Identifying disease sensitive and quantitative trait-relevant biomarkers from multidimensional heterogeneous imaging genetics data via sparse multimodal multitask learning

Hua Wang1, Feiping Nie1, Heng Huang1,*, Shannon L. Risacher2, Andrew J. Saykin2 and Li Shen2,∗

1Department of Computer Science and Engineering, University of Texas at Arlington, TX 76019 and 2Department of Radiology and Imaging Sciences, Indiana University School of Medicine, Indianapolis, IN 46202, USA

ABSTRACT

Motivation: Recent advances in brain imaging and high-throughput genotyping techniques enable new approaches to study the influence of genetic and anatomical variations on brain functions and disorders. Traditional association studies typically perform independent and pairwise analysis among neuroimaging measures, cognitive scores and disease status, and ignore the important underlying interacting relationships between these units.

Results: To overcome this limitation, in this article, we propose a new sparse multimodal multitask learning method to reveal complex relationships from gene to brain to symptom. Our main contributions are three-fold: (i) introducing combined structured sparsity regularizations into multimodal multitask learning to integrate multimodal heterogeneous imaging genetics data and identify multimodal biomarkers; (ii) utilizing a joint classification and regression learning model to identify disease-sensitive and cognition-relevant biomarkers; (iii) deriving a new efficient optimization algorithm to solve our non-smooth objective function and providing rigorous theoretical analysis on the global optimum convergency. Using the imaging genetics data from the Alzheimer’s Disease Neuroimaging Initiative database, the effectiveness of the proposed method is demonstrated by clearly improved performance on predicting both cognitive scores and disease status. The identified multimodal biomarkers could predict not only disease status but also cognitive function to help elucidate the biological pathway from gene to brain structure and function, and to cognition and disease.

Availability: Software is publicly available at: http://ranger.uta.edu/~heng/multimodal/

Contact: heng@uta.edu; shenli@iupui.edu

1 INTRODUCTION

Recent advances in acquiring multimodal brain imaging and genome-wide array data provide exciting new opportunities to study the influence of genetic variation on brain structure and function. Research in this emerging field, known as imaging genetics, holds great promise for a system biology of the brain to better understand complex neurobiological systems, from genetic determinants to cellular processes to the complex interplay of brain structure, function, behavior and cognition. Analysis of these multimodal datasets will facilitate early diagnosis, deepen mechanistic understanding and improved treatment of brain disorders.

Machine learning methods have been widely employed to predict Alzheimer’s disease (AD) status using imaging genetics measures (Batmanghelich et al., 2009; Fan et al., 2008; Risacher et al., 2009; Shen et al., 2010a). Since AD is a neurodegenerative disorder characterized by progressive impairment of memory and other cognitive functions, regression models have also been investigated to predict clinical scores from structural, such as magnetic resonance imaging (MRI), and/or molecular, such as fluorodeoxyglucose positron emission tomography (FDG-PET), neuroimaging data (Stonnington et al., 2010; Walhovd et al., 2010). For example, Walhovd et al. (2010) performed stepwise regression in a pairwise fashion to relate each of MRI and FDG-PET measures of eight candidate regions to each of four Rey’s Auditory Verbal Learning Test (RAVLT) memory scores. This univariate approach, however, did not consider either interrelated structures within imaging data or those within cognitive data. Using relevance vector regression (Stonnington et al., 2010) jointly analyzed the voxel-based morphometry (VBM) features extracted from the entire brain to predict each selected clinical score, while the investigations of different clinical scores are independent from each other.

One goal of imaging genetics is to identify genetic risk factors and/or imaging biomarkers via intermediate quantitative traits (QTs, e.g. cognitive memory scores used in this article) on the chain from gene to brain to symptom. Thus, both disease classification and QT prediction are important machine learning tasks. Prior imaging genetics research typically employs a two-step procedure for identifying risk factors and biomarkers: one first determines disease-relevant QTs, and then detects the biomarkers associated with these QTs. Since a QT could be related to many genetic or imaging markers on different pathways that are not all disease specific (e.g. QT 2 and Gene 3 in Fig. 1), an ideal scenario would be to discover only those markers associated with both QT and disease status for a better understanding of the underlying biological pathway specific to the disease.

On the other hand, identifying genetic and phenotypic biomarkers from large-scale multidimensional heterogeneous data is an important biomedical and biological research topic. Unlike simple feature selection working on a single data source, multimodal learning describes the setting of learning from data where observations are represented by multiple types of feature sets. Many multimodal methods have been developed for classification and clustering purposes, such as co-training (Abney, 2002; Brefeld and Scheffer, 2004; Ghani, 2002; Nigam et al., 2002), Brefeld and Scheffer (2004), Ghani (2002), Nigam et al. (2002).
The structured sparsity is usually obtained through different sparse simultaneous biomarkers correlated to memory scores and disease categories and regression multitask learning model is utilized to select the score as a regression task and select biomarkers that tend to play an individual QT (memory score), we consider predicting each memory related methods that mainly find the biomarkers correlated to each (Tibshirani, 1996), group LASSO (Yuan and Lin, 2006) and other biomarkers from multiple data sources. Different to LASSO heterogeneous genetic and phenotypic data effectively and limitation often yields inadequate performance. When they are combined with the features from other sources. This i.e. all features from the same data source are weighted equally, models train a single weight for all features from the same modality, heterogeneous data and select multitype features. However, such and Ong, 2007) have been recently studied and employed to integrate types: (i) The flat sparsity is often achieved by imposing non-smooth norms as regularizers in the optimization methods such as Lasso (Tibshirani, 1996) and group Lasso (Yuan and Lin, 2006), proximal methods (Beck and Teboulle, 2009). (ii) The structured sparsity is usually obtained through different sparse regularizers such as $\ell_1$-norm (Kim and Xing, 2010), group LASSO, $\ell_2$-norm (Fan et al., 2004), linear gradient search (Fan et al., 2004), proximal methods (Bach et al., 2009), multiview clustering (Bickel and Scheffler, 2008; Dhillon et al., 2003). However, they typically assume that the multimodal feature sets are conditionally independent, which does not hold in many real-world applications such as imaging genetics. Considering different representations give rise to different kernel functions, several Multiple Kernel Learning (MKL) approaches (Micchelli et al., 2006; Suykens et al., 2002; Ye et al., 2004; Zien et al., 2000) have been recently studied and employed to integrate heterogeneous data and select multitype features. However, such classification and regression multitask learning model is utilized to select the biomarkers correlated to memory scores and disease categories simultaneously.

In multitask learning, given a set of input variables (i.e. features such as SNPs and MRI/PET measures), we are interested in learning a set of related models (e.g. relations between genetic/imaging markers and cognitive scores) to predict multiple outcomes (i.e. tasks such as predicting cognitive scores and disease status). Because these tasks are relevant, they share a common input space. As a result, it is desirable to learn all the models jointly rather than treating each task as independent and fitting each model separately, such as Lasso (Tibshirani, 1996) and group Lasso (Yuan and Lin, 2006). Such multitask learning can discover regular patterns (because significant patterns in a single task could be outliers for other tasks) and potentially increase the predictive power.
In this article, we write matrices as uppercase letters and vectors as boldface lowercase letters. Given a matrix \( W = [w_{ij}] \), its \( i \)-th row and \( j \)-th column are denoted as \( w_i \) and \( w_j \), respectively. The \( \ell_{2,1} \) norm of the matrix \( W \) is defined as \( ||W||_{2,1} = \sum_{j=1}^{n} ||w_j||_2 \) (also denoted as \( \ell_{1,2} \)-norm by other researchers).

### 2.1 Heterogeneous data integration via combined structured sparse regularizations

First, we will systematically propose our new multimodal learning method to integrate and select the genetic and phenotypic biomarkers from large-scale heterogeneous data. In the supervised learning setting, we are given \( n \) training samples \( \{(x_i, y_i)\}_{i=1}^{n} \), where \( x_i = (x_i^1, \ldots, x_i^d)^T \in \mathbb{R}^d \) is the input vector including all features from a total of \( k \) different modalities and each modality \( j \) has \( d_j \) features \( d = \sum_{j=1}^{k} d_j \). \( y_i \in \mathbb{R}^q \) is the class label vector of data point \( x_i \) (only one element in \( y_i \) is 1, and others are zeros), where \( c \) is the number of classes (tasks). Let \( X = [x_1, \ldots, x_n] \in \mathbb{R}^{d \times n} \) and \( Y = [y_1, \ldots, y_n] \in \mathbb{R}^{q \times n} \). Different to MKL, we directly learn a \( d \times c \) parameter matrix as:

\[
W = \begin{bmatrix} w_{1,1} & \cdots & w_{1,c} \\ \vdots & \ddots & \vdots \\ w_{d,1} & \cdots & w_{d,c} \end{bmatrix} \in \mathbb{R}^{d \times c},
\]

(1)

where \( w_{q, j} \in \mathbb{R}^{d_j} \) indicates the weights of all features in the \( q \)-th modality with respect to the \( p \)-th task (class). Typically, we can use a convex loss function \( \mathcal{L}(X, W) \) to measure the loss incurred by \( W \) on the training samples. Compared with MKL approaches that learn one weight for one kernel matrix representing one modality, our method will learn the weight for each feature to capture the local feature importance. Since the features come from heterogeneous data sources, we impose the regularizer \( \mathcal{R}(W) \) to capture the interrelationships of modalities and features as:

\[
\min_{W} \mathcal{L}(X, W) + \gamma \mathcal{R}(W),
\]

(2)

where \( \gamma \) is a trade-off parameter. In heterogeneous data fusion, from multiview perspective of view, the features of a specific view (modality) can be more or less discriminative for different tasks (classes). Thus, we propose a new group \( \ell_1 \)-norm (\( G_1 \)-norm) as a regularization term in Equation (2), which is defined over \( W \) as following:

\[
||W||_{G_1} = \sum_{i=1}^{c} \sum_{j=1}^{k} ||w_{i,j}||_2,
\]

(3)

which is illustrated by the blue circles in Figure 3. Then the Equation (2) becomes:

\[
\min_{W} \mathcal{L}(X, W) + \gamma_1 ||W||_{G_1},
\]

(4)

Since the group \( \ell_1 \)-norm uses \( \ell_2 \)-norm within each modality and \( \ell_1 \)-norm between modalities, it enforces the sparsity between different modalities, i.e. if one modality of features are not discriminative for certain tasks, the objective in Equation (4) will assign zeros (in ideal case, usually they are very small values) to them for corresponding tasks; otherwise, their weights are large. This new group \( \ell_{1,2} \)-norm regularizer captures the global relationships between data modalities.

However, in certain cases, even if most features in one modality are not discriminative for the classification or regression tasks, a small number of features in the same modality can still be highly discriminative. From the multitask learning point of view, such important features should be shared by all/most tasks. Thus, we add an additional \( \ell_{2,1} \)-norm regularizer into Equation (4) as:

\[
\min_{W} \mathcal{L}(X, W) + \gamma_1 ||W||_{G_1} + \gamma_2 ||W||_{2,1},
\]

(5)

The \( \ell_{2,1} \)-norm was popularly used in multitask feature selection challenges because it can be efficiently implemented via convex optimization techniques. The \( \ell_{2,1} \)-norm regularizer impose the sparsity between all features and non-sparsity between tasks, the features that are discriminative for all tasks will get large weights.

![Illustration of the feature weight matrix \( W \).](image)

Our regularization items consider the heterogeneous features from both group-wise and individual viewpoints. Figure [4] visualizes the matrix \( W \) as a demonstration. In Figure 4 the elements with deep blue color have large values. The group \( \ell_1 \)-norm emphasizes the group-wise weights learning corresponding to each task and the \( \ell_{2,1} \)-norm accentuates the individual weight learning cross multiple tasks. Through the combined regularizations, for each task (class), many features (not all of them) in the discriminative modalities and a small number of features (may not be none) in the non-discriminative modalities will learn large weights as the important and discriminative features.

The multidimensional data integration has been increasingly important to many biological and biomedical studies. So far, the MKL methods are most widely used. Due to the learning model deficiency, the MKL methods cannot explore both modality-wise importance and individual importance of features simultaneously. Our new structured sparse multimodal learning method integrates the multidimensional data in a more efficient and effective way. The loss function \( \mathcal{L}(X, W) \) in Equation (4) can be replace by either least square loss function or logistic regression loss function to perform regression/classification tasks.

### 2.2 Joint disease classification and QT regression

Since we are interested in identifying the disease-sensitive and QT-relevant biomarkers, we consider performing both logistic regression for classifying disease status and multivariate regression for predicting cognitive memory scores simultaneously (Wang et al., 2011). A similar model was used in Yang et al. (2009) for...
heterogeneous multitask learning. Regular multitask learning only considers homogeneous tasks such as regression or classification individually. Joint classification and regression can be regarded as a learning paradigm for handling heterogeneous tasks.

First, logistic regression is used for disease classification, which minimizes the following loss function:

$$L_1(W) = \sum_{i=1}^{n} \sum_{k=1}^{c} \left( y_{ik} \log \frac{\sum_{l=1}^{c} W_{i}^{l} x_{i}^T - y_{ik} W_{i}^{T} x_{i}}{1 + \sum_{l=1}^{c} W_{i}^{l} x_{i}^T} \right)$$

(6)

Here, we perform three binary classification tasks for the following three diagnostic groups respectively (c1 = 3): AD, mild cognitive impairment (MCI), and health control (HC).

Second, we use the traditional multivariate least squares regression model to predict memory scores. Under the regression matrix $P \in \mathbb{R}^{d \times c}$, the least squares loss is defined by

$$L_2(P) = \|X^T P - Z\|_F^2,$$  

(7)

where $X$ is the data points matrix, $P$ is the coefficient matrix of regression with $c_2$ tasks, the label matrix $Z = \begin{bmatrix} (x_1) \ T \ (x_2) \ T \ \cdots \ (x_c) \ T \end{bmatrix} \in \mathbb{R}^{d \times c_2}$.

We perform the joint classification and regression tasks, the disease-sensitive and QTR-relevant biomarker identification task can be formulated as the following objective:

$$\min_{W} \sum_{i=1}^{n} c_1 \left( y_{k} \log \sum_{l=1}^{c} W_{i}^{l} x_{i}^T - y_{k} W_{i}^{T} x_{i} \right) + \frac{1}{2} \|X^T P - Z\|_F^2 \|V\|_{2,1},$$

(8)

where $V = [P] \in \mathbb{R}^{d \times (c_1 + c_2)}$. As a result, the identified biomarkers will be correlated to memory scores and also be discriminative to disease categories.

Since the objective in Equation (8) is a non-smooth problem and cannot be easily solved in general, we derive a new efficient algorithm to solve this problem in the next subsection.

2.3 Optimization algorithm

We take the derivatives of Equation (8) with respect to $W$ and $P$, respectively, and set them to zeros, we have

$$\frac{\partial C_1(W)}{\partial W} + 2\gamma_1 \sum_{i=1}^{n} D_W x_i + 2\gamma_2 D_W P = 0,$$  

(9)

$$2XX^T P - 2XXZ + 2\gamma_1 \sum_{i=1}^{n} D_W x_i + 2\gamma_2 D_W P = 0,$$  

(10)

where $D_W(1 \leq i \leq c_1 + c_2)$ is a block diagonal matrix with the $k$-th diagonal block as $\frac{1}{X_i^T X_i} I$ ($Q_d$ is a $Q_d$ by $Q_d$ identity matrix), $D$ is a diagonal matrix with the $k$-th diagonal element as $\frac{1}{X_i^T X_i}$. Since $D_W(1 \leq i \leq c_1 + c_2)$ and $D$ depend on $V = [P]$, they are also unknown variables to be optimized. In this article, we provide an iterative algorithm to solve Equation (8). First, we guess a random solution $V \in \mathbb{R}^{c_1 + c_2}$, then we substitute the matrices of $D_W(1 \leq i \leq c_1 + c_2)$ and $D$ according to the current solution $V$. After obtaining the $D_W(1 \leq i \leq c_1 + c_2)$ and $D$, we can update the solution $V = [P]$ based on Equation (8). Specifically, the $i$-th column of $P$ is updated by $P_i = (XX^T + \gamma_1 D_W + \gamma_2 D)^{-1} X_i$. We cannot update $W$ with a closed form solution based on Equation (8), but we can obtain the updated $W$ by the Newton’s method. According to Equation (8), we need to solve the following problem:

$$\min_{W} L_1(W) + \gamma_1 \sum_{i=1}^{n} w_i^T D_W w_i + \gamma_2 Tr(W^T D W).$$  

(11)

Similar to the traditional method in the logistic regression [Krishnapuram et al. 2005; Lee et al. 2004], we can use the Newton’s method to obtain the solution $W$.

For the first term, the traditional logistic regression derivatives can be applied to get the first-and second-order derivatives [Lee et al. 2004].

For the second term, the first-and second-order derivatives are

$$\frac{\partial C_1}{\partial W} = 2D_W(u, w)W_{up},$$

(12)

$$\frac{\partial C_1}{\partial W} = 2D_W(u, w)W_{up},$$

(13)

where $D_W(v, w)$ is the $u$-th diagonal element of $D_W$. For the third term, the first-and second-order derivatives are

$$\frac{\partial Tr(W^T D W)}{\partial W_{up}} = 2D_W(u, w)W_{up},$$

(14)

After obtaining the updated solution $V = [P]$, we can calculate the new matrices $D_W(1 \leq i \leq c_1 + c_2)$ and $D$. This procedure is repeated until the algorithm converges. The detailed algorithm is listed in Algorithm 1. We will prove that the above algorithm will converge to the global optimum.

2.4 Algorithm analysis

To prove the convergence of the proposed algorithm, we need a lemma as follows.

**Lemma 1.** For any vectors $v$ and $v_0$, we have the following inequality:

$$\|v\|_2^2 \geq \frac{\|v\|_2^2 - \|v_0\|_2^2}{2\|v_0\|_2^2} \|v_0\|_2^2.$$  

**Proof.** Obviously, $-\frac{1}{2}\|v\|_2^2 + \|v_0\|_2^2 \leq 0$, so we have

$$(-\|v\|_2^2 + \|v_0\|_2^2) \leq 0 \Rightarrow \|v\|_2^2 \|v_0\|_2^2 - \|v\|_2^2 \|v_0\|_2^2 \leq \|v\|_2^2 \|v_0\|_2^2 \Rightarrow \|v\|_2^2 \leq \frac{\|v\|_2^2 \|v_0\|_2^2}{2\|v_0\|_2^2}$$

(14)

which completes the proof. □

Then we prove the convergence of the algorithm, which is described in the following theorem.

**Theorem 1.** The algorithm decreases the objective value of problem (8) in each iteration.
we have two following inequalities:

\[ \| x \|_F^2 \leq \sum_{i=1}^{K} \| x_i \|_2^2 \]

According to Step 4, we have:

\[ \| x \|_2^2 \leq \sum_{i=1}^{K} \| x_i \|_2^2 \]

Based on the definitions of \( D_j(1 \leq j \leq c_2) \) and \( D_i \), and Lemma 1, we have the following inequalities:

\[ \sum_{i=1}^{K} \| x_i \|_2^2 - \sum_{i=1}^{K} \| x_i \|_2^2 \leq \sum_{i=1}^{K} \| x_i \|_2^2 - \sum_{i=1}^{K} \| x_i \|_2^2 \]

\[ \sum_{i=1}^{K} \| x_i \|_2^2 - \sum_{i=1}^{K} \| x_i \|_2^2 \leq \sum_{i=1}^{K} \| x_i \|_2^2 - \sum_{i=1}^{K} \| x_i \|_2^2 \]

\[ \sum_{i=1}^{K} \| x_i \|_2^2 - \sum_{i=1}^{K} \| x_i \|_2^2 \leq \sum_{i=1}^{K} \| x_i \|_2^2 - \sum_{i=1}^{K} \| x_i \|_2^2 \]

Algorithm 1 An efficient iterative algorithm to solve the optimization problem in Equation (9).

**Input:**
\[ X = \{x_1, x_2, \ldots, x_c\} \in \mathbb{R}^{d \times c}, \quad Y = \left[ (y_1)^T, (y_2)^T, \ldots, (y_c)^T \right] \in \{0, 1\}^{c \times c} \quad \text{and} \quad Z = \left[ (z_1)^T, (z_2)^T, \ldots, (z_c)^T \right] \in \mathbb{R}^{d \times c}. \]

**Output:** \( W \in \mathbb{R}^{d \times c} \) and \( P \in \mathbb{R}^{d \times c} \).

1. Initialize \( W \in \mathbb{R}^{d \times c} \), \( P \in \mathbb{R}^{d \times c} \). Let \( V = [W, P] \in \mathbb{R}^{d(d+1)\times c} \).

**repeat**

2. Calculate the block diagonal matrices \( D_j(1 \leq j \leq c_2) \), where the \( k \)-th diagonal block of \( D_k \) is \( \frac{1}{\| x \|_2^2} \).

3. Update \( w \) by \( w \leftarrow B^{-1}w \), where the \( (u,p) \)-th element of \( w \) is given by \( (d*u-1)+u \leq u \leq u+d \), \( 1 \leq p \leq c_1 \), and \( \frac{1}{\| x \|_2^2} \).

4. Update the \( i \)-th column of \( P \) by \( p_i = (XX^T + \gamma_1 D_1 + \gamma_2 D_2)^{-1}x_i \).

5. Update the \( V \) by \( V = [W, P] \).

**until** Converges

then by adding Equations (15) (16) in both sides, we arrive at

\[ \mathcal{L}(\tilde{W}) + \mathcal{L}(\tilde{P}) + \gamma_1 \sum_{i=1}^{c_1+c_2} \sum_{k=1}^{d} \| x_k \|_2^2 + \gamma_2 \sum_{k=1}^{d} \| x_k \|_2^2 \]

\[ \leq \mathcal{L}(W) + \mathcal{L}(P) + \gamma_1 \sum_{i=1}^{c_1+c_2} \sum_{k=1}^{d} \| x_k \|_2^2 + \gamma_2 \sum_{k=1}^{d} \| x_k \|_2^2. \]

Therefore, the algorithm decreases the objective value of problem 8 in each iteration.

In the convergence, \( W, P, D_j(1 \leq j \leq c_2) \) and \( D \) satisfy the Equation (9). As the Equation (9) is a convex problem, satisfying the Equation (9) indicates that \( V = [W, P] \) is a global optimum solution to the Equation (9). Therefore, the Algorithm 1 will converge to the global optimum of the Equation (9). Since our algorithm has the closed form solution in each iteration, the convergence is very fast.

3 **EMPIRICAL STUDIES AND DISCUSSIONS**

Data used in the preparation of this article were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database. One goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical
and neuropsychological assessment can be combined to measure the progression of MCI and early AD. For up-to-date information, see www.adni-info.org. Following a prior imaging genetics study (Shen et al., 2010a), 733 non-Hispanic Caucasian participants were included in this study. We empirically evaluate the proposed method by applying it to the ADNI cohort, where a wide range of multimodal biomarkers are examined and selected to predict memory performance measured by five RAVLT scores and classify participants into HC, MCI and AD.

3.1 Experimental design

Overall setting: Our primary goal is to identify relevant genetic and imaging biomarkers that can classify disease status and predict memory scores (Fig. 2). We describe our genotyping, imaging and memory data in Section 3.1; present the identified biomarkers in Section 3.2; discuss the disease classification in Section 3.3 and demonstrate the memory score prediction in Section 3.4.

Genotyping data: The single-nucleotide polymorphism (SNP) data (Saykin et al. 2010b) were genotyped using the Human 610-Quad BeadChip (Illumina, Inc., San Diego, CA, USA). Among all SNPs, only SNPs belonging to the top 40 AD candidate genes listed on the AlzGene database (www.alzgene.org) as of June 10, 2010, were selected after the standard quality control (QC) and imputation steps. The QC criteria for the SNP data include (i) call rate check per subject and per SNP marker, (ii) gender check, (iii) sibling pair identification, (iv) the Hardy–Weinberg equilibrium test, (v) marker removal by the minor allele frequency and (vi) population stratification. The quality-controlled SNPs were then imputed using the MaCH software to estimate the missing genotypes. After that, the Illumina annotation information based on the Genome build 36.2 was used to select a subset of SNPs, belonging or proximal to the top 40 AD candidate genes. This procedure yielded 1224 SNPs, which were annotated with 37 genes (Wang et al., 2012). For the remaining 3 genes, no SNPs were available on the genotyping chip.

Imaging biomarkers: In this study, we use the baseline structural MRI and molecular FDG-PET scans, from which we extract imaging biomarkers. Two widely employed automated MRI analysis techniques were used to process and extract imaging genotypes across the brain from all baseline scans of ADNI participants as previously described (Shen et al., 2010a). First, voxel-based morphometry (VBM) (Ashburner and Friston, 2000) was performed to define global gray matter (GM) density maps and extract local GM density values for 86 target regions (Fig. 4a). Second, automated parcellation via FreeSurfer V4 (Fischl et al., 2004) was conducted to define 56 volumetric and cortical thickness values (Fig. 4b) and to extract total intracranial volume (ICV). Further information about these measures is available in Shen et al. (2010a). All these measures were adjusted for the baseline age, gender, education, handedness and baseline ICV using the regression weights derived from the healthy control participants. For PET images, following Landau et al. (2009), mean glucose metabolism (CMglu) measures of 26 regions of interest (ROIs) in the Montreal Neurological Institute (MNI) brain atlas space were employed in this study (Fig. 4c).

Memory data: The cognitive measures we use to test the proposed method are the baseline RAVLT memory scores from all ADNI participants. The standard RAVLT format starts with a list of 15 unrelated words (List A) repeated over five different trials and participants are asked to repeat. Then the examiner presents a second list of 15 words (List B), and the participant is asked to remember as many words as possible from List A, without reading it again. Trial 7, termed as 30 min recall, requests the participant again to recall as many words as possible from List B, without reading it again. Trial 7, termed as 30 min recall, requests the participant again to recall as many words as possible from List B, without reading it again. Trial 7, termed as 30 min recall, requests the participant again to recall as many words as possible from List B, without reading it again.}

<table>
<thead>
<tr>
<th>Task ID</th>
<th>Description of RAVLT scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOTAL</td>
<td>Total score of the first 5 learning trials</td>
</tr>
<tr>
<td>TOT6</td>
<td>Trial 6 total number of words recalled</td>
</tr>
<tr>
<td>TOTB</td>
<td>List B total number of words recalled</td>
</tr>
<tr>
<td>T50</td>
<td>30 minute delay total number of words recalled</td>
</tr>
<tr>
<td>RECOG</td>
<td>30 minute delay recognition score</td>
</tr>
</tbody>
</table>

Fig. 4. Weight maps for multimodal data: (a) VBM measures from MRI, (b) FreeSurfer measures from MRI, (c) glucose metabolism from FDG-PET, and (d) top SNP findings. Weights for disease classification were labeled as Diag-L (left side), Diag-R (right side) or Diag, and weights for RAVLT regression were labeled as A VLT-L, A VLT-R or A VLT. In (a–c), weights were normalized by dividing the corresponding threshold used for feature selection, and thus all selected features had normalized weights ≥1 and were marked with ‘x’. In (d), only top SNPs were shown, weights were normalized by dividing the weight of the 10th top SNP, and the top 10 SNPs for either classification or regression task had normalized weights ≥1 and were marked with ‘x’.
**3.2 Biomarker identifications**

The proposed heterogeneous multitask learning scheme aims to identify genetic and phenotypic biomarkers that are associated with both cognition (e.g. RAVLT in this study) and disease status in a joint regression and classification framework. Here we first examine the identified biomarkers. Shown in Figure 4 is a summarization of selected features for all four data types, where the regression/classification weights are color-coded for each feature and each task.

In Figure 4b, many VBM measures are selected to be associated with disease status, which is in accordance with known global brain atrophy pattern in AD. The VBM measures associated with RAVLT scores seem to be a subset of those disease-sensitive markers, showing a specific memory circuitry contributing to the disease, as well as suggesting that the disease is implicated by not only this memory function but also other complicated factors. Evidently, the proposed method could have a potential to offer deep mechanistic understandings. Shown in Figure 4c is a comparison between RAVLT relevant markers and AD-relevant markers and their associated weights mapped onto a standard brain space.

Figure 4b shows the identified markers from the FreeSurfer data. In this case, a small set of markers are discovered. These markers, such as hippocampal volume, amygdala volume and entorhinal cortex thickness, are all well-known AD-relevant markers, showing the effectiveness of the proposed method. These markers are also shown to be associated with both AD and RAVLT. The PDC-PET findings (Fig. 2b) are also interesting and promising. The AD-relevant biomarkers include angular, hippocampus, middle temporal and post cingulate regions, which agrees with prior findings e.g. 

As to the genetic, only top findings are shown in Figure 4b. The APOE E4 SNP (rs429358), the best known AD risk factor, shows the strongest link to both disease status and RAVLT scores. A few other important AD genes, including recently discovered and replicated PICALM and BIN1, are also included in the results. For those newly identified SNPs, further investigation in independent cohorts should be warranted.

**3.3 Improved disease classification**

We classify the selected participants of ADNI cohort using the proposed methods by integrating the four different types of data.

---

**Table 2. Multimodal feature sets as predictors in multiview learning**

<table>
<thead>
<tr>
<th>View ID (feature set ID)</th>
<th>Modality</th>
<th>No. of features</th>
</tr>
</thead>
<tbody>
<tr>
<td>VBM</td>
<td>MRI</td>
<td>86</td>
</tr>
<tr>
<td>FreeSurfer</td>
<td>MRI</td>
<td>56</td>
</tr>
<tr>
<td>FDG-PET</td>
<td>FDG-PET</td>
<td>26</td>
</tr>
<tr>
<td>SNP</td>
<td>Genetics</td>
<td>1244</td>
</tr>
</tbody>
</table>

**Participant selection:** In this study, we included only participants with no missing data for all above four types (views) of features and cognitive scores, which resulted in a set of 345 subjects (83 HC, 174 MCI and 88 AD). The feature sets extracted from baseline multimodal data of these subjects are summarized in Table 2.

---

**Fig. 5.** VBM weights of joint regression of AVLT scores and classification of disease status were mapped onto brain (a) Overall weights for disease classification, (b) Overall weights for AVLT regression
Table 3. Classification performance comparison between the proposed method and related methods for distinguishing HC, MCI and AD

<table>
<thead>
<tr>
<th>Methods</th>
<th>Accuracy (mean ± SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM (SNP)</td>
<td>0.561 ± 0.026</td>
</tr>
<tr>
<td>SVM (FreeSurfer)</td>
<td>0.573 ± 0.012</td>
</tr>
<tr>
<td>SVM (VBM)</td>
<td>0.541 ± 0.032</td>
</tr>
<tr>
<td>SVM (PET)</td>
<td>0.535 ± 0.026</td>
</tr>
<tr>
<td>SVM (all)</td>
<td>0.575 ± 0.019</td>
</tr>
<tr>
<td>HML (all)</td>
<td>0.638 ± 0.019</td>
</tr>
<tr>
<td>SVM $\ell_1$ MKL method</td>
<td>0.624 ± 0.031</td>
</tr>
<tr>
<td>SVM $\ell_2$ MKL method</td>
<td>0.593 ± 0.042</td>
</tr>
<tr>
<td>SVM $\ell_3$ MKL method</td>
<td>0.561 ± 0.037</td>
</tr>
<tr>
<td>LSSVM $\ell_0$ MKL method</td>
<td>0.614 ± 0.031</td>
</tr>
<tr>
<td>LSSVM $\ell_1$ MKL method</td>
<td>0.585 ± 0.018</td>
</tr>
<tr>
<td>LSSVM $\ell_2$ MKL method</td>
<td>0.577 ± 0.033</td>
</tr>
<tr>
<td>Our method (SNP)</td>
<td>0.673 ± 0.021</td>
</tr>
<tr>
<td>Our method (FreeSurfer)</td>
<td>0.688 ± 0.020</td>
</tr>
<tr>
<td>Our method (VBM)</td>
<td>0.669 ± 0.031</td>
</tr>
<tr>
<td>Our method (PET)</td>
<td>0.621 ± 0.028</td>
</tr>
<tr>
<td>Our method</td>
<td>0.726 ± 0.032</td>
</tr>
</tbody>
</table>

Table 4. Comparison of memory prediction performance measured by average RMSEs (smaller is better)

<table>
<thead>
<tr>
<th>Test case</th>
<th>TOTAL</th>
<th>TOT6</th>
<th>TOTB</th>
<th>T30</th>
<th>RECOG</th>
</tr>
</thead>
<tbody>
<tr>
<td>MRV (SNP)</td>
<td>6.153</td>
<td>2.476</td>
<td>2.168</td>
<td>2.201</td>
<td>3.483</td>
</tr>
<tr>
<td>MRV (FreeSurfer)</td>
<td>5.928</td>
<td>2.235</td>
<td>2.039</td>
<td>2.088</td>
<td>3.339</td>
</tr>
<tr>
<td>MRV (VBM)</td>
<td>6.095</td>
<td>2.289</td>
<td>2.142</td>
<td>2.137</td>
<td>3.394</td>
</tr>
<tr>
<td>MRV (PET)</td>
<td>6.246</td>
<td>2.514</td>
<td>2.237</td>
<td>2.215</td>
<td>3.615</td>
</tr>
<tr>
<td>MRV (all)</td>
<td>5.909</td>
<td>2.232</td>
<td>1.992</td>
<td>2.032</td>
<td>3.306</td>
</tr>
<tr>
<td>Ridge (SNP)</td>
<td>6.076</td>
<td>2.416</td>
<td>2.147</td>
<td>2.117</td>
<td>3.368</td>
</tr>
<tr>
<td>Ridge (FreeSurfer)</td>
<td>5.757</td>
<td>2.203</td>
<td>2.004</td>
<td>2.017</td>
<td>3.237</td>
</tr>
<tr>
<td>Ridge (VBM)</td>
<td>5.976</td>
<td>2.147</td>
<td>2.038</td>
<td>2.129</td>
<td>3.249</td>
</tr>
<tr>
<td>Ridge (PET)</td>
<td>6.153</td>
<td>2.443</td>
<td>2.186</td>
<td>2.107</td>
<td>3.515</td>
</tr>
<tr>
<td>Ridge (all)</td>
<td>5.704</td>
<td>2.143</td>
<td>1.989</td>
<td>1.994</td>
<td>3.193</td>
</tr>
<tr>
<td>Our method (SNP)</td>
<td>5.991</td>
<td>2.201</td>
<td>2.008</td>
<td>2.001</td>
<td>3.107</td>
</tr>
<tr>
<td>Our method (FreeSurfer)</td>
<td>5.601</td>
<td>2.106</td>
<td>1.947</td>
<td>1.886</td>
<td>3.015</td>
</tr>
<tr>
<td>Our method (VBM)</td>
<td>5.715</td>
<td>2.011</td>
<td>1.899</td>
<td>1.974</td>
<td>3.041</td>
</tr>
<tr>
<td>Our method (PET)</td>
<td>6.013</td>
<td>2.241</td>
<td>2.017</td>
<td>2.017</td>
<td>3.331</td>
</tr>
<tr>
<td>Our method (all)</td>
<td>5.506</td>
<td>1.984</td>
<td>1.886</td>
<td>1.841</td>
<td>2.989</td>
</tr>
</tbody>
</table>

From Table 4 we can see that the proposed method always has better memory prediction performance. Among the test cases, the FreeSurfer imaging measures and VBM imaging measure have similar predictive power, which are better than those of PET imaging measures and SNP features. In general, combining the four types of features are better than only using one type of data. Since our method adaptively weight each type of data and each feature inside a type of data, it has the least regression error when using all available input data. These results, again, demonstrated the usefulness of our method and data integration in early AD diagnosis.

3.4 Improved memory performance prediction

Now we evaluate the memory performance prediction capability of the proposed method. Since the cognitive scores are continuous, we evaluate the proposed method via regression and compare it to two baseline methods, i.e. multivariate linear regression (MRV) and ridge regression. Since both MRV and ridge regression are for single-type input data, we conduct regression on each of the four types of features and a simple concatenation of them. Similarly, we also predict memory performance by our method on the same test conditions. When multiple-type input data are used, as demonstrated in Section 3.2, our method automatically and adaptively select the prominent biomarkers for regression. For each test case, we conduct standard 5-fold cross-validation and report the average results. For each of the five trials, within the training data, an internal 5-fold cross-validation is performed to fine tune the parameters in the range of $10^{-5}, 10^{-4}, \ldots, 10^5, 10^6$ for both ridge regression and our method. For our method, in each trial, from the learned coefficient matrix we sum the absolute values of the coefficients of a single feature over all the tasks as the overall weight, from which we pick up the features with non-zero weights (i.e. $w > 10^{-5}$) to predict regression responses for test data. The performance assessed by root mean square error (RMSE), a widely used measurement for statistical regression analysis, are reported in Table 4.

4 CONCLUSIONS

We proposed a novel sparse multimodal multitask learning method to identify the disease-sensitive biomarkers via integrating heterogeneous imaging genetics data. We utilized the joint classification and regression learning model to identify the disease-sensitive and QT-relevant biomarkers. We introduced a novel combined structured sparsity regularization to integrate heterogeneous imaging genetics data, and derived a new efficient optimization algorithm to solve our non-smooth objective function and followed with the rigorous theoretical analysis on the global convergency. The empirical results showed our method improved both memory scores prediction and disease classification accuracy.

ACKNOWLEDGEMENT

Data used in preparation of this article were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.
Funding. [This research was supported by National Science Foundation Grants CCF-0830780, CCF-0917274, DMS-0915228, and IIS-1111795] at UTA; and by [National Science Foundation Grant IIS-111735] to The National Institute of Health Grants UL1 RR025761, U01 AG024904, NIA RC2 AG036535, NIA R01 AG19771, and NIA R01 AG10133-18S1] at IU.

Data collection and sharing for this project was funded by the Alzheimer’s Disease Neuroimaging Initiative (ADNI) ([National Institutes of Health Grant U01 AG024904], ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott; Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; Amorfix AG19771, and NIA P30 AG10133-18S1] at IU.


Dhillon, I. S. et al. (2003) Information-theoretic clustering. In ACM SIGKDD (Special Interest Group on Knowledge Discovery and Data Mining) International Conference on Knowledge Discovery and Data Mining, pp. 89–98.


H. Wang et al.


