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

Permalink
https://escholarship.org/uc/item/6fw5h15n

Journal
Bioinformatics, 28(12)

ISSN
1367-4811

Authors
Wang, Hua
Nie, Feiping
Huang, Heng
et al.

Publication Date
2012-06-09

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

Hua Wang 1, Feiping Nie 1, Heng Huang 1,*, Shannon L. Risacher 2, Andrew J. Saykin 2 and Li Shen 1,*,†; For the Alzheimer's Disease Neuroimaging Initiative 1
1 Department of Computer Science and Engineering, University of Texas at Arlington, TX 76019 and 2 Department 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 multidimensional 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 convergence.

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, and ignore the important underlying interacting relationships between these units.

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

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, Pan et al. 2008, Huang et al. 2009, Shen et al. 2010). 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 (Schreiber 2002, Brefeld and Scheffer 2004, Schölkopf and Platt 2002).
The structured sparsity is usually obtained through different sparse problems. From the view of sparsity organization, we have two by imposing non-smooth norms as regularizers in the optimization
Kim and Xing, 2010; Micchelli
and applied to multitask learning models (Argyriou
simultaneously.
biomarkers correlated to memory scores and disease categories
and regression multitask learning model is utilized to select the
important role in influencing multiple tasks. A joint classification
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
biomarkers from multiple data sources. Different to LASSO
efficiently to identify disease-sensitive and cognition-relevant
sparse multimodal multitask learning algorithm that integrates
heterogeneous data and select multitype features. However, such
models train a single weight for all features from the same modality,
i.e. all features from the same data source are weighted equally,
when they are combined with the features from other sources. This
limitation often yields inadequate performance.

To address the above challenges, we propose a new sparse multimodal multitask learning algorithm that integrates heterogeneous genetic and phenotypic data effectively and efficiently to identify disease-sensitive and cognition-relevant biomarkers from multiple data sources. Different to LASSO
Tibshirani, 1996), group LASSO (Yuan and Lin, 2006; Ye et al., 2008), and other related methods that mainly find the biomarkers correlated to each individual QT (memory score), we consider predicting each memory score as a regression task and select biomarkers that tend to play an important role in influencing multiple tasks. A joint classification
and regression multitask learning model is utilized to select the
biomarkers correlated to memory scores and disease categories
simultaneously.

Sparsity regularizations have recently been widely investigated and applied to multitask learning models (Argyriou et al., 2006;
Kim and Xing, 2010; Micchelli et al., 2010; Sun et al., 2010). Sparse representations are typically achieved by imposing non-smooth norms as regularizers in the optimization
problems. From the view of sparsity organization, we have two
types: (i) The flat sparsity is often achieved by \( \ell_0 \)-norm or \( \ell_1 \)-norm regularizer or trace norm in matrix/sensor completion. Optimization techniques include LARS (Efron et al., 2004), linear gradient search (Lu et al., 2009), 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; Obozinski et al., 2010; Sun et al., 2009), \( \ell_2,0 \)-norm (Sun et al., 2010), \( \ell_{\infty,0} \)-norm (Quattoni et al., 2008) (also denoted as \( \ell_{1,0} \)-norm in different papers) and group \( \ell_1 \)-norm (Ye and Liu, 2008) which can be solved by methods in (Micchelli et al., 2010) and
Argyriou et al., 2009). We propose a new combined structured sparse
regularization to integrate features from different modalities and to
learn a weight for each feature leading to a more flexible scheme for
feature selection in data integration, which is illustrated in Figure 3.
In our combined structured sparse regularization, the group \( \ell_1 \)-norm regularization (blue circles in Fig. 3) learns the feature global
importance, i.e. the modal-wise feature importance of every data
modality on each class (task), and the \( \ell_2,1 \)-norm regularization (red circles in Fig. 3) explores the feature local importance, i.e. the
importance of each feature for multiple classes/tasks. The proposed method is applied to identify AD-sensitive biomarkers associated
with the cognitive scores by integrating heterogeneous genetic and
phenotypic data (as shown in Fig. 3). Our empirical results yield
clearly improved performance on predicting both cognitive scores
and disease status.

2 IDENTIFYING DISEASE SENSITIVE AND QT-RELEVANT BIOMARKERS FROM HETEROGENEOUS IMAGING GENETICS DATA

Pairwise univariate correlation analysis can quickly provide important association information between genetic/phenotypic data
data and QTs. However, it treats the features and the QTs as independent and isolated units, therefore the underlying interacting
relationships between the units might be lost. We propose a new
sparse multimodal multitask learning model to reveal genetic
phenotypic biomarkers, which are disease sensitive and QT-relevant, by simultaneously and systematically taking into account an ensemble of SNPs (single nucleotide polymorphism) and phenotypic signatures and jointly performing two heterogeneous tasks, i.e.
biomarker-to-QT regression and biomarker-to-disease classification.
The QTs studied in this article are the cognitive scores.

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,
2009). Such multitask learning can discover robust 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 \( ||W||_{2,1} = \sum_{j=1}^{n} ||w^i||_2 \) (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 \( x_i = (x_1^i, \ldots, x_d^i)^T \in \mathbb{R}^d \) is the input vector including all features from a total of \( k \) different modalities and each modality \( j \) has \( d_j \) features (\( d = \sum_{j=1}^{k} d_j \), \( y_i \) 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 \( Y = [y_1, \ldots, y_n] \in \mathbb{R}^{c \times n} \). Different to MKL, we directly learn a \( d \times c \) parameter matrix as:

\[
W = \begin{bmatrix} w_1^1 & \ldots & w_1^c \\
\vdots & \ddots & \vdots \\
 w_n^1 & \ldots & w_n^c \end{bmatrix} \in \mathbb{R}^{d \times nc},
\]

where \( w^i_j \in \mathbb{R}^{d_j} \) 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 weights of each feature to capture the local feature importance. Since the features come from heterogeneous data sources, we impose the regularizer \( R(W) \) to capture the interrelationships of modalities and features as:

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

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 (4), which is defined over \( W \) as following:

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

which is illustrated by the blue circles in Figure 3. Then the Equation (4) becomes:

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

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 \( \ell_1 \)-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:

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

The \( \ell_2,1 \)-norm was popularly used in multitask feature selection (Argyriou et al., 2008; Obozinski et al., 2010). Since 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.

Our regularization items consider the heterogeneity features from both group-wise and individual viewpoints. Figure 4 visualizes the matrix \( W^T \) 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., 2014). A similar model was used in (Yang et al., 2008) for
heterogeneous multitask learning. Regular multitask learning only considers homogenous 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:

\[ \mathcal{L}_1(W) = \sum_i^{n} \sum_{k=1}^{c_i} \left( y_{ik} \log \sum_{l=1}^{c_i} e^{w_i^T x_{il}} - y_{ik} w_i^T x_{ik} \right) \]

Here, we perform three binary classification tasks for the following three diagnostic groups respectively (\(c_i=3\)): 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 \(P \in \mathbb{R}^{d \times c_i}\), the least squares loss is defined by

\[ \mathcal{L}_2(P) = \|X^T P - Z\|_F^2 \]

where \(X\) is the data points matrix, \(P\) is the coefficient matrix of regression with \(c_2\) tasks, the label matrix \(Z = \begin{bmatrix} (z_1^T)^T, (z_2^T)^T, \ldots, (z_{c_2}^T)^T \end{bmatrix} \in \mathbb{R}^{n \times c_2}\).

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

\[ \min_V \sum_i^{n} \sum_{k=1}^{c_i} \left( y_{ik} \log \sum_{l=1}^{c_i} e^{w_i^T x_{il}} - y_{ik} w_i^T x_{ik} \right) + \|X^T P - Z\|_F^2 \|

\]

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

Since the objective in Equation (3) 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 (3) with respect to \(W\) and \(P\) respectively, and set them to zeros, we have

\[ \frac{\partial \mathcal{L}_1(W)}{\partial W} + 2 \gamma_1 \sum_{i=1}^{n} D_i W_i + 2 \gamma_2 D P = 0, \]

\[ 2X^T P - 2XZ + 2\gamma_1 \sum_{i=1}^{n} D_i P_i + 2 \gamma_2 D P = 0, \]

where \(D_i(1 \leq i \leq c_1 + c_2)\) is a block diagonal matrix with the \(k\)-th diagonal block as \(D_k \in \mathbb{R}^{d_k \times d_k}\), \(P\) is a diagonal matrix with the \(k\)-th diagonal element as \(1/d_k\).

Since \(D_i(1 \leq i \leq c_1 + c_2)\) and \(D\) depend on \(V = \begin{bmatrix} P \end{bmatrix}\), they are also unknown variables to be optimized. In this article, we provide an iterative algorithm to solve Equation (3). First, we guess a random solution \(V = \begin{bmatrix} P \end{bmatrix} \in \mathbb{R}^{d \times (c_1 + c_2)}\), then we calculate the matrices \(D_i(1 \leq i \leq c_1 + c_2)\) and \(D\) according to the current solution \(V\). After obtaining the \(D_i(1 \leq i \leq c_1 + c_2)\) and \(D\), we can update the solution \(V = \begin{bmatrix} P \end{bmatrix}\) based on Equation (3). Specifically, the \(j\)-th column of \(P\) is updated by \(W_j = (X^T + \gamma_1 D_i + \gamma_2 D)^{-1} X_k\), we cannot update \(W\) with a closed form solution based on Equation (3), but we can obtained the updated \(W\) by the Newton’s method. According to Equation (3), we need to solve the following problem:

\[ \min_W \mathcal{L}_1(W) + \gamma_1 \sum_{i=1}^{n} w_i^T D_i w_i + \gamma_2 Tr(W^T DW) \]

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

\[ \frac{\partial}{\partial w_{ip}} = 2D_p(u, w)W_{ip}, \]

\[ \frac{\partial}{\partial w_{iqp}} = 2D_p(u, w)W_{iqp}, \]

where \(D_p(u, w)\) is the \(u\)-th diagonal element of \(D_p\).

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

\[ \frac{\partial}{\partial w_{ip}} = 2D_p(u, w)W_{ip}, \]

\[ \frac{\partial}{\partial w_{ip}} = 2D_p(u, w)W_{ip}, \]

After obtaining the updated solution \(V = \begin{bmatrix} P \end{bmatrix}\), we can calculate the new matrices \(D_i(1 \leq i \leq c_1 + c_2)\) and \(D\). This procedure is repeated until the algorithm converges. The detailed algorithm is listed in Algorithm 4. 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 \(w\), we have the following inequality:

\[ \|v\|_2^2 - \frac{\|v\|_2^2}{2\|w\|_2} \leq \|v\|_2 - \frac{\|v\|_2}{2\|w\|_2}, \]

**Proof.** Obviously, \(-(\|v\|_2^2 - \|w\|_2^2)^2 \geq 0\), so we have

\[ (\|v\|_2 - \|w\|_2)^2 \leq 0 \Rightarrow (\|v\|_2 - \|w\|_2)^2 \leq \|v\|_2^2 \]

\[ \Rightarrow \|v\|_2 - \frac{\|v\|_2^2}{2\|w\|_2} \leq \|v\|_2 - \frac{\|v\|_2}{2\|w\|_2}, \]

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 (3) in each iteration.
we have two following inequalities:

\[\sum_{i=1}^{c_1} \sum_{j=1}^{k} w_i^T D_j w_j + \gamma_2 \text{Tr}(W^T D W)\]

\[\leq \mathcal{L}(W) + \gamma_1 \sum_{i=1}^{c_1} \sum_{j=1}^{k} (\tilde{w}_i^T D_j \tilde{w}_j + \gamma_2 \text{Tr}(W^T D W))\]

According to Step 4, we have:

\[\left| X^T P - Y \right|^2_F + \gamma_1 \sum_{i=1}^{c_1} \sum_{j=1}^{k} (\tilde{w}_i^T D_j \tilde{w}_j + \gamma_2 \text{Tr}(P^T D P))\]

\[\leq \left| X^T P - Y \right|^2_F + \gamma_1 \sum_{i=1}^{c_1} \sum_{j=1}^{k} (\tilde{w}_i^T D_j \tilde{w}_j + \gamma_2 \text{Tr}(P^T D P)).\]

Based on the definitions of \(D_j(1 \leq i \leq c_1 + c_2)\) and \(D_i\), and Lemma 1, we have two following inequalities:

\[\sum_{i=1}^{c_1} \sum_{j=1}^{k} \| \tilde{w}_i^T D_j \tilde{w}_j \|_2 - \gamma_1 \sum_{i=1}^{c_1} \sum_{j=1}^{k} \tilde{w}_i^T D_j \tilde{w}_j \leq \sum_{i=1}^{c_1} \sum_{j=1}^{k} \| \tilde{w}_i^T D_j \|_2 - \gamma_1 \sum_{i=1}^{c_1} \sum_{j=1}^{k} \tilde{w}_i^T D_j \tilde{w}_j\]

\[\Rightarrow \gamma_1 \sum_{i=1}^{c_1} \sum_{j=1}^{k} \| \tilde{w}_i^T D_j \|_2 - \gamma_1 \sum_{i=1}^{c_1} \sum_{j=1}^{k} \tilde{w}_i^T D_j \tilde{w}_j \leq \gamma_1 \sum_{i=1}^{c_1} \sum_{j=1}^{k} \| \tilde{w}_i^T D_j \|_2 - \gamma_1 \sum_{i=1}^{c_1} \sum_{j=1}^{k} \tilde{w}_i^T D_j \tilde{w}_j,\]

\[\Rightarrow \text{and} \]

\[\sum_{i=1}^{d} \| \tilde{v}_1^T D_j \tilde{v}_1 \|_2 - \sum_{i=1}^{d} \| \tilde{v}_1^T D_j \|_2 = \sum_{i=1}^{d} \tilde{v}_1^T D_j \tilde{v}_1 \]

\[\Rightarrow \gamma_2 \sum_{i=1}^{d} \tilde{v}_1^T D_j \tilde{v}_1 \quad \text{and} \quad \gamma_2 \sum_{i=1}^{d} \| \tilde{v}_1^T D_j \|_2 \]

\[\leq \gamma_2 \sum_{i=1}^{d} \| \tilde{v}_1^T D_j \|_2 - \gamma_2 \text{Tr}(V^T D V)\]

\[\leq \gamma_2 \sum_{i=1}^{d} \| \tilde{v}_1^T D_j \|_2 - \gamma_2 \text{Tr}(V^T D V).\]

Note that the following two equalities:

\[\sum_{i=1}^{c_1} \sum_{j=1}^{k} \tilde{w}_i^T D_j \tilde{w}_j = \sum_{i=1}^{c_1} \sum_{j=1}^{k} \tilde{w}_i^T D_j \tilde{w}_j + \sum_{i=1}^{c_1} \sum_{j=1}^{k} \tilde{w}_i^T D_j \tilde{w}_j,\]

\[\text{Tr}(V^T D V) = \text{Tr}(W^T D W) + \text{Tr}(P^T D P).\]
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., 2010b), 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. 4). We describe our genotyping, imaging and memory data in Section 3.2. First, voxel-based morphometry (VBM) (Ashburner and Friston, 2000) was performed as previously described (Shen et al., 2010b) across the brain from all baseline scans of ADNI participants. 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., 2010b). 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., 2002) 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 space were employed in this study (Fig. 4b).

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 A, without reading it again. Trial 7, termed as 30 min recall, is administered in the same way as Trial 6, but after a 30 min delay. Finally, a recognition test with 30 words read aloud, requesting the participant to indicate whether or not each word is on List A. The RAVLT has proven useful in evaluating verbal learning and memory. Table 1 summarizes five RAVLT scores used in our experiments.

Table 1. RAVLT cognitive measures as responses in multitask learning

<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>TOT8</td>
<td>List 8 total number of words recalled</td>
</tr>
<tr>
<td>T30</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’.

H. Wang et al.
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 summarized selection of features for all four data types, where the regression/classification weights are color-mapped for each feature and each task.

In Figure 4a, 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 4b 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 findings (Fig. 4c). They 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. Landau et al. (2009). Again, a subset of these markers are also relevant to RAVLT scores.

As to the genetics, only top findings are shown in Figure 4d. 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.

### 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>

### 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.
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.

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.
Multidimensional imaging genetics data integration

Funding: [This research was supported by National Science Foundation Grants CCF-0830780, CCF-0917274, DMS-0915228, and IIS-1117965] at UTA; and by [National Science Foundation Grant IIS-1117335, National Institutes of Health Grants U1L RR025761, U01 AG024904, NIA RC2 AG036355, NIA R01 AG19771, and NIA P30 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 Life Sciences Ltd.; AstraZeneca; Bayer HealthCare; BioClinica, Inc.; Biogen Idec Inc.; Bruker; CalTech; Covance; Eisai Inc.; Elan Pharmaceuticals Inc.; Eli Lilly and Company; F Hoffmann-La Roche Ltd and its affiliate company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Mesoscale Diagnostics, LLC.; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer’s Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles. This research was also supported by NIH [P30 AG010129, K01 AG030514] and the Dana Foundation.

Conflict of Interest: none declared.

REFERENCES


Dhillon,I.S. et al. (2003) Information-theoretic covering. 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.
