Available online at www.sciencedirect.com
ScienceDirect
Magnetic Resonance Imaging 26 (2008) 921–934
MAGNETIC RESONANCE IMAGING

Multivariate analysis of fMRI time series: classification and regression of brain responses using machine learning

Elia Formisano*, Federico De Martino, Giancarlo Valente

Faculty of Psychology, Department of Cognitive Neuroscience, University of Maastricht, 6200 MD Maastricht, The Netherlands

Received 31 October 2007; accepted 14 January 2008

Abstract

Machine learning and pattern recognition techniques are being increasingly employed in functional magnetic resonance imaging (fMRI) data analysis. By taking into account the full spatial pattern of brain activity measured simultaneously at many locations, these methods allow detecting subtle, non-strictly localized effects that may remain invisible to the conventional analysis with univariate statistical methods. In typical fMRI applications, pattern recognition algorithms "learn" a functional relationship between brain response patterns and a perceptual, cognitive or behavioral state of a subject expressed in terms of a label, which may assume discrete (*classification*) or continuous (*regression*) values. This learned functional relationship is then used to predict the unseen labels from a new data set ("brain reading"). In this article, we describe the mathematical foundations of machine learning applications in fMRI. We focus on two methods, support vector machines and relevance vector machines, which are respectively suited for the classification and regression of fMRI patterns. Furthermore, by means of several examples and applications, we illustrate and discuss the methodological challenges of using machine learning algorithms in the context of fMRI data analysis.

Keywords: Functional MRI; Machine learning; Pattern recognition; Multivariate classification; Multivariate regression

1. Introduction

A wide range of multivariate statistical methods are being increasingly applied to the analysis of functional magnetic resonance imaging (fMRI) time series. These methodological developments are providing cognitive neuroscientists with the opportunity of tackling new research questions that are relevant for our understanding of the functional organization of the brain and that cannot be addressed using the widely employed univariate statistical methods. In this latter approach, a spatially invariant model of the expected blood oxygenation level-dependent (BOLD) response is fitted independently at each voxel's time course, and the differences between estimated activation levels during two or more experimental conditions are tested [1]. Together with methods for mitigating the problem of performing a large number of tests, this *massively univariate* analysis produces statistical maps of response differences, highlighting brain locations that are "selective" or "specialized" for a certain stimulus dimension, i.e., voxels or regions of interest (ROIs) that respond more vigorously to a sensory, motor or cognitive stimulus compared to one or more appropriate control conditions [1]. This analytical strategy focuses on the mapping of a "stimulus–single location response" type of relation; however, it allows studying neither the stimulus (in-) dependent interactions between locations (functional and effective connectivity, see Ref. [2]) nor the effects which are reflected in a "stimulus–multiple locations response" type of relation. Whereas a variety of methods for the investigation of functional [2–7] and effective (e.g., Refs. [8–10]) connectivity exists and has gained a conspicuous tradition in functional neuroimaging, only recently have methods been proposed for analyzing the relation between a stimulus and the responses simultaneously measured at many locations (spatial response patterns or multivoxel response patterns).

* Corresponding author. Faculty of Psychology, Department of Cognitive Neurosciences, University of Maastricht, P.O. Box 616, 6200 MD Maastricht, The Netherlands. Tel.: +31 43 3884040; fax: +31 43 3884125.
E-mail address: e.formisano@psychology.unimaas.nl (E. Formisano).

0730-725X/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.mri.2008.01.052

This approach, which has been named multivoxel pattern (MVP) analysis [11] or, more evocatively, "brain reading" (see below), was initiated with a landmark fMRI study of the object-vision pathway [12]. In this study, Haxby et al. [12] demonstrated that spatial (multivoxel) patterns of BOLD responses evoked by a visual stimulus are informative with respect to the perceptual or cognitive state of a subject. Participants were presented with various visual stimuli from different object categories (including faces, chairs, bottles); measured data were split in half, and the spatial patterns of responses in the ventrotemporal cortex (the visual "what" stream) were estimated for each category and for each half of the data. By comparing the spatial correlation between all patterns obtained from the first half of the data with those obtained from the second half of the data, Haxby et al. demonstrated that perceiving each object category was associated with a distinct spatial pattern of responses and, thus, that these patterns could be used to "decode" the perceptual or cognitive state of the subjects. These results did not change when the same analysis was performed after excluding the regions of maximal responses (e.g., after excluding the BOLD signals in the fusiform face area for the "face" category). Importantly, these findings show that information on the perceived "object category" is entailed not only in the maximally responsive regions but also in spatially wide and distributed pattern of non maximal responses in the entire ventrotemporal cortex. Note that information in these latter responses is ignored in the conventional, subtraction-based approach that is aimed at detecting voxel-by-voxel statistically significant activation level differences and, thus, that only looks at the "tip of the iceberg" of the overall information content carried by the measured response patterns.

Following the study by Haxby et al. [12], several other groups examined, with increased methodological sophistication, the relation between sensory or cognitive stimuli and the spatial patterns of the measured response, and obtained remarkable results [13–20]. The methods employed derive mostly from statistical pattern recognition and machine learning and range from linear discriminant analysis [13,20], linear [14–19] and nonlinear [15] support vector machine (SVM) and Gaussian Naïve Bayes classifiers [16].

A major advantage of these methods compared to the conventional univariate statistical analysis is their increased sensitivity in discriminating perceptual and cognitive states. Statistical pattern recognition exploits and integrates the information available at many spatial locations, thus allowing the detection of perceptual and cognitive differences that may produce only weak single-voxel effects. This integration of information across multiple locations may range from considering only a small neighborhood of adjacent spatial locations (locally multivariate analysis, see, e.g., Ref. [20]) to jointly considering spatially remote regions or even voxels across the whole brain. Note that this integration of information is substantially different from spatial smoothing, commonly used in fMRI analysis. Spatial smoothing, indeed, increases sensitivity only when two conditions differ in terms of their regional mean activation levels. In these cases, the signal differences in the local neighborhood of a position of interest are all in the same direction and are enhanced by spatial averaging. Conversely, if two conditions differ in terms of their fine-grained spatial activation patterns, spatial smoothing has a destructive effect and cancels out the discriminative information, which can be detected by pattern recognition methods.

Applications and findings of "brain reading" have been recently reviewed [11,21]. These reviews and articles also consider the neural mechanisms that potentially underlie the observation of these spatially distributed effects and discuss implications, potential pitfalls and interpretational issues of using these methods in the context of cognitive neurosciences. In the present article, we will focus on the mathematical foundations and the methodological aspects of using machine learning algorithms for classifying fMRI patterns. Furthermore, we extend the use of machine learning to the regression of fMRI responses, which refers to the learning and prediction of a functional relationship between brain response patterns and a perceptual, cognitive or behavioral state of a subject that can be expressed in terms of a continuous label (in contrast to classification where the labels assume discrete values). Below, after a joint formulation of the problem of classifying and regressing fMRI responses, the two approaches will be considered separately and specific methodological aspects will be illustrated using own examples.

2. Classification and regression of fMRI responses

2.1. Problem formulation

Consider $\mathbf{D}$, $\mathbf{D}'$ as two data sets from a generic fMRI experiment, and $\mathbf{t}$, $\mathbf{t}'$ as the labels that describe the sensory, motor or cognitive stimulation associated with $\mathbf{D}$ and $\mathbf{D}'$, respectively. Pattern analysis algorithms aim at "learning" a functional relationship between data $\mathbf{D}$ (*training data set*) and label $\mathbf{t}$ in order to predict the unseen labels $\mathbf{t}'$ from the new data set $\mathbf{D}'$ (*test data set*). During this prediction, the algorithm blindly decodes the brain activation $\mathbf{D}'$ into the corresponding stimulus condition or cognitive state of the subject (as described by $\mathbf{t}'$), hence the definition of brain reading.

More formally, the problem that pattern analysis algorithms try to solve can be described as the learning of a function:

$$f = f(\mathbf{D}, \mathbf{t}, \boldsymbol{\theta}) \tag{1}$$

where $\boldsymbol{\theta}$ denotes the set of adjustable model parameters that may be estimated during the training phase. The estimated function can be used on a new data set to predict unseen labels:

$$\hat{\mathbf{t}}' = f(\mathbf{D}', \mathbf{D}, \mathbf{t}, \boldsymbol{\theta}) \tag{2}$$

where $\hat{\mathbf{t}}'$ denotes an estimate of the labels $\mathbf{t}'$. Labels may assume discrete values and describe the stimulus or subject

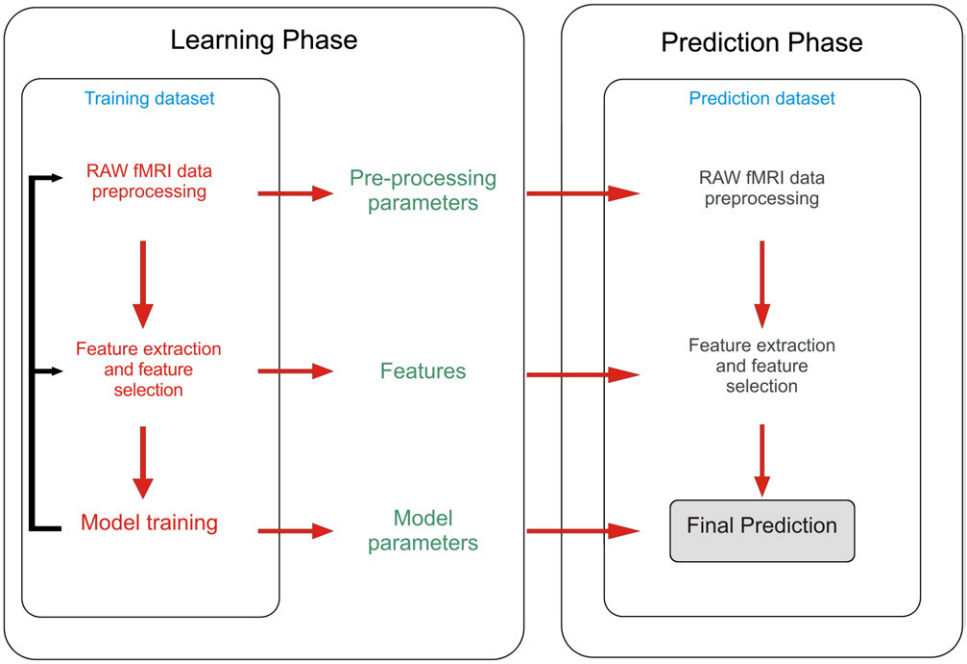

Fig. 1. Main steps of a generic pattern recognition algorithm as used in fMRI data analysis. In the training phase (left), the raw fMRI data are preprocessed and *relevant* features are selected from the data prior to *model training*. Prediction (right) is performed using the trained model on a new data set, after this latter has been preprocessed in the same way and reduced to same features as in the training.

state during the acquisition of a set of images; for instance in a blocked (or event-related) designed measurement with two conditions, labels will be t_i=+1 for all images collected under condition 1 and t_i=−1 for all images collected under condition 2. As described in section 3, this prediction problem is referred to as "classification." In other cases, the stimulus or subject state may be described using labels with continuous values. As described in section 4, the prediction problem is then referred to as "regression."

Fig. 1 summarizes in a block diagram the steps that are generally performed in applying pattern recognition algorithms in fMRI. In the training phase, the raw fMRI data are preprocessed in order to reduce the effects of noise. This step can be usually performed with well-known functional neuroimaging analysis software tools (AFNI, BrainVoyager, FSL, SPM). Prior to model training, the second step is the selection of *relevant* features from the data. This step aims at reducing the complexity of the data set and increasing the capabilities of the prediction scheme. During *model training*, several processing and feature extraction/selection may be explored; the learning phase is therefore depicted with feedbacks, indicating that preprocessing, feature extraction and model estimation cannot be regarded as completely separated parts of the training phase. Once the training phase is completed, a set of preprocessing parameters, a set of features and a set of model parameters is available, and the prediction is performed using the trained model on a new data set, after this latter has been preprocessed in the same way.

2.2. Performance metrics, cross-validation and model selection

Together with the model [Eq. (1)], a suitable performance metric has to be defined. This performance indicates how good the classification/regression is, and it can be expressed as an error (or loss) function ε. In very general words, the aim of a prediction algorithm is to learn a model on the training data set **D** that gives the minimum error on an unseen data set test **D′**. Several models and several performance metrics can be introduced to perform this operation [22], some of which will be presented in the following sections.

There are several ways to find suitable values for the model parameters, given the training data set. One of the most employed approaches is to split the training data set in two disjoint parts. One of the two is used as the traditional training set, while the other is used as a *validation* data set, to evaluate the generalization error. Sometimes, the data set is divided into *m* disjoint parts and a model is trained *m* times, each time leaving a part as a validation set. The estimated performance is then defined as the mean of the errors of the models on the different data sets (*m-fold cross-validation*) [22].

Bayesian methods usually do not require cross-validation to compare different model parameter choices (although sometimes it is also used in Bayesian frameworks for model selection [23]). In Bayesian analysis, model comparison involves the use of probabilities of the choice of a suitable model [22,24]. Given a set of models i=1,…L that we wish to compare using the observed data **D**, then it

is possible to compare these models by means of their posterior distribution:

$$p(M_i|\mathbf{D}) = \frac{P(\mathbf{D}|M_i)P(M_i)}{P(\mathbf{D})} \propto P(\mathbf{D}|M_i)P(M_i). \quad (3)$$

Where the data-dependent term $P(\mathbf{D}|M_i)$, also known as model evidence, can be seen as a likelihood function over the models space and $P(M_i)$ denotes the prior probability of the model. The normalizing factor $P(\mathbf{D})$ can be discarded while comparing models. Once the posterior probability in [Eq. (3)] has been estimated, it can be used in two ways: 1) considering a *mixture distribution*, using a linear combination of all the models, weighted by their probability; 2) performing *model selection*, that is, selecting the most probable model, given the data.

3. Classification of fMRI responses

In the context of classification of fMRI responses, let us consider a matrix of the multivoxel response patterns $\mathbf{X_c}$ ($n\times v$) derived from data $\mathbf{D}$ by estimating the responses to the stimulus presentations 1..n at voxels 1..v. Let us further suppose that the stimuli belong to one of two possible categories and, thus, we can define a vector of labels $\mathbf{t}$ with t_i=+1 for class 1 and t_i=−1 for class 2.

The training problem [Eq. (1)], in the simple case of *linear* classification of fMRI responses, can be formulized as the one of finding a discriminant function:

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}+b \quad (4)$$

where $\mathbf{w}$ ($1\times v$) is a weight vector and b is a bias term or threshold weight. The decision rule implemented by a binary (two classes) linear discriminant is, for a given sample $\mathbf{x_i}$, ($1\times v$):

$$\begin{cases} X_i\in\text{class 1 if } f(X_i)>0 \\ X_i\in\text{class 2 if } f(X_i)<0 \end{cases} \quad (5)$$

thus, the learning problem reduces to finding $\mathbf{w}$ and b such that:

$$t_i\left(\mathbf{w}^T\mathbf{x_i}+b\right)>0 \quad i=1,...n. \quad (6)$$

Assuming that the distribution of the classes is multinormal and that the features are statistically independent with the same variance (i.e., the covariance of the classes is the same and diagonal), the optimal linear discriminant is defined by the minimum distance (Euclidean) criterion. This criterion states that any new sample x is classified looking at its distance from the mean of the classes, as defined by the samples in the training set. This operation results in a weighting vector and bias defined as:

$$\begin{cases} \mathbf{w} = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1 \\ b = \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \end{cases} \quad (7)$$

where $\boldsymbol{\mu}_1$ ($1\times v$) and $\boldsymbol{\mu}_2$ ($1\times v$) are the means of the two classes computed from the training set. The minimum distance classifier criterion is easily modified to take into account classes with identical but nondiagonal covariances [22].

Analytical solutions to the classification problem can be also found when no a-priori assumptions are available on the distribution of the classes and the only interest is to maximize generalization performances (i.e., the classification of unseen patterns).

SVMs provide one such solution and are based on the maximization of the distance of the separating hyperplane to the nearest training sample. Such distance is called the *margin*. This optimization problem in the simple case in which the patterns are linearly separable can be formulated as:

$$\min_{w,b} J(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} \quad (8)$$

subject to:

$$t_i\left(\mathbf{w}^T\mathbf{x_i}+b\right)\geq 1,\ i=1,...,n. \quad (9)$$

Solution is obtained by applying Lagrangian methods [25]. The training patterns for which

$$t_i\left(\mathbf{w}^T\mathbf{x_{sv}}+b\right) = 1 \quad (10)$$

are called support vectors and are the examples in the training set whose correct classification is most difficult.

For the more general case of nonseparable classes (i.e., classes with overlapping distributions), the formulation of the SVM can be modified in order to account for misclassification errors introducing additional slack variables ξ_i, i=1,...,n. The optimization problem becomes:

$$\min_{w,b,\xi} J(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + a\sum_{i=1}^{l}\ \xi_i \quad (11)$$

subject to:

$$t_i\left(\mathbf{w}^T\mathbf{x_i}+b\right)\geq 1-\xi_i,\ i=1,...,n \quad (12)$$

and:

$$\xi_i\geq 0,\ i=1,...,l \quad (13)$$

where a is a positive real constant [26]. In the classical SVM formulation [Eqs. (8) and (9); (11–13)], the optimal boundary between different classes is obtained considering only the support vectors. Model selection (i.e., the parameter a) is generally performed using cross-validation techniques, i.e., further splitting the training data set in *training* and *validation* sets (see Section 2.2).

Least-squares SVMs (ls-SVM) are variations of the classical formulation [26]. In this formulation, each training point is weighted in order to obtain the distinguishing hypersurface (hyperplane).

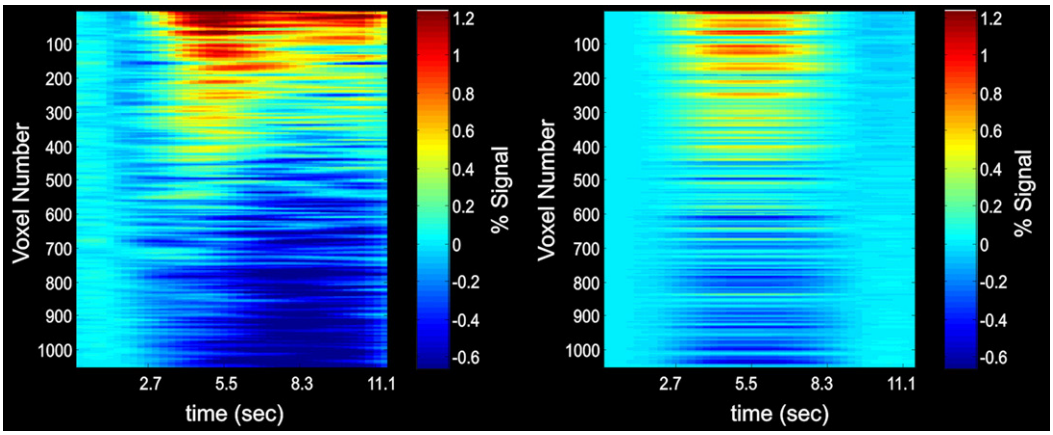

Fig. 2. Effects of single trial pre-processing in the context of multi-voxel pattern analysis. The unpreprocessed response to a single trial in a region of 1000 voxels is depicted on the left using a BOLD image plot. The response of the same voxels to the same trial after preprocessing is depicted on the right. Note the removal of the (linear) trend present in many of the unpreprocessed voxels' responses.

SVMs (and ls-SVMs) have been modified in order to find nonlinear division boundaries. In this case, the data are first projected to some other Euclidean space H, using a nonlinear transformation, which we call φ:

$$\varphi : \Re^d \rightarrow H. \tag{14}$$

A linear solution is found in the space H. as in Eqs. (8) and (9). This corresponds to finding a discriminant function such that:

$$t_i\left(\mathbf{w}^T\varphi(\mathbf{x_i}) + b\right) > 0 \quad \forall i. \tag{15}$$

Classification of new samples is obtained evaluating:

$$sign\left(\mathbf{w}^T\varphi(\mathbf{x}_{new}) + b\right) \tag{16}$$

The use of Kernel methods (see Ref. [25] for details) allows replacing the function ϕ (used for nonlinear extensions) with the kernel matrix $\mathbf{K}$. One example is the radial basis function (RBF) kernel the expression of which is given by:

$$\mathbf{K}\left(\mathbf{x}_i, \mathbf{x}_j\right) = e^{-\left\|\mathrm{x}_i - \mathrm{x}_j\right\| / 2\sigma 2} \tag{17}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ refer to rows of the matrix $\mathbf{X_c}$.

It is important to note that, *for linear classifiers*, the optimized weight vector $\mathbf{w}$ provides a measure of discrimination ability of each single feature, as estimated with the training set. This vector $\mathbf{w}$, whose dimensionality is equal to the number of voxels, can thus be seen a *discriminative map* indicating the locations of the brain that are most informative with respect to the classification. If a nonlinear kernel is used for the classification maps are not easily obtained, as in this case, the decision boundary in the original space is nonlinear, and thus, its normal vector ($\mathbf{w}$) may assume different orientations in different points of the hypersurface.

In general, the application of machine learning techniques to fMRI entails four steps (see Ref. [11] and Fig. 1). First, the condition-related brain activity is extracted from preprocessed fMRI time series. Second, the set of features (voxels) that will enter the multivariate analysis are selected. These two steps define the matrix $\mathbf{X_c}$. 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. Fourth, a decision rule is applied and the accuracy of classification of unseen patterns is evaluated.

Generally the performances of classification are evaluated in cross-validation, thus repeating Steps 2, 3 and 4 with different partitions of the data.

In the following, we describe the initial steps of feature extraction and selection and present examples of classification techniques in fMRI.

3.1. Preprocessing and feature extraction

In applying pattern classification algorithms, brain activity measured in response to a stimulus or a cognitive state is represented as a point in a multidimensional space of voxels (MVP), i.e., as a row vector in the matrix $\mathbf{X_c}$. It is thus necessary, as a first step, to obtain a good estimate of "activity" at each voxel forming this vector. In block-designed experiments, the intensity at a single acquisition volume (TR) [13,17] can provide reliable estimates of the activity level. Note, however, that, as the estimated patterns from the same block can be highly correlated, this solution requires careful splitting of the data for assessing of the generalization performances (i.e., using volumes from the same blocks for both training and testing might result in overestimating accuracy levels). An alternative is to use the average intensity in multiple TRs [14,18]; a drawback of this method, however, is a reduction in the number of samples available for training and testing.

In the case of slow event-related design, such measures might not provide reliable estimates of the activity patterns. A possible solution is represented by considering, at each

stimulus presentation, a trial T_i (i=1…n), formed by N_{pre} and N_{post} temporal samples (before and after stimulus onset respectively) of the time course of activity. A trial estimate of the response at every voxel V_j (j=1…v) is then obtained by fitting a general linear model (GLM) with one predictor coding for the trial response. The trial-response predictor is obtained by convolution of a boxcar with a double-gamma hemodynamic response function (HRF) [27]. At every voxel, the corresponding regressor coefficient (beta) is taken to represent the trial response. To account for intertrial variability, the GLM estimation may be repeated (at each voxel) several times varying the dispersion or other parameters of the HRF and the best-fitting beta is selected as the representative of 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. The need of an activity estimate at each single trial makes the use of classification techniques challenging in the case of fast event-related designs.

In our experience, appropriate preprocessing at the single trial level can improve the activity estimates. In particular, removing within trial linear trends, e.g., by introducing an additional linear predictor in the GLM-based estimation procedure previously described may produce considerable improvements as shown in Fig. 2.

3.2. Feature selection

While classification of fMRI patterns can be *massively* multivariate and consider all brain voxels simultaneously (whole-brain approach, see Ref. [17]), several strategies can be employed that reduce the dimensionality of the multi-voxel space [13,16,18].

In fact, whole-brain approaches may be problematic when the aim of the analysis is the fine-grained discrimination between perceptual states [14,21]. In these cases, the proportion of voxels that convey the discriminative information is expected to be small, and thus, whole brain approaches seem suboptimal. Machine learning algorithms are known to degrade their performances when faced with many irrelevant features (voxels) (overfitting, [28,35]), 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.

A priori anatomical assumptions can be used to limit the analysis to a subset of voxels from one ROI ([13–15]). However, 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.

Alternatively, feature selection strategies based on univariate measures have been suggested. In particular, introducing the hypothesis that interesting patterns consist of voxels having a significant stimulus-related BOLD response compared to baseline levels justifies the reduction of the number of features based on the univariate selection of "active" voxels (F test). Furthermore, it simplifies the interpretation of the results as the analysis is restricted to voxels showing neurophysiologically understood responses. A more restrictive form of univariate feature selection is based on the further selection of the voxels based on their univariate "discrimination" ability (t test, Wilcoxon) [13,16,18]. Low-resolution representations of the original patterns obtained by averaging neighboring features in a predefined radius have also been used to reduce the dimensionality of the classification problem [16]. Any such method of voxel selection, though, does not consider the inherent multivariate nature of the fMRI data.

An interesting alternative to univariate feature selection is the local multivariate search approach proposed in Ref. [20]. 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, suboptimal 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).

Multivariate feature selection strategies can be summarized in three categories, multivariate filters, wrappers and embedded methods [28,35]. 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 the 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. Multivariate feature selection in the context of fMRI pattern recognition is mainly limited by computational complexity and time. An approach based on recursive feature elimination (RFE, [29]) and ls-SVM has been proposed by De Martino et al. (under revision) [30]. This method uses the classifier recursively and a feature-ranking criteria based on the absolute discriminative map ($|\mathbf{w}|$) to eliminate irrelevant voxels and estimate informative spatial patterns. To avoid overfitting, feature selection (univariate or multivariate) must always be performed on the training data set alone, and test data should not enter in any form in this procedure.

3.3. Classification of auditory evoked fMRI patterns

As an example, we will present here an application of pattern classification to an auditory fMRI experiment on the representation of sounds from various categories [31]. Functional data were collected at 3T (Siemens Allegra) with a T2-weighted gradient-echo, EPI sequence (23 axial slices, TR 3.6 s; FOV 256×256; matrix size 128×128, voxel

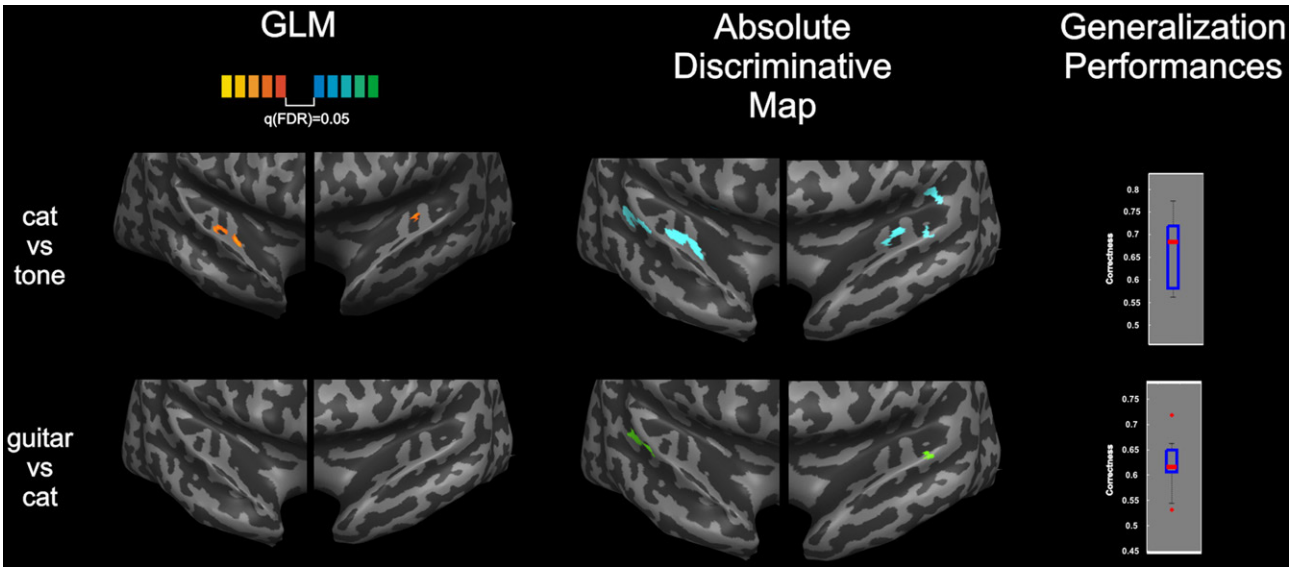

Fig. 3. Results from the pattern recognition analysis of two binary comparisons (cat vs. tone, top; guitar vs. cat, bottom) in the context of an fMRI study of the neural representation of sound categories. Absolute discriminitave maps (right) show the 20% most discriminative voxels and are compared to statistical parametric maps obtained using GLM analysis (voxelwise minimum t obtained in the two runs) and thresholded using FDR (q=0.05). Performances of the classifier indicate the percent of correct classification of single trials.

size=2×2×2 mm^3). Anatomical images were obtained using a high resolution (1×1×1 mm), T1-weighted sequence.

Stimuli consisted of 800-ms sounds of four different categories [cats, girls (singing female voices), guitars and tones). The sounds were matched not only in length and root mean square 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 14,440 ms. Stimulation blocks were followed by blocks of silence lasting 14,440 ms. Each run consisted of 15 trials per condition presented in a pseudorandom order and lasted 30 min approximately. Results presented in this article were obtained using two functional runs of one subject and focus only on two of the six possible binary comparisons, namely, cat vs. tone and guitar vs. cat (30 per condition). For each trial, activity was estimated at every voxel fitting a hemodynamic response function and correcting for within-trial linear drift.

Following an initial feature selection using univariate measures of activity (F test), a combination of linear ls-SVM and RFE was used to classify trials belonging to the two conditions following an n-fold cross validation (n=10) scheme to assess generalization performances. Fig. 3 shows some illustrative results from this pattern recognition analysis. A full account of these results is reported in [31]. Absolute discriminative maps showing the best 20% of the discriminative voxels are compared to statistical parametric maps obtained using GLM analysis (voxelwise minimum t obtained in the two runs) and corrected using false discovery rate (FDR, q=0.05). Performances of the classifier are reported in terms of the percent of correct classification of single trials (cat vs. tone=66%; guitar vs. cat=62%, median across the 10 folds). In the first comparison (cat vs. tone, upper panel), the univariate GLM-based contrast reveals significant differences of activation levels along the superior temporal gyrus. As expected, the same clusters highlighted by the univariate analysis are included in the discriminative maps using our recursive SVM approach. In the other comparison (guitar vs. cat, lower panel), the univariate GLM-based contrast fails to reveal any significant difference. Nevertheless, our recursive SVM approach reaches similar generalization performances as in the previous comparison and consistent activation patterns, indicating that the examined spatial patterns carry discriminative information between the two sound categories.

3.4. Classification of independent components fingerprints

A different application of machine learning techniques to the analysis of fMRI data is the one of automatically detecting and classifying the nature of independent components (ICs) extracted from the time series [32]. In this case, independent component analysis [3] and the extraction of the IC-*fingerprints* are the preliminary steps required to classify the components using a machine learning algorithm (non-linear ls-SVM, Fig. 4). The classifier is trained to distinguish among six different classes: (1) BOLD, (2) motion artifact (MOT), (3) large vessels artifact (VESSEL), (4) spatially distributed noise (SDN), (5) high temporal frequency noise (tHFN) and (6) EPI susceptibility artifact (EPI). This approach has been applied in several studies. In a visual experiment of structure from motion [33], 60 components were extracted in each of 14 experimental runs (seven subjects; two runs per subject). The ls-SVM classifier was

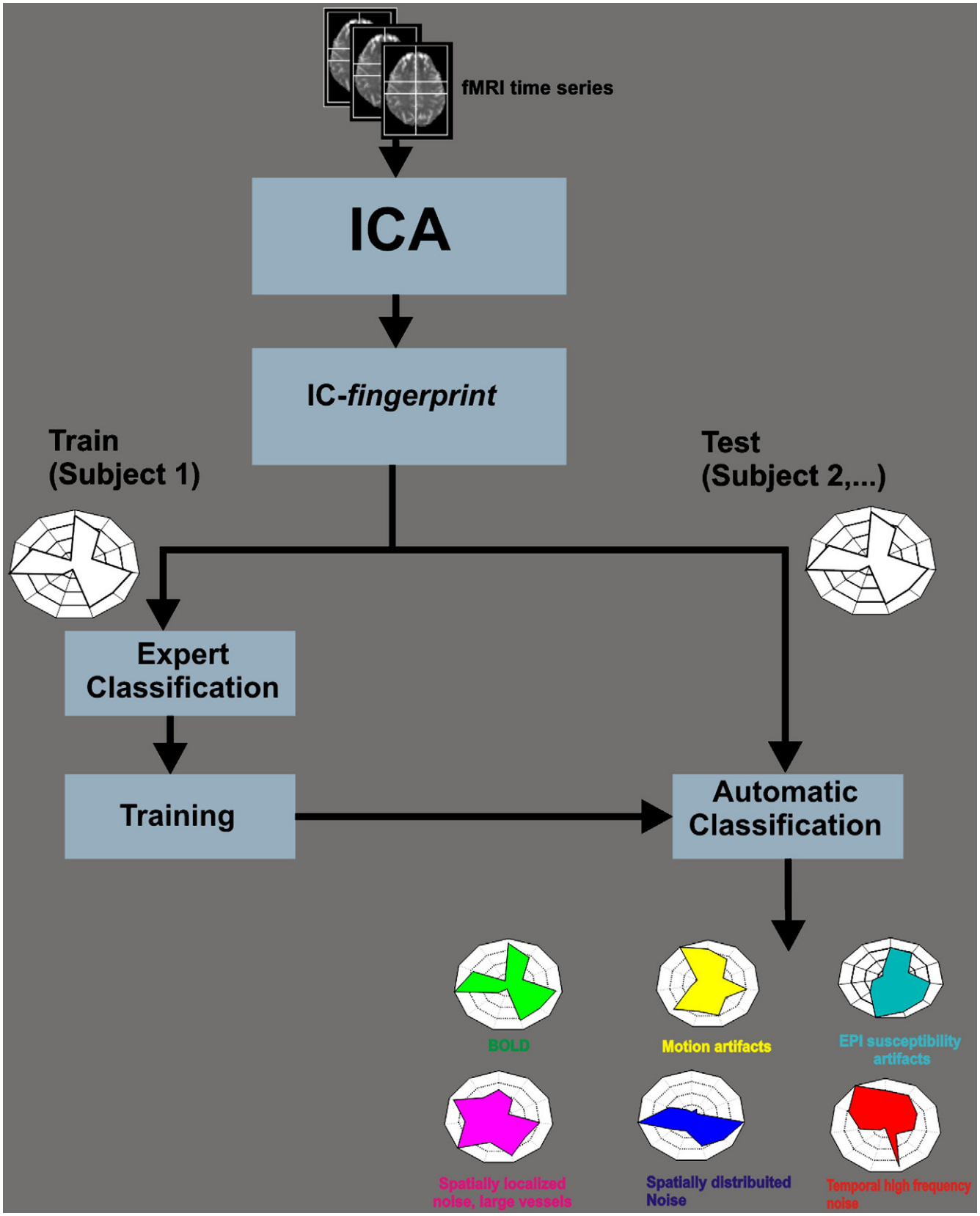

Fig. 4. Pattern recognition analysis applied to the classification of fMRI independent component *fingerprints*. After feature extraction (ICA and characterization of the *fingerprint*) the ICs from one run are labeled by an expert and used for training. IC *fingerprints* from new data (same or different experiments) are automatically classified in one of six classes.

Table 1
Performances of the ls-SVM based classifier on the classification of IC-*fingerprints* when compared with the expert classification

	BOLD		MOT	EPI	VESSEL	SDN	tHFN
True-positive rate	Task-related	Others					
	100% (100%)	90% (84%)	35% (48%)	61% (82%)	61% (60%)	100% (83%)	100% (74%)
False-positive rate	0% (0%)	4% (6%)	0% (5%)	2% (3%)	2% (10%)	4% (3%)	3% (1%)

For comparison, performances of a linear discriminant classifier are reported in parenthesis.
True-positive rate is defined as the ratio between the number of components correctly assigned to a class and the total number of components in that class.
False-positive rate is defined as the ratio between the components incorrectly assigned to a class and the total number of components not belonging to that class.

trained on the ICs extracted from the first run of one subject labeled in one of the six different classes by an expert. The remaining 13 data sets were automatically classified, and the resulting labels were compared to expert classification of the same ICs. Performances in terms of true-positive and false-positive rates are presented in Table 1. The automatic classifier reached a 94% true-positive rate for the class BOLD, signifying that, in 94% of the cases, the classification algorithm and the human expert identically labeled these components. Restricting the comparison to the task-related ICs (i.e., to the ICs with a clear interpretation), the overlap between expert and automatic classification increased to 100%. Similarly, a 100% true-positive rate was achieved for the classification of ICs in classes SDN and tHFN. For the classes EPI and VESSEL, the true-positive rate was 61%. Performances dropped for the class MOT (35% true-positive rate). The false-positive rate was computed as the ratio between the ICs incorrectly assigned to a class and the total number of components not belonging to that class. For each class, this rate was below 4%. Table 1 also reports the classification performance obtained using a linear discriminant analysis. When compared with the expert classification on the basis of the false-positive rate, we observe that, except for the MOTION and EPI class, the linear discriminant classifier performed slightly worse than the ls-SVM classifier. It is noteworthy that this same classifier has been successfully used for classifying, without retraining, the ICs obtained from fMRI data sets collected with different designs, field strength and acquisition parameters (see Refs. [32,34]) indicating the generality of the proposed approach.

4. Regression of fMRI responses

In the context of regression, the whole fMRI time series is employed, and the matrix of voxel patterns X thus corresponds to the data set D (that can be regarded as a $N \times v$ matrix, with N being the number of volumes and v the number of voxels). The data set D can therefore be seen as a collection of N pairs $(\mathbf{x_i}, t_i)$, denoting with $\mathbf{x_i}$ a sample vector of dimension v and with t_i the corresponding one-dimensional label.

Referring to Section 2 and Fig. 1, we will describe, in more detail, the steps that are usually performed for regression studies in fMRI pattern analysis. In Sections 4.1 and 4.2, the preprocessing and feature extraction/selection step commonly employed will be described, while in Section 4.3, some principles for designing and estimating models will be illustrated.

4.1. Preprocessing

The preprocessing of the fMRI data set usually consists of different steps. The first steps that are usually performed on the data sets are slice scan time correction, motion correction and a linear detrending of the time series. Particular attention must be paid to the coregistration of different data sets, due to the multivariate nature of the patterns employed for the analysis.

Other preprocessing steps, involving spatiotemporal filtering, can be considered separately from the previous steps, as they may be adjusted repeatedly according to the training of the model. Spatial smoothing tends to reduce the effects of isotropic noise in the data by performing an average of neighboring voxels. However, a strong amount of smoothing may reduce the effects of including multivariate information in the data. In the same way, temporal filtering helps reducing the information content of the data in those bands that are not relevant for the problem studied (i.e., that are not present in the labels). Again, an excessive amount of smoothing may suppress also information useful for the prediction. For these reasons, a safer way to address the issue of spatiotemporal filtering as a preprocessing strategy for fMRI brain reading is to evaluate different amounts of filtering on the data set and to choose the ones that provide the highest performance metric. Several filtering strategies are usually available, ranging from simple box-car smoothing functions to multiresolution approaches.

4.2. Feature extraction and selection

Given the preprocessed data set, the aim of this part is to move from the fMRI data set to a different representation of the data that may be more suitable for the regression scheme implemented. As pointed out in Ref. [22], there is no clear-cut conceptual boundary between the feature extraction and the pattern classification algorithm. In extreme cases, there may be a classification scheme so powerful that does not need any feature extraction, while there may be an ideal feature extraction scheme such that the work of the classifier becomes trivial.

In the context of regression for fMRI brain reading, different features can be considered, according to the model

employed. For instance, while using Kernel methods (that are not directly affected by the large dimensionality of the fMRI data), all the voxels' time courses can be considered as features. However, in many cases, some dimension reduction scheme may be employed. In fact, reducing the number of features (dimensions) helps in data understanding, reduces the memory storage requirements and training times and mitigates the effects of the *curse of dimensionality* [24], improving overall performances [35].

Feature selection algorithms employed in regression for brain reading are usually based on ranking methods, i.e., choosing a subset of the dimensions after ranking all of them according to some information criterion, like linear correlation or mutual information. The number of considered features may be fixed in advance, or it may be evaluated according to the model estimation procedure.

Another interesting strategy is the use of unsupervised learning approaches, like clustering and principal and independent component analysis (PCA/ICA) [22,36], to reduce the dimension of the data set and to consider more relevant representation of the time series. In the case of clustering, feature space dimension is reduced by grouping features together using a *similarity* criterion. In PCA, the features are projected into an orthogonal space where the maximum variance is explained, while ICA projects them on a space where they are maximally independent.

4.3. Model estimation

The main part of any pattern analysis approach is the use of a suitable model and its estimation. Let us assume that the steps described in Sections 4.1 and 4.2 have been performed. The main idea is to find an underlying generative model for the data that allows making prediction. The design of such a model is a crucial part of the learning process: a model which is too simple may fail to grasp the variability of the data, while a model which is too complex may fit also the noise component in the data, in case of relatively small training data set size (*overfitting*). Training data set size, model complexity, and performance metric are interconnected, and they must be taken into account while implementing a regression scheme.

Linear models have been extensively employed in machine learning for fMRI data analysis. There are many reasons for this choice. The high number of spatial points in fMRI data set, in fact, poses some challenges to the model estimation. If no feature reduction (Section 4.2) is performed, the dimension D of the feature space (i.e., the number of voxels) is usually considerably higher than the amount n of samples, and therefore, the effects of the curse of dimensionality are not negligible.

Moreover, it is possible to assess the relative "importance" of the single features (or dimensions of $\mathbf{x}$) helping improve the understanding of the relevance of different regions of the brain. With linear models, this step is straightforward and considerably easier when compared to nonlinear models, leading to a clearer interpretation of the ongoing neural processes.

A standard linear model has the following form:

$$t = y(\mathbf{x},\mathbf{w}) + \varepsilon \tag{18}$$

where $y(\mathbf{x}, \mathbf{w})$ is the deterministic input-output mapping part, and ε accounts for the noise in the measurements. The deterministic mapping can be modeled as [24]:

$$y(\mathbf{x},\mathbf{w}) = w_0 + w_1 x_1 + \ldots + w_D x_D = \mathbf{w}^{\mathbf{T}}\tilde{\mathbf{x}} \tag{19}$$

where $\mathbf{x}=(x_1,\ldots,x_D)^T$ denotes as usual the training data set (defined over a D-dimensional space), $\tilde{\mathbf{x}}=(1,\mathbf{x}^{\mathbf{T}})^{\mathbf{T}}$ and the D+1-dimensional vector $\mathbf{w}$ indicates the weights of the linear model (with w_0 indicating the bias term). This model, simply known as *linear regression*, is widely employed in fMRI data analysis.

However, for the purpose of generalization, we will consider now a mapping

$$\phi(\mathbf{x}) = (\phi_0(\mathbf{x}), \phi_1(\mathbf{x}), \ldots, \phi_{M-1}(\mathbf{x}))^T \tag{20}$$

with φ: $\Re^D \rightarrow \Re^M$, mapping the D-dimensional space of $\mathbf{x}$ into an M-dimensional one. ϕ can be, for instance, a linear polynomial, or RBF. Eq. (19) then becomes:

$$y(\mathbf{x},\mathbf{w}) = \mathbf{w}^{\mathbf{T}}\phi(\tilde{\mathbf{x}}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) \tag{21}$$

where this time, $\mathbf{w}$ is an M-dimensional vector of parameters.

The aim of the estimation procedure is to find the "best" model parameters (i.e., the "best" $\mathbf{w}$). Defining which is the best model is anyway not trivial, as the performance is evaluated in terms of an unknown data set. One criterion could be to maximize the fit of the model to the training data. Anyway, this procedure may be dangerous, as complex models may fit also the noise term.

One common error function in the case of regression is the sum of squares:

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N} \left\{t_n - \mathbf{w}^T\phi(\mathbf{x}_n)\right\}^2. \tag{22}$$

The minimization of this function (setting its gradient to zero) leads to the following estimate of the model parameters:

$$\tilde{\mathbf{w}} = \left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^T\mathbf{t} \tag{23}$$

where $\mathbf{\Phi}=(\phi(\mathbf{x}_1),\ldots,\phi(\mathbf{x}_N))^T$. It can be shown that the least-squares solution corresponds to the projection of the target t onto the subspace generated by the columns of $\mathbf{\Phi}$ [24]. As discussed previously, a perfect fit on the training data set may not be optimal for generalization purposes; therefore, some *regularization* coefficients are introduced, to control

for the smoothness of the estimate. The new error function will then be:

$$E(\mathbf{w}) = E_D(\mathbf{w}) + \lambda E_W(\mathbf{w}) \tag{24}$$

where E_D is the same as in Eq. (22) and E_W is the regularization term. A simple form of regularizing term is the following [24]:

$$E_W(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} \tag{25}$$

that leads to the solution:

$$\tilde{\mathbf{w}} = \left(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\mathbf{I}\right)^{-1}\boldsymbol{\Phi}^T\mathbf{t} \tag{26}$$

that is sometimes called *ridge regression* solution. Regularization is particularly effective in training on small data sets (reducing the model complexity and subsequently the risk of overfitting), but one has to employ a suitable value for the weighting coefficient λ. One way to set this parameter is to perform cross-validation choosing the parameter that gives the highest generalization on the validation set(s).

It is possible to derive the solutions encountered in Eqs. (23) and (26) from a probabilistic model. Consider Eq. (18), and assume that the noise term ε follows an independent, identically distributed Gaussian distribution with zero mean and precision (inverse of the variance) equal to β, that is, $p(\varepsilon|\beta)=\mathrm{N}(0,\beta^{-1})$. Considering the assumption of independent and identically distributed (i.i.d.) random variables, it follows that

$$p(\mathbf{t}|\mathbf{x},\mathbf{w},\beta) = \prod_{n=1}^{N} N\left(\mathbf{w}^{\mathrm{T}}/(\mathbf{x}_{\mathrm{n}}), \mathrm{b}^{-1}\right) \tag{27}$$

where with $\mathbf{t}$ we denote the vector of all the N targets.

Using Bayes' rule, it is possible to express the probability of the model parameters:

$$p(\mathbf{w}|\mathbf{t},\mathbf{x},\beta) = \frac{p(\mathbf{t}|\mathbf{w},\mathbf{x},\beta)p(\mathbf{w})}{p(\mathbf{t}|\mathbf{x},\beta)} \tag{28}$$

that is:

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{mariginal likelihood}} \tag{29}$$

where the prior term contains all the information one has on the model parameters. For a review of Bayesian methods refer to Refs. [22,24].

A Gaussian prior is commonly employed in regression.

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{m}_0, \boldsymbol{\Sigma}_0) \tag{30}$$

This has some computational advantages, as the posterior distribution [Eq. (30)] is Gaussian as well. A fully Bayesian perspective on the problem does not require the use of point estimate on the model. In fact, the prediction is performed averaging across *all* the possible models weighted by their probabilities. In other words, considering a new data point $\mathbf{x}^*$ then the predicted value t^* will be distributed according to the following:

$$p\left(t^*|\mathbf{x}^*,\mathbf{x},\mathbf{t}\right) = \int p\left(t^*|\mathbf{x}^*,\mathbf{w}\right)p(\mathbf{w}|\mathbf{x},\mathbf{t})d\mathbf{w} \tag{31}$$

that, in the case of Gaussian prior, is still has a Gaussian distribution [23,24].

4.4. An example: PBAIC 2006 competition

As an example we will illustrate an application of regression in fMRI brain reading using the PBAIC 2006 competition data set [37] (http://www.ebc.pitt.edu/). This data set is particularly useful in evaluating and comparing different learning algorithms. The data set consists of three subjects with three functional runs and an anatomical scan each. During the experiment, the subjects viewed freely an audiovisual movie (taken from episodes of a popular TV series). After the scanning sessions, the subjects were asked to rate the movie in terms of the content of different features (e.g. the presence of faces, motion, language, etc.), which produced continuous labels associated with the collected time series. The labels relative to the first two acquisition runs (subsampled at TR level and convolved with an HRF estimate) were provided together with the functional data. The third run (*test set*) was provided without the labels. The aim of the project was thus to train models on the first two runs and make predictions on the last run. The measure of accuracy was the correlation between the labels predicted on the basis of the fMRI data analysis and ones obtained from the subjects. The overall score consisted of an average of the score on each rating of each subject, after a Fisher Z-score transformation. Several types of preprocessing, feature extraction and learning algorithms, have been proposed (see http://www.ebc.pitt.edu/).

In this section, we will show an example with the application of the relevance vector machine (RVM) [38] to predict a selected rating ("language"), illustrating the approach using the two training runs of a single subject (Subject 1). Generalization performances are estimating considering the accuracies obtained predicting the second run after training on the first run and vice versa. The steps that are performed involve the preprocessing of the functional time series, the training of the model and the final prediction.

4.4.1. Preprocessing

Several preprocessing steps have been performed using Brainvoyager QX (Brain Innovation, Maastricht, The Netherlands). Those steps involved slice scan time correction, head movement correction with alignment across different runs and linear detrending of the functional time series. A coregistration of the functional to the anatomical scan (in Talairach space) was performed. We considered a volume mask and restricted our analysis only to the voxels within the brain. At the end of the preprocessing step, it is

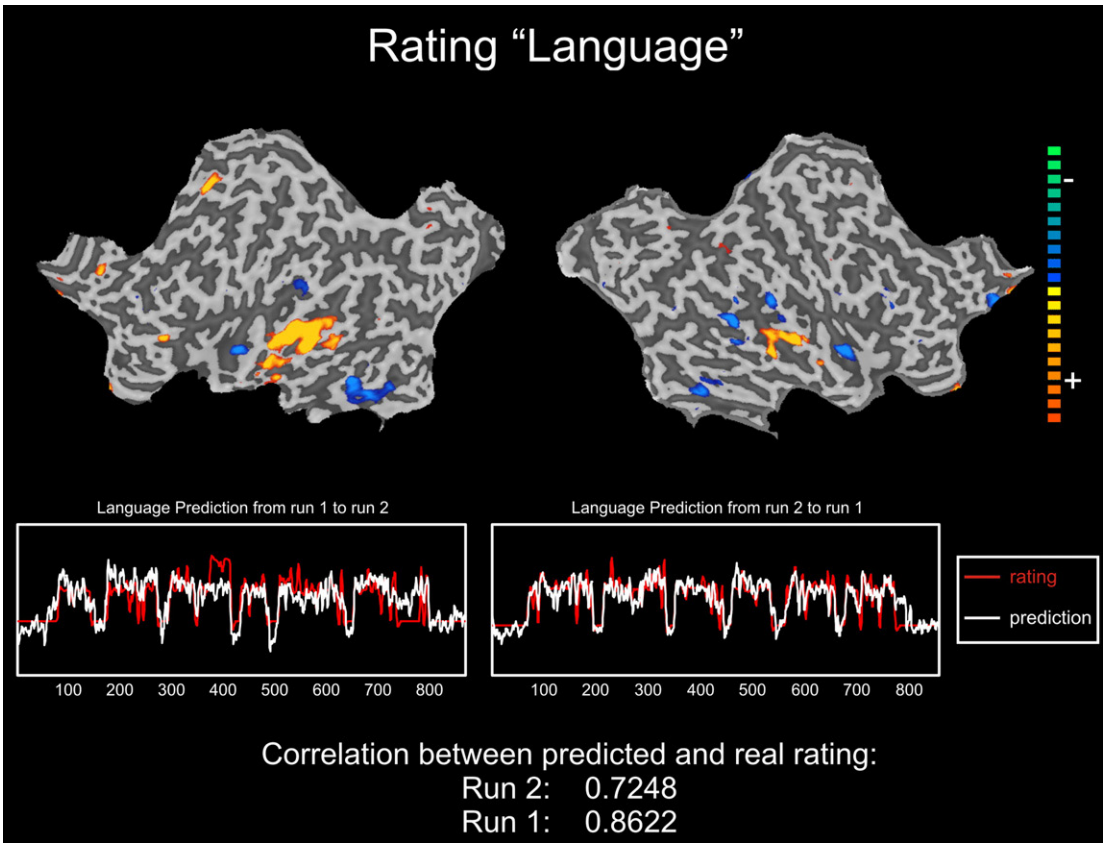

Fig. 5. An example of prediction results on the PBAIC 2006 competition data set (language rating). Predicted ratings and original ratings are presented on the bottom together with their correlation. The predictive map obtained from the training of the same rating on three subjects is presented on the top.

possible to consider for the subsequent analyses the two data sets $\mathbf{X}_1$ ($n_1 \times v$, where n_1 denotes the number of time points and v the number of voxels) and $\mathbf{X}_2$ ($n_2 \times v$) together with their labels $\mathbf{t}_1$ ($n_1 \times 1$) and $\mathbf{t}_2$ ($n_2 \times 1$).

4.4.2. Training and prediction

The model employed for this analysis is the RVM [38]. Similar to SVM [39], this method is based on the linear combination of kernel functions, with one kernel associated with each data point (in the training data set). Compared to SVM, RVM provides, in many applications a much sparser model, typically an order of magnitude more compact, with little or no reduction of generalization error. Furthermore, no parameter has to be estimated in cross-validation (like C and ε in SVM) [38].

The kernel can be defined starting from a nonlinear feature space mapping $\phi(\mathbf{x})$, as in Eq. (20):

$$k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^T \phi(\mathbf{x}'). \tag{32}$$

The model, with N+1 parameters, can be written as:

$$y(\mathbf{x}, \mathbf{w}) = \sum_{n=1}^{N} w_n k(\mathbf{x}, \mathbf{x}_n) + b \tag{33}$$

with b being the bias term and $k(\mathbf{x}, \mathbf{x}_n)$ the kernel function centered on $\mathbf{x}_n$. The assumption of Gaussian noise, Eq. (27), is used also for this algorithm. In RVM, the prior on the model weights $\mathbf{w}$ is:

$$p(\mathbf{w}|\alpha) = \prod_{i=1}^{N+1} N\left(0, \mathrm{a}_i^{-1}\right) \tag{34}$$

with a *hyperparameter* α_i for each model weight. It can be shown that posterior distribution of the weights is again Gaussian [38], with mean and covariances given by:

$$\mathbf{m} = \beta \boldsymbol{\Sigma} \boldsymbol{\Phi}^T \mathbf{t} \tag{35}$$

$$\boldsymbol{\Sigma} = \left(\mathbf{A} + \beta \boldsymbol{\Phi}^T \boldsymbol{\Phi}\right)^{-1} \tag{36}$$

with $\boldsymbol{\Phi}$ being the design matrix as in Eq. (23) and $\mathbf{A}$=diag (α_i). The estimation of α incorporates the *automatic relevance determination* [40,41]. In fact, during the training phase, many hyperparameters α_i will grow to infinity, so that the corresponding model weight w_i will have a posterior distribution concentrated around zero. In other words, only the model weights (and therefore, the functions associated with these parameters) that are "relevant" given the training data will remain, pruning out the unnecessary ones and leading to a sparse model. *Relevance vectors* can be seen as similar to the *support vectors* in the SVM formulation.

The values of α and β are determined using type II maximum likelihood (known also as *evidence approximation*) [24,38]. Once these parameters have been estimated, the prediction over a new data point $\mathbf{x}^*$ can be done as in Eq. (31), having a predictive distribution that is still Gaussian, with mean and variance given by [24,38]:

$$\mathbf{m}(\mathbf{x}^*) = \mathbf{m}^T \phi(\mathbf{x}^*) \tag{37}$$

$$\sigma^2(\mathbf{x}^*) = (\beta)^{-1} + \phi(\mathbf{x}^*)^T \boldsymbol{\Sigma} \phi(\mathbf{x}^*). \tag{38}$$

Without loss of generality, we refer to the training of the model on run 1. Suppose that both the functional time series and the rating have zero mean (note that the evaluation metric is based on Pearson correlation). Following the formulation proposed in Eq. (33), and considering a linear kernel, we have the following model:

$$y(\mathbf{X}_1, \mathbf{w}) = \mathbf{X}_1 \mathbf{X}_1^{\mathrm{T}} \mathbf{w} = \mathbf{K}\mathbf{w} \tag{39}$$

with $\mathbf{w}$ ($n_1 \times 1$) being the model weights vector and $\mathbf{K}=\mathbf{X}_1\mathbf{X}_1^{\mathrm{T}}$ ($n_1 \times n_1$) the linear kernel constructed from the starting data set.

The RVM training aims at finding an estimate of the posterior distribution of the weights $\mathbf{w}$ [see Eq. (28)]. This posterior distribution can be then used to perform predictions on a new data set (run 2) by means of Eq. (31). Denoting with $\tilde{\mathbf{w}}$ the estimated posterior mean, then the mean of the predictive distribution [Eq. (37)] is then:

$$\tilde{\mathbf{t}}_2 = \mathbf{X}_2 \mathbf{X}_1^{\mathrm{T}} \tilde{\mathbf{w}} \tag{40}$$

where $\tilde{\mathbf{t}}_2 (n_2 \times 1)$ is the estimate of the ratings on the second data set. It is possible, considering Eq. (40), to express the prediction in terms of maps:

$$\tilde{\mathbf{t}}_2 = \mathbf{X}_2 \tilde{\mathbf{M}} \tag{41}$$

with

$$\tilde{\mathbf{M}} = \mathbf{X}_1^{\mathrm{T}} \tilde{\mathbf{w}} \tag{42}$$

where $\tilde{\mathbf{M}}$(v×1) can be interpreted as a map of relative contribution of the different voxels to the final prediction.

Fig. 5 (lower panel) illustrates the predictions obtained for the rating language. Predicted ratings (white line) as obtained by Eq. (41) closely reflected the real ratings (red line), as indicated by the very high correlation values obtained both in the prediction of run 2 from run 1 and of run 1 from run 2. Fig. 5 (upper panel) also illustrates the predictive maps $\tilde{\mathbf{M}}$ [Eq. (42)], averaged across the three subjects and the two runs. In accordance with results from many neuroimaging studies of language processing, the most prominent clusters in the predictive maps are located bilaterally along the superior temporal sulcus and gyrus with some lateralization on the left hemisphere.

5. Conclusions

In this article, we have described the mathematical foundations of machine learning and pattern recognition techniques as employed in functional MRI data analysis. In the context of classification of fMRI responses, we have illustrated the use of SVMs. Whole-brain applications of these methods pose a relevant methodological challenge in terms of optimal feature selection. Albeit computationally demanding, the combination of SVM with recursive feature elimination and other multivariate feature-ranking approaches may provide a valuable solution. In the context of regression of fMRI responses, we have illustrated the use of RVM, a model of sparse Bayesian learning, for predicting labels obtained by continuous subjective ratings of a movie. RVM appears to be a very promising tool, opening new avenues in the analysis of very complex data sets, as those obtained with naturalistic and "real life" stimulation.

Acknowledgments

The authors are grateful to Martin Frost for comments on the manuscript. Financial support from NWO (MaGW-VIDI grant 452-04-330) to EF and from the Experience Based Cognition project (University of Pittsburgh) to EF and GV is gratefully acknowledged. Data used in this article to illustrate the regression of fMRI responses are part of the data set made available by the Pittsburgh Brain Activity Interpretation Competition (PBAIC) 2006, organized by Walter Schneider (University of Pittsburgh).

References

[1] Friston KJ, Holmes AP, Worsley KJ, Poline JB, Frith CD, Frackowiak RSJ. Statistical parametric maps in functional imaging: a general linear approach. Hum Brain Mapp 1995;2:189–210.
[2] Friston KJ. Functional and effective connectivity in neuroimaging: a synthesis. Hum Brain Mapp 1994;2:56–78.
[3] McKeown MJ, Makeig S, Brown GG, Jung T, Kindermann SS, Bell AJ, et al. Analysis of fMRI data by blind source separation into independent spatial components. Hum Brain Mapp 1998;6:160–88.
[4] Calhoun VD, Adali T, Pearlson GD, Pekar JJ. Spatial and temporal independent component analysis of functional MRI data containing a pair of task-related waveforms. Hum Brain Mapp 2001;13(1):43–53.
[5] Beckmann CF, Smith SM. Probabilistic independent component analysis for functional magnetic resonance imaging. IEEE Trans Med Imaging 2004;23(2):137–52.
[6] Formisano E, Esposito F, Di Salle F, Goebel R. Cortex-based independent component analysis of fMRI time series. Hum Brain Mapp 2004;22(10):1493–504.
[7] Smolders A, De Martino F, Staëren N, Scheunders P, Sijbers J, Goebel R, et al. Dissecting cognitive stages with time-resolved fMRI data: a comparison of fuzzy clustering and independent component analysis. Magn Reson Imaging 2007;25:860–8.
[8] McIntosh AR, Gonzalez-Lima F. Structural equation modeling and its application to network analysis in functional brain imaging. Hum Brain Mapp 1994;2:2–22.
[9] Friston KJ, Harrison L, Penny W. Dynamic causal modelling. Neuroimage 2003;19(4):1273–302.

[10] Roebroeck A, Formisano E, Goebel R. Mapping directed influence over the brain using Granger causality and fMRI. Neuroimage 2005; 25:230–42.
[11] Norman KA, Polyn SM, Detre GJ, Haxby JV. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn Sci 2006;10(9):424–30.
[12] Haxby JV, Gobbini MI, Furey ML, Ishai A, Aschouten JL, Pietrini P. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 2001;293(5539):2425–30.
[13] Haynes JD, Rees G. Predicting the orientation of invisible stimuli from activity in human primary visual cortex. Nat Neurosci 2005;8(5):686–91.
[14] Kamitani Y, Tong F. Decoding the visual and subjective contents of the human brain. Nat Neurosci 2005;8(5):679–85.
[15] Cox D, Savoy R. Functional magnetic resonance (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. Neuroimage 2003;19(2):261–70.
[16] Mitchell TM, Hutchinson R, Niculescu RS, Pereira F, Wang X. Learning to decode cognitive states from brain images. Mach Learn 2004;57:145–75.
[17] Mourao-Miranda J, Bokde AL, Born C, Hampel H, Stetter M. Classifying brain states and determining the discriminating activation patterns: support vector machine on functional MRI data. Neuroimage 2005;28(4):980–95.
[18] Mourao-Miranda J, Reynaud E, McGlone F, Calvert G, Brammer M. The impact of temporal compression and space selection on SVM analysis of single-subject and multi-subject fMRI data. Neuroimage 2006;33(4):1055–65.
[19] LaConte S, Strother S, Cherkassky V, Anderson J, Hu X. Support vector machines for temporal classification of block design fMRI data. Neuroimage 2005;26(2):317–29.
[20] Kreiegeskorte N, Goebel R, Bandettini P. Information-based functional brain mapping. Proc Natl Acad Sci U S A 2006;103(10):3863–8.
[21] Haynes JD, Rees G. Decoding mental states from brain activity in humans. Nat Rev Neurosci 2006;7(7):523–34.
[22] Duda RO, Hart PE, Stork DG. Pattern classification. 2nd ed. New York: John Wiley & Sons; 2001.
[23] Rasmussen CE, Williams CKI. Gaussian processes for machine learning. Cambridge (MA): MIT Press; 2006.
[24] Bishop CM. Pattern recognition and machine learning. New York: Springer; 2006.
[25] Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods. New York: Cambridge University Press; 2000.
[26] Suykens JAK, Van Gestel T, De Barbanter J, De Moor B, Vanderwalle J. Least squares support vector machines. Singapore: World Scientific Publishing; 2002.
[27] Friston KJ, Fletcher P, Josephs O, Holmes A, Rugg MD, Turner R. Event-related fMRI: characterizing differential responses. Neuroimage 1998;7(1):30–40.
[28] Kohavi R, John G. Wrappers for feature selection. Artif Intell 1997;97 (1-2):273–324.
[29] Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn 2002;46: 389–422.
[30] De Martino F, Valente G, Staëren N, Ashburner J, Goebel R, Formisano E. Combining multivariate voxel selection and Support Vector Machines for mapping and classification of fMRI spatial patterns (Under revision).
[31] Staëren N, Renvall H, De Martino F, Goebel R, Formisano E. Distributed representations of sound categories in the human auditory cortex (submitted).
[32] De Martino F, Gentile F, Esposito F, Balsi M, Di Salle F, Goebel R, et al. Classification of fMRI independent components using IC-*fingerprints* and support vector machine classifiers. Neuroimage 2007;34(1):177–94.
[33] Kriegeskorte N, Sorger B, Naumer M, Schwarzbach J, van den Boogert E, Hussy W, et al. Human cortical object recognition from a visual motion flowfield. J Neurosci 2003;23:1451–63.
[34] Rodionov R, De Martino F, Laufs H, Carmichael DW, Formisano E, Walker M, et al. Independent component analysis of interictal fMRI in focal epilepsy: comparison with general linear model-based EEG-correlated fMRI. Neuroimage 2007;38(3):488–550.
[35] Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003;3:1157–82.
[36] Hyvärinen A, Karhunen J, Oja E. Independent Component Analysis. New York: John Wiley & Sons; 2001.
[37] Editorial. What's on your mind? Nat Neurosci 2006;9(8):981.
[38] Tipping ME. Sparse Bayesian learning and the relevance vector machine. J Mach Learn Res 2001;1:211–44.
[39] Vapnik VN. The nature of statistical learning theory. New York: Springer-Verlag; 1995.
[40] MacKay DJC. Bayesian methods for backpropagation networks. In: Domany E, van Hemmen JL, Schulten K, editors. Models of neural networks III. New York: Springer-Verlag; 1994. p. 211–54. ch. 6.
[41] Neal RM. Bayesian learning for neural networks. New York: Springer; 1996.