Review

Sparse representation based biomarker selection for schizophrenia with integrated analysis of fMRI and SNPs

Hongbao Cao [a], Junbo Duan [b,c], Dongdong Lin [b], Yin Yao Shugart [a], Vince Calhoun [d,e], Yu-Ping Wang [b,c,*]

[a] Unit on Statistical Genomics, Intramural Program of Research, National Institute of Mental Health, NIH, Bethesda, 20852, USA
[b] Department of Biomedical Engineering, Tulane University, New Orleans, LA, USA
[c] Department of Biostatistics & Bioinformatics, Tulane University, New Orleans, LA, USA
[d] The Mind Research Network, Albuquerque, NM, USA
[e] Department of Electrical and Computer Engineering at the University of New Mexico, Albuquerque, NM, USA

ARTICLE INFO

ABSTRACT

Integrative analysis of multiple data types can take advantage of their complementary information and therefore may provide higher power to identify potential biomarkers that would be missed using individual data analysis. Due to different natures of diverse data modality, data integration is challenging. Here we address the data integration problem by developing a generalized sparse model (GSM) using weighting factors to integrate multi-modality data for biomarker selection. As an example, we applied the GSM model to a joint analysis of two types of schizophrenia data sets: 759,075 SNPs and 153,594 functional magnetic resonance imaging (fMRI) voxels in 208 subjects (92 cases/116 controls). To solve this small-sample–large-variable problem, we developed a novel sparse representation based variable selection (SRVS) algorithm, with the primary aim to identify biomarkers associated with schizophrenia. To validate the effectiveness of the selected variables, we performed multivariate classification followed by a ten-fold cross validation. We compared our proposed SRVS algorithm with an earlier sparse model based variable selection algorithm for integrated analysis. In addition, we compared with the traditional statistics method for uni-variant data analysis (Chi-squared test for SNP data and ANOVA for fMRI data). Results showed that our proposed SRVS method can identify novel biomarkers that show stronger capability in distinguishing schizophrenia patients from healthy controls. Moreover, better classification ratios were achieved using biomarkers from both types of data, suggesting the importance of integrative analysis.

© 2014 Elsevier Inc. All rights reserved.

Contents

* Corresponding author at: Department of Biomedical Engineering, Tulane University, New Orleans, LA, USA.
  E-mail address: wyp@tulane.edu (Y.-P. Wang).

Introduction

Schizophrenia has been hypothesized to arise from a number of genetic factors and environmental effects. To date, many studies have investigated the role of critical genes or single nucleotide polymorphisms (SNP) associated with schizophrenia. Many genes of great significance have been identified as potential causal genetic makers for schizophrenia, such as G72/G30 on chromosomes 13q, DISC1, GRIK3, EFNA5, AKAP5 and CACNG2 (Badner and Gershon, 2002; Callicott et al., 2005; Sutrala et al., 2008). Besides genetic studies, functional magnetic resonance imaging (fMRI) is another widely used tool for the study of schizophrenia in that it has the ability to identify both structural and functional abnormalities in the brain regions of schizophrenia patients (Meda et al., 2008; Szycik et al., 2009). Therefore, the identification of biomarkers from joint analysis of fMRI and SNP data is of tremendous importance for disease diagnosis and treatment (Lin et al., 2011; Liu et al., 2009).

In this paper we proposed a generalized sparse model (GSM) to integrate multi-modality data (e.g., SNP and fMRI data) for biomarker selection. Sparse representation, particularly compressive sensing, received a great attention in recent years (Cao et al., 2012a, 2012b; Donoho and Elad, 2003; Gribonval and Nielsen, 2003; Kidron et al., 2007; Tang et al., 2011; Tropp et al., 2004). For example, Kidron et al. used sparse regression for cross-modal localizations of sound-related region in the video (Kidron et al., 2007). We recently developed the sparse representation-based classification algorithms for sub-typing of leukemia from gene expression data (Tang et al., 2011), for chromosome image segmentation (Cao et al., 2012a) and for integrative analysis of gene copy number variation and gene expression data (Cao et al., 2012b).

Traditionally, the GSM can be solved by many existing algorithms, such as Homotopy method (Donoho and Tsaig, 2008), orthogonal matching pursuit (OMP) algorithm (Cai and Wang, 2011; Davis et al., 1997; Tropp, 2004), single best replacement (SBR) algorithm (Soussen et al., 2011), and FOCUSS method (Cotter et al., 2005). However, in compressive sensing theory, the exact signal recovery of an s-sparse signal typically requires a large number of samples (Davenport et al., 2011). Here the s-sparse signal refers to a vector having at most s number of nonzero entries. Those entries with high amplitudes correspond to the variables to be selected (Cai and Wang, 2011). When the number of measurements $n \gg m$, (m is the number of samples), it will be difficult for the exact signal recovery (Davenport et al., 2011; Hsu et al., 2009). One most often used conditions for exact signal recovery is the restricted isometry property (RIP) (Davenport et al., 2011). However, whether a measurement matrix satisfies the RIP condition is hard to verify in practice. Another method is using the coherence of a matrix X (Candès and Tao, 2006; Donoho, 2004), which is often required to be small (e.g. $O(\frac{1}{s})$). Moreover, when a matrix X satisfies the signal recovery condition, the number of signals to be recovered or variables to be selected using those traditional sparse representation methods will generally be equal to or less than the number of samples (Li et al., 2009). To address this problem, Li et al. proposed a sparse representation based variable selection method, aiming to achieve a sparse solution for the GSM when the sample number is large (e.g., larger than the number of variables to be selected; Li et al., 2009).

In this work and in many other practical cases, the number of samples (92 cases/116 controls) is far less than the number of variables (i.e., 759,075 SNPs and 153,594 fMRI voxels). As a consequence, the small coherence condition of the data matrix is hard to be satisfied (Hsu et al., 2009), and thus directly using existing compressive sensing methods may fail. To overcome the difficulty caused by this large-n–small-m problem, we proposed a novel sparse representation based variable selection (SRVS) algorithm, which can select the significant variables regardless of the coherence condition of the measurement matrix. Moreover, the proposed SRVS algorithm has been proven to have multi-resolution properties that select variables at different significance levels. Instead of solving the GSM directly, the proposed SRVS algorithm solves sub-matrixes based L_p norm minimization problems and generates a sparse solution for the GSM. In a preliminary work (Cao et al., 2012c), we studied the orthogonal matching pursuit (OMP) based SRVS algorithm. Our preliminary results documented that, even for small number of samples, the SRVS is capable of identifying a number of biomarkers for schizophrenia, leading to improved identification accuracy.

Here, we extend the work by applying our proposed SRVS algorithm to GSM with a more general penalization term ($L_p(0 \le p \le 1)$norms), aiming to identify more effective joint biomarkers for schizophrenia. Specifically, we tested and compared three models with $p = 0, 0.5, 1$. For the $L_p(0 \le p \le 1)$ based model, we proved that the proposed SRVS method can identify significant variables at different significant levels, and recover signals with large probability regardless of the coherence of the measurement matrix. We also showed the convergence and effectiveness of the proposed SRVS algorithm. After that, we applied the SRVS to the GSM model integrating 759,175 SNPs and 153,594 fMRI voxels in 208 subjects (92 cases and 116 controls) for the identification of biomarkers for schizophrenia. To test the predictive power of the biomarkers or variables selected, we used the selected variables to distinguish schizophrenia patients from healthy controls followed by a 10-fold cross-validation. We evaluated the three models with different penalization terms (i.e., L_p norm, $p = 0, 0.5$ and 1) and compared them with the biomarker selection approach proposed by Li et al. (2009) for integrated analysis. In addition, we compared our method with the traditional statistical methods for uni-type data analysis (i.e. Chi-squared test for SNP data and ANOVA for fMRI data).

Materials and methods

The proposed variable selection approach includes three steps, as shown in Fig. 1: 1.) Data combination. The GSM model is proposed to combine two types of data. 2.) Variable selection. The SRVS algorithm is proposed to solve the sparse linear system in GSM. 3.) Validation of the selected variables. We employed a multi-classification approach to test the effectiveness of the selected variables. We used cross validation to select the optimal parameters used in the GSM.

A sparse model for data combination

The sparse representation of a signal can be modeled as

$$y = X\delta + \epsilon, \tag{1}$$

where $y \in R^{m \times 1}$ is the observation vector; $X \in R^{m \times n}$ represents the measurement matrix; and $\epsilon \in R^{m \times 1}$ is the measurement error or
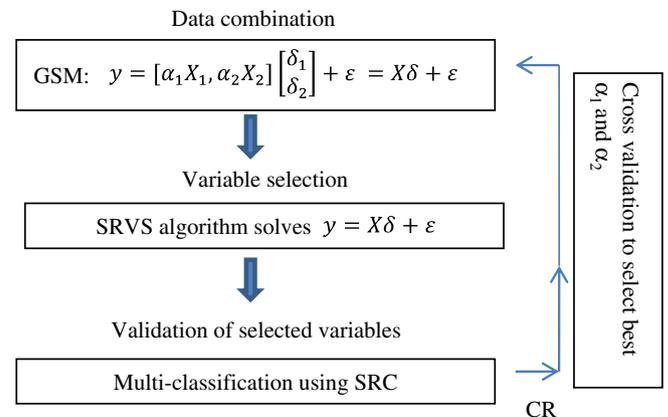
Fig. 1. The flowchart of our proposed variable selection for integrative analysis of two types of data.

noise. The goal of sparse representation is to recover the unknown sparse vector $\delta \in R^{n \times 1}$ from y and X, and the non-zero entries of δ correspond to selected variables/columns in X.

In case of representing multi-modality data (e.g. SNP data and fMRI data from same subject groups), we propose a generalized sparse model (GSM) in Eq. (2).

$$y = [\alpha_1 X_1, \alpha_2 X_2]\begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix} + \varepsilon = X\delta + \varepsilon, \tag{2}$$

where $y \in R^{m \times 1}$ is the observation vector (phenotypes of the subjects; e.g., 1 for disease case; 0 for healthy control); $X_1 \in R^{m \times n_1}$ and $X_2 \in R^{m \times n_2}$ are the measurements of two different data types (e.g., numerical SNPs values (0, 1, 2) and fMRI voxel values) with m samples, and n_1 (or n_2) features in each sample; each column is normalized to have unit L_2 norm; $X = [\alpha_1 X_1, \alpha_2 X_2] \in R^{m \times n}$; $\alpha_1 + \alpha_2 = 1$, and $\alpha_1, \alpha_2 > 0$ are the weight factors for the two types of data; $\varepsilon \in R^{m \times 1}$ is the measurement error. Then, the problem of variable selection becomes identifying the unknown sparse vector $\delta = \begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix} \in R^{n \times 1}$ from y and X, where $\delta_1 \in R^{n_1 \times 1}$, $\delta_2 \in R^{n_2 \times 1}$, and $n = n_1 + n_2$. To determine optimal weighting factors α_1 and α_2, cross validation can be used, i.e. the weighting factors that generates the best classification ratio (CR).

Variable selection with SRVS

By integrating multi-modality data, the GSM model given by Eq. (2) offers the potential to detect more significant and reliable biomarkers (Cao et al., 2012d; Liu et al., 2009; Rhodes and Chinnaiyan, 2005). However, in genomic and bio-imaging data analysis (e.g. SNP and fMRI), usually $n \gg m$ and the linear system defined by Eq. (2) is underdetermined, and the solution of the system is not unique. To overcome the problem, a sparse constraint is usually imposed on the model. An example is given in Eq. (3) by using the L_0 norm based penalty (Cai and Wang, 2011; Davis et al., 1997; Soussen et al., 2011; Tropp, 2004), which measures the number of nonzero elements.

$$\min \|\delta\|_0 \text{ subject to } \|y - X\delta\|_2 \leq \varepsilon. \tag{3}$$

But it is unfortunate that this penalty results in a combinatorial problem with NP-hard complexity. Thus, L_1 norm penalty (Donoho and Tsaig, 2008) is introduced instead:

$$\min \|\delta\|_1 \text{ subject to } \|y - X\delta\|_2 \leq \varepsilon. \tag{4}$$

Detailed discussions on the differences between L_0 and L_1 norm penalties can be found in Sharon et al. (2007) and Donoho et al. (2006). In recent years, L_p norm penalty ($0 < p < 1$) was also studied (Cotter et al., 2005; Xu et al., 2012), which can lead to a more sparse solution. The model is formulated as

$$\min \|\delta\|_p \text{ subject to } \|y - X\delta\|_2 \leq \varepsilon. \tag{5}$$

Several algorithms (Cotter et al., 2005; Foucart and Lai, 2009; Wang et al., 2011) have been proposed.

Nevertheless, when applying model (5) to signal recovery/variable selection, the measurement matrix $X \in R^{m \times n}$ is usually required to satisfy the RIP condition for exact signal recovery (Candès and Tao, 2006; Davenport et al., 2011; Donoho, 2004; Hsu et al., 2009). The RIP is defined as follows.

A matrix $X \in R^{m \times n}$ is said to satisfy the restricted isometry property (RIP) of order s if there exists a $\tau_s \in (0, 1)$ such that

$$(1 - \tau_s)\|\delta\|_2^2 \leq \|X\delta\|_2^2 \leq (1 + \tau_s)\|\delta\|_2^2 \tag{6}$$

holds for all s-sparse vector $\delta \in R^n$.

When a matrix $X \in R^{m \times n}$ satisfies the RIP of order s, the s-sparse vector $\delta \in R^N$ can be recovered from the m samples. One necessary condition for $X \in R^{m \times n}$ to satisfy RIP is that m and n satisfy Eq. (7) (Davenport et al., 2011).

Theorem. If $X \in R^{m \times n}$ satisfies an RIP of order $2s$ with $\varepsilon_{2s} \in (0, \frac{1}{2}]$, then

$$m \geq cs \log\left(\frac{n}{s}\right), \tag{7}$$

where $c = \frac{1}{2\log(\sqrt{24}+1)} \approx 0.28$ (Davenport et al., 2011). Thus for a data sample with the number of m, the following condition has to be satisfied

$$n \leq se^{m/cs} \tag{8}$$

for exact recovery of s-sparse vector $\delta \in R^N$. In genomic or medical imaging data analysis, we can assume $m \leq s$, i.e., the number of biomarkers to be detected is generally larger than the number of sample m (Li et al., 2004, 2009). Then Eq. (8) can be simplified as

$$n \leq me^{1/c} \approx 35m. \tag{9}$$

Eq. (9) suggests that, for a given sample size m, the number of n in the measurements matrix $X \in R^{m \times n}$ should be less than $35m$ so that the sparse solution can be obtained.

However, in practice, this condition cannot be satisfied, because the number of features (SNPs/fMRI voxels) is often greatly larger than that of the sample.

Eq. (9) describes a necessary condition for a matrix $X \in R^{m \times n}$ to satisfy signal recovery requirement. Since it's difficult to verify if a matrix $X \in R^{m \times n}$ satisfies the RIP, the coherence of X is used instead (Donoho, 2004). The coherence $\mu(X)$ is defined as:

$$\mu(X) = \max_{1 \leq i < j \leq n} \left| \left(X^T X\right)_{i,j} \right| / \sqrt{\left| \left(X^T X\right)_{i,i} \right| \left| \left(X^T X\right)_{j,j} \right|}. \tag{10}$$

The coherence given by Eq. (10) is always within the following range: $\mu(X) \in \left[\sqrt{\frac{n-m}{m(n-1)}}, 1\right]$; the lower bound is known as the Welch bound (Davenport et al., 2011). Some reconstruction algorithms require a strong condition of bounded coherence (Donoho, 2004; Hsu et al., 2009):

$$\mu(X) \leq O\left(\frac{1}{s}\right). \tag{11}$$

However, this small coherence condition is hard to be satisfied by our problem. In fact, when $n \gg m$, it is inevitable for some columns within the small-m–large-n measurement matrix to have big coherence (Candès and Tao, 2006). In our study, $n = 759{,}075 + 153{,}594$ (SNPs + fMRI voxels) and $m = 208$ (92 cases/116 controls). Therefore, we proposed the SRVS algorithm to find an approximate solution of the GSM model given by Eq. (2). The SRVS algorithm is described as follows, and its properties are presented in Appendix A.

SRVS Algorithm http://hongbaocao.weebly.com/software-for-download.html

1. Initialize $\delta^{(0)} = 0$;
2. For Step l, randomly choose k columns from $X = \{x_1, \ldots, x_n\} \in R^{m \times n}$ to construct a $m \times k$ sub-matrix denoted as $X_l \in R^{m \times k}$, and denote the index vector of the selected columns as $I_l \in \{1, 2, 3\ldots\}$;
3. Given the sub-matrix X_l, solve the following L_p minimization problem to get the optimal sparse solution $\delta_l \in R^{k \times 1}$

$$\min \|\delta_l\|_p \text{ subject to } \|y - X_l \delta_l\|_2 \leq \varepsilon; \tag{12}$$

4. Update $\delta^{(l)} \in R^{n \times 1}$ with δ_l: $\delta^{(l)}(I_l) = \delta^{(l-1)}(I_l) + \delta_l$; where $\delta^{(l)}(I_l)$ and $\delta^{(l-1)}(I_l)$ denote the I_l th entries in $\delta^{(l)}$ and $\delta^{(l-1)}$, respectively;

5. If a stopping rule is not satisfied, update $l = l + 1$ and go to Step 2. Otherwise, set $\delta = \delta^{(l)}/l$ and break. The non-zero entries in δ correspond to the column vectors selected, i.e., variable selection.

In Step 2, one way to achieve random selection of k columns from X is to shuffle the data with the Fisher–Yates algorithm (Fisher and Yates, 1948), and then use a window of length k to select variables randomly (Cao et al., 2012c). It should be noted that in each iteration, a different sub-matrix X_l will be randomly selected (totally $\binom{n}{k}$ possible combinations), which is not a simple split of the X into several sub-sets.

In Step 3, there are many well-established methods for solving the L_p minimization problem, such as the Homotopy algorithm (Donoho and Tsaig, 2008) for $p = 1$, orthogonal matching pursuit (OMP) algorithm (Cai and Wang, 2011; Davis et al., 1997; Tropp, 2004) and single best replacement (SBR) algorithm (Soussen et al., 2011) for $p = 0$ and the FOCUSS method for $0 \leq p \leq 1$ (Cotter et al., 2005).

In Step 5, we set the following two stop rules: 1.) $\|\delta^{(l)}/l - \delta^{(l-1)}/(l-1)\|_2 < \alpha$, where α is a predefined threshold; 2.) The probability that each pair of column vectors in X has been compared should be greater than $1-p_{stop}$. The algorithm terminates when both rules are satisfied, which decides the total number of iterations. In this work, we set $\alpha = 0.01$ and $p_{stop} = 1e^{-4}$. At those stop rules, the total number of iterations was around 200 for the simulation data with $n = 1e^{-6}$ features and 300 for the real data sets (759,075 SNPs and 153,594 fMRI voxels) tested in this work. The effect of stop rules on the number of iterations will be evaluated in *Sec. B* of Appendix A, where the convergence of the algorithm is also proved.

We present the discussion and proof of the properties of the proposed SRVS algorithm in Appendix A, including: 1.) The independence on coherence condition of the data matrix X; 2.) Convergence and effectiveness of SRVS; 3.) Multi-resolution property of SRVS; and 4.) Sparsity control using ε. The Matlab based software toolbox for the proposed SRVS algorithm is available online: http://hongbaocao.weebly.com/software-for-download.html.

Validation of selected variables

To test the detective power of the selected biomarkers (SNPs/fMRI voxels), we performed a multivariate classification followed by a ten-fold cross-validation to classify schizophrenia patients from health controls. Results from four models were compared: SRVS algorithm with different L_p norm penalties ($p = 0, 0.5, 1$) and Li et al.'s method (Li et al., 2009).

Furthermore, we compared several classifiers, including sparse representation-based classifier (SRC), fuzzy c-means (FCM) classifier, and support vector machine (SVM) based classifier, and the SRC gives the best performance (see Appendix B, Fig. B.1). The SRC has been proven effective for many tasks such as face recognition (Wright et al., 2009), speech recognition (Gemmeke et al., 2011), signal classification for brain computer interface (Shin et al., 2012), and image classification (Cao et al., 2012a). We provide the results in Appendix B, Fig. B.1. Here, we describe the SRC algorithm as follows.

Sparse Representation-based Classification (SRC) algorithm:

1. Inputs: a matrix of training samples $A = [A_1, A_2, ..., A_c] \in R^{n \times s}$ for c classes; and a test sample $s_t \in R^n$.
2. Normalize the columns of A to have unit L_2-norm;
3. Solve the L_1 norm minimization problem: $\hat{x} = \arg \min x_1$, subject to $Ax = s_j$;
4. Calculate the residuals $r_i(s_t) = s_t - A\delta_i(x)_2$ for $i = 1, ..., c$;
5. ClassID $(s_t) = \arg \min_i r_i(s_t)$.

The inputs of the SRC algorithm include 1. $s_t \in R^n$, the feature vector from the subject t; 2. $A \in R^{n \times s}$, feature vectors from $c = 2$ cluster groups in a total of s samples/subjects; $\delta_i(*)$ is a $R^s \to R^s$ transformation function, which selects the coefficients associated with the i-th class. The output is the ClassID of the subject t.

In each run of the 10-fold cross-validation, 90% subjects from both cases and controls were randomly selected for variable/biomarker selection, while the rest were used for testing. For each method, we carried out 100 runs and the average of the classification ratios was used as the final identification accuracy.

We also used the cross-validation to determine the optimal weighting factors in Eq. (2). For different pairs of weighting factors, different variable groups will be selected, resulting in different classification ratios. Therefore, using the cross validation stated above, we can select the best weighting factors that lead to the highest classification ratio.

To test the effectiveness of integrative analysis, we compared our method with two traditional statistical methods for uni-type data analysis (i.e. Chi-squared test for SNP data and ANOVA for fMRI data). We provide the top 200 selected SNPs and fMRI voxels and the classification ratios in Appendix C and Appendix D, respectively.

Results

This section is organized as follows. We first describe in the Data collection section the data we tested. Then we present in the Variable selection with generalized sparse model section the variables (SNPs/fMRI voxels) selected using GSM with different weighting factors. After that we compare in the Comparison of the variables selected using different methods section the variables selected with different models (SRVS with three different penalties, Li et al.'s method). Finally, in the Cross validation for the selection of weighting factor section, we provide the cross validation results for the selection of weighting factors.

Data c =ollection

In this study, participant recruitment and data collection were conducted by the Mind Clinical Imaging Consortium (MCIC). Two types of data (SNP and fMRI) were collected from 208 subjects including 96 schizophrenia patients (age: 34 ± 1, 22 females) and 112 healthy controls (age: 32 ± 1, 44 females). All of them provided written informed consents. Healthy participants were free of any medical, neurological or psychiatric illnesses and had no history of substance abuse. By the clinical interview of patients for DSM IV-TR Disorders (Kumari et al., 2012; Pascual-Leone et al., 2002) or the comprehensive assessment of symptoms and history, patients met criteria for DSM-IV-TR schizophrenia (Meier et al., 2008; Onitsuka et al., 2004). Antipsychotic history was collected as part of the psychiatric assessment.

fMRI data collecting and preprocessing

The fMRI data were collected during a sensorimotor task, a block-design motor response to auditory stimulation. During the on-block, 200 ms tones presented a 500 ms stimulus onset asynchrony (SOA). A total of 16 different tones were presented in each on-block, with frequency ranging from 236 Hz to 1318 Hz. The fMRI images were acquired on Siemens 3 T Trio Scanners and a 1.5 T Sonata with echo-planar imaging (EPI) sequences using the following parameters (TR = 2000 ms, TE = 30 ms (3.0 T)/40 ms (1.5 T), field of view = 22 cm, slice thickness = 4 mm, 1 mm skip, 27 slices, acquisition matrix = 64×64, flip angle = 90°.) Four scanners were used and we have roughly equal numbers of patients and controls at all sites. Data were pre-processed in SPM5 (http://www.fil.ion.ucl.ac.uk/spm) and were realigned, spatially normalized and re-sliced to $3 \times 3 \times 3$ mm^3, smoothed with a $10 \times 10 \times 10$ mm^3 Gaussian kernel to reduce spatial noise, and analyzed by multiple regression considering the stimulus and their temporal derivatives plus an intercept term as repressors. Finally the stimulus-on versus stimulus-off contrast images were extracted with $53 \times 63 \times 46$ voxels and all of the voxels with missing measurements were excluded.

SNPs data

A blood sample was obtained from each participant and DNA was extracted. Genotyping for all participants was performed at the Mind Research Network using the Illumina Infinium HumanOmni1-Quad assay covering 1,140,419 SNP loci. Bead Studio was used to make the final genotype calls. Next, the PLINK software package (http://pngu. mgh.harvard.edu/~purcell/plink) was used to perform a series of standard quality control procedures, resulting in the final dataset spanning 759,075 SNP loci. Each SNP was categorized into three clusters based on their genotype and was represented with discrete numbers: 0 for 'BB' (no minor allele), 1 for 'AB' (one minor allele) and 2 for 'AA' (two minor alleles).

Variable selection with generalized sparse model

Based on the generalized sparse model (Eq. (2)), we applied the proposed SRVS algorithm to select biomarkers for schizophrenia from the combination of two data sets (SNP data and fMRI data), where the weight factors α_1 and α_2 ($\alpha_1 + \alpha_2 = 1$) reflect the level of contribution from SNP and fMRI data set respectively. When the weight factor $\alpha_1 = 1$ or $\alpha_2 = 1$, the variable selection is performed only on one type of data. We tested the range of α_1 from 0.3 to 0.6, and used a step length of 0.02 with a total of 16 different trials and we set $k = 0.05n$. To test the most important biomarkers, we selected 200 biomarkers in each trial by using our proposed SRVS method in three models with different L_p norms ($p = 0, 0.5, 1$). We also compared with Li et al.'s method (Li et al., 2009). Fig. 2 shows the plot of the number of SNPs and fMRI voxels selected against weight factor α_1 for these four models. As shown in Fig. 2, the weight factor has similar effects on the variables selected with the four models. It was interesting to see that even though the number of SNPs was much larger than that of fMRI voxels (759,075 vs. 153,594), similar number of variables was selected from both data sets when weight factor α_1 was around 0.4 (0.38 for SRVS method with $L_{1/2}$ norms, 0.46 for SRVS method with L_0 norms, 0.47 for SRVS method with L_1 norms, and 0.47 for Li et al.'s method).

In addition, from Fig. 2 we can see that when α_1 took a smaller value, only a few SNPs were selected. Those SNPs can be viewed as the most important biomarkers since they were identified in both two data sets consisting of SNPs with small weight. When α_1 took a large value (α_2 was small), only a few fMRI voxels were selected. For the same reason, these voxels should be the most important ones. To further understand the relationships between the groups of variables selected in each trail, we analyzed the newly selected variables with the decrease of the corresponding weight factor, as shown in Fig. 3.

In Fig. 3, the newly selected variables shown in each trial have no overlap with variables from any other trials. When the weight factors have larger values (0.6 for SNP data set and 0.7 for fMRI data set), the selected groups have relatively larger size, and the variables were mostly from one type of data. Those were the variables that can be identified when using one type of data alone for the analysis. With the decrease of the weight factor, fewer new variables were detected. However, those variables should not be viewed as less significant than those selected with bigger weight factors, since they were selected over variables from both types of data with smaller weights.

a) SRVS method with $L_{1/2}$ norms

b) SRVS method with L_0 norms

c) SRVS method with L_1 norms

d) Li et al's SLR Method
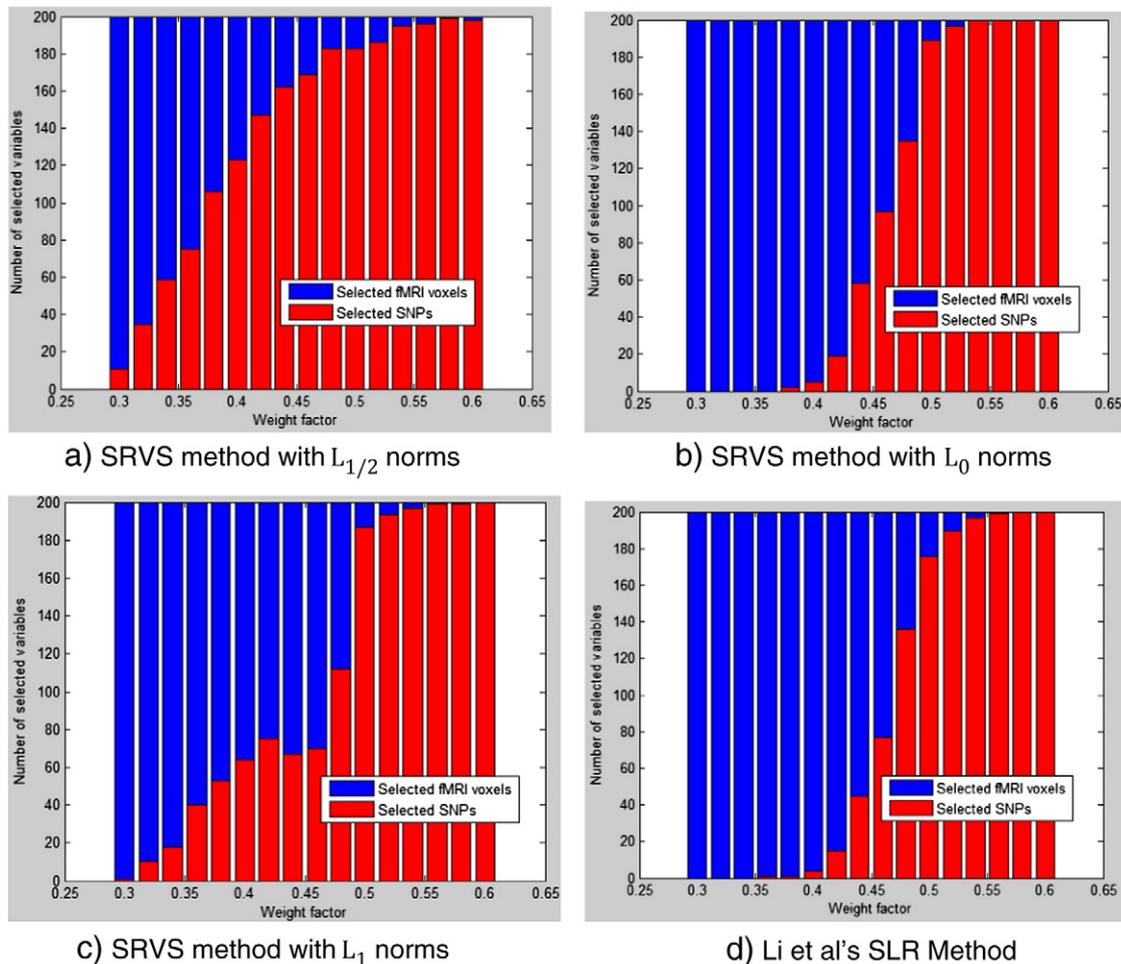
Fig. 2. Variable selection with generalized sparse model using different models, where the number of selected fMRI voxels is in red color and the number of selected SNPs is in blue color. The 'weight factor' in the plots refers to the weight factor α_1 (for SNP data set), and the weight factor $\alpha_2 = 1 - \alpha_1$ (for fMRI data set). (a) SRVS method with $L_{1/2}$ norms; (b) SRVS method with L_0 norms; (c) SRVS method with L_1 norms; (d) Li et. al's SLR method.
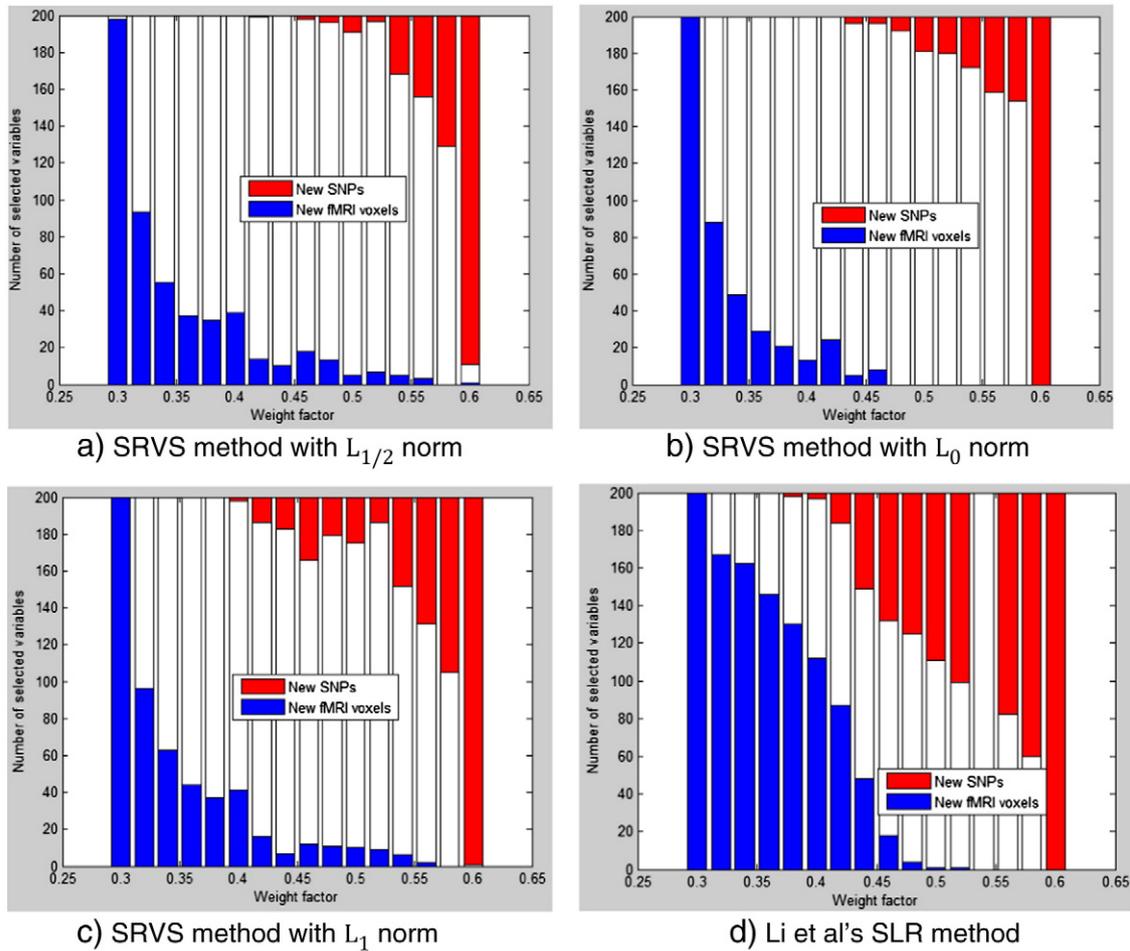
Fig. 3. The newly selected variables in each trial with the decrease of the corresponding weight factor. The 'weight factor' in the plots refers to the weight factor α_1, and the weight factor $\alpha_2 = 1 - \alpha_1$. (a) SRVS method with $L_{1/2}$ norms; (b) SRVS method with L_0 norms; (c) SRVS method with L_1 norms; (d) Li et al.'s SLR method.

Comparison of the variables selected using different methods

We further compared the selected variables (SNPs/fMRI voxels) using different methods: SRVS with (L_0, L_1, $L_{1/2}$) and Li et al.'s method, as shown in Table 1. For the 16 trials with 200 variables selected in each trial, there were totally 3200 variables. However, as shown in Fig. 3, only a few of new variables were selected in each run, resulting in less number of final selected variables (807, 888, 1092 and 1939 for the four models, respectively). The overlaps among the selected variable groups using SRVS models with different L_p norm penalties are around 50% (458, 447 and 514 as shown in Table 1 and Fig. 4). Totally 349 variables are selected by all those three models. When compared with Li et al.'s method, only a small percentage (<10%) overlapped with those

of the SRVS models with different L_p norms (67, 87 and 79 respectively). There were totally 48 variables selected by all the four models. We provided the first 50 SNPs and the corresponding genes identified by the four method in Table A.1.

We also compared our selected genes with the top ranked 45 schizophrenia genes reported in (http://www.szgene.org/default.asp) (see Table A.2). We selected 200 variables in each trial and identified 4 to 5 reported genes by using our proposed SRVS methods with L_0 and L_1

Table 1
The comparison of the variable numbers (SNPs/fMRI voxels) selected by the four Models: SRVS with (L_0, L_1, $L_{1/2}$) and Li et al.'s method.

	SRVS ($L_{1/2}$)	SRVS (L_0)	SRVS (L_1)	Li et al.'s method	Three SRVS methods
SRVS ($L_{1/2}$)	807	458	447	67	349
SRVS (L_0)	/	888	514	87	
SRVS (L_1)	/	/	1092	79	
Li et al's method	/	/	/	1939	/
All four methods	48				
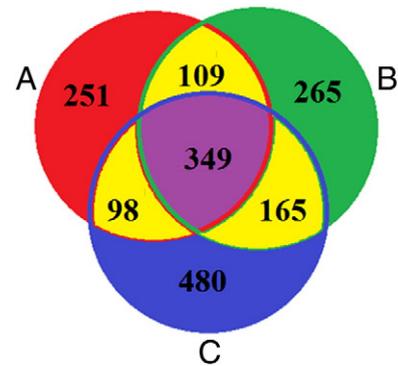
Fig. 4. Comparison of the selected variables (SNPs/fMRI voxels) using a Venn diagram. A, B and C are the variables selected using SRVS with $L_{1/2}$, L_0 and L_1 norm penalties, respectively.

norm penalties, and by using Li et al.'s method, as shown in Table 2. Our proposed SRVS algorithm with $L_{1/2}$ norm identified 6 reported genes. The genes/SNPs identified by each model were different. It should be noted that even though the OPCML gene was identified by all the four models, the corresponding SNPs, according to which the gene was identified, were different. If we select more variables in each run, corresponding to larger *s*-sparsity, more significant variables can be selected. This has been reported in our previous work, in which we selected around 800 variables in each of the 16 trials (α_1 is from 0.3 to 0.6; step length = 0.02), and 20 reported genes (e.g. PRSS16, NOTCH4, PDE4B, TCF4) (Cao et al., 2012c).

We also compared the fMRI voxels selected by our proposed SRVS method with those of using Li et al.'s method (Li et al., 2009), as shown in Fig. 5. It is evident that the voxels selected by the SRVS method tended to cluster together at specific regions such as temporal lobe, lateral frontal lobe, occipital lobe, and motor cortex, which are schizophrenia related brain regions (Kumari et al., 2012; Onitsuka et al., 2004; Pascual-Leone et al., 2002). However, the voxels selected by Li et al.'s method tend to be small regions scattered over the whole brain. This may be due to the fact that the voxels within the same brain region may not be simultaneously detected using their method.

The above comparisons show big differences among the biomarkers selected using the four models. To compare and the test the effectiveness of those different groups of biomarkers, they were used for the classification of schizophrenia patients from normal controls and the results were provided in the Cross validation for the selection of weighting factor section.

Cross validation for the selection of weighting factor

We used the SRC as a classifier in classification with a ten-fold cross validation to test the predictive power of the variables/biomarkers selected in the four models (SRVS with $L_{0.5}$ penalty; SRVS with L_0 penalty; SRVS with L_1 penalty; and Li et al.'s method). We also used cross-validation to select the best weighting factors for the GSM (i.e., the weighting factors corresponding to the highest classification accuracy). Fig. 6(a) shows the results of the ten-fold cross validation for the 16 trials given in Fig. 2. It can be seen from Fig. 6(b) that our proposed SRVS methods in all three models with different L_p norm penalties provide much higher classification ratios than that of Li et al.'s method (Li et al., 2009) (p - value < 1e^{-8}). In addition, the SRVS method with the $L_{1/2}$ norm penalty gives the highest classification ratio among four tested models. This is consistent with the fact that, the $L_{1/2}$ norm based sparse models provide the best data fitting and visual modeling among all the L_p norms ($p \in (0, 1]$) as demonstrated by Xu et al.'s work (Xu et al., 2012).

The cross validation results determined that, for the four models tested, the best weighting factors are: 1.) SRVS with $L_{1/2}$ penalty, $\alpha_1 = 0.58$, CR = 89.7%; 2.) SRVS with L_0 penalty, $\alpha_1 = 0.48$, CR = 81.5%; 3.) SRVS with L_1 penalty, $\alpha_1 = 0.38$, CR = 82.1%; and 4.) Li et al.'s method, $\alpha_1 = 0.46$, CR = 62.1%. It should be noted that the best CRs and the corresponding optimum weighting factors were achieved with biomarkers selected from both types of data.

When compared with the traditional statistics method for uni-type data analysis (Chi-squared test for SNP data and ANOVA for fMRI data), results showed that using selected top (200–1000) SNPs alone can reach identification accuracy of (83.11 ± 1.32)%, while using top (200–1000) fMRI voxels alone the accuracy was (63.13 ± 0.74)%. Please refer to Appendices C and D for the selected top features and the classification results.

Discussion and conclusion

This work aimed at biomarker identification using multi-modality data, i.e., SNP and fMRI data. To achieve this goal, we proposed generalized sparse model (GSM) solved by a novel SRVS algorithm. The selected biomarkers were then tested by applying to the classification of schizophrenia patients with a ten-fold cross validation.

The GSM given in Eq. (2) uses multi-modality data for integrative analysis to detect biomarkers that cannot be identified using one type of data alone. The two weighting factors in GSM represent the level of contribution from different data types, and the best values can be determined by cross-validation. As shown in the Cross validation for the selection of weighting factor section, the best weighting factors for all the four models are between [0.38, 0.58];using the combination of both data sets at these values can lead to highest classification ratios. This demonstrates the advantage of integrating multi-modality data for the diagnosis of schizophrenia. In addition, when compared to the uni-type data analysis, our proposed SRVS method with $L_{1/2}$ norm led to significantly higher identification accuracy (p-value < 0.001) than that from using one type of biomarkers alone (i.e., only using SNP or fMRI data for classification). This further demonstrates the advantage of using multiple data modalities.

The ten-fold validation results showed that features selected using our proposed SRVS algorithm gave higher classification accuracy than that of Li et al.'s method (see Fig. 6(b), p-value < 1e^{-8}), indicating its effectiveness when the sample numbers are much smaller than the number of variables. The comparison results indicate that, even though Li's method is valid for data of large sample size, our proposed SRVS is more suitable for processing data of small sample size.

For the data set of small-*m*-large-*n*, traditional sparse model may fail. By randomly sampling the columns of the original measurement matrix into smaller sub-matrixes, however, our proposed SRVS method can overcome the difficulty. We proved that, a significant variable from the original data set can be selected with high probability, regardless of the coherence conditions (Appendix A, Sec. A). Moreover, we showed the convergence and effectiveness of the proposed SRVS method in Appendix A, Sec. B. As shown in Fig. 6(a), voxels from the same clusters were identified simultaneously. Those variables may not be recovered with the traditional sparse models (e.g., the method used by Li et al. (2009)).

One advantage of our proposed SRVS algorithm is its multi-resolution properties. We showed in Appendix A, Sec. C that variables selected using larger window length *k* will be subset of those selected with smaller *k*. When *k* = *n*, the model reduces to traditional sparse model as given by Eq. (5), and no more than *m* biomarkers can be identified (Cai and Wang, 2011; Li et al., 2009). When using two different

Table 2
The comparison with the reported first 45 Schizophrenia genes (http://www.szgene.org/default.asp).

SRVS (L_0)		SRVS ($L_{1/2}$)		SRVS (L_1)		Li's method	
Genes	SNPs	Genes	SNPs	Genes	SNPs	Genes	SNPs
PDE4B	rs10846559	DRD2	rs10800893	HIST1H2BJ	rs11220916	PRSS16	rs13399561
NRG1	rs12097254	NRG1	rs16956192	DRD2	rs16828456	DAOA	rs16869700
PLXNA2	rs4811326	RGS4	rs1293448	NRG1	rs10846559	RPP21	rs1836942
OPCML	rs3026883	PPP3CC	rs6637088	PLXNA2	rs4632116	NRG1	rs10833482
/	/	PLXNA2	rs4072729	OPCML	rs11807403	OPCML	rs1745939
/	/	OPCML	rs11772714	/	/	/	/
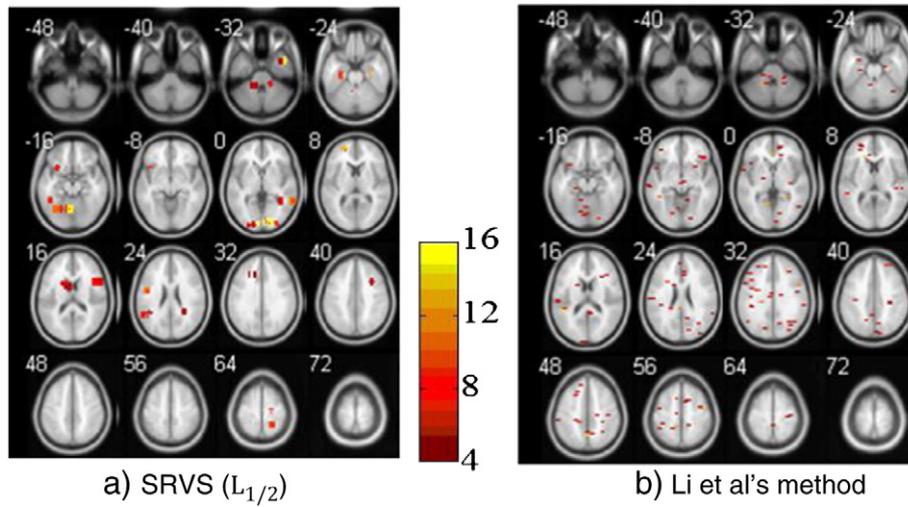
Fig. 5. A comparison of the selected fMRI voxels between SRVS ($L_{1/2}$) and Li et al.'s method (Li et al., 2009). The value of a voxel represents the frequency that it has been selected in the 16 trials. (a) SRVS ($L_{1/2}$); (b) Li et al's method.

parameters $k_1 > k_2$, a larger group of variables will be selected by using k_2, which include the variables selected using k_1. Thus, as long as the parameter $n > k > m$ (e.g., $k = 0.05n$), the same top (n/k) × m variables will always be selected. We provided the discussion of the relationship of window length k and the number of variables to be selected in Appendix A, *Sec. D*. In addition, those variables can be ranked in the order of their significance (i.e., the amplitudes of the corresponding entries in the solution δ; Cai and Wang, 2011; see Appendix A, Fig. A 3). Then the residual ε can be used to determine how many variables should be selected (i.e., sparsity control; see Appendix A, *Sec. E.*).

Our multivariate classification results show that the variables selected by using our proposed SRVS algorithm, especially with $L_{1/2}$ norm based penalty, generated highest classification accuracy in discriminating schizophrenia patients from healthy controls. This suggests that $L_{1/2}$ norm may be the best choice as penalization term for the proposed SRVS method. However, the multivariate classification approach is not an extra validation step, but rather a way to determine how predictive the selected variables are as selected by the SRVS approach. For those variables that are not reported by the previous studies, further validation approaches should be conducted to test their significance. Moreover, the SRVS method proposed in this work is a data driven method, which do not directly interpret the physiological meaning of the selected variables. It was suggested by Haufe et al.'s work that signals detected using general linear model may involve noise (Haufe et al., 2014). Therefore, those variables that are not reported by the previous studies may need further validation approaches using independent data sets to test their physiological significance.

In summary, we presented an effective multi-modality data integration method for biomarker selection. Using the method, we are able to integrate different types of data with large number of variables but small number of samples. However, due to the limited sample size, further biological experimental work is needed to validate the biomarkers identified in the paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.neuroimage.2014.01.021.
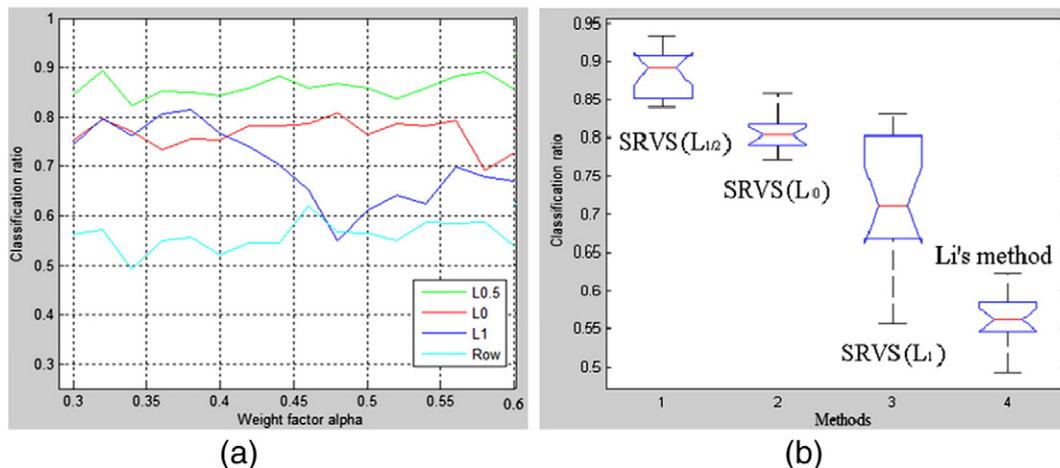
Fig. 6. A comparison of classification results of using four sparse models. Panel (a) gives the classification ratio of differentiating schizophrenia from healthy controls using four models with different weight factors; panel (b) is the box plot generated with ANOVA analysis of the classification ratios using four different models.

References

Badner, J.A., Gershon, E.S., 2002. Meta-analysis of whole-genome linkage scans of bipolar disorder and schizophrenia. Mol. Psychiatry 7 (4), 405–411.

Cai, T.T., Wang, L., 2011. Orthogonal matching pursuit for sparse signal recovery. IEEE Trans. Inf. Theory 57 (7), 1–26.

Callicott, J.H., Straub, R.E., Pezawas, L., Egan, M.F., Mattay, V.S., Hariri, A.R., Verchinski, B.A., Meyer-Lindenberg, A., Balkissoon, R., Kolachana, B., Goldberg, T.E., Weinberger, D.R., Jun. 2005. Variation in DISC1 affects hippocampal structure and function and increases risk for schizophrenia. Proc. Natl. Acad. Sci. U. S. A. 102 (24), 8627–8632.

Candès, E., Tao, T., 2006. Near optimal signal recovery from random projections: universal encoding strategies? IEEE Trans. Inf. Theory 52 (12), 5406–5425 (Dec.).

Cao, H., Deng, H., Li, M., Wang, Y.P., 2012a. Classification of multicolor fluorescence in-situ hybridization (M-FISH) images with sparse representation. IEEE Trans. Nano Biosci. 11 (2), 111–118.

Cao, H., Duan, J., Lin, D., Wang, Y.P., 2012b. Sparse representation based clustering for integrated analysis of gene copy number variation and gene expression data. IJCA 19 (2), 131 (June).

Cao, H., Duan, J., Lin, D., Calhoun, V., Wang, Y.P., 2012c. Biomarker identification for diagnosis of schizophrenia with integrated analysis of fMRI and SNPs, bioinformatics and biomedicine (BIBM). 2012 IEEE International Conference on. IEEE, Philadelphia, PA, USA, pp. 1–6. http://dx.doi.org/10.1109/BIBM.2012.6392674 (Oct. 4-7).

Cao, H., Lei, S., Deng, H.W., Wang, Y.P., 2012d. Identification of genes for complex diseases using integrated analysis of multiple types of genomic data. PLoS One 7 (9), e42755.

Cotter, S.F., Rao, B.D., Engan, K., Kreutz-Delgado, K., 2005. Sparse solutions to linear inverse problems with multiple measurement vectors. IEEE Trans. Signal Process. 53 (7), 2477–2488.

Davenport, M., Duarte, M., Hegde, C., Baraniuk, R., 2011. Introduction to compressive sensing. Connexions Web site http://cnx.org/content/m37172/1.7/ (Apr 10).

Davis, G., Mallat, S., Avellaneda, M., 1997. Greedy adaptive approximation. J. Constr. Approx. 13 (1), 57–98.

Donoho, D.L., 2004. Compressed sensing. Technical ReportStanford University.

Donoho, D.L., Elad, M., 2003. Maximal sparsity representation via L1 minimization. Proc. Natl. Acad. Sci. U. S. A. 100, 2197–2202 (Mar).

Donoho, D.L., Tsaig, Y., 2008. Fast solution of L1-norm minimization problems when the solution may be sparse. IEEE Trans. Inf. Theory 54, 4789–4812 (Nov).

Donoho, D.L., Elad, M., Temlyakov, V.N., 2006. Stable recovery of sparse over complete representations in the presence of noise. IEEE Trans. Inf. Theory 52 (1), 6–18.

Fisher, R.A., Yates, F., 1948. Statistical Tables for Biological, Agricultural and Medical Research, 3rd ed. Oliver & Boyd, London, pp. 26–27 (OCLC 14222135).

Foucart, S., Lai, M.J., 2009. Sparsest Solutions of Underdetermined Linear Systems via ℓq minimization for $0 < q \leq 1$. Appl. Comput. Harmon. Anal. 26 (3), 395–407 (May).

Gemmeke, J.F., Virtanen, T., Hurmalainen, A., 2011. Exemplar-based sparse representations for noise robust automatic speech recognition. IEEE Trans. Audio Speech Lang. Process. 19 (7), 2067–2080.

Gribonval, R., Nielsen, M., 2003. Sparse decompositions in unions of bases. IEEE Trans. Inf. Theory 49 (12), 3320–3325 (Dec.).

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.D., Blankertz, B., Bießmann, F., 2014. On the interpretation of weight vectors of linear models in multivariate neuroimaging. NeuroImage 87c, 96–110.

Hsu, D., Kakade, S., Langford, J., Zhang, T., 2009. Multi-label prediction via compressed sensing. Neural Information Processing Systems (NIPS).

Kidron, E., Schechner, Y.Y., Elad, M., 2007. Cross-modal localization via sparsity. IEEE Trans. Signal Process. 55 (4), 1390–1404 (Apr).

Kumari, V., Gray, J.A., Honey, G.D., Soni, W., Bullmore, E.T., Williams, S.C., Ng, V.W., Vythelingum, G.N., Simmons, A., Suckling, J., Corr, P.J., Sharma, T., 2012. Procedural learning in schizophrenia: a functional magnetic resonance imaging investigation. Schizophr. Res. 57 (1), 97–107 (Sep.).

Li, Y., Cichocki, A., Amari, S., 2004. Analysis of sparse representation and blind source separation. Neural Comput. 16 (6), 1193–1234.

Li, Y., Namburi, P., Yu, Z., Guan, C., Feng, J., Gu, Z., 2009. Voxel selection in fMRI data analysis based on sparse representation. IEEE Trans. Biomed. Eng. 56 (10), 2439–2451.

Lin, D., Cao, H., Calhoun, V.D., Wang, Y., 2011. Classification of schizophrenia patients with combined analysis of SNP and fMRI data based on sparse representation, BIBM 2011. IEEE International Conference on, Atlanta, GA, USA. IEEE.

Liu, J., Pearlson, G., Windemuth, A., Ruano, G., Perrone-Bizzozero, N.I., Calhoun, V.D., Jan. 2009. Combining fMRI and SNP data to investigate connections between brain function and genetics using parallel ICA. Hum. Brain Mapp. 30 (1), 241–255.

Meda, S.A., Bhattarai, M., Morris, N.A., Astur, R.S., Calhoun, V.D., Mathalon, D.H., Kiehl, K.A., Pearlson, G.D., 2008. An fMRI study of working memory in first-degree unaffected relatives of schizophrenia patients. Schizophr. Res. 104 (1–3), 85–95 (Sep).

Meier, L., Geer, S., Bühlmann, P., 2008. The group lasso for logistic regression. J. R. Stat. Soc. Ser. B 70 (1), 53–71.

Onitsuka, T., Shenton, M.E., Salisbury, D.F., Dickey, C.C., Kasai, K., Toner, S.K., Frumin, M., Kikinis, R., Jolesz, F.A., McCarley, R.W., 2004. Middle and inferior temporal gyrus gray matter volume abnormalities in chronic schizophrenia: an MRI study. Am. J. Psychiatry 161 (9), 1603–1611 (Sep).

Pascual-Leone, A., Manoach, D.S., Birnbaum, R., Goff, D.C., 2002. Motor cortical excitability in schizophrenia. Biol. Psychiatry 52 (1), 24–31 (Jul).

Rhodes, D.R., Chinnaiyan, A.M., 2005. Integrative analysis of the cancer transcriptome. Nat. Genet. 37, 31–37 (Jun).

Sharon, Y., Wright, J., Ma, Y., 2007. Computation and relaxation of conditions for equivalence between l1 and l0 minimization. UIUC, Tech. Rep. UILU-ENG-07-2008.

Shin, Y., Lee, S., Lee, J., Lee, H.-N., 2012. Sparse representation-based classification scheme for motor imagery-based brain–computer interface systems. J. Neural Eng. 9 (5) (1:12).

Soussen, C., Idier, J., Brie, D., Duan, J., 2011. From Bernoulli–Gaussian deconvolution to sparse signal restoration. IEEE Trans. Signal Process. 59 (10), 4572–4584.

Sutrala, S.R., Norton, N., Williams, N.M., Buckland, P.R., 2008. Gene copy number variation in schizophrenia. Am. J. Med. Genet. B Neuropsychiatr. Genet. 147B (5), 606–611 (Jul 5).

Szycik, G.R., Münte, T.F., Dillo, W., Mohammadi, B., Samii, A., Emrich, H.M., Dietrich, D.E., 2009. Audiovisual integration of speech is disturbed in schizophrenia: an fMRI study. Schizophr. Res. 110 (1–3), 111–118 (May).

Tang, W., Cao, H., Duan, J., Wang, Y.P., 2011. A compressed sensing based approach for subtyping of leukemia from gene expression data. J. Bioinforma. Comput. Biol. 9 (5), 631–645.

Tropp, J.A., 2004. Greed is good: algorithmic results for sparse approximation. IEEE Trans. Inf. Theory 50 (10), 2231–2242.

Wang, J., Yang, C.Y., Chen, B., 2011. Sparse Signal Recovery Based on lq ($0 < q = 1$) Minimization, 2011 International Conference on Multimedia and Signal Processing, Guilin, Guangxi China (ISBN: 978-0-7695-4356-7 May).

Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y., 2009. Robust face recognition via sparse representation. IEEE Trans. Pattern Anal. Intell. Mach. 31 (2), 210–227.

Xu, Z., Chang, X., Xu, F., 2012. L-1/2 regularization: a thresholding representation theory and a fast solver. IEEE Trans. Neural Netw. Learn. Syst. 23 (7), 1013–1027.