Tools of the Trade

Revealing representational content with pattern-information fMRI—an introductory guide

Marieke Mur,[1,2] Peter A. Bandettini,[1,3] and Nikolaus Kriegeskorte[1]

[1]Section on Functional Imaging Methods, Laboratory of Brain and Cognition, National Institute of Mental Health, National Institutes of Health, Bethesda, MD, USA, [2]Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, The Netherlands, and [3]Functional Magnetic Resonance Imaging Facility, National Institute of Mental Health, National Institutes of Health, Bethesda, MD, USA

Conventional statistical analysis methods for functional magnetic resonance imaging (fMRI) data are very successful at detecting brain regions that are activated as a whole during specific mental activities. The overall activation of a region is usually taken to indicate *involvement* of the region in the task. However, such activation analysis does not consider the multivoxel patterns of activity within a brain region. These patterns of activity, which are thought to reflect neuronal population codes, can be investigated by pattern-information analysis. In this framework, a region's multivariate pattern information is taken to indicate *representational content*. This tutorial introduction motivates pattern-information analysis, explains its underlying assumptions, introduces the most widespread methods in an intuitive way, and outlines the basic sequence of analysis steps.

INTRODUCTION

Conventional statistical analysis of functional magnetic resonance imaging (fMRI) data focuses on finding macroscopic brain regions that are involved in specific mental activities (Friston et al., 1994, 1995a,b; Worsley and Friston, 1995). In order to find and characterize brain regions that become activated as a whole, data is usually spatially smoothed and activity is averaged across voxels within a region of interest (ROI). These analysis steps increase sensitivity to spatially extended activations, but result in loss of sensitivity to fine-grained spatial-pattern information. In recent years, there has been a growing interest in going beyond *activation* assessment and analyzing fMRI data for the *information* carried by fine-grained patterns of activity within each functional region (Norman et al., 2006; Haynes and Rees, 2006; Kriegeskorte and Bandettini, 2007a). The goal of this tutorial paper is to motivate the use of pattern-information analysis and to provide a step-by-step introduction on how to implement this method.

A region's *involvement* in task processing versus its *representational content*

Conventional analysis focuses on regions that become activated as a whole during the performance of a specific task.

This motivates spatial smoothing of the data and averaging of activity across an ROI. Since this approach focuses on activations (in the sense of blobs consisting of multiple voxels all showing effects in the same direction) we refer to it as activation-based analysis. Activation-based analysis aims to detect regional-average activation differences and infer *involvement* of the region in a specific mental function. Pattern-information analysis, by contrast, aims to detect activity-pattern differences and infer *representational content* (see Table 1, Figure 1).

Regional activity patterns can reflect the neural population code (for a striking example, see Kamitani and Tong, 2005). However, fine-grained pattern differences go undetected in activation-based analysis unless the regional-average activation also differs (see Figure 1). Pattern-information analysis is suited for detecting pattern changes even if they occur in the absence of regional-average activation changes. For example, a recent study using pattern-information analysis showed that perceptually discriminable speech sounds elicit different patterns of activity in right auditory cortex (Raizada et al., 2008). The speech sounds elicited similar regional-average activation, but the patterns were statistically discriminable.
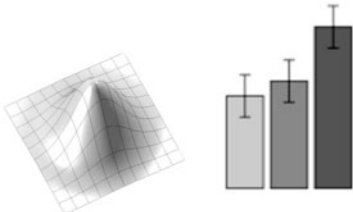
Scope and limitations

The use of pattern-information analysis is not restricted to investigating functional regions defined by activation-based analysis. It can also be used to investigate patterns of activity

Table 1 Overview of activation-based and pattern-information analysis

	Activation-based analysis	Pattern-information analysis
Goal of the analysis	Investigating the *involvement* of regions in a specific mental activity	Investigating the *representational content* of regions
Experimental contrast	Difference between mental activity *including* component of interest and mental activity *excluding* component of interest	Difference between representation of object 1 and representation of object 2
Analytical comparison	Compare spatial-average activation across conditions	Compare patterns of activity across conditions
Spatial resolution	Benefits of high-resolution imaging will be limited if data are smoothed	Fine-grained spatial information provided by high-resolution imaging is used effectively
Statistical methods	• Spatial smoothing • Combine single-voxel signals by smoothing and averaging activity within ROI • Univariate analysis • Group analysis in common stereotactic space	• No spatial smoothing • Combine single-voxel signals by computing multivariate statistics • Multivariate analysis (typically linear discriminant analysis) • Single-subject analysis in native subject space • Group analysis in common stereotactic space at the pattern-information level

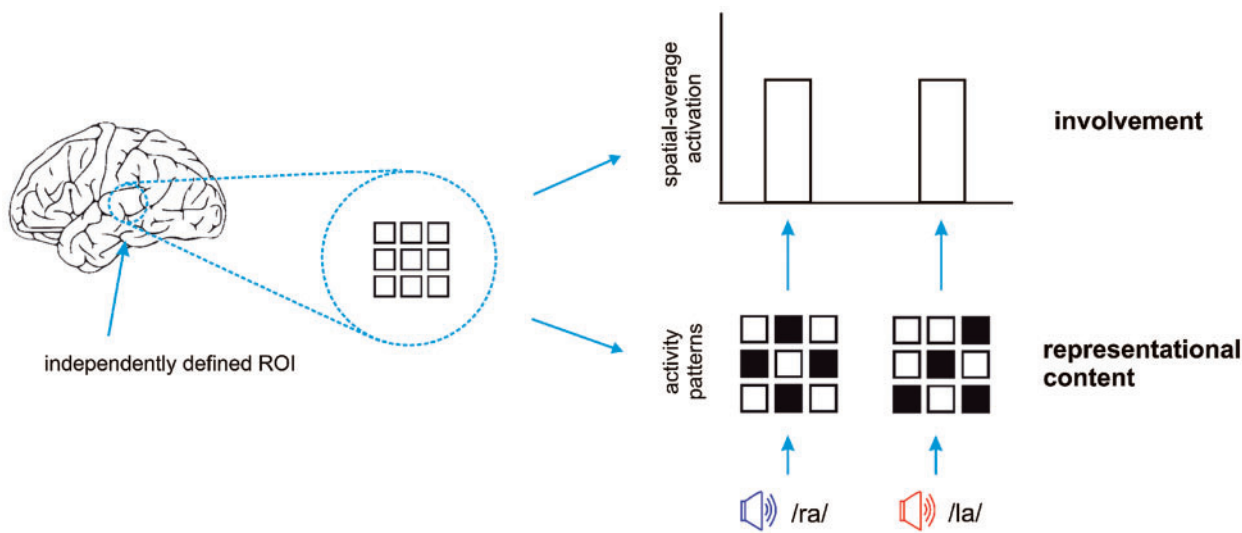Images in this table are reprinted with permission from Kriegeskorte and Bandettini (2007b).

Fig. 1 Activation indicates *involvement*, pattern-information indicates *representational content*. A specific ROI can show the same spatial-average activation resulting from different patterns encoding different representational content. This figure shows a hypothetical ROI consisting of nine voxels. The ROI's multivoxel pattern of activity is different for /ra/ than /la/ speech sounds, but these different patterns result in the same spatial-average activation. This difference will go undetected by conventional activation-based analysis. Pattern-information analysis can be used to show that an ROI's multivoxel activity pattern differs significantly across conditions, i.e. that the region contains information about the experimental conditions. Differences in multivoxel patterns across conditions can be interpreted as reflecting differences in underlying neuronal population activity. This figure has been adapted with permission from Raizada et al. (2008).

across more widely distributed sets of voxels (e.g. Haxby et al., 2001; Carlson et al., 2003) or to *define* functional regions by mapping the whole volume for effects using a multivariate searchlight (''information-based brain mapping'', Kriegeskorte et al., 2006, 2007). The change that activation-based analysis is sensitive to—all voxels changing their activity in *the same direction*—can be viewed as a special case of the changes that pattern-information analysis can detect: any change of the pattern, including spatial-mean activity changes as well as pattern changes where the spatial-mean is unaffected. This general sensitivity makes pattern-information analysis a powerful statistical tool.

With many successful applications in neuroimaging, the approach has gained momentum in recent years (e.g. Haxby et al., 2001; Carlson et al., 2003; Cox and Savoy, 2003; Friston et al., 2008; Hanson et al., 2004; Kamitani and Tong, 2005; Haynes and Rees, 2005; Haynes et al., 2007; Kriegeskorte et al., 2007; Kriegeskorte et al., 2008a; Mourao-Miranda et al., 2005; Mitchell et al., 2008; O'Toole et al., 2005; Pereira et al., 2008; Raizada et al., 2008). Note that related multivariate methods as well as prediction frameworks have been explored before in neuroimaging analysis (Strother et al., 2002; Worsley et al., 1997), but with different conceptual goals.

The blood-oxygen-level-dependent (BOLD) fMRI signal provides a complex reflection of underlying neural activity and is affected by noise (Boynton et al., 1996; Logothetis, 2008). As a consequence, interpretation of the BOLD fMRI signal in terms of underlying neural activity requires caution. The BOLD fMRI contrast has been shown to reflect stimulus-driven neural activity (Logothetis et al., 2001). Although the fine-grained activity patterns measured by fMRI may not precisely reflect neural activity patterns because of hemodynamic blurring and distortion, a change of signal (patterns) across conditions can be interpreted as a change of neural population activity.

Pattern-information fMRI is fundamentally limited by the amount of information about the neural population codes that can be provided by fMRI. Voxel resolution is one such limitation, thus motivating the use of high-resolution fMRI in conjunction with pattern-information analysis (Kriegeskorte and Bandettini, 2007a; Kriegeskorte et al., 2007). A technique that also targets the representational content of functional regions and that is not limited by voxel resolution is fMRI adaptation (Grill-Spector and Malach, 2001). This approach can potentially resolve sub-voxel representations by inferring neural selectivity from fMRI adaptation responses. However, the interpretation of positive findings ("release from adaptation") in terms of neural population selectivity relies on assumptions that have been questioned by recent experimental results (Tolias et al., 2005; Sawamura et al., 2006; Krekelberg et al., 2006). These results showed that release from adaptation does not necessarily reflect selectivity of the underlying neural population as measured by classical electrophysiological methods. Other explanations, e.g. attentional effects or carry-over of effects from connected regions (Tolias et al., 2005; Krekelberg et al., 2006), can account for release from adaptation as well. While the fMRI adaptation paradigm compares activation between pairs of either different or repeated stimuli and then *infers* single-stimulus selectivity from these activation differences, pattern-information fMRI follows the simpler logic of contrasting experimental conditions directly to determine if there is an effect on the dependent variable: the activity pattern within an ROI. Although its sensitivity is limited by the measurement technique of fMRI, a positive result, i.e. statistically distinct activity patterns, provides strong evidence for a difference between the underlying neural activity patterns in the region. It has recently been shown that it is possible to combine pattern-information fMRI and fMRI adaptation in a single experiment and simultaneously estimate activity patterns and adaptation effects (Aguirre, 2007).

Study design

Both event-related and block designs can be used in combination with pattern-information analysis. The choice will largely be based on similar considerations as for studies using activation-based analyses. Block designs yield a higher functional contrast-to-noise ratio than event-related designs. This holds both for constant inter-stimulus-interval (ISI) event-related designs (Bandettini and Cox, 2000) and jittered rapid event-related designs (Birn et al., 2002). This implies that block designs will generally yield better estimates of the average response pattern (i.e. the centroid) than event-related designs. This is especially useful for discriminating a small number of conditions (e.g. Haxby et al., 2001). However, event-related designs can be preferable for psychological reasons as they are less predictable and can reduce habituation effects. Moreover, event-related designs can accommodate a larger number of conditions (Kriegeskorte et al., 2008b). Another advantage of particular importance to information-based analysis is that they yield more independent data points than block designs and can therefore yield a better estimate of the shape of each condition's multivariate response distribution. This can improve classification performance and, thus, increase sensitivity in detecting pattern information. On the other hand the condition-mean pattern estimates (centroids) will typically be somewhat noisier. It should also be noted that rapid-event related designs involve temporally overlapping hemodynamic responses. The effects of temporal overlap can be accounted for using the same design optimization techniques that have proven useful for activation-based studies.

Imaging parameters

Most pattern-information analyses so far have utilized lower-resolution fMRI data (see Haxby et al., 2001; Kamitani and Tong, 2005; Haynes and Rees, 2005), indicating that larger-scale patterns—even if dominated by vascular effects—can contain a considerable amount of information even about quite fine-grained neuronal patterns (consider Kamitani and Tong, 2005). If information on a fine spatial scale is of interest, high-resolution fMRI (Kriegeskorte et al., 2007) might be a better choice. However, the tradeoff between the functional-contrast-to-noise ratio and the resolution has to be carefully considered (Kriegeskorte and Bandettini, 2007a). A voxel size of about 2 mm in each dimensions appears to be a reasonable compromise at 3 Tesla.

TESTING FOR PATTERN INFORMATION

In this section, we describe how to test for a multivariate activity-pattern difference. A significant pattern difference implies that the condition can be decoded (with some accuracy above chance level) from the activity patterns. In other words, it implies pattern-information about the experimental condition.

A wide variety of multivariate methods can be used for pattern-information analysis. All these methods aim to determine whether the patterns of activity associated with different conditions are statistically discriminable (i.e. significantly different). As in conventional analysis, every activity pattern we estimate from the data results from a combination of true effects and noise. Noise is always present and will make every pattern unique (just as in a univariate t-test there is always a small difference between the estimates of the two means to be compared, even if the null hypothesis is true). We need to determine whether the patterns associated with, say, condition A and condition B, are more different than expected under the null hypothesis of equal activity patterns in both conditions. Under the null hypothesis, any differences between the pattern estimates would be produced by noise alone.

Univariate data is usually analyzed using a *t*-test or analysis of variance (ANOVA). For multivariate data, the equivalent method would be a multivariate analysis of (co)variance (MANOVA). However, this method assumes that the distribution of the residuals is multivariate normal, an assumption that might not hold for fMRI data. This is one reason why most of the cited studies approach pattern analysis as a classification problem: If we can classify the experimental conditions (which elicit the representational states we are interested in) on the basis of the activity patterns better than chance, this indicates that the response pattern carries information about the experimental conditions. This approach has been referred to as "brain reading" (Cox and Savoy, 2003) or "decoding".

Linear classification is the most widespread and successful pattern-information analysis in neuroimaging so far

Multivoxel patterns of activity can be viewed as points in a multidimensional space (with as many dimensions as voxels). Consider the simple case of patterns based on activity of only two voxels. Each pattern can then be thought of as a point on a plane, where the activity in each voxel determines one of the coordinates (Figure 2). One way to classify these patterns is to construct a line that separates the patterns belonging to condition A from the patterns belonging to condition B (solid green lines in Figure 2). Patterns on one side of the line will be classified as condition A, patterns
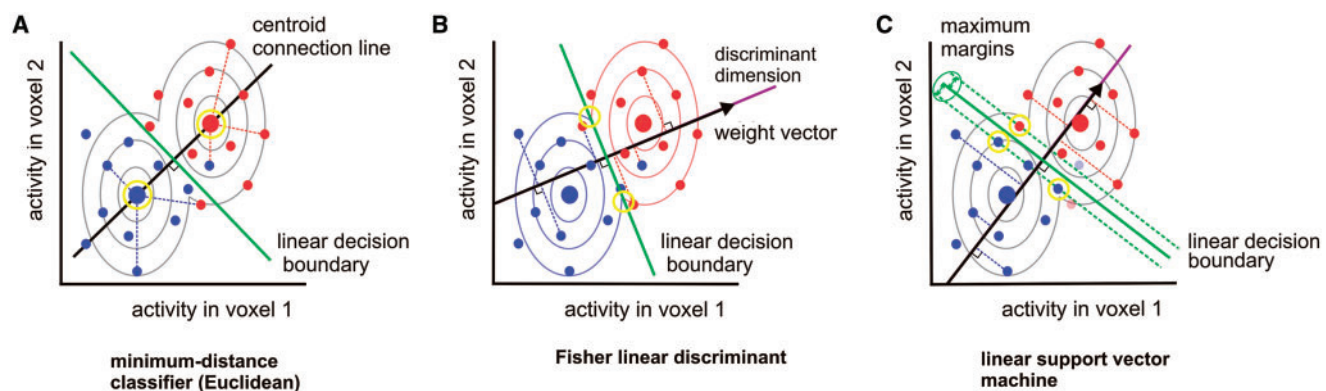

Fig. 2 Linear classification methods all define a linear decision boundary, but the boundary is placed slightly differently. This is shown for a given set of hypothetical activity patterns. The blue dots represent activity patterns for one experimental condition (e.g. the speech sound /ra/), the red dots represent activity patterns for a second condition (e.g. the speech sound /ra/). For simplicity, the displayed activity patterns are based on activity of only two voxels. Nevertheless, the classification methods generalize to higher-dimensional voxel spaces. The ellipses in the background of each panel are iso-probability-density contours describing the bivariate normal distribution of the activity patterns for each condition. The yellow circles indicate the geometrical features that define the linear decision boundary (green) for each classifier. (**A**) Minimum-distance classifier. This classifier first determines the centroids of the two multivariate distributions (large dots). Each activity pattern is then classified as the condition whose centroid it is closest to in multivariate space (using Euclidean distance here, as shown by the dotted lines). This implies a linear decision boundary (i.e. a hyperplane) orthogonal to the centroid connection line, equally dividing the distance between the two centroids. (**B**) Fisher linear discriminant analysis (FLDA). Response patterns are projected onto a linear discriminant dimension by weighting each voxel's activity in order to maximize the ratio of between-condition and within-condition variance. The voxel weights define a weight vector that points in the direction of the linear discriminant dimension. The patterns (i.e. the data points) are orthogonally projected onto the discriminant dimension and a threshold is used for classification. This implies a linear decision boundary (i.e. a hyperplane) orthogonal to the linear discriminant dimension. (**C**) Linear SVM. Same description as FLDA, except for the way the voxel weights are computed. The voxel weights computed by linear SVM are set to yield a linear decision boundary that maximizes the margin (i.e. the distance of the nearest data point to the decision boundary). To make this intuitive, we can imagine starting with a decision boundary that perfectly classifies the training set, then widening the margin equally on both sides while adjusting the angle and position of the decision boundary, until the margin cannot be widened anymore without including one of the training data points. The response patterns closest to the decision boundary (points in yellow circles) then define the margins and the decision boundary halfway in-between the margins. These points are therefore called "support vectors". In order to handle overlapping distributions, SVM algorithms are typically set to allow for a few misclassifications on the training set (see the two transparent points in our hypothetical example).

on the other side will be classified as condition B. For more than two voxels, the plane becomes a higher-dimensional space and the decision line generalizes to a linear decision boundary (also called a decision hyperplane). Classifiers that use a linear decision boundary are referred to as *linear* or *hyperplane* classifiers. Linear classification is the most widespread and successful tool for pattern-information analysis in neuroimaging so far.[1] A good introductory textbook on the mathematics of pattern classification is Duda et al. (2001).

The three most widespread linear classification methods in pattern-information fMRI (Figure 2) are the minimum-distance classifier (e.g. Haxby et al., 2001), Fisher linear discriminant analysis (FLDA) (e.g. Carlson et al., 2003) and the linear support vector machine (SVM) (e.g. Cox and Savoy, 2003). Each of these methods places the linear decision boundary slightly differently (solid green lines in Figure 2).

These methods will perform optimally under different assumptions about the distribution of the response patterns. In practice, they tend to perform somewhat similarly on fMRI data and there is no strong evidence to date suggesting a general superiority of any one of them in this context (but see Ku et al., 2008; Mourao-Miranda et al., 2005). Importantly the differences concern the *sensitivity* for detection of pattern information, not the *specificity* (i.e. the false-positives rate for detecting information). Thus, any of the methods can provide a valid statistical test of pattern-information when correctly applied.

Subtle differences between linear classifiers

In this section we provide a conceptual description of the three methods to give the interested reader an intuitive sense of how the linear decision boundary is placed in each method (solid green lines in Figure 2).

The minimum-distance classifier assigns each activity pattern to the condition whose centroid (multivariate mean) it is closest to in multivariate space. This results in a linear decision boundary orthogonal to the centroid connection line and equally dividing the distance between the two centroids (Figure 2A)—assuming that the multivariate distance is simply measured as the length of a straight line connecting the two points (i.e. the Euclidean distance). Using Euclidean distance, this method performs optimally when the distributions associated with the two conditions are identical (homoscedasticity) and isotropic (i.e. they fall off in the same way in all directions of multivariate space). Alternatively, the correlation of the patterns across voxels can be used to compare patterns. A correlation-based distance can be obtained as 1-r, where r is the correlation

coefficient. Minimum-distance classification using the correlation distance is equivalent to the method used by Haxby et al. (2001). Note that using pattern correlation renders the analysis insensitive to regional-average differences (activation effects), which may be desirable. With either distance measure, the minimum-distance classifier implies a linear decision boundary.

Unlike minimum-distance classification, FLDA (Figure 2B) takes the covariance structure of the data into account. FLDA is equivalent to modeling each condition's distribution as a multivariate normal distribution (with a covariance estimate pooled across the two conditions) and classifying each pattern as the condition that has the greater probability density at that point in the space. As a consequence, FLDA performs optimally when the distributions associated with the two conditions actually are approximately multivariate normal[2] (but not necessarily isotropic) and have the same covariance structure (homoscedasticity).

Linear SVM does not assume multivariate normality. Instead it searches for a linear decision boundary that not only discriminates the two sets of points but also has the maximum margin (greatest distance to the nearest points on both sides; Figure 2C). The response patterns on the margins are referred to as the "support vectors", because they "support" the margins and define the decision hyperplane. In other words, linear SVM only uses the most informative subset of data (the support vectors) for constructing the boundary. A linear SVM decision boundary will not change when data points (response patterns) far away from the boundary are moved—as long as the support vectors do not change. In contrast, an FLDA or minimum-distance-classifier decision boundary will move when any data point is shifted.

Mathematically, the linear decision boundary is defined by a vector w that points orthogonal to it in multivariate activity-pattern space and by a parameter that shifts it to the best location. We can think of each linear classifier as using a different rule for determining the vector w and the shift parameter. For a given linear decision boundary, we can use the vector w to determine which side a pattern falls on. To this end, we compute a weighted sum (also called a linear combination) of the voxel responses using the entries of the vector w as the weights, which is why w is also known as the weight vector.[3] Geometrically, computing a weighted sum of voxel responses corresponds to orthogonally projecting an activity pattern (point in multivariate space) onto a

---

[1]　Nonlinear classification algorithms have also been used for pattern-information analysis (e.g. Cox and Savoy, 2003; LaConte et al., 2005). These algorithms can capture more complicated class boundaries than linear classifiers. However, non-linear classification methods are more prone to overfit the data than linear classification methods. Overfitting is a particularly severe problem in fMRI because the number of data points (condition repetitions or time points) is typically not very large in relation to the number of ROI voxels. Overfitting leads to lower generalization performance (i.e. lower accuracy on the test data set) and a decrease in power for detecting linear pattern effects (STEP 5).

[2]　Note that, in contrast to MANOVA, the specificity of FLDA is not dependent on the assumption of multivariate normality of the residuals because classification algorithms use independent data sets for training and testing. Strong violations of multivariate normality will affect sensitivity, but not specificity, so a test of pattern information is valid.

[3]　Intuitively, we would like to weight each voxel by how well its activity discriminates the two conditions. This could be achieved by using the *t*-values for the contrast between these two conditions (A-B) as weights. This means that a voxel responding more to condition A than B (positive *t*-value) will be given a positive weight, and a voxel responding more to condition B than A (negative *t*-value) will be given a negative weight. A voxel that responds similarly to A and B will be given a weight close to zero. The methods for voxel weighting shown in Figure 2b and c are mathematically more complex, but conceptually similar to using contrast *t*-values as voxel weights.

linear discriminant dimension, which is a line in multivariate space. (These orthogonal projections are denoted by dashed lines in Figure 2B and C.) The weight vector points in the direction of the discriminant dimension, i.e. orthogonal to the decision boundary. We can apply a decision threshold to the weighted sums for all patterns so as to classify the patterns with the greatest accuracy. The threshold defines the shift of the decision boundary to the best location (Figure 2).

For the minimum-distance classifier, w is the difference between the centroids. For FLDA, w is the weight vector that maximizes the ratio of between-condition and within-condition variances (this constitutes an alternative but equivalent definition of FLDA to the one given above). For the linear SVM, *w* depends on the support vectors as determined by the training algorithm.

None of these methods is superior in general. Minimum-distance classification is expected to perform better than FLDA when its assumption of isotropic distributions is actually true. FLDA is expected to perform better than linear SVM when the data are actually multivariate normal or approximately so. Actual performance will crucially depend on the amount of data available, with limited amounts of data and greater numbers of voxels favoring simpler classification methods. Minimum-distance classification is the most conceptually simple, statistically stable, and computationally efficient method. FLDA is sensitive to the covariance structure of the data, but requires more data to capitalize on this advantage. FLDA also requires slightly more computation. Compared to linear SVM, FLDA is more computationally efficient and arguably more straightforward, conceptually as well as mathematically. However, linear SVM handles limited data in high-dimensional spaces naturally and gracefully, whereas FLDA might require a regularized covariance estimate (Ledoit and Wolf, 2003).

PATTERN-INFORMATION ANALYSIS: STEP-BY-STEP

In this section, we provide a step-by-step description of the methods for extracting patterns of activity from fMRI data and for analyzing these patterns. These steps are summarized in Figure 3.

STEP 1: Data splitting and preprocessing

Before analysis, the data should be split into an independent training and test set to ensure unbiased testing results. The training data set should be used for voxel selection (STEP 3) and classifier training (STEP 4). Both of these steps involve voxel weighting, either binary (voxel selection) or continuous (classifier training). Voxel weighting can bias testing results if performed on the same data and therefore it is crucial to use an independent data set for classifier testing (STEP 5). To make sure the data are independent, the two sets should be based on different scanner runs (e.g. even and odd runs) that use independent stimulus sequences. One option is to split the data into two halves. However, the training data set is generally chosen to be larger than the

test set in order to obtain stable voxel weights. Efficient use of the data can be achieved by cross-validation: divide the data into a number of independent subsets (e.g. single runs in your experiment), use all but one subset as training data and use the left out subset as test data; then repeat this procedure until each subset has been used as test data once. Performance on the different subsets is combined to obtain overall classifier performance. Ideally, preprocessing should be performed separately for training and test data sets so as to avoid introducing dependencies between the data sets. Preprocessing steps are the same as in activation-based analysis (i.e. slice-scan-time correction, motion correction, trend removal). In order to preserve fine-grained pattern information, spatial smoothing of the data should be omitted or strongly reduced.

STEP 2: Estimating the single-subject activity patterns

Previous studies have used several methods to estimate single-subject activity patterns. For block designs or slow event-related designs, where BOLD responses to different conditions do not overlap in time, it is possible to stay close to the raw data and use single-volume signal intensity values (Polyn et al., 2005) or temporally averaged normalized signal intensity values as patterns of activity (e.g. Kamitani and Tong, 2005). Single-subject patterns can also be estimated by univariate analysis at each voxel using the general linear model (GLM) (Friston et al., 1994, 1995a,b; Worsley and Friston, 1995). This is useful, in particular, for rapid event-related designs (e.g. Kriegeskorte et al., 2007, 2008a, 2008b) because of the hemodynamic response overlap, but has also been used in combination with block designs (e.g. Haxby et al., 2001). An advantage of using the GLM is the possibility to include motion and trend predictors in the model in order to obtain better estimates. Each condition or each example belonging to a condition (if estimating the shape of the response distribution) is entered as a predictor in the model. This part of the analysis is identical to activation-based analysis and will yield a beta-value for each predictor and voxel. The beta-values for one predictor across voxels form the pattern of activity for a specific condition. Pattern estimation yields a set of training patterns and a set of test patterns. In order to preserve fine-grained subject-specific information, the patterns should not be averaged across subjects. Therefore, pattern-information analysis is performed in native subject space. Group analysis can be performed as a second-level analysis based on pattern-information ROI estimates or pattern-information maps (Kriegeskorte et al., 2006, 2007).

STEP 3: Selecting the voxels

Once activity values are computed, the next step is to decide which voxels to include for pattern-information analysis. These voxels are selected using the training data set or another data set independent from the test set (e.g. anatomical data or functional data from a separate block-localizer
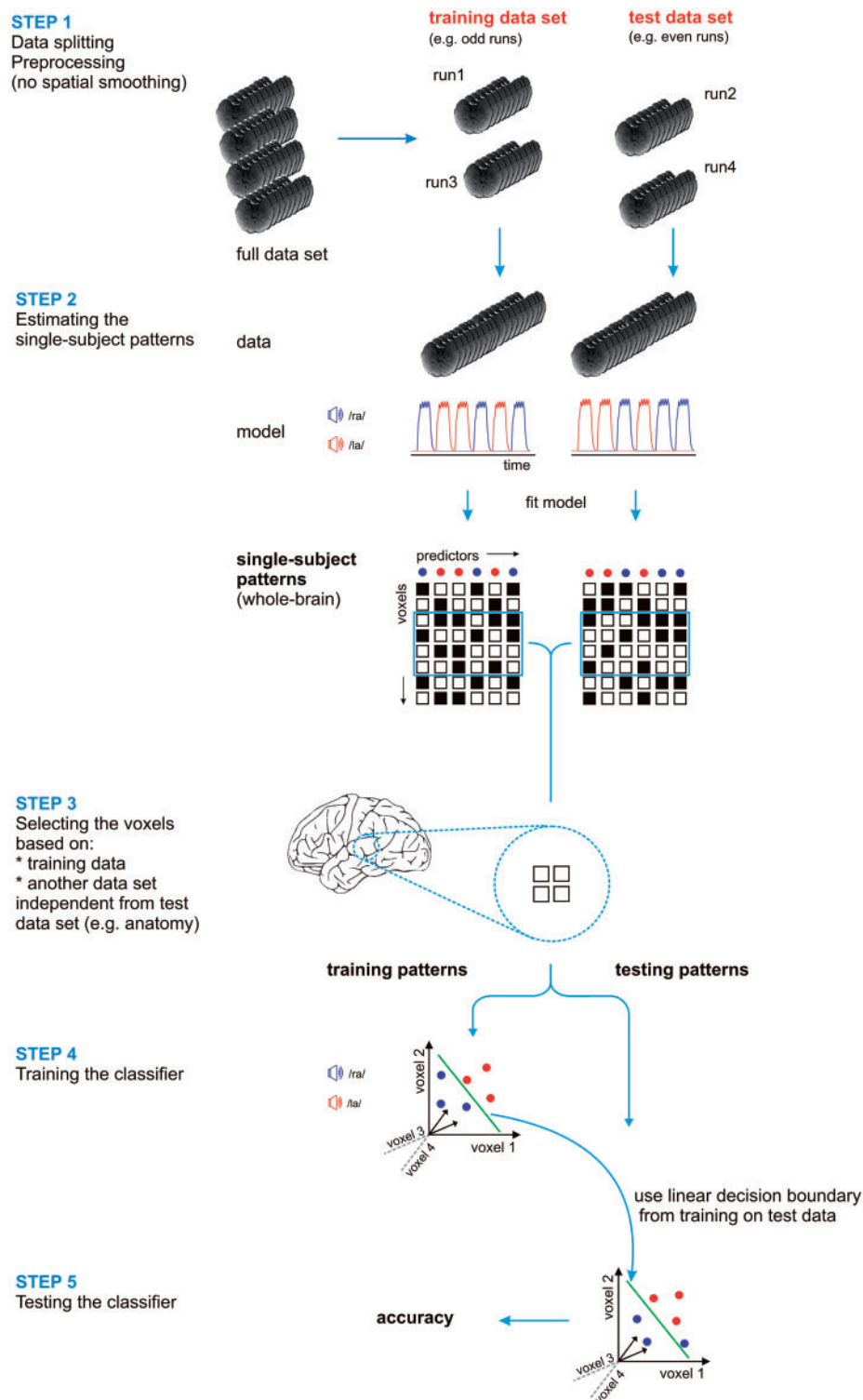
Fig. 3 Pattern-information analysis: step-by-step. Schematic illustration of the five steps of pattern-information analysis as described in the text. First, data are split into a training and a test data set and preprocessed separately. Then, single-subject patterns of activity are estimated from the data using univariate analysis (GLM) at each voxel. This results in whole-brain activity patterns consisting of beta-estimates. Black boxes indicate activated voxels; white boxes indicate nonactivated voxels. Note that activity levels are continuous in analysis and only stated as binary here for simplicity. There will be as many patterns as there are predictors (conditions) in the model. Pattern-estimation is done separately for the training and test data set. The third step consists of selecting voxels for pattern-information analysis. This can be done based on anatomy, function or both. For simplicity, the shown example region consists of four voxels only. Voxel selection should be based on the training data set or another data set that is independent from the test data set in order to prevent biased testing results. This also applies to STEP 4: voxel weighting should be performed on the training data set to prevent biased testing results. Voxels are weighted in order to maximize discriminability of the patterns belonging to the two conditions. The voxel weights computed in STEP 4 can then be tested on the test data set in STEP 5. If the weights capture true differences between the two conditions, good performance (classification accuracy) on the training data set will generalize to the test data set. Performance significantly better than chance indicates that the ROI contains information about the experimental conditions, i.e. the representational content of the region differs across conditions. The image for STEP 3 has been adapted with permission from Raizada et al. (2008).

experiment). One option would be to analyze the patterns of activity in a specific ROI. If defined by activation-based analysis, ROIs will be spatially contiguous sets of voxels, but they do not have to be. For example, to investigate object-category discrimination, the most visually responsive voxels in object-selective cortex could be selected for subsequent analysis, irrespective of whether these voxels are adjacent or not. A computationally more demanding option would be to analyze the pattern of activity across all brain voxels. This might increase informational content, but it will definitely also add substantial amounts of noise. Typically there will a decrease in performance as the number of voxels becomes very large. Possible solutions include selecting fewer voxels and transforming the original voxel space into a lower dimensional space (dimensionality reduction). Voxels can also be selected using information-based brain mapping (Kriegeskorte et al., 2006, 2007). This can be seen as the multivariate equivalent of univariate statistical parametric mapping (SPM) (Friston et al., 1995b).

STEP 4: Training the classifier

To investigate whether a region's pattern of activity discriminates two conditions, we first use the training data set to determine a set of weights (one for each voxel) that linearly combines the voxel responses in such a way as to maximize the difference between the two conditions (classifier training). We described three different linear classifiers that can be used for pattern-information analysis: the minimum-distance classifier, FLDA, and linear SVM. These may differ in sensitivity, depending on factors including the brain region, experimental events, the amount of data available, and the number of voxels in the ROI. Any of the three methods can provide a valid test of pattern-information.[4]

Most classifiers can also be trained on data from multi-condition experiments (Pereira et al., 2008). However, multi-class discriminations are often approached as a combination of multiple two-class discriminations. This approach is motivated by the fact that two-class discriminations are generally of neuroscientific interest, even if an experiment contains more than two conditions. For a detailed overview on using linear classification algorithms in neuroimaging, and their mathematical descriptions, see Pereira et al. (2008). Several pattern analysis toolboxes are listed in the reference section of this paper.

STEP 5: Testing the classifier

The weights computed during training are set to yield optimal classification performance on the training data set. To test whether good classification performance generalizes (i.e. is not based largely on noise present in the training data set), the weights are applied to an independent test data set. Performance of the classifier on the test data set can be measured by percent correct classification (accuracy). The

null hypothesis is that the classifier performs at chance level. To test whether classification accuracy is significantly better than chance, we can use a chi-square test (or a Monte-Carlo method in case of few observations). If the statistical test shows a significant result, this indicates that the region's response contains information about the experimental conditions.[5] Another way to test the classifier is to perform a univariate t-test on the projected test patterns (Kriegeskorte et al., 2007). As described above, projection (voxel weighting) converts the activity patterns into one-dimensional values. These values can then be analyzed by a conventional univariate *t*-test. Similar to a classification accuracy that is significantly better than chance, a significant *t*-value for the difference between the two conditions would indicate that the region's response contains information about the experimental conditions.

CONCLUSION

Pattern-information analysis investigates the representational content of a region by analyzing the information carried by a region's pattern of activity. This information would not be detected by conventional activation-based analysis and can significantly contribute to our understanding of neural representations of mental content. In combination with high-resolution fMRI, pattern-information analysis can detect fine-grained activity-pattern information. The most popular method is linear classification, which analyzes a region's activity patterns by means of a weighted sum of the single-voxel responses, with the weights chosen to maximally discriminate different conditions. Statistical inference is performed on a data set independent of that used for ROI definition and voxel weighting so as to prevent statistical circularity.

The conceptual appeal of pattern-information fMRI is that it allows us to "look into" the regions and investigate their representational content. Recent neuroscientific successes in the domain of sensation and perception suggest that higher-order cognitive functions in the domain of social and cognitive neuroscience might also benefit from the pattern-information approach.

Pattern-information analysis toolboxes

AFNI 3dsvm plug-in (http://www.cpu.bcm.edu/laconte/3dsvm.html)

Princeton MVPA toolbox (http://www.csbmb.princeton.edu/mvpa/)

PyMVPA toolbox (http://pkg-exppsy.alioth.debian.org/pymvpa/)

LIBSVM toolbox (http://www.csie.ntu.edu.tw/~cjlin/libsvm)

---

4   If more than one method is used, all results should be reported. (Picking the significant result among different analyses would require correction for multiple comparisons.)

5   In addition to the overall accuracy, we can examine the frequency of all four possible classifier outcomes (true/false positives, true/false negatives). This is important, in particular, when the frequencies of the two conditions are not equal.

REFERENCES

Aguirre, G.K. (2007). Continuous carry-over designs for fMRI. *Neuroimage*, *35*, 1480–94.

Bandettini, P.A., Cox, R.W. (2000). Event-related fMRI contrast when using constant interstimulus interval: Theory and experiment. *Magnetic Resonance in Medicine*, *43*, 540–8.

Birn, R.M., Cox, R.W., Bandettini, P.A. (2002). Detection versus estimation in event-related fMRI: Choosing the optimal stimulus timing. *Neuroimage*, *15*, 252–64.

Boynton, G.M., Engel, S.A., Glover, G.H., Heeger, D.J. (1996). Linear systems analysis of functional magnetic resonance imaging in human V1. *The Journal of Neuroscience*, *16*, 4207–21.

Carlson, T.A., Schrater, P., He, S. (2003). Patterns of activity in the categorical representations of objects. *Journal of Cognitive Neuroscience*, *15*, 704–17.

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*, 261–70.

Duda, R.O., Hart, P.E., Stork, D.G. (2001). *Pattern Classification*. New York, NY: John Wiley and Sons.

Friston, K., Chu, C., Mourao-Miranda, J., Hulme, O., Rees, G., Penny, W., Ashburner, J. (2008). Bayesian decoding of brain images. *Neuroimage*, *39*, 181–205.

Friston, K.J., Holmes, A.P., Poline, J.-B., Grasby, P.J., Williams, S.C.R., Frackowiak, R.S.J., Turner, R. (1995a). Analysis of fMRI time-series revisited. *Neuroimage*, *2*, 45–3.

Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.-P., Frith, C.D., Frackowiak, R.S.J. (1995b). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, *2*, 189–210.

Friston, K.J., Jezzard, P., Turner, R. (1994). Analysis of functional MRI time-series. *Human Brain Mapping*, *1*, 153–71.

Grill-Spector, K., Malach, R. (2001). fMRI-adaptation: a tool for studying the functional properties of human cortical neurons. *Acta Psychologica*, *107*, 293–321.

Hanson, S.J., Matsuka, T., Haxby, J.V. (2004). Combinatorial codes in ventral temporal lobe for object recognition: Haxby (2001) revisited: is there a ''face'' area? *Neuroimage*, *23*, 156–66.

Haxby, J.V., Gobbini, M.I., Fury, M., Ishai, A., Schouten, J.L., Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*, 2425–30.

Haynes, J.-D., Rees, G. (2005). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature Neuroscience*, *8*, 686–91.

Haynes, J.-D., Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, *7*, 523–34.

Haynes, J.-D., Sakai, K., Rees, , G. Gilbert, S., Frith, C., Passingham, R.E. (2007). Reading hidden intentions in the human brain. *Current Biology*, *17*, 323–8.

Kamitani, Y., Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, *8*, 679–85.

Krekelberg, B., Boynton, G.M., Van Wezel, R.J.A. (2006). Adaptation: from single cells to BOLD signals. *Trends in Neurosciences*, *29*, 250–6.

Kriegeskorte, N., Bandettini, P. (2007a). Analyzing for information, not activation, to exploit high-resolution fMRI. *Neuroimage*, *38*, 649–62.

Kriegeskorte, N., Bandettini, P. (2007b). Combining the tools: Activation- and information-based fMRI analysis. *Neuroimage*, *38*, 666–8.

Kriegeskorte, N., Formisano, E., Sorger, B., Goebel, R. (2007). Individual faces elicit distinct response patterns in human anterior temporal cortex. *Proceedings of the National Academy of Sciences USA*, *104*, 20600–5.

Kriegeskorte, N., Goebel, R., Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences USA*, *103*, 3863–8.

Kriegeskorte, N., Mur, M., Bandettini, P.A. (2008b). Representational similarity analysis – connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* doi:10.3389/neuro.06.004.2008.

Kriegeskorte, N., Mur, M., Ruff, D.A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., Bandettini, P.A. (2008a). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* in press.

Ku, S.P., Gretton, 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. *Magnetic Resonance Imaging*, *26*, 1007–14.

LaConte, S., Strother, S., Cherkassky, V., Anderson, J., Hu, X. (2005). Support vector machines for temporal classification of block design fMRI data. *Neuroimage*, *26*, 317–29.

Ledoit, O., Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, *10*, 603–21.

Logothetis, N.K. (2008). What we can do and what we cannot do with fMRI. *Nature Reviews*, *453*, 869–78.

Logothetis, N.K., Pauls, J., Augath, M., Trinath, T., Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. *Nature*, *412*, 150–7.

Mitchell, T.M., Shinkareva, S.V., Carlson, A., Chang, K.-M., Malave, V.L., Mason, R.A., Just, M.A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, *320*, 1191–5.

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 fMRI data. *Neuroimage*, *28*, 980–95.

Norman, K.A., Polyn, S.M., Detre, G.J., Haxby, J.V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*, 424–30.

O'Toole, A.J., Jiang, F., Abdi, H., Haxby, J.V. (2005). Partially distributed representations of objects and faces in ventral temporal cortex. *Journal of Cognitive Neuroscience*, *17*, 580–90.

Pereira, F., Mitchell, T., Botvinick, M. (2008). Machine learning classifiers and fMRI: a tutorial overview. *Neuroimage*, manuscript submitted (in press).

Polyn, S.M., Natu, V.S., Cohen, J.D., Norman, K.A. (2005). Category-specific cortical activity precedes retrieval during memory search. *Science*, *310*, 1963–6.

Raizada, R.D.S., Tsao, F.M., Liu, H.M., Kuhl, P.K. (2008). Quantifying the adequacy of neural representations for a cross-language phonetic discrimination task: prediction of individual differences. *Cerebral Cortex*, manuscript submitted.

Sawamura, H., Orban, GA., Vogels, R. (2006). Selectivity of neuronal adaptation does not match response selectivity: a single-cell study of the fMRI adaptation paradigm. *Neuron*, *49*, 307–18.

Strother, S.C., Anderson, J., Hansen, L.K., Kjems, U., Kustra, R., Sidtis, J., Frutiger, S., Muley, S., LaConte, S., Rottenberg, D. (2002). The quantitative evaluation of functional neuroimaging experiments: The NPAIRS data analysis framework. *Neuroimage*, *15*, 747–71.

Tolias, AS., Keliris, GA., Smirnkakis, SM., Logothetis, NK. (2005). Neurons in macaque area V4 acquire directional tuning after adaptation to motion stimuli. *Nature Neuroscience*, *8*, 591–3.

Worsley, K.J., Friston, K.J. (1995). Analysis of fMRI time-series revisited – again. *Neuroimage*, *2*, 173–81.

Worsley, K.J., Poline, J.-B., Friston, K.J., Evans, A.C. (1997). Characterizing the response of PET and fMRI data using multivariate linear models. *Neuroimage*, *6*, 305–19.