NeuroImage xxx (2010) xxx–xxx

Contents lists available at ScienceDirect

NeuroImage

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

Sparse logistic regression for whole-brain classification of fMRI data

Srikanth Ryali [a,*], Kaustubh Supekar [b,c], Daniel A. Abrams [a], Vinod Menon [a,d]

[a] *Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Stanford, CA 94305, USA*
[b] *Graduate Program in Biomedical Informatics, Stanford University School of Medicine, Stanford, CA 94305, USA*
[c] *Center for Biomedical Informatics Research, Stanford University School of Medicine, Stanford, CA 94305, USA*
[d] *Program in Neuroscience, Stanford University School of Medicine, Stanford, CA 94305, USA*

ARTICLE INFO

Article history:
Received 6 August 2009
Revised 9 February 2010
Accepted 16 February 2010
Available online xxxx

Keywords:
Classification
Logistic regression
Regularization

ABSTRACT

Multivariate pattern recognition methods are increasingly being used to identify multiregional brain activity patterns that collectively discriminate one cognitive condition or experimental group from another, using fMRI data. The performance of these methods is often limited because the number of regions considered in the analysis of fMRI data is large compared to the number of observations (trials or participants). Existing methods that aim to tackle this dimensionality problem are less than optimal because they either over-fit the data or are computationally intractable. Here, we describe a novel method based on logistic regression using a combination of L1 and L2 norm regularization that more accurately estimates discriminative brain regions across multiple conditions or groups. The L1 norm, computed using a fast estimation procedure, ensures a fast, sparse and generalizable solution; the L2 norm ensures that correlated brain regions are included in the resulting solution, a critical aspect of fMRI data analysis often overlooked by existing methods. We first evaluate the performance of our method on simulated data and then examine its effectiveness in discriminating between well-matched music and speech stimuli. We also compared our procedures with other methods which use either L1-norm regularization alone or support vector machine-based feature elimination. On simulated data, our methods performed significantly better than existing methods across a wide range of contrast-to-noise ratios and feature prevalence rates. On experimental fMRI data, our methods were more effective in selectively isolating a distributed fronto-temporal network that distinguished between brain regions known to be involved in speech and music processing. These findings suggest that our method is not only computationally efficient, but it also achieves the twin objectives of identifying relevant discriminative brain regions and accurately classifying fMRI data.

© 2010 Elsevier Inc. All rights reserved.

Introduction

Multivariate pattern recognition (MPR) methods are rapidly becoming a popular tool for analyzing fMRI data (Cox and Savoy, 2003; De Martino et al., 2008; Haynes et al., 2007; Kriegeskorte et al., 2006; Mourao-Miranda et al., 2005; Pereira et al., 2009). These methods use fMRI data to detect activity patterns in brain regions that collectively discriminate one cognitive condition or participant group from another. Most fMRI studies that use MPR methods restrict the analysis to specific brain regions of interest (ROI) (Cox and Savoy, 2003; Haynes et al., 2007), however this approach is problematic if the ROIs are not known *a priori*. In these cases, a data-driven approach that incorporates multiple brain regions is desirable for several reasons. For one, it is possible that no single brain region can accurately discriminate given a set of experimental stimuli, task conditions or participant groups, and simultaneously incorporating multiple brain regions may be necessary to describe the distributed networks sub serving differential brain processes. Therefore, the MPR method used in fMRI data analysis should, ideally, consider activity patterns in all brain regions, and identify the subset of regions that discriminates between experimental conditions in an unbiased manner. Hereafter, we refer to MPR methods that include activity patterns across the entire brain as “whole-brain classifiers.”

Designing a whole-brain classifier presents a number of technical challenges since the number of regions considered in the analysis of fMRI data (“features”) is large compared to the number of observations (trials or participants). Typically, this results in over-fitting of the data, leading to high classification accuracies for data used in designing the classifier, but poor classification accuracies for independent “test” data. Furthermore, a common characteristic of fMRI data is that the number of brain regions involved in a given cognitive

Abbreviations: MPR, Multivariate pattern recognition methods; CNR, Contrast-to-noise ratio; IRWLS, Iterated readjusted weighted least squares; LR1, Logistic regression with L1 norm; LR12, Logistic regression with L1 and L2 norms; LR12-UST, LR12 with universal soft thresholding; SVM-RFE, Support vector machine with recursive feature elimination

[*] Corresponding author. Department of Psychiatry and Behavioral Sciences, 780 Welch Rd, Room 201, Stanford University School of Medicine, Stanford, CA 94305-5778, USA. Fax: +1 650 736 7200.
E-mail address: sryali@stanford.edu (S. Ryali).

1053-8119/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2010.02.040

task is typically small relative to the total number of brain regions. Selecting the brain regions that are most relevant in discriminating cognitive tasks/condition overcomes the problem of over-fitting and improves the generalization performance of the classifier. Furthermore, identifying these relevant regions is also critical for understanding which brain regions can discriminate between stimulus conditions. Taken together, the problem of whole-brain classification can be distilled to two key problems: (1) feature selection, or selection of only those relevant regions that discriminate between cognitive conditions, and (2) designing a classifier using these selected regions.

The problem of feature selection has been extensively studied by the machine learning community (Kohavi, 1997). The overall goal of feature selection is to identify subsets of features that are most useful in discriminating two or more conditions of interest. Existing methods for feature selection can be grouped in two categories: filter and wrapper (Guyon, 2003; Kohavi, 1997). In the filter strategy, features are selected independent of classification, and the selected features are then used in designing the classifier. The features are ranked based on univariate scores such as correlation or mutual information between a feature and an experimental manipulation. This strategy has been implemented in a number of fMRI studies (Haynes and Rees, 2005; Mitchell et al., 2004; Mourao-Miranda et al., 2006). A limitation of the filter strategy is that this method applies only univariate measures and therefore does not consider the relationships between features while selecting them. This is a major limitation since fMRI data is inherently multivariate, with strong spatial correlation between neighboring voxels. Furthermore, this method does not consider classifier performance in selecting features. In contrast, the wrapper strategy utilizes methods in which features are selected that maximize the performance of the classifier. The selected features are then used in designing the classifier, as in the support vector machine-based recursive feature elimination algorithm (SVM-RFE) developed by Guyon et al. (2002) and Guyon (2003). This method has been applied for feature selection and classification of fMRI data by De Martino et al. (2008). A weakness of this approach is that thresholds used to select features are arbitrary and different datasets may require different settings of thresholds (De Martino et al., 2008).

An alternative strategy was recently proposed to simultaneously address the problem of feature selection and classifier design (Krishnapuram et al., 2005; Tipping, 2001; Zou and Hastie, 2005). In this strategy, feature selection is included as part of the classifier design, ensuring efficient use of data and faster computation time since the classifier does not need to be repeatedly trained during feature selection. In this approach, regularization is used to prevent over-fitting of the data and thereby improve generalizability of the classifier. Regularization-based approaches have been successfully applied to problems such as EEG/MEG source localization (Phillips et al., 2002), classification of multi sensor EEG data (van Gerven et al., 2009) and gene selection in micro data analysis (Zou and Hastie, 2005). Moreover, these approaches are well-suited for the analysis of fMRI data which, as mentioned earlier, is characterized by a large number of features and limited training data. SVM based feature selection using L1, L2 or L0 regularization methods was also proposed in the literature (Bi et al., 2003; Perkins et al., 2003; Weston et al., 2003).

Here, we present a novel method LR12, based on logistic regression with a combination of L1 and L2 norm regularization to accurately estimate discriminative brain regions from whole-brain fMRI data. The use of L1 norm regularization results in sparse solutions, thereby helping in feature selection. However, when features are highly correlated, as in fMRI data, using only L1 norm regularization selects only a subset of relevant features. Using L2 norm regularization in addition to L1 helps in selecting all correlated and relevant voxels. Furthermore, our method uses a novel and fast component-wise update procedure to estimate discriminative brain regions; this procedure is used to maximize the logistic regression cost function that includes L1 and L2 norm regularization (Krishnapuram et al., 2005). The L1 norm and fast estimation procedure ensure rapid computation and a generalizable solution. The L2 norm provides additional benefit by including correlated brain regions in the solution, a critical step often overlooked by existing methods. We first evaluate the performance of our LR12 method, on simulated data and then examine its effectiveness in discriminating between well-matched music and speech stimuli. We also compared our procedures with other logistic regression methods and SVM-RFE.

Methods

Logistic regression with regularization

Logistic regression fits a separating hyper plane that is a linear function of input features between two conditions or classes. Here, we interchangeably use the terms conditions and class labels. Given a set of training data, the goal is (1) to estimate the hyper plane that accurately predicts the class label of a new example and (2) identify a subset of the features that is most informative about the class distinction. Let $x = [x_1, x_2, \ldots, x_p]^t \in \boldsymbol{R}^p$ be a vector of input features (voxels) and y (y is a binary variable which is either 0 or 1) be its class label. Let $\boldsymbol{D} = \{(x^i, y^i)\}$, $i = 1, 2, \ldots, N$ be a set of N training examples. Under the logistic regression framework, the probability that the i-th example belongs to class-1 is defined as

$$P\left(y^i = 1 \mid x^i, \boldsymbol{\theta}\right) = \boldsymbol{h}_{\boldsymbol{\theta}}\left(x^i\right) \tag{1}$$

where, $\boldsymbol{h}_{\boldsymbol{\theta}}(\boldsymbol{x})$ is a logistic function given by $\dfrac{1}{\exp(-\boldsymbol{\theta}^t x)}$ and $\boldsymbol{\theta} \epsilon \boldsymbol{R}^{\boldsymbol{p}}$ is a vector of weights associated with each feature. These weights are estimated from the training data $\boldsymbol{D}$ by using the maximum likelihood method wherein the following log-likelihood is maximized

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log P\left(y^i \mid x^i, \boldsymbol{\theta}\right). \tag{2}$$

The above cost function results in a solution that accurately predicts the class label of a new example. In the context of fMRI analysis, the prediction accuracy of this solution is limited because the number of features (voxels) is far greater than the number of observations ($p \gg N$). To overcome this problem, regularization can be applied by assuming a prior on the weights. In an ideal case, the regularization should force the weights to be large for features which are sensitive to class labels and exactly zero for other features. Such a constraint achieves the twin objectives of classifier design with good prediction accuracy and the automatic detection of relevant features, which is very important for interpreting brain imaging data.

A commonly used Gaussian prior on weights lead to L2 regularization and the corresponding cost function to be maximized is

$$L_g(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log P\left(y^i \mid x^i, \boldsymbol{\theta}\right) - \gamma \boldsymbol{\theta}^t \boldsymbol{\theta} \tag{3}$$

where, γ controls the degree of regularization. Maximizing this cost function results in a regularized solution wherein the magnitudes of weights corresponding to irrelevant features are reduced to small values but not exactly to zero. This cost function is also concave, which can be optimized using the conventional iterated readjusted weighted least squares (IRWLS). Another commonly used prior is the Laplacian, a sparsity promoting prior, which has been used successfully in regression analysis (Tibshirani, 1996). This prior makes weights corresponding to irrelevant features to be exactly zero. The cost function that needs to be maximized in this case is

$$L_\iota(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log P\left(y^i \mid x^i, \boldsymbol{\theta}\right) - \gamma |\boldsymbol{\theta}|_1 \tag{4}$$

Please cite this article as: Ryali, S., et al., Sparse logistic regression for whole-brain classification of fMRI data, NeuroImage (2010), doi:10.1016/j.neuroimage.2010.02.040

$$|\boldsymbol{\theta}|_1 = \sum_{k=1}^{p} |\boldsymbol{\theta}(\boldsymbol{k})| \tag{5}$$

where, the operator | | returns the absolute value of $|\boldsymbol{\theta}(\boldsymbol{k})|$. This cost function is also concave, but cannot be optimized using IRWLS since it is not differentiable at the origin. Optimizing this cost function results in a sparse solution when the features are uncorrelated. In the case of correlated features, which are the case in fMRI data wherein the adjacent voxels are highly correlated, only a subset of these correlated features is selected. However in the context of fMRI, we require all the regions (or features) to be selected which differentiate the two class conditions. This grouping effect can be introduced by combining L1 and L2 regularizations (Zou and Hastie, 2005). The required cost function to be maximized in this case is now

$$L_{\iota g}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log P\left(y^i \,|\, \boldsymbol{x}^i, \boldsymbol{\theta}\right) - y_1|\boldsymbol{\theta}|_1 - \gamma_2 \boldsymbol{\theta}^t \boldsymbol{\theta} \tag{6}$$

where the parameters γ_1 and γ_2 respectively control the degrees of L1 and L2 norm regularization. Maximizing this cost function results in a sparse solution even when the features are correlated. In the following section, we describe a novel bound optimization method we developed to maximize $L_{\iota g}(\boldsymbol{\theta})$. This bound optimization method does not require computing the inverse of the Hessian matrix at each iteration and has been applied successfully to maximize both $L_g(\boldsymbol{\theta})$ and $L_\iota(\boldsymbol{\theta})$ (Krishnapuram et al., 2005). It can be easily scaled to applications such as whole-brain classification where the feature dimension is very high.

Bound optimization

Let $L(\boldsymbol{\theta})$ be the cost function to be maximized. In the bound optimization approach, $L(\boldsymbol{\theta})$ is optimized by iteratively maximizing a surrogate function Q,

$$\hat{\boldsymbol{\theta}}^{k+1} = \arg\, maxQ\left(\boldsymbol{\theta} \,|\, \hat{\boldsymbol{\theta}}^k\right) \tag{7}$$

where, $\hat{\theta}^k$ is the solution at k-th iteration. This procedure monotonically increases the cost function at each iteration if Q satisfies the condition that $L(\boldsymbol{\theta}) - Q(\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^k)$ attains its minimum at $\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^k$ (Krishnapuram et al., 2005).

When $L(\boldsymbol{\theta})$ is concave, surrogate function $Q(\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^k)$ can be constructed by using a bound on the Hessian matrix $H(\boldsymbol{\theta})$. If there exists a nonnegative matrix $\boldsymbol{B}$ such that $H(\boldsymbol{\theta}) - \boldsymbol{B}$ is nonnegative then it can be shown that

$$Q\left(\boldsymbol{\theta} \,|\, \hat{\boldsymbol{\theta}}^k\right) = \boldsymbol{\theta}^t \left(g\left(\hat{\boldsymbol{\theta}}^k\right) - \boldsymbol{B}\hat{\boldsymbol{\theta}}^k\right) - \frac{1}{2}\boldsymbol{\theta}^t \boldsymbol{B}\boldsymbol{\theta} \tag{8}$$

is a valid surrogate function. $g(\boldsymbol{\theta})$ denotes the gradient of $L(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. The matrix $\boldsymbol{B}$ is given by (Krishnapuram et al., 2005)

$$\boldsymbol{B} = -0.25 \sum_{i=1}^{N} \left(x_i x_i^t\right) \tag{9}$$

The component-wise update procedure can be used to maximize Q. Specifically, the surrogate function Q is maximized with respect to one of the components of $\boldsymbol{\theta}$ while fixing the remaining components to their current values. This procedure avoids the inversion of the Hessian matrix. Since the cost function is concave in parameters, the global optimal solution is guaranteed. Most importantly, this approach can be used for both L1 and L2 regularizations and the combination of both. For joint regularization of L1 and L2, the surrogate cost function of $L_{\iota g}$ $(\boldsymbol{\theta})$ to be maximized is

$$Q\left(\boldsymbol{\theta} \,|\, \hat{\boldsymbol{\theta}}^k\right) = \boldsymbol{\theta}^t \left(g\left(\hat{\boldsymbol{\theta}}^k\right) - B\hat{\boldsymbol{\theta}}^k\right) + \frac{1}{2}\boldsymbol{\theta}^t B\boldsymbol{\theta} - \gamma_1|\boldsymbol{\theta}|_1 - \gamma_2 \boldsymbol{\theta}^t \boldsymbol{\theta}. \tag{10}$$

The update rule for the m-th component of θ is given by

$$\hat{\boldsymbol{\theta}}_m^{k+1} = soft\left(\frac{B_{mm}}{B_{mm} - \gamma_2}\,\hat{\boldsymbol{\theta}}_m^k - \frac{g_{m\left(\hat{\boldsymbol{\theta}}^k\right)}}{B_{mm} - \gamma_2}; \frac{-\gamma_1}{B_{mm} - \gamma_2}\right). \tag{11}$$

Here, only the m-th component of $\boldsymbol{\theta}$ is updated while all other components are held at their values in the previous iteration. $\boldsymbol{B}_{mm}$ denotes the m-th diagonal entry of $\boldsymbol{B}$, $\boldsymbol{g}_{\boldsymbol{m}(\hat{\boldsymbol{\theta}}^k)}$ is the m-th element of the gradient vector, $\mathbf{g}(\hat{\theta}^k)$, and

$$soft(\alpha, \delta) = sign(\alpha) \max\{0, |\alpha| - \delta\} \tag{12}$$

is a soft threshold function. This update equation ensures that the value of Q is non-decreasing at each iteration and is sufficient to guarantee monotonicity of the procedure.

Choice of γ_1 and γ_2

In Eq. (10), the parameters $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ respectively control the degree of L1 and L2 regularizations. The performance of the classifier and selection of features depends on the choice of these parameters. These parameters were derived from the data using a combination of grid search and a three-way cross validation procedure. This procedure consists of two nested loops. In the outer loop, the data was split into N_1 ($N_1 = 10$) folds. One fold was used as test data for estimating the generalizability of the classifier and was involved neither in determining the weights of the classifier nor in the estimation of the parameters. In the inner loop, the remaining $N_1 - 1$ folds were further divided into N_2 ($N_2 = 10$) folds. $N_2 - 1$ folds were used as the training data and the remaining fold was used as the validation data. For each combination of γ_1 and γ_2, we obtained the discriminative weights using the training data and estimate the class labels of the validation data. We repeated the above procedure N_2 times by leaving a different fold as validation and the remaining folds as the train data. We obtained the average classification accuracy of the classifier across the N_2 folds for every combination of γ_1 and γ_2. We chose that combination of γ_1 and γ_2 for which this accuracy was maximum. We then obtained the discriminative weights by training the classifier using all the N_2 folds with the optimal parameters obtained above. We estimated the class labels of the test data which was left out in the outer loop using these discriminative weights. We repeated the above procedure N_1 times by leaving a different fold as the test data. We estimated class labels of the test data at each of the N_1 folds and computed average classification accuracy obtained at each fold, termed here as the cross validation accuracy (CVA). We then computed the final discriminative weights using all the data with average parameters obtained in N_1 folds and evaluated the performance metrics such as sensitivity, false positive rates and accuracy in feature selection, described below, based on these weights. In the gird search, the value of γ_1 was varied logarithmically from 2^{-2} to 2^5 in steps of 2 and γ_2 is varied logarithmically between 10^{-1} to 10^4 in steps of 10. The optimal values are searched in a logarithmical grid to cover a wide range of values.

Feature selection using SVM (SVM-RFE)

Feature selection using SVM based recursive feature elimination was developed by Guyon et al. (2002). This method was applied for feature selection in fMRI by (De Martino et al., 2008). Feature selection and generalizability of this approach was estimated using the two-level cross validation procedure described in (De Martino et al., 2008). In this procedure, the data was divided in to N_1 ($N_1 = 10$) folds. One fold was used as test data which was used only to estimate the generalizability of the classifier and does not influence the computation of discriminative maps. The remaining $N_1 - 1$ folds were used as the training data. This training data was further divided into N_2 ($N_2 = 5$) splits. The discriminative weights were obtained by training the linear SVM classifier using $N_2 - 1$ splits leaving out one split. The above procedure was repeated N_2 times by leaving out one split at a time. Average absolute discriminative weights were then computed using the N_2 discriminative weights obtained above. Recursive feature elimination (RFE) was performed R ($R = 10$) times based on these average weights. At each feature selection level, voxels corresponding to the smallest rankings were discarded and the remaining voxels were used to train the classifier at next level. In our implementation we discarded 10% of the lowest ranking weights at each RFE level. The generalization performance at this feature selection was assessed using the test data which was left out. The entire procedure was repeated N_1 times by leaving out different fold as test data. Final generalization performances and discriminative maps of each RFE level were obtained as the average over N_1 folds. We selected the RFE level for which the generalization performance (CVA) was highest. To compute the performance metrics such as sensitivity, false positive rates and accuracy in feature selection, we used the discriminating weights computed in the following two ways. In the first approach, we used the average discriminative maps obtained as the average over N_1 folds at the RFE level at which the CVA was highest. This approach was also taken in (De Martino et al., 2008). Here, we refer to this approach as SVM-RFE1. In the second approach, we retrain the classifier using the entire dataset and obtain the discriminating weights by applying RFE up to the level at which CVA was maximum. We refer to the performance evaluation by this approach as SVM-RFE2. This approach of obtaining discriminating maps is similar to the one employed in sparse logistic regression method. We have reported the results obtained by this approach in addition to the first approach to have a fair comparison with sparse logistic regression methods.

Initial voxel reduction

De Martino et al. (2008) reported that cross validation accuracy improved with initial voxel reduction, particularly at lower CNRs (De Martino et al., 2008). To further examine this issue, we used a similar voxel reduction method and selected a subset of the most activated voxels in both classes. We applied this procedure to examine how the performance of these methods improves with respect to the case where no initial voxel reduction is applied. For initial voxel reduction, we applied the same univariate activation based method used by De Martino et al. (2008)). In this method, the voxels were sorted independently using a scoring function and the union of top N' voxels per class were selected. The score for v-th voxel in i-th class ($S_i(\nu)$) is defined as:

$$S_i(\nu) = \frac{\mu_i(\nu)}{\sqrt{\frac{\sigma_i^2}{n_i}}} \tag{13}$$

where, $\mu_i(\nu)$ and σ_i^2 are the mean and variance of v-th voxel computed across n_i observations in i-th class. Note that this initial voxel reduction was performed only on the training data in the cross validation procedure and no test data was used in this step.

Evaluation of classifier performance

The performance of the classifier on simulated datasets in selecting relevant features was assessed by computing the sensitivity, false positive rate, accuracy in feature selection and the 10-fold cross validation accuracy (CVA) based on the optimal parameters $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ in sparse logistic regression method and the best feature elimination level in SVM-RFE method. The performance metrics such as sensitivity, false positive rate and accuracy were computed as follows:

$$\textit{sensitivity} = \frac{TP}{TP + FN} \tag{14}$$

$$\textit{false positive rate} = \frac{FP}{TN + FP} \tag{15}$$

$$\textit{accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{16}$$

where, TP is the number of true positives, TN is the number of true negatives, FN is the number of false negatives and FP is the number of false positives. TP, FP, TN and FN were determined as follows:

(a) TP: By counting the number of non-zero discriminative weights in the discriminative regions of the simulated data.
(b) FP: By counting the number of non-zero discriminative weights in the non-discriminative regions of the simulated data.
(c) TN: By counting the number of discriminative weights which are exactly zero in the non-discriminative regions of the simulated data.
(d) FN: By counting the number of discriminative weights which are exactly zero in the discriminative regions of the simulated data.

In single subject analysis, CVA accuracy can be evaluated by training a classifier over several experimental runs. In group-level analysis, CVA can be evaluated across subjects performing two different experiments. The latter procedure was used in this study. Here we refer to logistic regression with L1-norm regularization as LR1 and logistic regression with both L1 and L2 norms as LR12. We also use a special case of LR12 where the parameter $\boldsymbol{\gamma}_2$ is set to a high value (10,000) and the optimal value of $\boldsymbol{\gamma}_1$ is found using the above cross validation procedure. We refer to this method as universal soft thresholding (LR12-UST) for reasons discussed elsewhere (Grosenick et al., 2008; Zou and Hastie, 2005).

Simulated data

The performance of LR12, LR12-UST, LR1 and SVM-RFE were assessed using simulated datasets. This data consists of two discriminating regions responding to two conditions but with different amplitudes. Simulated datasets were constructed by creating summary statistics (Z-scores) of fMRI time-series data and then adding signals in multiple predefined regions using procedures similar to those described by De Martino et al. (2008) and Wang (2009). The datasets were created at various contrast-to-noise ratios (CNRs) and prevalence rates.

In this simulated dataset, we create two classes (or conditions) with two non-overlapping discriminatory regions. This simulation method is similar to that used in (De Martino et al., 2008) but with the following extensions:

(a) We directly simulate summary statistics (Z-scores) rather than voxel time-series.
(b) We generated datasets with several different prevalence rates, rather than using just one rate.
(c) We introduce spatial correlations in both discriminating as is typically the case in fMRI data.
(d) The distribution of voxels within the "activated" regions was simulated with spatially contiguous correlations. In typical fMRI data, clusters of contiguous voxels respond to a condition.

Therefore, we simulate the voxels responding to these conditions as spatially correlated contiguous voxels.

We created discriminative regions in the following way: In region-1, the level of activations in class-1 is greater than that in class-2. In region-2, the levels of activation in class-1 are less than that in class-2. The differences between the levels are simulated in such a way that a fixed contrast-to-noise ratio is satisfied. Since in actual fMRI data, adjacent active voxels are spatially correlated, in our simulations we introduced spatial correlations among the discriminating voxels.

We simulated high spatial correlation (Pearson correlation coefficient $\rho=0.7$) for voxels in region-1, class-1; medium correlation ($\rho=0.5$) in region-1, class-2. Conversely, high spatial correlation ($\rho=0.7$) for voxels in region-2, class-2; medium correlation ($\rho=0.5$) in region-2, class-1. The other non-discriminating voxels in both classes have no spatial correlation.

More specifically, for class-1, region-1, s-th observation for i-th feature $X_i^{(s)}$ was simulated as follows:

$$X_i^{(s)} = Z_1 + \varepsilon_i \; i = 1 \ldots p_1, s = 1 \ldots 25 \quad (17)$$

where, Z_1 is chosen as 1 and p_1 is the number of discriminating features in class-1, region-1. ε_i, $i=1,\ldots,p_1$ were generated using Matlab's *mvnrnd* function where in the correlation between $\varepsilon_i^{\prime}s$ was set to 0.7 and variance of each ε_i was set to 1.

For class-2, region-1, s-th observation for i-th feature $X_i^{(s)}$ was simulated as follows:

$$X_i^{(s)} = Z_2 + \varepsilon_i \; i = 1 \ldots p_1, s = 1 \ldots 25 \quad (18)$$

where, Z_2 is chosen such that certain CNR is satisfied. Here, CNR is defined as

$$CNR = \frac{|Z_1 - Z_2|}{\sigma} \quad (19)$$

where, σ is noise variance which is set to 1. In class-2, $\varepsilon_i^{\prime}s$ were generated such that the spatial correlation between them was 0.5 and variance was 1. In non-discriminating regions, data was generated such that there is no correlation between voxels.

$$X_i^{(s)} = \epsilon_i, \; s = 1 \ldots 50 \quad (20)$$

$$\epsilon_i \sim N(0,1). \quad (21)$$

Data was generated similarly in region-2 in both classes but with the difference that the spatial correlations and activation levels in region-2, class-2 was greater than that in region-1, class-1 as mentioned earlier. We chose to introduce different spatial correlations in the same region across two classes in order to simulate spatially discriminative patterns in the data in addition to the discriminative features with respect to activation levels. We generated 25 observations ($s=25$) in each class.

The datasets were generated with CNR = 0.1, 0.3, 0.5, 0.75, 1, and 1.5 and for each CNR we generated datasets with feature prevalence rates of 0.5%, 1%, 2.5%, 5%, 10% 20%, 30%, 40% and 50%. The total number of voxels for each dataset was 40,000. Here, we define prevalence rate as the percentage of discriminating voxels in both regions compared to actual number of voxels.

Experimental data

We examined the performance of each method on fMRI data acquired from 20 participants during an auditory experiment involving music and speech stimuli. Music stimuli consisted of three familiar and three unfamiliar symphonic excerpts composed during the Classical or Romantic period, and speech stimuli were familiar and unfamiliar speeches (e.g., Martin Luther King, President Roosevelt) selected from a compilation of famous speeches of the 20th century (Various, 1991). All music and speech stimuli were digitized at 22,050 Hz sampling rate in 16-bit. A pilot study in a separate group of participants was used to select music and speech samples that were matched for emotional content, attention, memory, subjective interest, level of arousal, and familiarity (Abrams et al., submitted for publication).

Each music and speech excerpt was 22–30 s in length. To present the stimuli to the participants in the scanner, we programmed two runs (one each for music, and speech) into Eprime V1.0 (Psychological Software Tools, 2002). We counterbalanced and randomized the order of the individual excerpts.

Participants were instructed to press a button on a magnetic scanner-compatible button box whenever a sound excerpt ended. Response times were measured from the beginning of the experiment and the beginning of the excerpt. The button box malfunctioned in eight of the scans and recorded no data, but because the main purpose of the button press was to ensure that participants were paying attention, we retained those scans, and they were not statistically different from the other scans. All participants reported listening attentively to the music and speech stimuli.

Images were acquired on a 3 T GE Signa scanner using a standard GE whole-head coil (software Lx 8.3). Images were acquired every 2 s in two runs, each lasting 8 min, 4 s. A custom-built head holder was used to prevent head movement during the scan. Twenty-eight axial slices (4.0 mm. thick, 1.0 mm skip) parallel to the AC/PC line and covering the whole brain were imaged with a temporal resolution of 2 s using a T2*-weighted gradient-echo spiral in-out pulse sequence (TR = 2000 ms, TE = 30 ms, flip angle = 80°, 262 time frames and 224 time frames, respectively, and 2 interleaves). The field of view was 200 × 200 mm, and the matrix size was 64 × 64, providing an in-plane spatial resolution of 3.125 mm. To reduce blurring and signal loss arising from field in homogeneities, an automated high-order shimming method based on spiral in-out acquisitions was used before acquiring functional MRI scans (Kim et al., 2000). Images were reconstructed, by gridding interpolation and inverse Fourier transform, for each time point into 64 × 64 × 28 image matrices (voxel size 3.125 × 3.125 × 5.0 mm). A linear shim correction was applied separately for each slice during reconstruction using a magnetic field map acquired automatically by the pulse sequence at the beginning of the scan (Glover and Lai, 1998).

To aid in localization of the functional data, a high-resolution T1-weighted spoiled grass gradient recalled (SPGR) inversion-recovery 3D MRI sequence was used with the following parameters: TR = 35 ms; TE = 6.0 ms; flip angle = 45 °; 24 cm field of view; 124 slices in coronal plane; 256 × 192 matrix; 2 averages, acquired resolution = 1.5 × 0.9 × 1.1 mm. The images were reconstructed as a 124 × 256 × 256 matrix with a 1.5 × 0.9 × 0.9-mm spatial resolution. Structural and functional images were acquired in the same scan session.

Data were pre-processed using SPM5 (www.fil.ion.ucl.ac.uk/spm). Images were corrected for movement using least squares minimization without higher order corrections for spin history, and were then normalized to stereotaxic MNI coordinates using nonlinear transformations (Friston et al., 1996). Images were then resampled every 2 mm using sinc interpolation and smoothed with a 4-mm Gaussian kernel to reduce spatial noise. *T*-scores (*T*-maps) for the contrasts [Music − Rest] and [Speech − Rest] were computed for each subject using a general linear model. The *T*-maps computed for these two contrasts were then used for classification.

Results

We first compare the performance of LR12, LR12-UST, LR1, and SVM-RFE on simulated datasets by evaluating the sensitivity, false positive rate, accuracy in feature selection and cross validation accuracy provided by each of these methods at various CNRs and feature prevalence rates. We then compare these methods on experimental data.

Performance on simulated dataset

Fig. 1 shows 10-fold cross validation accuracies obtained using LR12, LR12-UST, LR1, SVM-RFE1 and SVM-RFE2 methods. For CNRs of 0.1 and 0.3, the CVAs obtained by these methods are only about chance level (0.5). The classification accuracies obtained by these methods improve for CNRs of 0.5 and above and are comparable.

Fig. 2 shows the accuracies in feature selection obtained by LR12, LR12-UST, LR1, SVM-RFE1 and SVM-RFE2 methods. Accuracies of LR12, LR12-UST and LR1 improved with the increase in CNRs. For CNRs of 0.5 and above, LR12 performed better than LR12-UST, LR1 and SVM-RFE at most of the prevalence rates. Between the two SVM-RFE methods, accuracies obtained by SVM-RFE2 were better than that achieved by SVM-RFE1.

Figs. 3 and 4 respectively show sensitivities and false positive rates obtained by each of the methods. For low CNRs of 0.1 and 0.3, the sensitivities obtained by all methods are poor. SVM-RFE1 shows higher sensitivity but with very high false positive rates as shown in Fig. 4. For CNRs of 0.5 and above, sparse logistic regression methods (LR12, LR12-UST and LR1) performed better than SVM-RFE1 and SVM-RFE2. Among the logistic regression methods, LR12 has higher overall performance with respect to accuracies in voxel selection as shown in Fig. 2. Between the two SVM-RFE methods, SVM-RFE1 resulted in higher sensitivities compared to SVM-RFE2 as shown in Fig. 3 but at the cost of higher false positives (Fig. 4).

Univariate methods based on general linear models are generally used to analyze fMRI data. These methods take only differences in activation levels in voxels between conditions while multivariate methods presented here consider both spatial and activation level differences in the data. In order to examine whether the conventional univariate approach is sensitive in finding discriminative voxels, we applied two-sample *T*-test on the simulated data at a *p*-value of 0.05, corrected for multiple comparisons using false discovery rate. The sensitivity of univariate two-sample *T*-test was poor compared to other methods at CNRs of 0.75 and below as shown in Fig. 3.

Effects of initial voxel reduction

We applied a voxel reduction step in conjunction with LR12 and SVM-RFE at a prevalence rate of 0.5%, identical to the rate used by De Martino et al. (2008). We selected a union of top 2000 voxels, corresponding to 10 times the number of discriminating voxels.

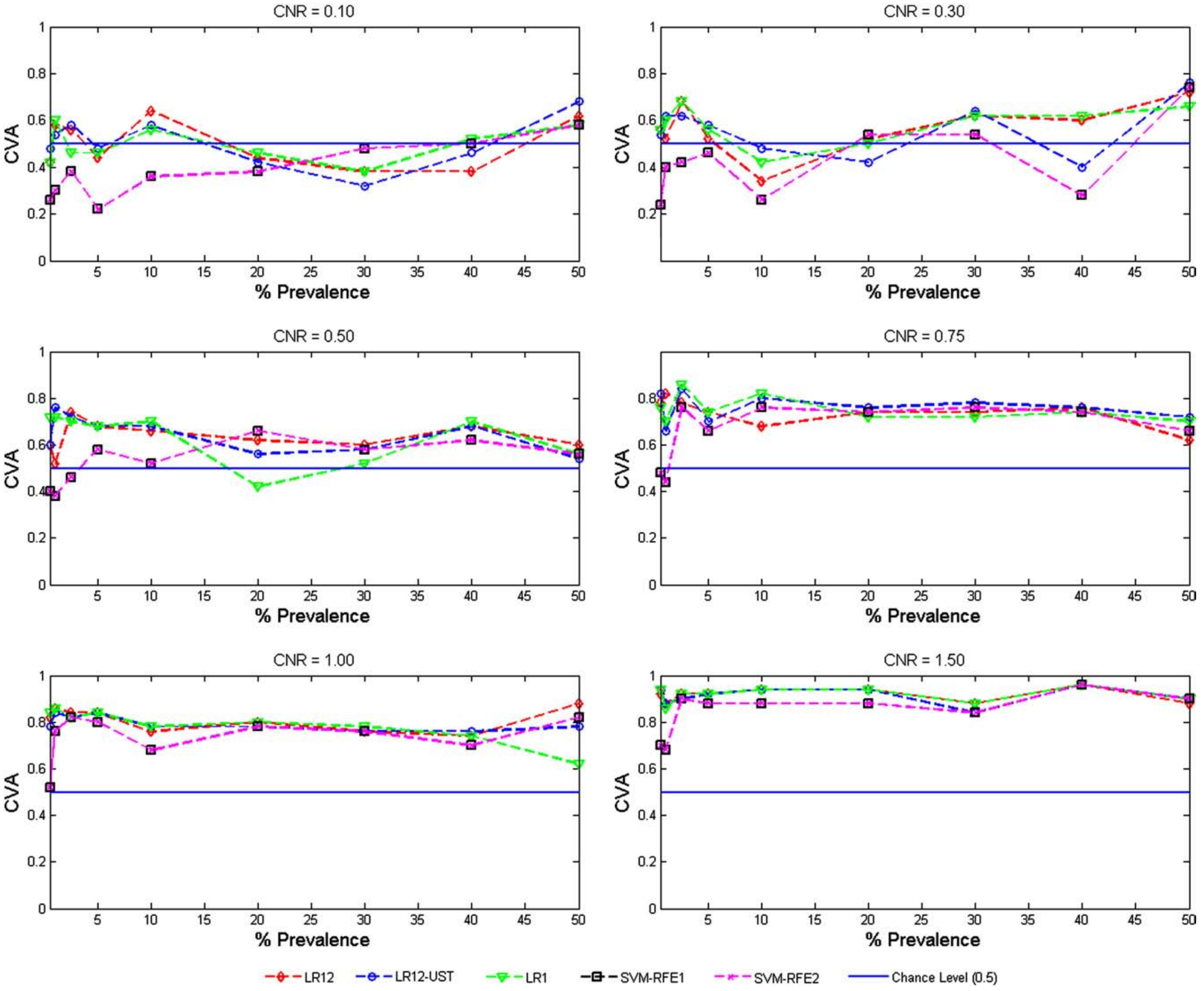

Fig. 1. 10-fold, 3-way, cross validation accuracy (CVA) obtained using LR12, LR12-UST, LR1, SVM-RFE1 and SVM-RFE2 at different CNRs and feature prevalence rates. Chance level is 0.5. CVAs are above chance level for only CNRs of 0.5 and above. CVAs obtained by all methods are comparable.

Please cite this article as: Ryali, S., et al., Sparse logistic regression for whole-brain classification of fMRI data, NeuroImage (2010), doi:10.1016/j.neuroimage.2010.02.040

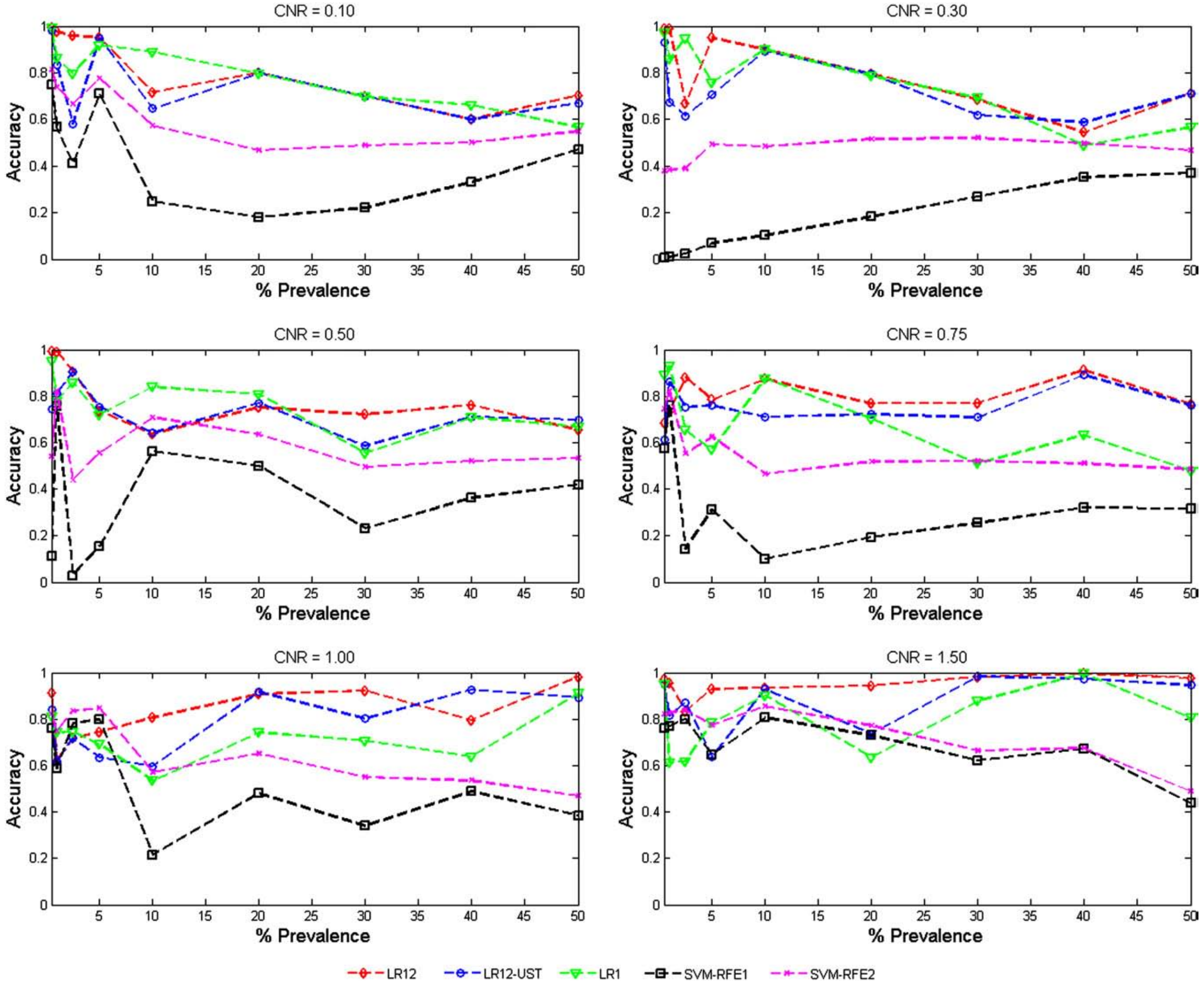

Fig. 2. Accuracy of feature selection obtained using LR12, LR12-UST, LR1, SVM-RFE1 and SVM-RFE2. LR12 has better accuracy compared to other methods for most CNRs and feature prevalence rates.

Table 1 compares the performance of these methods with and without voxel reduction step.

Table 1A shows that CVA improved with voxel reduction step for both LR12 and SVM-RFE at 0.5% prevalence rate, particularly at low CNRs (0.1–0.5). The improvement is more significant for LR12 at lower CNRs. The CVAs achieved by LR12 are higher than that of SVM-RFE with and without voxel reduction. Table 1B, C and D shows accuracies, sensitivity and false positive rates in voxel selection with and without voxel reduction step at 0.5% prevalence rate. In this case, the performance of SVM-RFE1 and SVM-RFE2 in voxel selection accuracy improved at both low and high CNRs (Table 1B) with voxel reduction. However, the sensitivity in voxel selection achieved by LR12 after voxel reduction is better than that of SVM-RFE1 and SVM-RFE2 (Table 1C) but at marginally higher false positives (Table 1D) for CNRs above 0.5. Although the false positive rates of SVM-RFE1 and SVM-RFE2 reduced with the initial voxel reduction (Table 1D) but their sensitivities decreased (Table 1C) compared to the case where there was no voxel reduction.

Performance on experimental fMRI data

We examined the performance of the four classification approaches on fMRI data from an auditory experiment examining neural processing of global acoustical differences between music and speech. Using the four classification methods, we quantified the cross validation accuracies for the Music versus Speech conditions. In addition to performing whole-brain analyses, we also performed the exact same analyses using a mask as a means of excluding deactivated voxels and including only those voxels which showed increased signal levels during music and/or speech stimuli (Supplemental Fig. S1).

LR12 and LR12-UST methods

LR12 and LR12-UST classified a distributed cortical network in the frontal, temporal, and parietal and occipital lobes, as shown in Figs. 5A and B, respectively. LR12 and LR12-UST methods identified nearly identical voxels throughout these cortical structures, with LR12-UST indentifying a slightly larger extent of voxels relative to LR12. Temporal lobe structures identified using these methods included large portions of bilateral anterior and posterior divisions of the middle and superior temporal gyri and temporal poles, as well as right-hemisphere planum temporale. Both methods also identified bilateral parahippocampal gyri, left-hemisphere hippocampus, amygdala, and putamen, as well as right-hemisphere insula. Frontal lobe structures identified using LR12 and LR12-UST methods included bilateral frontal orbital cortex (BA 47), frontal poles, and post-central gyri. In the parietal lobe, LR12 and LR12-UST methods identified bilateral angular gyri as well as a number of occipital cortical regions, including the occipital pole and inferior and superior lateral occipital

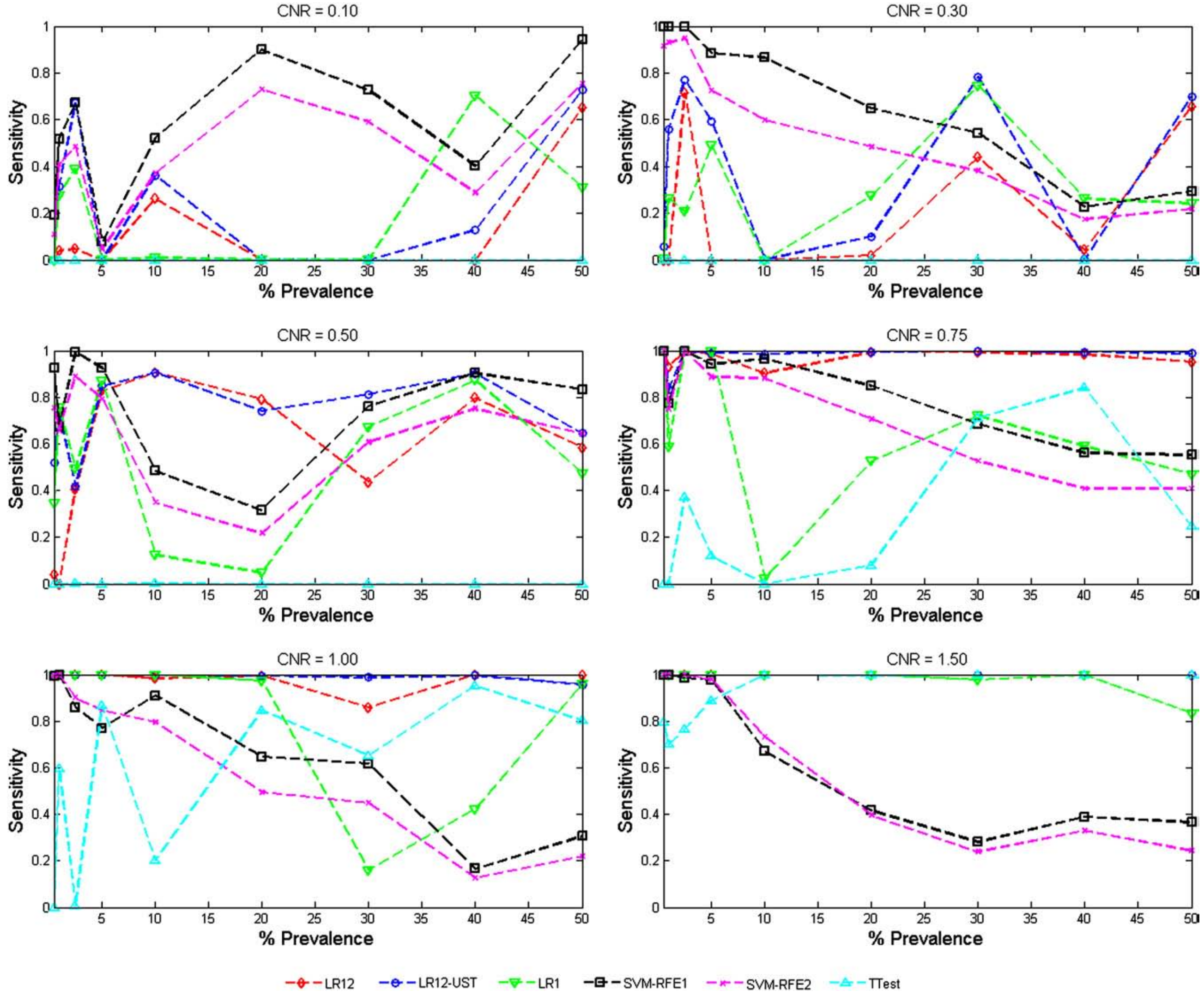

Fig. 3. Sensitivity of feature selection obtained using LR12, LR12-UST, LR1, SVM-RFE1, SVM-RFE2 and univariate *T*-test. LR12 has better sensitivity compared to other methods for most CNRs (in particular for high CNRs) and feature prevalence rates. The sensitivity of univariate *T*-test (at *p*-value of 0.05, FDR corrected) is poor for CNRs below 0.75 compared to other methods.

cortex bilaterally. Finally, discriminating voxels were also found in anterior and posterior cingulate and paracingulate cortex in the left-hemisphere as well as the cerebellum and brainstem. The cross validation accuracies obtained by LR12 and LR12-UST were 58.67% and 58.33% respectively in classifying music versus speech.

LR1 method

Brain regions that LR1 discriminated were extremely focal (Fig. 5C). This method revealed an extremely small collection of voxels in the left-hemisphere posterior middle temporal gyrus, inferior lateral occipital cortex, and cerebellum. Discriminated voxels in the left-hemisphere were sparse, where fewer than 5 voxels were selected in each of these left-hemisphere brain regions; LR1 did not identify any voxels in the right-hemisphere. This method revealed substantially fewer voxels than any of the other classification methods. The cross validation accuracy in classifying music versus speech by this method was 51.66%.

SVM-RFE method

The SVM-RFE1 (Fig. 5D) and SVM-RFE2 (Fig. 5E) methods were considerably less selective compared to the other methods. Not only did SVM-RFE1 and SVM-RFE2 identify all of the cortical and subcortical structures revealed using both LR12 and LR12-UST methods, they also identified a large number of additional voxels throughout the brain. The additional structures identified by SVM-RFE1 and SVM-RFE2 covered a large extent of the cortex, including many voxels in white matter areas of the brain. Compared to L1, LR12 and LR12-UST methods, the SVM-RFE methods were far less specific. The cross validation accuracy in classifying music versus speech by these methods was 54%. Note that CVAs obtained by SVM-RFE1 and SVM-RFE2 were exactly the same. They differ only with respect to the discriminative map computations.

LR1, LR12, LR12-UST and SVM-RFE methods using a functional mask

In addition to performing whole-brain analyses, we also performed the exact same analyses using a functional mask as a means of excluding deactivated voxels and including only those voxels which showed increased signal levels during music and/or speech stimuli (Supplemental Fig. S1). Similar to results from the whole-brain analysis, results varied considerably among the classification methods, with LR1 showing a relatively sparse collection of voxels, LR12 and LR12-UST methods showing intermediate specificity, and SVM-RFE1 and SVM-RFE2 showing less specificity compared to the other methods. Between SVM-RFE1 and SVM-RFE2 methods, SVM-RFE2 was more specific, while SVM-RFE1 showed nearly every voxel within the masked brain regions as discriminating voxels. Furthermore, the

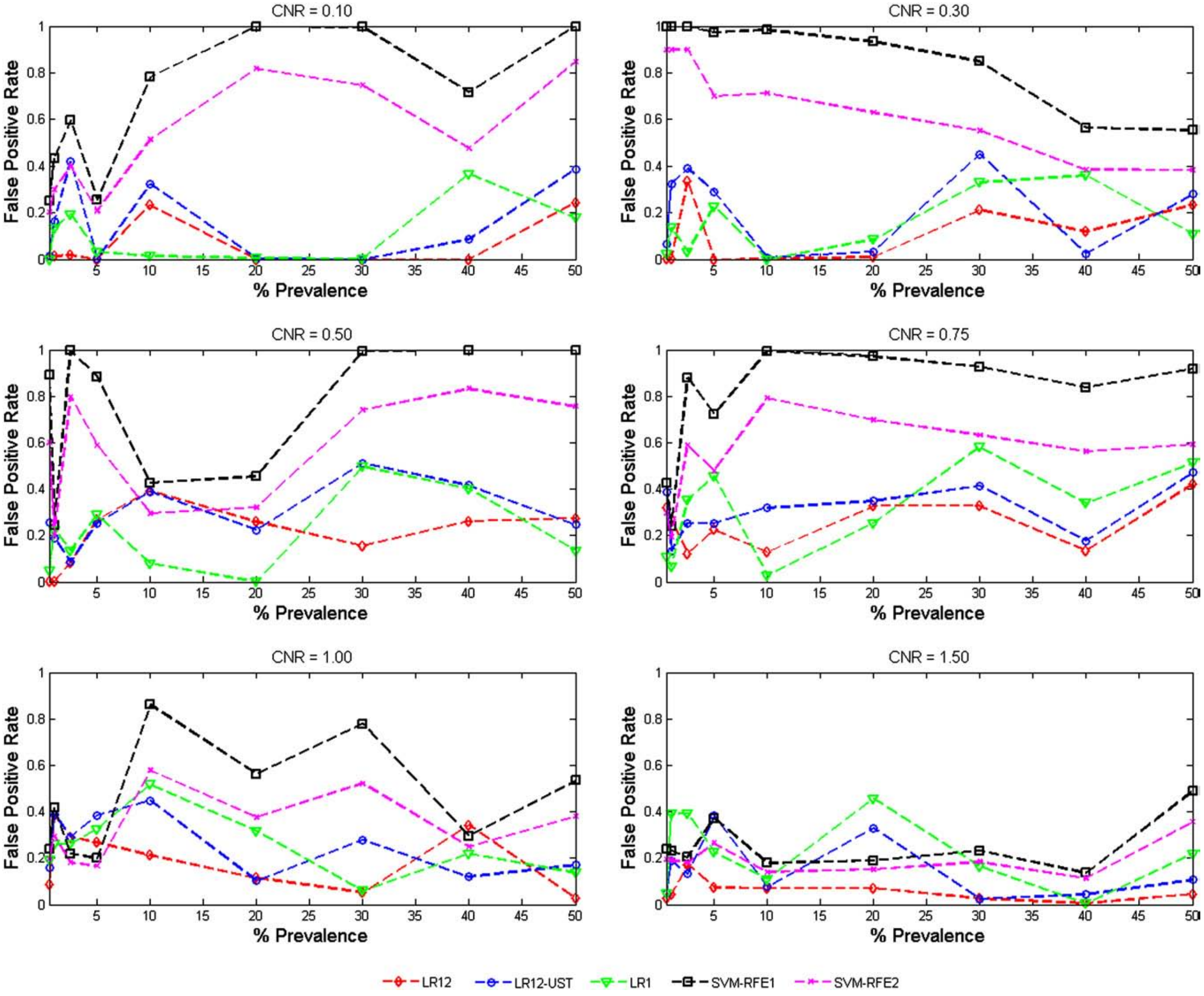

Fig. 4. False positive rates in feature selection obtained using LR12, LR12-UST, SVM-RFE1 and SVM-RFE2. False positive rates of LR12 are lower compared to other methods for most CNRs and prevalence rates.

LR12-UST method again showed a slightly larger extent of voxels compared to LR12. Both LR12 and LR12-UST methods indentified voxels in bilateral superior and middle temporal cortex, medial temporal lobe structures, frontal orbital cortex (BA 47) and frontal pole, and the cerebellum and brainstem. The cross validation accuracies provided by LR12, LR12-UST, LR1 and SVM-RFE were respectively 67.33%, 62.6%, 70.67% and 70% (again, CVAs obtained by SVM-RFE1 and SVM-RFE2 were exactly the same).

Discussion

We developed a novel whole-brain classification algorithm based on logistic regression for analysis of functional imaging data. Our LR12 method incorporates L1 and L2 norm regularization to achieve optimal feature selection in the presence of highly correlated features. This method provides three key improvements over existing methods: first, LR12 method can be scaled to whole-brain analysis; second, the method provides a data-driven mechanism to eliminate voxels which do not discriminate between two classes, while retaining voxels which can distinguish between the two classes of stimuli; and third, LR12 does not depend on any preset parameters. Critically, comparison of our classification algorithm with LR12-UST, LR1 and SVM-RFE on simulated datasets revealed superior performance in terms of accuracy of feature selection at various CNRs and feature prevalence rates. On the experimental data, LR12 was more effective in selectively isolating a distributed fronto-temporal network that distinguished between brain regions known to be involved in speech and music processing.

Advantages of LR12 method for classification of fMRI data

We used the bound optimization strategy along with the component-wise update procedure employed in Krishnapuram et al. (2005) in order to achieve computationally feasible whole-brain classification. This approach could be applied to LR12-, LR12-UST- and LR1-based methods. In comparison, existing methods that use the IRWLS optimization on small ROI data (Yamashita et al., 2008) cannot be scaled for whole-brain analysis. The reason IRWLS cannot be scaled is that it requires computation and inversion of a Hessian matrix, whose size is the same as the number of voxels at each iteration. This is computationally intractable. Our simulations show, for the first time, that using bound optimization along with component-wise update procedure is highly suited for fMRI data classification.

Our LR12 method incorporates both L1 and L2 norm regularizations. This combination of L1 and L2 norm regularization helps in determining the spatially correlated regions in the brain which discriminate between conditions. The degrees of these regularizations (γ_1 and γ_2) that need to be used for achieving this purpose is

Table 1
Cross validation accuracy (CVA) (A) and accuracy (B) and sensitivity (C) and false rate positive (D) in voxel selection obtained using LR12, SVM-RFE1 and SVM-RFE2 with and without an initial reduction step at a prevalence rate of 0.5%. Note that CVAs obtained by SVM-RFE1 and SVM-RFE2 were exactly the same. They differ only with respect to the discriminative map computations.

(A) Cross validation accuracy (CVA)				
	No voxel reduction		Voxel reduction	
CNR	LR12	SVM-RFE1/2	LR12	SVM-RFE1/2
0.1	0.42	0.26	0.66	0.5
0.3	0.56	0.24	0.66	0.5
0.5	0.6	0.4	0.74	0.62
0.75	0.78	0.48	0.78	0.56
1.0	0.82	0.52	0.8	0.62
1.5	0.92	0.7	0.94	0.86

	No voxel reduction			Voxel reduction		
CNR	LR12	SVM-RFE1	SVM-RFE2	LR12	SVM-RFE1	SVM-RFE2
(B) Accuracy						
0.1	0.99	0.75	0.81	0.92	0.92	0.9
0.3	0.99	0.005	0.37	0.92	0.9	0.93
0.5	0.99	0.11	0.53	0.92	0.96	0.95
0.75	0.68	0.58	0.74	0.93	0.96	0.94
1.0	0.91	0.76	0.82	0.93	0.96	0.96
1.5	0.97	0.76	0.82	0.93	0.97	0.96
(C) Sensitivity						
0.1	0	0.19	0.11	0.03	0.02	0.025
0.3	0	1.0	0.915	0.07	0.075	0.055
0.5	0.04	0.93	0.755	0.4	0.17	0.245
0.75	1.0	1.0	1.0	0.94	0.44	0.825
1	0.99	0.99	0.99	0.96	0.54	0.755
1.5	1.0	1.0	1.0	1.0	0.8	0.99
(D) False positive rate (FPR)						
0.1	5.03E−05	0.25	0.2	0.08	0.07	0.1
0.3	0.004	0.99	0.9	0.08	0.1	0.07
0.5	0.001	0.89	0.6	0.07	0.04	0.05
0.75	0.32	0.43	0.3	0.07	0.03	0.06
1.0	0.08	0.24	0.2	0.07	0.03	0.04
1.5	0.03	0.24	0.2	0.068	0.028	0.03

estimated directly from the data using a combination of grid search and cross validation procedure. Therefore, unlike the other methods, this approach does not require any arbitrary preset parameters for feature selection.

Another advantage of the LR12 algorithm is that it allows the user to select useful priors during whole-brain analysis. For example, we can incorporate spatial priors to account for neighborhood information and correlated activity around each voxel. By introducing such priors, we can avoid isolated features and noise which are frequently encountered in approaches that use the search-light algorithm or even the general linear model. Such spatial constraints can easily be incorporated in our framework by modifying the cost function in Eq. (10).

Comparison with LR1 and LR12-UST

Performance comparison

LR1 and LR12-UST are special cases of LR12. In LR1 $\gamma_2 = 0$ and in LR12-UST γ_2 is set to 10^4 while in LR12, both the parameters (γ_1 and γ_2) are optimized. LR12 resulted in higher overall accuracy in feature selection at most of the prevalence rates, and for CNRs of 0.5 and above. For low CNRs (0.1 and 0.3), all three sparse logistic regression methods and SVM-RFE resulted in CVAs at or below chance level (0.5). In the case of LR1, the sensitivity of feature selection is not consistent, as shown in Fig. 3. On the other hand, LR12-UST resulted in high sensitivity in voxel selection (Fig. 3) but false positive rates were also higher (Fig. 4). The performance of these methods can be attributed to inclusion or exclusion of L2-norm penalty. When the discriminating voxels are spatially correlated, LR1, which did not include L2-norm, selected only a subset of these voxels, resulting in decreased sensitivity. LR12-UST, which uses a fixed L2-norm regularization, resulted in higher sensitivity as well as higher false positive rates since the regularization parameter (γ_2) was not optimized in this case. On the other hand, LR12 resulted in higher accuracy in selecting relevant features compared to LR1 and LR12-UST because it optimizes both L1 and L2 norm regularization parameters.

Our findings are consistent with evidence from previous linear regression literature (Zou and Hastie, 2005), LR1 yielded sparse solutions with variable sensitivity. This can be explained by the fact that L1 norm regularization facilitates sparse solutions and serves as a powerful method for feature selection when features are uncorrelated. However, for datasets in which features are correlated, such as fMRI data, methods based on L1 norm regularization select only a subset of correlated features. This phenomenon was first observed in Lasso (least absolute shrinkage and selection operator) (Tibshirani, 1996), which is an L1 regularized method for linear regression methods. Zou and Hastie (2005) developed a method called Elastic-net, an extension of Lasso that introduces a combination of L1 and L2 norm regularization. It was shown that the Elastic-net method is effective in selecting an entire group of relevant and highly correlated features and that introducing L2 norm regularization is crucial for the selection of relevant features. Carroll et al. (2009) used Elastic-net for

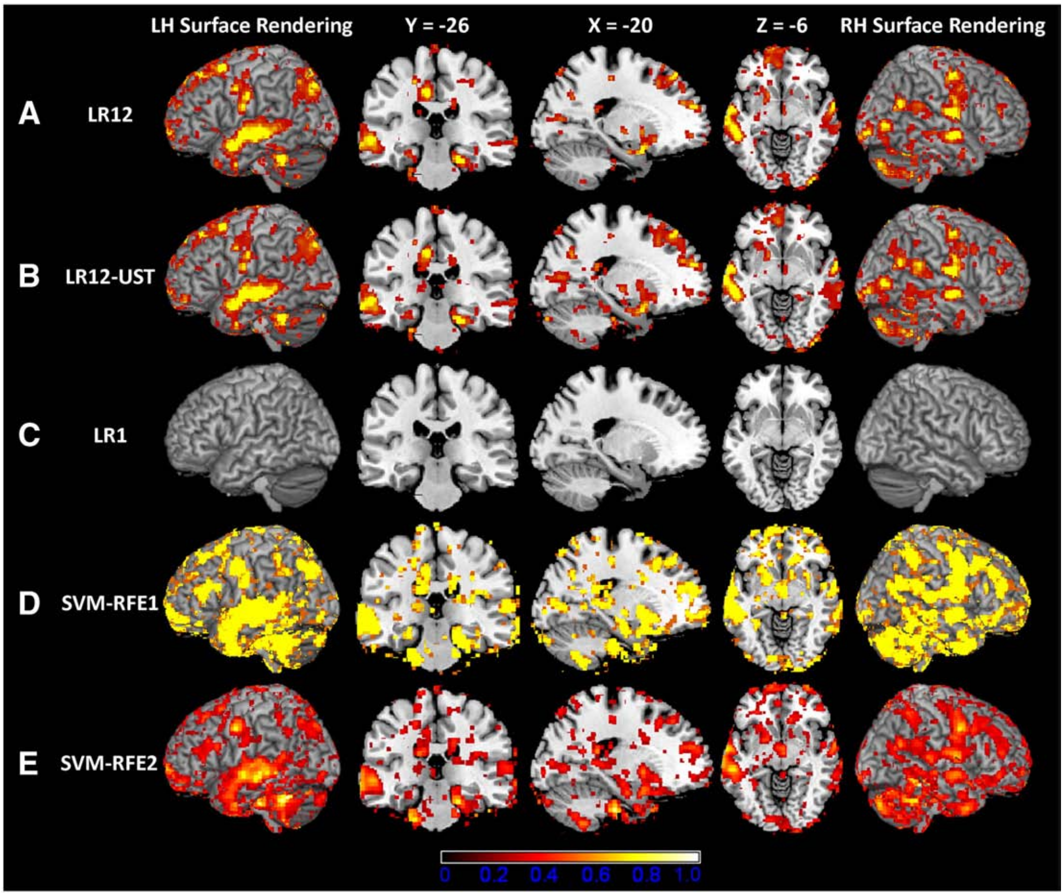

Fig. 5. Brain areas that discriminated between speech and music stimuli using LR12, LR12-UST, LR1, SVM-RFE1 and SVM-RFE2 methods (rows A–E). Surface renderings (left and rightmost columns) and sections are shown for each method. Note the increasing spatial extent of brain voxels that discriminated across conditions. SVM-RFE1 was highly non-selective in the sense that many voxels were chosen by the classifier.

fMRI data analysis. Our method extends Elastic-net, which was designed for regression analysis, to classification problems.

Comparison with SVM-RFE

Performance comparison

The cross validation accuracies obtained by SVM-RFE and sparse logistic regression methods are comparable. For low CNRs of 0.1 and 0.3, all the classification methods resulted in CVAs at or below chance levels. For CNRs of 0.5 and above, both LR12 and SVM-RFE resulted in CVAs above chance level (Fig. 1). The poor CVAs achieved by the classification methods at low CNRs can be attributed to the small differences between discriminating regions across the conditions. Under such conditions, the data is difficult to classify because of which all the methods resulted in below chance level CVAs. At CNRs of 0.5 and above, the discriminability of the spatial patterns between the classes improved thereby facilitating the classification of the data better.

In terms of accuracy, sensitivity and false positive rates in feature selection, LR12 performed better than SVM-RFE methods as shown in Figs. 2–4. The sensitivity of feature selection by SVM-RFE1 and SVM-RFE2 (Fig. 3) is greater than that of LR12 at low CNRs (0.1 and 0.3) but this is accompanied by more false positives (Fig. 4). As a result, the accuracy of feature selection is better in LR12 compared to SVM-RFE1 and SVM-RFE2 (Fig. 2). At CNRs of 0.5 and above, the overall accuracy of LR12 feature selection is higher than that of SVM-RFE1 and SVM-RFE2. The sensitivity of SVM-RFE1 and SVM-RFE2 decreases with an increase in prevalence rates as shown in Fig. 3, particularly at high CNRs; at the same time, false positive rates are greater in SVM-RFE1 and SVM-RFE2 compared to LR12 (Fig. 4).

Among SVM-RFE methods, SVM-RFE1 resulted in better sensitivity than SVM-RFE2 but the false positive rates obtained by SVM-RFE1 are higher than that of SVM-RFE2. This can be attributed to the way the final discriminative weights are computed. In SVM-RFE1 method, the final discriminative weights were computed as the average of discriminative weights obtained in each fold. In SVM-RFE2 method, the final discriminative weights were computed using the entire dataset by applying RFE at level at which CVA is maximum. The discriminating maps obtained at each fold may have false positives occurring at different locations. Therefore, averaging across folds inflates the false positive rate and therefore the results obtained by this approach are difficult to interpret. We also observed this fact when we applied this procedure of averaging of discriminative weights across folds on our sparse logistic regression methods (data not shown). Moreover, it is a very common practice in the machine learning literature, to obtain the discriminative weights on the entire dataset after having estimated the unknown parameters with a cross validation procedure (Hastie et al., 2001). Accordingly, the final discriminative weights in our method were computed using the entire dataset.

Critically, however, false positive rates in SVM-RFE1 and SVM-RFE2 are higher compared to LR12. This can be attributed to L1-norm regularization used in LR12 which drives the small magnitude weights to exactly zero. However, the SVM-RFE method drives the weights corresponding to non-discriminative voxels to small values but not exactly to zero. Therefore, additional thresholding of weights is required to prune out these false positives. In general, it is not

Please cite this article as: Ryali, S., et al., Sparse logistic regression for whole-brain classification of fMRI data, NeuroImage (2010), doi:10.1016/j.neuroimage.2010.02.040

straightforward to choose optimal thresholds without impairing the performance of the classifier.

Issue of free parameters in SVM-RFE

In our simulations with SVM-RFE, we removed 10% of the smallest weights at each recursive step. This is an arbitrary threshold but one that is necessary for any implementation of SVM-RFE (De Martino et al., 2008). The choice of this threshold may influence the performance of SVM-RFE and similar methods. Our analysis suggests that this threshold may not be optimal for many CNRs and prevalence rates. For example, the sensitivity of SVM-RFE1 and SVM-RFE2 decreased with the increase in prevalence rate even at the higher CNRs (1.0 and 1.5). This performance may be improved if this threshold is chosen appropriately for each prevalence rate. For these reasons, LR12 methods developed here may be preferable to methods with free parameters.

Initial voxel reduction

Previous studies have suggested that the performance of SVM-RFE improves with initial feature selection (De Martino et al., 2008). This is typically implemented by retaining voxels having high-level activations. Specifically, top N' ($=10\times$discriminative voxels) voxels rank ordered according to scoring function are retained in our analysis. We examined how the performance of LR12 and SVM-RFE was affected by feature selection. Here we chose to describe the comparative results at 0.5% prevalence rate because at higher prevalence rates (>0.5%) the number of voxels retained post voxel reduction is comparatively high; thereby making the voxel reduction less effective (De Martino et al., 2008). We found that classification accuracies (CVAs) and accuracy in feature selection improved with voxel reduction in both LR12 and SVM-RFE, as shown in Table 1. This improvement maybe due to the fact that the number of discriminative voxels, compared to the non-discriminative voxels was low (prevalence rate = 0.5%) and the voxel reduction step removes a large number of non-discriminative voxels. Notably, the SVM-RFE accuracy values showed significant improvement (Table 1B). However, the sensitivity in voxel selection achieved by LR12 after voxel reduction is better than that of SVM-RFE1 and SVM-RFE2 after voxel reduction (Table 1C) but at marginally higher false positives (Table 1D) for CNRs above 0.3. Surprisingly, the sensitivity of both SVM-RFE1 and SVM-RFE2 decreased with voxel reduction (Table 1C), although the false positive rates reduced with this step (Table 1D). Among the SVM-RFE methods, the decrease in sensitivity of SVM-RFE1 is greater than that of SVM-RFE2. This result is puzzling because one would expect SVM-RFE1 to achieve higher sensitivity because of the way the discriminating weights were computed. The reasons for this behavior of SVM-RFE1 with initial voxel reduction need to be investigated further. In contrast, the sensitivity of LR12 improved marginally for CNRs of 0.1, 0.3 and 0.5 and remained almost the same for CNRs of 0.75, 1 and 1.5 (Table 1C) with the initial voxel reduction. Therefore, these results suggest that the SVM-RFE approach, unlike LR12, is sensitive to the selection of initial voxels. Another critical limitation of this approach is that the number of voxels to be used in the classification must be chosen on the basis of an arbitrary threshold. However, the main objective of our study was to develop a fully multivariate feature selection method without the need for ad hoc procedures for feature selection. Moreover, in actual fMRI data, the number of voxels that can be discarded in such an initial voxel reduction step is clearly not known *a priori*. Furthermore, the discriminating features do not necessarily have to be the most highly strongly activated voxels.

Performance on experimental fMRI data

We applied LR12, LR12-UST, LR1 and SVM-RFE (SVM-RFE1 and SVM-RFE2) methods to an fMRI data involving well-matched speech and music stimuli. We hypothesized that music and speech stimuli would be distinguished by discrete but distributed structures largely confined to the temporal and frontal lobes which have previously been implicated in speech and music processing (Formisano et al., 2008; Friederici et al., 2003; Koelsch et al., 2002; Levitin and Menon, 2003; Tervaniemi et al., 2006). LR12 revealed distributed clusters in temporal and frontal lobe regions previously implicated in speech and music processing; LR12-UST results were nearly identical to the LR12 results, with the addition of a small number of voxels extending beyond those identified by LR12; LR1 showed a sparse pattern with an extremely small number of discriminatory voxels; both SVM-RFE methods exhibited very little specificity, and revealed a diffuse network of cortical and subcortical structures underlying speech and music acoustics.

While the ground truth in this data set is not known, these classification results are consistent with findings from our simulations. Results on both datasets demonstrate a continuum of anatomical specificity across the four classification methods with LR1 being the most anatomically specific and SVM-RFE methods being the most anatomically diffuse. Furthermore, results from the LR12 and LR12-UST methods are consistent with our knowledge of the auditory system and differential processing of speech and music stimuli as they identified a number of key auditory structures thought to be sensitive to both acoustical differences in the posterior temporal cortex (Formisano et al., 2008; Tervaniemi et al., 2006), as well as areas within the anterior temporal (Humphries et al., 2005; Rogalsky and Hickok, 2008) and prefrontal (Levitin and Menon, 2003; Tervaniemi et al., 2006) cortex thought to be sensitive to phrase- and sentence-level processing of music and speech stimuli. Our data suggest that methods based on LR12 and LR12-UST achieve a balance between sparse and diffuse discriminatory classification of auditory stimuli. The LR12 algorithms developed here are also highly computationally efficient compared to search-light algorithms (Haynes et al., 2007; Kriegeskorte et al., 2006) that can take several days to classify whole-brain data on a standard lab computer: analysis using the LR12 algorithm typically takes only about 3–4 h for a sample size of 20 subjects. The LR12-UST algorithm is even faster and it typically takes less than an hour to classify whole-brain data.

Conclusions

We developed a new method for whole-brain classification based on a combination of L1 and L2 norm regularization. Our method provides a completely data-driven and computationally efficient approach for both accurate feature selection and classification of whole-brain fMRI data. Critically, it does not require user-specified thresholds for feature selection as in recursive feature elimination method. In the case of fMRI data, where voxels are spatially correlated, the combination of L1 and L2 norm regularization provides a reliable feature selection. The identification of correlated features that discriminate between the experimental manipulations of interest is very important for the interpretability of the fMRI classification results. More importantly, extensive simulations indicated that methods based on LR12 had significantly higher accuracy in feature selection than other methods for a wide range of CNRs and feature prevalence rates. On experimental fMRI data, LR12 was more effective in selectively isolating a distributed fronto-temporal network that distinguished between brain regions known to be involved in speech and music processing.

Acknowledgments

We thank Dr. Lucina Uddin, Dr. Elena Rykhlevskaia and Rohan Dixit for reviewing the manuscript and for the insightful comments. We also thank the reviewers for their comments and suggestions, which resulted in an improvement of the manuscript. This research was supported by the National Institutes of Health (R01 HD047520, R01 HD045914, and NS058899), the National Science Foundation (BCS/DRL 0449927) and the Lucas Foundation.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2010.02.040.

References

Abrams, D.A., Bhatara, A.K., Ryali, S., Balaban, E., Levitin, D.J., Menon, V., submitted for publication. Music and speech structure engage shared brain resources but elicit different activity patterns.

Bi, J., Bennett, K., Embrechts, M., Breneman, C., Song, M., 2003. Dimensionality reduction via sparse support vector machines. J. Mach. Learn. Res. 1229–1243.

Carroll, M.K., Cecchi, G.A., Rish, I., Garg, R., Rao, A.R., 2009. Prediction and interpretation of distributed neural activity with sparse models. Neuroimage 44, 112–122.

Cox, D.D., Savoy, R.L., 2003. Functional magnetic resonance imaging (fMRI) “brain reading”: detecting and classifying distributed patterns of fMRI activity in human visual cortex. Neuroimage 19, 261–270.

De Martino, F., Valente, G., Staeren, N., Ashburner, J., Goebel, R., Formisano, E., 2008. Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. Neuroimage 43, 44–58.

Formisano, E., De Martino, F., Bonte, M., Goebel, R., 2008. “Who” is saying “what”? Brain-based decoding of human voice and speech. Science 322, 970–973.

Friederici, A.D., Ruschemeyer, S.A., Hahne, A., Fiebach, C.J., 2003. The role of left inferior frontal and superior temporal cortex in sentence comprehension: localizing syntactic and semantic processes. Cereb. Cortex 13, 170–177.

Friston, K.J., Williams, S., Howard, R., Frackowiak, R.S., Turner, R., 1996. Movement-related effects in fMRI time-series. Magn. Reson. Med. 25, 346–355.

Glover, G.H., Lai, S., 1998. Self-navigated spiral fMRI: interleaved versus single-shot. Magn. Reson. Med. 39, 361–368.

Grosenick, L., Greer, S., Knutson, B., 2008. Interpretable classifiers for FMRI improve prediction of purchases. IEEE Trans. Neural Syst. Rehabil. Eng. 16, 539–548.

Guyon, I.E.A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182.

Guyon, I.W.J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Mach. Learn. 46, 389–422.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning: Data mining, Inference and Prediction. Springer.

Haynes, J.D., Rees, G., 2005. Predicting the orientation of invisible stimuli from activity in human primary visual cortex. Nat. Neurosci. 8, 686–691.

Haynes, J.D., Sakai, K., Rees, G., Gilbert, S., Frith, C., Passingham, R.E., 2007. Reading hidden intentions in the human brain. Curr. Biol. 17, 323–328.

Humphries, C., Love, T., Swinney, D., Hickok, G., 2005. Response of anterior temporal cortex to syntactic and prosodic manipulations during sentence processing. Hum. Brain Mapp. 26, 128–138.

Kim, S.H., Adalsteinsson, E., Glover, G.H., Spielman, S., 2000. SVD regularization algorithm for improved high-order shimming. Proceedings of the 8th Annual Meeting of ISMRM, Denver.

Koelsch, S., Gunter, T.C., von Cramon, D.Y., Zysset, S., Lohmann, G., Friederici, A.D., 2002. Bach speaks: a cortical “language-network” serves the processing of music. NeuroImage 17, 956–966.

Kohavi, R.J.G., 1997. Wrappers for feature selection. Artif. Intell. 97, 273–324.

Kriegeskorte, N., Goebel, R., Bandettini, P., 2006. Information-based functional brain mapping. Proc. Natl. Acad. Sci. U. S. A. 103, 3863–3868.

Krishnapuram, B., Carin, L., Figueiredo, M.A., Hartemink, A.J., 2005. Sparse multinomial logistic regression: fast algorithms and generalization bounds. IEEE Trans. Pattern Anal. Mach. Intell. 27, 957–968.

Levitin, D.J., Menon, V., 2003. Musical structure is processed in “language” areas of the brain: a possible role for Brodmann Area 47 in temporal coherence. NeuroImage 20, 2142–2152.

Mitchell, T.M.H.R., Niculescu, R.S., Pereira, F., Wang, X., 2004. Learning to decode cognitive states from brain images. Mach. Learn. 57, 145–175.

Mourao-Miranda, J., Bokde, A.L., Born, C., Hampel, H., Stetter, M., 2005. Classifying brain states and determining the discriminating activation patterns: Support Vector Machine on functional MRI data. Neuroimage 28, 980–995.

Mourao-Miranda, J., Reynaud, E., McGlone, F., Calvert, G., Brammer, M., 2006. The impact of temporal compression and space selection on SVM analysis of single-subject and multi-subject fMRI data. Neuroimage 33, 1055–1065.

Pereira, F., Mitchell, T., Botvinick, M., 2009. Machine learning classifiers and fMRI: a tutorial overview. Neuroimage 45, S199–209.

Perkins, S., Lacker, K., Theiler, J., 2003. Grafting: fast incremental feature selection by gradient descent in function space. J. Mach. Learn. Res. 1333–1356.

Phillips, C., Rugg, M.D., Fristont, K.J., 2002. Systematic regularization of linear inverse solutions of the EEG source localization problem. Neuroimage 17, 287–301.

Rogalsky, C., Hickok, G., 2008. Selective attention to semantic and syntactic features modulates sentence processing networks in anterior temporal cortex. Cereb. Cortex 19 (4), 786–796.

Tervaniemi, M., Szameitat, A.J., Kruck, S., Schroger, E., Alter, K., De Baene, W., Friederici, A.D., 2006. From air oscillations to music and speech: functional magnetic resonance imaging evidence for fine-tuned neural networks in audition. J. Neurosci. 26, 8647–8652.

Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. B 58, 267–288.

Tipping, M., 2001. Sparse Bayesian learning and relevant vector machine. J. Mach. Learn. Res. 1, 211–244.

van Gerven, M., Hesse, C., Jensen, O., Heskes, T., 2009. Interpreting single trial data using groupwise regularisation. Neuroimage 46 (3), 665–676.

Various, 1991. On great speeches of the 20th century [CD]. Los Angeles: Rhino Records.

Wang, Z., 2009. A hybrid SVM-GLM approach for fMRI data analysis. Neuroimage 46 (3), 608–615.

Weston, J., Elisseff, A., Schoelkopf, B., Tipping, M., 2003. Use of the zero norm with linear models and kernel methods. J. Mach. Learn. Res. 3, 1439–1461.

Yamashita, O., Sato, M.A., Yoshioka, T., Tong, F., Kamitani, Y., 2008. Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns. Neuroimage 42, 1414–1429.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. B Stat. Methedol. 67, 301–320.