Published in final edited form as:
J Neurosci Methods. 2014 November 30; 0: 69–78. doi:10.1016/j.jneumeth.2014.09.001.

Sparse models for correlative and integrative analysis of imaging and genetic data

Dongdong Lin[1,2], **Hongbao Cao**[3], **Vince D. Calhoun**[4,5], and **Yu-Ping Wang**[1,2]

Dongdong Lin: dlin5@tulane.edu; Hongbao Cao: hongbao.cao@nih.gov; Vince D. Calhoun: vcalhoun@unm.edu; Yu-Ping Wang: wyp@tulane.edu

[1]Biomedical Engineering Department, Tulane University, New Orleans, LA, USA

[2]Center of Genomics and Bioinformatics, Tulane University, New Orleans, LA, USA

[3]Unit on Statistical Genomics, Intramural Program of Research, National Institute of Mental Health, NIH, Bethesda 20852, USA

[4]The Mind Research Network & LBERI, Albuquerque, NM 87106, USA

[5]Department of Electrical and Computer Engineering, University of New Mexico, Albuquerque, NM, 87131, USA

Abstract

The development of advanced medical imaging technologies and high-throughput genomic measurements has enhanced our ability to understand their interplay as well as their relationship with human behavior by integrating these two types of datasets. However, the high dimensionality and heterogeneity of these datasets presents a challenge to conventional statistical methods; there is a high demand for the development of both correlative and integrative analysis approaches. Here, we review our recent work on developing sparse representation based approaches to address this challenge. We show how sparse models are applied to the correlation and integration of imaging and genetic data for biomarker identification. We present examples on how these approaches are used for the detection of risk genes and classification of complex diseases such as schizophrenia. Finally, we discuss future directions on the integration of multiple imaging and genomic datasets including their interactions such as epistasis.

Keywords

Imaging genetics; sparse modeling; correspondence analysis; integration; classification

Correspondence to: Yu-Ping Wang, wyp@tulane.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could a3ect the content, and all legal disclaimers that apply to the journal pertain.

1. INTRODUCTION

In the past decades, increasing development of medical imaging and genomic techniques provides new opportunities to study tissue structure, function, and genetic variation as well as their relationship with human behavioral components, e.g., cognitive phenotypes and psychiatric disorders. For example, medical imaging measurements such as structural magnetic resonance imaging (sMRI), functional MRI (fMRI), diffusion tensor imaging (DTI) and positron emission tomography (PET), provide quantitative measurements of structural information at brain tissue level, dynamic blood oxygenation-level dependent (BOLD) response of neural activity and brain structural and functional connectivity. In addition, genetic measurements such as single polymorphism (SNP), gene expression, copy number variation (CNV) and proteomics can reveal structural and functional variations at molecular level. The goal of imaging genetics is to identify genetic factors that influence the intermediate quantitative measurements from anatomical or functional images, and the cognition and psychiatric disorders in humans. Rasetti and Weinberger et al [3] described a cascade of imaging genetic studies, in which mutations start from genetic level to cellular processes, to the system level, e.g., brain structure, function and integrity, and eventually to human behaviors. Numerous examples like this demonstrate that the fusion of imaging and genetics will facilitate the understanding of the pathophysiology, the diagnosis of complex and heritable psychiatric disorders and the optimization of treatments in a personalized manner.

In recent imaging genetic studies, neural imaging endophenotypes (or intermediate phenotype) derived from diverse medical images, are commonly used for genetic analysis. By the definition of ‘endophenotype’ by psychiatric geneticists [4, 5], an endophenotype should: 1) be associated with the illness/disorder of interest; 2) be heritable; 3) be state-independent; 4) exist temporally before the onset of the clinical illness in the pathophysiological pathway to the emergence of the clinical syndrome; 5) be found with higher frequency in healthy relatives of illness/disorder than in the general population. Each of these criteria is based on the hypothesis that the effects of susceptible genes will be more penetrated to the endophenotyes. The endophenotypes are considered to be closer to the biology of genetic function than the diagnostic results from self-reported and questionnaire-based clinical assessments [5, 6], which can boost the causal variants detection power. Some quantitative endophenotypes derived from brain imaging are reproducible and reliable with high heritability, and can accommodate highly heterogeneous symptoms from patients in the same group. For these reasons, obtaining reliable and heritable endophenotypes is critical for imaging genetics studies.

Many endophenotypes have been used in imaging genetic studies such as voxel-, vertex-, surface- or connection-based measures from structural, functional or diffusion images, respectively. For example, on structural MRI, the volumetric measures of total cerebral and gray and white matters [7, 8], cortical thickness and cortical area [9] have been studied as quantitative traits. On rest- or task- functional MRI images, there are functional endophenotypes utilized such as the extent of activation or deactivation for each voxel responding to the task-related stimuli, t-test contrast map, and functional brain connectivity [10]. On DTI images, brain integrity (e.g., fractional anisotropy and mean diffusivity), measures of coherent direction of axons (e.g., radial diffusivity and axial diffusivity) [11, 12] and the anatomical connectivity such as the measurement of fiber density or integrity have also been explored. Moreover, due to the high resolution of brain imaging (e.g., structural MRI), we can analyze the genetic influence on diverse imaging endophenotypes across the entire brain, which facilitates the understanding of underlying neurobiological mechanism of psychiatric disorders.

Fig. 1 shows an integrative approach for combining imaging and genetics techniques for biomarker detection, from which multiple psychiatric disorders or subtypes can be better classified. We will elaborate on this approach in the following three aspects: **1)** Between modality analysis to explore the correspondence or association between imaging and genetic data. One type of data is taken as an endophenotype (e.g., brain structure [9], fiber integrity [11], functional connectivity and network [13]) to find the correlated or associated variables in the other data (e.g., genotype [14]). The imaging endophenotypes are usually used as quantitative traits to explore the potential genetic risk factors. As reviewed in [15, 16], the research in this area has evolved from candidate approaches to the whole genome and whole brain investigation, and many complicated models are proposed to account for the increasing number of variables. **2**) Integration of imaging and genetic data for biomarker identification. A variety of medical imaging modalities (e.g., sMRI, fMRI and DTI) provide different insights on the change of brain or neuron activity at tissue-level, while genetic data (e.g., SNP, mRNA expression, DNA methylation and proteomics) measure different layers of genetic information at the molecular level. These different types of data are complementary, so combing these multiple modalities is likely to facilitate a better identification of biomarkers and a more comprehensive diagnosis of complex diseases. **3**) Sub-typing or classification of diseases by using biomarkers extracted from multimodal data. The identified biomarkers from imaging and genetic data contain complementary information about multiple psychiatric disorders (e.g., schizophrenia, bipolar and unipolar disorders). By using these biomarkers as features and input into a linear or non-linear classifier, we can achieve better disease classification, translating into more accurate diagnosis and ideally a clinical impact.

Many processing strategies and analysis approaches have been proposed to combine imaging and genetic information. For example, as reviewed in [17], between modality analysis methods can be categorized into univariate and multivariate imaging genetic analysis. Voxel-wise genome-wide association study (vGWAS [18]) and voxel-wise gene-wide study (vGeneWAS [19]) have been used to screen each pair of SNP/gene and voxel in maps of regional brain volume under the control of multiple comparisons. Canonical correlation analysis (CCA [20, 21]), partial least square(PLS [22, 23]) and parallel ICA [24] have been applied to extract a pair of correlated latent variables from imaging and genetic datasets. Kernel machine based [25] and Bayes methods [26] have also been proposed for imaging genetics analysis. For biomarker identification, joint ICA [27], multi-set CCA [28], multi-table PLS [29] and multi-task learning methods[30] have been used in modeling multimodal imaging, genetic and human behavioral data. Based on the identified biomarkers, a variety of classifiers such as support vector machine [31, 32] and multiple kernel learning [33, 34] have been applied to the classification of complex diseases.

Despite the success of these methods in the analysis of imaging and genetic data, there are still challenges due to the high dimensionality and heterogeneity of these datasets. For example, many conventional statistical methods such as CCA, PLS and ICA perform poorly for data with smaller sample size but with larger number of features/variables (e.g., voxels and SNPs). A dimensional reduction approach, e.g., principle component analysis or univariate test, is often used, which, however, may cause the loss of useful information. Statistical tests such as vGWAS [18] and vGeneWAS [19] may not have enough power due to a large number of multiple comparisons. In addition, high collinearity among whole brain or genome-wide variables may introduce instability and computational problems in the model (e.g., non-invertible of matrix, over-fitting). Sparse representation, a powerful method recently developed in statistics and signal processing, has found successful applications in a broad of fields such as bioinformatics [35], remote sensing [36] and image processing [37]. We have been developing sparse models for imaging genetics study, which demonstrate great advantages for the identification of biomarkers, leading to more accurate classification of diseases than many existing approaches.

In this review, we will review recent developments of sparse representation based methods for imaging genetics, with a focus on our own work. We will show how a variety of sparse models are developed and applied to the correlation and integration of imaging and genomic data. Finally, we conclude the review with a discussion of future research directions.

2. SPARSE MODELS FOR [CORRELATION/ASSOCIATION] ANALYSIS OF IMAIGING AND GENETIC DATA

Fig. 2 shows three types of sparse models in general for the correlation/association analysis between imaging and genetic data. In the first approach (Fig. 2(A)), sparse penalties such as lasso [38], group lasso [39], fuse lasso [40, 41] and sparse group lasso [42], have been used in the model to perform feature selection in a dataset (e.g., SNP data). A small number of SNPs (i.e., non-zero entries in the vector) are identified to be associated with specific imaging phenotypes. In the second and third approaches, the two-block methods (Fig. 2(B)) and sparse multivariate regression models (Fig. 2(C)) are used. Both are multivariate models to account for pleiotropic effects [43, 44] (i.e., genetic variants associated with multiple imaging quantitative traits) and the covariance structure among imaging endophenotypes [45, 46]. Two block methods include sparse CCA [47, 48], regularized kernel CCA [49] and sparse PLS correlation (PLSC) [50, 51], which are usually used for correspondence analysis between two modalities. Floch et al. [50] compared the performance of detecting the correlation between fMRI imaging and SNP data with different methods such as univariate approach, PLSC, sparse PLSC, regularized kernel CCA, and their combinations with PCA and pre-filters. The results show the best performance of using filtering in combination with sparse PLSC. Sparse multivariate regression is another way to study the association between multiple imaging and genetic factors, which includes sparse multi-task regression [52], sparse reduced rank regression(sRRR) [53–55], collaborative sRRR [56] and sparse PLS regression [57]. These different models incorporate prior knowledge with different ways, leading to different results. In the following, we present several models for the correlation or association analysis of imaging and genomic data, with a focus on our own work.

2.1 Sparse models for two block analysis

Due to high dimensionality and small sample size of neuroimaging and genetic datasets, there is usually an issue of high collinearity in the data. To this end, several types of regularized CCA and PLS approaches have been proposed, which enforce different sparse penalties (e.g., lasso, elastic net and sparse group lasso) on the loading vectors into the model. We present a unified framework [35]to formulate sparse CCA models as in Eq.1:

$$\min_{\boldsymbol{u},\boldsymbol{v}} - \boldsymbol{u}^t \sum\nolimits_{\boldsymbol{XY}} \boldsymbol{v} + \lambda_1 \|\boldsymbol{u}\|_G + \tau_1 \|u\|_1 + \lambda_2 \|v\|_G + \tau_2 \|\boldsymbol{v}\|_1 \quad s.t.\, \boldsymbol{u}^t \sum\nolimits_{\boldsymbol{XX}} \boldsymbol{u} \le 1, \boldsymbol{v}^t \sum\nolimits_{\boldsymbol{YY}} \boldsymbol{v} \le 1 \tag{1}$$

where ***X***,***Y*** are the two data matrices; ***u*** and ***v*** are the loading vectors constrained by sparse terms; $\|\boldsymbol{u}\|_1$and $\|\boldsymbol{v}\|_1$ are the $l-1$ norm based lasso penalty to perform the selection of individual variables/features in two datasets respectively; and $\|\boldsymbol{u}\|_G = \sum_{l=1}^{L} \omega_l \|\boldsymbol{u}_l\|_2, \|\boldsymbol{v}\|_G = \sum_{h=1}^{H} \mu_h \|\boldsymbol{v}_h\|_2$ are the group penalties to account for joint effects of features within the same group in two data sets respectively. The group penalty uses the non-diffentialbility of $\|\boldsymbol{u}_l\|_2$ (or $\|\boldsymbol{v}_h\|_2$) at $\boldsymbol{u}_l$=0 ($\boldsymbol{v}_h$=0) to set the coefficients of the group to 0, such that the entire group of features will be removed to achieve group sparsity.

The above model is realistic for the analysis of many cases, e.g., multiple SNPs rather than individual SNPs from the same gene usually function together as a group to be associated with a disease. This general formulation includes a variety of existing sparse CCA models as specific examples [47, 58], e.g., CCA-*l*1 or CCA-elastic net (λ_1=0, λ_2=0) and CCA-group lasso (τ_1=0, τ_2=0) models. The solution of Eq.1 usually involves with the inversion of the covariance matrices, which might be non-invertible because of the high dimensionality of data. In the simplified case when the covariance matrix is diagonal, sparse CCA model also reduces to sparse PLS model. A simulation has been performed to evaluate the efficiency of CCA-sparse group lasso model. Fig. 3(a) shows the results of recovered loading vectors ***u*** and ***v*** by CCA-*l*1, CCA-group and CCA-sparse group models respectively. It can be seen that the CCA-sparse group model can better estimate true u and ***v*** than CCA-*l*1, and CCA-group model. Fig. 3(b) compares the accuracy of recovering loading vectors from three methods with respect to different noise levels corresponding to different degrees of correlations between the two data sets. The result shows that the CCA-group model can recover the most correlated variables but give the highest total discordance. The CCA-sparse group model has a comparable recovering accuracy as the CCA-group model but with much less total discordance, especially when the noise level decreases. We have also applied these methods to fMRI data and SNP data to identify significant sets of SNPs correlated with brain region activity [59].

2.2 Sparse multivariate regression models

Multivariate regression model (Fig. 2(c)) is another popular approach for correlation analysis between imaging and genomic data, i.e., associating the entire genetic variants with whole brain imaging measurements. A regression model with regularization on the coefficient matrix can be described by the following formula:

$$\min_{\boldsymbol{A}} \|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{A}\|_F^2+\phi(\boldsymbol{A}) \quad (2)$$

where $\boldsymbol{X}$, $\boldsymbol{Y}$ are the two data matrices (e.g., genomics and imaging datasets) with $n\times p$ and $n\times q$ (p,q≫n) dimensions respectively; and $\boldsymbol{A}$ is $p\times q$ coefficient matrix penalized by function $\varphi(\cdot)$. There are several models that vary in the use of penalized functions.

A sparse multi-task learning method was applied to MRI and SNP datasets in Wang et al [52] with a combination of penalty terms on $\boldsymbol{A}$, $\phi(A)=\lambda_1\sum_{i=1}^{K}\|\boldsymbol{A}_{[i]}\|_F+\lambda_2\sum_{j=1}^{q}\|\boldsymbol{A}_j\|_1$, where $\|\boldsymbol{A}_{[i]}\|_F$ indicates a collaborative group lasso penalization on submatrix to account for the group structure of SNPs, e.g., multiple SNPs with linkage disequilibrium (LD) or from the same gene; $\|\boldsymbol{A}_j\|_1$ is a row sparse penalty on each individual SNP. The simulation experiment shows better performance of this method than multivariate linear regression, ridge regression and multitask feature learning models in predicting SNPs using longitudinal data.

Sparse low rank regression model was developed to consider the correlations among features in $\boldsymbol{Y}$ matrix. A low rank regularization was imposed on the coefficient matrix $\boldsymbol{A}$ due to the collinearity in $\boldsymbol{Y}$. Vounou et al [53] proposed a sparse reduced rank regression method (sRRR) by decomposing coefficient matrix into two full-rank matrices, i.e., $A = PQ$, $P \in R^{p\times r}$, $Q \in R^{r\times q}$, where r is the rank of $\boldsymbol{A}$ ($r < p, q$). At each stage sRRR performs unit rank decomposition to extract a pair of column vector P_k, Q_k, k=1,2,…r, which are also constrained to be sparse (i.e., with the use of $\|P_k\|_1$, $\|Q_k\|_1$). The sRRR was tested to have higher sensitivity than univariate method. It was applied to a whole brain and whole genome data set to identify those genetic variants associated with Alzheimer-related imaging biomarkers. A pathway sparse reduced rank (PsRRR) model [60], based on sRRR, was developed to consider the joint effects of SNPs within the same pathway by enforcing a group lasso penalty (i.e., $\sum_{g=1}^{G}\|P_k^g\|_2$), followed by a SNP-level selection at each pathway. A unit rank decomposition method is needed to estimate rank r, which cannot perform feature selection across ranks. Instead of decomposing $\boldsymbol{A}$ into two matrices, Wang et al [55] applied a trace-norm penalty $\|\cdot\|$, to enforce the low rank of $\boldsymbol{A}$ and combined it with group lasso penalty $\|\cdot\|_{2,1}$. The model is applied to the longitudinal imaging genetic data to identify imaging biomarkers associated with a set of SNPs. To further consider the group effects of SNPs and gene-gene interactions, we proposed a collaborative sparse reduce rank regression (c-sRRR) [56] to incorporate protein-protein interaction information into the model for the grouping of SNPs. The method can perform bi-level selection at both SNP- and module-levels. Several top gene modules are identified to be associated with functional network from postcentral and precentralgyri. The genes from these modules are enriched in some susceptible schizophrenia-related pathways [56].

In addition to the sparse models discussed above, there are other sparse multivariate regression analysis methods for integrative analysis of imaging and genetic data. For example, Liu et al [61] applied an overlapped group fused lasso to incorporate genetic information into the multivariate regression model to identify the brain connectivity pattern in resting fMRI data.

3. SPARSE MODEL FOR INTEGRATION OF IMAGING AND GENETIC DATA

Integrative analysis of multiple datasets can combine complementary information from each individual data and therefore may provide higher power to identify potential biomarkers that would otherwise be missed by using individual data alone. Due to different characteristics of diverse data modality (e.g., resolution, size and format), data integration is challenging. Wang et. al [62] proposed a sparse multimodal multitask learning method to combine sMRI, PET and GWAS data to identify genetic and phenotypic biomarkers associated with Alzheimer disease. In the model, the variables from each modality were combined equally and group lasso penalty was applied to identify those variables shared by multiple tasks (e.g., disease diagnosis, quantitative trait association). We address the data integration problem by developing a generalized sparse model (GSM) with weighting factors to account for the contribution of different data in data combination, as shown in Fig. 4. To solve the small-sample-large-variable problem, we developed a novel sparse representation based variable selection (SRVS) algorithm, which is described as the following,

For the purpose of illustration, we consider joint analysis of two types of data (which can be easily generalized to multiple data):

$$Y=[\alpha_1 A_1, \alpha_2 A_2]\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}+\varepsilon=AX+\varepsilon, \quad (3)$$

where $Y \in R^{m\times 1}$ is the observation vector (phenotypes of the subjects or disease status); $A_1 \in R^{m\times n1}$ and $A_2 \in R^{m\times n2}$ are the measurements of two different data types; $A = [\alpha_1 A_1, \alpha_2 A_2] \in R^{m\times n}$, where $\alpha_1 + \alpha_2 = 1$, and α_1, $\alpha_2 > 0$ are the weight factors for the two types of data; and $\varepsilon \in R^{m\times 1}$ is the measurement error due to noise. The biomarker/variable selection is formulated as the solution of sparse vector $X \in R^{n\times 1}$ based on $\boldsymbol{Y}$ and $\boldsymbol{A}$, where the sparsity (i.e., the number of non-zero elements) indicates the significant markers/variables selected, as illustrated in Fig. 4.

For SNPs and fMRI imaging data, the number of samples is significantly less than that of variables, i.e., $m \ll n$. In this case, the conventional sparse recovery algorithms will often fail. In order to overcome such a difficulty, we find the approximate solution to the problem by using the following approximation

$$min \sum_{i=1}^{t} \|X_i\|_p \quad subject\, to \quad \|Y-AX\|_2 \le \varepsilon \quad (4)$$

where X_i is the approximation to X. It can be solved with the minimization of the following functions:

$$min_X \, \|AX-Y\|^2+\lambda P(X) \quad (5)$$

where $P(X)=\sum_{i=1}^{t}\|X_i\|_p$ will make the solution to be sparse; λ is a regularization parameter, whose choice depends on the noise level. We have proposed an iterative procedure for the solution and we called this method as sparse representation based variable selection (SRVS) [2].

Using the model, we obtained encouraging result [2] for the selection of both fMRI imaging and genetic biomarkers associated with schizophrenia on the analysis of 208 subjects (92 cases and 116 controls). Fig. 5(a) shows the results of identified brain regions with SRVS method in comparison with Li et al.'s sparse regression method [1] (Fig. 5(b)); our method tends to find regions with voxels clustered together that have been verified before. We evaluate the performance of the method for differentiating healthy controls and schizophrenia patients [2, 63] by comparing the SRVS models regularized with three L_p norms (p = 0,0.5,1). They all give higher classification ratios than that of Li et al.'s method [1] ($p<1E-11$) and the sparse model regularized with $L_{1/2}$ norm generates the highest classification ratio (Fig. 5(c)).

4. APPLICATION TO DETECTION OF GENES AND DISEASE DIAGNOSIS

We demonstrate the applications of above described sparse models to imaging genomics with the following two examples: detection of genes and diagnosis of diseases.

4.1 Detection of risk genes with correlative models

Our prosed correlative or integrative model has found significant application to the detection of risk genes and signaling pathways. For example, based on the two pairs of canonical variates using the group sparse CCA model (Eq.1 of Sec.2.1), we found some regions of interest (ROI) in the brain correlated with a set of genes, contributing to the cause of schizophrenia. For each gene-ROI, gene-gene and ROI-ROI correlation, 10000 permutations were performed to test the significance. As shown in Fig. 6(a), Gene ERBB4, NRG1, MAGI2 and GABRG2 show correlations with some ROIs such as 3, 4, 7, 8 (the index of ROI is defined by the automated anatomic labelling (aal) template [64]) with correlations of ρ=0.1822, 0.1483, 0.1335, 0.1782($p<0.004$), respectively. These ROIs mostly consist of superior, middle, medial front gyrus and precentral gyrus located at frontal lobe containing primary motor cortex and were suspected to have abnormal changes in schizophrenia patients [65, 66]. ROIs 75, 76, 77, 11 and 13(not shown) are also found to be correlated with gene ERBB4, NRG1, and MAGI2 with the correlation value of 0.1822, 0.1649 and 0.1541 respectively. These ROIs are mainly located at thalamus (right), which plays a critical role in coordinating the pass of information between brain regions. Many studies show the association between dysfunction of thalamus with schizophrenia [27, 67, 68]. These three genes are also correlated with each other, which may have the similar effects on the ROIs. In addition, ROIs 91,92,99,100,103,111,112 are correlated with many genes (*e.g.*, GRIN2B, CHL1, ERBB4, NRG1, FOXP2, MAGI1, GABRG2, MAGI2). These ROIs are located at declive of cerebellum and culmen. Their relationship with schizophrenia is not clear yet but several previous publications have reported significant difference of these regions between normal control and schizophrenia patients [69] and there were models of schizophrenia on the importance of the cerebellum [70].

4.2 Better diagnosis with integrated information

Many studies have demonstrated that the integration of fMRI and SNP information will give a more comprehensive analysis of schizophrenia. Based on the biomarkers identified by the correlative and integrative models, we can use them to improve the diagnosis of schizophrenia in combination with appropriate classifiers. For example, Yang et al. [32]proposed a hybrid machine learning method to identify schizophrenia patients from healthy controls combining genetic and fMRI data, and this method achieved better classification accuracy than using either data alone. Castro et al.[33] proposed a multiple kernel learning approach that employed both the phase and magnitude of complex-valued fMRI data for the classification, showing improved classification accuracy compared to using only the magnitude of fMRI data.

We proposed a sparse representation clustering (SRC) model [2, 63] to simultaneously select SNPs and fMRI voxels as biomarkers for schizophrenia and then input the detected biomarkers as features into a classifier such as support vector machine (SVM). The accuracy of classification was further validated by the cross-validation. This approach also offers an objective evaluation of the identified biomarkers as described in Sec. 2 and 3. Fig. 7 gives a demonstration that the combination of the identified SNPs and fMRI voxels can lead to a better classification of schizophrenia patients from healthy controls in a study [2, 63]. When the integrative model was tested on 20 schizophrenia patients and 20 healthy controls, the use of both imaging and SNPs markers can improve the accuracy from 67.5% (i.e., using fMRI only) to 92.5% [71]. This indicates that by combining SNPs and imaging voxels, the complementary information from both data can increase the classification accuracy, leading to improved diagnosis.

In addition to feature selection based on sparse models described in Sec. 2 and 3, it should be noted that sparse representations can be directly applied to classification or subtyping of diseases. We have recently developed sparse models based classifiers to chromosome classification [72] and subtyping of leukemia [73] and gliomas [74], showing that sparse models provide a more robust approach in classifying image data containing variations, with better accuracy than many other classifiers such as SVM and Bayesian classifiers [75]. We have recently extended the model [72] to account for the correlation within the neighborhood of an image pixel [37] for further improvement of the classification. We are currently also applying these models for the classification of multiple psychiatric disorders (e.g., schizophrenia, bipolar and unipolar disorders).

5. CONCLUSION AND FUTURE DIRECTION

Imaging genetics is a relatively new and promising field which aims to explore genetic effects on the neurobiology and etiology of brain structure and function, and human behavior and psychiatric disorders. A number of heritable imaging endophenotypes can improve our understanding of genetic influence on different brain regions. The availability of multiscale and multimodal imaging genetics data bring about computational challenges for developing powerful, efficient and biologically interpretable methods.

There have been many proposed approaches for the integration of imaging and genomic information. Methods and tools such as ICA, CCA, pICA have been developed by us, with many successful applications. Despite this fact, there are still many challenges to overcome in modeling the complex data. Sparse representations provide a powerful and flexible way to model the multi-scale heterogeneous information. We have shown the advantage of sparse models in integrating imaging and genomic information in the following aspects: 1) Imaging genetics data usually contain a smaller sample size than that of features; enforcing sparsity can overcome the difficulty of many existing models such as CCA and regression analysis in modeling these data. 2) Imaging genomic data often exhibit many characteristics such as their correlations. Sparse models provide a powerful approach to consider this information into the integration. 3) Biological knowledge databases such as PPI contain rich information but is often overlooked. Sparse models provide a flexible way to incorporate this prior knowledge into the model. A problem with sparse models is that they are often computationally expensive. However, many effective algorithms are available to overcome such a difficulty.

The field of imaging genomics is progressing rapidly and poses significant challenges. Most integrative analysis methods in imaging genetics focus on two modalities but there are many other types of data available. Besides the GWAS data with SNPs, there are other levels of genetic data such as copy number variation, gene expression, microRNA, DNA methylation and proteomics. These various types of data have been widely studied in the study of complex diseases including cancers, e.g., TCGA(http://cancergenome.nih.gov/). Many advanced methods were proposed for integrating these multiple omic datasets [76]. The integration of other types of data will be able to provide a more comprehensive insight on genetic mechanisms underlying complex diseases. In addition, other types of neuroimaging measurements, e.g., sMRI, fMRI, PET, EEG and EPR, have also been available but limited work existed to integrate these multimodal imaging datasets as reviewed in[77]. For example, Sui et al.[78] proposed a mCCA+jICA method to integrate fMRI, DTI with DNA methylation to identify potential brain abnormalities in schizophrenia. It is both promising and pressing to develop multi-way integrative method to systematically combine these multiple types of imaging and genetic datasets to facilitate the understanding of brain activity and human behavior and their underlying biological mechanisms.

Sparse models have also shown promise for the study of epistasis, e.g., SNP-SNP interaction and gene-gene interaction. Epistasis, as an important factor for explaining the heritability of common traits, has been widely studied in genetic data analysis[79, 80]. Some work has reported and verified the effect of interactions on imaging traits as reviewed in [81]. The methods used for identifying these interactions are mainly based on exhaustive pair-wise searching. For example, Hibar et. al [82] applied an iterative sure independence screening (SIS) algorithm to search SNP-SNP interactions and found a significant interaction associated with temporal lobe volume. Meda et al. [83] identified 109 SNP-SNP interactions associated with right hippocampal atrophy and 125 for right entorhinal cortex atrophy. Despite of the crucial role of epistasis in explaining part of missing heritability, the extremely high computational burden of exhaustive searching imposes a challenge for the study. Sparse representation can provide much effective way for epistasis identification from such a large number of interactions. For examples, adaptive mix lasso model[84], adaptive group lasso[85]and more generally penalized multiple regressions[86] have the promise for large scale interaction analysis. Further effort is needed for the developments of sparse methods for detecting epistasis factors, which can be further incorporated into imaging genetics study.

In our current work, we mainly provide approaches for the prediction of schizophrenia patients from controls. However it is more interesting clinically to consider other psychiatric disorders such as bipolar disorder and unipolar disorders. In some cases (e.g. psychotic bipolar disorder and schizophrenia) these multiple mental illnesses share clinically overlapping symptoms, providing a way to study the diagnostic utility of our methods. Clinically unipolar and bipolar can sometimes be difficult to distinguish initially because individuals with bipolar sometimes have more depressive episodes than manic episodes and depressive episodes tend to be longer and more unpleasant for patients than manic episodes. To date, there is no biological marker to distinguish these two disorders. NIMH’s Research Domain Criteria (RDoC) Project [87] aims to define basic dimensions of function that cut across disorders as traditionally defined and can be studied across multiple units of analysis, from genes to neural circuits to behaviors. In such a context, the integration of imaging and genetics as well as other sources of biological information has the potential to better discriminate clinically subtle subgroups and hopefully to eventually translate into individualized treatments. As demonstrated in our examples, sparse models provide a powerful approach for the detection of biomarkers by integrating multiple imaging and genetic information, which can better utilize their specific characteristics and incorporate biological knowledge. We will apply these models to look not only within the diagnostic category, but also to evaluate the degree to which the biologic data is consistent with the symptom-based categories. We are currently developing novel sparse models to help identify risk genes and/or genomic regions underlying these multiple psychiatric disorders, and better distinguish these clinically cryptic subgroups. Based on these discoveries, personalized and optimal treatments can be done according to their different genetic make-ups.

Acknowledgments

This work is partially supported by NIH grants (NIBIB 2R01EB000840, COBRE 5P20RR021938/P20GM103472, R01GM109068, and R01MH104680).

References

1. Li Y, Namburi P, Yu Z, Guan C, Feng J, Gu Z. Voxel selection in FMRI data analysis based on sparse representation. IEEE transactions on bio-medical engineering. 2009; 56(10):2439–2451. [PubMed: 19567340]
2. Cao H, Duan J, Lin D, Calhoun V, Wang YP. Integrating fMRI and SNP data for biomarker identification for schizophrenia with a sparse representation based variable selection method. Bmc Med Genomics. 2013; 6 (Suppl 3):S2. [PubMed: 24565219]
3. Rasetti R, Weinberger DR. Intermediate phenotypes in psychiatric disorders. Curr Opin Genet Dev. 2011; 21(3):340–348. [PubMed: 21376566]
4. Shen L, Thompson PM, Potkin SG, Bertram L, Farrer LA, Foroud TM, Green RC, Hu X, Huentelman MJ, Kim S. Genetic analysis of quantitative phenotypes in AD and MCI: imaging, cognition and biomarkers. Brain imaging and behavior. 2013:1–25. [PubMed: 22660945]

5. Gottesman II, Gould TD. The endophenotype concept in psychiatry: etymology and strategic intentions. American Journal of Psychiatry. 2003; 160(4):636–645. [PubMed: 12668349]

6. Meyer-Lindenberg A, Weinberger DR. Intermediate phenotypes and genetic mechanisms of psychiatric disorders. Nature Reviews Neuroscience. 2006; 7(10):818–827.

7. Nymberg C, Jia T, Ruggeri B, Schumann G. Analytical strategies for large imaging genetic datasets: experiences from the IMAGEN study. Ann N Y Acad Sci. 2013; 1282:92–106. [PubMed: 23488575]

8. Baaré WF, Pol HEH, Boomsma DI, Posthuma D, de Geus EJ, Schnack HG, van Haren NE, van Oel CJ, Kahn RS. Quantitative genetic modeling of variation in human brain morphology. Cerebral Cortex. 2001; 11(9):816–824. [PubMed: 11532887]

9. Winkler AM, Kochunov P, Blangero J, Almasy L, Zilles K, Fox PT, Duggirala R, Glahn DC. Cortical thickness or grey matter volume? The importance of selecting the phenotype for imaging genetics studies. Neuroimage. 2010; 53(3):1135–1146. [PubMed: 20006715]

10. van den Heuvel MP, van Soelen IL, Stam CJ, Kahn RS, Boomsma DI, Hulshoff Pol HE. Genetic control of functional brain network efficiency in children. European Neuropsychopharmacology. 2013; 23(1):19–23. [PubMed: 22819192]

11. Jahanshad N, Kochunov PV, Sprooten E, Mandl RC, Nichols TE, Almasy L, Blangero J, Brouwer RM, Curran JE, de Zubicaray GI. Multi-site genetic analysis of diffusion images and voxelwise heritability analysis: A pilot project of the ENIGMA–DTI working group. Neuroimage. 2013; 81:455–469. [PubMed: 23629049]

12. Kochunov P, Glahn DC, Lancaster JL, Winkler AM, Smith S, Thompson PM, Almasy L, Duggirala R, Fox PT, Blangero J. Genetics of microstructure of cerebral white matter using diffusion tensor imaging. Neuroimage. 2010; 53(3):1109–1116. [PubMed: 20117221]

13. Thompson PM, Ge T, Glahn DC, Jahanshad N, Nichols TE. Genetics of the connectome. Neuroimage. 2013; 80:475–488. [PubMed: 23707675]

14. Filippini N, MacIntosh BJ, Hough MG, Goodwin GM, Frisoni GB, Smith SM, Matthews PM, Beckmann CF, Mackay CE. Distinct patterns of brain activity in young carriers of the APOE-epsilon4 allele. Proc Natl Acad Sci U S A. 2009; 106(17):7209–7214. [PubMed: 19357304]

15. Liu J, Calhoun VD. A review of multivariate analyses in imaging genetics. Front Neuroinform. 2014; 8:29. [PubMed: 24723883]

16. Ge T, Schumann G, Feng J. Imaging genetics—towards discovery neuroscience. Quantitative Biology. 2013:1–19. [PubMed: 24619230]

17. Hibar DP, Kohannim O, Stein JL, Chiang MC, Thompson PM. Multilocus genetic analysis of brain images. Front Genet. 2011; 2:73. [PubMed: 22303368]

18. Stein JL, Hua X, Lee S, Ho AJ, Leow AD, Toga AW, Saykin AJ, Shen L, Foroud T, Pankratz N. Voxelwise genome-wide association study (vGWAS). Neuroimage. 2010; 53(3):1160–1174. [PubMed: 20171287]

19. Hibar DP, Stein JL, Kohannim O, Jahanshad N, Saykin AJ, Shen L, Kim S, Pankratz N, Foroud T, Huentelman MJ. Voxelwise gene-wide association study (vGeneWAS): multivariate gene-based association testing in 731 elderly subjects. Neuroimage. 2011; 56(4):1875–1891. [PubMed: 21497199]

20. Correa NM, Li YO, Adali T, Calhoun VD. Canonical Correlation Analysis for Feature-Based Fusion of Biomedical Imaging Modalities and Its Application to Detection of Associative Networks in Schizophrenia. IEEE J Sel Top Signal Process. 2008; 2(6):998–1007. [PubMed: 19834573]

21. Sui J, Adali T, Pearlson G, Yang H, Sponheim SR, White T, Calhoun VD. A CCA+ICA based model for multi-task brain imaging data fusion and its application to schizophrenia. Neuroimage. 2010; 51 (1):123–134. [PubMed: 20114081]

22. Wold S, Martens H, Wold H. The Multivariate Calibration-Problem in Chemistry Solved by the Pls Method. Lect Notes Math. 1983; 973:286–293.

23. Krishnan A, Williams LJ, McIntosh AR, Abdi H. Partial Least Squares (PLS) methods for neuroimaging: a tutorial and review. Neuroimage. 2011; 56(2):455–475. [PubMed: 20656037]

24. Liu J, Pearlson G, Windemuth A, Ruano G, Perrone-Bizzozero NI, Calhoun V. Combining fMRI and SNP data to investigate connections between brain function and genetics using parallel ICA. Human brain mapping. 2009; 30(1):241–255. [PubMed: 18072279]

25. Ge T, Feng J, Hibar DP, Thompson PM, Nichols TE. Increasing power for voxel-wise genome-wide association studies: the random field theory, least square kernel machines and fast permutation procedures. Neuroimage. 2012; 63(2):858–873. [PubMed: 22800732]

26. Stingo FC, Guindani M, Vannucci M, Calhoun VD. An Integrative Bayesian Modeling Approach to Imaging Genetics. J Am Stat Assoc. 2013; 108(503)

27. Sui J, Pearlson G, Caprihan A, Adali T, Kiehl KA, Liu J, Yamamoto J, Calhoun VD. Discriminating schizophrenia and bipolar disorder by fusing fMRI and DTI in a multimodal CCA+ joint ICA model. Neuroimage. 2011; 57(3):839–855. [PubMed: 21640835]

28. Correa NM, Eichele T, Adali T, Li YO, Calhoun VD. Multi-set canonical correlation analysis for the fusion of concurrent single trial ERP and functional MRI. Neuroimage. 2010; 50(4):1438–1445. [PubMed: 20100584]

29. Caplan JB, McIntosh AR, De Rosa E. Two distinct functional networks for successful resolution of proactive interference. Cereb Cortex. 2007; 17(7):1650–1663. [PubMed: 16968868]

30. Zhou J, Chen J, Ye J. Clustered Multi-Task Learning Via Alternating Structure Optimization. NIPS: 2011. 2011:702–710.

31. Mourão-Miranda J, Bokde AL, Born C, Hampel H, Stetter M. Classifying brain states and determining the discriminating activation patterns: support vector machine on functional MRI data. Neuroimage. 2005; 28(4):980–995. [PubMed: 16275139]

32. Yang HH, Liu JY, Sui J, Pearlson G, Calhoun VD. A hybrid machine learning method for fusing fMRI and genetic data: combining both improves classification of schizophrenia. Front Hum Neurosci. 2010; 4

33. Castro E, Gomez-Verdejo V, Martinez-Ramon M, Kiehl KA, Calhoun VD. A multiple kernel learning approach to perform classification of groups from complex-valued fMRI data analysis: application to schizophrenia. Neuroimage. 2014; 87:1–17. [PubMed: 24225489]

34. Ji S, Sun L, Jin R, Ye J. Multi-label Multiple Kernel Learning. NIPS: 2008. 2008:777–784.

35. Lin D, Zhang J, Li J, Calhoun VD, Deng HW, Wang YP. Group sparse canonical correlation analysis for genomic data integration. Bmc Bioinformatics. 2013; 14:245. [PubMed: 23937249]

36. Chen Y, Nasrabadi NM, Tran TD. Hyperspectral Image Classification Using Dictionary-Based Sparse Representation. Ieee T Geosci Remote. 2011; 49(10):3973–3985.

37. Li J, Lin D, Cao H, Wang YP. An improved sparse representation model with structural information for Multicolour Fluorescence In-Situ Hybridization (M-FISH) image classification. BMC Syst Biol. 2013; 7 (Suppl 4):S5. [PubMed: 24565230]

38. Kohannim O, Hibar DP, Stein JL, Jahanshad N, Hua X, Rajagopalan P, Toga AW, Jack CR Jr, Weiner MW, de Zubicaray GI, et al. Discovery and Replication of Gene Influences on Brain Structure Using LASSO Regression. Front Neurosci. 2012; 6:115. [PubMed: 22888310]

39. Silver M, Montana G. Initiative AsDN. Fast identification of biological pathways associated with a quantitative trait using group lasso with overlaps. Stat Appl Genet Mol. 2012; 11(1):7.

40. Yang S, Pan Z, Shen X, Wonka P, Ye J. Fused multiple graphical lasso. 2012 arXiv preprint arXiv: 12092139.

41. Chen X, Kim S, Lin Q, Carbonell JG, Xing EP. Graph-structured multi-task regression and an efficient optimization method for general fused Lasso. 2010 arXiv preprint arXiv:10053579.

42. Silver M, Chen P, Li R, Cheng C-Y, Wong T-Y, Tai E-S, Teo Y-Y, Montana G. Pathways-Driven Sparse Regression Identifies Pathways and Genes Associated with High-Density Lipoprotein Cholesterol in Two Asian Cohorts. PLoS genetics. 2013; 9(11):e1003939. [PubMed: 24278029]

43. Andreassen OA, Djurovic S, Thompson WK, Schork AJ, Kendler KS, O'Donovan MC, Rujescu D, Werge T, van de Bunt M, Morris AP, et al. Improved detection of common variants associated with schizophrenia by leveraging pleiotropy with cardiovascular-disease risk factors. Am J Hum Genet. 2013; 92(2):197–209. [PubMed: 23375658]

44. Mier D, Kirsch P, Meyer-Lindenberg A. Neural substrates of pleiotropic action of genetic variation in COMT: a meta-analysis. Mol Psychiatry. 2010; 15(9):918–927. [PubMed: 19417742]

45. Alexander-Bloch A, Raznahan A, Bullmore E, Giedd J. The convergence of maturational change and structural covariance in human cortical networks. J Neurosci. 2013; 33(7):2889–2899. [PubMed: 23407947]

46. Bullmore E, Sporns O. Complex brain networks: graph theoretical analysis of structural and functional systems. Nat Rev Neurosci. 2009; 10(3):186–198. [PubMed: 19190637]

47. Le Cao KA, Martin PGP, Robert-Granie C, Besse P. Sparse canonical methods for biological data integration: application to a cross-platform study. Bmc Bioinformatics. 2009; 10:34. [PubMed: 19171069]

48. Lin D, Calhoun VD, Wang YP. Correspondence between fMRI and SNP data by group sparse canonical correlation analysis. Med Image Anal. 2013

49. Waaijenborg S, Zwinderman AH. Correlating multiple SNPs and multiple disease phenotypes: penalized non-linear canonical correlation analysis. Bioinformatics. 2009; 25(21):2764–2771. [PubMed: 19689958]

50. Le Floch E, Guillemot V, Frouin V, Pinel P, Lalanne C, Trinchera L, Tenenhaus A, Moreno A, Zilbovicius M, Bourgeron T, et al. Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse Partial Least Squares. Neuroimage. 2012; 63(1):11–24. [PubMed: 22781162]

51. Le Cao KA, Rossouw D, Robert-Granie C, Besse P. A sparse PLS for variable selection when integrating omics data. Stat Appl Genet Mol Biol. 2008; 7:Article 35. [PubMed: 19049491]

52. Wang H, Nie F, Huang H, Kim S, Nho K, Risacher SL, Saykin AJ, Shen L. Alzheimer's Disease Neuroimaging I. Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the ADNI cohort. Bioinformatics. 2012; 28 (2): 229–237. [PubMed: 22155867]

53. Vounou M, Nichols TE, Montana G. Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach. Neuroimage. 2010; 53 (3): 1147–1159. [PubMed: 20624472]

54. Vounou M, Janousova E, Wolz R, Stein JL, Thompson PM, Rueckert D, Montana G, Initia ADN. Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer's disease. Neuroimage. 2012; 60(1):700–716. [PubMed: 22209813]

55. Wang H, Nie F, Huang H, Yan J, Kim S, Nho K, Risacher SL, Saykin AJ, Shen L. From phenotype to genotype: an association study of longitudinal phenotypic markers to Alzheimer's disease relevant SNPs. Bioinformatics. 2012; 28(18):i619–i625. [PubMed: 22962490]

56. Lin, D.; Calhoun, V.; Deng, HW.; Wang, YP. Network-based investigation of genomic modules associated with functional brain network in schizophrenia. 2013 IEEE International Conference on Bioinformatics and Biomedicine (BIBM13); Shanghai, China. Dec. 17–21; IEEE; 2013.

57. Chun H, Keleş S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2010; 72(1):3–25.

58. Witten DM, Tibshirani RJ. Extensions of Sparse Canonical Correlation Analysis with Applications to Genomic Data. Statistical applications in genetics and molecular biology. 2009; 8(1)

59. Lin D, Calhoun V, Wang YP. Correspondence between fMRI and SNP data by group sparse canonical correlation analysis. Medical image analysis. 2013

60. Silver M, Janousova E, Hua X, Thompson PM, Montana G, Initiative aTAsDN. Identification of gene pathways implicated in Alzheimer's disease using longitudinal imaging phenotypes with sparse regression. Neuroimage. 2012; 63:1681–1694. [PubMed: 22982105]

61. Liu A, Chen X, Wang ZJ, Xu Q, Appel-Cresswell S, McKeown MJ. A genetically informed, group FMRI connectivity modeling approach: application to schizophrenia. IEEE Trans Biomed Eng. 2014; 61(3):946–956. [PubMed: 24557696]

62. Wang H, Nie F, Huang H, Risacher SL, Saykin AJ, Shen L. Identifying disease sensitive and quantitative trait-relevant biomarkers from multidimensional heterogeneous imaging genetics data via sparse multimodal multitask learning. Bioinformatics. 2012; 28(12):i127–i136. [PubMed: 22689752]

63. Cao, H.; Calhoun, V.; Wang, YP. Biomarker Identification for Diagnosis of Schizophrenia with Integrated Analysis of fMRI and SNPs. 2012 IEEE International Conference on Bioinformatics

and Biomedicine (BIBM12); Philadelphia, PA. 2013; IEEE; 2013. Selected as an invited paper for BMC Medical Genomics special issue

64. Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, Mazoyer B, Joliot M. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage. 2002; 15(1):273–289. [PubMed: 11771995]
65. Kiehl KA, Stevens MC, Celone K, Kurtz M, Krystal JH. Abnormal hemodynamics in schizophrenia during an auditory oddball task. Biol Psychiatry. 2005; 57(9):1029–1040. [PubMed: 15860344]
66. Honey GD, Pomarol-Clotet E, Corlett PR, Honey RA, McKenna PJ, Bullmore ET, Fletcher PC. Functional dysconnectivity in schizophrenia associated with attentional modulation of motor function. Brain. 2005; 128(Pt 11):2597–2611. [PubMed: 16183659]
67. Clinton SM, Meador-Woodruff JH. Thalamic dysfunction in schizophrenia: neurochemical, neuropathological, and in vivo imaging abnormalities. Schizophr Res. 2004; 69(2–3):237–253. [PubMed: 15469196]
68. Kiehl KA, Liddle PF. An event-related functional magnetic resonance imaging study of an auditory oddball task in schizophrenia. Schizophr Res. 2001; 48(2–3):159–171. [PubMed: 11295369]
69. Kim DI, Mathalon DH, Ford JM, Mannell M, Turner JA, Brown GG, Belger A, Gollub R, Lauriello J, Wible C, et al. Auditory oddball deficits in schizophrenia: an independent component analysis of the fMRI multisite function BIRN study. Schizophr Bull. 2009; 35(1):67–81. [PubMed: 19074498]
70. Andreasen NC, Pierson R. The role of the cerebellum in schizophrenia. Biol Psychiatry. 2008; 64 (2):81–88. [PubMed: 18395701]
71. Lin, D.; Calhoun, V.; Wang, YP. Integrating of SNPs and fMRI data for improved classification of schizophrenia. 2011 IEEE International Conference on Bioinformatics and Biomedicine (BIBM11); Atlanta, GA. 2011.
72. Cao HB, Deng HW, Wang YP. Segmentation of M-FISH Images for Improved Classification of Chromosomes With an Adaptive Fuzzy C-means Clustering Algorithm. Ieee T Fuzzy Syst. 2012; 20 (1):1–8.
73. Tang W, Cao H, Wang YP. A Compressive Sensing Method for Subtyping of Leukemia with Gene Expression Analysis Data. Journal of bioinformatics and computational biology. 2011; 9(5)
74. Tang W, Zhang J, Duan J, Wang YP. Subtyping of Glioma by Combining Gene Expression and CNVs Data Based on a Compressive Sensing Approach. Medical Advancements in Genetic Engineering. 2012; 1(1)
75. Cao HB, Deng HW, Li M, Wang YP. Classification of multicolor fluorescence in-situ hybridization (M-FISH) images with sparse representation. Ieee T Nanobiosci. 2012; 11(2):111–118.
76. Kristensen VN, Lingjærde OC, Russnes HG, Vollan HKM, Frigessi A, Børresen-Dale A-L. Principles and methods of integrative genomic analyses in cancer. Nat Rev Cancer. 2014; 14(5): 299–313. [PubMed: 24759209]
77. Sui J, Yu Q, He H, Pearlson GD, Calhoun VD. A selective review of multimodal fusion methods in schizophrenia. Front Hum Neurosci. 2012; 6:27. [PubMed: 22375114]
78. Sui J, He H, Liu J, Yu Q, Adali T, Pearlson GD, Calhoun VD. Three-way FMRI-DTI-methylation data fusion based on mCCA+jICA and its application to schizophrenia. Conf Proc IEEE Eng Med Biol Soc. 2012; 2012:2692–2695. [PubMed: 23366480]
79. Cordell HJ. Epistasis: what it means, what it doesn’t mean, and statistical methods to detect it in humans. Human molecular genetics. 2002; 11(20):2463–2468. [PubMed: 12351582]
80. Pan, Q.; Hu, T.; Moore, JH. Springer. Genome-Wide Association Studies and Genomic Prediction. 2013. Epistasis, complexity, and multifactor dimensionality reduction; p. 465-477.
81. Birnbaum R, Weinberger DR. Functional neuroimaging and schizophrenia: a view towards effective connectivity modeling and polygenic risk. Dialogues Clin Neurosci. 2013; 15(3):279–289. [PubMed: 24174900]

82. Hibar DP, Stein JL, Jahanshad N, Kohannim O, Toga AW, McMahon KL, de Zubicaray GI, Montgomery GW, Martin NG, Wright MJ, et al. Exhaustive search of the SNP-sNP interactome identifies epistatic effects on brain volume in two cohorts. Med Image Comput Comput Assist Interv. 2013; 16(Pt 3):600–607. [PubMed: 24505811]

83. Meda SA, Koran M, Pryweller JR, Vega JN, Thornton-Wells T. Genetic interactions associated with 12-month atrophy in hippocampus and entorhinal cortex in Alzheimer's Disease Neuroimaging Initiative. Neurobiology of aging. 2012; 30(1):e10.

84. Wang D, El-Basyoni IS, Baenziger PS, Crossa J, Eskridge K, Dweikat I. Prediction of genetic values of quantitative traits with epistatic effects in plant breeding populations. Heredity. 2012; 109(5):313–319. [PubMed: 22892636]

85. Yang C, Wan X, Yang Q, Xue H, Yu W. Identifying main effects and epistatic interactions from large-scale SNP data via adaptive group Lasso. Bmc Bioinformatics. 2010; 11(Suppl 1):S18. [PubMed: 20122189]

86. Hoffman GE, Logsdon BA, Mezey JG. PUMA: a unified framework for penalized multiple regression analysis of GWAS data. PLoS computational biology. 2013; 9(6):e1003101. [PubMed: 23825936]

87. Simmons JM, Quinn KJ. The NIMH Research Domain Criteria (RDoC) Project: implications for genetics research. Mammalian genome : official journal of the International Mammalian Genome Society. 2014; 25(1–2):23–31. [PubMed: 24085332]

Highlights

- We review integration methods for imaging and genomic data analysis.
- We focus on our efforts in developing sparse models for imaging and genomic data integration.
- We show real examples on applications of sparse models to detecting genes and diseases diagnosis.
- We give a perspective on future research directions in imaging genomics.

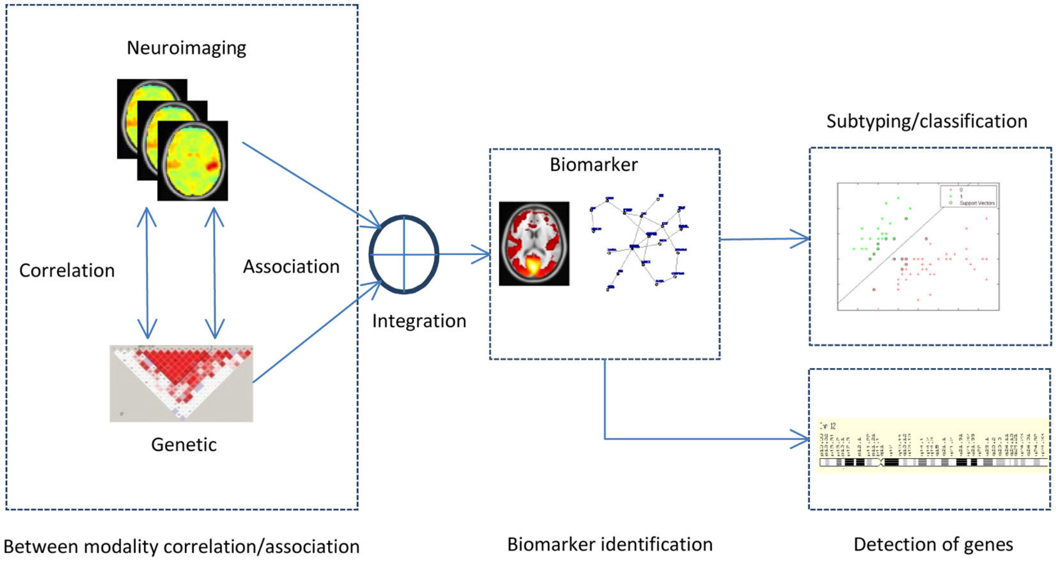

Figure 1.
The illustration of an integrative approach for imaging genetics study. Here, correlation/association is to find genetic biomarkers correlated/associated with imaging endophenotypes extracted from medical images and related to the disease of interest. Different from the correlation/association analysis, integration analysis aims to integrate complementary imaging and genetic data for identifying disease-related biomarkers and their interactions (e.g., genetic variants and their network, risk brain regions and their connectivity). The identified biomarkers will be then used for the detection of genes and/or the diagnosis/subtyping of complex mental illnesses.

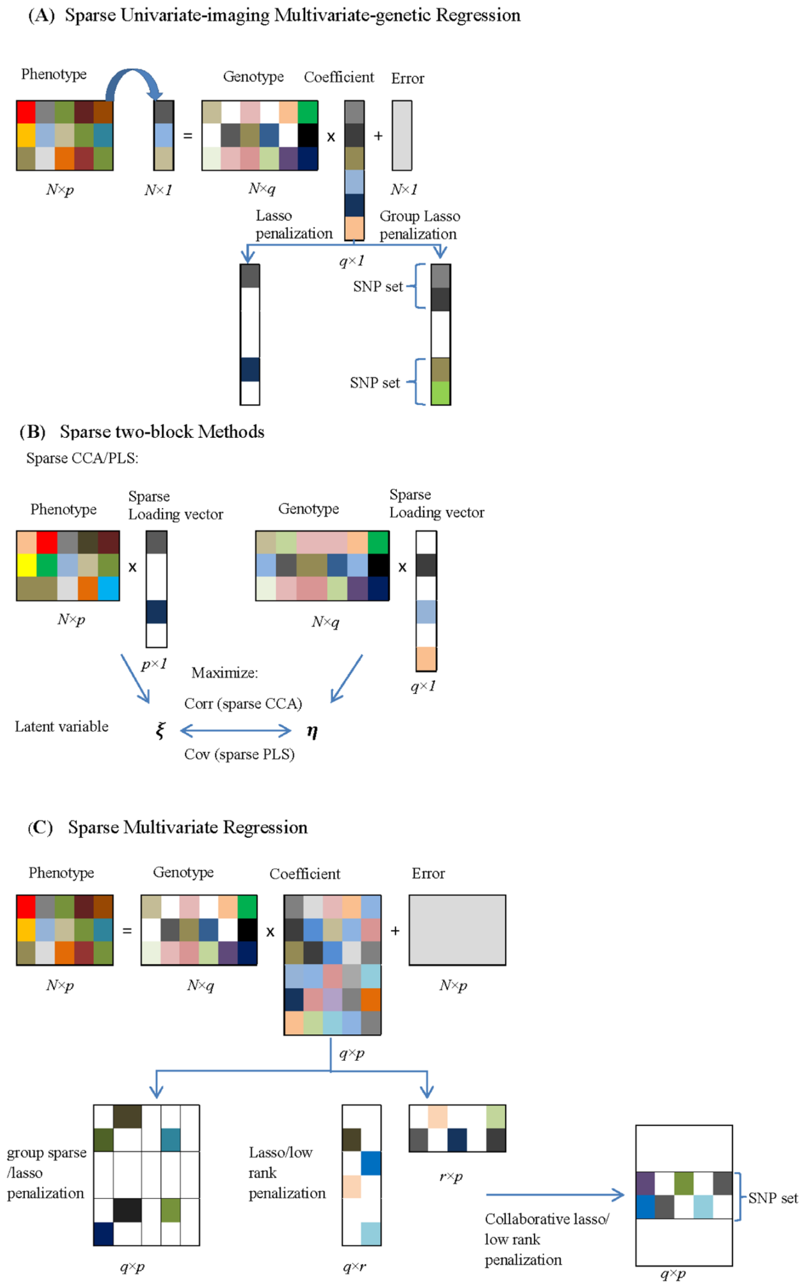

Figure 2.

An illustration of three types of analysis methods with sparse penalizations for imaging genetics study. (A). Sparse univariate-imaging and multivariate-genetic regression is a sparse multiple regression model used to identify multiple genetic factors associated with a single imaging endopheontype. (B). Sparse two-block methods include a number of correspondence analysis methods using sparse latent variable models, i.e., sCCA, sparse kernel CCA and sPLS. These methods can be used to explore a pair of correlated latent variables, which consists of linear or non-linear combination of sparse features from each dataset. (C). Sparse multivariate regression is a regression based model to identify a set of genetic factors associated with a set of imaging endophenotypes, represented by the coefficient matrix penalized with a variety of sparse terms.

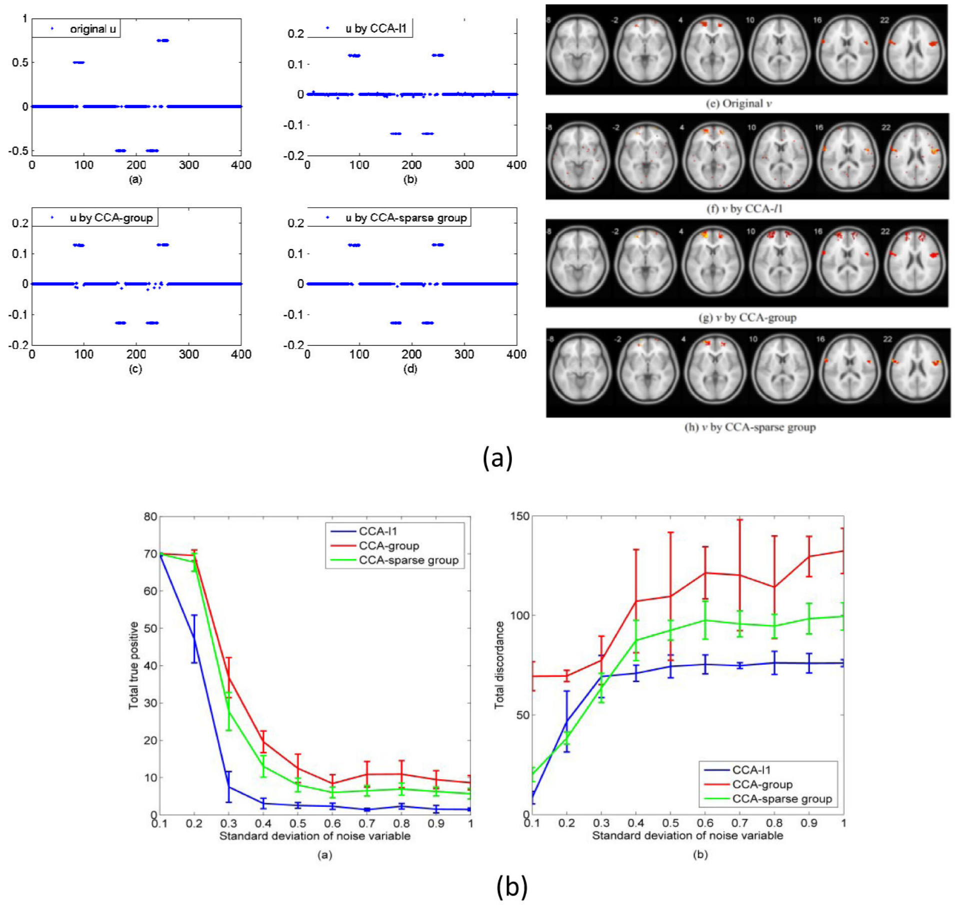

Figure 3.

A comparison of group sparse CCA with the other sparse CCA models (e.g., CCA-lasso, CCA-group lasso).(a) The accuracy of recovering the loading vectors (i.e., ***u*** and ***v*** for simulated genetic and imaging data respectively) by three different models. (b) The total true positives and total discordance with respect to different correlation values between imaging and genetic datasets.

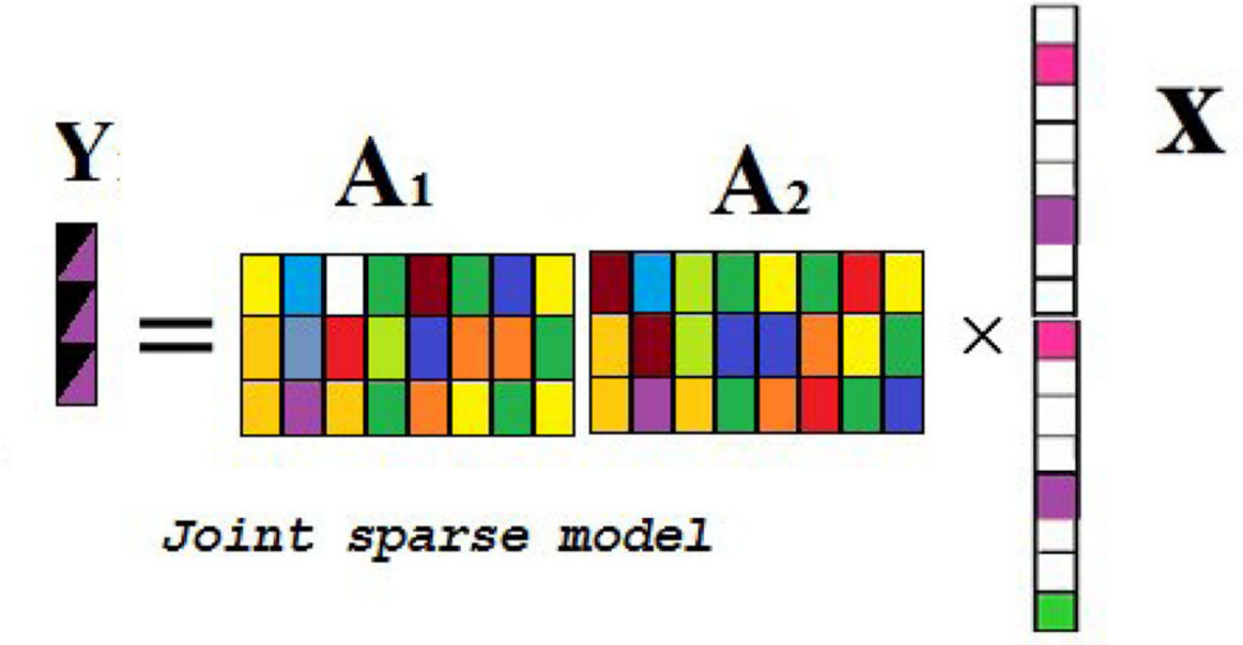

Figure 4.

The representation of phentoypes or disease states by the fusion of two datasets A_1 and $A_{2.}$ with a parallel sparse model, where the correlative information is represented simultaneously by non-zero entries in a sparse vector. The model can be easily extended to include multiple data sets.

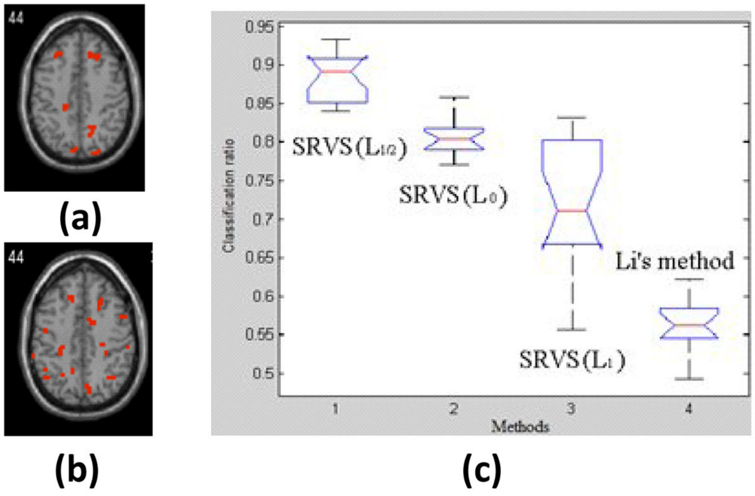

Figure 5.

Our sparse model based variable selection (SRVS) method can better identify abnormal brain region (a) than Li et al.'s method [1] (b) (i.e., yielding more clustered regions that have been validated before), and give better classification accuracy for discriminating schizophrenia patients and health controls (c).

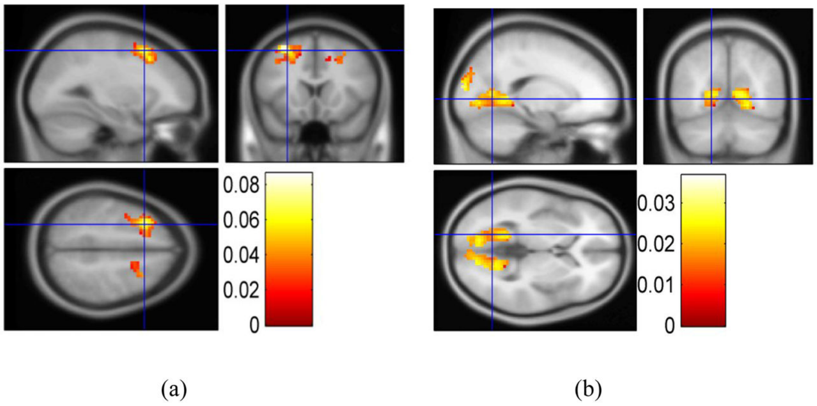

Figure 6.

(a). ROIs 3,4,7,8 which are identified to be mainly correlated with genes ERBB4, NRG1, MAGI1 and GABRG2. (b) ROIs 40,43,44,45,46,47,48,49,50,51,52,53,54,55,56,67,68 which are found to correlate with Gene ERBB4, CHL1, GRM3 and MAGI2. Note: the index of ROI is given by the aal template [64].

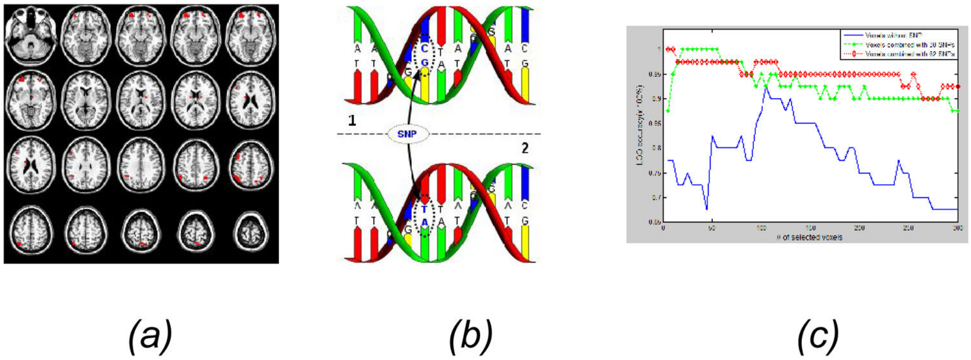

Figure 7.

Integration of fMRI (a) and SNP (b) with our proposed sparse models[2]results in better classification of schizophrenia and healthy controls than using just fMRI data or SNPs only. In (c), blue line is the classification result without using combined data, while the red and green lines are the results by combing fMRI with SNP markers with different numbers, showing higher classification accuracy.