NeuroImage 65 (2013) 69–82

Contents lists available at SciVerse ScienceDirect

NeuroImage

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

Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): Random permutations and cluster size control

Johannes Stelzer [a,*], Yi Chen [a,b], Robert Turner [a]

[a] Max-Planck-Institute for Human Cognitive and Brain Sciences, Leipzig, Germany
[b] Bernstein Center for Computational Neuroscience Berlin and Charité - Universitätsmedizin Berlin, Germany

ARTICLE INFO

Article history:
Accepted 25 September 2012
Available online 4 October 2012

Keywords:
MVPA
fMRI
Statistics
Multiple testing
Cluster size control
Second level analysis

ABSTRACT

An ever-increasing number of functional magnetic resonance imaging (fMRI) studies are now using information-based multi-voxel pattern analysis (MVPA) techniques to decode mental states. In doing so, they achieve a significantly greater sensitivity compared to when they use univariate frameworks. However, the new brain-decoding methods have also posed new challenges for analysis and statistical inference on the group level. We discuss why the usual procedure of performing *t*-tests on accuracy maps across subjects in order to produce a group statistic is inappropriate. We propose a solution to this problem for local MVPA approaches, which achieves higher sensitivity than other procedures. Our method uses random permutation tests on the single-subject level, and then combines the results on the group level with a bootstrap method. To preserve the spatial dependency induced by local MVPA methods, we generate a random permutation set and keep it fixed across all locations. This enables us to later apply a cluster size control for the multiple testing problem. More specifically, we explicitly compute the distribution of cluster sizes and use this to determine the *p*-values for each cluster. Using a volumetric searchlight decoding procedure, we demonstrate the validity and sensitivity of our approach using both simulated and real fMRI data sets. In comparison to the standard *t*-test procedure implemented in SPM8, our results showed a higher sensitivity. We discuss the theoretical applicability and the practical advantages of our approach, and outline its generalization to other local MVPA methods, such as surface decoding techniques.

Introduction

In this paper, we present a non-parametric approach for dealing with group-level analysis in local multi-voxel pattern analysis (MVPA) methods based on classification.

Local MVPA, or information-based brain mapping, aims to assess the task-related neural information at every location in the brain, by analyzing the signal patterns extracted from a spatial neighborhood (searchlight) centered at the location (Chen et al., 2011; Kriegeskorte et al., 2006). A convenient way to estimate the information contained in these regions is classification-based MVPA. This involves training a classifier on a subset of the data and predicting the class labels of another, yet unseen subset of the data. Thereby, the *generalizability* of the classifier is assessed (Kriegeskorte, 2011). The average percentage of correctly predicted labels, known as the decoding accuracy, is taken as an indicator of the information content of the searchlight volume. Customarily, the accuracy is mapped to the central voxel of the searchlight. The repetition of this procedure for all searchlight locations in the brain mask, results in a three-dimensional accuracy *map*, which reflects the spatial distribution of information decodable from the functional brain images.

In the context of neuroscientific studies, the decoding accuracies themselves are usually not of primary interest. Instead, the *statistical significance* of the decoding accuracy on the group level is of relevance. For this, it is common practice (e.g., Bode and Haynes, 2009; Carlin et al., 2012; Kahnt et al., 2010) to estimate a group-level statistic by performing a voxel-wise *t*-test against the theoretical chance level (e.g., an accuracy of 0.5 in a two class paradigm) using the accuracy maps of all subjects. Finally, the multiple testing problem is corrected at the cluster level with family-wise error (FWE) or false-discovery rate (FDR) methods.

This commonly practiced group statistic procedure is, however, questionable for various reasons; particularly, the low number of observations and the non-gaussianity of the probability distribution of accuracy. As a consequence, several assumptions of the *t*-statistics are not met, rendering the procedure invalid from a theoretical point of view. We will demonstrate that *t*-test procedures (for decoding studies) are problematic not only from a strict theoretical perspective, but also from practical considerations: using simulations, we will show that *t*-test procedures, which implement cluster control with FDR, exhibit exceedingly high levels of false positivity. Alternatively, the evaluation of a classifier's performance can be modeled as Bernoulli trials (Pereira and Botvinick, 2011) and this leads to a binomial test for the subsequent statistical inference. However, the dependency between the cross-validation folds causes a

* Corresponding author at: Max Planck Institute for Human Cognitive and Brain Sciences, Department of Neurophysics, Stephanstrasse 1a, D-04103 Leipzig, Germany.
E-mail address: stelzer@cbs.mpg.de (J. Stelzer).

1053-8119/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.neuroimage.2012.09.063

problem here, because the aggregated performance over all the cross-validation folds is modeled as the performance of a single classifier. This approximation is only valid if the underlying binomial variables from each cross-validation are assumed to be independent. Because the cross-validation procedure introduces a correlation between the binomial variables, the assumption of independency is challenged. Moreover, it is unclear to which extent the deviation of the accuracy distribution from Gaussian and binomial distribution has an effect on the group-level statistical test.

To overcome these potential pitfalls, and as a non-parametric alternative, permutation tests can be used to assess statistical significance. Permutation tests for fMRI analysis were pioneered more than 15 years ago (Arndt et al., 1996; Holmes et al., 1996). An excellent primer on the topic is found in Nichols and Holmes (2002). Permutation tests rely on minimal assumptions (Good, 2006) and their use in the context of classification has been verified theoretically (Golland and Fischl, 2003). These tests have now also been adapted to fMRI studies using classification-based MVPA (Chen et al., 2011; Pereira and Botvinick, 2011) and MVPA toolboxes (Hanke et al., 2009).

The idea behind permutation tests for classification is to estimate the dependency between class labels and observations. In a general sense, the null hypothesis for permutation tests is defined as an independency between class labels and observations. To approximate the probability distribution of accuracy under this null hypothesis empirically, a large number of permutations are applied on the class labels and the corresponding accuracies are estimated. The probability or significance level for rejecting the null hypothesis is then evaluated by comparing the original accuracy against the accumulated empirical distribution.

The practice of permutation tests, in the context of fMRI decoding studies on the group level, however, is not directly evident. On the one hand, the permutation test applied on the single subject level produces the significance of labeling for each subject, while the assessment of statistical significance is more desirable on the group level. On the other hand, the number of (independent) observations in fMRI studies is small, which greatly limits the number of available permutations and thus the precision of statistical significance evaluation. To account for this, we combine the permutation tests on the single subject level with a bootstrapping procedure on the group level: instead of evaluating the significance of labeling on the single subject level, we first generate subject-wise the empirical null distribution. Next, a bootstrapping procedure is used to build an empirical null distribution of the mean accuracy across subjects, allowing voxel-wise estimation of significance. This procedure allows us to overcome the limitation of both the potentially small number of available permutations and the computational resources.

The voxel-wise null hypothesis, given our proposed method is defined as the following: the mean group accuracy follows an empirical group null distribution of accuracy values (which is derived by permutation and bootstrapping methods). In other words, the null hypothesis for our method states that there is no class information present at a group level and, hence the classifiers behaved randomly. In contrast, the (voxel-wise) null hypothesis for *t*-based methods is defined such that the sample of single subject accuracies stems from a normal distribution centered at an accuracy value of 0.5.

In whole brain analysis approaches, voxel-wise statistics are often not of ultimate interest, because the statistical testing procedures are applied many times (about 50,000 locations for whole brain data at 3 T and 500,000 at 7 T). Hence, both *t*-based methods and permutation test procedures are subject to the multiple testing problem. Setting an arbitrary threshold on statistical significance level could be questionable; due to the sheer number of statistical tests a large number of false positives may arise if low thresholds are applied, while high thresholds may obscure true effects. On the other hand, the statistical tests at proximate spatial locations are known to be interdependent, mainly due to two facts: first, the BOLD effects of interest are spatially widespread over several voxels (see Chumbley and Friston, 2009); second, local MVPA approaches introduce spatial correlations in the analyzing procedure. For instance, the voxels extracted by spherical searchlights largely overlap at adjacent locations.

To address this multiple testing problem, we followed Nichols and Holmes (2002) and used a cluster size inference on the group level. The basic idea of cluster size inference is to exploit the fact that the probability of two voxels exceeding a given voxel threshold and *simultaneously being contiguous* is smaller than the chance of one sole voxel surpassing a threshold (Forman et al., 1995). In this approach, the fundamental units of interest are therefore regions and not voxels (Heller et al., 2006). Furthermore, cluster-based approaches have been demonstrated to be statistically more powerful than voxel-based tests (Hayasaka and Nichols, 2003). It is important to emphasize that a cluster size inference applicable for *local* MVPA is required to account not only for the spatial correlations due to the BOLD effect, but also for spatial correlations from the analyzing procedure (e.g., searchlight). Furthermore, it should be mentioned that the latter source of correlation strongly depends on the location and local information content.

Hence, we ultimately consider cluster-wise statistics instead of voxel-wise statistics; we are interested in how likely it is to observe a *cluster* of voxels which surpass a certain voxel-wise threshold. In commonly practiced *t*-based methods, the probability of the occurrence of a cluster (voxels being contiguous and surpassing a *t*-based voxel-wise threshold) is usually derived by random field methods in combination with FWE or FDR corrections. Most critically, the underlying smoothness of the accuracy maps has to be estimated for this. In our non-parametric approach, we empirically construct a cluster size distribution and use FDR corrections on the cluster level. Our procedure implicitly implements the smoothness of the accuracy maps and renders the estimation of the latter obsolete.

To conclude, our study aims to provide a framework for non-parametric inference for classification-based decoding on a group level, which controls for multiple testing. Using a volumetric searchlight technique, we demonstrate the benefits of our method:

- We will show the *sensitivity* of our approach by using a simulation where the information content's size and distribution are known. This allows us to quantify and thus compare the detection rate.
- We will demonstrate the *validity* of our method using a large number of null simulations (i.e., simulations where the actual null hypothesis holds up). Every significant cluster found here can, thus, be ascribed to false positivity.
- We will display the *applicability* and sensitivity of our method using a real fMRI data set.

For all data sets, we compare our method to the common practice of conducting *t*-tests with multiple testing corrections.

Materials and methods

In this section we want to illuminate the methods used in this paper. The methods section is divided into two parts, one constitutes the raw data generation used (two simulation approaches and one fMRI experiment). The second part describes the statistical analysis of the raw data.

Simulated data: comparison to standard methods

To compare the statistical sensitivity and the special accuracy between our proposed method and *t*-based approaches, we created a data simulation with information deposited at known locations. We used Matlab (MathWorks, Inc.) to generate 12 data sets representing single "subjects" and processed these in the same way that we processed the fMRI data set. Each data set consisted of 16 volumes in two conditions (eight for *A* and eight for *B*, representing eight runs with two conditions). The volumes comprised blocks the size of $108 \times 17 \times 17$ voxels (we chose this format for illustrative reasons). For condition *A*, the volumes were filled with uniformly distributed random numbers at an interval of [0,1]. We created the volumes for condition *B*

in the same way, but added an offset at specific spatial locations. Therefore, only in these locations was information about the condition present. The spatial locations were five cubes with an edge length of six voxels, which were aligned in the middle of the volumes and were 12 voxels apart. Within the cubes, we added an offset between 0.15 and 0.2 to the uniform noise (with the most left cube having an offset of 0.15 and the most right at 0.2). Furthermore, we applied two information degradation procedures: in order to account for inter-session variability present in real data, we randomly subtracted a value between 0% and 50% of the offset of the corresponding cube for each of the eight volumes. Additionally, to account for the inter-subject variability, such as anatomical differences, we randomly subtracted 0–50% of the offset of the corresponding cube for each data set, i.e., we subtracted the same percentage for all eight volumes. In other words, the information content in the cubes depended on the session, the virtual subject and the position of the cube, with the most left cube having an offset between 0 and 0.15 and the most right cube having an offset between 0 and 0.20.

Note that the simulation was used as raw data for the classifier, i.e., we treated the simulation identical to if it were an fMRI data set. All subsequent analysis was carried out using our proposed methods described in later paragraphs of the method section (searchlight decoding, permutation tests, group inference and cluster statistics) and standard methods described in the end of the methods section (searchlight decoding, standard *t*-tests, and multiple comparison on cluster level).

Simulated null data — validation

For validation, we used Matlab (MathWorks, Inc.) to generate 1000 group data sets. Each of these group data sets consisted of 10 "single subject" data sets. Note that the number of subjects used in this simulation is smaller than in the simulation above, for computational reasons. Each of the single subject data sets consisted of 10 volumes. The volumes comprised blocks of $30\times30\times30$ voxels. All volumes were filled with noise drawn from a uniform distribution in the interval [0,1]. In contrast to the previous simulation (for method comparison), there was no information deposited at any location, as the intention of this simulation was to empirically validate the false-positive rate.

We analyzed each group data set the same way as we did in the previous simulation (the simulation for comparison with the standard methods), while varying the cluster threshold between $p_{\text{cluster}}=0$ and $p_{\text{cluster}}=0.15$ in equidistant steps. Hence we could obtain the total number of clusters independent on the cluster threshold for both our proposed method and the standard *t*-test methods. For our proposed method, we only analyzed the first 100 group data sets (due to the large computation time), while we analyzed all 1000 group data sets for the standard *t*-test methods.

Furthermore, we investigated the effects of severe undersampling of the permutation space. As the number of data volumes on a single subject level becomes big, the number of available permutations becomes huge. Therefore, we depicted three datasets consisting of 12 virtual subjects and one single location (i.e., one single searchlight position) with a varying number of data points (observations) per virtual subject (80, 120 and 160 observations). Our choice of the number of data points is not motivated by first principles; we chose multiple large set sizes which allow different numbers of possible permutations. All data sets lacked information about the condition, thus were null data sets. We analyzed the four data sets with our proposed method (excluding cluster statistics) for different numbers of permutations on single subject level.

Experimental design (fMRI data set)

Fourteen healthy subjects participated in the study ($M=27.7$, all right-handed). The subjects were paid for their participation and gave written consent. The data sets of two subjects were incomplete and were discarded from further analysis, leaving a total of 12 participants.

The experiment consisted of 80 trials, which were split into two runs. Each trial lasted 19.2 s and was separated by a variable inter-trial interval of between 9.4 and 12.2 s. Participants were instructed to tap, with their right index finger, in time with four isochronous experimental conditions: a *discrete auditory* pacing sequence (50 ms sine beeps at 1350 Hz), a *continuous auditory* pacing sequence (pitch sweeps between 1350 Hz and 450 Hz, $T_{cycle}=600$ ms), a *discrete visual* sequence (a white bar flashed for 50 ms over a black background), and a *continuous visual* sequence (a white bar moving up and down). In the fMRI study, each of the four conditions was presented with a slow and a fast variant (inter-stimulus-interval = 400 ms or 600 ms). For this study, we collapsed the slow and fast variants of each condition together. The presentation of the pacing sequences was randomized using a computer and the software Presentation (Neurobehavioral Systems), which also recorded the tap timing.

Data acquisition and preprocessing (fMRI data set)

Functional MRI data (gradient EPI) was collected on a Siemens 3 T system (Trio) with a standard head coil. The scans contained 36 axial slices covering the whole brain (TR = 2000 ms, TE = 24 ms, slice thickness 4 mm with 1 mm gap, in plane resolution 3×3 mm^2). A sagittal T1-weighted anatomical scan was obtained from our database for all subjects (3 T Siemens Trio system, TR = 1300 ms, TE = 3.93 ms, FOV = 256×240 mm^2, 64 slices, slice thickness 1 mm, in plane resolution 1×1 mm^2).

We split the functional data into five parts, which had the same number of trials per condition. The data was then corrected for head motion, coregistered to the anatomical scan and spatially normalized to the MNI305 space (preprocessing performed in SPM8, Wellcome Department of Imaging Neuroscience, Institute of Neurology, London, UK). After this, a standard hemodynamic response function model was fitted to the data to estimate the statistical parameters (scaling parameters, beta values) for each of the four experimental conditions, resulting in one beta map for every part per experimental condition. The differences between different statistical inference methods are mostly illustrated in the results from the two visual conditions (discrete vs. continuous, in total 10 beta maps), which are presented in the main text. We include the results from the two auditory conditions in the supplementary materials, which show very similar though slightly weaker effects.

Searchlight decoding

We used the standard spherical searchlight approach (Bode and Haynes, 2009; Kriegeskorte et al., 2006) with a searchlight diameter of five voxels. The data generally consisted of $2\times k$ 3D volumes, where $2\times k$ is the number of observations and k is the number of temporal subdivisions of the data (in the fMRI example above, $k=5$). The observations themselves were always split into two experimental conditions of the same count. For each location within the brain mask, we extracted the voxels contained in the searchlight sphere of all $2\times k$ observations. This subset of voxels was put into a linear support vector machine (Chang and Lin, 2011), performing a k-fold LOOCV procedure: each cross-validation step used $2\times(k-1)$ observations ($k-1$ from each condition) as training set and two observations (one from each condition) as test set (see Fig. 2). Over the course of k cross-validation folds, the classifier was trained on the training set and the labels of the unseen test set were predicted. The average accuracy (i.e., the percentage of correctly predicted labels) of the k cross-validation steps was then mapped onto the center of the location. The procedure was applied to all locations of a whole brain mask and resulted in a map of decoding accuracies. To investigate the effects of searchlight diameter, we applied several different diameter values for our first simulation (where information was present in cubes). We used a total of five diameters: three, five, seven, nine and eleven voxels. The searchlight volumes to these diameters were 19

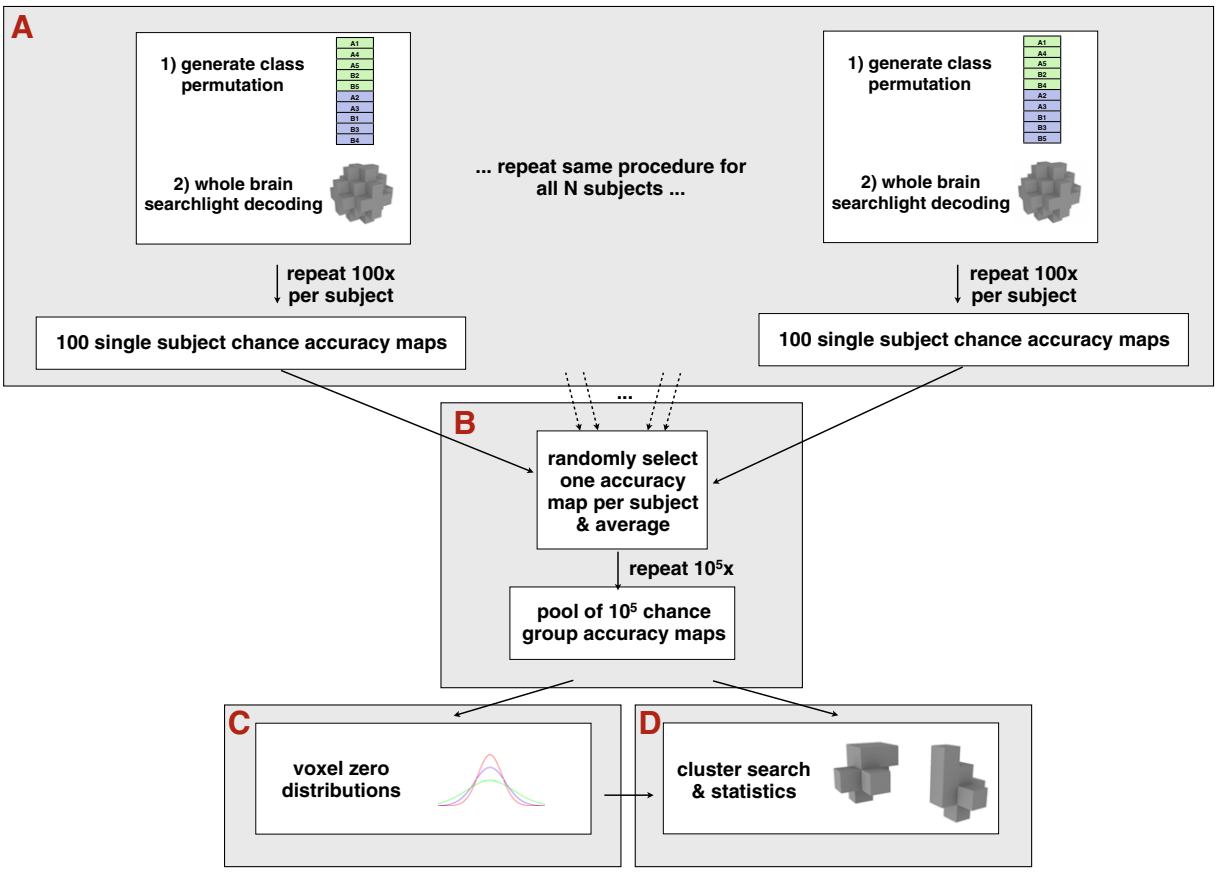


Fig. 1. Schematic diagram of the steps performed for the cluster level thresholding. (A) Whole brain searchlight decoding using a permutation of the training label on a single-subject level. The whole procedure was repeated M times ($M \geq 100$) per subject for all N subjects, resulting in a pool of $N \times M$ single-subject chance accuracy maps. (B) One chance accuracy map was randomly selected per subject for 10^5 times. The selection of N maps (for N subjects) was averaged to one group accuracy map, resulting in a pool of 10^5 group chance accuracy maps. (C) Voxel-wise obtained empirical chance distributions, given the pool of 10^5 group chance accuracy maps. (D) The empirical chance distributions allowed the determination of a voxel-wise threshold which was used for a cluster search in the group chance accuracy maps.

(D = 3), 57 (D = 5), 171 (D = 7), 365 (D = 9), and 691 (D = 11) voxels, respectively.

Additionally, we implemented an alternative searchlight technique, which maps the accuracies to every voxel contained in the searchlight (Björnsdotter et al., 2011). Because voxels might participate in multiple searchlights, the final accuracy is defined as the mean accuracy of all searchlights in which a voxel took part. The results for this procedure can be found in the supplementary materials.

Permutation tests on single subject level

On the single subject level, we employed permutation tests. It is important to highlight that the permutation tests in our method do *not* serve as a direct means to infer statistical significance on a single subject level, rather they should be regarded as an intermediate module for subsequent group recombination and inference methods (see Fig. 1 for an overview). The reasoning behind permutation tests on a single subject level is to assess chance distributions and chance accuracy maps (Chen et al., 2011; Golland and Fischl, 2003). For this, we created a random permutation of the observation order and applied it to the data set (see Fig. 2 for an overview of the permutation scheme).

Using the searchlight approach, the classifier was trained and tested for all locations. The resulting "chance" map of decoding accuracies was saved. Most crucially it should be noted that, on the one hand, we used one fixed permutation for each searchlight course (which preserved spatial correlations) and, on the other hand, the relationship between class label and data point remained constant for all cross-validation steps (which preserved the correlation between the cross-validation steps). Furthermore, we would like to highlight that the proposed permutation scheme avoids bias due to an uneven class distribution in the test or training set.

The procedure of generating a permutation and subsequent searchlight classification using permuted labels was repeated 100 times per subject. This is due to computational reasons, as the unrestricted permutation space contains 3.6×10^6 permutations and an exhaustive enumeration is not feasible.

Group methods for statistical inference

On the group level, we recombined the single subject accuracy maps gained from the permuted classification into group accuracy maps. For this, we randomly drew (with replacement) one of the 100 chance accuracy maps for each subject, and averaged this selection to one permuted group accuracy map (equivalent to a Monte Carlo technique). This procedure was repeated 10^5 times resulting in 10^5 permuted group accuracy maps.

For each voxel position, we constructed the empirical chance distribution using the 10^5 instances of permuted group accuracy maps (Chen et al., 2011). Hence, this allowed an estimation of the histogram of chance accuracy values. Next, we determined a (voxel-wise) accuracy level, which corresponds to a *low probability* if it were retrieved by chance. This was done by finding the accuracy for which the right-tailed area of the normalized histogram of voxel-wise accuracies was below 0.001. This accuracy was noted as threshold accuracy; the procedure of finding this threshold accuracy was repeated for each voxel. In a statistical notion, the threshold accuracy marks the p-value ($p_0 = 10^{-3}$), given the null

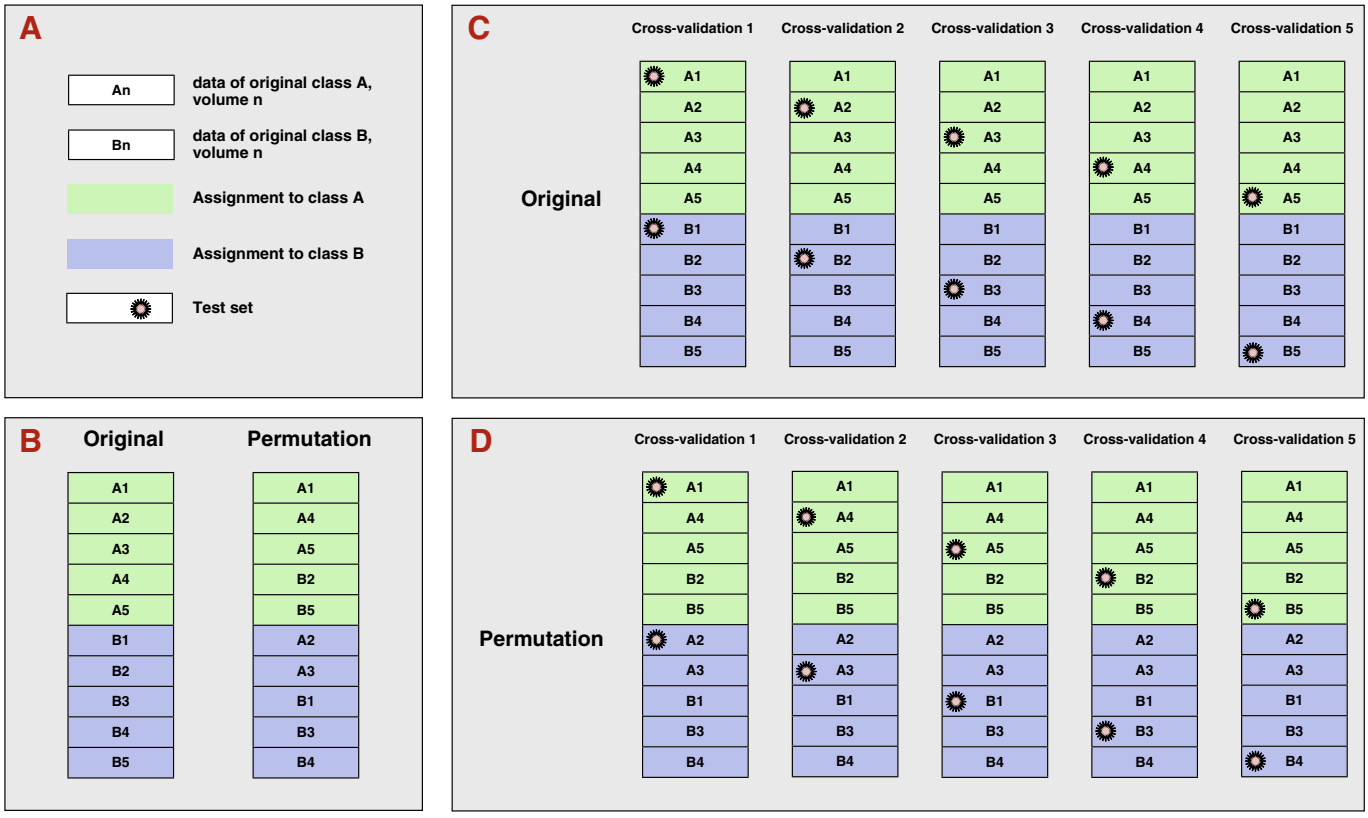


Fig. 2. Schematic overview of the permutation procedure. (A) Legend: A box with An *n* represents data point *n* of the original class A (Bn *n* represents data point *n* of original class B). The color shading stands for the class assignment (green for A and blue for B). The little star in the corner of the box marks this data point as the test set, which is unseen in the classifier training and the labels of these marked data points are predicted by the classifier. (B) In the original data set the data points of class A and B are assigned to class A and B respectively. In the permuted data set on the right the data points are shuffled. (C) Full cross-validation scheme of the original data set: in each cross-validation scheme, the classifier is trained on a subset of the data points and tested on another (unseen) data set (the latter is marked with stars). (D) Full cross-validation scheme for the permuted data set. Note that the sample count of the train and test subset is always balanced between the two classes and that the relationship between data sample and label remains fixed for all cross-validation steps.

hypothesis that the mean group accuracy was obtained by pure chance. Note that by virtue of this procedure, we are effectively applying a fixed-effects analysis. We saved both the threshold map and the distribution for use in the cluster search algorithm.

Cluster search algorithm

To perform a cluster search, we used a 6-connectivity scheme; two voxels were considered connected if they shared a face, but not an edge or a vertex. Instead of joining a connected voxel if it surpassed a fixed global threshold, we used the threshold map that was acquired from the empirical chance distribution functions: a connected voxel was joined to a cluster only if its accuracy exceeded the accuracy corresponding to a *p*-value of $p = 10^{-3}$ in the threshold map, i.e., if its accuracy was below 10^{-3} against the empirical chance level of this specific location. The choice of this value is based on common practice. To investigate the influence of this initial parameter, we varied it using a total of six different voxel thresholds ($p_1 = 0.05$, $p_2 = 0.01$, $p_3 = 0.005$, $p_4 = 0.001$, $p_5 = 0.0005$, $p_6 = 0.0001$) in our first simulation (where information was deposited). Furthermore, to exclude effects of our choice of the cluster search parameters, we implemented an 18-connectivity scheme in the cluster search. In that search, voxels which shared either a face or an edge, but not a vertex, were considered connected. The effect of the different connectivity schemes was also tested for the first simulation.

Cluster-size statistics

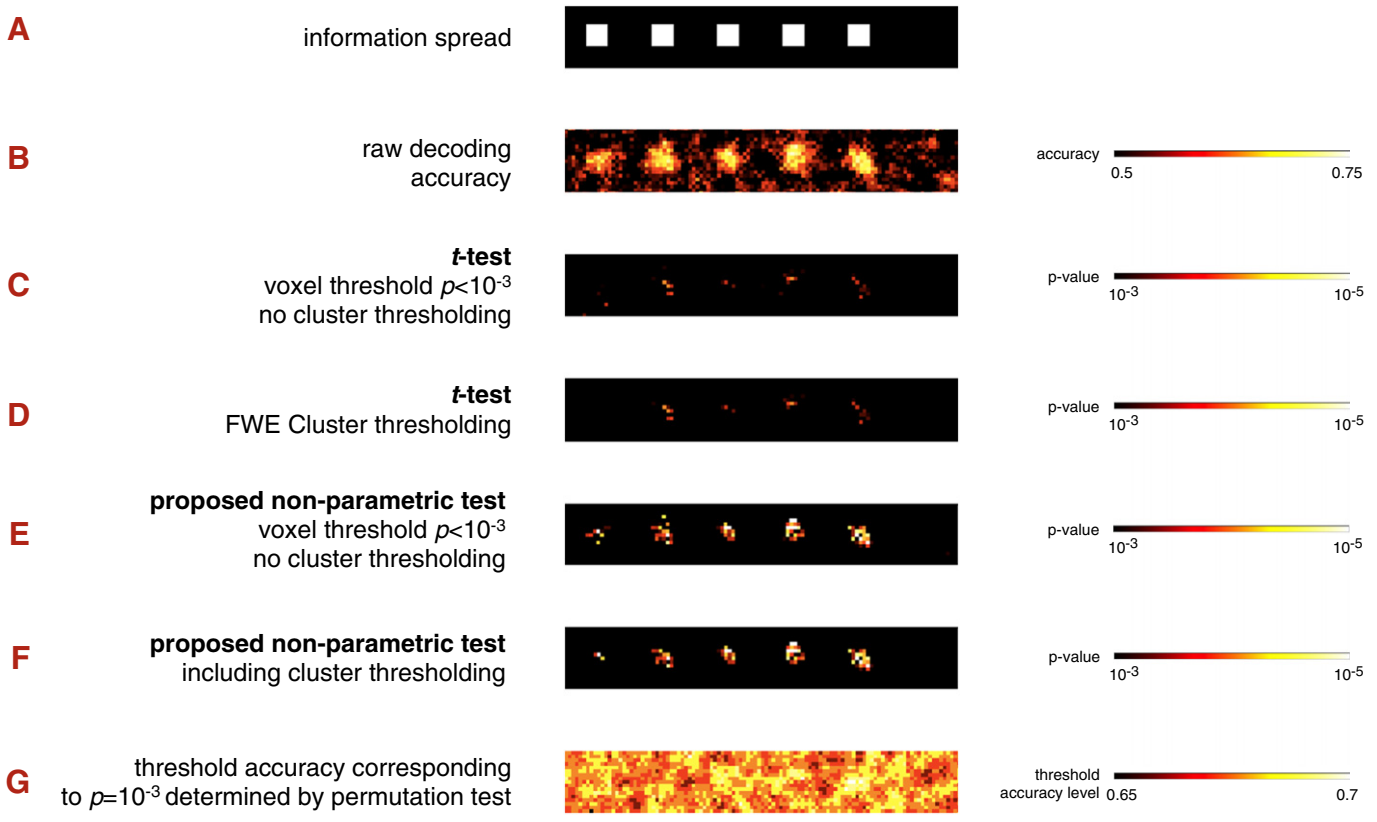
We applied a cluster search using the above algorithm in the 10^5 permuted group accuracy maps and collected the occurring cluster sizes. For each of the permuted group accuracy maps we recorded all occurring cluster sizes. We applied the same cluster search to the original group accuracy map and gathered the present cluster sizes. Given the two cluster size records (one from the chance population and one from the original data), it is possible to compute the chance level for the occurrence of a discovered cluster size in the original data: a cluster with the size *s* is computed to have a *p*-value of

$$p_{\text{cluster}} = \sum_{s' > s}^{\infty} H_{\text{cluster}}(s')$$

where H_{cluster} is the normalized histogram of cluster sizes in the chance record (occurrence divided by the total number of detected clusters). Hence, we can assign a *p*-value to each cluster size and introduce a threshold of cluster size for reaching significance. To correct for multiple comparisons at cluster level, we implemented a step-down FDR method (Benjamini, 1999) on all cluster *p*-values. A cluster size threshold was then applied to the original group accuracy maps, yielding filtered accuracy maps. The voxel-wise *p*-values for these clusters were defined as

$$p_{\text{voxel}} = \sum_{a' > a}^{1} H_{\text{voxel}}(a')$$

where *a* is the original accuracy of the voxel and H_{voxel} is the normalized chance distribution at the specified location.



H

cluster level *p*-value (FWE)	0.068	0	0	0	0	standard statistics in SPM8
cluster level *p*-value (FDR)	0.008	0	0	0	0	
minimum voxel-wise *p*-value (FWE)	1	0.957	0.165	0.746	0.098	
minimum voxel-wise *p*-value (FDR)	0.877	0.635	0.558	0.635	0.558	
cluster level *p*-value	0.0023	0	0	0	0	cluster size inference
minimum voxel-wise *p*-value (Local)	$<10^{-5}$	$<10^{-5}$	$<10^{-5}$	$<10^{-5}$	$<10^{-5}$	
minimum voxel-wise *p*-value (Global)	3×10^{-6}	8×10^{-8}	4×10^{-8}	4×10^{-8}	3×10^{-10}	

Fig. 3. Simulated data set with the respective classification results and statistics. (A) Information spread of the raw data (which served as the input for the classifier) over all sessions and virtual subjects. Note the five cubes where information was stored and the variation of information content: the first cube from the left has the smallest information content, progressing to the last on the right where the information was at the maximum. (B) Mean decoding accuracy map over all 12 virtual subjects. (C) Standard voxel-wise *t*-test without multiple testing correction, carried out with SPM8. The voxel threshold was set to $p=10^{-3}$. (D) The same *t*-test as in C, but corrected for multiple testing using FWE cluster thresholding. (E) Results of the new method proposed based on permutation and bootstrapping methods. The map displays the *p*-values without cluster size thresholding, i.e. without multiple testing correction. (F) Results of the new method proposed based on permutation and bootstrapping methods and cluster size control. The map displays the *p*-values and already implements a multiple testing correction. (G) Threshold map for the cluster search algorithm. The map displays an inhomogeneity of the local chance distribution. (H) Table with the results for each of the five cubes (represented as columns) using different statistical measures (represented as rows). The first four rows display standard methods, the last three rows indicate our new proposed method. Furthermore, we added the minimum *p*-values for both methods, which represent the smallest *p*-value within each cluster.

Standard t-statistics

We compared our proposed method to the common practice of performing *t*-tests on the accuracy against chance level and subsequent FDR or FWE correction. Note that no smoothing was applied here. The second-level statistics were carried out in SPM8, where we calculated a *t*-test with an uncorrected threshold of $p<10^{-3}$ and thresholded this map using a FDR or FWE correction on a cluster level.

Results

We applied our method to multiple simulated data sets and a real fMRI data set. In the first simulation, we were able to precisely define informative regions, and assess their information content more accurately, compared to standard procedures. The second simulation allowed us to validate our method using multiple null data sets with no information content. In both the simulations and the real data set, we compared the results obtained by our method with the commonly practiced *t*-test based second-level analysis in SPM8.

Simulated data set (comparison to standard methods)

The cluster distribution of the 10^5 chance group accuracy maps is shown in Fig. 4. The red dashed line indicates the cluster size at the $p=0.05$ level, where the right tail area of the (normalized) cluster size distribution is smaller than 0.05. For this *p*-value (or smaller), clusters need to have an extent of five voxels or more.

A detailed comparison between our proposed method and standard statistics is found in Fig. 3. Here, we show one slice of the simulated data, which includes the five cubes where information had been entered. The arrangement of the cubes is displayed in Fig. 3A. The white regions indicate the five informative cubes and the black regions indicate the noise background. The raw group accuracy map, which is the average of the 12 single-subject accuracy maps, is depicted in Fig. 3B.

Upon visual inspection, it is possible to locate the five informative regions, while in some other areas (especially on the right side) the noise background appears with a very similar structure. Fig. 3C shows the standard *t*-test against chance level, carried out in SPM8. The voxel threshold was set to $p = 10^{-3}$ without further correction. When a cluster-size FWE threshold in SPM8 is applied, this statistical map is further reduced, as shown in Fig. 3D. Here, we increased the necessary cluster extent so that the smallest cluster remaining passes the cluster-wise FWE threshold of $p_{cluster} < 0.05$. Fig. 3F demonstrates the new cluster size control proposed: Only clusters of a size of six voxels or more were included, everything below that size was filtered out. The map shows the voxel-wise *p*-values derived from the chance distribution of each location. The uncorrected *p*-value map, i.e. with no cluster thresholding applied is depicted in Fig. 3E. The threshold map for a voxel-wise $p = 10^{-3}$ level is displayed in Fig. 3G, revealing the spatial inhomogeneity found in the FWHMs of the chance distribution. Fig. 3H shows the detailed statistics for both the standard and our proposed approach. We included statistics on both a cluster level and on the voxel level. For our framework, we included two measures for *p*-values. The local one indicates the probability of the achieved accuracy on this specific location. The global *p*-value stands for the probability of the achieved accuracy given the joint chance distribution from all locations, in a notion of a maximum statistic.

Out of the five informative regions in our data simulation, four (FWE controlled) or all five (FDR controlled) were revealed using SPM8 and *t*-tests. When our proposed method was used, all five informative regions could be decoded. The total informative area principally contained $6 \times 6 \times 6 \times 5 = 1080$ voxels in total. The standard SPM8 method with FWE correction on cluster level labeled approximately 11% of this volume as significant (127 voxels), while our approach determined that approximately 24% of the informative volume was significant (258 voxels). This represents an increase of over 100% in terms of sensitivity.

Influence of the initial voxel threshold

We varied the initial voxel threshold, that is, the threshold for a voxel to be counted as belonging to a cluster, to investigate its impact on the cluster statistics. We used a total of six different voxel thresholds ($p_1 = 0.05$, $p_2 = 0.01$, $p_3 = 0.005$, $p_4 = 0.001$, $p_5 = 0.0005$, $p_6 = 0.0001$) and performed both our permutation method and the standard *t*-test approaches. Next, we considered all voxels that had been declared significant and checked which fraction of these significant voxels were located inside the originally informative regions, and which were outside. The results are displayed in Fig. 5. Note that significant voxels located outside the informative regions should not be directly considered false positive in a strict sense, because the searchlight center may be outside of an informative region, while its borders overlap with informative regions (i.e., the searchlight acts as spatial filter).

Fig. 5 depicts the results of the variation. For very low thresholds (p_1, p_2), the amount of significant voxels in informative regions is comparable for all three implementations (permutation cluster, *t*-test and FDR, *t*-test and FWE). However, the voxels declared significant in non-informative regions by *t*-tests for the first threshold is very high. For every higher threshold (p_2 to p_6), the total number of voxels declared significant that are located in informative regions is larger when our method is used. The number of significant voxels outside of informative regions is lower for our technique for the four lowest thresholds (p_1 to p_4), for

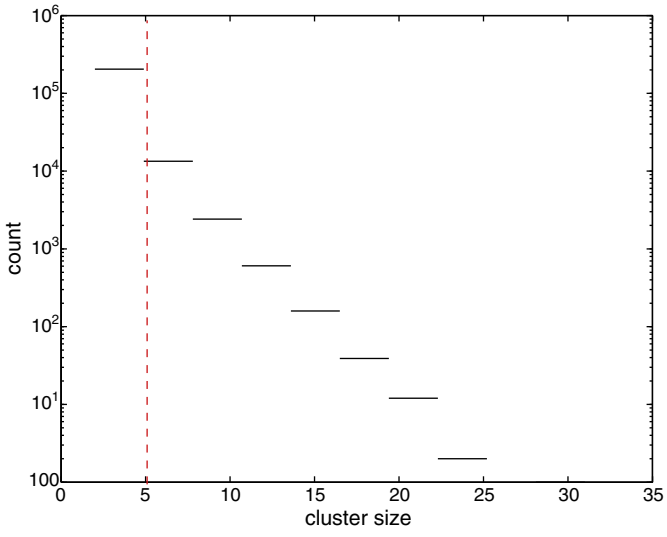


Fig. 4. Cluster size histogram from the simulated data set. The red line marks the $p = 10^{-3}$ percentile of the cluster size distribution (cluster size = 62 voxels).

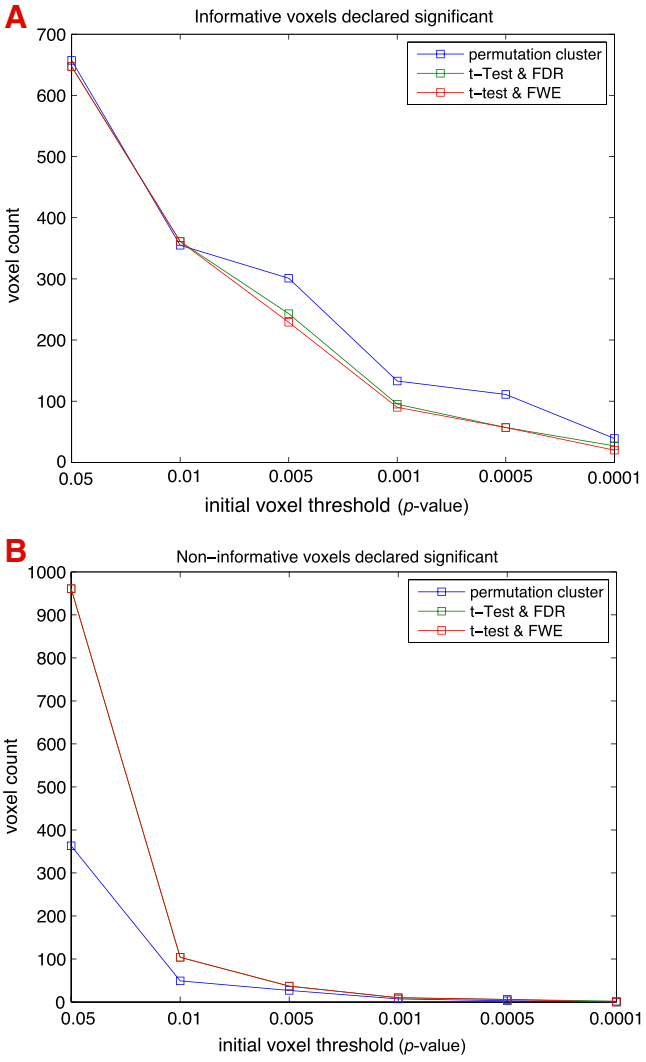


Fig. 5. Dependency of the initial voxel threshold (simulation). (A) Number of voxels declared significant, which were located within the informative regions. With higher thresholds (i.e., smaller initial voxel *p*-values), the number of voxels declared significant decreases. For *p* values smaller than $p = 0.005$, the proposed method declares more voxels significant as compared to the *t*-test based frameworks. (B) Number of voxels declared significant in non-informative regions. Also here, for higher thresholds (smaller *p*-values) less voxels are declared significant. For all thresholds, the proposed method declares a smaller fraction of voxels significant in the non-informative regions. Note that in this particular simulation, the number of voxels declared significant outside the informative regions is the same for both *t*-test methods. Furthermore it is noteworthy, that in the *t*-test implementations the larger fraction of voxels declared significant actually resides outside of the informative regions.

the highest thresholds the number of voxels is slightly higher for our method.

Influence of searchlight diameter

We varied the searchlight diameter over five values: three, five, seven, nine and eleven voxels were used. Both our proposed method and standard *t*-based methods were analyzed with the different diameters. For this we calculated the number of voxels declared significant inside, and outside the informative regions. Note that as the searchlight diameter increases, this rather strict measure becomes less sensible, because the searchlight itself acts as spatial filter. In other words, when larger searchlights are employed, the likelihood of mapping accuracy values *outside* the informative regions increases, because a sufficient fraction of the searchlight might lie *inside* informative regions while the *center is not*. To compensate for this, we included an additional measure: where the informative regions were extended by the searchlight diameter. The results are depicted in Fig. 6. Throughout all searchlight diameters, our method defines a larger number of voxels as being significant. For small searchlight volumes (three and five voxels) a volumetric gain of more than 100% can be achieved, for larger diameters the gain is about 50–70% (Fig. 6A). The number of voxels declared significant outside the informative cubes is similar for both methods if small searchlight diameters are used. For larger diameters, the amount of significant voxels outside the informative cubes becomes larger, especially for our proposed method (Fig. 6B). However, as discussed above, for larger searchlight diameters the strict borders of informative regions can be considered to be less meaningful. The results for the extended borders are depicted in Figs. 6C and D. The gain in performance for our method is similar here, and all methods are well behaved; the volume outside of the extended informative regions is negligible or zero.

Influence of connectivity scheme

By default, we applied a 6-connectivity scheme for our cluster search. To investigate potential influences regarding the connectivity scheme, we computed the cluster statistics using an 18-connectivity scheme. Next we considered all voxels declared significant inside and outside the informative regions for both our proposed method and *t*-based statistics. Using the 18-connectivity scheme, 288 voxels inside the informative regions were declared significant (256 voxels with 6-connectivity), while 21 voxels outside the informative regions

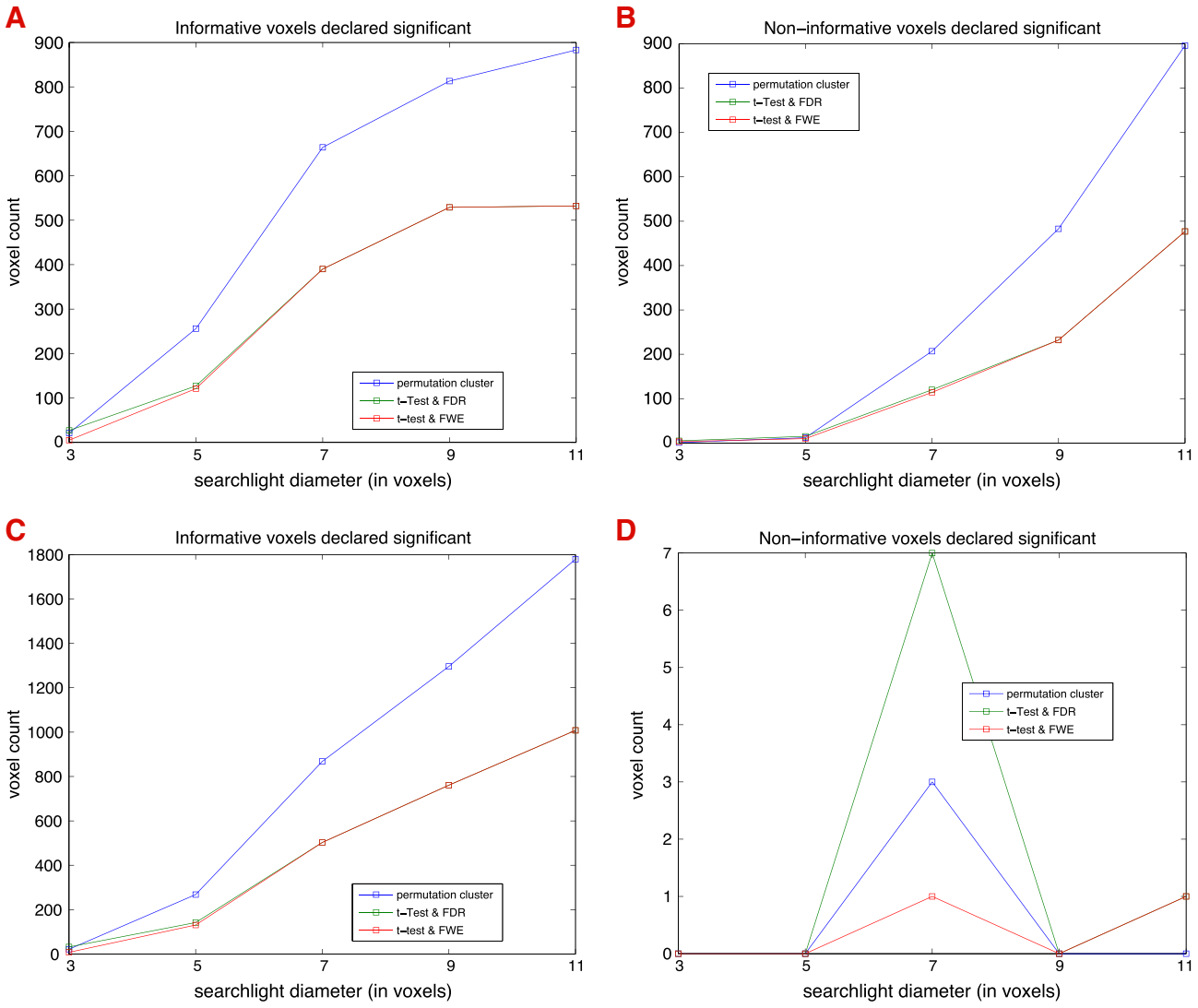


Fig. 6. Influence of searchlight diameter. We computed the number of voxels within and outside the informative regions for five different searchlight diameters. (A) Number of voxels inside the informative cubes declared significant. (B) Number of voxels outside the informative cubes declared significant. (C) Number of voxels inside the *extended* informative cubes. We redefined the informative areas by extending the informative cubes with the searchlight diameter. (D) Number of voxels inside the *extended* informative cubes.

were declared significant (12 voxels with 6-connectivity). Overall, both connectivity schemes showed strikingly similar results.

Simulated null data sets (validation)

We analyzed the simulations with a set of 16 equidistant cluster thresholds ranging from $p_{cluster} = 0.0$ to 0.15, and noted the actual observed number of clusters in the simulations for each threshold. As an initial voxel-threshold we set $p_{voxel} = 10^{-3}$. The number of observed clusters (for each threshold) can then be directly related to the expected number of clusters, which corresponds to the cluster threshold (e.g., for a false-positivity threshold on cluster level of $p_{cluster} = 0.05$, it would be expected to find five clusters in 100 simulated data sets or equivalently 50 clusters in 1000 simulated data sets). In Fig. 7A, we show the ratio between observed and expected clusters for 100 simulated data sets for our proposed framework. The number of observed clusters is well below the number of expected clusters for any given threshold. This indicates that our proposed method can be regarded as rather conservative. The standard *t*-test procedure results for 1000 simulations are displayed in Fig. 7B. For higher thresholds, the FWE approximates the dashed line very closely, indicating that the false-positivity is at the expected rate. However, when using FDR correction on the cluster level for the commonly practiced *t*-test frameworks, the observed clusters in the data exceed what is expected in a rigorously type I error controlled scenario by a factor more than three.

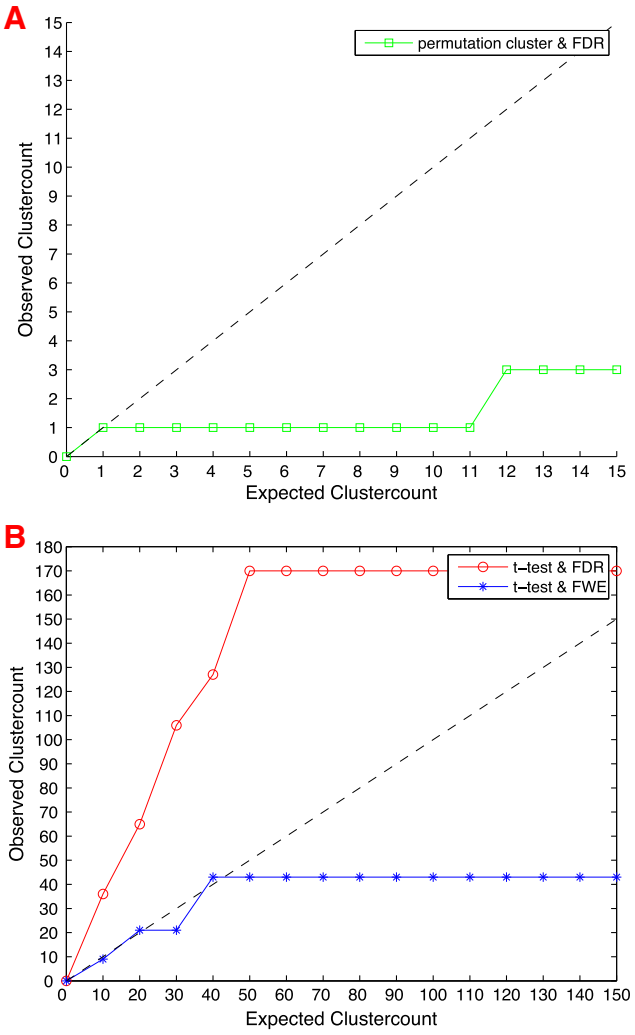


Fig. 7. False positivity diagrams for classification of simulated data sets containing no information. The observed cluster count is the number of clusters found in all simulations for a given false-positivity level. The expected cluster count is the false-positivity level multiplied by the total number of simulations. The dashed line displays an empirical false-positivity level corresponding exactly to the theoretical false-positivity. Data points found above the dashed line indicate a higher empirical false-positivity than expected by the cluster threshold. (A) False-positivity of our proposed method. In total, 100 simulations were considered here. (B) False-positivity of the *t*-test methods, controlled for FDR or FWE on cluster level. 1000 simulations were considered here.

Undersampling of the permutation space

Using three data sets with differing number of data points (80, 120, 160), which only had one location (i.e., one searchlight position), we were able to study the effects of undersampling of the permutation space. For the three data sets there are, respectively, 10^{118}, 10^{157} and 10^{198} permutations available, rendering an extensive search unfeasible. We applied 10^1, 10^2, 10^3 and 10^4 permutations on the single subject level, for each of which we created 10^5 group recombinations, resulting in four distributions (per data set). We plotted these distributions in Fig. 8. For each of the three data sets, the distributions seem to converge with only 100 permutations on single subject level.

fMRI data set

The chance distributions obtained by the permutation test procedure reveal a spatial inhomogeneity: The FWHM of the voxel chance distribution differs greatly across locations. Fig. 9 shows the accuracy for which the area of the (normalized) chance distribution is <0.1%, therefore depicting the accuracy level for $p < 10^{-3}$ (this map served as threshold map for the cluster search). Additionally, in Fig. 9, we display the null distributions for three distinct locations.

Fig. 10 shows the distribution of cluster sizes found in the 10^5 permuted group accuracy maps. The red dotted line marks the $p = 0.05$ level, where the right-tail area of the (normalized) cluster size distribution is smaller than 0.05. The corresponding cluster size is required to be bigger than 48 voxels in this data set for this level of significance.

We display the raw decoding accuracy in Fig. 11A. The cluster-size controlled *p*-value maps are displayed in Fig. 11B for the conventional *t*-test method implementing FWE correction on cluster level. Our proposed method is depicted in Fig. 11C.

Both our proposed method and *t*-test methods determine large parts of the occipital cortex to be significant. However, while *t*-based methods only reveal early visual areas (V1/V2) and V5, our method finds additional informative regions in secondary visual areas and the superior parietal lobule. Secondary visual areas are known to be implicated in discriminating directional from unidirectional information (Sato et al., 2009), and also when spatial attention is shifted (Prado and Weissman, 2011; Thakral and Slotnick, 2009). Areas in the superior parietal lobule have been shown to be implicated in spatial attentional selection (Krueger et al., 2007) and visual imagery of complex hand movements (Guillot et al., 2009). Hence, we conclude that the effects found in these additional regions are consistent with the related cognitive model and imply an increased statistical power of the proposed method.

When using *t*-test implementations with FWE correction (on cluster level, $p_{cluster} < 0.05$), 2548 voxels were labeled significant. Our proposed method labeled 3661 voxels significant. We found 2269 voxels labeled as significant by both methods, leaving 255 voxels identified when using *t*-test and FWE frameworks solely and 1393 voxels by our proposed method solely.

Discussion

We present a group analysis method tailored for decoding studies based on local MVPA (such as the searchlight approach). Our method incorporates non-parametric statistics and provides a solution for the multiple testing problem based on a cluster size thresholding. In the following sections we want to argue our concerns about *t*-based frameworks and discuss our proposed method.

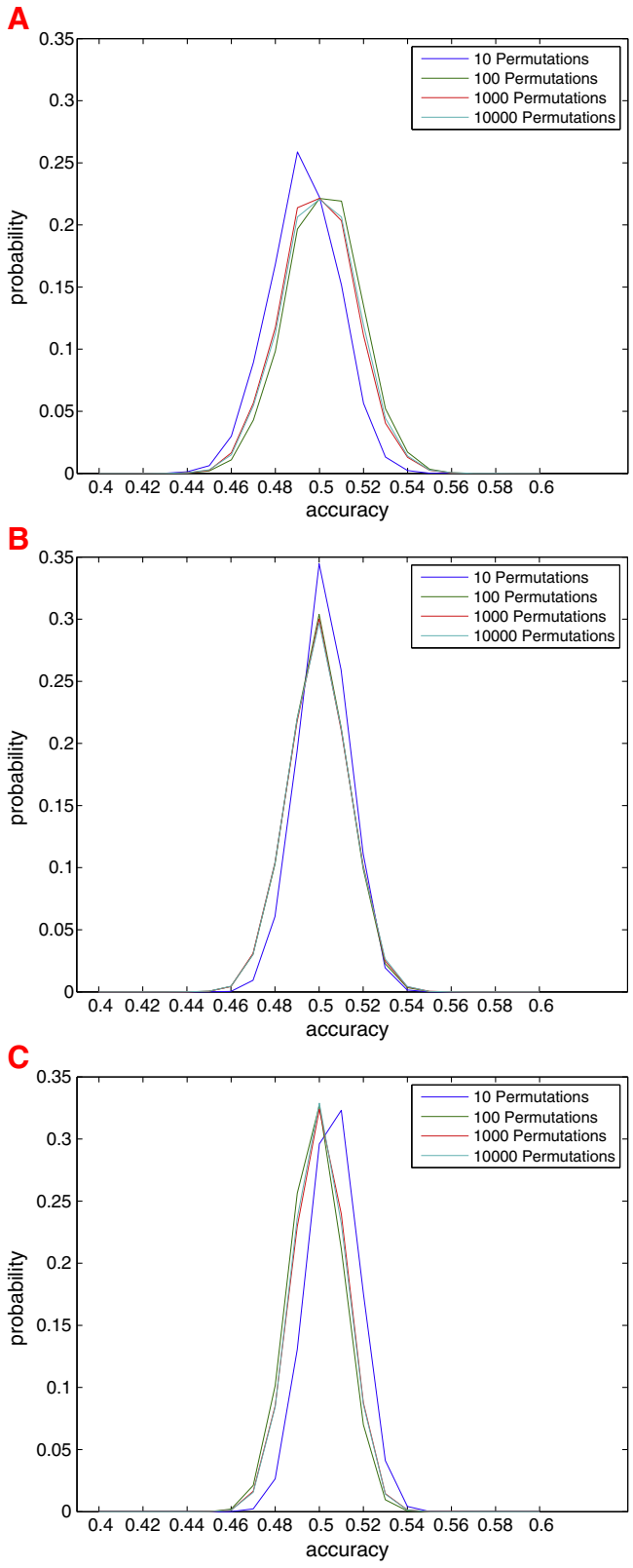


Fig. 8. Influence of undersampling of the permutation space. For three depicted data sets, we applied 10^1, 10^2, 10^3 and 10^4 permutations on the single subject level. We depict the histogram of accuracies after the recombination (i.e., the histogram of 10^5 accuracies). (A) Data set consisting of 80 observations (making 10^{118} permutations possible). (B) Data set consisting of 120 observations (making 10^{157} permutations possible). (C) Data set consisting of 160 observations (making 10^{198} permutations possible). In all cases, the distribution converges already with a small number (more than 100) permutations applied.

Pitfalls of t-based frameworks in decoding studies

The Student's *t*-test represents a commonly practiced method for determining the probability of a decoding result on the group level stands. Importantly, *t*-tests impose certain assumptions on the data. For instance, the samples need to be distributed normally, particularly if the sample size is small. Furthermore, the underlying distribution from which samples are drawn should be continuous.

Both requirements are problematic if a *t*-test is performed for a second-level group analysis of decoding accuracies. In general, decoding accuracies are not normally distributed, because the unknown distribution of decoding accuracies is generally skewed and long-tailed. In practice, the distribution depends heavily on the classifier used and the input data itself. Moreover, the samples are not drawn from a continuous distribution: The indicator function, which maps the number of correctly predicted labels to an accuracy value between [0,1], can only take certain values: for k cross-validation steps and a test set of size t, only $k \times t + 1$ different values between [0,1] can be taken.

Furthermore, the high variance on the single subject accuracies depicts a problem for *t*-based frameworks. One of the most critical assumptions of any classification-based method is that the observations are drawn independently from the data set (Langford, 2006). This imposes a problem for the typical fMRI data set, given the severe temporal contamination of subsequent scans due to autocorrelation (Zarahn et al., 1997). Therefore, the prerequisite of independence has to be approximated by taking into account well-separated groups of scans or their statistical estimation parameters (e.g., from a general linear model on separate runs). Hence, the ultimate number of observations available for classification is greatly limited. This limitation of samples imposes severe challenges, as demonstrated in the recent work by Isaksson et al. (2008). Most noteworthy, an inverse relation between the number of samples and the variance of accuracies is found: The fewer samples used, the larger the variance in the estimated accuracies. The variance is also driven by the size of the test set, as shown in Wickenberg-Bolin et al. (2006): The smaller the size of the test set, the greater the variance in the estimated accuracy. It is important to emphasize that the variance of the estimation of accuracies must not be confused with the true underlying variance of the performance of the classifier (caused, for example, by inter-session and inter-subject variability and generally non-observable in a real data set). However, simulations that enable a measurement of true performance, clearly demonstrate that the variance is, in fact, dominated by the effects of a small sample size (Isaksson et al., 2008). Furthermore, the indicator function mapping the number of correctly predicted labels to an accuracy value between [0,1] is of a discrete nature, which additionally increases the variance for small data sets (Hefny and Atiya, 2010).

For these reasons, a statistical inference method for classification-based decoding in fMRI should not heavily rely on the variance of the decoding accuracies on a single-subject level. However, the commonly practiced *t*-tests on single subject accuracies fundamentally implement the variance, as the square root of the variance enters the denominator of the *t*-formula.

Independent Bernoulli trials as a parametric alternative

As a parametric alternative to the *t*-test, stands the derivation of a theoretical null distribution by modeling the classifier as Bernoulli trial (Pereira and Botvinick, 2011). This procedure appears straight-forward, because the null hypothesis was defined to be the absence of class information in the data. In other words, if there is no class information present, the classifier practically has to *guess* the labels of the test set. More precisely, in a situation where n labels are estimated by a single classifier, it is possible to compute a theoretical null distribution by assuming n independent Bernoulli trials. The number of correctly estimated labels of the n trials can thus be represented by the binomial

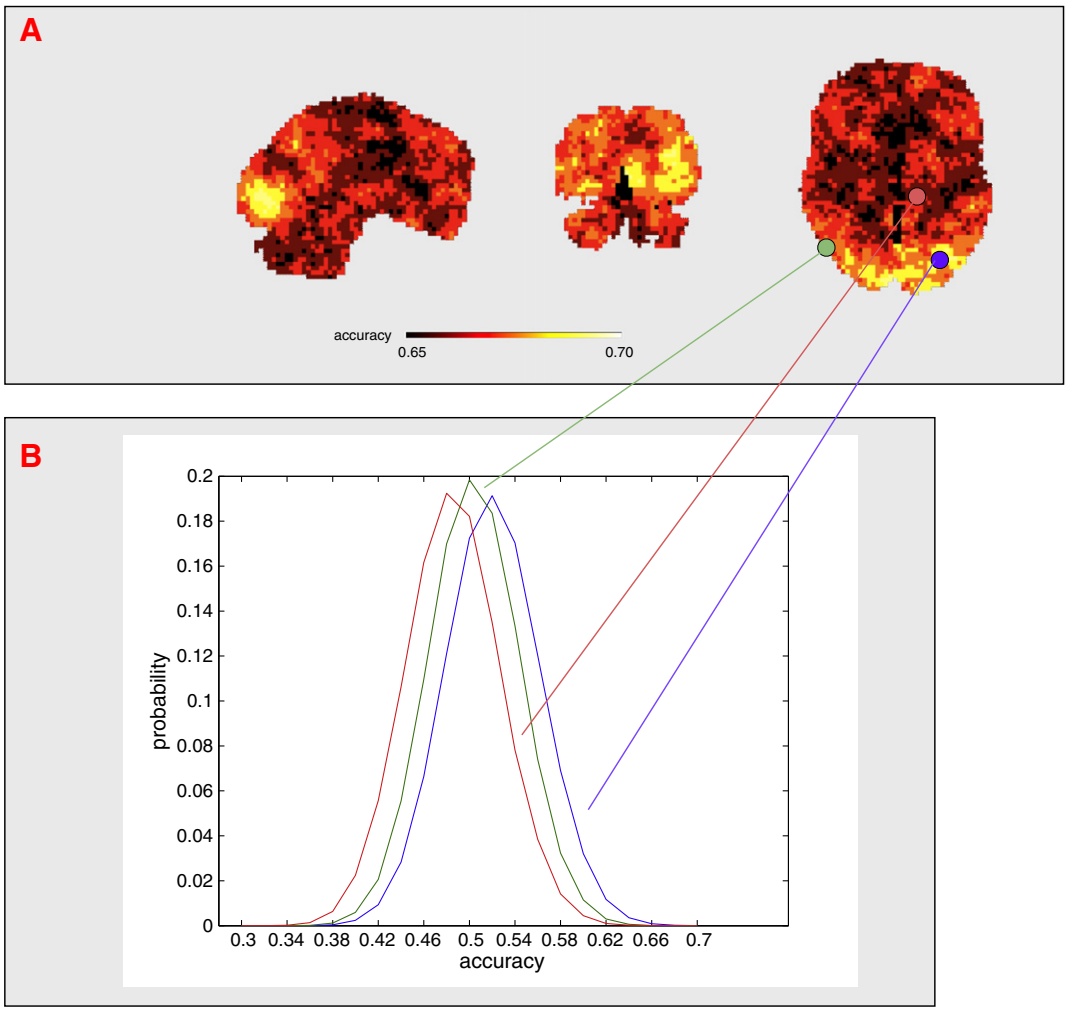


Fig. 9. Threshold map representing a p-value of 10^{-3} obtained using voxel-wise empirical chance distributions and three voxel-wise histograms of the group chance maps. (A) Inhomogeneous map which indicates a spatial inhomogeneity of the underlying chance distribution. (B) Three locations and their respective voxel-wise chance distribution. The chance distributions of these three locations differ considerably.

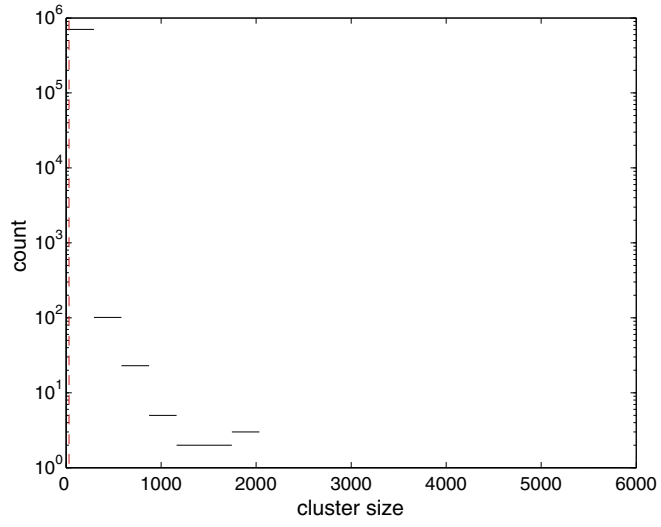


Fig. 10. Cluster size histogram from the fMRI data set. The red line marks the $p=0.05$ percentile of the cluster size distribution (cluster size = 48 voxels).

random variable X, which depends on the parameters (n, p). The probability mass function of X then is:

$$p_X(c) = \binom{n}{c} p^c (1-p)^{n-c} \quad c = 0, 1, \ldots, n$$

where c is the number of successful trials (i.e., correctly estimated labels) and p is usually set to 1/2 in a two-class paradigm.

If k leave-one-out cross-validations with the same number of test labels are applied, *each* cross-validation fold yields *one* binomial random variable, hence the whole cross-validation process yields k identically distributed binomial random variables $X_1, X_2, \ldots, X_k$, each with the parameters (n, p). As the accuracy is estimated on the total number of correctly classified samples over all the cross-validation folds, the sum of these k variables is computed, which under the assumption of independency, is again a binomial random variable with the parameters $(n \times k, p)$ (Ibe, 2005):

$$p_Z(c) = \binom{n \cdot k}{c} p^c (1-p)^{n \cdot k - c} \quad c = 0, 1, \ldots, n \cdot k.$$

Hence, the k cross-validations can be treated as one *single* classifier, which estimates $n \times k$ labels—under the assumption of independence of the respective binomial random variables.

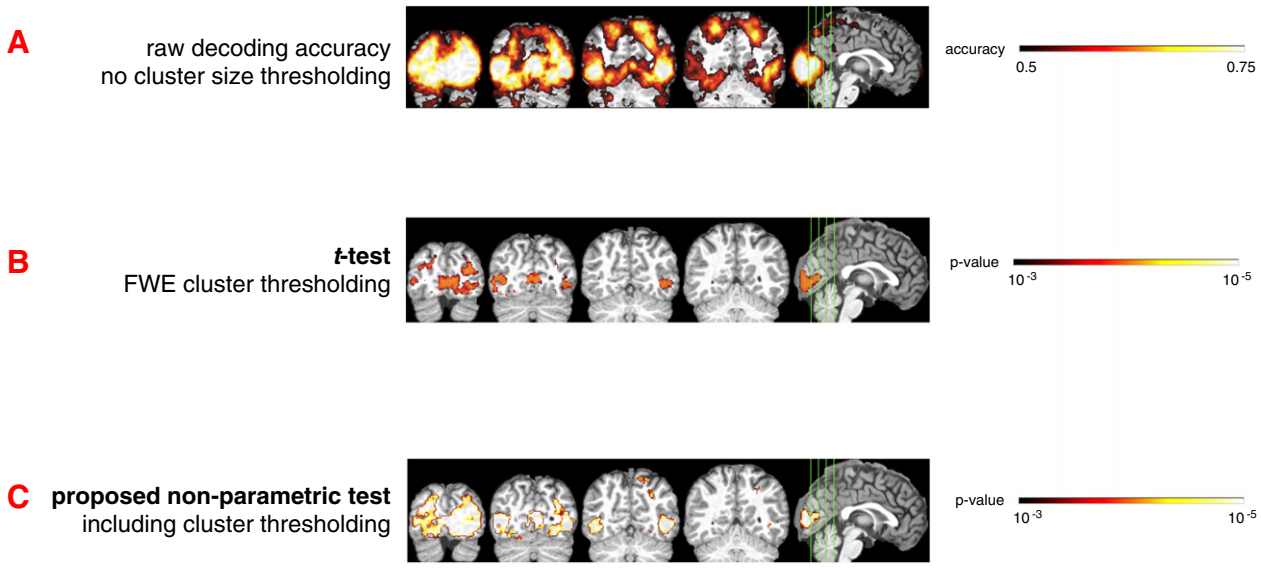


Fig. 11. Comparison of the classification results between our proposed method and standard correction methods. (A) Raw decoding accuracy map. (B) The results of a standard *t*-test, thresholded at cluster level using FWE methods implemented in SPM8. (C) Corresponding voxel-wise *p*-values of the classification result when our proposed cluster thresholding is applied.

The independence of test examples in each cross-validation fold, however, does not automatically assure independence of the binomial random variables $X_1, X_2, \ldots, X_k$. Most importantly, the correlations between training sets and testing sets in different folds cause the binomial random variables to be *correlated* to each other. This correlation violates the earlier assumption of independence, and makes the above used procedure for summing independent binomial variables formally incorrect.

To investigate this issue empirically, we simulated different scenarios (results and methods can be found in the supplementary materials), where we left the product $n \times k$ constant (k was the number of cross-validations, n the size of the test set). By varying the number of cross-validations k, we manipulated the degree of correlation and hence dependency between cross-validations (n was varied accordingly). As the product $n \times k$ was the same for all simulations, the simulations would be expected to return identical null distributions—if the assumption of independence was fulfilled. The point of the simulation is to show that the above approximation for summing binomial variables can only be applied if independency is given, and that the empirical deviation to the approximation indeed depends on the degree of correlation between the binomial variables.

The empirical results reveal that the deviation from the theoretically derived sum Z monotonically depends on the degree of correlation between the binomial random variables $X_1, X_2, \ldots, X_k$: the higher the correlation between cross-validation folds, the larger the deviation to the single classifier approximation. On the contrary, we showed that in case of no cross-validation and a true single classifier estimating of $n \times k$ labels, the binomial model fits exactly.

Moreover, the variance of the empirical null distributions depends on the degree of correlation between cross-validation folds; the higher the correlation, the broader the distribution. When applied in statistical inference, the smaller variance of the null distribution from the binomial model has an effect of overestimating the *p*-values. Effectively, smaller *p*-values will be reported from the binomial model, than from the empirical ones, as reflected in studies (e.g., Pereira and Botvinick, 2011). Therefore, adopting the above binomial model in case of correlated cross-validation principally increases the false positivity.

Non-parametric approaches in the context of decoding

Nonparametric tests, such as the permutation test, are exact regardless of whether the underlying distribution is normal or not, because permutation tests rely on minimal assumptions (Good, 2006), and most crucially exchangeability of the data points. Note that this assumption is imposed not only for permutation tests but also for classification methods in general (Langford, 2006). The theoretical applicability of permutation tests for classification-based methods has been demonstrated and is well established (Golland and Fischl, 2003).

A limitation for the usage in fMRI studies is that the number of permutations on single subject level is small. This is due to the low number of available observations (i.e., independent data points).

Our method circumvents the low number of available permutations on the single subject level by applying bootstrapping methods (Efron and Tibshirani, 1993) to construct a group statistics. This boost of the chance pool overcomes the limitation of the number of observations and allowed to establish 10^5 random group maps (thus the lowest *p*-value could be defined at $p = 10^{-5}$).

The proposed framework outlined in this paper suffers from the effects of small sample size to a much smaller extent than *t*-based methods. This is particularly the case because all statistical assessments are based on the *group* level, because only the means over *all* subjects are regarded (i.e., N sets of small samples are regarded in combination, given N subjects).

Our method aims to incorporate spatial correlations, regardless of whether they are already present in the raw data or introduced by the local MVPA method itself. Furthermore, we avoid potential biases due to uneven distribution of samples across classes (i.e., if one class dominates the training or test subset) and maintain the correlation between the cross-validations by leaving the permutation fixed for all cross-validation folds. This is achieved by applying a permutation to the order of the observations in the data. Each permutation is then held fixed for all locations and cross-validations (see Fig. 2).

Comparison between t-based and non-parametric approaches

When comparing our proposed framework to the commonly practiced *t*-based methods, we found our method to have a much higher statistical sensitivity. This was reflected by an increase of about 50% in terms of detection volume, i.e., to total volume labeled significant (when compared to *t*-test with cluster control using FWE). We demonstrated this high sensitivity both when our method was applied to simulated data and to real fMRI data. In the former case, we were able to explicitly deposit information at *known* locations and examine

to what extent the different approaches were able to decode information situated in these informative regions. In doing so we could demonstrate that the plus in detection volume of our method was indeed located in the informative regions. At the same time, the number of non-informative voxels labeled significant was smaller compared to the corresponding number of significant and non-informative voxels in *t*-based methods. The non-informative voxels had been adjacent to the informative regions, thus these voxels don't truly reflect false positives (because the searchlight approach acts as spatial filter). Nevertheless, this finding implies that our method has a higher spatial accuracy in detection of truly informative regions.

Furthermore, when using the null simulation for validation, we revealed that the common practice of *t*-test and subsequent FDR correction on a cluster level produces an unacceptably high rate of false positives. Given widely used parameters (initial voxel threshold $p_{voxel} = 10^{-3}$ and $p_{cluster} = 0.05$), the empirically observed number of clusters labeled significant surpasses the expected number of cluster of a factor greater than three, assuming a type I error control. Consequently, it seems possible that *t*-tests with cluster level control using FDR do represent a defective way to control for false-positivity. Because FDR procedures regulate the expected proportion of incorrectly rejected null hypothesis, they are commonly regarded to be less conservative than Bonferroni correction methods. In our null simulations we created a situation where the number of (uncorrected) significant clusters could become large while *none* of them actually reflect a true signal. Given this, FDR procedures applied here might be prone to false positivity, because *every* rejection of the null hypothesis inherently stands for a false positive result. In experimental environments, principally there is no direct access to the distribution of true signals. Therefore, we regard it remarkably problematic to apply cluster FDR procedures on *t*-test results in the context of decoding.

Threshold map procedure and general considerations

Our approach incorporates the spatial inhomogeneity in the null distributions in the cluster search algorithm. If the algorithm used a constant threshold level (e.g., one global accuracy level), the resulting local cluster sizes would be *overestimated* in the case of a broad null distribution (and *underestimated* if the local chance distribution were narrow). The resulting cluster size distribution would therefore be biased, depending on the spatial inhomogeneity and the choice of global threshold. Effectively, the threshold map procedure used in our method is equivalent to performing a permutation test on every voxel and applying a threshold (because only voxels exceeding a certain statistical threshold defined by the permutation distribution do surpass). Naturally, the width of the permutation distribution depends on several factors such as the actual information content; in the presence of information the width and possibly location of the permutation distribution is increased or shifted to the right and hence the threshold at this location is more conservative. Nevertheless, in the presence of information, the permutation test procedure does not become overly conservative, because the sensitivity of the permutation framework still appears to be superior compared to parametric methods. The same holds for the opposite case where no information is present. Here, the permutation distribution does not become more open to false-positives, as we demonstrated in our simulation for validation.

Both the commonly practiced *t*-test frameworks and our permutation method essentially incorporate both the initial voxel-threshold and the cluster-level threshold as free parameters. The former cannot be deduced and principally underlies a certain arbitrariness of the experimenter. On the one side of the spectrum, thresholds that are too high (i.e., low *p* values) drastically reduce the sensitivity; while on the other side thresholds that are too low (i.e., high *p*-values) hurt the localization of the truly informative regions. In terms of a worst-case scenario, lowering the initial voxel threshold might even lead to the merging of otherwise separated clusters. Therefore, as a compromise we recommend a *p* value between 0.005 and 0.001 as an initial voxel-threshold.

Advanced local MVPA applications

We implemented our approach in the simplest local MVPA application, the volumetric searchlight method. However, our proposal can easily be applied to any local MVPA application, which produces maps of decoding results. For example, our method can be easily implemented in surface-based techniques such as surface decoding (Chen et al., 2011).

Computation time

The computational cost of random permutation procedure increases linearly with the number of permutations, compared to that of the cross-validation on the same sample set. The independence of computation between different permutations admits a straightforward parallelization. With the real fMRI data set presented above, the computation time for the entire statistical evaluation was about seven hours on an 8-core Sempron 2.0 GHz server. Note, however, that when the number of samples grows large, both the complexity of the cross-validation and the number of permutations for a reasonable non-parametric distribution approximation increase supra-linearly. It appears that the number of permutations necessary on the single subject level is not very large, because the null distributions already converge with hundreds of permutations. Nevertheless, we are working on very fast C/C++ implementation that employs massive parallelism in the framework of our in-house fMRI analysis software LIPSIA (Lohmann et al., 2001), to address this problem. The software will be available for download on our website (http://www.cbs.mpg.de/institute/software/lipsia/download.html) and will reduce the computation time of the full procedure significantly; targeting to a time frame considerably smaller than an hour.

Conclusion

We demonstrate a procedure for second-level (group) analysis for local MVPA methods which, compared to other approaches, is based on less critical assumptions and features a higher sensitivity. Furthermore, we show that, fMRI decoding studies implementing a *t*-test-based second-level analysis discard results due to an inappropriate statistical procedure. Given our proposed method, the fraction of locations containing information being discarded is considerably smaller.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.neuroimage.2012.09.063.

References

Arndt, S., Cizadlo, T., Andreasen, N.C., Heckel, D., Gold, S., O'Leary, D.S., 1996. Tests for comparing images based on randomization and permutation methods. J. Cereb. Blood Flow Metab. 16, 1271–1279.

Benjamini, Y., 1999. A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. Journal of Statistical Planning and Inference http://dx.doi.org/10.1016/S0378-3758(99)00040-3 (Journal of Statistical Planning and Inference | ScienceDirect.com.).

Björnsdotter, M., Rylander, K., Wessberg, J., 2011. A Monte Carlo method for locally multivariate brain mapping. NeuroImage 56 (2), 508–516 http://dx.doi.org/10.1016/j.neuroimage.2010.07.044.

Bode, S., Haynes, J.D., 2009. Decoding sequential stages of task preparation in the human brain. NeuroImage 45, 606–613.

Carlin, J.D., Rowe, J.B., Kriegeskorte, N., Thompson, R., Calder, A.J., 2012. Direction-Sensitive Codes for Observed Head Turns in Human Superior Temporal Sulcus. Cereb. Cortex 22 (4), 735–744.

Chang, Chih-Chung, Lin, Chih-Jen, 2011. LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. 2, 27:1–27:27 (Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm).

Chen, Y., Namburi, P., Elliott, L.T., Heinzle, J., Soon, C.S., Chee, M.W., Haynes, J.D., 2011. Cortical surface-based searchlight decoding. NeuroImage 56, 582–592.

Chumbley, J.R., Friston, K.J., 2009. False discovery rate revisited: FDR and topological inference using Gaussian random fields. NeuroImage 44, 62–70.
Efron, A., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapmann & Hall.
Forman, S.D., Cohen, J.D., Fitzgerald, M., Eddy, W.F., Mintun, M.A., Noll, D.C., 1995. Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold. Magn. Reson. Med. 33, 636–647.
Golland, P., Fischl, B., 2003. Permutation tests for classification: towards statistical significance in image-based studies. Inf. Process. Med. Imaging 18, 330–341.
Good, P., 2006. Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses, 2nd edition.
Guillot, A., Collet, C., Nguyen, V.A., Malouin, F., Richards, C., Doyon, J., 2009. Brain activity during visual versus kinesthetic imagery: an fMRI study. Hum. Brain Mapp. 30 (7), 2157–2172 http://dx.doi.org/10.1002/hbm.20658.
Hanke, M., Halchenko, Y.O., Sederberg, P.B., Hanson, S.J., Haxby, J.V., Pollmann, S., 2009. PyMVPA: a Python Toolbox for Multivariate Pattern Analysis of fMRI data. Neuroinformatics 7 (1), 37–53 http://dx.doi.org/10.1007/s12021-008-9041-y.
Hayasaka, S., Nichols, T.E., 2003. Validating cluster size inference: random field and permutation methods. NeuroImage 20, 2343–2356.
Hefny, A., Atiya, A.F., 2010. A New Monte Carlo-Based Error Rate Estimator.
Heller, R., Stanley, D., Yekutieli, D., Rubin, N., Benjamini, Y., 2006. Cluster-based analysis of FMRI data. NeuroImage 33, 599–608.
Holmes, A.P., Blair, R.C., Watson, J.D., Ford, I., 1996. Nonparametric analysis of statistic images from functional mapping experiments. J. Cereb. Blood Flow Metab. 16, 7–22.
Ibe, O., 2005. Fundamentals of Applied Probability and Random Processes. Academic Press.
Isaksson, A., Wallman, M., Göransson, H., Gustafsson, M.G., 2008. Cross-validation and bootstrapping are unreliable in small sample classification. Pattern Recognit. Lett. 29 (14), 1960–1965.
Kahnt, T., Heinzle, J., Park, S.Q., Haynes, J.D., 2010. The neural code of reward anticipation in human orbitofrontal cortex. Proc. Natl. Acad. Sci. U. S. A. 107, 6010–6015.
Kriegeskorte, N., 2011. Pattern-information analysis: from stimulus decoding to computational-model testing. NeuroImage 56, 411–421.
Kriegeskorte, N., Goebel, R., Bandettini, P., 2006. Information-based functional brain mapping. Proc. Natl. Acad. Sci. U. S. A. 103, 3863–3868.
Krueger, F., Fischer, R., Heinecke, A., Hagendorf, H., 2007. An fMRI investigation into the neural mechanisms of spatial attentional selection in a location-based negative priming task. Brain Res. 1174, 110–119 http://dx.doi.org/10.1016/j.brainres.2007.08.016.
Langford, J., 2005. Tutorial on practical prediction theory for classification. J. Mach. Learn. Res. 6, 273–306.
Lohmann, G., Müller, K., Bosch, V., Mentzel, H., Hessler, S., Chen, L., Zysset, S., et al., 2001. LIPSIA—a new software system for the evaluation of functional magnetic resonance images of the human brain. Comput. Med. Imaging Graph. 25 (6), 449–457.
Nichols, T.E., Holmes, A.P., 2002. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum. Brain Mapp. 15, 1–25.
Pereira, F., Botvinick, M., 2011. Information mapping with pattern classifiers: a comparative study. NeuroImage 56 (2), 476–496.
Prado, J., Weissman, D.H., 2011. Spatial attention influences trial-by-trial relationships between response time and functional connectivity in the visual cortex. NeuroImage 54 (1), 465–473 http://dx.doi.org/10.1016/j.neuroimage.2010.08.038.
Sato, W., Kochiyama, T., Uono, S., Yoshikawa, S., 2009. Commonalities in the neural mechanisms underlying automatic attentional shifts by gaze, gestures, and symbols. NeuroImage 45 (3), 984–992 http://dx.doi.org/10.1016/j.neuroimage.2008.12.052.
Thakral, P.P., Slotnick, S.D., 2009. The role of parietal cortex during sustained visual spatial attention. Brain Res. 1302(C), 157–166 http://dx.doi.org/10.1016/j.brainres.2009.09.031.
Wickenberg-Bolin, U., Goransson, H., Fryknas, M., Gustafsson, M.G., Isaksson, A., 2006. Improved variance estimation of classification performance via reduction of bias caused by small sample size. BMC Bioinforma. 7, 127.
Zarahn, E., Aguirre, G.K., D'Esposito, M., 1997. Empirical analyses of BOLD fMRI statistics. I. Spatially unsmoothed data collected under null-hypothesis conditions. NeuroImage 5, 179–197.