

available at www.sciencedirect.comwww.elsevier.com/locate/brainres

**BRAIN
RESEARCH**

Research Report
An introduction to anatomical ROI-based fMRI classification analysis
Joset A. Etzel*, Valeria Gazzola, Christian Keysers

 BCN Neuroimaging Center, University of Groningen, Department of Neuroscience, University Medical Center Groningen, The Netherlands

ARTICLE INFO
Article history:

Accepted 23 May 2009

Available online 6 June 2009

Keywords:

fMRI

Region of interest

Multivariate analysis

Classification analysis

Multivoxel pattern analysis

ABSTRACT

Modern cognitive neuroscience often thinks at the interface between anatomy and function, hypothesizing that one structure is important for a task while another is not. A flexible and sensitive way to test such hypotheses is to evaluate the pattern of activity in the specific structures using multivariate classification techniques. These methods consider the activation patterns across groups of voxels, and so are consistent with current theories of how information is encoded in the brain: that the pattern of activity in brain areas is more important than the activity of single neurons or voxels. Classification techniques can identify many types of activation patterns, and patterns unique to each subject or shared across subjects. This paper is an introduction to applying classification methods to functional magnetic resonance imaging (fMRI) data, particularly for region of interest (ROI) based hypotheses. The first section describes the main steps required for such analyses while the second illustrates these steps using a simple example.

© 2009 Published by Elsevier B.V.

1. Introduction

Mapping the human brain is a deeply multidisciplinary enterprise. Over time, neuroanatomists have developed sophisticated ways of mapping the brain, moving from dividing the brain into lobes and sulci to mapping cytoarchitectonic and connectivity features. The development of neuroimaging methods such as functional magnetic resonance imaging (fMRI) has allowed neurophysiologists to measure from many thousands of locations in the brain while subjects perform tasks. Drawing from these sources of information, modern cognitive neuroscience, especially social neuroscience, often thinks at the interface between anatomy and function, such as by hypothesizing that one Brodmann Area (BA) is important for a task while another is not. The most intuitive strategy for testing these types of hypotheses is to focus on the relevant regions of interest (ROIs).

Most methods for fMRI analysis do not focus on ROIs but rather explore the entire brain. These methods range from mass-univariate approaches which test each voxel separately (e.g. statistical parametric mapping (SPM); Friston et al., 1994), through “searchlight” approaches which multivariately examine the information in small groups of voxels centered on each voxel in the brain (e.g. Kriegeskorte et al., 2006; Nandy and Cordes, 2004), to classification methods that analyze the brain as a whole (e.g. Hanson and Halchenko, 2008; Mourao-Miranda et al., 2005). These methods can provide information about ROIs, such as by examining where clusters of activated voxels occur. This can lead to ambiguous interpretation however, because of problems such as evaluating activations which overlap several ROIs or only a small part of a ROI. ROI analysis is often carried out by summarizing the activity of a ROI into a single value (e.g. mean), the timecourse of which is analyzed (Friston et al., 2006; Saxe et al., 2006). This method

* Corresponding author. BCN Neuroimaging Center, University Medical Center Groningen, Antonius Deusinglaan 2, Room 125, 9713AW Groningen, The Netherlands. Fax: +31 50 363 8875.

E-mail address: j.a.etzel@med.umcg.nl (J.A. Etzel).

unambiguously focuses on the ROIs, but at the expense of discarding information.

Most neuroscientists suspect that information in the brain is stored in the *patterns* of activity across groups of neurons — patterns that are not considered by some of these analysis methods. Multivariate machine learning methods (also called “classification analysis” or “multivoxel pattern analysis”) can detect patterns across voxels in fMRI data (Haynes and Rees, 2006; Norman et al., 2006; O’Toole et al., 2007), and so may provide more detailed information about the activity in each ROI. Classification analysis frames hypotheses in terms of separating pairs (or more) of conditions. For example, I could predict that a region’s activity will be higher when listening to speech by an angry than a relaxed speaker. Restated, I hypothesize that classification of speaker emotion will be significantly above chance in the region. Note that the restated hypothesis is that a different *pattern of activity* exists in the voxels making up the region, not that the “activation level” is different. This enables us to move away from interpreting BOLD patterns in terms of “activated” voxels, a term which implies that these neurons are spiking more than others, which may or may not be correct (Devor et al., 2008; Logothetis, 2008), to take a more abstract view: we can conclude that consistent patterns arise in a particular ROI, without basing such conclusions on the precise neuronal mechanism creating the patterns.

This ROI-based approach has been used for a wide range of hypotheses in many fields, using both multivariate classification and mass-univariate techniques. For example, to test simulation (common coding) theories multivariate pattern classifiers were trained during action perception and tested during action execution (Dinstein et al., 2008; Etzel et al., 2008). Significant cross-modal classification indicates that a ROI contains information about which action was executed or perceived, and that this information is represented in similar patterns. In other areas, classification and correlation analyses have been used with functionally-defined ROIs to test the processing of object and location visual information (Schwarzlose et al., 2008), and to identify the speaker and content of heard speech (Formisano et al., 2008).

This paper describes a method for ROI-based classification analysis, and assumes familiarity with fMRI research, but not necessarily machine learning techniques. The first section describes the analysis technique, followed by an example analysis and a discussion of analytical issues. We assert that classification methods are suitable for addressing anatomical ROI-based hypotheses, but do not mean to imply that it is the only appropriate method, nor that ROI-based analysis is always best. Rather, it is a powerful tool that should be considered, while simultaneously recognizing how it fits into the wide array of available techniques (see also Kriegeskorte and Bandettini, 2007). Indeed, some research questions may require a combination of methods to be fully addressed.

2. Analysis technique

This section describes a method for ROI-based classification analysis. For additional details and background many excellent introductions to fMRI experimental design and analysis

are available (e.g. Huettel et al., 2004; Strother, 2006), as are reviews of multivoxel pattern analysis (Norman et al., 2006; O’Toole et al., 2007). Other introductions to classification analysis are also available (e.g. Mitchell et al., 2004; Mur et al., 2009; Pereira et al., 2009). fMRI analysis is viewed here as a classification problem: “can the activation pattern of the voxels in this ROI show whether the subject was performing task *a* or *b*?” The analysis produces a measure summarizing the activity in each ROI: the accuracy. If a ROI classifies with chance accuracy (i.e. 0.50 for two classes) we have no evidence that it discriminates between the stimuli categories.

2.1. Experiment design

Although many experiments produce data which can be analyzed with multiple techniques, ROI-based classification analysis is especially suited for experiments in which the subject was in discrete states and the hypotheses involve localized differences between these states. For example, we could ask, “does the activity in BA44 classify whether the subjects were viewing hand or foot action movies?” or “does the activity in the primary somatosensory cortex classify whether the subjects believed they were being touched by someone attractive or unattractive?” Experiments which describe functional or anatomical relationships between regions, involve extended time periods, or without anatomically localized hypotheses may find different analysis techniques more useful, such as functional connectivity analysis (Buchel and Friston, 1997; Friston et al., 1993), Granger causality mapping (Goebel et al., 2003; Roebroeck et al., 2005), or mass-univariate analyses (Friston et al., 1994).

If classification analysis is chosen during experimental design steps can be taken to optimize the data. First, the classifications should be planned, and conditions which will be classified randomly presented within the same run. Otherwise there is a risk of confounding with time, so that it will be impossible to tell whether the classification is capturing brain differences between the conditions or based on slight differences between runs (e.g. from scanner drift). This time confound could make it difficult to conduct a permutation significance test (see below).

Second, the experiment should be structured so that enough examples (repetitions) of each condition will be collected, and that the examples can be considered independent. As many examples of each condition as possible should be collected to improve classifier performance. But in practice, the number of stimulus presentations will usually be strictly limited by other considerations, such as to restrict experiment length or avoid habituation. In our experience at least ten examples of each condition are needed, although classification may sometimes be possible with fewer. The number of examples obtained from an experiment is also dependent on the temporal compression method (see below), which should also be considered during design.

Additionally, the volume acquisition timing should be considered. Interleaved stimulus presentation timing (i.e. jittering the stimulus timing relative to the volume acquisition) is recommended to improve temporal resolution in mass-univariate analysis (Huettel et al., 2004). Classification analyses may benefit from time-locking the volume acquisition to

the stimulus presentation timing however, depending on the method of temporal compression (interleaved stimulus presentation may present few problems if a model-fitting method is used). Event-related designs should be considered, as they can allow more unique examples of each condition. Regardless, sufficient time between examples must be included for the hemodynamic response to return to baseline, or at least pass the refractory period (6 s minimum; [Huettel et al., 2004](#)).

Finally, it can be helpful to include positive (classifications that should be possible) and negative (classifications which should be impossible) controls in the experiment. These can strengthen the interpretation that classifications are based on interesting activity patterns, rather than scanner drift, subject movement, or some other confound.

2.2. Data preprocessing

Many of the preprocessing steps required for mass-univariate analysis are also required for ROI-based classification analysis. fMRI volumes generally need to be realigned (to correct for subject movement) and slice-time corrected (depending on scan parameters). Normalization may also be needed, for ROI definition or to allow between-subjects classification (see below). Spatial smoothing is sometimes useful, though risks blurring anatomical boundaries ([Brett et al., 2002](#)), which may be of particular concern for small ROIs. Spatial smoothing may be of particular benefit for between-subjects analysis, by reducing differences between subjects' activity patterns.

Temporal compression is often useful for classification analyses. It combines several volumes collected near in time to make a single summary volume. This step reduces the time confound that occurs because of the BOLD response's delayed onset (volumes collected at different times after stimulus onset will contain a different amount of activity), and often improves signal-to-noise and classification accuracy ([Mourao-Miranda et al., 2006](#)). Also, reducing time dependence makes it possible to consider the examples interchangeable and independent, which enables construction of training and testing sets and allows determination of significance by permutation testing (see discussion below).

The best temporal compression method depends on the experiment design and hypotheses. In general, it may be best to create one summary volume per block in a block design, and one summary volume per event in an event-related design. Temporal compression is often performed by averaging volumes falling within a fixed time window (e.g. [Cox and Savoy, 2003](#); [Kamitani and Tong, 2006](#); [Shinkareva et al., 2008](#)), or by fitting a GLM with separate parameter estimates for each block (often called "parameter estimate images", e.g. [Davatzikos et al., 2005](#); [Etzel et al., 2008](#); [Soon et al., 2008](#)). Classification has also been performed on single volumes, without temporal compression (e.g. [LaConte et al., 2005](#); [Mitchell et al., 2004](#)).

Many data sets will require voxel-wise detrending to remove low-frequency trends introduced by unavoidable factors such as scanner drift and physiological processes. The need for detrending also depends on the data processing pipeline; for example the models used to create parameter estimate images remove many types of trends. Various techniques exist for detrending, such as mean-subtraction, linear regression, and filtering ([Macey et al., 2004](#); [Strother,](#)

[2006](#); [Tanabe et al., 2002](#)). Additional preprocessing may be needed in some cases, such as denoising ([Ku et al., 2008](#); [Thomas et al., 2002](#); [Tohka et al., 2008](#)), or to derive a measure for analysis (instead of analyzing summary volumes). For example, the signal in a baseline condition could be subtracted from the active conditions (e.g. [Mitchell et al., 2004](#)).

2.3. Regions of interest (ROIs)

The ROIs should be selected during experimental design, based on the experimental hypotheses. We focus on anatomical ROIs, but functionally-defined ROIs ([Saxe et al., 2006](#)) could also be used. In this technique a functional localizer task is added during scanning (e.g. viewing faces), and the voxels involved in this task are found with a mass-univariate analysis. These voxels are then defined as a ROI, with a unique ROI for each subject. It is important for the validity of the analysis that the method for choosing the ROIs be determined during design, independently of the classification itself. Otherwise there is a risk of biasing the results, such as by choosing small ROIs on the results of a whole-brain analysis using the same classification.

When anatomical ROIs have visible anatomical landmarks maximum accuracy may be obtained by drawing the ROIs for each subject without normalization ([Brett et al., 2002](#); [Nieto-Castanon et al., 2003](#); [Saxe et al., 2006](#)). When individual delineation is not possible the use of maximum probability maps calculated from probabilistic cytoarchitectonic maps, such as those in the SPM Anatomy Toolbox ([Eickhoff et al., 2005](#)) may be the most anatomically accurate method available at this time ([Devlin and Poldrack, 2007](#); [Eickhoff et al., 2006](#); [Poldrack, 2007](#)).

In our opinion it is generally preferable to perform classification using all voxels in each ROI, as this does not restrict the classifiers to specific spatial patterns (e.g. clusters of a particular radius). Also, including all voxels allows conclusions to be drawn about each region as a whole, such as "significant classification was possible in the right, but not the left, secondary somatosensory cortex." However, including all voxels may make analysis difficult since anatomical structures vary in size. ROI size is of concern because it has been observed in machine learning that classifier performance may become poor if there are many more dimensions than examples (i.e. more voxels than volumes, related to the "curse of dimensionality," e.g. [Hastie et al., 2001](#); [O'Toole et al., 2007](#)).

The maximum number of voxels which should be in any ROI varies with classifier, number of examples, and difference in signal between the examples, as well as available computational power. We are not aware of any theoretically-derived guidelines for fMRI classification analysis specifically; in our experience 500 voxels is a reasonable upper limit if only around 10 examples of each class are available per subject; much larger ROIs sometimes classify well. [Ku et al. \(2008\)](#) and [Cox and Savoy \(2003\)](#) investigated how classification accuracy varied with number of voxels, and found (using linear support vector machines) that accuracy increased as more voxels were included up to a certain number of voxels (100 for [Cox and Savoy \(2003\)](#), 400 for [Ku et al. \(2008\)](#)), then leveled off. If this finding is generalizable it suggests that, at least within a

reasonable range, classification accuracy may not suffer as additional voxels are added, implying that analyzing ROIs of a range of sizes is justifiable. Practical limits to ROI size exist, however: a ROI can contain so many voxels that computer memory and run time become limiting.

If a ROI contains too many voxels, there are several options for reducing the number of voxels it contains. The voxels could be made larger by rescaling, which reduces the number of voxels, but risks blurring small-scale activation patterns. When possible, the voxel size should be kept similar to that during scanning to maximize spatial resolution. Conversely, the ROI could be made smaller, such as by using more stringent anatomical criteria; if the SPM Anatomy Toolbox maximum probability maps were used a higher likelihood threshold could be assigned. Also, the number of voxels in each ROI could be reduced by selecting a subset of voxels from each ROI (“feature selection”). Many feature selection methods have been developed, aimed at identifying the “best” voxels with the goal of maximizing accuracy (e.g. De Martino et al., 2008; Ku et al., 2008; Mitchell et al., 2004; Pessoa and Padmala, 2007; Shinkarova et al., 2008). Alternatively a dimension reduction method could be applied to the voxels in each ROI, such as singular value decomposition (Brett et al., 2002; Ku et al., 2008; Mourao-Miranda et al., 2006; O’Toole et al., 2007). Care must be taken with these methods to ensure that the analysis will be interpretable (described further in the Discussion).

2.4. Classification

Classification analyses can be performed within- and between-subjects, the choice of which relies upon assumptions about the activity patterns. Within-subjects analyses allow the activity pattern in each ROI to each type of stimuli to be different for each subject, while classification in between-subjects analysis is only possible if the activity patterns are similar in all subjects.

Multivariate analyses have been performed on fMRI data using different classifiers (reviewed in Ku et al., 2008; Mitchell et al., 2004; O’Toole et al., 2007), and there is not clear guidance for which classifiers are most suitable for particular analyses. Linear support vector machines (SVMs) have been successfully used with many fMRI data sets (e.g. Cox and Savoy, 2003; Ku et al., 2008; Mitchell et al., 2004), and are good classifiers to start with, if there is no reason to use a particular classifier. A mathematical description of SVMs is beyond the scope of this paper; the interested reader is referred to the machine learning literature for more details (e.g. Breiman, 2001a; Burges, 1998; Joachims, 1999; Schölkopf and Smola, 2002; Vapnik, 2000; Witten and Frank, 2005). Other techniques used with fMRI data include Fisher’s Linear Discriminant and Bayesian methods (e.g. Carlson et al., 2003; Cox and Savoy, 2003; Friston et al., 2008; Hampton and O’Doherty, 2007; Mitchell et al., 2004). Several different classifiers may be tried on a data set to see if certain families of classifiers work better.

The choice of classifier may be influenced by the number of classes to be separated. Many classification algorithms, including SVMs, are designed to separate two categories (“binary classification”), although some can handle more (e.g. Random Forests; Breiman, 2001b). Techniques exist to

allow classification of more classes with binary classifiers, often by combining the results of multiple classifiers, each of which separate two classes (Hsu and Lin, 2002).

Regardless of the classifier used, classification is performed by splitting the data into testing and training sets. A classifier is trained using the training set data, and its accuracy when classifying the data in the testing set is the classification accuracy (Breiman, 2001a; Witten and Frank, 2005). Within-subjects classification is done by training classifiers for each individual subject: the classifier is trained on a portion of the subject’s data and tested on the remaining samples. The accuracy is then averaged across subjects for the classification accuracy of the ROI, which can be considered analogous to performing a second-level mass-univariate analysis. Between-subjects classification is performed by training a classifier on the data from all but a few subjects, then testing it on the omitted subjects’ data.

It is usually possible to divide (“partition”) the data into training and testing sets multiple ways. For instance, suppose we are performing a within-subjects analysis and for each subject and ROI there are 15 examples of each of the stimuli. We could make a training set with three examples of each stimulus type as the testing set, using the remaining 12 as the training data. There are multiple ways to choose the three examples to use in the testing set (${}_{15}C_3$ of each stimulus type). The usual solution is to perform a “stratified-fold cross-validation” by repeatedly splitting the data into training and testing sets (ensuring that each class is equally represented in each) and averaging the results. For this example, we would perform a five-fold cross-validation by classifying five times, each time randomly selecting three examples of each stimuli type to be in the test set, so that each example was in a test set once. These five accuracies are then averaged. As there is more than one way to create the test sets the entire procedure should be repeated (usually ten times; Witten and Frank, 2005), so that there are different test sets in each repetition. Another common strategy is to split the data by run, with one run serving as the test set and the others the training.

The testing and training sets need to be as independent as possible to avoid inflating accuracy rates. For example, if single volumes are classified, volumes from within the same block will be very similar, and so should not be included in both the training and testing sets. This problem is minimized if temporal compression has been performed so that there is time for the hemodynamic response to diminish between summary volumes. The number of examples to include in the testing and training sets is not easy to determine. It is generally desired to maximize the training set size, but if the testing set is too small the variance of the accuracies may be high, reducing significance (Golland and Fischl, 2003; Mukherjee et al., 2003). If only a small number of examples are available (e.g. 10), using one example of each class as the testing set may be the best option, but as the number of examples increases, so should the size of the testing set.

For between-subjects analyses the training and testing data sets are often constructed by splitting the data by subject: the data for one subject is used for testing and data from the remaining subjects for training. In this case there is only one

way of partitioning the data so cross-validation is not required. If other methods are used, such as using multiple subjects in the test set or pooling data from multiple subjects, appropriate cross-validation procedures must be employed.

2.5. Significance testing

After calculating the accuracy of each ROI for each classification it is necessary to determine if the accuracies are significantly above chance. Both parametric (e.g. t-tests) and nonparametric (permutation testing) methods have been used successfully; only permutation methods will be described here. A statistical treatment of these methods is beyond the scope of this paper; for background on parametric testing see introductory statistical texts (e.g. Greer and Mulhern, 2002; Petrie and Sabin, 2000); for an introduction to permutation testing see (Edgington, 1995; Good, 2001); while for a statistical discussion of permutation methods for classification analyses see (Golland and Fischl, 2003; Mukherjee et al., 2003). Regardless of the significance testing method multiple comparisons corrections are needed when multiple ROIs, classifiers, or classifications are included.

A permutation test is based upon determining the range of classification accuracies when there is no relationship between the voxel activity patterns and the stimulus labels, which is ensured by randomly permuting the labels. If there is no relationship between the activity pattern in the data and the labels classification should not be possible. Not every classifier will perform exactly at chance, however: some will be slightly better or worse. This range can be used to calculate the significance of the true classification accuracy. This test has the null hypothesis that there is no difference between the stimulus labels; that the “labeling is arbitrary” (Nichols and Holmes, 2001). Restated, the test evaluates if the classification accuracy is due to chance: what is the likelihood that random-labeled datasets will be classified as accurately as the real dataset?

For within-subjects analysis the test requires constructing permuted-label data files by randomizing the stimulus type labels within each subject, using the same randomization for all subjects. The analogous procedure is followed for a between-subjects test. The classification accuracy is then determined for each permuted-label data set in the same manner as the actual data sets. The significance level is obtained from permutation tests by comparing the accuracy of the actual and permuted data sets: if the accuracy of the actual labeling is greater than 95% of the permuted-label accuracies in a single permutation test, the difference is significant at the 0.05 level.

The validity of a permutation test is based on random assignment (Edgington, 1995; Good, 2001); in this case, condition label exchangeability. fMRI data may violate label exchangeability however, due to temporal autocorrelation (Nichols and Holmes, 2001) or factors which result in stimulus-linked confounds (e.g. if subjects moved more in one type of trial), so the validity of exchangeability must be considered for each experiment. In general, label exchangeability can be justified if random stimulus presentation ordering within each run and temporal compression are used.

2.6. Localization test

The localization test evaluates how unusual it is to have a group of voxels of a particular size with a particular accuracy in the volumes. This test partially guards against the risk that the volumes varied systematically, enabling classification to be based on this difference rather than true activation patterns. The localization test is performed by comparing the classification accuracy of each ROI with the accuracy of size-matched subsets of ipsilateral *other* voxels, which are from brain structures which should not be able to classify the stimuli. Ideally, *other* areas would be anatomical structures with the same physical dimensions as each ROI, but with activity not predicted to vary with the experimental tasks. This is not possible, given the wide variation in size and shape of brain structures, so in practice the strategy is to identify the largest possible group of structures *a priori*. It is likely that some voxels with true classification information will be found in these structures, even though none were expected. This works against the experimenter in that it will make it harder to reject the null hypothesis that a ROI is better than *other*. The presence of these voxels does not reduce the validity of the test if the ROI's accuracy is higher, but rather reinforces that it is unusual to have a group of voxels with the observed accuracy. Note that it is essential that *other* areas are identified *a priori*; choosing ROIs with low accuracy after analysis introduces bias.

The localization test is performed by calculating the classification accuracy of many random subsets of *other*, each subset matched in size to an ipsilateral ROI, and then calculating the proportion of subsets classifying more accurately than the ROI. This is similar to the permutation test procedure, except that the accuracy is calculated for *other* subsets, rather than label rearrangements. It follows that the localization test can be seen as resulting in a *p*-value for the null hypothesis that the group of voxels making up the ROI classifies no better than equal-sized groups of voxels from *other* (*other* is treated as the null set). There will generally be too many possibilities to compute all possible subsets, but sufficient power should be possible by using at least 1000 random subsets (Edgington, 1995; Nichols and Holmes, 2001).

It will not always be possible to identify *other* areas, particularly when the experimental tasks are complex. For example, if the stimuli involve emotional videos to which the subject makes a motor response, identification of areas which should not be involved may be impossible. Extra care must be taken in these cases to ensure that biases are not present in the data which could drive the classification. Viewing and analyzing the data in multiple ways (e.g. with mass-univariate analysis) is always important to insure data quality, but especially important when the localization test is not possible.

3. Example fMRI data analysis

This example analysis will illustrate how these strategies and steps are used in practice. To keep the focus on the analysis technique rather than the experimental question a “toy” example of little direct interest will be used. In practice more ROIs and classifications will be included, but the analysis steps

will largely be the same. A small portion of the data collected during a study (Etzel, Valchev, et al., in preparation) will be described and analyzed here. This experiment was performed to test specific hypotheses, not the example which is presented here. However, to make this example easier to follow the design and analysis will be described as if the experiment was performed for this purpose.

3.1. Hypothesis and design

Suppose that we want to test the hypothesis that the auditory cortex has different activity patterns when listening to different sounds. Suppose further that rather than testing the difference in activity to unique sounds we are interested in the responses to two types of sounds: hand action sounds (e.g. the sound of ripping paper) and phase-scrambled versions of these sounds (“controls”), which are unrecognizable but have the same global frequency composition as the hand action sounds. Since we are interested in auditory cortex activity to these two classes of stimuli we need to play the subjects multiple examples of each type. We think that the pattern of auditory cortex activity may be different in every subject, so we will perform a within-subjects test.

A block or event-related design could be used for this experiment. In an event-related design we each sound presentation is separate, allowing classification between sounds in the same category (e.g. does the response to these two hand action sounds differ?). Since our hypothesis involves the difference between sound classes we can use a block design, in which we play multiple examples of the same sound type in a short time, perhaps increasing the signal.

The sound stimuli and scanning parameters used here were the same as those in Gazzola et al. (2006), which gives additional detail. In brief, five different sounds of each type (“hand action” or “controls”) were used, presented in blocks of three stimuli of the same type. The stimuli were randomly presented in a sparse block design with three runs, with multiple blocks of each type in each run. To ensure that the sounds were clearly audible sparse sampling was used: sounds were presented in the 4.1 second silent interval between the acquisition of volumes (acquisition time 1.5 s), using a T2* weighted acquisition at 3 T (TE=30 ms, TR=5.6 s, TA=1.5 s, 25 axial slices, 4.5 mm thick, 3.5×3.5 mm in plane resolution). 24 blocks of each type were presented. To ensure that subjects were listening closely to the sounds the subjects performed an odd-ball detection task (the odd-ball was a mixed block, such as a hand action sound followed by two control sounds), and indicated the odd-ball by pushing a button. Blocks containing an odd-ball, button push, or timing

error were not analyzed, so the number of usable blocks per subject ranged from 14 to 24.

Sixteen healthy males ranging in age from 19 to 33 years (mean=22.6, sd=3.9) participated in the experiment, which was approved by the Medisch Ethische Toetsingscommissie of the University Medical Center Groningen. One subject was excluded from analysis due to excessive movement during scanning.

3.2. ROIs

For this example we hypothesized that the auditory cortex has a different activity pattern when listening to hand action sounds and control sounds. This gives us one ROI: the auditory cortex, separately in each hemisphere. We can include a second ROI, the amygdala, as a control. We do not expect that the amygdala will be able to classify the sounds, as they don't differ emotionally. Since the amygdala is approximately the same size as the auditory cortex it can provide protection against spurious results. Finally, we can choose an *other* area to serve as the basis for the localization test (described previously). The primary visual cortex is not expected to be able to classify the sounds, since no characteristic visual stimuli accompanied them.

After choosing the ROIs a method to mark which voxels in each subject fall in each area (“creation of a ROI mask”) must be selected. Since we want to perform classification within-subjects we need to identify the same voxels in each subject, so we need to normalize and use a single ROI mask. We used the maximum probability maps based on the probabilistic cytoarchitectonic maps from the SPM Anatomy Toolbox (Eickhoff et al., 2005, 2006) to create the masks; the Toolbox areas used for each ROI are listed in Table 1.

3.3. Data preprocessing

In fMRI analyses the raw volumes need to be preprocessed to prepare them for analysis. In this case we want to realign the volumes (for motion correction) and normalize them (since the probabilistic maps were used to create ROI masks), which was done with SPM2 (Wellcome Department of Imaging Neuroscience, London, UK). During normalization the voxels were resized to 4×4×4 mm, close to the resolution at which the volumes were acquired. No spatial smoothing was applied. These steps create an image for each volume collected during the experiment. Further analysis was performed in R (R Development Core Team, 2008), so the images were converted to text using the AnalyzeFMRI R package (<http://www.cran.r-project.org/>).

Table 1 – Number of voxels (4×4×4 mm) and the SPM Anatomy Toolbox areas used to create each ROI.

ROI	SPM Anatomy Toolbox areas	Abbreviation	Number of voxels	
			Left	Right
Auditory	TE 1.0, TE 1.1, TE 1.2	aud	56	65
Amygdala	Amyg. (CM), Amyg. (LB), Amyg. (SF)	amyg	64	56
Other	BA17, BA18, hOC5	other	565	430

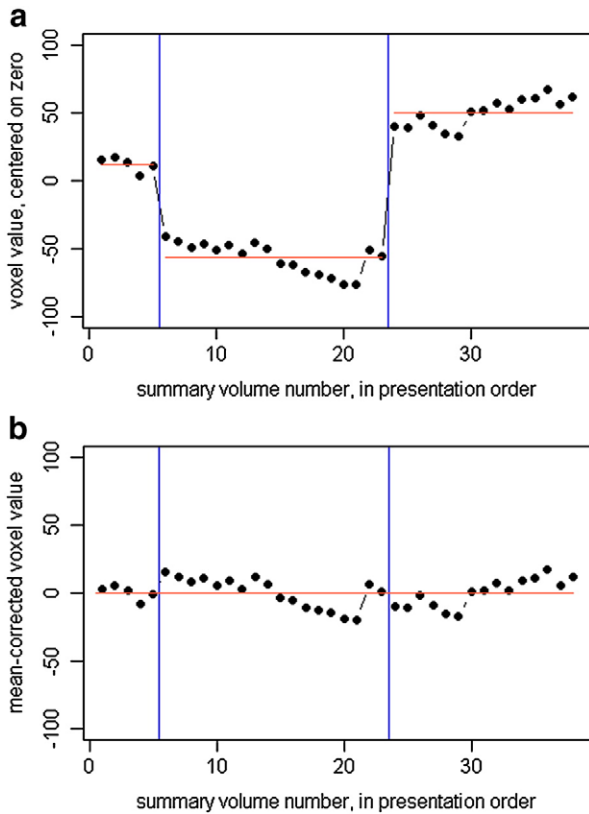


Fig. 1 – Voxel time course for one randomly-selected voxel and subject before (a) and after (b) mean correction. Each dot gives the voxel’s value in a different summary volume, in presentation order. The vertical blue lines indicate breaks between the runs. Pane a shows the time course before detrending. The voxel values were centered on zero for plotting; the raw values range from 950 to 1100. The horizontal red lines show the mean for the values in each run. Pane b shows the same time course after mean correction.

This experiment used a block design, so temporal compression was performed to create one summary volume per block by averaging (for each subject and each voxel individually) the volumes collected immediately following the presentation of each sound (three volumes per block). Alternatively, we could have fit a GLM and used the resulting parameter estimate images; there is currently no way to predict which method is best for a particular case. Next, the data files were divided using the ROI masks to create separate files for each ROI. Then, voxels were removed which had zero variance in any individual subject; voxels with zero variance cannot contribute to a classification and interfere with normalization. The number of voxels remaining in each ROI (Table 1) is small enough (565 maximum) that all voxels in each ROI can be used, without the need for further data reduction.

Finally, the need for detrending was assessed by plotting randomly-selected voxel time courses. Changes in voxel intensities were often noticed at the boundaries between runs (an example voxel is plotted in Fig. 1, pane a), indicating

that detrending was necessary. In this case a simple detrending was employed: subtracting the mean from the activity value for each voxel, separately for each run and subject (Fig. 1, pane b).

3.4. Classification

Support vector machines (SVMs) were chosen for the classifiers for this example, since they often work well with fMRI data. All classifications were performed in R using the svm command in the e1071 package with a linear kernel, the cost parameter fixed at 1, and default scaling (both voxels and examples to zero mean and unit variance). These choices are commonly used with fMRI data (e.g. Haynes et al., 2007; Mourao-Miranda et al., 2007; Pessoa and Padmala, 2007). SVMs generally perform better with scaled

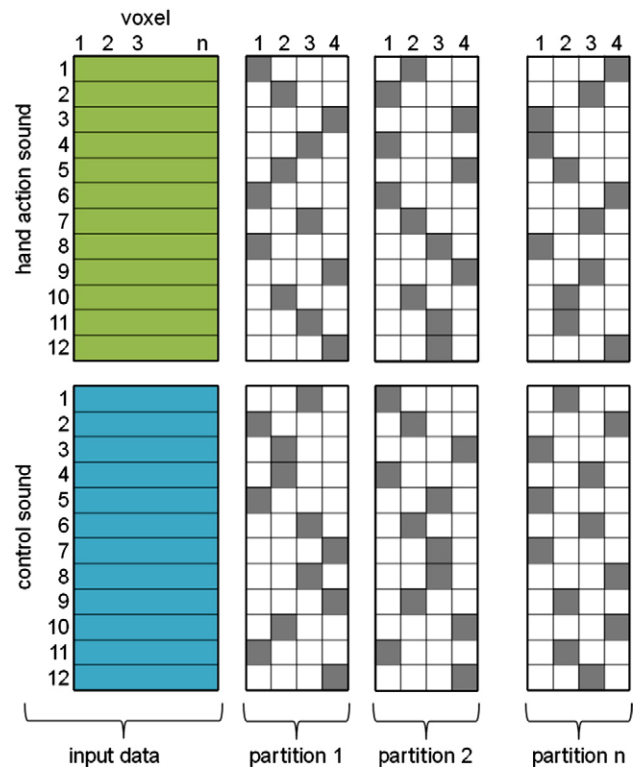


Fig. 2 – Illustration of the data partitioning for a hypothetical within-subjects analysis of one subject with 12 examples of control and hand action sounds. For clarity, only 12 control and hand action sounds are shown here, with 4-fold cross-validation; 24 examples of each type are available in the example analysis, and eight-fold cross-validation was performed. The “input data” section shows the structure of the data file for one subject and ROI: each example is a row, with the value for each voxel in the ROI in separate columns. Three data partitions are shown here, with the four test sets making up each partition in separate columns, and the volumes to include in the test set shaded. For example, the first test set in the first partition is made of the 1st, 6th, and 8th action sound and the 2nd, 5th, and 11th control sound. Each control and action sound is in one test set in each partition.

Table 2 – Classification accuracy (accur.) and *p*-value (*p*) for each partition.

Partition	aud L		aud R		amyg L		amyg R		other L		other R	
	accur.	<i>p</i>	accur.	<i>p</i>	accur.	<i>p</i>	accur.	<i>p</i>	accur.	<i>p</i>	accur.	<i>p</i>
1	0.748	0.001	0.763	0.001	0.521	0.226	0.546	0.044	0.543	0.052	0.552	0.031
2	0.725	0.001	0.755	0.001	0.496	0.579	0.544	0.043	0.548	0.038	0.542	0.062
3	0.756	0.001	0.734	0.001	0.496	0.578	0.554	0.019	0.535	0.091	0.548	0.034
4	0.751	0.001	0.762	0.001	0.506	0.372	0.523	0.189	0.553	0.017	0.533	0.093
5	0.717	0.001	0.757	0.001	0.510	0.333	0.544	0.052	0.551	0.026	0.525	0.149
6	0.734	0.001	0.781	0.001	0.534	0.114	0.562	0.008	0.570	0.003	0.543	0.073
7	0.715	0.001	0.766	0.001	0.513	0.304	0.536	0.088	0.572	0.002	0.560	0.016
8	0.730	0.001	0.765	0.001	0.516	0.251	0.545	0.049	0.565	0.005	0.557	0.012
9	0.731	0.001	0.779	0.001	0.516	0.262	0.552	0.027	0.539	0.067	0.527	0.165
10	0.758	0.001	0.773	0.001	0.517	0.227	0.574	0.001	0.561	0.008	0.537	0.067

data; it can be helpful to evaluate the impact of different types of scaling on each data set (i.e. scale only voxels, only examples, or both).

For most subjects 24 examples of each type of sound were available (each “example” is a summary volume calculated from one block). This is sufficient to allow more than one of each class for the test set; in this case we chose to include three of each class. This is a reasonable balance between test and training set size for fMRI data, and in our experience a range of test set sizes will produce similar results. This is eight-fold stratified cross-validation: test sets were constructed by randomly selecting three of each class of stimuli, which can be done eight times for subjects with 24 examples of each class (i.e. the data was divided into eight groups of three examples; see Fig. 2). In subjects with fewer examples, fewer test sets were made (e.g. only five for the subject with only 14 usable examples). If the number of examples was not divisible by three, the last test set contained fewer examples (i.e. only one or two of each class). The same test sets were used in each subject (i.e. the first, fifth and sixth examples of the control sounds were put into the test sets for each subject). For some subjects the number of examples in each case was not balanced (e.g. 24 hand sounds, but only 22 control sounds). In these cases a random subset of the larger class was used (e.g. 22 of the hand sounds were chose at random to use for the classification).

There are many ways to divide (“partition”) a set of 24 examples into groups of three, so the entire procedure (choosing test sets and the examples to use in the classification, if needed) was repeated ten times to ensure that the accuracy is not dependent on the particular examples chosen. The classification, permutation test, and localization tests were performed for each partition.

3.5. Results

This example involves one hypothesis: the auditory cortex has a different pattern of activity when listening to different types of sounds. This was translated to a specific classification: does the auditory cortex classify the hand action and control sounds? We did a within-subjects test with eight-fold stratified cross-validation, repeating the analysis ten times. This results in an accuracy for each partition and ROI, as shown in Table 2, though these would not usually be reported. The accuracy and *p*-values vary across the partitions. In our experience this range is typical, and illustrates why precise accuracy estimates require repeating the cross-validation. A very large range in accuracies across partitions suggests that the results are unusually unstable, and should be checked for errors or further pre-processing considered.

After calculating the accuracy for each partition the overall results (Table 3) can be calculated, by averaging the partition accuracies. The auditory cortex (bilaterally) could classify the stimuli (accuracy of 0.74 and 0.76), but the amygdala and *other* could not (accuracy of 0.51 to 0.55). The significance of the classification accuracies was calculated by permutation testing. 1000 permutations were performed, so the minimum *p*-value possible is 0.001 (meaning that the true labeling was classified better than all relabeled data sets), which was obtained by both the left and right auditory cortex. Multiple comparison correction should take the ROIs into account, but not necessarily control areas (*other* and amygdala). Since we considered the auditory cortex on the left and right side we should correct for two ROIs; using a Bonferroni correction gives a significance cutoff of $0.5/2=0.025$.

Table 3 – Overall classification accuracy and classification and localization test *p*-values for each ROI.

ROI	Accuracy	Accuracy <i>p</i> -value	Localization <i>p</i> -value
aud L	0.736	0.001	0.001
aud R	0.764	0.001	0.001
amyg L	0.513	0.324	0.812
amyg R	0.548	0.052	0.161
<i>other</i> L	0.554	0.031	NA
<i>other</i> R	0.542	0.070	NA

In more complex analyses it is often useful to perform statistical tests on the accuracy levels to evaluate the relationships between the ROIs or classifications (if multiple classifications were performed in the same experiment). For example, if lateralization were of interest in this example a paired t-test could be used to test whether the left auditory cortex classified more accurately than the right. More complex relationships can be tested with other methods, such as repeated-measures ANOVA to test for differences between ROIs, or mixed models to evaluate a design with multiple classifications, ROIs, and laterality.

Since we were able to define an *other* area we performed a localization test (Table 3), which was performed for each partition and subject separately. The localization test is a quality check which suggests how unusual it is, for these volumes, to have a group of voxels with a particular accuracy. If the classification was based on a systematic error or artifact an unusual pattern of results would be expected, such as wide subset accuracy variability. That did not happen here; rather, voxels from the auditory ROI were able to classify the stimuli more accurately than any (of 1000) same-sized subsets from the *other* areas, suggesting that there was something “unusual” in the activity pattern of the auditory ROI. The amygdala’s accuracy was not unusual, however, but instead consistent with a region unable to classify, reinforcing the previous finding of non-significant classification accuracy.

In addition to the group-level results, since a within-subjects analysis was performed we can examine each subject’s accuracy (Table 4), which may be informative for some experiments, and is useful for exploring the consistency of the results. Significance is calculated for each subject using the same permuted-label data sets and partitions used to calculate the overall results. When calculating the overall results the accuracy of each permuted data set is averaged across subjects, so that the permutation test is the proportion of permuted data sets with an across-subjects average accuracy higher than the real data. For a single-subject permutation test each subject’s accuracy is compared with

the accuracy of his permuted data sets. In this example, classification was well above chance in the auditory ROIs in most subjects, but was near or slightly below in two subjects (4 and 7).

From these results we can conclude that our hypothesis was correct: significant classification of hand action and control sounds was possible in the auditory cortex, bilaterally. This means that the pattern of activity in the voxels making up the auditory cortex was different when the subjects were listening to hand action sounds and control sounds. The significant localization test result, but poor classification in the amygdala, makes us confident in the results. As we performed a within-subjects analysis, we cannot conclude that the activity patterns are the same in all subjects, but we can conclude that the significant classification is possible in nearly all subjects.

4. Discussion

This paper describes how to use multivariate machine learning methods to conduct anatomical ROI-based analyses. This strategy draws from key tenets of contemporary neuroscience: that information in the brain is encoded in the pattern of activity across large populations of neurons (Georgopoulos et al., 1986), and that brain activity is sometimes plastic: brain area representations changing with experience (Logothetis et al., 1995). Accordingly, it is reasonable to assume that while the same cytoarchitectonic brain region might always react to the same class of stimuli, each person’s activity pattern may differ. Within-subjects analysis is compatible with this speculation, as classifiers consider each participant separately, and so can detect idiosyncratic patterns. By contrast, traditional mass-univariate analyses examine whether each voxel shows similar changes in every subject. Multivariate methods are therefore suited to test many theories of brain function, and since these methods are now mature enough to be productively used in fMRI research, we recommend their application to the many

Table 4 – Classification accuracy (accur.) and *p*-values for each individual subject (sub.).

sub.	aud L		aud R		amyg L		amyg R		other L		other R	
	accur.	<i>p</i> -value	accur.	<i>p</i> -value	accur.	<i>p</i> -value	accur.	<i>p</i> -value	accur.	<i>p</i> -value	accur.	<i>p</i> -value
1	0.972	0.001	0.818	0.008	0.497	0.571	0.62	0.193	0.517	0.459	0.743	0.039
2	0.747	0.016	0.753	0.026	0.4	0.797	0.458	0.676	0.464	0.675	0.411	0.813
3	0.725	0.012	0.904	0.001	0.487	0.578	0.577	0.247	0.469	0.647	0.463	0.672
4	0.612	0.136	0.481	0.593	0.43	0.766	0.556	0.331	0.645	0.09	0.564	0.284
5	0.823	0.002	0.894	0.001	0.623	0.13	0.627	0.112	0.646	0.099	0.604	0.179
6	0.74	0.015	0.854	0.001	0.592	0.186	0.56	0.317	0.644	0.097	0.642	0.099
7	0.46	0.694	0.526	0.458	0.343	0.957	0.431	0.762	0.405	0.839	0.417	0.815
8	0.874	0.001	0.87	0.001	0.547	0.339	0.581	0.228	0.561	0.31	0.504	0.5
9	0.68	0.046	0.819	0.002	0.489	0.563	0.502	0.524	0.493	0.546	0.585	0.215
10	0.926	0.001	0.848	0.001	0.792	0.004	0.607	0.148	0.655	0.099	0.693	0.029
11	0.593	0.237	0.669	0.076	0.474	0.628	0.567	0.29	0.576	0.286	0.483	0.608
12	0.696	0.034	0.627	0.112	0.473	0.615	0.467	0.636	0.435	0.735	0.46	0.664
13	0.669	0.054	0.729	0.02	0.481	0.626	0.533	0.409	0.552	0.325	0.515	0.48
15	0.789	0.003	0.917	0.001	0.561	0.275	0.652	0.084	0.671	0.048	0.684	0.036
16	0.792	0.002	0.738	0.025	0.454	0.697	0.481	0.619	0.531	0.445	0.396	0.839

social neuroscience hypotheses that can be reformulated as ROI-based classification questions. Some aspects of this analysis method remain to be fully understood, however. Several of these issues are especially relevant for multivariate ROI-based analysis, while others apply both to multivariate analysis and fMRI research in general. We will briefly review these issues below.

4.1. ROIs of very different sizes

Due to anatomy anatomical ROIs will vary in size, sometimes dramatically (e.g. the auditory cortex is about one-tenth the size of the premotor cortex). As previously discussed, there is some evidence that classification accuracy behaves asymptotically as more voxels are added, but this has not been thoroughly explored, particularly in a ROI-based context. General guidelines for the number of voxels a ROI should contain would be helpful for interpretation, but may be impossible due to the range of classifiers, preprocessing, and paradigms used.

The evaluation of ROIs of different sizes is difficult because of the complex relationship between voxel count and classification accuracy. Large ROIs, because they contain more voxels, may contain more voxels with classifiable information, and so classify more accurately. Or, large ROIs, because they contain more voxels, could classify less accurately, because the classifier cannot find the informative patterns. It is important to keep these potential confounds in mind when comparing the accuracy of ROIs with very different sizes. It may therefore be better in some cases to restrict strong conclusions to the presence or absence of significant classification in ROIs, instead of directly comparing the accuracy of very different-sized ROIs.

4.2. Sub-ROI information

The method described in this paper provides information at the ROI level (e.g. “significant classification was possible in the right amygdala”). This is useful in many contexts, and sufficient to answer many experimental hypotheses. Additional anatomical detail (“which part of the amygdala?”) may be useful in some situations, however. Different strategies can provide this detail, depending on the particular question. In some situations the creation of additional, smaller ROIs may be helpful, although this risks making correction for multiple comparisons difficult. If the goal is to insure that a few voxels are not responsible for the ROI’s classification the classification can be repeated while leaving out individual voxels or small groups of voxels (e.g. Cox and Savoy, 2003), or the voxel weights could be mapped (if relevant for the classifier used; e.g. Mourao-Miranda et al., 2005). However, if the primary goal of analysis is to find very small-scale patterns a technique aimed at uncovering such patterns is more appropriate (e.g. searchlights (Kriegeskorte et al., 2006) or SPM (Friston et al., 1994)).

4.3. Classification accuracy levels

While there are multiple accepted methods to determine the statistical significance of an accuracy level, determining the

minimum accuracy which should be considered *important* is a much more difficult question, and one that perhaps cannot be satisfactorily answered at the present time. In mass-univariate fMRI analysis very small effects (<1% signal change) are accepted as evidence, as long as the difference is significant at $p < 0.05$ corrected for multiple comparisons. Should classification accuracy be examined using the same standard (important if statistically significant), or should a certain accuracy level be required? Should findings with low, but highly significant accuracy, be considered less important than those with higher accuracy?

4.4. Use of feature selection

This paper only briefly discussed feature selection beyond ROI definition (using ROIs is itself a feature selection method). It may sometimes be possible to obtain higher classification accuracy using only a subset of the voxels in each ROI (employing additional feature selection), rather than all voxels. If feature selection is used within ROIs, what proportion of the voxels can be omitted before it is unreasonable to claim that “the ROI” has significant classification accuracy? Additionally, sometimes multivariate analysis is performed by selecting voxels from the entire brain, rather than restricting the voxels to ROIs. These results are often interpreted like mass-univariate analyses: checking to see where the voxels resulting in high classification accuracy fall, requiring decisions to be made on issues such as minimum cluster size and clusters which overlap multiple brain structures. How should significant classification in voxels scattered across the brain, overlapping several ROIs, or encompassing only a small fraction of a ROI be interpreted?

4.5. Anatomical alignment

The quality of anatomical alignment between subjects is a problem not just for multivariate ROI-based analyses, but for all fMRI research: we are often not certain that a voxel represents the same anatomical region in all subjects even after normalization (Brett et al., 2002; Devlin and Poldrack, 2007). Multivariate ROI analysis may be less reliant on precise anatomical alignment, particularly in within-subjects analyses when anatomical (or functional) ROIs are defined on each individual subject, since only the classification accuracies are compared across subjects. Even when within-subjects analyses are performed on normalized data with one set of ROI masks the alignment assumption is somewhat weak: we assume that the ROI mask contains the ROI for each subject, but not that each voxel *within* the ROI is equivalent. Between-subjects analysis assumes that the voxels are equivalent in each subject, and may require smoothing if this is not achieved.

Acknowledgments

The research was supported by a Nederlandse Organisatie voor Wetenschappelijk Onderzoek VIDI and a Marie Curie Excellence grant to Christian Keysers. Computer time for the example analysis was provided by the Centre for High-

Performance Computing and Visualisation at the University of Groningen (The Netherlands) on the HPC Cluster.

REFERENCES

- Breiman, L., 2001a. Statistical modeling: the two cultures. *Stat. Sci.* 16, 199–215.
- Breiman, L., 2001b. Random forests. *Mach. Learn.* 45, 5–32.
- Brett, M., Johnsrude, I.S., Owen, A.M., 2002. The problem of functional localization in the human brain. *Nat. Rev. Neurosci.* 3, 243–249.
- Buchel, C., Friston, K.J., 1997. Modulation of connectivity in visual pathways by attention: cortical interactions evaluated with structural equation modelling and fMRI. *Cereb. Cortex* 7, 768–778.
- Burges, C.J.C., 1998. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* 2, 121–167.
- Carlson, T.A., Schrater, P., He, S., 2003. Patterns of activity in the categorical representations of objects. *J. Cogn. Neurosci.* 15, 704–717.
- Cox, D.D., Savoy, R.L., 2003. Functional magnetic resonance imaging (fMRI) “brain reading”: detecting and classifying distributed patterns of fMRI activity in human visual cortex. *Neuroimage* 19, 261–270.
- Davatzikos, C., Ruparel, K., Fan, Y., Shen, D.G., Acharyya, M., Loughhead, J.W., Gur, R.C., Langleben, D.D., 2005. Classifying spatial patterns of brain activity with machine learning methods: application to lie detection. *Neuroimage* 28, 663–668.
- De Martino, F., Valente, G., Staeren, N., Ashburner, J., Goebel, R., Formisano, E., 2008. Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. *Neuroimage* 43, 44–58.
- Devlin, J.T., Poldrack, R.A., 2007. In praise of tedious anatomy. *Neuroimage* 37, 1033–1041.
- Devor, A., Hillman, E.M.C., Tian, P., Waeber, C., Teng, I.C., Ruvinskaya, L., Shalinsky, M.H., Zhu, H., Haslinger, R.H., Narayanan, S.N., Ulbert, I., Dunn, A.K., Lo, E.H., Rosen, B.R., Dale, A.M., Kleinfeld, D., Boas, D.A., 2008. Stimulus-induced changes in blood flow and 2-deoxyglucose uptake dissociate in ipsilateral somatosensory cortex. *J. Neurosci.* 28, 14347–14357.
- Dinstein, I., Gardner, J.L., Jazayeri, M., Heeger, D.J., 2008. Executed and observed movements have different distributed representations in human aIPS. *J. Neurosci.* 28, 11231–11239.
- Edgington, E.S., 1995. *Randomization Tests*. Marcel Dekker, New York.
- Eickhoff, S.B., Stephan, K.E., Mohlberg, H., Grefkes, C., Fink, G.R., Amunts, K., Zilles, K., 2005. A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *Neuroimage* 25, 1325–1335.
- Eickhoff, S.B., Heim, S., Zilles, K., Amunts, K., 2006. Testing anatomically specified hypotheses in functional imaging using cytoarchitectonic maps. *Neuroimage* 32, 570–582.
- Etzel, J.A., Gazzola, V., Keysers, C., 2008. Testing simulation theory with cross-modal multivariate classification of fMRI data. *PLoS One* 3, e3690.
- Formisano, E., De Martino, F., Valente, G., 2008. Multivariate analysis of fMRI time series: classification and regression of brain responses using machine learning. *Magn. Reson. Imaging* 26, 921–934.
- Friston, K.J., Frith, C.D., Liddle, P.F., Frackowiak, R.S.J., 1993. Functional connectivity: the principal-component analysis of large(PET) data sets. *J. Cereb. Blood Flow Metab.* 13, 5–14.
- Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.P., Frith, C.D., Frackowiak, R.S.J., 1994. Statistical parametric maps in functional imaging: a general linear approach. *Hum. Brain Mapp.* 2, 189–210.
- Friston, K.J., Rotshtein, P., Geng, J.J., Sterzer, P., Henson, R.N., 2006. A critique of functional localisers. *Neuroimage* 30, 1077–1087.
- Friston, K.J., Chu, C., Mourao-Miranda, J., Hulme, O., Rees, G., Penny, W., Ashburner, J., 2008. Bayesian decoding of brain images. *Neuroimage* 39, 181–205.
- Gazzola, V., Aziz-Zadeh, L., Keysers, C., 2006. Empathy and the somatotopic auditory mirror system in humans. *Curr. Biol.* 16, 1824–1829.
- Georgopoulos, A.P., Schwartz, A.B., Kettner, R.E., 1986. Neuronal population coding of movement direction. *Science* 233, 1416–1419.
- Goebel, R., Roebroeck, A., Kim, D.-S., Formisano, E., 2003. Investigating directed cortical interactions in time-resolved fMRI data using vector autoregressive modeling and Granger causality mapping. *Magn. Reson. Imaging* 21, 1251–1261.
- Golland, P., Fischl, B., 2003. Permutation tests for classification: towards statistical significance in image-based studies. In: Taylor, C.J., Noble, J.A. (Eds.), *Information Processing in Medical Imaging. Lecture Notes in Computer Science*, Vol. 2732/2003. Springer, Berlin, pp. 330–341.
- Good, P.I., 2001. *Resampling methods: a practical guide to data analysis*, Vol. Birkhauser, Boston, Basel, Berlin.
- Greer, B., Mulhern, G., 2002. *Making Sense of Data and Statistics in Psychology*. Palgrave.
- Hampton, A.N., O’Doherty, J.P., 2007. Decoding the neural substrates of reward-related decision making with functional MRI. *Proc. Natl. Acad. Sci. U. S. A.* 104, 1377–1382.
- Hanson, S.J., Halchenko, Y.O., 2008. Brain reading using full brain support vector machines for object recognition: there is no “face” identification area. *Neural Comput.* 20, 486–503.
- Hastie, T., Tibshirani, R., Friedman, J., 2001. *The Elements of Statistical Learning*. Springer.
- Haynes, J.-D., Rees, G., 2006. Decoding mental states from brain activity in humans. *Nat. Rev. Neurosci.* 7, 523–534.
- Haynes, J.-D., Sakai, K., Rees, G., Gilbert, S., Frith, C., Passingham, R.E., 2007. Reading hidden intentions in the human brain. *Curr. Biol.* 17, 323–328.
- Hsu, C.W., Lin, C.J., 2002. A comparison of methods for multiclass support vector machines. *IEEE Trans. Neural Netw.* 13, 415–425.
- Huettel, S.A., Song, A.W., McCarthy, G., 2004. *Functional Magnetic Resonance Imaging*. Sinauer Associates, Inc, Sunderland, Massachusetts USA.
- Joaquims, T., 1999. Making large-scale support vector machine learning practical. In: Scholkopf, B., Burges, C.J.C., Smola, A.J. (Eds.), *Advances in Kernel Methods: Support Vector Learning*. MIT Press, Cambridge, MA, pp. 169–184.
- Kamitani, Y., Tong, F., 2006. Decoding seen and attended motion directions from activity in the human visual cortex. *Curr. Biol.* 16, 1096–1102.
- Kriegeskorte, N., Goebel, R., Bandettini, P., 2006. Information-based functional brain mapping. *PNAS* 103, 3863–3868.
- Kriegeskorte, N., Bandettini, P., 2007. Combining the tools: activation- and information-based fMRI analysis. *Neuroimage* 38, 666–668.
- Ku, S.-p., Grettton, A., Macke, J., Logothetis, N.K., 2008. Comparison of pattern recognition methods in classifying high-resolution BOLD signals obtained at high magnetic field in monkeys. *Magn. Reson. Imaging* 26, 1007–1014.
- LaConte, S., Strother, S., Cherkassky, V., Anderson, J., Hu, X., 2005. Support vector machines for temporal classification of block design fMRI data. *Neuroimage* 26, 317–329.
- Logothetis, N.K., Pauls, J., Poggio, T., 1995. Shape representation in the inferior temporal cortex of monkeys. *Curr. Biol.* 5, 552–563.

- Logothetis, N.K., 2008. What we can do and what we cannot do with fMRI. *Nature* 453, 869–878.
- Macey, P.M., Macey, K.E., Kumar, R., Harper, R.M., 2004. A method for removal of global effects from fMRI time series. *Neuroimage* 22, 360–366.
- Mitchell, T.M., Hutchinson, R., Niculescu, R.S., Pereira, F., Wang, X., 2004. Learning to decode cognitive states from brain images. *Mach. Learn.* 57, 145–175.
- Mourao-Miranda, J., Bokde, A.L.W., Born, C., Hampel, H., Stetter, M., 2005. Classifying brain states and determining the discriminating activation patterns: support vector machine on functional MRI data. *Neuroimage* 28, 980–995.
- Mourao-Miranda, J., Reynaud, E., McGlone, F., Calvert, G., Brammer, M., 2006. The impact of temporal compression and space selection on SVM analysis of single-subject and multi-subject fMRI data. *Neuroimage* 33, 1055–1065.
- Mourao-Miranda, J., Friston, K.J., Brammer, M., 2007. Dynamic discrimination analysis: a spatial-temporal SVM. *Neuroimage* 36, 88–99.
- Mukherjee, S., Golland, P., Panchenko, D., 2003. Permutation tests for classification. AI Memo 2003-019. Vol. Massachusetts Institute of Technology Computer Science and Artificial Intelligence Laboratory.
- Mur, M., Bandettini, P.A., Kriegeskorte, N., 2009. Revealing representational content with pattern-information fMRI—an introductory guide. *Soc. Cogn. Affect. Neurosci.* nsn044.
- Nandy, R., Cordes, D., 2004. Improving the spatial specificity of canonical correlation analysis in fMRI. *Magn. Reson. Med.* 52, 947–952.
- Nichols, T.E., Holmes, A.P., 2001. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum. Brain Mapp.* 15, 1–25.
- Nieto-Castanon, A., Ghosh, S.S., Tourville, J.A., Guenther, F.H., 2003. Region of interest based analysis of functional imaging data. *Neuroimage* 19, 1303–1316.
- Norman, K.A., Polyn, S.M., Detre, G.J., Haxby, J.V., 2006. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cogn. Sci.* 10, 424–430.
- O’Toole, A.J., Jiang, F., Abdi, H., Penard, N., Dunlop, J.P., Parent, M.A., 2007. Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data. *J. Cogn. Neurosci.* 19, 1735–1752.
- Pereira, F., Mitchell, T., Botvinick, M., 2009. Machine learning classifiers and fMRI: a tutorial overview. *Neuroimage* 45, S199–S209.
- Pessoa, L., Padmala, S., 2007. Decoding near-threshold perception of fear from distributed single-trial brain activation. *Cereb. Cortex* 17, 691–701.
- Petrie, A., Sabin, C., 2000. *Medical Statistics at a Glance*. Blackwell Science, Oxford.
- Poldrack, R.A., 2007. Region of interest analysis for fMRI. *Soc. Cogn. Affect. Neurosci.* 2, 67–70.
- R Development Core Team, 2008. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Roebroeck, A., Formisano, E., Goebel, R., 2005. Mapping directed influence over the brain using Granger causality and fMRI. *Neuroimage* 25, 230–242.
- Saxe, R., Brett, M., Kanwisher, N., 2006. Divide and conquer: a defense of functional localizers. *Neuroimage* 30, 1088–1096.
- Schölkopf, B., Smola, A.J., 2002. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA.
- Schwarzlose, R.F., Swisher, J.D., Dang, S., Kanwisher, N., 2008. The distribution of category and location information across object-selective regions in human visual cortex. *Proc. Natl. Acad. Sci.* 105, 4447–4452.
- Shinkareva, S.V., Mason, R.A., Malave, V.L., Wang, W., Mitchell, T.M., Just, M.A., 2008. Using fMRI brain activation to identify cognitive states associated with perception of tools and dwellings. *PLoS One* 3, e1394.
- Soon, C.S., Brass, M., Heinze, H.J., Haynes, J.D., 2008. Unconscious determinants of free decisions in the human brain. *Nat. Neurosci.* 11, 543–545.
- Strother, S.C., 2006. Evaluating fMRI preprocessing pipelines. *IEEE Eng. Med. Biol. Mag.* 25, 27–41.
- Tanabe, J., Miller, D., Tregellas, J., Freedman, R., Meyer, F.G., 2002. Comparison of detrending methods for optimal fMRI preprocessing. *Neuroimage* 15, 902–907.
- Thomas, C.G., Harshman, R.A., Menon, R.S., 2002. Noise reduction in BOLD-based fMRI using component analysis. *Neuroimage* 17, 1521–1537.
- Tohka, J., Foerde, K., Aron, A.R., Tom, S.M., Toga, A.W., Poldrack, R.A., 2008. Automatic independent component labeling for artifact removal in fMRI. *Neuroimage* 39, 1227–1245.
- Vapnik, V.N., 2000. *The Nature of Statistical Learning Theory*. Springer, New York, NY.
- Witten, I.H., Frank, E., 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers.