Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns

NeuroImage 43 (2008) 44–58

Contents lists available at ScienceDirect

NeuroImage

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

Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns

Federico De Martino [a,*], Giancarlo Valente [a], Noël Staeren [a], John Ashburner [b], Rainer Goebel [a], Elia Formisano [a]

[a] Department of Cognitive Neurosciences, Faculty of Psychology, University of Maastricht, Maastricht, Postbus 616, 6200 MD, Maastricht, The Netherlands
[b] Wellcome Trust Centre for Neuroimaging, University College London, UK

ARTICLE INFO

Article history:
Received 26 September 2007
Revised 8 May 2008
Accepted 26 June 2008
Available online 11 July 2008

ABSTRACT

In functional brain mapping, pattern recognition methods allow detecting multivoxel patterns of brain activation which are informative with respect to a subject's perceptual or cognitive state. The sensitivity of these methods, however, is greatly reduced when the proportion of voxels that convey the discriminative information is small compared to the total number of measured voxels. To reduce this dimensionality problem, previous studies employed univariate voxel selection or region-of-interest-based strategies as a preceding step to the application of machine learning algorithms.

Here we employ a strategy for classifying functional imaging data based on a multivariate feature selection algorithm, Recursive Feature Elimination (RFE) that uses the training algorithm (support vector machine) recursively to eliminate irrelevant voxels and estimate informative spatial patterns. Generalization performances on test data increases while features/voxels are pruned based on their discrimination ability. In this article we evaluate RFE in terms of sensitivity of discriminative maps (Receiver Operative Characteristic analysis) and generalization performances and compare it to previously used univariate voxel selection strategies based on activation and discrimination measures. Using simulated fMRI data, we show that the recursive approach is suitable for mapping discriminative patterns and that the combination of an initial univariate activation-based (*F*-test) reduction of voxels and multivariate recursive feature elimination produces the best results, especially when differences between conditions have a low contrast-to-noise ratio.

Furthermore, we apply our method to high resolution ($2 \times 2 \times 2\text{mm}^3$) data from an auditory fMRI experiment in which subjects were stimulated with sounds from four different categories. With these real data, our recursive algorithm proves able to detect and accurately classify multivoxel spatial patterns, highlighting the role of the superior temporal gyrus in encoding the information of sound categories. In line with the simulation results, our method outperforms univariate statistical analysis and statistical learning without feature selection.

Introduction

Machine learning and pattern recognition techniques are being increasingly used in fMRI data analysis. These methods allow detecting subtle, non-strictly localized effects that may remain invisible to the conventional analysis with univariate statistics (Haxby et al., 2001, Norman et al., 2006, Haynes and Rees, 2006). In contrast to these latter approaches, machine learning techniques take into account the full spatial pattern of brain activity, measured simultaneously at many locations, and exploit the inherent multivariate nature of fMRI data.

The application of machine learning techniques to fMRI has been referred to as multivoxel pattern analysis (MVPA) and it generally entails four steps (Norman et al., 2006). First, the set of voxels that will enter the multivariate analysis is selected. With respect to this, the analysis may be *massively* multivariate and consider all brain voxels simultaneously (whole-brain approach, Mourao-Miranda et al., 2005) or may be limited to a subset of voxels from one region-of-interest (ROI) (Cox and Savoy, 2003, Haynes and Rees, 2005, Kamitani and Tong, 2005), in which case the dimensionality of the multivariate space is greatly reduced. Second, stimulus-evoked brain activity is represented as a point in a multidimensional space, i.e. as the pattern of intensity values at selected voxels (multivoxel patterns, MVP). In order to represent the brain response to a stimulus or cognitive state any estimate of activation at the selected voxels can be used, such as the intensity at a single acquisition volume (TR) (Haynes and Rees, 2005, Mourao-Miranda et al., 2005) or the average intensity in multiple TRs (Kamitani and Tong, 2005, Mourao-Miranda et al., 2006). Third, using a subset of trials, a classifier is trained and the optimal separating boundary (hypersurface) between different conditions in this multidimensional space is defined. Several methods including

* Corresponding author. Fax: +31 43 3884125.
E-mail address: f.demartino@psychology.unimaas.nl (F. De Martino).

1053-8119/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2008.06.037

Support Vector Machines (SVMs) (Cox and Savoy, 2003, Mitchell et al., 2004, Mourao-Miranda et al., 2005, LaConte et al., 2005, Kamitani and Tong, 2005), linear disciminant analysis (LDA) (O'Toole et al., 2005, Haynes and Rees, 2005, Kriegeskorte et al., 2006) Gaussian Naïve Bayes (GNB) (Mitchell et al., 2004) and Neural Networks (Hanson et al., 2004) classifiers have been used for this purpose. During training, a map coding for the relative contribution of each voxel to the discrimination of conditions (discriminative map) can be directly obtained for all linear classifiers (Mourao-Miranda et al., 2005). Fourth, the capability of the trained classifier to accurately discriminate the experimental conditions when presented with new data (i.e. trial responses not used during training) is tested (generalization).

This article deals with issues concerning the first point, i.e. the initial selection of the set of voxels, with the aim of optimizing the performance of the multivoxel pattern analysis. For consistency with the pattern recognition literature the voxels of an fMRI data set are also referred to as "features".

Whole-brain approaches are appealing in that they do not require *a priori* hypothesis on the location of the relevant voxels, which can be determined *post-hoc* from the *discriminative maps*. These approaches seem most appropriate when the discrimination of perceptual or cognitive states is reflected by widely distributed activation patterns that extend and include various and separated brain regions. However, whole-brain approaches may be problematic when the aim of the analysis is the fine-grained discrimination between perceptual states (Haynes and Rees, 2005, Kamitani and Tong, 2005). In fact, in these cases the proportion of voxels that convey the discriminative information is expected to be small and thus whole-brain approaches seem sub-optimal. Machine learning algorithms are known to degrade their performances when faced with many irrelevant features (overfitting, Kohavi and John, 1997, Guyon and Elisseef, 2003, Norman et al., 2006), especially, when the number of training samples is rather limited as in typical fMRI studies. Thus selection of an adequate subset of features/voxels is of critical importance in order to obtain classifiers with good generalization performance.

Restricting the multivariate analysis to an anatomically or functionally pre-defined subset of voxels can be seen as a solution to this feature selection problem. This solution is affected by all limitations of ROI-based approaches, which only allow testing a limited set of spatial hypotheses and cannot be used when the aim of the study is the localization of those voxels forming discriminative patterns. An interesting alternative is the local multivariate search approach proposed by Kriegeskorte et al., (2006). This method relies on the assumption that the discriminative information is encoded in neighboring voxels within a "searchlight" of specified radius. Such locally-distributed analysis might be, however, sub-optimal when no hypothesis is available on the size of the neighborhood and might fail to detect discriminative patterns jointly encoded by distant regions (e.g. bilateral activation patterns).

The main limitation of whole-brain MVPA is its computational complexity since the number of voxels is very large in comparison to the number of trials in a typical fMRI acquisition (Norman et al., 2006). In pattern recognition approaches, feature selection strategies are usually employed prior to the analysis in order to reduce the dimensionality and to preserve sensitivity to small effects. In previous neuroimaging applications, machine learning algorithms have been combined with univariate feature selection strategies (Mitchell et al., 2004, Mourao-Miranda et al., 2006). Both the activation level (*F*-test) or the discrimination ability (*t*-test) have been used as univariate ranking criteria for voxel selection. Any such method of voxel selection, though, does not consider the inherent multivariate nature of the fMRI data.

Multivariate feature selection strategies can be summarized in three categories, multivariate filters, wrappers and embedded methods (for a review see Kohavi and John, 1997 and Guyon and Elisseef, 2003). Filter methods are applied previous to the classifier and thus do not make use of the classifier performance to evaluate the feature subset. Wrappers and embedded methods, on the other end, use the classifier to find the best feature subset. Wrappers consider any classifier as a black box and make use of different search engines in the feature space to find the subset that maximizes generalization performances. Embedded methods instead incorporate feature selection as a part of the training process.

Here we consider an approach to fMRI MVPA that ensures high sensitivity to fine-grained spatial discriminative patterns, while preserving the appealing properties of whole-brain analysis. This approach combines a wrapper method (Recursive Feature Elimination) and SVMs to perform fMRI MVPA. Recursive Feature Elimination (RFE) has been compared to other multivariate feature selection strategies (Rakotomamonjy, 2003) and has been successfully applied to gene selection and sample classification in combination with Support Vector Machine classifiers (Guyon et al., 2002). The recursive nature of the algorithm makes RFE computationally feasible in fMRI MVPA where the number of features can reach 300,000 cortical voxels, as in the case of whole-brain high resolution ($2 \times 2 \times 2\text{mm}^3$) acquisitions. In a recent publication Hanson and Halchenko, (2008) introduced the combination of RFE and SVMs to fMRI multivoxel pattern recognition analysis. The authors showed that removing iteratively irrelevant voxels improves generalization performances in discriminating visual stimuli (Faces and Houses) during two different tasks (1-back recognition detection task, oddball).

In the present article, we evaluate and compare the performances of RFE, of different univariate feature filter methods (activation- and discrimination-based), and of their combination on simulated fMRI data. For each method, sensitivity analysis (ROC analysis) and generalization performance are computed at different levels of functional signal-to-noise (SNR, activation level with respect to the baseline) and functional contrast-to-noise (CNR, differences between activation levels in two conditions). We show that, especially in the case of low SNR and/or CNR, the combination of univariate activation-based voxel reduction and RFE outperforms all other methods.

We also apply our method to a real data set obtained in an experiment with auditory stimulation in which sounds from four different categories were presented. The results of the analysis on this fMRI data set confirm the expectations from the simulations and show that the combination of activation-based univariate feature selection and RFE provides the highest generalization performance.

Methods

General description of the approach

Fig. 1 illustrates schematically the proposed approach, which consists of an *N*-fold cross-validation procedure with two nested cycles. At each fold k ($k = 1,\ldots,N$), trials from the fMRI time series are divided into training (Train_k) and test (Test_k) set, with the latter only used to assess generalization performance. The training trials (Train_k) are further partitioned in several (*RL*) splits. For each of the splits, an SVM classifier is trained and discriminative weights are calculated on a subset of trials ($\text{Train}_{\text{RFEi}}$). Multivariate feature selection using Recursive Feature Elimination (RFE, red dashed box in Fig. 1) is performed R times based on the ranking of the average absolute discriminative weights of L consecutive trainings. At each feature selection, voxels corresponding to the smallest ranking are discarded; voxels with the highest discriminative values are used for training in the next iteration. Generalization performances corresponding to the current feature selection level are assessed using the external test trials (Test_k). Note that these trials do not enter the training and thus influence neither the decision boundary nor the discriminative weights nor the recursive selection. Final generalization performances and discriminative maps of each RFE level are obtained as the average over the N folds.

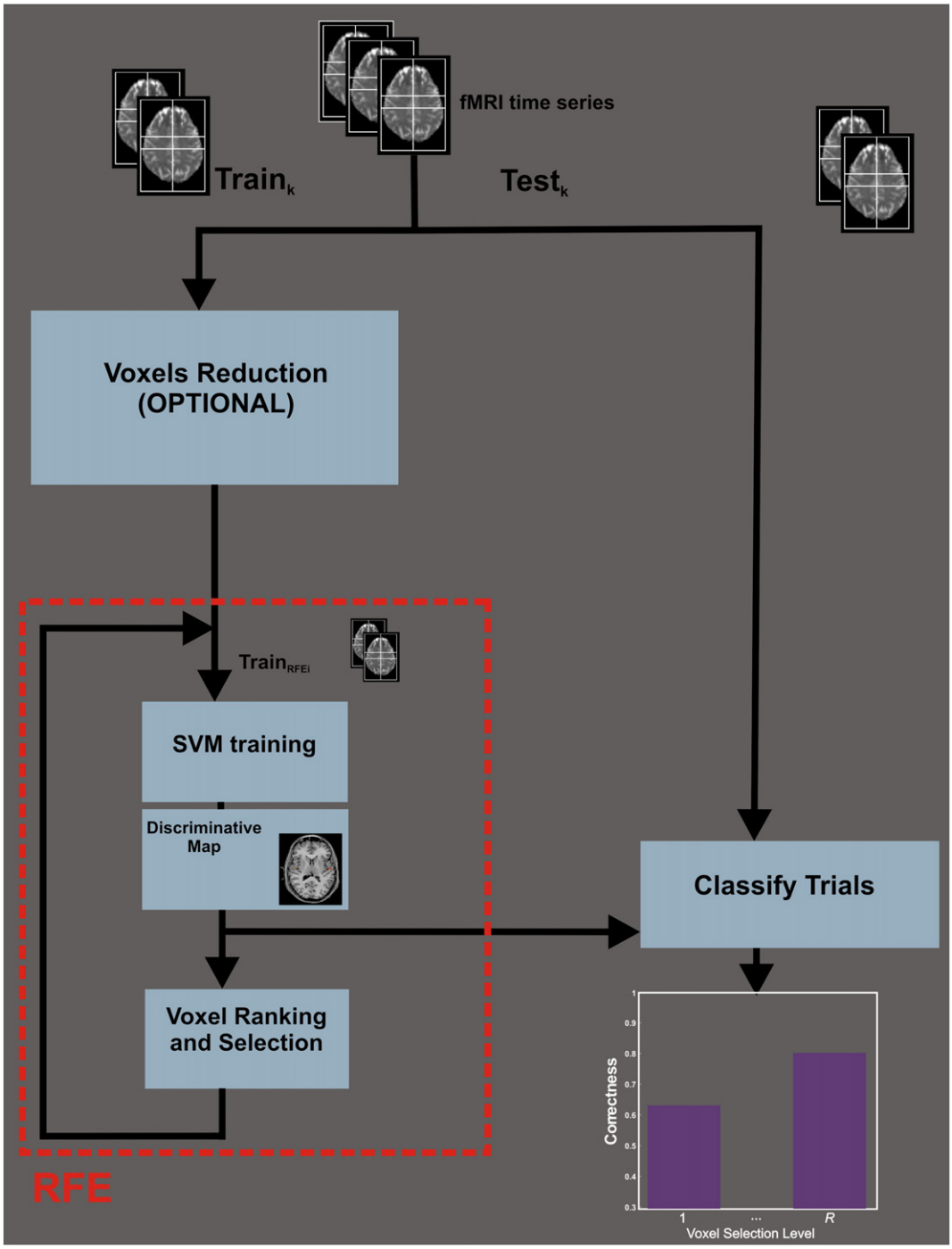

Fig. 1. General description of the proposed SVM/RFE iterative procedure to brain mapping. After single trial-response estimation functional time series are divided in training and test data sets (Train$_k$; Test$_k$). An optional step of voxel reduction can be performed prior to RFE using only the training data (Train$_k$). For each voxel selection level the recursive procedure (RFE; red dashed box in figure) consists of two steps. First an SVM classifier is trained on a subset of the training data (Train$_{RFEi}$) using the current set of voxels. Second a set of voxels is discarded according to their discriminative weights as estimated during training. Test data (Test$_k$) are classified at each iteration and generalization performances are assessed.

Fig. 1 also indicates that prior to RFE an additional, preliminary step of univariate voxel reduction can be used. This step consists in selecting a subset of voxel, based on univariate statistics computed on the training data (Train$_k$).

Single trial-response estimation

We estimated the multivoxel pattern of intensities forming the input to the SVM classifier in the following way. At each stimulus presentation, a trial $t(t = 1,\ldots,T)$ is formed considering N_{pre} and N_{post} temporal samples (before and after stimulus onset respectively) of the pre-processed (see below) time course of activity. A trial estimate of the response at every voxel $v(v = 1,\ldots,V)$ is then obtained by fitting a General Linear Model (GLM) with one predictor coding for the trial response and one linear predictor accounting for a within-trial linear trend. The trial-response predictor is obtained by convolution of a boxcar with a double-gamma hemodynamic response function (HRF) (Friston et al., 1998). At every voxel, the

corresponding regressor coefficient (beta) is taken to represent the trial response.

To account for BOLD response variability we repeated the estimation procedure (at each voxel) changing the time-to-peak (4 to 6s) of the modeled HRF response. The best-fitting beta (minimum p-value) was selected as representative for the trial response.

The outlined procedure is designed for slow event related designs or block designs in which the responses to contiguous trials are not overlapping in time (see below for a discussion of rapid event related designs). The result is a matrix $\mathbf{M}$ ($T \times V$), whose element $m_{t,v}$ is the response estimate at trial t and voxel v. This matrix is partitioned in training and testing matrices ($\mathbf{M}_{\text{train}}$ and $\mathbf{M}_{\text{test}}$) which are used in the rest of the analysis.

Activation- and discrimination-based univariate feature selection

In previous fMRI applications of pattern recognition methods (Mitchell et al., 2004, Haynes and Rees, 2005, Mourao-Miranda et al., 2006), univariate feature selection strategies have been suggested for reducing the dimensionality of the multivoxel space (i.e. number of columns in matrix M).

Consider a training set defined as:

$$\{\mathbf{m}_i, c_i\}(i = 1, \ldots, T_{\text{train}}); c_i \in \{-1, +1\}; \mathbf{m}_i \in \Re^v, \tag{1}$$

where $\mathbf{m}_i$ is one row of matrix $\mathbf{M}_{\text{train}}$ and represents a trial in the V dimensional space of the voxels, whose class c_i is known (e.g. the two stimulus conditions).

Introducing the hypothesis that interesting patterns consist of voxels that show a significant stimulus-related BOLD response to any of the two conditions compared to baseline levels justifies the reduction of the number of features based on the univariate selection of 'active' voxels. Furthermore, it simplifies the interpretation of the results as the analysis is restricted to voxels showing neurophysiologically understood responses (but see Haynes et al., 2007 and Discussion).

From the values c_i and $m_{i,v}$ (v=1,…,V) we can compute the following scoring functions:

$$S_{A,+1}(v) = \frac{\overline{m_{+1,v}}}{\sqrt{\frac{\sigma^2_{+1,v}}{n_{+1}}}},\ S_{A,+1}(v) = \frac{\overline{m_{-1,v}}}{\sqrt{\frac{\sigma^2_{-1,v}}{n_{-1}}}}, \tag{2}$$

where $\overline{m_{+1,v}}, \sigma^2_{+1,v}\left(\overline{m_{-1,v}}, \sigma^2_{-1,v}\right)$ indicate an estimate of mean response and variance calculated over the n_{+1} (n_{-1}) trials of condition $c = +1$ ($c = -1$) at voxel v.

In the present paper the use of univariate activation-based ranking is twofold. First, it is a feature selection step to be compared with recursive feature elimination (univAct), in which case we used as ranking criteria the mean between $S_{A,+1}$ and $S_{A,-1}$. Second, we use activation-based ranking as a method of initial univariate feature reduction in order to select a subset of voxels on which the iterative feature selection procedure is subsequently applied (univActRed). In the latter case we sorted the voxels independently using $S_{A,+1}$ and $S_{A,-1}$ and selected the union of the first V' voxels per condition.

A more restrictive form of univariate feature selection is based on the selection of voxels that show a significant difference between the two conditions (Mitchell et al., 2004, Haynes and Rees, 2005, Mourao-Miranda et al., 2006). As measures of discrimination ability a parametric (t) or non-parametric (Wilcoxon) statistical test can be used.

From the values c_i and $m_{v,i}$ (v=1,…,V) we can compute the following scoring functions:

$$S_T(v) = \frac{\overline{m_{+1,v}} - \overline{m_{-1,v}}}{\sqrt{\frac{\sigma^2_{+1,v}}{n_{+1}} + \frac{\sigma^2_{-1,v}}{n_{-1}}}}, \tag{3}$$

$$S_W(v) = \left| \frac{\frac{R_{+1,v}}{n_{+1}}}{n_{-1}} - 1 \right|, \tag{4}$$

where $\overline{m_{+1,v}}, \sigma^2_{+1,v}$ and $R_{+1,v}(\overline{m_{-1,v}}, \sigma^2_{-1,v}, R_{-1,v})$ indicate an estimate of mean response, variance and sum or ranks calculated over the n_{+1} (n_{-1}) trials of condition $c = +1 (c = -1)$ at voxel v.

The univariate discrimination-based selection (univT; univW) is obtained sorting the voxels according to S_T or S_w and selecting the first V' voxels.

It is important to underline the necessity of performing any sort of initial feature selection (activation- or discrimination-based) only using the training data in order to reduce potential biases in the evaluation of generalization performances.

To better quantify generalization abilities of the univariately selected voxels the scoring functions are computed in cross-validation, i.e. further splitting the training data in different subsets, computing voxel-by-voxel scores on the different sub-splits and then averaging the different scores.

In order to compare univariate feature selection to multivariate feature selection implemented using RFE, we matched the number of voxels selected with the different univariate methods (univT; univW; univAct) to the number of voxels selected by RFE at the different iterations. Furthermore we evaluate the impact of an initial univariate voxel reduction based on activation (univActRed), both on multivariate (RFE) and univariate (univT; univW; univAct) feature selection methods.

Recursive feature elimination

Activation- and discrimination-based feature filtering consider each voxel independently, and thus do not take into account the intrinsic multivariate nature of the fMRI data. Wrapper methods, such as RFE, constitute a multivariate alternative to classical feature filtering and use the classifier itself to discard irrelevant features. Our implementation of RFE can be described with the following pseudo-code:

while (~ stop)

1. Train SVM$\left(\mathbf{M}_{\text{Train}_{\text{RFE}_i}}, \text{Labels}_{\text{RFE}_i}\right) i = 1, \ldots, L$
2. Compute the scoring function: $S_{\text{RFE}}(v) = \frac{\sum_{i=1}^{L} |w_i(v)|}{L}$
3. Sort V based on $S_{\text{RFE}}(v)$
4. Eliminate features with smallest scores

end

where $\text{Labels}_{\text{RFE}i}$ are the trials' classes in the training set $\text{Train}_{\text{RFE}i}$, and $w_i(v)$ is the discriminative weight for voxel v as obtained from the SVM training (see below). Before feature selection the SVM is trained multiple times on different training data sets ($\text{Train}_{\text{RFE}i}$; i = 1,…,L). To perform feature selection the scoring function $S_{\text{RFE}}(v)$ is used.

The backward elimination procedure used to search the multi-dimensional space needs a stopping criterion to be defined. One possible solution is to terminate the algorithm based on the generalization performances (e.g. performance drop compared to previous iteration). A more conservative choice is to proceed from the original feature set to the empty set or a set of desired dimensionality, which in cases of high dimensional feature spaces can be very time consuming (Guyon et al., 2002). When the latter is chosen as a stopping criteria the total number of iterations is controlled by the number of voxels discarded at each iteration and the best feature set is selected *post-hoc* based on the highest generalization performances.

Linear support vector machines (binary classification)

Let us consider a training set as in (1). In the general case of overlapping classes (i.e. non-linearly separable classes) the problem of

finding the optimal separating hyperplane (defined by the normal w and the distance to the origin of the multidimensional space b) that maximizes the distance to the nearest training points of the two classes is defined as:

$$\min_{w,b,\xi} J(w) = \frac{1}{2} w^T w + a \sum_{n_i}^{l} \xi_i \tag{5}$$

subject to:

$$c_i(\mathbf{w}^T \mathbf{m}_i + b) \geq 1 - \xi_i, \; i = 1, \ldots, T_{\text{train}} \tag{6}$$

and:

$$\xi_i \geq 0, \; i = 1, \ldots, T_{\text{train}} \tag{7}$$

where $\xi_i, i = 1, \ldots, T_{\text{train}}$ are slack variables that account for training errors and a is a positive real constant (Suykens et al., 2002). The solution is obtained using Lagrangian methods (Cristianini and Shawe-Taylor, 2000). Classification of new trials $\mathbf{m}_{new}$ is obtained by evaluating:

$$\text{sign}(\mathbf{w}^T \mathbf{m}_{\text{new}} + b) \tag{8}$$

An absolute discriminative map can be obtained considering the vector $|\mathbf{w}|$ (i.e. voxels that contribute the most to the discrimination of the two classes are represented by high values of $|\mathbf{w}|$).

In the present paper, we use a variant of SVM known as ls-SVM. In the classical SVM formulation of Eq. (5–7) the optimal boundary between different classes is obtained by considering only the training point falling on the separating hyperplane (i.e. support vectors). In ls-SVM each training point is weighted in order to obtain the distinguishing hypersurface (hyperplane). The optimization problem for the general case of non-separable classes is defined as:

$$\min_{w,b,e} J(\mathbf{w}) = \frac{1}{2} \mathbf{w}^T \mathbf{w} + \gamma \frac{1}{2} \sum_{i=l}^{1} e_i^2 \tag{9}$$

subject to:

$$c_i(\mathbf{w}^T \mathbf{m}_i + b) = 1 - e_i, i = 1, \ldots, T_{train} \tag{10}$$

where is γ a positive real constant (Suykens et al., 2002).

By changing the inequalities constraints (SVM, Eq. 6–7) into the equality constraints (ls-SVM, Eq. 10), training of ls-SVMs is simplified and is reduced to solving a set of linear equations instead of a quadratic programming problem. Ls-SVMs have been benchmarked on a series of typical classification problems and have been proved to outperform other classification techniques (e.g. Linear Discriminant Analysis, Quadratic Discriminant Analysis) and achieved higher or comparable classification accuracies if compared to classical SVM (Suykens et al., 2002). The link of ls-SVMs and other classification techniques such as Kernel Fischer Linear Discriminant Analysis has been described in Suykens et al. (2002).

fMRI data

Simulated time series

We simulated fMRI time series according to a design with two conditions with 30 trials per condition and each trial lasting 14,440ms

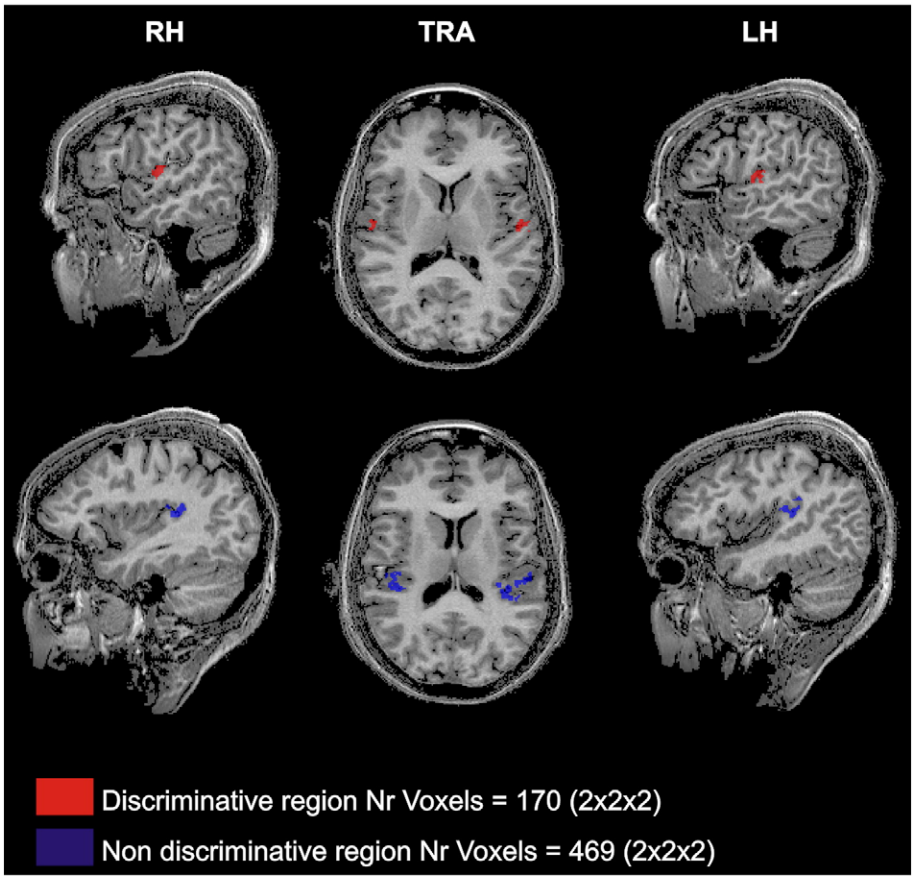

Fig. 2. The simulated ROIs projected in the volume of a subject. In red the discriminative ROIs (170 2 × 2 × 2 voxels); in blue the active but non-discriminative ROIs (469 2 × 2 × 2 voxels).

(block design). The functional time series had a simulated TR of 3610ms and functional voxel resolution of 2 × 2 × 2mm^3. These parameters were used in order to match experimental design and acquisition parameters used in the real fMRI data set also presented in this paper.

The discriminative voxels (170 in total) were confined to two realistically shaped regions (Fig. 2 top row) and belonged to one of two populations (condition1 > condition2; condition2 > condition1) whose spatial distribution was random within the regions. We also

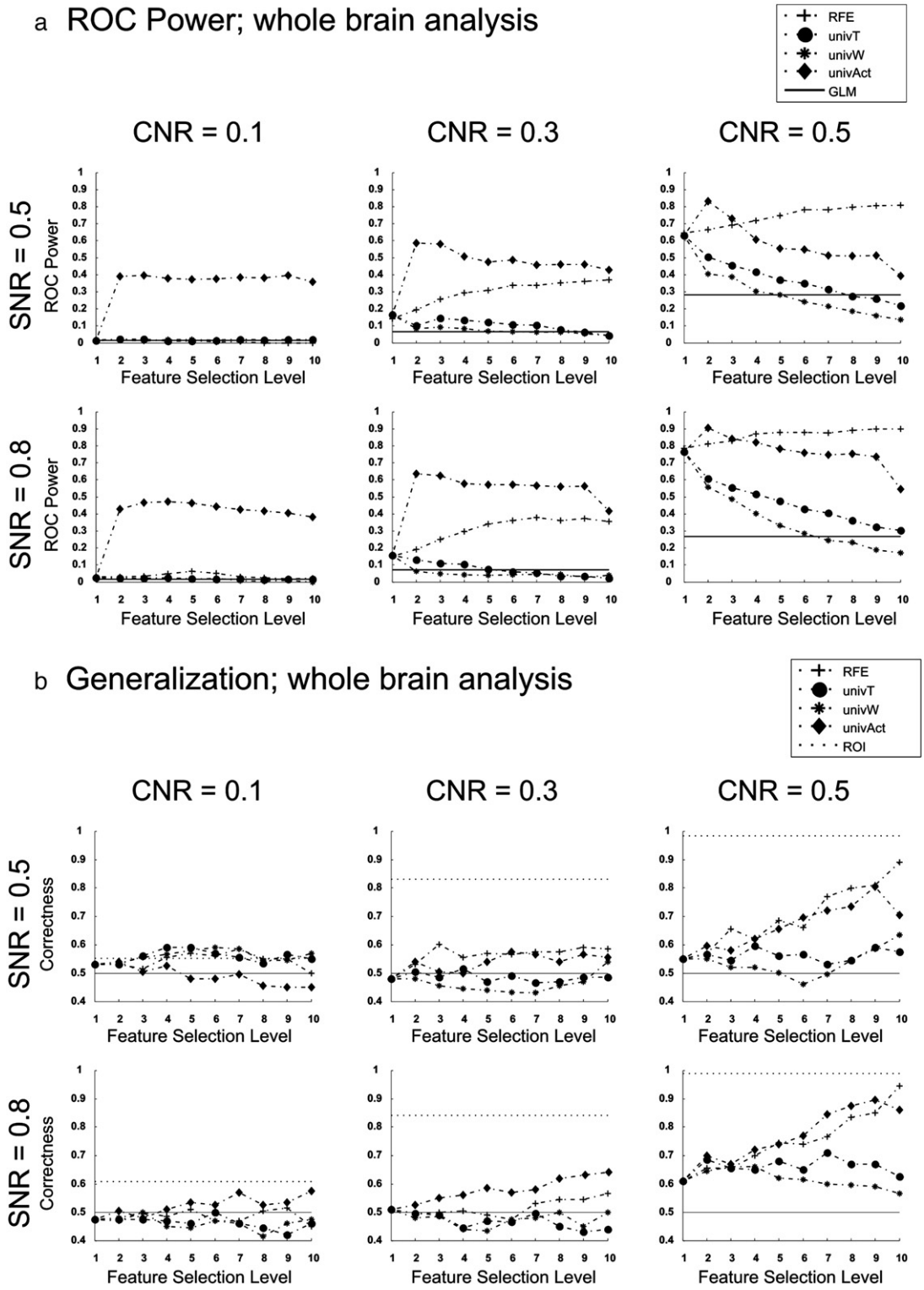

Fig. 3. Results obtained on the whole-brain analysis using different feature selection strategies at different CNR and SNR values (varBOLD fixed at 10% of the maximum response). (a) Receiver Operative Characteristic (ROC) power (defined as the area of the ROC curve in the false positive range of [0; 0.01]) of SVM based maps obtained for different numbers of selected voxels starting from the whole brain. Compared feature selection schemes are: 1) RFE; 2) univariate discrimination-based selection (univT; univW); and 3) univariate activation-based selection (univAct). The bold line represents the sensitivity analysis of conventional statistical parametric maps obtained using the GLM. (b) Generalization performances of SVM based classifier are plotted for a different number of selected voxels starting from the whole brain. Compared feature selection schemes are: 1) RFE; 2) univariate discrimination-based selection (univT; univW); and 3) univariate activation-based selection (univAct). Different methods are compared to classification obtained using only the discriminative voxels (dotted lines). The bold line represents the chance level (0.5).

simulated neighboring regions (469 voxels in total) that responded to both stimulation conditions without carrying specific discriminative information (Fig. 2 bottom row).

The simulated BOLD responses were obtained by convolving the simulated stimulus with a standard hemodynamic response function modeled using a double-gamma function (Friston et al., 1998). At each voxel the simulated activations were added to temporally autocorrelated noise obtained as:

$$R_a(t) = \rho_k R_0(t-1) + \sqrt{1-\rho_k^2 R_0(t)}, \quad (11)$$

where R_0 is random Gaussian noise and $\rho_k \sim N(0.5,0.1)$ controls the amount of autocorrelation at voxel k.

We simulated the data at the level of the original time series and not at the level of the matrix (**M**) as it allows us to examine also the influence of the trial estimation step. For each active voxel we varied the signal-to-noise ratio (i.e. the response amplitude compared to the noise standard deviation; SNR), the contrast-to-noise ratio (i.e. the response differences compared to the noise standard deviation; CNR) and the variability of the BOLD responses to trials of the same stimulus condition (varBOLD), the latter defined in terms of percent of variability compared to the maximum response. Three different levels for SNR [0.3; 0.5; 0.8], CNR [0.1; 0.3; 0.5] and varBOLD [10%; 20%; 60%] were used to produce 27 simulated fMRI data sets.

Simulated functional time series were used to test the performances of a purely multivariate feature selection strategy (RFE). RFE was compared, matching the number of selected voxels, to the performances of SVM based classification preceded by purely univariate feature selection strategies based on *t*-test or Wilcoxon (univT; univW) and univariate activation-based selection (univAct).

Furthermore using univariate activation-based ranking as initial voxels reduction strategy (i.e. selecting the union of the 2000 most active voxels for both conditions; univActRed) we evaluated its impact on multivariate feature selection (univActRed+RFE). Matching the number of selected voxels, we compared univActRed+RFE to methods in which the same initial activation-based reduction was followed by different univariate selection strategies (univActRed+univT; univActRed+univW) and to the case in which only activation-based selection was used (univActRed+univAct).

To assess generalization performances we followed an *N*-fold cross-validation scheme leaving out ten trials for each fold. The twenty remaining trials were further split fifty times (each time leaving out five trials) and an ls-SVM classifier was trained on each split. RFE was performed ten times (R=10) based on the average absolute discriminative weights (**|w|**) of five consecutive trainings (L=5). The entire procedure was repeated ten times (i.e. ten folds) changing training and testing sets. Finally, for each RFE level, generalization performances and discriminative maps were obtained as the average of levels and maps obtained in the ten folds. The best RFE level refers to the level with the highest average generalization.

For all tested feature selection methods, the number of discarded voxels at every step was computed so that the size of the final feature set equals the number of simulated discriminative features.

We quantitatively assessed the differences in sensitivity between the various methods using Receiver Operative Characteristics (ROCs) curves computed based on the absolute discriminative maps obtained from each analysis. As a figure of merit we computed the area under the curve in the false positive rate interval [0, 0.01] (Skudlarski et al., 1999, Sorenson and Wang, 1996, Fadili et al., 2000). The sensitivity of the maps obtained using the different MVPA methods was compared to a conventional univariate analysis (GLM, *t*-test) in which the entire data sets were used and a design matrix consisting of three predictors (i.e. one for each condition and one accounting for a linear trend) was fitted to the data. The resulting absolute *t*-maps (one for each simulated data set) were used to compute ROC curves and consequently the area under the curve in the false positive rate interval [0, 0.01].

Real data

We examined the performances of our approach on real data using a time series from an auditory experiment on sound categorization performed in a 3T system (Siemens Allegra). Functional runs consisted of 23 axial slices obtained with a T2-weighted gradient echo, EPI sequence (TR 3.6s; FOV256 × 256; matrix size 128 × 128, voxel size = 2 × 2 × 2mm^3). Anatomical images were obtained using a high resolution (1 × 1 × 1mm), T1-weighted sequence.

Stimuli consisted of 800ms tonal sounds of four different categories (cats, girls [singing female voices], guitars and tones). The sounds were matched not only in length and RMS power but also in the temporal profile of the fundamental frequency, such that the perceptual pitch could be considered identical across categories. Stimuli were presented in blocks of four during silent periods between TRs, each block lasting 14,440ms. Stimulation blocks were followed by blocks of silence lasting 14,440ms. Each run consisted of 15 trials per condition presented in a pseudo-random order and lasted 30min approximately. Results presented in this article were obtained using two functional runs of one subject.

The fMRI data sets were subjected to a series of pre-processing operations. (1) Slice-scan-time correction was performed by resampling the time courses with linear interpolation such that all voxels in a given volume represent the signal at the same point in time. (2) Head movements were detected and automatically corrected by minimizing the sum of squares of the voxel-wise intensity differences between each volume and the first volume of the run. Each volume was then resampled in three-dimensional space according to the optimal parameters using trilinear interpolation. (3) Temporal high-pass filtering was performed to remove temporal drifts of a frequency below seven cycles per run. (4) Temporal low pass filtering was performed using a Gaussian kernel with FWHM of two data points. (5) After co-registration to the anatomical images collected in the same session the functional volumes were projected into Talairach space. (6) Moderate spatial smoothing with a Gaussian kernel of FWHM of 3mm was performed on the volume time series.

After pre-processing, the two functional time series were used for the SVM based analysis as described in Fig. 1, which produced a total of 30 trials per condition.

In particular we employed a purely multivariate feature selection strategy (RFE). RFE was compared, matching the number of selected voxels, to the performances of SVM based classification preceded by purely univariate feature selection strategies based on *t*-test or Wilcoxon (univT; univW) and univariate activation-based selection (univAct).

Using univariate activation-based ranking as initial voxel reduction strategy (i.e. selecting the union of the 2000 most active voxels for all conditions; univActRed) we evaluated its impact on multivariate feature selection (univActRed + RFE). Matching the number of selected voxels we compared univActRed + RFE to methods in which the same initial activation-based reduction was followed by different univariate selection strategies (univActRed + univT; univActRed + univW) and to the case in which only activation-based selection was used (univActRed+univAct).

The cross-validation scheme we used for the real data analysis followed the procedure used for the analysis of the simulated data (ten folds each time leaving ten trials out; 50s level splits; ten feature selection levels [R=10, L=5]; percentage of discarded voxels per feature selection step).

The same data set was also subjected to conventional univariate statistical analysis using BrainVoyager QX (Brain Innovation, Maastricht, The Netherlands). For all six possible contrasts, statistical parametric maps were computed searching for voxels that discriminated between conditions consistently in the two functional runs (conjunction analysis; Nichols et al., 2005) and were thresholded using false discovery rate (FDR, q = 0.05).

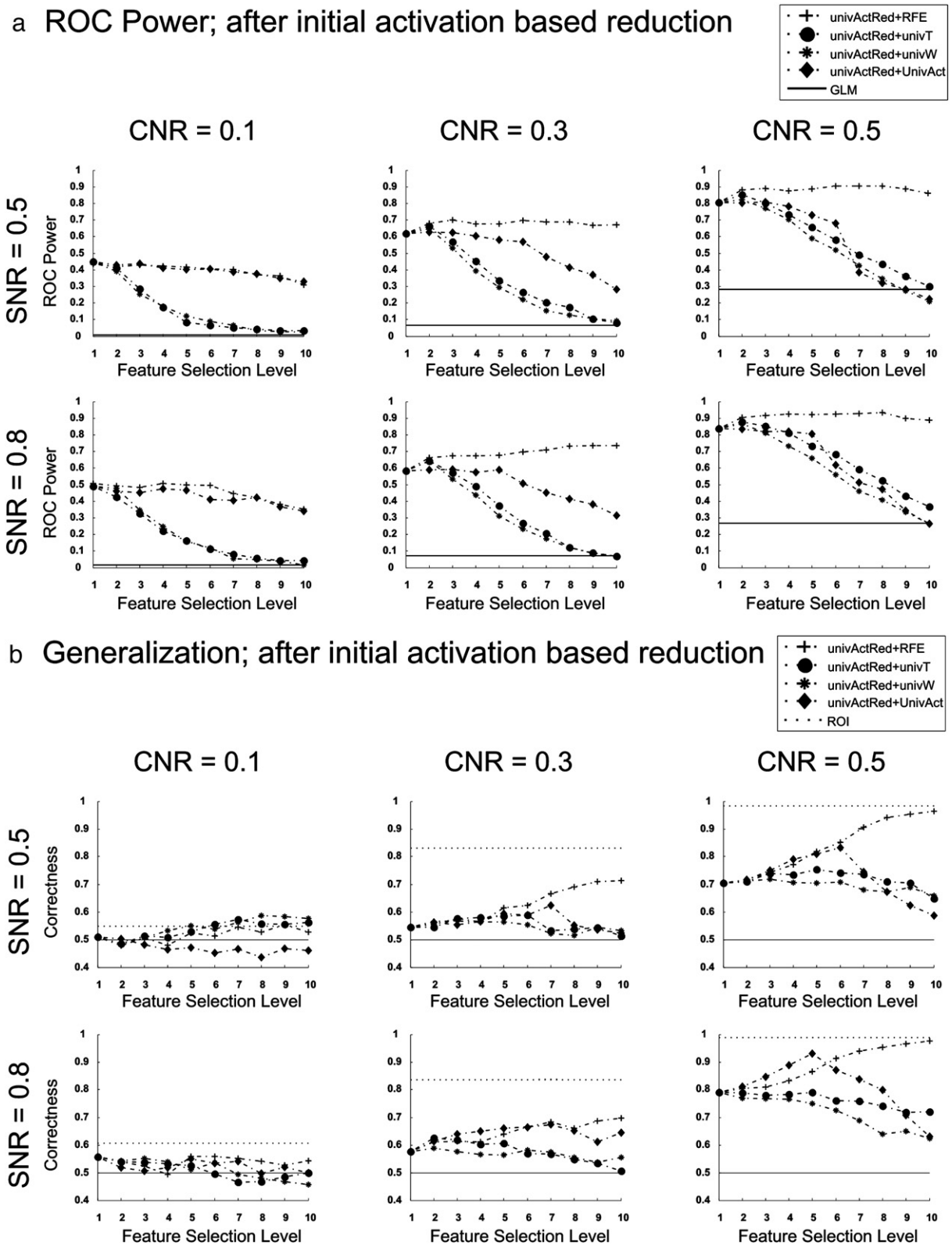

Fig. 4. Results obtained from the combination of an initial univariate activation-based voxel reduction and different subsequent voxel selection strategies. Results are reported at different CNR and SNR values (varBOLD fixed at 10% of the maximum response). (a) Receiver Operative Characteristic (ROC) power (defined as the area of the ROC curve in the false positive range of [0; 0.01]) of SVM based maps obtained for different numbers of selected voxels starting from a subset of voxels selected using activation-based reduction (univActRed). Compared feature selection schemes are: 1) RFE; 2) univariate discrimination-based selection (univT; univW); and 3) univariate activation-based selection (univAct). The bold line represents the sensitivity analysis of conventional statistical parametric maps obtained using the GLM. (b) Generalization performances of SVM based classifier are plotted for a different number of selected voxels starting from a subset selected using univariate activation-based reduction (univActRed). Compared feature selection schemes are: 1) RFE; 2) univariate discrimination-based selection (univT; univW); and 3) univariate activation-based selection (univAct). Different methods are compared to classification obtained using only the discriminative voxels (dotted lines). The bold line represents the chance level (0.5).

Results

We compared different voxel selection methods in terms of their sensitivity to the true discriminative voxels (ROC analysis; Fig. 3a, whole-brain analysis; Fig. 4a, after initial activation-based voxel reduction) and generalization performances (Fig. 3b, whole-brain analysis; Fig. 4b, after initial activation-based voxel reduction). In what follows we detail the results obtained on the simulated and real fMRI data. For comparison, ROC power obtained using conventional univariate statistical parametric mapping (GLM, bold line in Figs. 3a and 4a), chance level (bold line in Figs. 3b and 4b) and generalization performances obtained using only the simulated discriminative voxels (ROI; dotted line in Figs. 3b and 4b) are reported.

Simulated fMRI data

Whole-brain analysis

When starting the analysis from the whole set of voxels ($V \approx 40000$), selecting voxels univariately based on their activation results in higher ROC power (Fig. 3a) and higher generalization performances (Fig. 3b) than univariate discrimination-based voxel selection. A purely multivariate voxel selection strategy (RFE) iteratively improves sensitivity to the discriminative pattern at CNR=0.3 and CNR=0.5 (Fig. 3a) and improves generalization at CNR=0.5 (Fig. 3b) but not at very low CNR levels (CNR=0.1). Compared to RFE, activation-based voxel selection has an advantage in terms of ROC power (Fig. 3a) at CNR=0.1 and CNR=0.3. On the contrary, at the highest CNR level (CNR=0.5) RFE outperforms activation-based voxel selection both in terms of sensitivity (Fig. 3a) and generalization (Fig. 3b). These results can be explained considering the nature of the simulated discriminative voxels, which are few compared to the entire set and present activation levels (depending on the SNR) above the baseline noise. At high CNR levels, the discriminative information is sufficient to drive the multivariate search despite the large number of irrelevant voxels. At lower CNR levels this is not the case and selecting univariately the voxels based on their activation proves to be more effective. This suggests the combination of univariate activation-based voxel selection and RFE as a promising strategy for MVPA.

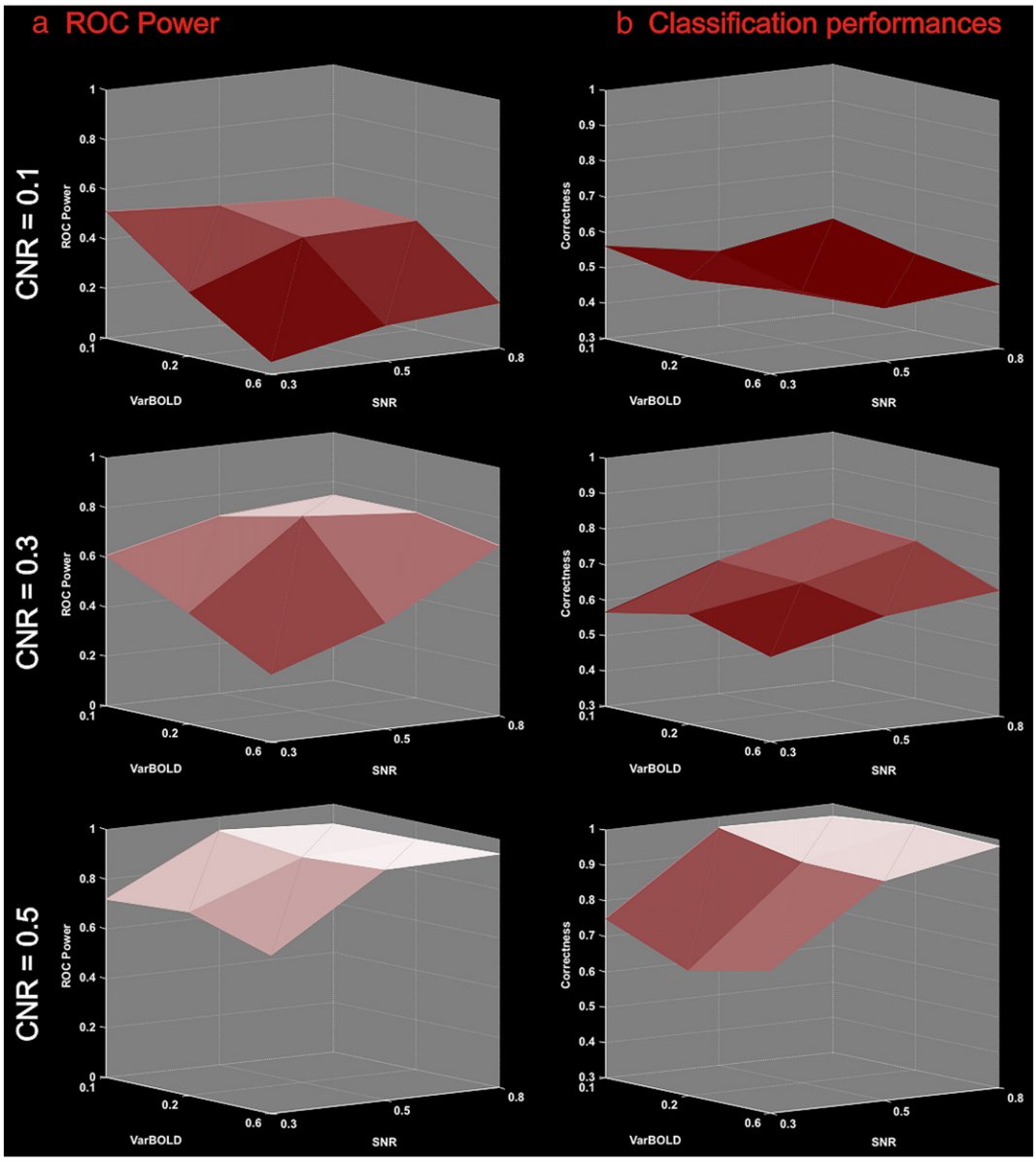

Fig. 5. ROC power (a) and generalization performances (b) obtained using univActRed + RFE at different SNRs and varBOLD levels reported for different CNR values.

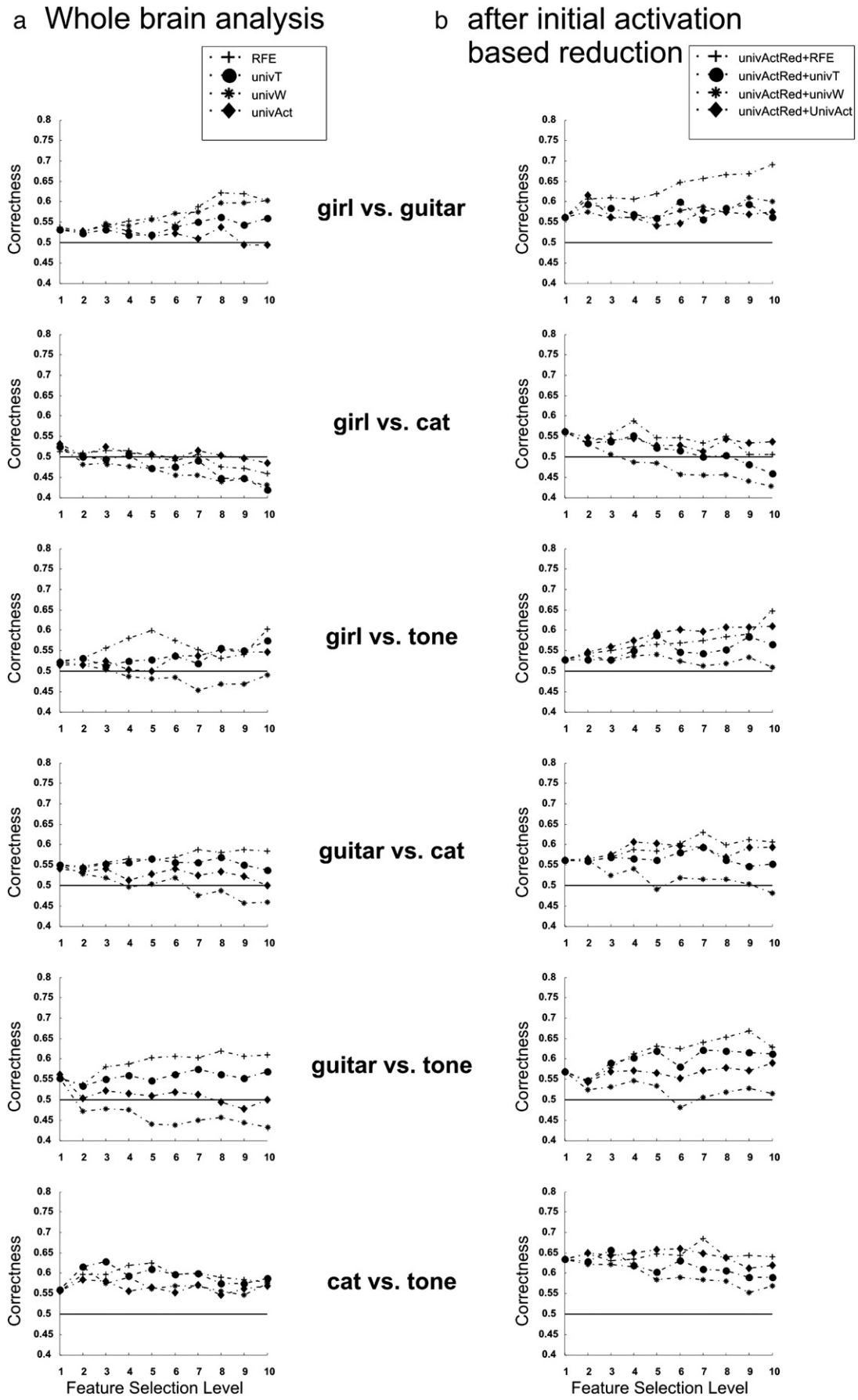

Fig. 6. Classification performances obtained on the real data set for each binary comparison in the discrimination of sound categories. (a) Performances of different feature selection schemes on whole-brain analysis. (b) Performances of different feature selection schemes after an initial univariate activation-based voxel reduction.

Combination of univariate and multivariate voxel selection

Fig. 4 shows that, after univariate activation-based voxel reduction, iteratively pruning the voxels based on their multivariate information (univActRed + RFE) clearly outperforms all other feature selection strategies both in terms of sensitivity (Fig. 4a) and generalization performances (Fig. 4b). This same strategy provides the highest performances also compared to whole-brain analysis (compare Figs. 4 and 3) and allows approaching close-to-optimal levels of classification, as defined by those obtained using only the discriminative voxels (dotted lines in Fig. 4b). Improvements are not observed only at CNR = 0.1, this can be due to the fact that at this CNR level, even using only the discriminative voxels, classification performances are close-to-chance level.

Performances decrease when the variability of the BOLD response increases

Fig. 5 shows univActRed+RFE ROC power and generalization at different SNR and varBOLD levels for different CNRs. Both sensitivity and classification performances are negatively affected by the variability of the BOLD response. As expected, the decrease in performances with increasing variability is stronger at lower CNRs and SNRs where the intra-class distance in the multivariate space is lower.

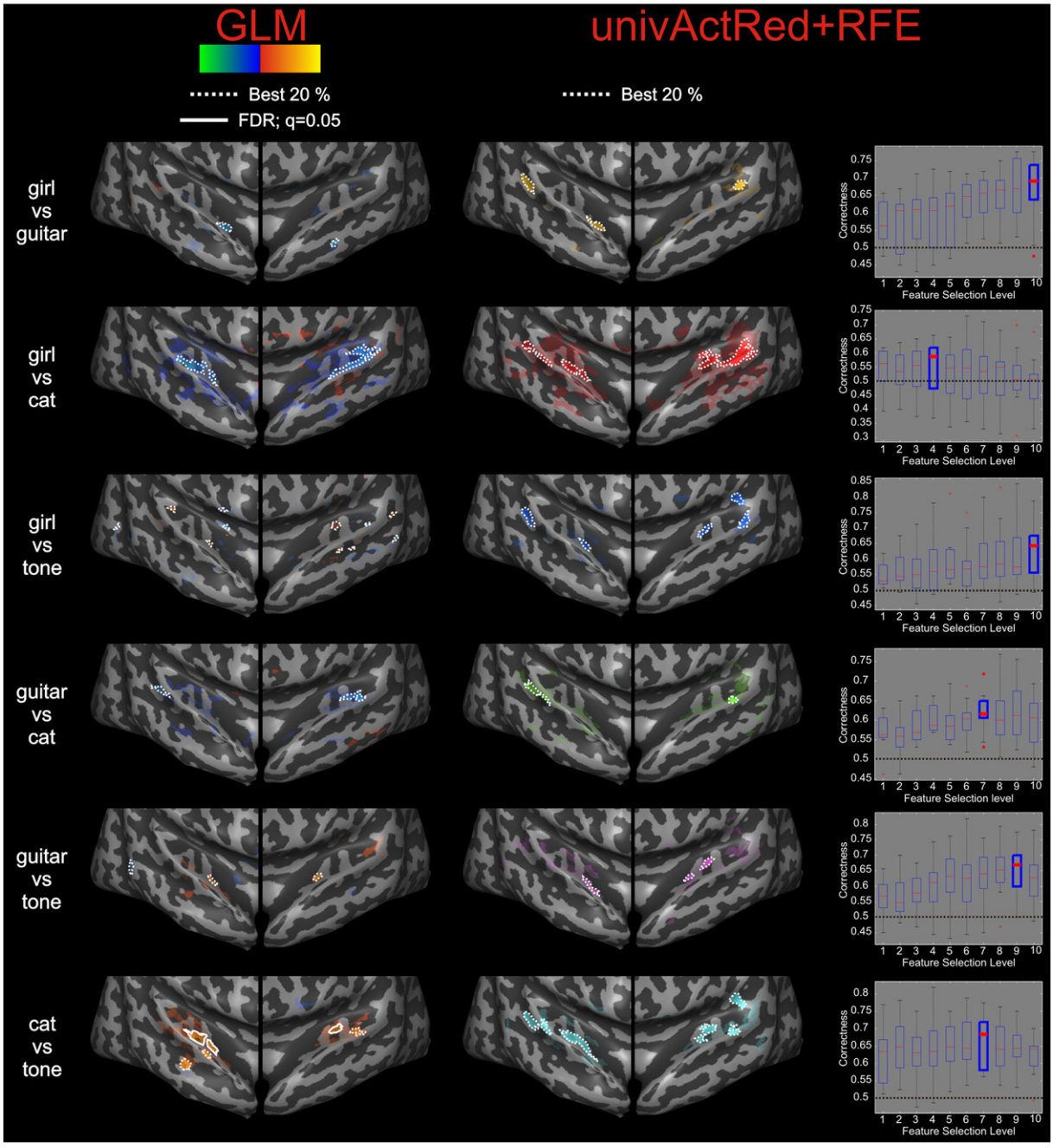

Fig. 7. Detailed results obtained on the real data set for each binary comparison in the discrimination of sound categories. Unthresholded (transparency coding) GLM maps (first column) are projected over the inflated cortex of the subject, the 20% of the voxels with the highest absolute *t*-values are shown within dotted lines and statistically significant voxels (thresholded with FDR; $q = 0.05$) within bold lines. Unthresholded (transparency coding) discriminative maps obtained using univActRed + RFE at the best feature selection level (second column) are projected over the inflated cortex of the subject, the best (most discriminative) 20% of the voxels are shown within dotted lines. The third column shows generalization results (median, lower, upper quartile and dispersion) obtained using univActRed + RFE at different feature selection levels, the best level (i.e. the highest generalization performances) is highlighted.

Real data

Figs. 6 and 7 show the results obtained using SVM based classification and different feature selection schemes on the real fMRI data.

RFE improves single trials classification performances. Fig. 6 shows the SVM based classification performances for the six possible contrasts at different feature selection levels for different feature selection methods. Both when starting from the whole brain (Fig. 6a) or after an initial voxel reduction based on single condition activation levels (Fig. 6b), the highest classification performances for each contrast are obtained using RFE. Improvements in classification of single trials are visible for each contrast especially when RFE follows an initial univariate feature selection based on activation measures (Fig. 6b). In Table 1 we report for each binary classification the percentage of correct classification obtained using univActRed + RFE, the best feature selection level (and the corresponding number of voxels used in the classification), and the size of the improvement (the highest classification performance minus classification performance at the initial feature selection step). For comparison we also report the percent difference in classification between univActRed + RFE and the closest performing method at the same feature selection level.

Fig. 7 shows detailed results obtained using classical univariate mapping (GLM, first column) and RFE after univariate activation-based voxel reduction (univActRed + RFE) (second column) for the six different discriminations.

Generalization performances of univActRed+RFE at different feature selection levels are reported as median, lower and upper quartile across the different iterations. The best feature selection level, as defined by the highest generalization performances, is highlighted and chance level is reported for comparison as a dashed line.

In the case of univActRed+RFE we report the entire map associated with the best feature selection level using a transparency coding scheme in which the size of a map value determines its transparency value (the lower the value, the more transparent it will be shown). The 20% of the voxels with the highest absolute discriminative weights are reported within dotted lines.

For comparison unthresholded GLM maps are reported with transparency coding and the 20% of the voxels with the highest absolute *t*-values are reported within dotted lines. Furthermore, for the GLM we show significant voxels as identified by conventional methods (FDR at $q = 0.05$).

The sixth contrast (cat vs. tone) shows significant bilateral univariate differences. The same areas are highlighted as most discriminative using univActRed + RFE, with the highest generalization performances (0.67) obtained at the seventh iteration. Similar generalization performances are obtained for the other contrasts (girl vs. guitar: 0.66; girl vs. cat: 0.58; girl vs. tone: 0.65; guitar vs. cat: 0.61; guitar vs. tone: 0.65) at different feature selection levels. Note that none of them show significant univariate differences. For all contrasts, maps obtained with univActRed + RFE highlight bilateral discriminative regions along the superior temporal gyrus, both anterior and posterior to the primary auditory regions located along the Heschl's gyrus.

Table 2
Generalization performances on the real data set reported together with estimated CNR, SNR, BOLD variability and sensitivity (i.e. ROC power for the false positive rate interval [0; 0.01])

	Girl/guitar	Girl/cat	Girl/tone	Guitar/cat	Guitar/tone	Cat/tone
SNR	0.26	0.28	0.25	0.30	0.25	0.26
CNR	0.32	0.26	0.30	0.31	0.30	0.34
VarBOLD	0.11	0.12	0.12	0.11	0.12	0.11
% Correct	66	58	65	61	65	67
ROC power	0.65	0.61	0.66	0.67	0.66	0.67

In order to compare the real data results with the simulation results we computed the SNR, CNR and BOLD variability of the real data for each contrast in the voxels that produced the highest generalization results. Using these computed values and the generalization performances we determined the closest simulation case and estimated from the corresponding parameter values the ROC power for the real data. The results are reported in Table 2. Accepting a false positive rate in the interval [0, 0.01] all contrasts are in the ROC power interval [0.61, 0.67].

Discussion

Differently from conventional univariate statistical analyses, machine learning techniques take advantage of the multivariate nature of the fMRI data and highlight maximally discriminative spatial patterns. While these methods offer a sensible advantage compared to conventional univariate mapping in the case of low contrast-to-noise scenarios, the main challenge in their application to fMRI is dealing with the large number of voxels in combination with a rather low number of trials of a typical scan. Performances of pattern recognition methods such as SVMs are in fact known to degrade with the increasing number of irrelevant features.

In the present article we have described and evaluated an approach for fMRI pattern discrimination analysis based on Support Vector Machines and a combination of univariate and multivariate feature selection strategies. Using this approach, the search for multivoxel discriminative patterns is iterative and data-driven, thus minimizing the number of required spatial assumptions on the location and extent of the patterns.

Compared to previous approaches employing whole-brain analyses (Mourao-Miranda et al., 2005), the evaluated method increases the sensitivity for the discriminative patterns, especially when they include a relatively small number of voxels compared to the whole data set (sparse discriminative patterns). This method can thus be seen as a useful solution when specific hypotheses on the localization (Haynes and Rees, 2005, Kamitani and Tong, 2005) and/or dimension (Kriegeskorte et al., 2006) of the spatial patterns are not available.

In our approach, the search of patterns is based on the Recursive Features Elimination (RFE) algorithm (Guyon et al., 2002), which

Table 1
Summary of results obtained on the real data analysis when RFE is applied after initial voxel reduction obtained using univariate activation-based ranking

	Girl/guitar	Girl/cat	Girl/tone	Guitar/cat	Guitar/tone	Cat/tone
Correct classification	66%	58%	65%	61%	65%	67%
Best level (univActRed + RFE)	10 (NrVox = 236)	4 (NrVox = 2002)	10 (NrVox = 236)	7 (NrVox = 687)	9 (NrVox = 337)	7 (NrVox = 687)
Improvement size (univActRed + RFE)	10%	2%	12%	6%	9%	3%
Difference to closest method	7% (univActRed + univW	3% (univActRed + univT)	5% (univActRed + univAct)	2% (univActRed + univAct)	5% (univActRed+ univT)	3% (univActRed+ univAct)

Improvement size is defined as the highest generalization performances minus the generalization performances at the first feature selection level. The best feature selection level (number of voxels used in the classification are reported in brackets) is defined as the level at which the highest classification performances are obtained, for each contrast. The difference to the closest performing method (reported in brackets) at the same feature selection level is also reported.

iteratively eliminates the least discriminative features based on multivariate information as detected by the classifier (Support Vector Machine) itself. RFE has been recently used for the analysis of fMRI data (Hanson and Halchenko, 2008), and has been proven to improve generalization performances in discriminating visual stimuli (Faces and Houses; block design) during two different tasks (1-back recognition detection task, oddball). Here we compared the performances of RFE and different univariate feature filter methods (activation- and discrimination-based) on simulated fMRI data. Furthermore we evaluated, on the same simulated fMRI data, the performances of a combined approach, univariate activation-based feature reduction and multivariate recursive feature elimination, to feature selection. As an illustrative example, we applied the proposed MVPA approach to the classification of fMRI responses elicited by auditory stimuli revealing overlapping, but distinguishable multivoxel patterns for different sound categories.

In a previous publication Carlson et al. (2003) used a "knock out" procedure to examine the degree of overlap in information between the representations of different object categories (chairs, faces and houses) in the visual cortex. In particular the procedure aimed to compare the reduction in classification performances of a stimulus category (e.g. chairs) in two cases: first, when the discriminant direction between a stimulus category and all other categories was removed (chairs vs. faces and houses) and second, when the discriminant direction of another stimulus was removed (faces vs. chairs and houses). This "virtual lesion" approach was implemented projecting the multivoxel patterns on the different category-specific discriminant hyperplanes (i.e. removing the direction of the category-specific maximum discrimination in the multivoxel space) and subsequently evaluating the performance losses for each category. The authors showed that removing a category-specific discriminant reduced the classification of all other categories only in part. These results suggest that there is not a complete overlap between the representations of the different object categories in the visual cortex. While the aim of the knock out procedure of Carlson et al. is to evaluate the similarity between the multi voxel patterns elicited by different stimulation conditions, RFE aims to optimize in a multivariate and data-driven way the discriminative information between different categories. In particular, while RFE removes at each iteration the least discriminant voxels the knock out procedure of Carlson et al. removes a direction in the feature space which is a weighted average of all voxels.

Results of our simulations show that the combination of RFE and univariate activation-based reduction of voxels ensures the highest sensitivity and generalization performances (see Fig. 4). In particular, when RFE is applied after an initial univariate activation-based voxel reduction there is a sensible advantage compared to the case the same initial voxel reduction is followed by univariate discrimination or activation-based selection, especially at very low CNRs (see Figs. 4a–b). This is a consequence of assuming that, at single voxel level, BOLD changes of a condition compared to the baseline are greater than BOLD differences between conditions (SNR > CNR), which appears as a realistic assumption in most fMRI studies. This result confirms previous comments on the use of univariate feature selection methods to MVPA (Mitchell et al., 2004, Mourao-Miranda et al., 2006). Because of the reduced sensitivity of univariate statistics at low SNRs/CNRs the most appropriate choice for combining univariate and multivariate features selection is to use rather liberal thresholding for univariate selection (which prevents the exclusion of potentially informative voxels) and further discard irrelevant voxels based on the multivariate scoring function.

Also alone, the recursive approach proves to be more sensitive than conventional univariate analysis (General Linear Model; Fig. 3a) and shows sensible improvements with a decreasing number of voxels (Figs. 3a–b). In our simulations, indeed, both ROC power and generalization performance increased with feature selection level, with the latter approaching optimal values, i.e. those obtained using only the simulated discriminative voxels (Fig. 4b). Note that the superiority of the combined approach (univActRed + RFE) compared to the purely multivariate approach (RFE) is due to the chosen strategy to use a constant value for the total number of feature selection steps in the different methods. Better performances might be obtained allowing the multivariate approach to exhaustively search the whole set of features with a larger number of smaller steps. However, this would require a much longer computational time.

Our simulations did not consider the case in which discriminative patterns are not represented by regions that do not show a global main effect of activation (Haynes et al., 2007), in which case using RFE without univariate activation-based pre-selection may prove to be more sensitive. More generally, available *a priori* information on the nature of the effects of interest (e.g. presence of a global main effect) or on its location may aid and guide the chosen feature selection strategy. As shown by the simulation results, perfect anatomical knowledge of the location of the discriminative patterns (ROI approach) proves to be the most sensitive method. Our RFE-based approach, on the other hand, allows searching for discriminative patterns in a more "data-driven" way with no initial assumption on their location. The analysis of the real data shows an illustrative case. For sound categorization, little is known on the exact localization of the discriminating patterns within the human auditory cortex and defining anatomical or functional landmarks is not straightforward. Using RFE we were able to map these discriminative patterns and improve the generalization performance compared to the initial anatomical selection of voxels.

Furthermore our simulations were limited to the case of two sparse and spatially distributed populations of voxels without explicit covariances between them (e.g. functional and effective connectivity). The observed superiority of the multivariate analysis compared to the mass-univariate GLM is thus due to the integration of weak univariate differences irrespective of the sign of the discriminative information, which also explains its advantage compared to conventional smoothing (Kreiegeskorte et al., 2006). In such cases, combining RFE with classifiers other than the ls-SVM (e.g. GNB) would produce similar results. More generally, however, the presence of functional and effective connectivity among voxels may affect the final outcome of our method. In fact, while the proposed RFE procedure can be applied to any algorithm, the weighting of individual features is algorithm-dependent and may be influenced by the way the classifier handles the covariance between the features.

We tested the same approaches on fMRI time series obtained in an auditory experiment with sounds from different categories. In line with the results from the simulations, RFE preceded by activation-based univariate voxel reduction selection (univActRed+RFE) produced the highest generalization performances and the recursive approach to feature elimination improved generalization performances in all contrasts (Fig. 6; Table 1). Classical univariate mapping failed to reveal significant discriminative regions for all contrasts except the sixth (cat vs. tone) (Fig. 7, first column). As expected, in this latter case, discriminative maps produced by univActRed+RFE overlapped with GLM contrast map and were accompanied by above chance generalization performances (cat vs. tone: 0.67). In all other contrasts, despite the lack of statistically univariately significant voxels in standard GLM analysis, our approach reached comparable generalization performances (girl vs. cat: 0.58; girl vs. guitar: 0.66; girl vs. tone: 0.65; guitar vs. cat: 0.61; guitar vs. tone: 0.65). The discriminative spatial patterns as highlighted by univActRed + RFE comprise multiple non-neighboring regions in the anterior and posterior portions of the superior temporal gyrus, in both the right and left auditory cortex. These results are consistent with the notion of a 'what' auditory processing stream originating in the superior temporal areas, anterior to the Heschl's gyrus (Belin and Zatorre

2000a, Belin et al., 2000b, Rauschecker and Tian, 2000, Lewis et al., 2005) and with recent fMRI studies which point to a relevant role of STS in the representation and processing of complex sounds (Warren et al., 2005). A full account of these results, including a group analysis, is given in Staeren et al. (2008).

One possible drawback of the application of RFE is the backward elimination strategy, which requires setting of several parameters, the most relevant being the number of iterations and the number of features to discard at each iteration. Searching exhaustively the whole feature set would require a large number of iteration with few discarded voxels at each iteration. Especially in fMRI, however, such an approach would result in very long computational time. In the present paper we selected a relatively small but practically feasible number of feature selection steps (10) and discarded, at each iteration, a fixed proportion of the current number of features computed based on the desired final set size. While arbitrary, this choice proved to be effective both in the case of the simulations and of the real data. It should be noted, however, that different data sets may require different settings. Other multivariate feature selection methods, such as the embedded algorithm by Rakotomamonjy (2003) suffer from similar problems, as the size of the discriminative feature set has to be chosen after the optimization is terminated. The application of these methods, thus, requires heuristic choices, compromising practical feasibility and optimal search criteria.

The single trial estimation procedure we outlined in this paper was designed for block or slow event related designs for which one may derive a response-pattern estimate for each block/event (or even TR). This is not possible with rapid event related designs. In these cases, applying MVPA requires subdividing the measurements in many subparts, each one including a sequence of trials that allows for estimating condition response patterns. Assuming linearity of BOLD responses, one obtains a response-pattern estimate from each of these sub-runs that can then be used as those obtained from a slow event related design. After trial estimation, application of RFE is identical as in blocked or slow event related designs.

Another methodological consideration regards the possibility of defining a statistical threshold for the maps produced by the SVM classifier. Statistical assessment of discriminative maps obtained by MVPA, however, is not simple and it requires assumption or estimation of a null hypothesis distribution for the voxels' discriminative weights.

Without assumptions on this distribution, thresholding might be performed using permutation testing (Mourao-Miranda et al., 2005, Wang et al., 2007). This approach allows estimating an empirical null hypothesis distribution for the discriminative contribution of each voxel. However, it is computationally very intense, especially considering that permutation-based tests should be performed, in our case, for each level of feature selection.

Alternatively, one may test the consistency of the discriminative contribution of a location (voxel) across subjects in a group analysis of the discriminative maps (Wang et al., 2007). This approach assumes that there is a spatial (anatomical) correspondence between the discriminative patterns in different subjects. We have implemented a method for group analysis that combines the anatomical cortex-based realignment of the subjects' brains with a random effect analysis of their discriminative maps. A full account of the results of group level statistical analysis performed on the maps obtained applying our approach to the problem of sound categorization is given in Staeren et al. (2008).

The ROC analysis performed to evaluate the sensitivity of our approach on the simulated data sets is independent of the threshold as the true positive and false positive ratios are computed for a range of thresholds. For real data, we report unthresholded maps at the best feature selection level, as defined by the highest generalization accuracies. Note that the voxels selected using the proposed iterative procedure provide significant generalization performances in the classification of new trials and the results reported in Table 2 indicate that the proposed approach is sensitive to the underlying discriminative patterns.

Conclusions

We illustrated different strategies to perform feature selection for pattern discrimination analysis of fMRI data and introduced a novel, data-driven feature selection strategy that uses multivariate information. Our results show that the combination of univariate (activation-based) and multivariate feature selection outperforms other techniques when no *a priori* information is available on the size and location of the pattern of interest.

The proposed method could be extended to the multivariate analysis of data from other imaging modalities (perfusion MRI, PET, optical imaging, EEG, MEG) or to their combination. In the latter, feature elimination could be applied in the multidimensional space of features extracted from different methods (e.g. voxels of fMRI and time points of EEG) and used to reveal the most discriminative set of multi modal features.

Acknowledgments

Financial support from NWO (MaGW-VIDI grant 452-04-330) to EF is gratefully acknowledged.

References

Belin, P., Zatorre, R.J., 2000a. 'What', 'where' and 'how' in auditory cortex. Nat. Neurosci. 3 (10), 965–966.
Belin, P., Zatorre, R.J., Lafaille, P., Ahad, P., Pike, B., 2000b. Voice-selective areas in human auditory cortex. Nature 403 (6767), 309–312.
Carlson, T.A., Schrater, P., He, S., 2003. Patterns of Activity in the Categorical Representation of Objects. J. Cog. Neorosci. 15 (5), 704–717.
Cox, D., Savoy, R., 2003. Functional magnetic resonance (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. Neuroimage 19 (2), 261–270.
Cristianini, Shawe, T., 2000. An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge University Press.
Fadili, M.J., Ruan, S., Bloyet, D., Mazoyer, B., 2000. A multi-step unsupervised fuzzy clustering analysis of fMRI time series. Hum. Brain Mapp. 10, 160–178.
Friston, K.J., Fletcher, P., Josephs, O., Holmes, A., Rugg, M.D., Turner, R., 1998. Event-related fMRI: characterizing differential responses. Neuroimage 7 (1), 30–40.
Guyon, I., Elisseef, A., 2003. An introduction to variable and feature selection. Journal of Machine Learning Research 3, 1157–1182.
Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Machine Learning 46, 389–422.
Hanson, S.J., Halchenko, Y., 2008. Brain reading using full brain support vector machines for object recognition: there is no face identification area. Neural Computation 20 (2), 486–503.
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 (1), 156–166.
Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Aschouten, J.L., Pietrini, P., 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293 (5539), 2425–2430.
Haynes, J.D., Rees, G., 2005. Predicting the orientation of invisible stimuli from activity in human primary visual cortex. Nat. Neurosci. 8 (5), 686–691.
Haynes, J.D., Rees, G., 2006. Decoding mental states from brain activity in humans. Nat. Rev. Neurosci. 7 (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. Current Biology 17, 323–328.
Kamitani, Y., Tong, F., 2005. Decoding the visual and subjective contents of the human brain. Nat. Neurosci. 8 (5), 679–685.
Kohavi, R., John, G., 1997. Wrappers for feature selection. Artificial Intelligence 97 (1–2), 273–324.
Kriegeskorte, N., Goebel, R., Bandettini, P., 2006. Information-based functional brain mapping. Proceedings of the National Academy of Sciences of the United States of America 103 (10), 3863–3868.
LaConte, S., Strother, S., Cherkassky, V., Anderson, J., Hu, X., 2005. Support vector machines for temporal classification of block design fMRI data. Neuroimage 26 (2), 317–329.
Lewis, J.W., Brefczynski, J.A., Phinney, R.E., Janik, J.J., DeYoe, E.A., 2005. Distinct cortical pathways for processing tool versus animal sounds. Journal of Neuroscience 25 (21), 5148–5158.
Mitchell, T.M., Hutchinson, R., Niculescu, R.S., Pereira, F., Wang, X., 2004. Learning to decode cognitive states from brain images. Machine Learning 57, 145–175.
Mourao-Miranda, J., Bokde, A.L., 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 (4), 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 (4), 1055–1065.
Nichols, T.E., Brett, M., Andersson, J., Wager, T., Poline, J.-B., 2005. Valid conjunction inference with the minimum statistic. NeuroImage 25, 653–660.
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. Sep. 10 (9), 424–430.
O Toole, A.J., Jang, F., Abdi, H., Haxby, J.V., 2005. Partially distributed representations of objects and faces in ventral temporal cortex. Journal of Cogn. Neurosci. 17 (4), 580–590.
Rakotomamonjy, A., 2003. Variable selection using SVM-based criteria. Journal of Machine Learning Research 3, 1357–1370.
Rauschecker, J.P., Tian, B., 2000. Mechanisms and streams for processing of "what" and "where" in auditory cortex. Proceedings of the National Academy of Sciences of the United States of America 97 (22), 11800–11806.
Skudlarski, P., Constable, R.T., Gore, J.C., 1999. ROC analysis of statistical methods used in fMRI: individual subjects. Neuroimage 9, 311–329.
Sorenson, J.A., Wang, X., 1996. ROC method for evaluation of fMRI techniques. Magn. Reson. Med. 36, 737–744.
Staeren, N., Renvall, H., De Martino, F., Goebel, R., Formisano, E., (2008). Sound categories are represented as distributed patterns in the human auditory cortex (submitted for publication).
Suykens, J.A.K., Van Gestel, T., De Barbanter, J., De Moor, B., Vanderwalle, J., 2002. Least Squares Support Vector Machines. World Scientific Publishing.
Wang, Z., Childress, A.R., Wang, J., Detre, J.A., 2007. Support vector machine learning-based fMRI data group analysis. Neuroimage, doi:10.1016/j.neuroimage.2007.03.072.
Warren, J.D., Jennings, A.R., Griffith, T.D., 2005. Analysis of the spectral envelope of sounds by the human brain. Neuroimage 24, 1052–1057.