. **9**(3), 156–164 (2000)
25. Maitra, R.: Assessing certainty of activation or inactivation in test–retest fMRI studies. NeuroImage **47**(1), 88–97 (2009)
26. Maitra, R., Roys, S.R., Gullapalli, R.P.: Test–retest reliability estimation of functional MRI data. Magn. Reson. Med. **48**(1), 62–70 (2002)
27. Manoach, D.S., Halpern, E.F., Kramer, T.S., Chang, Y.C., Goff, D.C., Rauch, S.L., Kennedy, D.N., Gollub, R.L.: Test–retest reliability of a functional MRI working memory paradigm in normal and schizophrenic subjects. Am. J. Psychol. **158**(6), 955–958 (2001)
28. Mourao-Miranda, J., Bokde, A.L.W., Born, C., Hampel, H., Stetter, M.: Classifying brain states and determining the discriminating activation patterns: Support Vector Machine on functional mri data. NeuroImage **28**(4), 980–995 (2005)
29. Mourao-Miranda, J., Reynaud, E., McGlone, F., Calvert, G., Brammer, M.: The impact of temporal compression and space selection on SVM analysis of single-subject and multi-subject fMRI data. NeuroImage **33**(4), 1055–1065 (2006)
30. Norman, K.A., Polyn, A.M., Detre, G.J., Haxby, J.V.: Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn. Sci. **10**, 424–430 (2006)
31. Pereira, F., Mitchell, T., Botvinick, M.: Machine learning classifiers and fMRI: a tutorial overview. NeuroImage **45**(1, Supplement 1), S199–S209 (2009)
32. Goebel, R., Esposito, F., Formisano, E.: Analysis of functional image analysis contest (FIAC) data with brainvoyager QX: From single-subject to cortically aligned group general linear model analysis and self-organizing group independent component analysis. Hum. Brain Mapp. **27**(5), 392–401 (2006)
33. Raemaekers, M., Vink, M., Zandbelt, B., van Wezel, R.J.A., Kahn, R.S., Ramsey, N.F.: Test–retest reliability of fmri activation during prosaccades and antisaccades. NeuroImage **36**(3), 532–542 (2007)
34. Robert, P., Escoufier, Y.: A unifying tool for linear multivariate statistical methods: The RV- coefficient. J. R. Stat. Soc. Ser.C (Appl. Stat.) **25**(3), 257–265 (1976)
35. Rombouts, S.A.R.B., Barkhof, F., Hoogenraad, F.G.C., Sprenger, M., Scheltens, P.: Within-subject reproducibility of visual activation patterns with functional magnetic resonance imaging using multislice echo planar imaging. Magn. Reson. Imag. **16**(2), 105–113 (1998)
36. Salli, E., Korvenoja, A., Visa, A., Katila, H.J., Aronen, T.: Reproducibility of fMRI: effect of the use of contextual information. NeuroImage **13**(3), 459–471 (2001)
37. Shrout, P.E., Fleiss, J.L.: Intraclass correlations—uses in assessing rater reliability. Psychol. Bull. **86**(2), 420–428 (1979)

38. Sneath, P.H.A., Sokal, R.R.: Numerical Taxonomy. W.H. Freeman & Co, New York (1973)
39. Specht, K., Willmes, K., Shah, N.J., Jancke, L.: Assessment of reliability in functional imaging studies. J. Magn. Reson. Imag. **17**(4), 463–471 (2003)
40. Thirion, B., Pinel, P., Meriaux, S., Roche, A., Dehaene, S., Poline, J.-B.: Analysis of a large fMRI cohort: statistical and methodological issues for group analyses. NeuroImage **35**(1), 105–120 (2007)
41. Voultsidou, M., Dodel, S., Herrmann, J.M.: Neural networks approach to clustering of activity in fMRI data. In: IEEE Trans. Med. Imag. **24**(8), 987–996 (2005)
42. Xing, E.P., Jordan, M.I., Karp, R.M.: Feature selection for high-dimensional genomic microarray data. In: Proceedings of the 18th International Conference on Machine Learning, (ICML2001) (2001)
43. Yang, J., Zhong, N., Liang, P., Wang, J., Yao, Y., Lu, S.: Brain activation detection by neighborhood one-class SVM. Cogn. Syst. Res. **11**(1), 16–24 (2010)
44. Yoo, S.-S., Fairneny, T., Chen, N.-K., Choo, S.-E., Panych, L.P., Park, H.-W. LeeS.-Y., Jolesz, F.A.: Brain-computer interface using fMRI: spatial navigation by thoughts. NeuroReport **15**(10), 1591–1595 (2004)