Randomized Structural Sparsity via Constrained Block Subsampling for Improved Sensitivity of Discriminative Voxel Identification

Yilun Wang\textsuperscript{a}, Junjie Zheng\textsuperscript{b}, Sheng Zhang\textsuperscript{a}, Xunjuan Duan\textsuperscript{b}, Huafu Chen\textsuperscript{b,∗}

\textsuperscript{a} Yilun Wang and Sheng Zhang are both with the School of Mathematical Sciences, University of Electronic Science and Technology of China, Chengdu, Sichuan, 611731 P. R. China.

\textsuperscript{b} Key laboratory for Neuroinformation of Ministry of Education, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, Sichuan, P. R. China.

Abstract

In this paper, we focus on the sparse regularization based voxel selection for functional Magnetic Resonance Imaging (fMRI, for short) brain data. The main difficulty lies in the extremely high dimensional voxel space and very few training samples, easily resulting in unstable feature selection results for many traditional multivariate feature selection methods. In order to deal with the difficulty, stability selection based on the common sparsity models have recently received great attentions, especially due to its finite sample control for certain error rates of false discoveries and transparent principle for choosing a proper amount of regularization. However, it fails to consider some prior or estimated incoherent structural information of these features and could lead to large false negative rates in order to keep small false positive rate. Correspondingly, we propose a new variant of stability selection named “Randomized Structural Sparsity”, which incorporates the idea of structural sparsity, and its advantages over the common stability selection is the better control of false negatives, yet keeping the control of false positives. Numerical experiments demonstrate that our method can achieve better controls of both false positives and false negatives than many alternatives.

Keywords: voxel selection, structural sparsity, stability selection,

∗Corresponding Author
Email address: chenhf@uestc.edu.cn (Huafu Chen)
1. Introduction

1.1. Problem Statement

Learning from neuroimaging data, also called brain reading or brain decoding, as a kind of pattern recognition, has led to impressive results, such as guessing which image a subject is looking at from his brain activity (Haxby et al., 2001), as well as medical diagnosis, e.g., finding out whether the person is a healthy control or a patient. While we will mainly consider the function MRI data, our algorithm might be applied to other kinds of modalities. Specifically, from viewpoint of pattern recognition, we consider each fMRI volume (referred to as a sample) consisting of $p$ voxels (features) as a vector in a $p$ dimensional feature space and each sample is labelled or classified reflecting different experimental conditions or brain states, and then the obtained relationship between the voxels and their corresponding labels are tested with another independent set of samples.

The above pattern recognition procedure typically consists of two important components, feature selection and classifier designing. While the prediction or classification accuracy of these designed classifiers are mainly considered in most existing literatures, we are focusing more on the first component, i.e. feature selection, which is aiming to accurately identify the true and comprehensive discriminative or called predictive features. This task is also called support identification, which is even a more important goal in many practical cases including the medical diagnosis where these selected voxels can be used as biomarker candidates (Guyon and Elisseeff, 2003). While these two goals are often consistent, they might be contrast with each other to certain degree (Li et al., 2009). Specifically, comprehensively selecting the most relevant variables or features is usually suboptimal for building a predictor, particularly if these features are redundant (Guyon and Elisseeff, 2003; Blum and Langley, 1997), which is quite true for fMRI data (Chu et al., 2012). In other words, classic feature selection methods aim at selecting a minimum subset of features to construct a classifier of the best predictive accuracy and often ignore “stability in the algorithm design (Yu et al., 2008). In addition, for the cases of high dimensional feature space and very few samples, feature selection is even harder in terms of controlling false positive
selections and false negative selections, especially when features are often of strong interrelations (Bühlmann et al., 2011). In this paper, we will be mainly considering finding or ranking all potentially relevant features, i.e. accurately identifying all the brain activation regions involved in the certain cognitive processing of an external stimuli or a given disease in the context of neuroimaging. Specifically, while the specificity is expected to be controlled with a given false positive rate, we expect the method to be among the most sensitive, i.e. detecting as many true positives as possible.

While the dominant approaches for finding the discriminative regions based on the neuroimaging data have been univariate due to their being directly testable, easily interpretable, and computationally tractable, the brain is intrinsically multivariate and correspondingly there exists a strong growth of interest in multivariate approaches for analysing brain activity patterns in recent years (Haxby et al., 2014). However, it has been argued that many existing multi-variable pattern analysis methods often fail to provide a reliable result as to support identification partially due to the inherent significant run to run variability in the decision space generated by the classifiers in terms of even very small changes to the training set (Anderson and Oates, 2010). It becomes even worse when the feature space is of high dimensionality, which refers to the situation where the number of to be estimated parameters is one or several orders of magnitude larger than the number of samples in the data (Bühlmann and Van De Geer, 2011). In such cases, non-degenerate classes will always be linearly separable (Cover, 1965), and a classifier can then trivially obtain 100% correct on any training set, without it necessarily having accurately captured all the discriminative features of the data (Raizada and Kriegeskorte, 2010). In other words, classic feature selection methods aim at selecting a minimum subset of features to construct a classifier of the best predictive accuracy and often ignore “stability” in the algorithm design (Yu et al., 2008; Bühlmann et al., 2011).

Correspondingly, in this paper, we aim to tackle the above “curse of dimensionality” difficulty and propose an efficient algorithmic framework of multivariate pattern recognition for stable, reliable and comprehensive support identification. Specifically, while the specificity is expected to be controlled with a given false positive rate, we expect the method to be among the most sensitive, i.e. detecting as many true positives as possible. Notice that the higher sensitivity is not necessarily reflected by the higher prediction accuracy as mentioned before, and therefore, we will mainly focus on the direct quality assessment of the selected discriminative voxels.
1.2. Advantages and Limitations of Sparsity Applied to Neuroimaging

In this paper, we consider the commonly used supervise learning to identify the discriminative brain voxels from the given labeled training fMRI data. While the classification problem is mainly considered, the regression problem can be treated in a similar way. For purpose of feature selection, the linear model is usually adopted as follows, though it might proceed with a nonlinear classifier, if the classification task will be considered.

\[ y = Xw + \epsilon \] (1)

where \( y \in \mathbb{R}^{n \times 1} \) is the binary classification information and \( X \in \mathbb{R}^{n \times p} \) is the given training fMRI data and \( w \in \mathbb{R}^{p \times 1} \) is the unknown weights reflecting the degree of importance of each voxel. As a multivariate inverse inference problem, identification of discriminative voxels is based on the values of the weight vector \( w \) and their importance is proportional to the absolute values of the components.

Considering that the common challenge in this field is the curse of dimensionality \( p \gg n \), we are focusing on the sparsity based voxel selection methods, because sparsity makes much sense that the most discriminative activated voxels are only a small portion of the whole brain voxels (Yamashita et al., 2008).

However, sparsity alone is not sufficient for making reasonable and stable inferences, in cases of very high voxel space and small number of training samples. In such cases, the plain sparse learning models often provide overly sparse and hard to interpret solutions where the selected voxels are often assigned to be individually scattered (Rasmussen et al., 2012), with either high false positive rate or false negative rate. Specifically, if there are a set of discriminative features are highly correlated to each other, then only a small part of representative voxels are selected, resulting into a big false negative rate. Thus we have to extend the plain sparse learning model to incorporate some important structural features of brain imaging data such as brain segregation and integration, in order for stable, reliable and interpretable results.

Specifically, the Functional MRI (fMRI) is based on Blood Oxygen-Level Dependent (BOLD) signal contrast to gives images of brain activity. It is the latest in a line of techniques to study human brain function that use brain perfusion as a proxy for brain function (Pekar, 2006), instead of making any direct measurements of brain function. An fMRI scanner measures the blood-oxygenation-level-dependent (BOLD) signal at all points in a 3-D grid, and
each grid of the 3-D image is known as a voxel. A typical fMRI dataset is composed of time series (BOLD signals) of thousands of voxels.

In brains of higher vertebrates, the functional segregation of local areas that differ in their anatomy and physiology contrasts sharply with their global integration during perception and behavior (Tononi et al., 1994). fMRI responses consisting of voxels whose three dimensional spatial arrangement respect the anatomy of the brain and the discriminative voxels are thus expected to have a specific localized spatial organization, which is important for the subsequent identification task performed by neuroscientists (Xiang et al., 2009). That is to say, sets of voxels allowing to discriminate between different brain states are expected to form distributed and localized areas (Tononi et al., 1994).

Correspondingly, we aim to propose an efficient algorithmic framework of sparsity based voxel selection models suitable for fMRI data, with the purpose to increase sensitivity with the tolerable false positives, by considering sparsity and clustering effect simultaneously. In addition, better balance between computational efficiency and statistical inference quality is also expected to be achieved (Chandrasekaran and Jordan, 2013).

1.3. Existing Extensions of the Plain Sparse Model

As mentioned above, two common hypothesis have been made for fMRI data analysis. One is the sparsity: few relevant most discriminative voxels/regions implied in the classification task; the other is compact structure: relevant discriminative voxels are grouped into several distributed clusters and the voxels within a cluster have similar behaviors and correspondingly strongly correlated. Thus making use of the above two hypothesis is very important and we will review some state of art existing works along this direction, considering that the classical sparse learning model is usually based on the \( \ell_1 \) regularization, which is only making use of the first sparsity hypothesis and discard the structure related assumption.

Elastic Net (Zou and Hastie (2005)) tries to make use of the voxel correlation by adding an \( \ell_2 \) regularization or called Tikhonov regularization to the classical \( \ell_1 \) penalty (Ryali et al. (2012a)) for establishing a network. Elastic net, selects all together a group of redundant features or none of them in what is known as the grouping effect, because strict convexity due to the help of the added \( \ell_2 \) term to deal with redundancy. Elastic net has been used for classification based on the fMRI data (Ryali et al. (2010)) and has achieved promising results. While the Elastic net might achieve improved
Correspondingly, another class of methods to more explicitly make use of the segregation and integration of the brain, are based on structured sparsity models (Bach et al., 2012b), which have been proposed to extend the well-known plain $\ell_1$ models such as (2) by enforcing more structured constraints on the solution, such as the discriminative voxels are grouped together in possibly few clusters (Baldassarre et al., 2012; Michel et al., 2011). Among them, one of the most popular structured sparsity models is the group sparsity regularization (Xiang et al., 2012; Liu and Ye, 2010; Yuan et al., 2013; Jacob et al., 2009), where the (possibly overlapping) groups have often been known as a prior information. For example, multi-task feature learning via efficient $\ell_{2,1}$-norm minimization has been proposed in (Liu et al., 2009a), and the group sparsity based classifiers has been applied to fMRI classification in (Ng and Abuhgharieh, 2011). However, in many cases, the grouping information is not available beforehand, ones can either use the anatomical regions as an approx (Batmanghelich et al., 2012), or use the data driven methods to obtain the grouping information such as the hierarchical agglomerative clustering (Ward hierarchical clustering, for example) and a top-down step to prune the generated tree of hierarchical clusters for obtaining the grouping information (Michel et al., 2012; Jenatton et al., 2012).

Another important class of methods for high dimensional data analysis is stable selection (Meinshausen and Bhlmann, 2010; Shah and Samworth, 2013), which as a special “ensemble learning” procedures following (Hastie et al., 2009), is an effective way to voxel selection and structure estimation based on subsamplings (bootstrapping would behave similarly). It aims to alleviate the disadvantage of LASSO, which either selected by chance non-informative regions, or even worse, neglected relevant regions that provide duplicate or redundant classification information (Mitchell et al., 2004; Li et al., 2012), partially because of the worrying instability and potential deceptiveness of the most informative voxel sets when information is non-local or distributed (Anderson and Oates, 2010; Poldrack, 2006). Correspondingly, one main advantage of stability selection is the control of false positives, i.e. it is able to obtain the selection probability threshold based on the theoretical bound on the expected number of false positives. In addition, stability selection makes the choice of sparsity penalty parameter do not matter much, and the stability selection has been applied to the pattern recognition based on brain MRI data and achieved better results than the plain $\ell_1$ models (Ye et al., 2012).
In order to make use of the assumption that these discriminative voxels are often spatially contiguous and results in distributed clusters, ones proposed to use the idea of common stability selection together with clustering (Gramfort et al., 2012; Gaël Varoquaux, 2012). However, their implementations required running a clustering during each resampling and their adopted “rescaling” for the columns of $X$ fails to consider the second hypothesis, i.e. the random reweighting during their implemented stability selection is voxel-wise and fails to consider the spacial contiguousness of the clustered discriminative voxels. Their evaluation of different methods are more on the prediction accuracy or classification accuracy, instead of the accuracy of the detected support.

1.4. Our focus and contributions

The key point of this paper is to achieve both computational efficiency and better inferential quality in terms of both smaller false positive rate and false negative rate, by making proper use of the above two hypothesis, especially that correlated discriminative voxels are often neighbored and locally clustered.

Firstly, we proposed a variant of stability selection based structural sparsity named “randomized structural sparsity”, particularly via the adoption of the “constrained block subsampling” technique for fMRI data, instead of single voxel-wise subsampling in the classical stability selection for the selection of the comprehensive set of discriminative voxels. The “block subsampling” aims to make use of the spatially contiguousness of discriminative voxels. The “constrained” means that we propose to make use of the prior parcelling information of brain or data driven parcelling information to separate or combined the obtained blocked subsamples. Secondly, a multi-stage procedure is proposed to gradually eliminate irrelevant voxels and reduce the sparsity penalty on the detected relevant voxels. Thirdly, we empirically showed that this “blocked” variant of stability selection can achieves significantly better sensitivity with the control of false positives for voxel selection, than other alternatives including the original stability selection. Specifically, we perform the comprehensive comparison with the mass univariate method GLM and the state of the art multivariate methods such as SVM, logistic regression in terms of better voxel selection performance, while most of the existing works have focused on the prediction or classification accuracy instead (Carroll et al., 2009; Charalampous et al., 2012; Michel et al., 2011),
which is not necessarily reflecting the sensitivity of feature selection, due to the redundancy of the discriminative voxels for our cases.

We need to point out that this new algorithm is beyond a simple sum of stability selection and structural stability, and it has the following important extra advantage. In many cases where the structural information such as clustering structures is only a rough approximation, i.e. there is block mischaracterization or the neighboring voxels in the same brain area might be highly correlated but not necessarily all discriminates, the subsampling scheme can help remedy it via supervised refines and still outlines the true shapes of the discriminative regions, as showed by numerical experiments.

The rest of the paper is organized as follows. In section 2 we introduce our new algorithm for stable voxel selection. In section 3 we demonstrate the advantages of our algorithm based on both synthetic data and real fMRI data in terms of both higher sensitivity and specificity. In section 4 a short summary of our work and some possible future research directions will be given.

\section{The Proposed Method}

\subsection{The Background and Motivation}

We aim to propose a new efficient and reliable algorithm to find out the true discriminative features for the available high dimensional training fMRI data with classification information. Denote an fMRI data matrix as $X \in \mathbb{R}^{n \times p}$ where $n$ is the number of samples and $p$ is the number of voxels with $n \ll p$, and corresponding classification information $y \in \mathbb{R}^{n \times 1}$. Here we only consider the binary classification and $y_i \in \{1, -1\}$. While our main ideas can be applied to most existing sparsity based multi-variate model in terms of either prediction or classification tasks, we take the following sparse logistic regression for classification as the example to show the existing difficulties and our corresponding efforts, in details.

\[
\min_{w} \|w\|_1 + \lambda \sum_{i=1}^{n} \log(1 + \exp(-y_i(X^T_i w + c)))
\]

where $X_i$ denotes the $i$-th row of $X \in \mathbb{R}^{n \times p}$, $y \in \mathbb{R}^{n \times 1}$ is the labeling information containing the classification information of each row of $X$; $w \in \mathbb{R}^{p \times 1}$ is the weight vector for the voxels; $c$ is the intercept (scalar). The voxels
corresponding to \( w_i \) of large absolute value will be considered as the discriminative voxels. When \( p \gg n \), the plain \( \ell_1 \) model [2] can identify at most only \( n \) variables before it saturates (Zou and Hastie, 2005).

Correspondingly, structured sparsity models beyond the plain \( \ell_1 \) models have been proposed to enforce more structured constraints on the solution (Bach et al., 2012b; Li et al., 2013; Mairal and Yu, 2013). As an important special case, the common way to make use of the clustering or grouping structure is to adopt the group sparsity induced norm Bach et al. (2012a), as follows.

\[
\min_w \sum_{g \in G} \| w_g \|_2 + \lambda \sum_{i=1}^{n} \log(1 + \exp(-y_i(w^T X_i + c))),
\]

where \( G \) is the grouping information. Compared with (2), the main difference is the regularization term and now we are using a mixed \( \ell_1/\ell_2 \) norm.

However, the group sparsity-induced norm regularized model (3) expects to improve the sensitivity than the plain \( \ell_1 \) model (2) due to the adopted mixed \( \ell_{2,1} \) norm if the grouping information \( G \) is reliable enough. The obtaining of appropriate \( G \) might be difficult in practice partially because most of methods of obtaining \( G \) are not incorporating the available classification or labelling information, and it might only be derived from either the prior anatomical knowledge or data driven methods based on the voxel correlation. In addition, just like the plain \( \ell_1 \) model, the difficulty of choosing a proper regularization parameter still exists. Moreover, due to the very high dimension and few training samples, the finite sample control of false positives is much important, but not achieved yet.

As mentioned before, an effective way to control the false positives when applying the sparsity regularization based models is stability selection (Meinshausen and Bhlmann, 2010), which have been applied onto voxel selection or connection selection (Ye et al., 2012; Cao et al., 2014; Ryali et al., 2012b). However, the control could be too conservative and cause a big false negative rate. In this paper, we aim to find a mechanism to incorporate the spatial structural knowledge of voxels into the stability selection framework, in order to achieve simultaneous low false positive rate and false negative rate.

### 2.2. The Key Component: Constrained Block Subsampling

In this paper, we propose to use the subsampling based stability selection (Beinrucker et al., 2012), rather than the original reweighting based
stability selection ([Meinshausen and Bhlmann, 2010]). It has been showed that the former generally likely yields an improvement over the latter whenever the latter itself improves over standalone pure ℓ¹ regularization model ([Beinrucker et al., 2012]). Moreover, subsampling is easier for parallelization.

Now let us explain the subsampling based stability in more details. For the training data matrix \( X \in \mathbb{R}^{n \times p} \), subsampling based stability selection consists in applying the baseline, i.e. the pure ℓ¹ regularization model such as (2), to random submatrices of \( X \) of size \([n/L] \times [p/V]\), where \([\cdot]\) is the round off to the nearest integer number, and returning those features having the largest selection frequency. The original stability selection ([Meinshausen and Bhlmann, 2010]) can be roughly considered as a special case, where \( L = 2 \) and \( V = 1 \), except that the original stability selection ([Meinshausen and Bhlmann, 2010]) reweighs each feature (voxel, here) by a random weight uniformly sampled in \([\alpha, 1]\) where \( \alpha \) is a positive number, and subsampling can be intuitively seen as a crude version of this by simply dropping out randomly a large part of the features ([Beinrucker et al., 2012]). Authors in ([Beinrucker et al., 2012]) has also showed that \( L \) reflects the bias/variance tradeoff: as \( L \) increases, the variance is reduced, but on the other hand we expect that the selection probabilities of relevant and noisy features become more similar and thus more difficult to separate since the subsample size is smaller.

For our problems, the relevant features (voxels, here) are intercorrelated and there will be “mutual masking” if the pure ℓ¹ model (2) is applied. That is to say, these features can not be simultaneously selected. While subsampling can help solve this “mutual masking” problem as explained in ([Beinrucker et al., 2012]) to certain degree, we propose to make use of the block subsampling ([Lahiri, 1999]), which tries to replicate the correlation by subsampling instead blocks of data, for better performance. The rationality of “blocking” exists in the assumption that the voxels are partitioned into spatially contiguous homogeneous subgroups. Here the “homogeneity” is defined by the strong correlations of the voxels.

Correspondingly, we prefer to incorporating the parcelling information of the brain into block subsampling, which we name as “constrained block subsampling”, where the “constrained” means that the parcelling information will be respected. The kind of partition information is based on either the prior anatomical knowledge of brain partition ([Tzourio-Mazoyer et al., 2002]), or the clustering results based on the fMRI data, as did in the structural sparsity model (3).

Specifically, for each cluster \( g \in \mathcal{G} \), it may consist of either only one or
several distributed localized brain regions or called partitions. That is to say, a cluster based on “homogenousness” could be distributed into several different brain areas \cite{Tzourio-Mazoyer2002}. In terms of brain network, these areas have strong functional connections, though we do not consider brain network in this paper. Correspondingly, after the common block subsamplings, the selected voxels from the same cluster will be considered as a group. In particular, the chosen voxels lying in a cluster \( g \in \mathcal{G} \) are noted as a set \( g' \subseteq g \). In addition, in order for every brain partition, especially those small ones to be sampled during the block subsampling, we borrow some idea of “proportionate stratified sampling” \cite{Sarndal1986}, i.e. the same sampling fraction is used within each partition. The purpose is to reduce the false negatives, especially when the sizes of different partitions are of quite a range. Correspondingly, ones can solve the following group-sparsity based recovery model.

\[
\min_{w'} \sum_{g' \subseteq g \in \mathcal{G}} \|w'_{g'}\|_2 + \lambda \sum_{i \in \mathcal{J}} \log(1 + \exp(-y_i(w'_{g'}^T X'_{i} + c))) \quad (4)
\]

where \( w' \) and \( X' \) are corresponding parts of \( w \) and \( X \), respectively, based on the selected voxels during the subsampling, and \( \mathcal{G} \) is a predefined partitions of the brain, based on the either biological knowledge or data driven learning or estimation such as clustering. \( \mathcal{J} \) is the set of the indices of the selected samples during the current subsampling.

Notice that while “constrained block subsamplings” respects the prior knowledge \( \mathcal{G} \), it also provides the flexibility that resulted discriminative regions can be of any shape, and the final selected voxels of each cluster can be only portion of all of it, because the randomness of the block subsampling in the different iterations of the stability selection, makes the selection frequency score can outline structures of the true discriminative regions, whose sizes may not be the same as the sizes of the original partitions defined by \( \mathcal{G} \). This kind of flexibility is important because the grouping information is usually obtained via certain clustering algorithms and only depending on the correlation information of voxels and therefore it might not exactly reflect the true shapes of the discriminative regions. Specifically, while the neighboring voxels belonging to the same brain area might be are highly correlated to each other, they are not necessarily all significantly discriminative voxels. In this cases, “constrained block subsamplings” plays an important role to control the false positives and we will show this point in the following synthetic example.
The block size might affect the performance of our algorithm \cite{Lahiri2001}. Given the number of blocks, there is an inherent trade-off in the choice of the size of blocks. When only very limited number of randomizations are allowed, big blocks will most likely not match the geometry of the true support and easily results in high false positives. But too small blocks is likely to result in high false negatives due to the likely ignorance of the local correlations of neighboring voxels. Thus a multi-stage scheme might be helpful to partially remedy this issue. Specifically, at each stage, enough number of randomization are performed until a stable result is obtained. Then the least nondiscriminatory voxels are eliminated. In addition, as this multi-stage procedure procedure, we can gradually reduce the block size from big value to small values for different stages.

In short, “constrained block subsampling” is the combination of structure sparsity and stability selection, where the structure information is mostly reflected in $G$, which is very important for a better result \cite{Yu2012,Huang2009}. In addition, this clustering information helps to reduce the possible combinations of nonzeros of $w$ for subsamplings and provide an efficient exploration of the solution space. Moreover, considering the spacial contiguousness structure of discriminative voxels, the block subsampling is preferred instead of traditional individual sampling. A similar idea has been used in the terms of identifying differences in brain networks based on the mass-univariate testings \cite{Zalesky2010} via cluster-based statistical methods considering that connections comprising the contrast or effect of interest are interconnected. It has showed its advantages by identifying an expansive dysconnected subnetwork which a mass-univariate analysis fails to find for schizophrenia.

For our case of small samples and very high dimensional feature space, we need to consider the bias-variance dilemma or bias-variance tradeoff \cite{Geman1992}. In general, we would like to pay a little bias to save a lot of variance, and dimensionality reduction can decrease variance by simplifying models \cite{James2013}. Correspondingly, while we can still use the (4) as the baseline subproblem for our stability selection framework, we prefer a simple “averaging” idea \cite{Varoquaux2012} applied to (4), because \cite{Park2007} has showed that when the variables or features were positively correlated, their average was a strong feature, yielding a fit with lower variance than the individual variables. Specifically, by averaged the voxels picked by the block subsampling lying in the same group as a single super-voxel, the
model (4), can be further reduced to the following low dimensional version

\[
\min_{\tilde{w}} \sum_{g' \subset g \in G} |\tilde{w}_{g'}| + \lambda \sum_{i \in J} \log(1 + \exp(-y_i(\tilde{w}^T\tilde{X}_i + c)))
\]

(5)

where \(\tilde{w} \in \mathbb{R}^q\), and \(q\) is the number of clusters. \(\tilde{w}_{g'}\) is an average of voxels in the subset \(g'\) of cluster \(g \in G\), and \(\tilde{X} \in \mathbb{R}^{[\alpha n] \times p}\) is the corresponding averaged \(X\). Thus the number of variables of the sparse recovery model (2) is greatly reduced to the number of clusters and the properties of the resulting new data matrix \(\tilde{X}\) is also improved due to de-correlation via the clustering. If the \(j\)-th column of \(\tilde{X}\) is selected due to the large magnitude of \(\tilde{w}_j\), then its represented picked blocked voxels lying in the group \(g^{(j)} \in G\) \((j = 1, 2, \ldots, q)\) of \(X\) are all counted to be selected, in the non-clustered space. Its corresponding score \(s_i\) will be updated \((i = 1, 2, \ldots, p)\). Notice that the “averaging” of sum-sampling is more than a simple spatial smoothing, due to different sum-sampling results of different stability selection iterations. Therefore, the boundaries of the detected discriminative regions can be trusted to certain accuracy.

2.3. Algorithmic framework

In this section, we will give an overall introduction to the algorithmic framework, which is a multi-stage scheme and the number of stages is denoted as \(N\). For each stage, it consists of three parts, i.e. “obtaining partitioning information”, “running stability selection based on constrained block subsampling”, and “removing the least important partitions”, respectively.

Specifically, we perform a clustering operation aiming to partition the voxels into many patches according to their strong local correlations. While we do not prefer any specific algorithm and ones can adopt k-means (Hartigan and Wong, 1979) and hierarchical clustering algorithms (Murtagh, 1985), it might be better that the clustering algorithm will take spacial contiguousness as well as other available brain segregation knowledge into consideration. We denote the set of the groups via the clustering algorithm as \(G\), whose cardinality is denoted as \(q\), which is usually much less than \(n\) and comparable with \(m\).

Now it comes to the “constrained block subsamplings” for reliable support identification. Denote the number of resamplings as \(K\). This blocked variant of stability selection is the same as the classical stability selection in terms of dealing with the samples, or the rows of the data matrix \(X\) by taking only a fraction of the training samples. Let the sub-sampling fraction be \(\alpha \in [0, 1]\) and the indices of selected rows are denoted as \(J\) and the cardinality of
\( J \) is \([\alpha n]\), where \([\cdot]\) is the round off to the nearest integer number. Then “constrained block subsamplings” are applied to the voxels, i.e. the columns of \( X \) as mention in last section.

Then “removing the least important partitions” is performed as follows. After stability selection, the algorithm check the average value of \( s_i \) for each \( g^{(j)} \in \mathcal{G} \) \((j = 1, 2, \ldots, q)\). If for certain \( g^{(j)} \), its averaged \( s_i \) value is almost or close to 0, we think \( g^{(j)} \) probably is not belonging to the true support and therefore we will discard it in the next stage, this way we have a reduced data matrix \( X \), which has a smaller number of columns. In the next stage, we can run the clustering again on the remaining voxels and repeat the randomization, and then remove more irrelevant voxels. These procedure is repeated and consists of a multi-stage coarse-to-fine greedy procedure, especially suitable for high dimensional problems.

The procedure of our algorithm is summarized in the following table, also plotted in Figure 1.

---

**The Algorithmic Framework of Constrained Blocked Stability Selection Method:**

**Inputs:**

1. Datasets \( X \in \mathbb{R}^{n \times p} \)
2. Label or classification information \( y \in \mathbb{R}^n \)
3. Sparse penalization parameter \( \lambda > 0 \)
4. Number of randomizations \( K \) for each stage; the number of stages of voxel elimination \( N \), sub-sampling fraction \( \alpha \in [0, 1] \) in terms of rows of \( X \); sub-sampling fraction \( \beta \in [0, 1] \) in terms of columns of \( X \);
5. Initialized stability scores: \( s_i = 0 \). \((i = 1, 2, \ldots, p)\)

**Output:** Stability scores \( s_i \) for each voxel. \((i = 1, 2, \ldots, p)\)

**for** \( j=1 \) to \( N \)

Obtain parcelling information of brain. For example, perform the clustering of voxels based on their spacial correlation and denote the number of clusters as \( q \)
for k=1 to K

1: Perform sub-sampling in terms of rows: $X \leftarrow X_{[J,:]}$, $y \leftarrow y_J$ where $J \subset \{1, 2, \ldots, n\}$, card($J$) = $\alpha n$, the updated $X \in R^{(\alpha n) \times p}$, and the updated $y \in R^{\alpha n}$.

2: Perform constrained block sub-sampling in terms of columns (voxels): $X' \leftarrow X_{[:,I]}$, where $I \subset \{1, 2, \ldots, p\}$, and card($I$) = $\beta p$

3: Use the current clustering, and calculate the mean of randomly picked voxels within each cluster: $\tilde{X} \leftarrow \text{mean}(X')$, $\tilde{X} \in R^{\alpha n \times q}$

4: Estimate $\tilde{w} \in R^q$ from $\tilde{X}$ and $y$ with sparse logistic regression

5: Set weights for the randomly picked voxels with estimated coefficients of the averaged voxels: $w^{(k)} \leftarrow \tilde{w}$, $w^{(k)} \in R^{\beta p}$

6: $s_i = s_i + 1$, if $i \in \text{supp}(w^{(k)})$, for $i = 1, 2, \ldots, p$.

end for

Remove the irrelevant blocks or contiguous regions of voxels (with small $s_i$) from consideration and correspondingly reduce the dimension of voxel space in next stage
end for

3. Numerical Experiments

In this paper, we will compare our algorithm with the classical univariate voxel selection method, the recently popular state-of-the-art multi-voxel pattern recognition methods including t-test, $\ell_2$-SVM, $\ell_2$ Logistic Regression, and $\ell_1$-SVM, $\ell_1$ Logistic Regression, Elastic Net, and randomized $\ell_1$ logistic regression. We will show that while our method have achieved better control of false positives, it also has smaller false negatives than other alternatives.

3.1. Brief Introduction to Alternative Algorithms

We first briefly review several state of the art algorithms adopted for support identification in this paper. Most of these algorithm have been implemented in LIBLINEAR ([Fan et al., 2008]) or SLEP (Sparse Learning with Efficient Projections) software ([Liu et al., 2009b]), except the two sample t-test, which is implemented as an internal function of MATLAB.
Figure 1: The procedure of our algorithm. The key component is “Constrained Block Subsampling”, which incorporates the estimated grouping information into the block subsampling.
3.1.1. Two sample t-test

A hypothesis test that is used to determine questions related to the mean in situations where data is collected from two random data samples. The two sample t-test is often used for evaluating the means of two variables or distinct groups, providing information as to whether the means between the two populations differs. In this paper, we are using the available MATLAB command \texttt{ttest}.

3.1.2. $\ell_2$-SVM (linear)

Support vector machine (SVM) belongs to supervised learning models used for either classification or regression analysis. In the training stage, an SVM training algorithm builds a model for the given set of training examples marked as belonging to one of two categories, for the purpose of prediction of the new examples into one category or the other, or identification of the involved features based on their trained weights (as did in this paper). The common used linear $\ell_1$-regularized $\ell_2$-loss SVM is as follows with the $\ell_2$ regularization on the weight vector $w$, whose obsolete value of each component represents the importance degree of the corresponding voxel. The notations here are following those of the above sections.

$$\min_w \frac{1}{2} \|w\|^2_2 + C \sum^n_{i=1} (\max(0, 1 - y_i w^T X_i))^2$$

3.1.3. $\ell_1$-SVM (linear)

If $\ell_1$-regularization is used, ones have the following $\ell_1$-regularized $\ell_2$-SVM, which generates a sparse solution $w$.

$$\min_w \|w\|_1 + C \sum^l_{i=1} (\max(0, 1 - y_i w^T X_i))^2$$

where $\| \cdot \|_1$ denotes the $\ell_1$ norm.

3.1.4. $\ell_2$-Logistic Regression

Logistic regression is a type of probabilistic statistical classification model, used to predict the outcome of a categorical dependent variable (i.e., a class label) based on the weighted combination of features. It shares much in common with SVM except that it uses different loss function. Logistic regression measures the relationship between a categorical dependent variable
and many independent feature variables, which are usually (but not necessarily) continuous, by using probability scores as the predicted values of the dependent variable. When the $\ell_2$ regularization is adopted, the corresponding $\ell_2$-logistic regression is as follows.

$$
\min_w \frac{1}{2} \|w\|^2_2 + \lambda \sum_{i=1}^m \log(1 + \exp(-y_i(w^T x_i + c)))
$$

where $c$ is the intercept (scalar).

### 3.1.5. $\ell_1$-Logistic Regression

If the $\ell_2$ regularization is adopted in the logistic regression, the corresponding $\ell_1$-logistic regression is as follows, which prefers a sparse solution.

$$
\min_w \|w\|_1 + \lambda \sum_{i=1}^m \log(1 + \exp(-y_i(w^T x_i + c)))
$$

where $c$ is the intercept (scalar).

### 3.1.6. Elastic Net

Elastic net is a hybrid of $\ell_2$ regularization and $\ell_1$ regularization and it is applied to logistic regression models here. While elastic net can generate reduced models by generating zero-valued coefficients of $w$. Empirical studies have suggested that the elastic net technique can outperform the $\ell_1$ model on data with highly correlated predictors. We will also demonstrate it in the following numerical experiments.

$$
\min_w \|w\|_1 + \beta \|w\|_2 + \lambda \sum_{i=1}^m \log(1 + \exp(-y_i(w^T x_i + c)))
$$

where $c$ is the intercept (scalar).

### 3.1.7. Randomized $\ell_1$-Logistic Regression

It is the stability selection [Meinshausen and Bhlmann, 2010] applied to the $\ell_1$-logistic regression model. Randomization is one key for stable results in the area of brain imaging of the strong correlation. Traditionally, the columns of $X$ corresponding to the voxels are reweighted by a random weight uniformly sampled in $[\gamma, 1]$ ($\gamma \geq 0$).
3.2. Settings of Algorithms

For our algorithm, either LIBLINEAR or SLEP can efficiently solve the model (3), which in fact in only a common $\ell_1$ model. In this paper we use the SLEP software, though LIBLINEAR can also be applied. The block size is not optimized in the following experiments via the probable prior knowledge of the discriminative regions, but still achieves an impressive performance. It was set to be $3 \times 3$ in synthetic data and $4 \times 4 \times 4$ in the real fMRI data experiment, respectively. We set the subsampling rate $\alpha = 0.5$ and $\beta = 0.1$. For our synthetic data and the real fMRI data, we set $N = 2, K = 20$ and correspondingly the total resampling times is $N \times K = 40$. For the selection of regularization parameters of the involved multivariate methods except our method, cross validation is used.

3.3. Evaluation Criteria

For each involved multi-variate pattern recognition methods, picking the discriminative voxels is based on the weight vector $w \in \mathbb{R}^p$ and the magnitude of each entry of the weight vector represents the significance of its corresponding voxel or feature in a classification or regression problem ([Li et al., 2009]). Correspondingly, thresholding has been popular ([Donoho and Jin, 2008]), i.e., voxels whose corresponding $w$ components of larger than a given threshold value can chosen as the discriminative voxels. While the threshold choice itself is a very important and complicate topic, we are not going to discuss it in depth here due to the length of the paper.

In this paper, we mainly aim to demonstrate that our method can achieve a better control of both false positives and false negatives than other alternatives. For this purpose, we uses the Receiver Operating Characteristic (ROC) curves to compare sensitivity versus specificity of these different support detection methods. While ROC is mostly used in terms of classification (prediction) accuracy, it is used to show the voxel selection accuracy, for the synthetic data, where the ground truth of discriminative voxels are known. Similar idea has been adopted in [Zalesky et al., 2010] for identifying network based features. The sensitivity or true positive rate (TPR), and specificity (SPC) or True Negative Rate, are defined as

$$TPR = \frac{TP}{TP + FN}$$

and

$$SPC = \frac{TN}{FP + TN},$$

19
respectively, where TP is true positive, TN is true negative, FP is false positive and FN is false negative, respectively. In practice, people usually use the False Positive Rate (FPR) which is defined as $1 - SPC$, instead of SPC, when drawing the ROC. The method of the largest true positive rate and smallest false positive rate is preferred.

We use the MATLAB command $[X,Y] = 	ext{perfcurve}(\text{LABELS},\text{SCORES}, \text{POSCLASS})$ to plot the ROC. Here “LABELS” is the true classification information of the voxels, and it is a binary vector where 0 represents not being the discriminative voxels and 1 represents being the discriminative voxels, respectively. SCORES here are the absolute value of the weight vector $\mathbf{w}$ of different multi-variate pattern recognition methods. For the two sample t-test, SCORES is $1 - p$-value. POSCLASS is 1 for our synthetic data.

As for the real data, we will see that the voxels selected by our method share much in common with those by the univariate two sample t-test and these voxels selected by the two methods are in the appropriate areas of the brain. However, our algorithm selects a combination of voxels that are generally distributed in wider brain areas than those selected by the two sample t-test, which makes sense because our method can also find out the multivariate patterns beyond the univariate patterns. While we do not have the ground truth, we would like to claim that these extra regions selected by our method are unlikely to be false positives. One reason is that number of selected voxels are determined by the two sample t-test, where the FDR (false discovery rate) control level is set to be 0.05, i.e. no more than 5% of your active voxels are false positives, and we have show that our method has a better false positive control than two sample t-test via the synthetic data.

3.4. Synthetic Data

We simulated simple case control analysis model and work on one slice brain image with a $(100 \times 100)$ grid 10000 voxels as the whole features. We generated 50 observations for each group, i.e. the control group and the case(patient) group.

There were five discriminated clustered features each of which contained $10 \times 10 = 100$ voxels, as showed in the first subplot of Figure 2 in white. The elements in first two clustered features: $x_{i,n}^{(k)} = k + \epsilon_{i,n}^{(k)}$ (case), $y_{j,n}^{(k)} = \eta_{j,n}^{(k)}$ (control), where $i, j = 1, 2, \ldots, 50$ representing the index of persons of each group, and $k = 1, 2$ representing the index of the first two clustered features, and $n = 1, 2, \ldots, 100$ representing the index of features of each cluster. $\epsilon_{i,n}^{(k)}$ and $\eta_{j,n}^{(k)}$ are Gaussian i.i.d distributed.
The elements in other three clustered features are spatially distributed patterns, which are Gaussian i.i.d distributed and constrained by linear condition: 
\[ x^{(k)}_{ij} = \epsilon^{(k)}_{ij}, \quad x^{(k)}_{jn} = \eta^{(k)}_{jn}, \quad \sum_{k=3}^{5} x^{(k)}_{ij} > 1 \text{ (case), and } \sum_{k=3}^{5} y^{(k)}_{jn} < 1 \text{ (control)}, \]
where \( i, j = 1, 2, \ldots, 50 \) representing the index of persons of each group, and \( k = 3, 4, 5 \) representing the index of the last three clustered features, and \( n = 1, 2, \ldots, 100 \) representing the index of features of each cluster. As above, \( \epsilon^{(k)}_{ij} \) and \( \eta^{(k)}_{jn} \) are also Gaussian i.i.d distributed. The features were spatially clustered in \( 10 \times 10 \) square grids. We also simulated the other voxels in matrix \( X \) as Gaussian noise. Notice that these are distributed multivariate discriminative patterns, each of which consists of 3 voxels from each of the last 3 clusters, respectively.

Figure 2: The left figure is the maps of estimated discriminative voxels by different methods on the synthetic data (unthreshold, i.e. the gray level is based on the absolute value of \( w \)): our method is the only one which can find out all the five discriminative regions accurately. Particularly, the randomized \( \ell_1 \) method which is based on the common stability selection takes more much resamplings than our method, and still achieves a less spatially contiguous results with missing the first discriminative block. The right figure is the ROC of different methods in terms of voxel selection accuracy. Our method can achieve almost perfect results with nearly no false positives and 100% true positive rate, much better than the alternatives. Notice that we did not plot the results of L1-SVM and L1-logistic because their results are too sparse.

Figure 2 is one typical result. From the left subfigure, we can see that the univariate method (two sample t-test) finds out the univariate discriminative voxels (the first 2 blocks) as expected, and fails to find out the rest 3 blocks, which represent the multivariate discriminative patterns, as showed in the 4-th subplot. As for the multi-variate pattern recognition methods, they may only find out part of all the discriminative regions. Specifically, as for \( \ell_2 \) logistic regression, it missed the first blocks. As for the \( \ell_1 \) logistic
regression and \( \ell_1 \)-SVM, they both return over-sparse solutions, which are hard to discriminate and interpret, as expected. The elastic net is able to find the contiguous discriminative regions via the adding of the \( \ell_2 \) norm, but it still can not find out all the discriminative regions, i.e. missing the first block. \( \ell_2 \)-SVM slightly disappoints me in this case, because it is only able to find the first 2 blocks and missing the rest 3 blocks, which are multi-voxel patterns. It shows that while SVM is a multi-variate classifier, it does not necessarily find out all the multi-variate patterns, which might not be very sensitive for SVM though they might be so for other multi-variate methods such as \( \ell_2 \) logistic regression. As we known from stability selection, the randomness can help remedy the problem of over-sparsity of the \( \ell_1 \) regularized methods such as \( \ell_1 \)-SVM or \( \ell_1 \)-logistic regression to some degree. However, the classical stability selection, can not return a satisfying contiguous results within 500 subsamplings, and especially the first block is missing. The “constrained block subsampling” is very important to get the comprehensive, reliable and clustered results as showed in the 9-th subplot which corresponds to our method. With 40 repetitions, our method can find out the discriminative regions with almost no either false positives or false negatives. In the right subfigure, we use ROC to further demonstrate the advantages of our method. Here the true positive rate and false positive rate in terms of voxel selection, rather than the classification or prediction accuracy, are showed. For this simple synthetic data, we can see our method achieve almost perfect result that with almost no false positives, a nearly 100% true positive rate. Even for more complicated cases as showed below in Figure 3, our method can still achieve a very good voxel selection performance.

Next we will consider a more complicated case where the true discriminative regions does not match the clustering (partitions) results well, for example, only part of a certain cluster belongs to the significant discriminative regions, and show that our method can still achieve very accurate results. For this purpose, we modified the way of generating synthetic data by adding highly correlated but non-discriminative voxels around the true features. We set five \( 20 \times 20 \) linear correlated clustered voxels in each 2d image. For each cluster, only 1/4 portion (10 \( \times \) 10) voxels in it have the discriminative power and the classification pattern of these features are same as those in previous synthetic data, as the upper left subplot of Figure 3.

See Figure 3, the upper left subfigure is the true discriminative regions. The upper right is the clustering results using k-means via the distance defined by correlation. Notice that the last 3 blocks, the true discriminative
regions are only the small portion (1/4) of the clusters generated by k-means. Thanks to the “constrained block subsampling”, our method can outline discriminative structures of a smaller size than the clusters based on only 40 subsamplings, as showed in the most right subplot. In addition, our method is easy to set a threshold value for voxel selection based on the $w$. We sort the absolute values of $w$ components as showed in the third subplot. The absolute values first gradually change then suddenly change fast and the value at the change point, surrounded by a red circle, is an appropriate value. This appearance of the inflexion point is not only because that our method is based on sparse regularization and therefore most of the components of the $w$ are forced to be close to 0, but also because that our method can achieve very high voxel selection accuracy for our synthetic data as showed in the last subfigure. Notice that the original stability selection (Randomized L1 logistic regression) does not have such an obvious threshold point and its has worse specificity and sensitivity than our method as showed in the last subfigure. Using the threshold value suggested by the changing point, we achieve a good support recovery as showed in the fifth subfigure. The same number of top weighted voxels of the randomized $\ell_1$ logistic regression are plotted in the fourth subfigure, which mostly misses the first discriminative block and introduces many false positives. Notice that our “constrained block subsampling” does not bring blurring to the final result, as also showed in the last subfigure.

3.5. Real fMRI Data

The synthetic data has demonstrated the advantages of our algorithm in terms of better sensitivity, i.e. being able to find out more or even all the discriminative regions keeping the control of false positives. We will further validate this new algorithm through the real fMRI data. We will show that our method can reliably find out more distributed discriminative brain regions than other alternatives, as showed by the above synthetic examples. While cross-validation method is suitable for settings the number of voxels for prediction or classification in fMRI data, we will using the following probability method in (Li et al., 2009) to set the number of selected voxels, which is good for the purpose of detection of localized regions in the brain that contain task or disease-relevant information.

We aim to identify the brain activation pattern of Chinese-chess problem-solving task in professional Chinese-chess grandmaster. Fourteen grandmaster and master-level Chinese-chess players were recruited and studied. 14
Figure 3: Result on the new Synthetic Data: Our method can outline discriminative structures of a smaller size ($10 \times 10$) than the cluster size ($20 \times 20$). In addition, as showed in the third subplot, an appropriate threshold value can be easily set by finding the inflexion point, which is marked by a red circle. Compared with the original stability selection, our method can achieve a better control of false positives and false negatives as showed in the last subfigure.
masters on Chinese chess were recruited and studied. All subjects were right-handed and with no history of psychiatric or neurological disorders. During the fMRI scanning, subjects were presented with two kinds of stimuli: a blank chessboard and patterns of Chinese chess spot game with checkmate problems. Each condition was presented for 20s, with a 2s-long break between. The block repeated nine times with different problems in each block. The break between each block is also 2s. There are 9 blocks in all. In consideration of the delay of Blood Oxygenation Level Dependent (BOLD) effect (Aguirre et al., 1998) and the condition that the master may solve the problem less than 20s, we just select the 4th-8th images of each state in each block for classification. That is, the number of observations of each subject for classification is 90, among which 45 are in blank states while the other 45 are in task states. We are using an averaged data among all the 14 grandmasters. Data Acquisition and Preprocessing Scanning was performed on a 3T Siemens Trio system at the MR Research Center of West China Hospital of Sichuan University, Chengdu, China. T2-weighted fMRI images were obtained via a gradient-echo echo-planar pulse sequence (TR, 2000ms; TE, 30ms; flip angle=90°; whole head; 30 axial slice, each 5mm sick (without gap); voxel size=3.75 × 3.75 × 5mm³). fMRI images were preprocessed using Statistical Parametric Mapping-8 (SPM8, Welcome Trust Centre for Neuroimaging, London, UK. [http://www.fil.ion.ucl.ac.uk/spm]. Spatial transformation, which included realignment and normalization, was performed using three-dimensional rigid body registration to for head motion. The realigned images were spatially normalized into a standard stereotaxic space at 2 × 2 × 2 mm³, using the Montreal Neurological Institute (MNI) echo-planar imaging (EPI) template. A spatial smoothing filter was employed for each brains three-dimensional volume by convolution with an isotropic Gaussian kernel (FWHM= 8 mm) to increase the MR signal-to-noise ratio. Then, for the fMRI time series of the task condition, a high-pass filter with a cut-off of 1/128 Hz was used to remove low-frequency noise. Among all 90 fMRI samples, each of them is of size 91 × 109 × 91.

Previous fMRI study of large-scale brain network of board game experts identified significantly active brain regions in domain-related task. The case-control analysis (game condition vs. blank board condition) results indicated Task-Evoked Activation mainly in dorsolateral prefrontal cortex and parietal cortex of cognitive network (Duan et al., 2012). Now we performed multivariate pattern analysis methods to identify discriminative brain functional activation regions in board game experts and compared the feature selection
power of them. For fair comparison, we try different combinations of setting number of selected voxels for all the involved methods. For two sample t-test method, the number is determined by setting the p-value $< 0.05$ in multiple hypothesis testing to correct for multiple comparisons, as commonly done. For the multivariate methods, we need to set the threshold value for controlling the false positives, where the voxels of corresponding weight is larger than the threshold value are selected.

As mentioned before, while the cross-validation method is widely used, it usually works well for the cases where there is a lot of training data available and prediction is the main purpose. For our purpose of localization of discriminative voxels, the use of probability based method is suggested in [Li et al., 2009]. We repeat the description of the probability method in [Li et al., 2009] as follows. Specifically, considering that the entries of $w$ are sparse, we assume that the probability distribution of the entries of $w$ is Laplacian. Using all entries of $w$ as samples, we estimate the mean, the variance, and the inverse cumulative distribution function $F^{-1}$ of this Laplacian distribution. We then define $R = i : |w_i| > \theta, i = 1, \ldots, p$, where $\theta$ is chosen as $F^{-1}(p_0)$, $p_0$ is a given probability (e.g., 0.975 or 0.99 in this paper).

We have plotted 4 figures for selected voxels of different methods, from Figure 4 to Figure 7. Due to the plain sparse models such as either $\ell_1$-SVM or $\ell_1$-Logistic Regression returning over-sparse solutions, which are hard to interpret and are not able to give contiguous discriminative regions, we are not going to plot the results of them. In all these figures, subfigure A is the results of t-test determined by setting the p-value $< 0.05$ in multiple hypothesis testing to correct for multiple comparisons. Subfigures B, C, D, and E are the results of multivariate methods of $\ell_2$-Logistic Regression, $\ell_2$-SVM, randomized $\ell_1$ logistic regression and our method, respectively, and the number of selected voxels may vary in different figures. In particular, in Figure 4 the number of selected voxels of different multivariate methods is the same as the t-test. In Figure 5 the number of selected voxels of different multivariate methods is set by the probability method, where the given probability is set to be 0.975, which is commonly used value (Li et al., 2009). In Figure 6 the number of selected voxels of different multivariate methods also is set by the probability method but in a more restrict way in terms of the given probability value being set to be a very tight value 0.99. In Figure 7 the number of selected voxels of all multivariate methods is the same as our method with the probability value being 0.99. We show all these results corresponding different probability values and selected voxel numbers.
to better demonstrate the advantages of our method in a more fair way. While the number of selected voxels may vary to certain degree in different figures, they have similar results and we can see the highest sensitivity of our method.

Specifically, all the multivariate pattern feature selection methods successfully identified at least partial task-related prefrontal and parietal lobe regions. These results briefly indicated co-working pattern of cognitive network and default mode network of human brain during board game task state. However, compared with most alternative algorithms besides the common stability selection, our proposed method were able to identify two more brain regions in medial prefrontal cortex and precuneus which are functional and structural central hub in default mode network. The common stability selection, i.e. randomized $\ell_1$ logistic regression is able to identify the medial prefrontal gyrus, but misses precuneus gyrus. Moreover, as showed in subfigure D, the common stability selection likely returns a slightly more scattered results, which does not match the second hypothesis about continuousness and compactness. Even worse, its scattered results bring much difficulty to tell the true positives and false positives. In addition, it required much more subsamplings, for example, 500 times here while our methods only takes 40 subsamplings. This result verifies one of the biggest advantages of our method, i.e. computational efficiency, which is especially important for extremely high dimensional problems, and better inference quality due to the incorporation of the prior information of the fMRI data. As mentioned before, this computational efficiency also comes from the even smaller size of the subproblem (5) due to the adoption of the averaging idea within a cluster.

4. Conclusion and Future work

Voxel selection is very important for both encoding and decoding of the fMRI data. In this paper we propose a simple and computationally efficient approach for data-driven voxel selection, which is also called support identification potentially for the following biomarker extraction (Orrù et al., 2012) in different literature. We propose to use the “constrained block subsampling” as a variant of the classical stability selection, applied to the existing sparse multi-variate classifiers such as $\ell_1$ logistic regression, in case of fMRI data, which has the strong correlations and distributed multivariate discriminative patterns. For better computational efficiency and inference quality, we can take data driven clustering (as done in this paper) or prior anatomical
Figure 4: Brain areas that discriminated between game condition vs. blank board condition using different methods (the number of selected voxels across different methods is the same as t-test). Our method (E) can find out more discriminative and interpretable structures. Randomized $\ell_1$ method (D) misses precuneus gyrus and the other alternative (A, B, and C) miss both the precuneus and medial prefrontal gyrus. In addition, our method requires much less subsamplings and achieves a more contiguous and comprehensive result than the randomized $\ell_1$ method.
Figure 5: Brain areas that discriminated between game condition vs. blank board condition using different methods (the number of selected voxels of multivariate methods is determined by setting the probability value to be 0.975). Our method (E) can find out more discriminative and interpretable structures. Randomized ℓ1 method (D) misses precuneus gyrus and the other alternative (A, B, and C) miss both the precuneus and medial prefrontal gyrus. In addition, our method requires much less subsamplings and achieves a more contiguous and comprehensive result than the randomized ℓ1 method.
Figure 6: Brain areas that discriminated between game condition vs. blank board condition using different methods (the number of selected voxels of multivariate methods is determined by set the probability value to be 0.99). Our method (E) can find out more discriminative and interpretable structures. Randomized ℓ₁ method (D) misses precuneus gyrus and the other alternative (A, B, and C) miss both the precuneus and medial prefrontal gyrus. In addition, our method requires much less subsamplings and achieves a more contiguous and comprehensive result than the randomized ℓ₁ method.
Figure 7: Brain areas that discriminated between game condition vs. blank board condition using different methods (the number of selected voxels of multivariate methods is the same as our method with the probability value being 0.99). Our method (E) can find out more discriminative and interpretable structures. Randomized \ell_1 method (D) misses precuneus gyrus and the other alternative (A, B, and C) miss both the precuneus and medial prefrontal gyrus. In addition, our method requires much less subsamplings and achieves a more contiguous and comprehensive result than the randomized \ell_1 method.
information into consideration and this is where the “constrained” comes from. Numerical experiments demonstrate that our method can achieve much better both specificity and sensitivity than other alternatives in terms of voxel selection accuracy rather than prediction or classification accuracy. The higher sensitivity brought by our method is of quite important for medical or cognitive analysis. In the future, we will further test the stability and reliability of our algorithm in terms of the more challenging multi-center data (Stöcker et al., 2005; Binder et al., 2011).

Acknowledgment

This work was supported by the Natural Science Foundation of China, Grant Nos. 61035006, 61125304, 11201054, 61273361, 91330201 and by the Specialized Research Fund for the Doctoral Program of Higher Education of China 20120185110028, and by the Fundamental Research Funds for the Central Universities ZYGX2012J118, ZYGX2013Z005 and by the Key Technology R&D Program of Sichuan Province 2012SZ0159.

References

Aguirre, G., Zarahn, E., D’esposito, M., 1998. The variability of human, bold hemodynamic responses. Neuroimage 8, 360–369.

Anderson, M.L., Oates, T., 2010. A critique of multi-voxel pattern analysis, in: Proceedings of the 32nd Annual Conference of the Cognitive Science Society.

Bach, F., Jenatton, R., Mairal, J., Obozinski, G., 2012a. Optimization with sparsity-inducing penalties. Found. Trends Mach. Learn. 4, 1–106. doi:10.1561/2200000015.

Bach, F., Jenatton, R., Mairal, J., Obozinski, G., 2012b. Structured sparsity through convex optimization. Statist. Sci. 27, 450C468.

Baldassarre, L., Mourao-Miranda, J., Pontil, M., 2012. Structured sparsity models for brain decoding from fmri data, in: Pattern Recognition in NeuroImaging (PRNI), 2012 International Workshop on, pp. 5–8.

Batmanghelich, N., Taskar, B., Davatzikos, C., 2012. Generative-discriminative basis learning for medical imaging. Medical Imaging, IEEE Transactions on 31, 51–69. doi:10.1109/TMI.2011.2162961.
Beinrucker, A., Dogan, u., Blanchard, G., 2012. A simple extension of stability feature selection, in: Pinz, A., Pock, T., Bischof, H., Leberl, F. (Eds.), Pattern Recognition. Springer Berlin Heidelberg. volume 7476 of Lecture Notes in Computer Science, pp. 256–265.

Binder, J.R., Gross, W.L., Allendorfer, J.B., Bonilha, L., Chapin, J., Edwards, J.C., Grabowski, T.J., Langfitt, J.T., Loring, D.W., Lowe, M.J., Koenig, K., Morgan, P.S., Ojemann, J.G., Rorden, C., Szaflarski, J.P., Tivarus, M.E., Weaver, K.E., 2011. Mapping anterior temporal lobe language areas with fMRI: A multicenter normative study. NeuroImage 54, 1465 – 1475. URL: http://www.sciencedirect.com/science/article/pii/S1053811910012486, doi: http://dx.doi.org/10.1016/j.neuroimage.2010.09.048.

Blum, A.L., Langley, P., 1997. Selection of relevant features and examples in machine learning. Artificial Intelligence 97, 245 – 271. URL: http://www.sciencedirect.com/science/article/pii/S0004370297000635, doi: http://dx.doi.org/10.1016/S0004-3702(97)00063-5.

Bühlmann, P., Bühlmann, P., Rütimann, P., Kalisch, M., 2011. Controlling false positive selections in high-dimensional regression and causal inference. Stat Methods Med Res online first. URL: http://dx.doi.org/10.1177/0962280211428371.

Bühlmann, P., Van De Geer, S., 2011. Statistics for high-dimensional data: methods, theory and applications. Springer.

Cao, H., Duan, J., Lin, D., Shugart, Y.Y., Calhoun, V., Wang, Y.P., 2014. Sparse representation-based biomarker selection for schizophrenia with integrated analysis of fMRI and {SNPs}. NeuroImage , –.

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. URL: http://www.sciencedirect.com/science/article/pii/S1053811908009415, doi: http://dx.doi.org/10.1016/j.neuroimage.2008.08.020.

Chandrasekaran, V., Jordan, M.I., 2013. Computational and statistical tradeoffs via convex relaxation. Proceedings of the National Academy of Sciences 110, E1181–E1190. doi: 10.1073/pnas.1302293110.
Charalampous, K., Kostavelis, I., Amanatiadis, A., Gasteratos, A., 2012. Sparse deep-learning algorithm for recognition and categorisation. Electronics Letters 48, 1265–1266.

Chu, C., Hsu, A.L., Chou, K.H., Bandettini, P., Lin, C., 2012. Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images. NeuroImage 60, 59–70. URL: http://dx.doi.org/10.1016/j.neuroimage.2011.11.066, doi:10.1016/j.neuroimage.2011.11.066.

Cover, T., 1965. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. IEEE transactions on electronic computers 14, 326–334.

Donoho, D., Jin, J., 2008. Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. Proceedings of the National Academy of Sciences 105, 14790–14795. doi:10.1073/pnas.0807471105.

Duan, X., Liao, W., Liang, D., Qiu, L., Gao, Q., Liu, C., Gong, Q., Chen, H., 2012. Large-scale brain networks in board game experts: Insights from a domain-related task and task-free resting state. PLoS ONE 7.

Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J., 2008. LIBLINEAR: A library for large linear classification. Journal of Machine Learning Research 9, 1871–1874.

Gaël Varoquaux, Alexandre Gramfort, B.T., 2012. Small-sample brain mapping: sparse recovery on spatially correlated designs with randomization and clustering, in: ICML.

Geman, S., Bienenstock, E., Doursat, R., 1992. Neural networks and the bias/variance dilemma. Neural Comput. 4, 1–58. URL: http://dx.doi.org/10.1162/neco.1992.4.1.1, doi:10.1162/neco.1992.4.1.1.

Gramfort, A., Varoquaux, G., Thirion, B., 2012. Beyond brain reading: Randomized sparsity and clustering to simultaneously predict and identify, in: Langs, G., Rish, I., Grosse-Wentrup, M., Murphy, B. (Eds.), Machine Learning and Interpretation in Neuroimaging. Springer Berlin Heidelberg. volume 7263 of Lecture Notes in Computer Science, pp. 9–16.
Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. Journal of Machine Learning Research 3, 1157–1182.

Hartigan, J.A., Wong, M.A., 1979. A k-means clustering algorithm. Applied Statistics 28, 100–108.

Hastie, T., Tibshirani, R., Friedman, J., 2009. Ensemble learning, in: The Elements of Statistical Learning. Springer New York. Springer Series in Statistics, pp. 605–624.

Haxby, J.V., Connolly, A.C., Guntupalli, J.S., 2014. Decoding neural representational spaces using multivariate pattern analysis. Annual Review of Neuroscience 37, 435–456.

Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P., 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293, 2425–2430.

Huang, J., Zhang, T., Metaxas, D., 2009. Learning with structured sparsity, in: Proceedings of the 26th Annual International Conference on Machine Learning, ACM, New York, NY, USA. pp. 417–424. URL: http://doi.acm.org/10.1145/1553374.1553429, doi:10.1145/1553374.1553429.

Jacob, L., Obozinski, G., Vert, J.P., 2009. Group lasso with overlap and graph lasso, in: Proceedings of the 26th Annual International Conference on Machine Learning, ACM, New York, NY, USA. pp. 433–440.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning with Applications in R. volume 103 of Springer Texts in Statistics. Springer.

Jenatton, R., Gramfort, A., Michel, V., Obozinski, G., Eger, E., Bach, F., Thirion, B., 2012. Multiscale mining of fmri data with hierarchical structured sparsity. SIAM J. Imaging Sciences , 835–856.

Lahiri, S., 2001. Effects of block lengths on the validity of block resampling methods. Probability Theory and Related Fields 121, 73–97. URL: http://dx.doi.org/10.1007/PL00008798, doi:10.1007/PL00008798.
Lahiri, S.N., 1999. Theoretical comparisons of block bootstrap methods. Ann. Statist. 27, 386–404.

Li, Y., Long, J., He, L., Lu, H., Gu, Z., et al, 2012. A sparse representation-based algorithm for pattern localization in brain imaging data analysis. PLoS ONE 7.

Li, Y., Namburi, P., Yu, Z., Guan, C., Feng, J., Gu, Z., 2009. Voxel selection in fMRI data analysis based on sparse representation. Biomedical Engineering, IEEE Transactions on 56, 2439–2451. doi:10.1109/TBME.2009.2025856

Li, Z., Liu, J., Yang, Y., Zhou, X., Lu, H., 2013. Clustering-guided sparse structural learning for unsupervised feature selection. IEEE Transactions on Knowledge and Data Engineering 99, 1. doi:http://doi.ieeecomputersociety.org/10.1109/TKDE.2013.65

Liu, J., Ji, S., Ye, J., 2009a. Multi-task feature learning via efficient $\ell_{2,1}$-norm minimization, in: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, AUAI Press, Arlington, Virginia, United States. pp. 339–348.

Liu, J., Ji, S., Ye, J., 2009b. SLEP: Sparse Learning with Efficient Projections. Arizona State University. URL: http://www.public.asu.edu/~jye02/Software/SLEP

Liu, J., Ye, J., 2010. Moreau-yosida regularization for grouped tree structure learning, in: Lafferty, J., Williams, C., Shawe-taylor, J., Zemel, R., Culotta, A. (Eds.), Advances in Neural Information Processing Systems 23, pp. 1459–1467.

Mairal, J., Yu, B., 2013. Supervised feature selection in graphs with path coding penalties and network flows. J. Mach. Learn. Res. 14, 2449–2485. URL: http://dl.acm.org/citation.cfm?id=2567709.2567740

Meinshausen, N., Bhlmann, P., 2010. Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72, 417–473. URL: http://dx.doi.org/10.1111/j.1467-9868.2010.00740.x, doi:10.1111/j.1467-9868.2010.00740.x

36
Michel, V., Gramfort, A., Varoquaux, G., Eger, E., Keribin, C., Thirion, B., 2012. A supervised clustering approach for fMRI-based inference of brain states. Pattern Recognition 45, 2041–2049.

Michel, V., Gramfort, A., Varoquaux, G., Eger, E., Thirion, B., 2011. Total variation regularization for fMRI-based prediction of behavior. Medical Imaging, IEEE Transactions on 30, 1328–1340. doi:10.1109/TMI.2011.2113378

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

Murtagh, F., 1985. A survey of algorithms for contiguity-constrained clustering and related problems. The Computer Journal 28, 82–88. doi:10.1093/comjnl/28.1.82

Ng, B., Abugharbieh, R., 2011. Generalized group sparse classifiers with application in fMRI brain decoding, in: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, Washington, DC, USA. pp. 1065–1071.

Orrù, G., Pettersson-Yeo, W., Marquand, A.F., Sartori, G., Mechelli, A., 2012. Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: A critical review. Neuroscience & Biobehavioral Reviews 36, 1140–1152. URL: http://www.sciencedirect.com/science/article/pii/S0149763412000139, doi:http://dx.doi.org/10.1016/j.neubiorev.2012.01.004

Park, M.Y., Hastie, T., Tibshirani, R., 2007. Averaged gene expressions for regression. Biostatistics 8, 212–227. doi:10.1093/biostatistics/kx1002

Pekar, J., 2006. A brief introduction to functional MRI. Engineering in Medicine and Biology Magazine, IEEE 25, 24–26. doi:10.1109/MEMB.2006.1607665

Poldrack, R.A., 2006. Can cognitive processes be inferred from neuroimaging data? Trends in cognitive sciences 10, 59.

Raizada, R.D.S., Kriegeskorte, N., 2010. Pattern-information fMRI: New questions which it opens up and challenges which face it. Int. J. Imaging

37
Rasmussen, P.M., Hansen, L.K., Madsen, K.H., Churchill, N.W., Strother, S.C., 2012. Model sparsity and brain pattern interpretation of classification models in neuroimaging. Pattern Recognition 45, 2085 – 2100.

Ryali, S., Chen, T., Supekar, K., Menon, V., 2012a. Estimation of functional connectivity in fmri data using stability selection-based sparse partial correlation with elastic net penalty. NeuroImage 59, 3852–3861.

Ryali, S., Chen, T., Supekar, K., Menon, V., 2012b. Estimation of functional connectivity in fmri data using stability selection-based sparse partial correlation with elastic net penalty. NeuroImage 59, 3852 – 3861. URL: http://www.sciencedirect.com/science/article/pii/S105381191101336X, doi:http://dx.doi.org/10.1016/j.neuroimage.2011.11.054.

Ryali, S., Supekar, K., Abrams, D.A., Menon, V., 2010. Sparse logistic regression for whole-brain classification of fmri data. NeuroImage 51, 752 – 764. URL: http://www.sciencedirect.com/science/article/pii/S1053811910002089, doi:http://dx.doi.org/10.1016/j.neuroimage.2010.02.040.

Särndal, C.E.e.a., 2003. Model Assisted Survey Sampling. Springer. chapter Stratified Sampling. pp. 100–109.

Shah, R.D., Samworth, R.J., 2013. Variable selection with error control: another look at stability selection. J. R. Statist. Soc. B 75, 55–80.

Stöcker, T., Schneider, F., Klein, M., Habel, U., Kellermann, T., Zilles, K., Shah, N.J., 2005. Automated quality assurance routines for fmri data applied to a multicenter study. Human Brain Mapping 25, 237–246. URL: http://dx.doi.org/10.1002/hbm.20096, doi:10.1002/hbm.20096.

Tononi, G., Sporns, O., Edelman, G.M., 1994. A measure for brain complexity: relating functional segregation and integration in the nervous system. Proceedings of the National Academy of Sciences 91, 5033–5037.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., 2002. Automated anatomical labeling of activations in {SPM} using a macroscopic anatomical parcellation
of the {MNI} {MRI} single-subject brain. NeuroImage 15, 273 – 289. URL: http://www.sciencedirect.com/science/article/pii/S1053811901909784, doi: http://dx.doi.org/10.1006/nimg.2001.0978.

de Vries, P., 1986. Stratified random sampling, in: Sampling Theory for Forest Inventory. Springer Berlin Heidelberg, pp. 31–55.

Xiang, S., Shen, X., Ye, J., 2012. Efficient sparse group feature selection via nonconvex optimization. CoRR abs/1205.5075.

Xiang, Z., Xi, Y., Hasson, U., Ramadge, P., 2009. Boosting with Spatial Regularization, in: Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C.K.I., Culotta, A. (Eds.), Advances in Neural Information Processing Systems 22, pp. 2107–2115.

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

Ye, J., Farnum, M., Yang, E., Verbeeck, R., Lobanov, V., Raghavan, N., Novak, G., DiBernardo, A., Narayan, V., for the Alzheimer’s Disease Neuroimaging Initiative, 2012. Sparse learning and stability selection for predicting MCI to AD conversion using baseline ADNI data. BMC Neurology 12, 46.

Yu, L., Ding, C., Loscalzo, S., 2008. Stable feature selection via dense feature groups, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA. pp. 803–811. URL: http://doi.acm.org/10.1145/1401890.1401986, doi:10.1145/1401890.1401986.

Yu, L., Sun, H., Barbot, J., Zheng, G., 2012. Bayesian compressive sensing for cluster structured sparse signals. Signal Processing 92, 259 – 269. URL: http://www.sciencedirect.com/science/article/pii/S0165168411002490, doi: http://dx.doi.org/10.1016/j.sigpro.2011.07.015.

Yuan, L., Liu, J., Ye, J., 2013. Efficient methods for overlapping group lasso. Pattern Analysis and Machine Intelligence, IEEE Transactions on PP, 1–1.
Zalesky, A., Fornito, A., Bullmore, E.T., 2010. Network-based statistic: Identifying differences in brain networks. NeuroImage 53, 1197 – 1207. URL: http://www.sciencedirect.com/science/article/pii/S1053811910008852, doi:http://dx.doi.org/10.1016/j.neuroimage.2010.06.041

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67, 301–320.