# Voxel Selection Framework in Multi-Voxel Pattern Analysis of fMRI Data for Prediction of Neural Response to Visual Stimuli

Chun-An Chou, Kittipat Kampa, Sonya H. Mehta, Rosalia F. Tungaraza, W. Art Chaovalitwongse*, *Senior Member, IEEE*, and Thomas J. Grabowski

***Abstract*—Multi-voxel pattern analysis (MVPA) of functional magnetic resonance imaging (fMRI) data is an emerging approach for probing the neural correlates of cognition. MVPA allows cognitive states to be modeled as distributed patterns of neural activity and classified according to stimulus conditions. In practice, building a robust, generalizable classification model can be challenging because the number of voxels (features) far exceeds the number of stimulus instances/data observations. To avoid model overfitting, there is a need to select informative voxels before building a classification model. In this paper, we propose a robust feature (voxel) selection framework using mutual information (MI) and partial least square regression (PLS) to establish an informativeness index for prioritizing selection of voxels based on the degree of their association to the experimental conditions. We evaluated the robustness of our proposed framework by assessing performance of standard classification algorithms, when combined with our feature selection approach, in a publicly-available fMRI dataset of object-level representation widely used to benchmark MVPA performance (Haxby, 2001). The computational results suggest that our feature selection framework based on MI and PLS drastically improves the classification accuracy relative to those previously reported in the literature. Our results also suggest that highly informative voxels may provide meaningful insight into the functional-anatomic relationship of brain activity and stimulus conditions.**

***Index Terms*—Classification, feature selection, functional magnetic resonance imaging (fMRI), information theory, multi-voxel pattern analysis (MVPA), partial least square (PLS), pattern recognition.**

Manuscript received August 12, 2013; accepted January 02, 2014. Date of publication January 09, 2014; date of current version March 31, 2014. This work was supported in part by the Research Foundation at Binghamton University under Grant (I920247) and in part by the Integrated Brain Imaging Center at University of Washington Medical Centerunder the RC4 Grant (NS073008, PI: Grabowski).. *Asterisk indicates corresponding author.*

C.-A. Chou is with the Department of Systems Science and Industrial Engineering, Binghamton University, the State University of New York, Vestal, NY 13902 USA (e-mail: cachou@binghamton.edu).

K. Kampa is with the Department of Industrial and Systems Engineering and the Integrated Brain Imaging Center, University of Washington, Seattle, WA 98195 USA.

S. H. Mehta is with the Departments of Radiology and Psychology, and the Integrated Brain Imaging Center, University of Washington, Seattle, WA 98195 USA.

R. F. Tungaraza is with the Department of Mathematics and Computer Science, Kalamazoo College, Kalamazoo, MI 49024 USA.

*W. A. Chaovalitwongse is with the Departments of Industrial and Systems Engineering and Radiology, and the Integrated Brain Imaging Center, University of Washington, Seattle, WA 98195 USA (e-mail: artchao@uw.edu).

T. J. Grabowski is with the Departments of Radiology and Neurology, and the Integrated Brain Imaging Center at University of Washington, Seattle, WA 98195 USA.

Digital Object Identifier 10.1109/TMI.2014.2298856

## I. Introduction

MULTI-VOXEL pattern analysis (MVPA) is an emerging approach for studying the relationship between cognition and brain activity measured by functional magnetic resonance imaging (fMRI). fMRI measures blood oxygenation level-dependent (BOLD) signal that arises from the interaction between blood flow (and blood oxygenation) and changes in neural activity [2]. In task-based studies, changes in neural activity are assumed to be experimentally induced, with fMRI time series data reflecting the response to stimuli at certain locations (i.e., voxels) across the brain. The hemodynamic response coupled to neural activity is slow (on the order of seconds) and systematic, allowing the observed BOLD signal change associated with a given stimulus (or stimulus block) to be adequately modeled by a canonical response function and summarized by a single value.

A common objective of MVPA is to build a pattern classification model of BOLD responses from the whole brain or restricted to *a priori* brain regions of interest (ROIs) to predict cognitive representations associated with experimental conditions (e.g., the responses to different categories of visual stimuli). One of the pioneering studies of MVPA was performed by Haxby *et al.* [1], where multi-voxel patterns in the ventral temporal (VT) cortex were investigated in response to different categories of visual stimuli. Although this MVPA study was originally cast as a straightforward classification problem, improving performance remains an open challenge because classification using fMRI data typically operates on a large set of voxels creating problems of overfitting and high computational complexity [3], [4]. Thus, it is important to reduce the dimensionality by selecting voxels with a high degree of association to the experimental conditions, called *informative voxels*, before building a classification model. Although classification algorithms like support vector machines (SVMs), neural networks (NNs), classification tree (CT) have a built-in feature weighting/selection and/or regularization, there is still a pressing need of more aggressive feature selection techniques that can discard noninformative voxels up front. There have been a number of feature selection methods developed in the neuroimaging applications [5]–[10], but almost all of them employ univariate feature selection strategies, whereas only a few studies focus on multivariate (multi-voxel) selection.

In this paper, we propose a computational framework for multi-voxel selection in MVPA to identify and prioritize voxels in the ROI (VT cortex) that are associated with object-level representations of concrete entities. To quantify such associations, we propose two metrics to score the voxels (called the "importance indexes" or "informativeness indexes"). The first metric is based on an information theoretic approach, called mutual information (MI), that will be used to evaluate the degree of association of each individual voxel to the experimental conditions in a univariate fashion. The second metric is based on a multivariate statistical measure, called partial least square (PLS) regression, that will be used to identify information latent in the multivariate relationships (patterns) of voxel activity for purposes of guiding selection of feature sets of informative voxels. In multi-class classification problems like the one used in this study, the class label (i.e., category of a stimulus) can be cast as a random variable that naturally lends itself to its dependency with BOLD responses of informative voxels. MI has been widely applied to feature ranking and selection problems [3] in many real-life applications such as fMRI analysis [11]–[13], image registration [14], gene expression [15], computer-aided diagnosis [16], etc. MI has advantages over traditional linear methods (e.g., Pearson's correlation) as it can capture a higher degree of dependency between variables through their joint and marginal probability distributions. Although MI has been previously used in neuroimaging [11]–[13], [17], to the best of our knowledge, the current study is among the first to apply MI as a criterion for feature selection in MVPA [18]. Unlike multivariate MI (between the class label and the joint feature set) [12], we propose using a univariate MI as a feature ranking measure to rank individual voxels independently. The latter approach can be seen as a more voxelwise specific, and less computationally expensive, version of the former. PLS regression is another approach that has been widely applied in feature selection and known to be effective in finding correlated variables in high dimensional data [19]–[24]. Although most frequently applied to industrial problems [25]–[29], it has been used in some neuroimaging studies [30]–[33]. PLS has been applied to task-related fMRI data mainly to identify changes and correlations in time-courses across all voxels (spatiotemporal analysis) without the assumption of the shape of hemodynamic response function (HRF) for the BOLD signal. In contrast to previous studies, this study focuses on identifying the correlation between voxel activity and presented stimuli [2], [34]. In other words, we propose to employ the PLS regression to identify the correlations between voxels and stimulus categories through the resulting regression coefficients. Based on MI and PLS measures, we demonstrate in this study that our feature set selection framework can improve the accuracy and interpretability of pattern analysis results.

The organization of the paper is as follows. In Section II, we present the background of MVPA of fMRI data, and review the concepts of MI and PLS applied in feature ranking and selection. In Section III, we describe the proposed feature selection framework to rank and select informative (important) voxels. In Section IV, we show the experimental results of our feature selection method on a benchmark dataset and compare them with classification results of other MVPA approaches in the literature. We also extend and apply our approach to additional multiple subjects. In Section V, we conclude this work with future directions of our research.

## II. Background

### A. Multi-Voxel Pattern Analysis of fMRI Data

Conventional analysis of fMRI data relies on univariate statistics (e.g., contrasting task conditions), in which the response at each brain location (operationally, each image voxel) is considered independently. However, recent studies have demonstrated that "mental representations" may in fact be embedded in a distributed neural population code captured in the activity pattern across multiple voxels [1], [4], [6]. MVPA research, adapted from machine learning and pattern recognition, has thus been applied to study the joint contribution of brain space activity (operationally, across multiple image voxels) to elucidate the neural basis of cognition.

MVPA approaches have most commonly been used to perform "cognitive state decoding" (i.e., classification of "cognitive representations" into discrete categories of stimulus conditions). This subset of MVPA entails several steps: 1) feature extraction, 2) feature selection, and 3) pattern classification, where *features* are commonly operationalized as *voxels*. Feature extraction is a procedure to characterize the temporally-evolving BOLD response to a stimulus at a voxel. The most common approach is to estimate the magnitude of the fMRI response for each stimulus event or block using an impulse or a box-car regressor convolved with a hemodynamic response function in the general linear model (GLM). By making some assumptions about the nature of the hemodynamic response, the approach reduces the observed temporally-evolving fMRI signal to a single value for each stimulus event or block. The rationale for this data reduction step is to improve characterization of the response so that extracted features reflect the underlying neural response versus noise. The approach is well-suited for block designs in which the majority of activity-related information is in the magnitude of the fMRI response. Nonetheless this GLM approach could result in the loss of feature information, and alternative methods for feature extraction are an important, but separate, area of research from feature selection. Feature selection is a procedure to identify and select the subset of voxels to use with the classification model. Scientifically, this step is extremely important as the selected features are interpreted as reflecting the anatomical regions supporting the cognitive processes being studied. Pattern classification is a procedure to train a classification algorithm to create a prediction/classification model that best separates the stimulus categories within the multidimensional space defined by the selected features (voxels). Linear classifiers are often preferred because of their scientific interpretability, with an above chance decoding accuracy indicating the presence of information at an "explicit" level (e.g., not requiring further nonlinear neural computation) [35].

Fig. 1 illustrates the feature extraction of the fMRI time series data in a region of interest (the VT cortex in this case). To characterize the temporally-evolving BOLD signal change in response to a stimulus, a general linear model (GLM) is applied, and coefficient parameters $\beta$ are estimated by fitting a

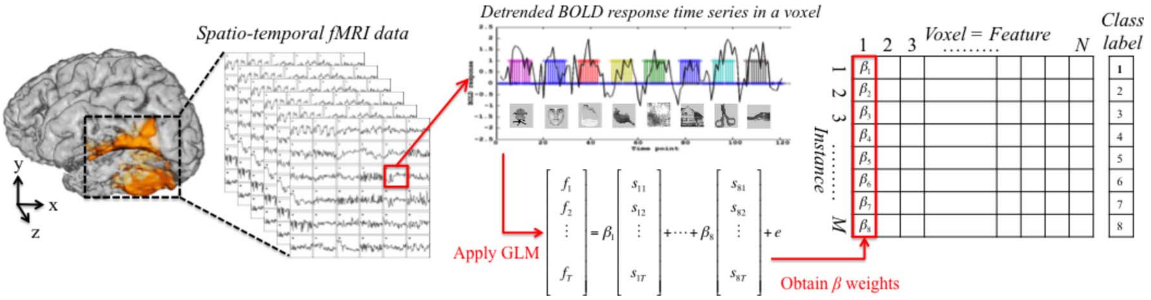

Spatio-temporal fMRI data
y
x
z
Detrended BOLD response time series in a voxel
Time point
Apply GLM
$f_1$, $f_2$, ⋮, $f_T$ = $\beta_1$ $s_{11}$, $s_{12}$, ⋮, $s_{1T}$ + ⋯ + $\beta_8$ $s_{81}$, $s_{82}$, ⋮, $s_{8T}$ + $e$
Obtain $\beta$ weights
Voxel = Feature
1 2 3 ........ N
Instance
1 2 3 ........ M
$\beta_1$ $\beta_2$ $\beta_3$ $\beta_4$ $\beta_5$ $\beta_6$ $\beta_7$ $\beta_8$
Class label
1 2 3 4 5 6 7 8

Fig. 1. Illustration of spatio-temporal fMRI data in the VT cortex with an example of a BOLD time series in response to eight different stimuli (images) at a single voxel. Collected fMRI data after preprocessing is transformed into an instance-by-feature data matrix. (Best viewed in color, available online)

GLM with different predictors for each stimulus block. In this study, the predictors (i.e., $s_{i1}, s_{i2}, \ldots, s_{iT}$ for stimulus condition $i = 1$ to 8, and BOLD responses at time 1 to $T$) were modeled with a boxcar convolved with a canonical HRF [2]. We used a double-Gamma function provided by SPM [34], with the default settings, as the HRF. The $\beta$ weights (parameters) are extracted for each run of the experiment, each generating a 3-D $\beta$ weight matrix for each voxel, which can be in turn transformed to a 2-D feature matrix. We denote this input feature matrix $X$, whose size is $M \times N$, where $M$ is the number of data instances (the total number of presented stimuli) and $N$ is the number of features (voxels). The element $x_{ij}$ of the data matrix $X$ represents the real-valued coefficient parameter $\beta$ of the $i$th data instance at the $j$th voxel. It is helpful to view $x_{ij}$ as the $i$th sample of the $j$th feature random variable $X_j$, the $j$th column of $X$. It is more convenient to treat $X_j$ as a random variable of the real-valued coefficient $\beta$ in relevant probabilistic measures. We denote class label $c_i \in \{1, \ldots, K\}$ (i.e., stimulus category), where $K$ is the total number of stimulus categories. For each data instance $i$, $c_i$ is known precisely according to the experiment design.

### B. *Mutual Information*

MI is a measure to quantify statistical dependence of two random variables [36]–[38]. In this study, MI is employed as a dependence measure between a feature ($\beta$ weight of individual voxel) and class label (stimulus condition), or between features. Consider MI between the class label random variable $C$ and the $j$th feature $X_j$, $MI(C; X_j)$ is calculated for all the features $j \in \{1, \ldots, N\}$, which is expressed by

$$MI(C; X_j) = \sum_{c=1}^{K} \int_{X_j} p(c, x_j) \log \left( \frac{p(c, x_j)}{p(c)p(x_j)} \right) dx_j \tag{1}$$

where $p(c, x_j)$ (short-handed for $p(C = c, X_j = x_j)$) is a joint distribution of both discrete and continuous variables $C$ and $X_j$, which can be calculated more conveniently using the chain rule $p(c, x_j) = p(c)p(x_j|c)$. The estimator $\hat{p}(x_j|c)$ of the conditional term $p(x_j|c)$ can be calculated by a kernel density estimation technique, called the *Parzen–Rosenblatt* window method

$$\hat{p}(x_j|c) = \left( \sum_{i=1}^{M} \delta_c(c_i) \mathcal{K} \left( \frac{x_j - x_{ij}}{w} \right) \right) \Bigg/ \left( w \sum_{i'=1}^{M} \delta_c(c_{i'}) \right) \tag{2}$$

where $\delta_x(y)$ is the Kronecker delta function, where $\delta_x(y) = 1$ when $x = y$ and 0 otherwise, $\mathcal{K}(\cdot)$ is the kernel, and $w$ is the bandwidth. In this particular case, the kernel function is chosen to be a Gaussian kernel, and thus the bandwidth becomes the standard deviation. $p(c)$ is derived from marginalizing out $X_j$, $p(c) = \int_{X_j} p(c)p(x_j|c)dx_j$, and $p(x_j)$ can be obtained from $p(x_j) = \sum_{c=1}^{K} p(c)p(x_j|c)$. A higher value of MI reflects more information at a voxel about the stimulus condition. In other words, a higher MI value implies a greater statistical dependency between the voxel and the stimulus condition.

### C. *Partial Least Square Regression*

The PLS regression can be applied to the matrix of our extracted $\beta$ weights as follows. Consider a data matrix $X$, whose size is $M \times N$, where $M$ is the number of data instances and $N$ is the number of features (voxels), and a matrix $Y$, whose size is $M \times K$, where $M$ is the number of data instances and $K$ is the number of dependent variables. Note that in this study $K = 1$ as the class label is only a 1-D vector (i.e., one of the eight categories). The PLS regression model is then expressed by $y_{ic} = b_{cj} x_{ij} + \epsilon_{ic}$, where $b_{cj}$ are the regression coefficients that take into account the correlations among all the features leading to a better prediction of the dependent variables and $\epsilon_{ic}$ are the residual errors. The objective of PLS is to construct a small number of independent, linear combinations of features. These new independent features, called PLS components, account for much of the variance present in the original features and in the dependent variables (i.e., class label). Typically, only three or four PLS components can be used to represent dozens or even hundreds of features. Independent PLS components that are linear combinations of features can be calculated by $t_{ia} = \sum_{j=1}^{N} w_{ja} x_{ij}$, where $w_{ja}$ represents the weight of feature $j$ in component $a$, which provides information about the way the variables combined themselves to generate $X$ and $Y$ [25]. Similarly, components are constructed for the dependent variables by $u_{ia} = \sum_{c=1}^{K} q_{ca} y_{ic}$, where $q_{ca}$ represents the weight of dependent variable $c$ in component $a$. The weight vectors $\mathbf{w}_a$ and $\mathbf{q}_a$ are selected to maximize the covariance of the PLS process and the components $\mathbf{t}_a$ and $\mathbf{u}_a$. Further, the weights are selected to yield orthogonal components; i.e., $T_a$'s are independent of one another and $U_a$'s are independent of one another [25], [39]. Loadings associated with $A$ independent components contain the regression coefficients of the columns of $X$ regressed on $\mathbf{t}_a$: $\hat{X}_a = T_a L_a$, where $L_a$ is defined as a loading vector. The greater the absolute value of the loading, the more effect the component has on the prediction of that process variable. These key parameters of weights and loadings can be calculated by means of the NIPALS algorithm [40]. Finally, the PLS regression coefficients are a function of the weights and loadings of the original features defined by

$$b_{cj} = \sum_{a \in A} q_{ca} w^{*}_{ja} \tag{3}$$

where $w^{*}_{ja} = w_{ja}(l_{ja} w_{ja})^{-1}$ is a modified weight emphasizing the effect of the features and leads to more stable predictions and feature selection procedures than $w_{ja}$. Further mathematical details of PLS can be referred to [41], [25], [42], [40].

## III. Voxel Selection Using MI and PLS

The voxel selection in MVPA for prediction/classification consists of the following steps: 1) extract the representative features (i.e., $\beta$ weights) from the original fMRI time series data by using GLM; 2) generate features' importance (also called informativeness) indexes with respect to the class information; 3) apply a voxel selection strategy proposed here to select the best voxel set; and 4) classify the testing instances using the selected voxel set. Voxels here are selected from the VT cortex ROI, which were provided as part of the public release of the data. The ROI were defined using combined anatomic and functional criteria to include only the voxels (within the anatomically-defined regions) exhibiting a significant response to object categories relative to scrambled images. Details are provided in the original article [1]. Because this work is focused on selecting informative voxels to enhance classification performance, we also refer the readers to the previous section and the literature [34], [2] for more detailed information about the data pre-processing and feature extraction of fMRI data from the ROI. In this section, we start our framework with the input feature matrix $X$ of representative HRF features (i.e., $\beta$ weights), each obtained from an individual voxel in the ROI. Note that the $\beta$ values of individual voxels are standardized with a mean equal to 0 and a standard deviation equal to 1.

### *Step 1. Generate MI-Based and PLS-Based Importance indexes*

We propose two importance indexes for voxels from MI and PLS, respectively, to quantify the importance of voxels for multi-class classification purpose. We calculate a MI value as an importance index (MI_based importance index) for a voxel with respect to the class label using (1), denoted by $MI_j$. The higher the value of MI is, the greater dependence between the $\beta$ weights of a voxel and the class label. With PLS regression, we consider the magnitude of PLS regression coefficient $b_{cj}$ for the use of an importance index (PLS_based importance index), which measures the intensity of the relationship between voxels and the class label [29]. The proposed importance index is the sum of the absolute values of $b_{cj}$ over the $K$ classes, expressed by

$$V_j = \frac{\sum_{c=1}^{K} |b_{cj}|}{\max_{j=1,\ldots,N} \sum_{c=1}^{K} |b_{cj}|} \tag{4}$$

where the PLS regression coefficient $b_{cj}$ is obtained by using (3). Both the MI_based and PLS_based values of individual voxels are normalized between 0 and 1 to have comparable values of importance across voxels.

**Algorithm 1** pseudocode for the maximum informativeness selection strategy (*maxI*)

```
for $\alpha \in [0, 1]$ do
    $S = \emptyset$.
    Select the top voxels $S$ from the original set $J$, whose importance indices ($MI_j$ or $V_j$) are greater than $\alpha$.
    Train the classification model $\Omega$ with the selected voxels $S$ and test for the validation dataset.
    Store the classification accuracy: $accur(\alpha)$ and $S(\alpha)$.
end for
Update the best calibration:
    $\alpha^* = arg\max_{\alpha}\{accur(\alpha)\}$.
Report the best classification performance:
    $accur\_best \leftarrow accur(\alpha^*)$ and $S^* \leftarrow S(\alpha^*)$.
```

### *Step 2. Voxel Selection Strategy*

To find important (i.e., "highly informative") voxels to be used in a classification model, we propose the maximum informativeness selection strategy, called *maxI*, that selects the set of voxels based on the MI and PLS informativeness measures. Because it is not intuitive to determine how many voxels should be selected, this strategy determines the best level of MI and PLS indexes, rather than the best number for voxels to be selected. The *maxI* strategy works as follows. We first define a threshold, $\alpha \in [0, 1]$, that controls the importance indexes of voxels to be selected in the classification model. All features (voxels) are sorted in a descending order, and subsequently only voxels whose MI_based or PLS_based importance indexes ($MI_j$ and $V_j$, respectively) exceed the threshold $\alpha$ are included in the classification model. To obtain an optimal value of $\alpha$, we first split a training dataset into a training subset and a validation subset in order to calibrate the best MI and PLS levels. A calibration procedure is iteratively carried out for a set of threshold values. The trained model with selected voxels is then applied to validation dataset to obtain the associated classification accuracies that are associated with individual $\alpha$ values. The best calibration $\alpha^*$ is determined according to the highest classification accuracy, and then the associated set of informative voxels will be used to construct a classification model that will be used to test unseen samples. The procedure of *maxI* is illustrated in Algorithm 1.

*Step 3. Prediction of Visual Stimuli With the Selected Voxels*

The last step of the procedure is to test and assess the classification performance. The best set of voxels $S^*$ from Step 2 is adopted and used in a classification model. In a testing dataset, the class (i.e., category) of each stimulus is known *a priori*. After the classification model is derived, the predicted class is compared to the actual class to obtain the classification accuracy. In this study we adopted three classification methods that are widely used in the MVPA literature.

*1) Gaussian Naive Bayes Classifier (GNB):* A naive Bayes classifier [43] is a probabilistic classifier based on Bayes' theorem with a strong assumption that each feature is independent, namely, "independent feature model."

*2) Logistic Regression Classifier (LR):* Logistic regression (LR) classifier is a linear discriminative classifier [44], whose underlying intuition is to separate the data instances into two groups by a hyperplane. LR works substantially well in the scenario where the feature-to-instance ratio is high, fitting well with the description of fMRI data in general. Although the LR's cost function is originally formulated for binary classification, the extension for multi-class classification can be done using a one-vs-all paradigm.

*3) Support Vector Machines (SVM):* Support vector machine (SVM) is arguably the most widely used algorithm to analyze fMRI data [7], [8], [45], [9], [10] because it performs well on the data with high feature-to-instance ratio. SVM is a linear discriminative classifier which not only separates data into two groups by a hyperplane, but also maximizes the margin between the two classes, and it is hence best described by a "maximum margin classifier." The notion of margin lends robustness and good generalization to SVM, making it one of the most popular classifiers. In this experiment, a multi-class SVM is implemented using a pair-wise strategy. For the sake of interpretability, a linear SVM is used throughout the experiment.

## IV. Experimental Results

### A. fMRI Data

In this study, we used the fMRI data of one subject from the study originally published by Haxby *et al.* [1]. The dataset, which is available on their research website (http://code.google.com/p/princeton-mvpa-toolbox/), has been a widely used to benchmark performance of MVPA techniques. This dataset consisted of 10 fMRI runs from a block-design experiment [1], with each run comprised of eight stimulus blocks. Each stimulus block displayed image exemplars from one of 8 different conceptual categories: 1) faces, 2) houses, 3) cats, 4) bottles, 5) scissors, 6) shoes, 7) chairs, and 8) "scrambled pictures." The order of the stimulus blocks were randomized across runs. One image of brain activity in the dataset (consisting of 64 × 64 × 40 voxels) was acquired every repetition time (TR) of 2.5 s, and there are a total of 9 TRs $(= 22.5/2.5)$ in each block. Thus there are a total of 720 data instances for the dataset (10 runs × 8 blocks × 9 TRs). In our study, we used the provided ROI in the VT cortex defined by functional anatomic criteria [1]. The ROI contains a total of 577 thresholded voxels. In addition, we extended our analysis to the whole brain space that covers a total of 43 193 voxels. The fMRI data were acquired on a GE 3T scanner.

The fMRI dataset underwent standard data preprocessing steps to improve spatial alignment and to attenuate noise of time series data, including motion correction and linear detrending. We also standardized (z-scored) the data by subtracting the mean and dividing by the standard deviation of the time series signal at each voxel. We refer interested readers to [2], [45] for more details. Subsequently, we resampled the data to generate new datasets, that respected the original experimental design, for testing purposes. To do so, we randomly shuffled the original data matrix ($\beta$ weights) by rows and randomly selected one from each of eight categories to form a new run. Each data instance kept its spatial correlations across all voxels. The final feature matrix that was extracted from the Haxby dataset contained a total of 80 data instances (10 runs × 8 blocks), each having 577 features (for all 577 voxels in the VT cortex) and 43 193 features (for all 43 193 voxels in the whole brain).

### B. Classification Power of Different Voxel Ranking Criteria

In this experiment, to demonstrate the classification power of our MI_based and PLS_based indexes, we compared the classification results using different criteria of feature ranking. Here we evaluated candidate voxel sets that were prioritized (ranked) by different criteria to see how well the classification performed based on those candidate sets. Because the optimal size of voxel sets for each criterion was not known, we employed a simple leave-one-run-out cross validation with varying sizes of voxel sets. Note that in these experiments, we used only training (nine folds) and testing (one fold) datasets without parameter calibrations. In each cross validation, we measured the classification performance (accuracy) when the ranked features (voxels) were added to the classifier $\Omega$ one by one until all the voxels were included. Classification accuracy is defined as the percentage of data instances whose categories were correctly predicted over the total testing data instances. The first $n$ voxels were input to the classifier $\Omega$, and the accuracy was calculated from the classification result. The experiment continued from $n = 1$ to $N$. In other words, each succeeding voxel set was constructed by adding the voxel with highest rank to the previous set, and its performance after each classification was measured. The classification/addition process was concluded when all voxels were included. The voxel ranking criteria that were evaluated are listed as follows:

- Ranking voxels in a descending order based on MI_based and PLS_based indexes: *MI_Descend* and *PLS_Descend*. In this approach, we started building our classification models from the most informative voxels and kept adding less informative voxels until all voxels had been added.
- Ranking voxels in an ascending order based on MI_based and PLS_based indexes: *MI_Ascend* and *PLS_Ascend*. This approach is used as a contrasting example to demonstrate the poor classification performance when we started building our classification models from the least informative voxels.
- Ranking voxels in a random order: *Shuffled*. This approach is used as a baseline. We randomly permuted the order of

TABLE I

CLASSIFICATION ACCURACIES OBTAINED FROM THREE CLASSIFIERS USING THE TOP $n$% OF VOXELS IN THE VT CORTEX RANKED BY *MI_DESCEND*, *PLS_DESCEND* AND *MI+PLS_DESCEND*

| | Percentage of voxels | 10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% | 90% | 100% |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | **No. of voxels** | 58 | 115 | 173 | 231 | 289 | 346 | 404 | 462 | 519 | 577 |
| **SVM** | **MI_Descend** | 0.86 ± 0.14 | 0.96 ± 0.08 | 0.95 ± 0.09 | 0.95 ± 0.09 | 0.94 ± 0.12 | 0.93 ± 0.11 | 0.89 ± 0.09 | 0.91 ± 0.08 | 0.91 ± 0.08 | 0.90 ± 0.10 |
| | **PLS_Descend** | 0.91 ± 0.08 | 0.95 ± 0.09 | 0.96 ± 0.06 | 0.91 ± 0.10 | 0.91 ± 0.08 | 0.90 ± 0.10 | 0.91 ± 0.08 | 0.90 ± 0.10 | 0.90 ± 0.10 | 0.90 ± 0.10 |
| | **MI+PLS_Descend** | 0.93 ± 0.11 | 0.96 ± 0.06 | 0.96 ± 0.06 | 0.95 ± 0.09 | 0.94 ± 0.11 | 0.89 ± 0.09 | 0.91 ± 0.08 | 0.91 ± 0.08 | 0.91 ± 0.08 | 0.90 ± 0.10 |
| **LR** | **MI_Descend** | 0.86 ± 0.11 | 0.94 ± 0.09 | 0.94 ± 0.07 | 0.96 ± 0.06 | 0.93 ± 0.06 | 0.93 ± 0.09 | 0.90 ± 0.11 | 0.93 ± 0.06 | 0.91 ± 0.08 | 0.88 ± 0.10 |
| | **PLS_Descend** | 0.94 ± 0.07 | 0.94 ± 0.07 | 0.93 ± 0.06 | 0.90 ± 0.08 | 0.90 ± 0.08 | 0.90 ± 0.08 | 0.91 ± 0.08 | 0.90 ± 0.10 | 0.94 ± 0.07 | 0.88 ± 0.10 |
| | **MI+PLS_Descend** | 0.94 ± 0.09 | 0.94 ± 0.07 | 0.93 ± 0.06 | 0.93 ± 0.09 | 0.91 ± 0.06 | 0.91 ± 0.06 | 0.90 ± 0.08 | 0.90 ± 0.10 | 0.89 ± 0.09 | 0.88 ± 0.10 |
| **GNB** | **MI_Descend** | 0.88 ± 0.12 | 0.83 ± 0.20 | 0.86 ± 0.12 | 0.86 ± 0.12 | 0.83 ± 0.11 | 0.79 ± 0.12 | 0.78 ± 0.11 | 0.78 ± 0.11 | 0.78 ± 0.11 | 0.76 ± 0.11 |
| | **PLS_Descend** | 0.81 ± 0.12 | 0.83 ± 0.12 | 0.81 ± 0.12 | 0.83 ± 0.12 | 0.79 ± 0.12 | 0.78 ± 0.11 | 0.79 ± 0.13 | 0.80 ± 0.13 | 0.79 ± 0.12 | 0.76 ± 0.11 |
| | **MI+PLS_Descend** | 0.89 ± 0.12 | 0.84 ± 0.10 | 0.84 ± 0.13 | 0.83 ± 0.13 | 0.84 ± 0.13 | 0.80 ± 0.13 | 0.80 ± 0.13 | 0.78 ± 0.11 | 0.76 ± 0.11 | 0.76 ± 0.11 |

the voxels and added voxels one-by-one to our classification models (in the similar fashion as the above two approaches). This approach is employed to demonstrate that our informativeness indexes used to prioritize (rank) the voxels were able to achieve significantly better accuracies compared to the baseline.

Fig. 3 shows the accuracy curves from three different classifiers, GNB, LR, and SVM, using the above-mentioned feature ranking criteria. The results illustrated in the figures indicate that the *MI_Descend* and *PLS_Descend* yielded higher classification accuracies in all cases with different percentages of included voxels, when compared to the *MI_Ascend* and *PLS_Ascend*. Among the three classifiers, SVM and LR yielded better accuracies than GB, with performance results not significantly different between MI_based and PLS_based rankings. The fact that the *MI_Ascend* and *PLS_Ascend* yielded the worst performance indicates that voxels with low MI or PLS values are noninformative with respect to the stimuli. Thus, these results demonstrate that both MI and PLS indexes are beneficial for feature ranking and removing noninformative voxels. It is interesting to see in all shuffle cases shown in Fig. 3 that the curves are almost monotonically increasing with respect to $n$, suggesting that when the voxels are selected randomly for each time, using more features can improve the performance. However, performance never reached the highest accuracy level obtained with the *MI_Descend* and *PLS_Descend* methods. Thus, these results demonstrate that both MI and PLS indexes are beneficial for feature ranking and selection of an optimally informative set of voxels.

Table I reports the classification accuracies of the top percentage of voxels in the ROI (e.g., 10%, 20%, ..., 100%) ranked by *MI_Descend* and *PLS_Descend* obtained by using three classification algorithms. In addition to individual MI and PLS indexes, we also introduced an index that takes into account the joint effect of both MI and PLS importance indexes, called *MI+PLS_Descend*, to rank voxels. These results suggest that the *MI_Descend* and *PLS_Descend* criteria of voxel ranking achieved high classification accuracies with fewer voxels. Specifically, as shown results in the table, using only top-ranked 20%–40% voxels achieved the best possible classification performance.

### *C. Prospective Classification Implementation and Evaluation*

It is important to note that the results in the previous subsection were based on the best possible parameter setting and the reported accuracies were an optimistic estimate (i.e., if one knew *a priori* what parameters to use to select the best possible voxel set that would result in an optimal performance). In this experiment, we performed a true prospective analysis of the 10-fold cross validation. Because we did not know before hand how many or percentage of top-ranked voxels to use as a voxel set, we used the training and validation datasets to identify the best voxel set, and employed our classifiers with only the best voxel set to the testing dataset. To find the best voxel set, we employed the *maxI* strategy.

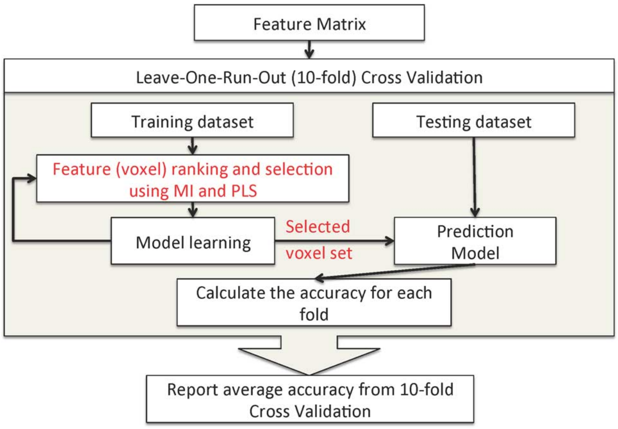

Feature Matrix
Leave-One-Run-Out (10-fold) Cross Validation
Training dataset
Testing dataset
Feature (voxel) ranking and selection using MI and PLS
Model learning
Selected voxel set
Prediction Model
Calculate the accuracy for each fold
Report average accuracy from 10-fold Cross Validation

Fig. 2. MVPA framework includes extracting representative feature values, generating importance indexes of features, selecting informative voxels, and training classification model.

Fig. 2 provides an overall computational procedure in our experiments in the subsequent subsections. A 10-fold cross-validation was employed to train and test the unbiased classification performance. Specifically, the dataset was divided into 10 folds (runs), each having eight data instances (one for each stimulus category) in each run. In each cross validation, one fold was retained as a testing dataset while the remaining nine folds were further divided into a validation dataset (three folds) and a training dataset (six folds). The training dataset was used to optimize the model parameter $\theta$ for the classifier $\Omega$ whereas the validation dataset was used to optimize the regularization parameter $\lambda$ and to calibrate the threshold $\alpha$ for the selection strategy. After all optimal parameters were identified, the classification model was then used to predict the stimulus categories of all eight data instances in the testing set. This cross-validation step was repeated 10 times until all folds were used as a testing set. The classification accuracy, as mentioned, refers to an average accuracy across 10 folds.

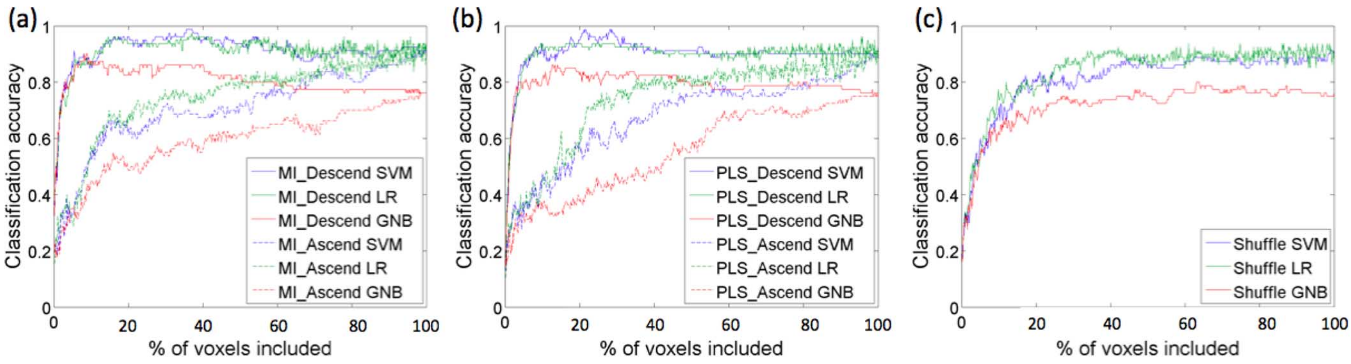


Fig. 3. Classification accuracies (on testing dataset) of SVM, LR, and GNB when voxels in VT cortex are ranked in descending and ascending orders with MI based importance index in (a) and PLS based importance index in (b), compared to when voxels are randomly ranked (shuffled) in (c). (Best viewed in color, available online)

TABLE II
CLASSIFICATION ACCURACIES FOR TRAINING AND TESTING DATA USING THE BEST CALIBRATION OF THE THRESHOLD $\alpha^*$

| MI_based Classifier | SVM | LR | GNB | |
|---|---|---|---|---|
| Best threshold ($\alpha^*$) | 0.50 | 0.49 | 0.52 | |
| Best Training Accuracy | 0.88 ± 0.08 | 0.94 ± 0.05 | 0.75 ± 0.11 | |
| Testing Accuracy | 0.88 ± 0.18 | 0.90 ± 0.10 | 0.75 ± 0.16 | |

| PLS_based Classifier | SVM | LR | GNB | PLS |
|---|---|---|---|---|
| Best threshold ($\alpha^*$) | 0.43 | 0.35 | 0.42 | |
| Best Training Accuracy | 0.87 ± 0.09 | 0.93 ± 0.04 | 0.78 ± 0.10 | 0.80 |
| Testing Accuracy | 0.93 ± 0.09 | 0.88 ± 0.08 | 0.70 ± 0.19 | 0.80 |

TABLE III
NUMBERS OF SELECTED VOXELS REPORTED FROM THE BEST CALIBRATION OF THE THRESHOLD $\alpha^*$

| | MI_based Classifier | | | PLS_based Classifier | | |
|---|---|---|---|---|---|---|
| Fold | SVM | LR | GNB | SVM | LR | GNB |
| 1 | 51 | 213 | 213 | 194 | 262 | 577 |
| 2 | 213 | 116 | 534 | 194 | 133 | 54 |
| 3 | 116 | 116 | 116 | 133 | 432 | 16 |
| 4 | 51 | 534 | 28 | 194 | 432 | 262 |
| 5 | 11 | 339 | 11 | 54 | 499 | 133 |
| 6 | 339 | 447 | 213 | 194 | 262 | 54 |
| 7 | 577 | 577 | 577 | 577 | 262 | 577 |
| 8 | 577 | 577 | 51 | 577 | 568 | 350 |
| 9 | 339 | 339 | 339 | 133 | 568 | 432 |
| 10 | 570 | 339 | 577 | 350 | 575 | 499 |
| Average | 284.4 | 359.7 | 265.9 | 260.0 | 399.3 | 295.4 |
| Intersection | 11 | 116 | 11 | 54 | 133 | 16 |
| Union | 577 | 577 | 577 | 577 | 575 | 577 |

Table II shows classification accuracies for training and testing datasets from the best calibration employing the *maxI* procedure to maximize the informativeness of selected voxels. Only voxels whose MI_based and PLS_based importance indexes, on average, exceed a threshold of $\alpha = 0.35 - 0.52$ are selected and yield the best classification accuracy.

Table III presents the distribution of the numbers of selected voxels as well as intersection and union sets of the selected voxels. Among the three classifiers, LR yielded a more consistent voxel selection across 10 folds even though the number of selected voxels was quite large. To investigate the robustness of voxel selection, we ran 10-fold cross validation 10 times using both the intersect and the union of voxel sets across 10 folds. The accuracy results are presented in Table IV. It is observed that LR yielded the highest accuracy using the intersection of selected voxel sets.

TABLE IV
AVERAGE ACCURACY ACROSS 10-FOLD CROSS VALIDATION USING BOTH THE INTERSECTION AND THE UNION OF VOXEL SETS WITH THE BEST CALIBRATED THRESHOLD $\alpha^*$

| | | MI_based Classifier | | |
|---|---|---|---|---|
| | | SVM | LR | GNB |
| Intersect. | No. of voxels | 11 | 116 | 11 |
| | Accuracy | 0.76 ± 0.13 | 0.94 ± 0.08 | 0.80 ± 0.13 |
| Union | No. of voxels | 577 | 577 | 577 |
| | Accuracy | 0.87 ± 0.11 | 0.90 ± 0.09 | 0.72 ± 0.14 |

| | | PLS_based Classifier | | |
|---|---|---|---|---|
| | | SVM | LR | GNB |
| Intersect. | No. of voxels | 54 | 133 | 16 |
| | Accuracy | 0.91 ± 0.13 | 0.94 ± 0.08 | 0.73 ± 0.12 |
| Union | No. of voxels | 577 | 575 | 577 |
| | Accuracy | 0.87 ± 0.11 | 0.89 ± 0.10 | 0.72 ± 0.14 |

### *D. Classification Results of Voxels From the Whole Brain*

We extended our experiment to evaluate the classification accuracy curves of the three classifiers using the *MI_Descend* when using all voxels in the whole brain. The results shown in Fig. 4 and Table V indicate that LR yields the best overall accuracy and maximum accuracy. The accuracies of all the classifiers grow steeply within the range of $250 \leq n \leq 500$ voxels and decrease after that. In Fig. 5, we also show the top-ranked 500 voxels based on MI importance index anatomically distributed in the whole brain, compared to the 577 voxels in the VT cortex. This suggests that the bottom half of ranked voxels are either noisy and/or nonrelevant to the stimuli. Importantly these results demonstrate that inclusion of noninformative voxels can substantially degrade classification accuracy, and the *MI_Descend* can reveal the optimal range of informative voxels, which is crucial for classification performance. When considering the variation of the accuracy curves especially at the right end of the curve, it can be inferred that LR is less robust (more sensitive to noise) among all three classifiers. That is because LR relies on a hyperplane whereas SVM takes margin into account and GNB models the class conditional directly. In fact, around the best voxel range $100 \leq n \leq 250$, SVM appears to perform slightly better than LR. GNB always performed worst. Additionally, we compared the performance of the three classifiers with the *MI_Descend* with that of the relevant approaches reported in [10] and we obtained overall better results.

TABLE V
COMPARISON OF ACCURACY FROM THE THREE CLASSIFIERS WHEN ALL THE VOXELS IN THE WHOLE BRAIN ARE RANKED USING THE *MI_DESCEND*. RESULTS REPORTED IN [10] ARE SHOWN IN THE LAST TWO ROWS OF THE TABLE. IN CASE WHERE THE NUMBER OF VOXELS USED IN THE ORIGINAL PAPER IS NOT IDENTICAL TO THAT IN OUR EXPERIMENT, WE PUT THE ORIGINAL VOXEL NUMBER IN THE PARENTHESIS

| **No. of voxels** | 5 | 20 | 40 | 60 | 100 | 200 | 500 | 1000 | 43193 | $best\_accur$ |
|---|---|---|---|---|---|---|---|---|---|---|
| **MI_Descend SVM** | 0.50 | 0.61 | 0.83 | 0.89 | 0.90 | 0.94 | 0.93 | 0.84 | 0.19 | 0.95 |
| **MI_Descend LR** | 0.48 | 0.73 | 0.82 | 0.94 | 0.94 | 0.93 | 0.89 | 0.86 | 0.26 | 0.96 |
| **MI_Descend GNB** | 0.50 | 0.60 | 0.66 | 0.73 | 0.71 | 0.78 | 0.78 | 0.69 | 0.14 | 0.79 |
| **SVM +** $w_{SVM}$ | 0.38 | 0.58 | 0.60 (33) | 0.63 (50) | 0.63 | 0.64 | n/a | 0.70 | n/a | 0.70 |
| **RFO +** $w_{SVM}$ | 0.40 | 0.59. | 0.62 (33) | 0.64 (50) | 0.69 | 0.71 | n/a | 0.72 | n/a | 0.72 |

TABLE VI
INTERSECTIONS OF THE VOXEL SETS SELECTED BY THE *MI_DESCEND* AND *PLS_DESCEND* METHODS AND THEIR ASSOCIATED CLASSIFICATION ACCURACIES FOR ALL SIX SUBJECTS FROM THE HAXBY STUDY [1]

| | **% of voxels** | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Subject 01** | **No. of Voxels** | 58 | 115 | 173 | 231 | 289 | 346 | 404 | 462 | 519 | 577 |
| | **No. of overlapping voxels** | 14 | 38 | 72 | 119 | 170 | 232 | 303 | 384 | 472 | 577 |
| | **SVM** | 0.83 ± 0.11 | 0.88 ± 0.13 | 0.92 ± 0.10 | 0.94 ± 0.08 | 0.91 ± 0.11 | 0.93 ± 0.10 | 0.89 ± 0.16 | 0.89 ± 0.1456 | 0.85 ± 0.16 | 0.88 ± 0.17 |
| | **LR** | 0.84 ± 0.08 | 0.88 ± 0.11 | 0.93 ± 0.08 | 0.92 ± 0.11 | 0.90 ± 0.12 | 0.92 ± 0.10 | 0.92 ± 0.10 | 0.92 ± 0.0973 | 0.93 ± 0.10 | 0.92 ± 0.10 |
| | **GNB** | 0.78 ± 0.13 | 0.84 ± 0.12 | 0.86 ± 0.14 | 0.88 ± 0.14 | 0.86 ± 0.16 | 0.85 ± 0.17 | 0.84 ± 0.17 | 0.84 ± 0.1696 | 0.83 ± 0.17 | 0.80 ± 0.20 |
| **Subject 02** | **No. of Voxels** | 46 | 93 | 139 | 186 | 232 | 278 | 325 | 371 | 418 | 464 |
| | **No. of overlapping voxels** | 6 | 23 | 52 | 89 | 134 | 184 | 240 | 302 | 379 | 464 |
| | **SVM** | 0.36 ± 0.11 | 0.52 ± 0.14 | 0.54 ± 0.20 | 0.69 ± 0.14 | 0.72 ± 0.14 | 0.73 ± 0.14 | 0.76 ± 0.14 | 0.78 ± 0.1083 | 0.76 ± 0.11 | 0.75 ± 0.11 |
| | **LR** | 0.38 ± 0.15 | 0.49 ± 0.10 | 0.60 ± 0.22 | 0.76 ± 0.08 | 0.77 ± 0.13 | 0.81 ± 0.13 | 0.76 ± 0.14 | 0.79 ± 0.1231 | 0.77 ± 0.10 | 0.78 ± 0.11 |
| | **GNB** | 0.28 ± 0.14 | 0.46 ± 0.12 | 0.50 ± 0.13 | 0.57 ± 0.16 | 0.64 ± 0.08 | 0.57 ± 0.12 | 0.59 ± 0.11 | 0.55 ± 0.1245 | 0.55 ± 0.11 | 0.56 ± 0.11 |
| **Subject 03** | **No. of Voxels** | 31 | 61 | 92 | 123 | 154 | 184 | 215 | 246 | 276 | 307 |
| | **No. of overlapping voxels** | 5 | 11 | 26 | 50 | 78 | 111 | 153 | 199 | 249 | 307 |
| | **SVM** | 0.45 ± 0.12 | 0.55 ± 0.22 | 0.69 ± 0.14 | 0.82 ± 0.14 | 0.83 ± 0.19 | 0.86 ± 0.16 | 0.82 ± 0.15 | 0.86 ± 0.1125 | 0.85 ± 0.14 | 0.85 ± 0.14 |
| | **LR** | 0.51 ± 0.19 | 0.58 ± 0.23 | 0.72 ± 0.17 | 0.85 ± 0.14 | 0.89 ± 0.10 | 0.90 ± 0.14 | 0.89 ± 0.14 | 0.94 ± 0.0843 | 0.91 ± 0.12 | 0.94 ± 0.10 |
| | **GNB** | 0.52 ± 0.15 | 0.61 ± 0.19 | 0.70 ± 0.18 | 0.71 ± 0.15 | 0.72 ± 0.16 | 0.69 ± 0.17 | 0.71 ± 0.17 | 0.71 ± 0.1539 | 0.69 ± 0.20 | 0.74 ± 0.16 |
| **Subject 04** | **No. of Voxels** | 68 | 135 | 203 | 270 | 338 | 405 | 472 | 540 | 608 | 675 |
| | **No. of overlapping voxels** | 14 | 42 | 82 | 131 | 194 | 263 | 345 | 440 | 550 | 675 |
| | **SVM** | 0.54 ± 0.14 | 0.66 ± 0.12 | 0.69 ± 0.16 | 0.75 ± 0.12 | 0.80 ± 0.12 | 0.76 ± 0.12 | 0.76 ± 0.16 | 0.74 ± 0.1456 | 0.72 ± 0.12 | 0.69 ± 0.16 |
| | **LR** | 0.57 ± 0.16 | 0.72 ± 0.12 | 0.75 ± 0.12 | 0.83 ± 0.11 | 0.79 ± 0.14 | 0.81 ± 0.10 | 0.80 ± 0.11 | 0.84 ± 0.1083 | 0.80 ± 0.11 | 0.83 ± 0.12 |
| | **GNB** | 0.54 ± 0.20 | 0.61 ± 0.17 | 0.67 ± 0.14 | 0.70 ± 0.15 | 0.64 ± 0.11 | 0.63 ± 0.11 | 0.59 ± 0.08 | 0.57 ± 0.1245 | 0.58 ± 0.13 | 0.56 ± 0.13 |
| **Subject 05** | **No. of Voxels** | 42 | 84 | 127 | 169 | 211 | 253 | 295 | 338 | 380 | 422 |
| | **No. of overlapping voxels** | 4 | 19 | 40 | 70 | 107 | 153 | 207 | 271 | 343 | 422 |
| | **SVM** | 0.35 ± 0.24 | 0.55 ± 0.28 | 0.64 ± 0.28 | 0.70 ± 0.31 | 0.66 ± 0.32 | 0.66 ± 0.29 | 0.66 ± 0.28 | 0.72 ± 0.3216 | 0.70 ± 0.33 | 0.67 ± 0.33 |
| | **LR** | 0.35 ± 0.15 | 0.52 ± 0.29 | 0.61 ± 0.24 | 0.66 ± 0.31 | 0.68 ± 0.34 | 0.68 ± 0.32 | 0.74 ± 0.31 | 0.76 ± 0.3375 | 0.77 ± 0.34 | 0.77 ± 0.34 |
| | **GNB** | 0.33 ± 0.20 | 0.52 ± 0.24 | 0.58 ± 0.26 | 0.60 ± 0.28 | 0.57 ± 0.28 | 0.60 ± 0.27 | 0.58 ± 0.25 | 0.63 ± 0.285 | 0.61 ± 0.28 | 0.64 ± 0.29 |
| **Subject 06** | **No. of Voxels** | 35 | 70 | 104 | 139 | 174 | 209 | 244 | 278 | 313 | 348 |
| | **No. of overlapping voxels** | 5 | 16 | 33 | 58 | 88 | 128 | 172 | 223 | 282 | 348 |
| | **SVM** | 0.50 ± 0.16 | 0.71 ± 0.19 | 0.80 ± 0.15 | 0.84 ± 0.12 | 0.83 ± 0.11 | 0.86 ± 0.15 | 0.85 ± 0.13 | 0.88 ± 0.141 | 0.85 ± 0.15 | 0.88 ± 0.13 |
| | **LR** | 0.55 ± 0.16 | 0.74 ± 0.16 | 0.81 ± 0.10 | 0.83 ± 0.11 | 0.88 ± 0.09 | 0.90 ± 0.10 | 0.90 ± 0.10 | 0.92 ± 0.0814 | 0.90 ± 0.10 | 0.91 ± 0.09 |
| | **GNB** | 0.49 ± 0.17 | 0.69 ± 0.16 | 0.71 ± 0.13 | 0.73 ± 0.13 | 0.75 ± 0.15 | 0.77 ± 0.15 | 0.77 ± 0.15 | 0.75 ± 0.1685 | 0.69 ± 0.20 | 0.71 ± 0.19 |

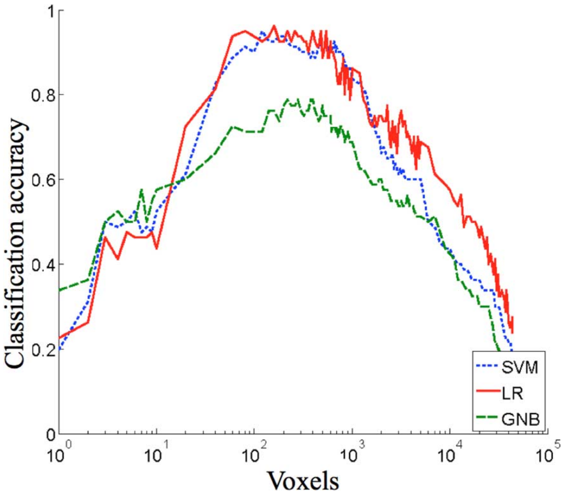


Fig. 4. Accuracy curves from three classifiers using the *MI_Descend* for examining the voxels in the whole brain. (Best viewed in color, available online)

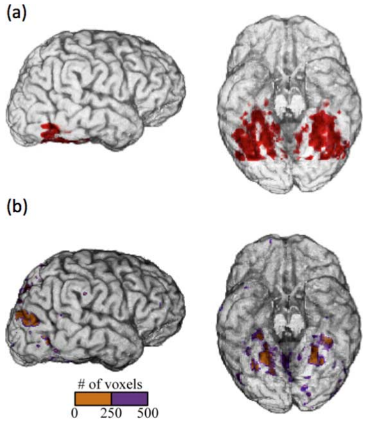
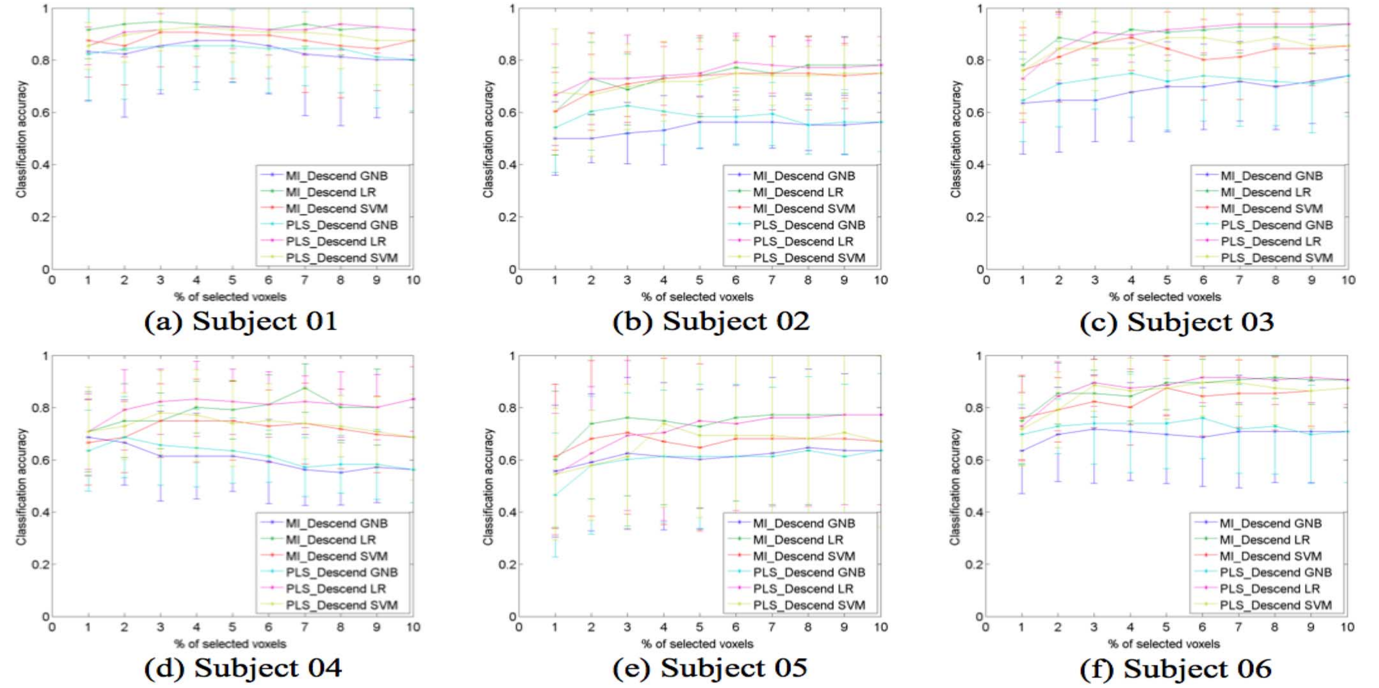


Fig. 5. Anatomical distribution of the top-ranked 500 voxels based on MI importance index in the whole brain (b), compared to the 577 voxels in the VT cortex ROI (a). (Best viewed in color, available online)

## E. Generalization to Multiple Subjects

To demonstrate that our technique is useful in general, we applied the proposed voxel ranking and selection approaches to all six subjects from the original Haxby study [1] and later used in [46], [47]. These fMRI datasets were downloaded from http://data.pymvpa.org/datasets/haxby2001/. There are six subjects named *Subject* 01, 02, 03, 04, 05, and 06. The data for these subjects differed from the benchmark dataset in that there are 12 (instead of 10) runs in total for each subject, with the data for *Subject* 01 an extended version of the benchmark dataset, which was used earlier in this paper. Note as reported, the ninth



(a) Subject 01 (b) Subject 02 (c) Subject 03

(d) Subject 04 (e) Subject 05 (f) Subject 06

Fig. 6. Classification performance (on testing data) from SVM, LR, and GNB when ranking voxels in VT cortex with the *MI_Descend* and *PLS_Descend* for additional multiple subjects under the same experiments in [1], [46], [47].

run of *Subject* 05 was corrupted and therefore not used for the analyses.

Fig. 6(a)–(f) presents the classification accuracies from the three classifiers with the ranking criteria *MI_Descend* and *PLS_Descend* for the six subjects. The performance among the classifiers are considerably consistent with the previous results. GNB yields the worst accuracy compared to the two other classifiers. The classification accuracies vary across the six subjects. Lower accuracies are observed in *Subject* 02 and *Subject* 05. Notably, accuracy does not increase drastically after the top 20%–30% of voxels are selected, and inclusion of additional voxels degrades performance somewhat. The distribution of intersection of the voxel sets selected by the two criteria *MI_Descend* and *PLS_Descend* of the top voxels is reported in Table VI. We observe that the percentage of intersection of the voxel sets increases steadily as the percentage of the voxels increases for the six subjects. The associated classification accuracies from the three classifiers are also presented in the bottom three rows of each subject's results.

## V. Conclusion

In this paper, we proposed a new voxel prioritization and selection framework that ranks voxels with respect to the degree of their association to the experimental conditions. Two metrics based on MI and PLS regression were proposed to score and establish informativeness indexes to select voxels, before building a classification model in MVPA. Using a benchmark fMRI dataset [1], we demonstrated the utility of our approach by assessing the impact of feature (i.e., voxel) selection on classification accuracy with three linear classifiers: LR, SVM, and GNB. Our results illustrated that the use of MI and PLS as importance indexes successfully improved overall classification performance. The proposed selection strategy guaranteed inclusion of the best selection of informative voxels in the classification model, leading to the highest classification accuracy when limiting the set of voxels under consideration to a functional-anatomic predefined region (i.e., VT cortex). We additionally showed our results from SVMs outperformed several machine learning approaches previously reported in the literature [10]. Within the VT cortex, classification accuracy plateaued with the inclusion of roughly the top 100 MI-ranked voxels, consistent with the idea of redundantly distributed encoding of object-level representations across voxels in this region. Consideration of voxels within the whole brain revealed substantial degradation of classification performance after inclusion of roughly the top 1000 MI-ranked voxels. These results underscore the impact of feature selection on the ability to decode cognitive states, and support functional localization of cognitive processes. Interestingly the top MI-ranked voxels from the whole brain analysis overlapped with those in the *a priori* functional anatomical defined region of interest. This correspondence suggests that feature ranking and selection can localize brain functionality and thereby improve the scientific interpretability of classification results.

In parallel, we have worked on developing a new clustering method to identify class-specific regions associated to different stimuli across multi-subjects. In ongoing work, the proposed approach will be applied to other relevant task-based datasets (e.g., lexical processing). We also plan to more extensively compare our MI_based and PLS_based feature selection approaches with other such methods, for example, a "stability score" [48] and the multivariate MI approach proposed in [12].

## ACKNOWLEDGMENT

The authors would like to thank the Haxby research group for sharing the datasets publicly for research purpose.

## REFERENCES

[1] J. V. Haxby, M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, and P. Pietrini, "Distributed and overlapping representations of faces and objects in ventral temporal cortex," *Science*, vol. 293, no. 5539, pp. 2425–2430, 2001.

[2] R. A. Poldrack, J. A. Mumford, and T. E. Nichols, *Handbook of Functional MRI Data Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[3] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Machine Learn. Res.*, vol. 3, pp. 1157–1182, 2003.

[4] K. A. Norman, S. M. Polyn, G. J. Detre, and J. V. Haxby, "Beyond mind-reading: Multi-voxel pattern analysis of fMRI data," *RENDS Cognitive Sci.*, vol. 10, no. 9, pp. 424–430, 2006.

[5] T. M. Mitchell, R. Hutchinson, R. S. Niculescu, F. Pereira, and X. Wang, "Learning to decode cognitive states from brain images," *Machine Learning*, vol. 57, pp. 145–175, 2004.

[6] J.-D. Haynes and G. Rees, "Decoding mental states from brain activity in humans," *Neuroscience*, vol. 7, pp. 523–534, 2006.

[7] J. Mourão-Miranda, A. L. Bokde, C. Born, H. Hampel, and M. Stetter, "Classifying brain states and determining the discriminating activation patterns: Support vector machine on functional MRI data," *Neuroimage*, vol. 28, no. 4, pp. 980–995, 2005.

[8] J. Mourão-Miranda, E. Reynaud, F. McGlone, G. Calvert, and M. Brammer, "The impact of temporal compression and space selection on SVM analysis of single-subject and multi-subject fMRI data," *Neuroimage*, vol. 33, no. 4, pp. 1055–1065, 2006.

[9] F. D. Martino, G. Valente, N. Staeren, J. Ashburner, R. Goebel, and E. Formisano, "Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns," *NeuroImage*, vol. 43, pp. 44–58, 2008.

[10] L. I. Kuncheva and J. J. Rodréguez, "Classifier ensembles for fMRI data analysis: An experiment," *Magn. Reson. Imag.*, vol. 28, pp. 583–593, 2010.

[11] A. Tsai, I. John, W. Fisher, C. Wible, I. William, M. Wells, J. Kim, and A. S. Willsky, "Analysis of functional MRI data using mutual information," in *Proc. 2nd Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 1999, pp. 473–480.

[12] V. Michel, C. Damon, and B. Thirion, "Mutual information-based feature selection enhances fMRI brain activity classification," in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2008, pp. 592–595.

[13] V. Gómez-Verdejo, M. Martínez-Ramón, J. Florensa-Vila, and A. Oliviero, "Analysis of fMRI time series with mutual information," *Med. Image Anal.*, vol. 16, no. 2, pp. 451–458, 2012.

[14] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever, "Mutual information based registration of medical images: A survey," *IEEE Trans. Med. Imag.*, vol. 22, no. 8, pp. 986–1004, Aug. 2003.

[15] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *J. Bioinformat. Computat. Biol.*, vol. 3, no. 2, pp. 185–205, 2005.

[16] G. D. Tourassia, E. D. Frederick, M. K. Markey, J. Carey, and E. Floyd, "Application of the mutual information criterion for feature selection in computer-aided diagnosis," *Med. Phys.*, vol. 28, no. 12, pp. 2394–2402, 2001.

[17] B. Afshin-Pour, H. Soltanian-Zadeh, G.-A. Hossein-Zadeh, C. L. Grady, and S. C. Strother, "A mutual information-based metric for evaluation of fMRI data-processing approaches," *Human Brain Mapp.*, vol. 32, no. 5, pp. 699–715, 2011.

[18] C.-A. Chou, K. B. Kampa, S. H. Mehta, R. F. Tungaraza, W. A. Chaovalitwongse, and T. J. Grabowski, "Information-theoretic based feature selection for multi-voxel pattern analysis of fMRI data," in *Brain Informatics*, F. Zanzotto, S. Tsumoto, N. Taatgen, and Y. Yao, Eds. Berlin, Germany: Springer, 2012, vol. 7670, pp. 196–208.

[19] F. Lindgren, P. Geladi, S. Räannar, and S. Wold, "Interactive variable selection (IVS) for PLS: Part 1. Theory and algorithms," *J. Chemometr.*, vol. 8, pp. 349–363, 1994.

[20] L. Breiman, "Better subset selection using the nonnegative garrote," *Technometrics*, vol. 37, pp. 373–384, 1995.

[21] L. Breiman, "Heuristics of instability and stabilization in model selection," *Ann. Stat.*, vol. 24, no. 6, pp. 2350–2383, 1996.

[22] M. Forina, C. Casolino, and C. Millan, "Iterative predictor weighting (IPW) PLS: A technique for the elimination of useless predictors in regression problems," *Chemometr. Intell. Lab. Syst.*, vol. 12, pp. 165–184, 1999.

[23] P. Bastien, V. E. Vinzi, and M. Tenenhaus, "PLS generalised linear regression," *Computat. Stat. Data Analysis*, vol. 48, pp. 17–46, 2005.

[24] A. Kondylis and J. Whittaker, "Adaptively preconditioned Krylov spaces to identify irrelevant predictors," *Chemometr. Intell. Lab. Syst.*, vol. 104, pp. 205–213, 2010.

[25] S. Wold, M. Sjöströma, and L. Erikssonb, "PLS-regression: A basic tool of chemometrics," *Chemometr. Intell. Lab. Syst.*, vol. 58, no. 2, pp. 109–130, 2001.

[26] N. Kettaneha, A. Berglundb, and S. Wold, "PCA and PLS in very large datasets," *Computat. Stat. Data Analysis*, vol. 48, pp. 69–85, 2005.

[27] P. R. Nelson, J. F. MacGregor, and P. A. Taylor, "The impact of missing measurements on PCA and PLS prediction and monitoring applications," *Chemometr. Intell. Lab. Syst.*, vol. 80, pp. 1–12, 2006.

[28] M. J. Anzanello, S. L. Albin, and W. A. Chaovalitwongse, "Selecting the best variables for classifying production batches into two quality classes," *Chemometr. Intell. Lab. Syst.*, vol. 97, pp. 111–117, 2009.

[29] M. J. Anzanello, S. L. Albin, and W. A. Chaovalitwongse, "Multicriteria variable selection for classification of production batches," *Eur. J. Operat. Res.*, vol. 218, pp. 97–105, 2012.

[30] A. R. Mcintosh, F. Bookstein, J. V. Haxby, and C. L. Grady, "Spatial pattern analysis of functional brain images using partial least squares," *Neuroimage*, vol. 3, pp. 143–157, 1996.

[31] A. R. McIntosh, W. Chau, and A. Protzner, "Spatiotemporal analysis of event-related fMRI data using partial least squares," *NeuroImage*, vol. 23, pp. 764–775, 2004.

[32] A. R. McIntosh and N. J. Lobaugh, "Partial least squares analysis of neuroimaging data: Applications and advances," *NeuroImage*, vol. 23, pp. S250–S263, 2004.

[33] A. Krishnan, L. J. Williams, A. R. McIntosh, and H. Abdi, "Partial least squares (PLS) methods for neuroimaging: A tutorial and review," *NeuroImage*, vol. 56, pp. 455–475, 2011.

[34] K. J. Friston, P. Fletcher, O. Josephs, A. Holmes, M. D. Rugg, and R. Turner, "Event-related fMRI: Characterizing differential responses," *NeuroImage*, vol. 7, pp. 30–40, 1998.

[35] N. Kriegeskorte, "Pattern-information analysis: From stimulus decoding to computational-model testing," *NeuroImage*, vol. 56, no. 2, pp. 411–421, 2011.

[36] C. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, 1948.

[37] S. Kullback, *Information Theory and Statistics*. Mineola, NY: Dover, 1997.

[38] J. A. T. Thomas and M. Cover, *Elements of Information Theory*. New York: Wiley, 2006.

[39] D. Xu and S. L. Albin, "Manufacturing start-up problem solved by mixed-integer quadratic programming and multivariate statistical modeling," *Int. J. Prod. Res.*, vol. 40, no. 3, pp. 625–640, 2002.

[40] H. Abdi, "Partial least squares (PLS) regression," in *Encyclopedia of Measurement and Statistics*, Thousand Oaks, CA, 2007, pp. 740–744.

[41] J. A. Westerhuis, T. Kourti, and J. F. MacGregor, "Analysis of multiblock and hierarquical PCA and PLS models," *J. Chemometr.*, vol. 12, pp. 301–321, 1998.

[42] S. Wold, J. Trygg, A. Berglund, and H. Antti, "Some recent developments in PLS modeling," *Chemometr. Intell. Lab. Syst.*, vol. 58, no. 2, pp. 131–150, 2001.

[43] T. M. Mitchell, *Machine Learning*. New York: McGraw Hill, 1997.

[44] A. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes," *Adv. Neural Inf. Process. Syst.*, vol. 14, p. 841, 2002.

[45] F. Pereira, T. Mitchell, and M. Botvinick, "Machine learning classifiers and fMRI: A tutorial overview," *NeuroImage*, vol. 45, pp. 199–209, 2009.

[46] S. J. Hanson, T. Matsuka, and J. V. Haxby, "Combinatorial codes in ventral temporal lobe for object recognition: Haxby (2001) revisited: Is there a "face" area?," *NeuroImage*, vol. 23, pp. 156–166, 2004.

[47] A. J. O'Toole, F. Jiang, H. Abdi, and J. V. Haxby, "Partially distributed representations of objects and faces in ventral temporal cortex," *J. Cognit. Neurosci.*, vol. 17, pp. 580–590, 2005.

[48] T. M. Mitchell, S. V. Shinkareva, A. Carlson, K.-M. Chang, V. L. Malave, R. A. Mason, and M. A. Just, "Predicting human brain activity associated with the meanings of nouns," *Science*, vol. 320, pp. 1191–1195, 2008.