Predicting Cognitive Declines Using Longitudinally Enriched Representations for Imaging Biomarkers*

Lyujian Lu, Hua Wang†, Saad Elbeleidy
Department of Computer Science,
Colorado School of Mines, Golden, Colorado 80401, U.S.A.
lyujianlu@mymail.mines.edu, huawangcs@gmail.com, selbeleidy@mymail.mines.edu
Feiping Nie
School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL),
Northwestern Polytechnical University, Xi'an 710072, P. R. China
feipingnie@gmail.com
for the Alzheimer's Disease Neuroimaging Initiative

Abstract

With rapid progress in high-throughput genotyping and neuroimaging, researches of complex brain disorders, such as Alzheimer's Disease (AD), have gained significant attention in recent years. Many prediction models have been studied to relate neuroimaging measures to cognitive status over the progressions when these disease develops. Missing data is one of the biggest challenge in accurate cognitive score prediction of subjects in longitudinal neuroimaging studies. To tackle this problem, in this paper we propose a novel formulation to learn an enriched representation for imaging biomarkers that can simultaneously capture both the information conveyed by baseline neuroimaging records and that by progressive variations of varied counts of available follow-up records over time. While the numbers of the brain scans of the participants vary, the learned biomarker representation for every participant is a fixed-length vector, which enable us to use traditional learning models to study AD developments. Our new objective is formulated to maximize the ratio of the summations of a number of ℓ_1-norm distances for improved robustness, which, though, is difficult to efficiently solve in general. Thus we derive a new efficient iterative solution algorithm and rigorously prove its convergence. We have performed extensive experiments on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset. A performance gain has been achieved to predict four different cognitive scores, when we compare the original baseline representations against the learned representations with enrichments. These promising empirical results have demonstrated improved performances of our new method that validate its effectiveness.

1. Introduction

As one of the most prevalent and severe type of neurodegenerative disorders, Alzheimer's Disease (AD) strongly impacts human memory, thinking and behavior, which is characterized by progressive impairment of memory and other cognitive capabilities, triggered by the damage of neurons. AD usually progresses along a temporal continuum, initially from a pre-clinical stage, subsequently to mild cognitive impairment (MCI) and ultimately deteriorating to AD [36]. According to a recent report [1], AD is the sixth leading cause of death in the United States. It is estimated that 5.7 million individuals are living with AD and this number is projected to grow to 13.8 million by mid-century, fueled in large part by the aging of the Baby Boom Generation. The number of AD sufferers worldwide is estimated to be 44 million now and 1 in 85 people will be affected by AD by 2050 [1].

With all these facts, AD has attracted growing attentions in recent years. Over the past decade, neuroimaging measures have been widely studied to predict disease status and/or cognitive performance [8, 30, 31, 24, 40, 29, 22, 38]. However, these approaches routinely perform standard re-

*Data used in preparation of this article were obtained from the ADNI database (adni.loni.usc.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: https://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

†Corresponding author. This work was supported in part by the National Science Foundation (NSF) under the grants of IIS 1652943, IIS 1849359 and CNS 1932482.

2575-7075/20/$31.00 ©2020 IEEE
DOI 10.1109/CVPR42600.2020.00488

gression and/or classification at all time points separately, which thereby ignore the longitudinal variations of brain phenotypes. Since AD is a progressive neurodegenerative disorder, it would be beneficial to explore the temporal relation among the longitudinal records of the biomarkers.

In the study of the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort, participants are followed up at various time points, including the baseline (BL), the 6th Month (M6), the 12th month (M12), the 18th month (M18), the 24th month (M24), and the 36th month (M36), which provides the possibility for developing more effective predictive models by using longitudinal data from multiple time points. To explore the temporal structure of brain phenotypes, longitudinal prediction models have been proposed [33, 32, 34, 35, 17, 5, 4, 3] in recent years. However, in these studies longitudinal information has been modeled as tensors, which inevitably complicates the problem. As a result, it is not easy to extend classical machine learning models that can only deal with vector or matrix data to study AD developments.

Missing data in medical records is another critical challenge when we study the longitudinal data. Higher mortality risk and cognitive impairment hinder older adults from staying in studies that require multiple visits and thus result in incomplete data [37, 39]. The missing imaging records in the longitudinal medical data lead to samples with varied lengths for different participants. To deal with this problem, many existing AD studies using longitudinal medical data choose to only use data samples with complete temporal records for model analysis and ignore those with few time points [33, 35, 17]. Apparently, discarding the samples with less temporal records could potentially ruin the data set. To address this, data imputation methods [14, 37, 39] have been proposed to handle the missing records of the longitudinal AD measures. With imputed data, regression and classification studies can be conducted. However, whether or not these data completion methods could preserve the longitudinal structure of neuroimaging measurements is still an under-explored topic in AD studies. What's worse, these missing data imputation methods could possibly introduce undesirable artifacts, thereby possibly further worsen the predictive power of the longitudinal learning models.

To solve the longitudinal prediction problem with incomplete temporal inputs, in this study we propose a novel formulation to learn an enriched biomarker representation which combines the baseline biomarker measurements and the dynamic temporal imaging measurements across the following time points. In our enriched biomarker representation learning framework, we use the biomarker records at all available time points (a subset of {M6, M12, M18, M24, M36}) of each participant, from which we learn a projection that can map the baseline record into a lower-dimensional fixed-length vector, regardless of the inconsistent sizes of the medical records of the participants in a data set. Armed with the fixed-length biomarker representations, we can take advantage of conventional regression and/or classification methods to predict the cognitive declines of AD patients.

In our proposed framework to learn the enriched sample representations, it first learns a projection from the available follow-up imaging records. It then applies the learned projection to the baseline neuroimaging record to compute a fixed-length enriched biomarker representation. Through these procedures, the learned representation simultaneously captures the information conveyed by both baseline neuroimaging record and the progressive summary of all available follow-up records of each participant. We further develop the proposed objective by replacing the squared ℓ_2-norm distances by the ℓ_1-norm distances in our formulation, to improve the robustness of the learned enriched representation against possible outlying samples caused by varied numbers of the brain scans of the participants in the studied cohort.

Despite its clear motivation to integrate the information from both baseline neuroimaging records and the available follow-up ones, the proposed objective ends up to be an optimization problem that simultaneously maximizes and minimizes the summations of a number of ℓ_1-norm distances. To solve this challenging optimization problem, we derive an efficient non-greedy iterative algorithm with theoretically guaranteed convergence.

Extensive experiments have been performed on the ADNI cohort that demonstrate the improved performance resulting from our new approach. We first compare the prediction power of the baseline biomarker representations against its enriched counterparts obtained by learning using five different broadly used prediction models: linear regression (LR), ridge regression (RR), Lasso [25], support vector regression (SVR) [23] and convolutional neural networks (CNN) [2]. We achieve a clear performance gain on the four cognitive scores on the voxel-based morphometry (VBM) biomarkers, which validate the effectiveness of the our proposed method.

In the remainder of this paper, we will first introduce the optimization objective of our new learning model to learn the projections to enrich the baseline imaging biomarker representations in Section 2, followed by the mathematical derivations of an iterative algorithm to solve the proposed objective and the convergence analysis of the algorithm in Section 3. Then we report the experimental results in our comprehensive empirical studies that support our hypothesis in Section 4. Finally, the paper is concluded in Section 5.

2. The objective of our new method

In this section, we will first formalize problem to learn the enriched neuroimaging biomarker representations,

where we will introduce the notations used in this paper. Then we will gradually develop the proposed objective to learn a single fixed-length vector representation that can simultaneously capture the information from both baseline neuroimaging record and progressive changes of follow-up records along all time points.

2.1. Notations

Throughout this paper, we will write matrices as bold uppercase letters and vectors as bold lowercase letters. The trace of the matrix $\mathbf{M} = [m_{ij}]$ is defined as $\mathrm{tr}(\mathbf{M}) = \sum_i m_i$. The ℓ_1-norm of a vector $\mathbf{v}$ is defined as $\|\mathbf{v}\|_1 = \sum_i |v_i|$ and the ℓ_2-norm of v is defined as $\|\mathbf{v}\|_2 = \sqrt{\sum_i v_i^2}$.

2.2. Problem formalization

In the task of predicting cognitive declines using the ADNI dataset, we aim at learning a fixed-length biomarker representation vector for every participant from both the baseline neuroimaging record and all available follow-up medical scans of the participant. We denote the neuroimaging measures of each participant as: $\mathcal{X} = \{\mathbf{x}, \mathbf{X}\}$. Here, $\mathbf{x} \in \Re^d$ is the biomarker representations of the participant at the baseline time point, where d denotes the number of the neuroimaging features; $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \Re^{d\times n}$ collects all available follow-up biomarker records at each time point in the later three years, where n denotes the number of available numbers of neuroimaging records of the studied participant.

Given the neuroimaging sample $\mathcal{X}$ of a participant in a studied cohort, we aim to learn an enriched representation of $\mathbf{y} = f(\mathcal{X})$ that captures information from both the baseline neuroimaging record and the progressive changes along each time point. To be more specific, first from $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \Re^{d\times n}$ we intend to learn a projection which summarizes the temporal variations of neuroimaging records along all time points that follow the baseline time point: $\mathbf{W} = g(\mathbf{X})$. Then by applying the learned projection $\mathbf{W}$ on the baseline neuroimaging record, we compute a single fixed-length vector representation as following:

$$\mathbf{y} = f(\mathcal{X}) = f(g(\mathbf{X}), \mathbf{x}). \tag{1}$$

We hope the learned representation $\mathbf{y}$ cna simultaneously capture the information conveyed by boht baseline neuroimaging record and the dynamic changes of follow-up neuroimaging records. Because such learned biomarker representations across samples of the dataset are of the same length, they can be readily used by traditional learning models for a variety of analysis tasks to study cognitive declines.

2.3. Representation learning through projections

In this subsection, we will develop the proposed objective to learn a new single fixed-length vector representation for the neuroimaging records. By integrating the baseline neuroimaging record and dynamic temporal changes of follow-up neuroimaging records, we aim to preserve the global and local consistencies among the neuroimaging records in the learned projected subspace.

Usually healthy participants (marked as healthy control (HC) in the ADNI dataset) and most patients diagnosed with impairment will remain cognitively stable within 4–6 years [6]. Namely, the neuroimaging measurements of the participants will not experience drastic changes over a short time. Thus, we aim to preserve this local consistency in the projected space via minimizing the local variance of records among nearby months in the projected subspace. Mathematically, we denote the K-nearest neighbors of $\mathbf{x}_i$ as $\mathcal{N}_i$ and the local mean vector of $\mathbf{x}_i$ as:

$$\overline{\mathbf{x}}_i = \frac{1}{K+1} \sum_{\mathbf{x}_j \in \{\mathcal{N}_i \cup \{\mathbf{x}_i\}\}} \mathbf{x}_j. \tag{2}$$

We can achieve the overall local consistency of the entire dataset by minimizing the following objective [27]:

$$\begin{aligned} \mathcal{J}_{\text{Local}}(\mathbf{W}) &= \mathrm{tr}\left(\mathbf{W}^T \mathbf{S}_L \mathbf{W}\right), \\ s.t.\ \ \mathbf{W}^T\mathbf{W} &= \mathbf{I}, \end{aligned} \tag{3}$$

where $\mathbf{S}_L$ is defined as:

$$\mathbf{S}_L = \sum_{i=1}^{n} \mathbf{S}_{Li}, \tag{4}$$

and

$$\mathbf{S}_{Li} = \sum_{\mathbf{x}_j \in \{\mathcal{N}_i \cup \{\mathbf{x}_i\}\}} (\mathbf{x}_j - \overline{\mathbf{x}}_i)(\mathbf{x}_j - \overline{\mathbf{x}}_i)^T. \tag{5}$$

Obviously, as discussed in our earlier work in [27], $\mathbf{S}_{Li}$ computes the local covariance matrix of the data points around $\mathbf{x}_i$. Thus minimizing $\mathrm{tr}\left(\mathbf{W}^T\mathbf{S}_{Li}\mathbf{W}\right)$ ensures the local consistency around $\mathbf{x}_i$ and minimizing $\mathcal{J}_{\text{Local}}$ in Eq. (3) ensures the overall local consistency around all data samples. The constant factor $\frac{1}{K+1}$ is omitted in Eqs. (3–5) for brevity.

Apart from taking advantage of the local consistency of the available neuroimaging records in the follow-up months, we further take into account the global structure of the neuroimaging records. Using a global projection learned by the principal component analysis (PCA) [10], we map the baseline measurements $\mathbf{x}$ in the high d-dimensional space into a vector $\mathbf{y}$ in a lower r-dimensional space by computing $\mathbf{y} = \mathbf{W}^T\mathbf{x}$, such that data in the projected space $\Re^r$ preserve as much information as possible:

$$\begin{aligned} \mathcal{J}_{\text{Global}}(\mathbf{W}) &= \mathbf{tr}\left(\mathbf{W}^T\mathbf{S}_G\mathbf{W}\right) \\ &= \sum_{i=1}^{n} \left\|\mathbf{W}^T(\mathbf{x}_i - \overline{\mathbf{x}})\right\|_2^2, \\ s.t.\ \ \mathbf{W}^T\mathbf{W} &= \mathbf{I}, \end{aligned} \tag{6}$$

where we compute:

$$\mathbf{S}_G = \sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}}) (\mathbf{x}_i - \overline{\mathbf{x}})^T, \tag{7}$$

which is the covariance matrix of input data $\mathbf{X}$ and we define:

$$\overline{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i, \tag{8}$$

which is the global mean of the input data $\mathbf{X}$. Again, the constant factor $\frac{1}{n}$ is also omitted for brevity.

To integrate the global and local consistencies of neuroimaging records by sing the trace ratio of matrices, we can formulate the new objective as:

$$\begin{aligned} \mathcal{J}_{\ell_2^2}(\mathbf{W}) &= \frac{\mathrm{tr}\left(\mathbf{W}^T \mathbf{S}_G \mathbf{W}\right)}{\mathrm{tr}\left(\mathbf{W}^T \mathbf{S}_L \mathbf{W}\right)} \\ &= \frac{\sum_{i=1}^{n} \left\| \mathbf{W}^T (\mathbf{x}_i - \overline{\mathbf{x}}) \right\|_2^2}{\sum_{i=1}^{n} \sum_{\mathbf{x}_j \in \{\mathcal{N}_i \cup \{\mathbf{x}_i\}\}} \left\| \mathbf{W}^T (\mathbf{x}_j - \overline{\mathbf{x}}_i) \right\|_2^2}, \\ s.t.\ \ & \mathbf{W}^T \mathbf{W} = \mathbf{I}. \end{aligned} \tag{9}$$

A critical problem of $\mathcal{J}_{\ell_2^2}(\mathbf{W})$ in Eq. (9) lies in that it computes the ratio of the summations of a number of squared ℓ_2-norm distances, which are notoriously known to be sensitive to both outlying samples and features [28, 20]. Thus we further rewrite the objective in Eq. (9) as follows:

$$\begin{aligned} \mathcal{J}_{\ell_1}(\mathbf{W}) &= \frac{\sum_{i=1}^{n} \left\| \mathbf{W}^T (\mathbf{x}_i - \overline{\mathbf{x}}) \right\|_1}{\sum_{i=1}^{n} \sum_{\mathbf{x}_j \in \{\mathcal{N}_i \cup \{\mathbf{x}_i\}\}} \left\| \mathbf{W}^T (\mathbf{x}_j - \overline{\mathbf{x}}_i) \right\|_1}, \\ s.t.\ \ & \mathbf{W}^T \mathbf{W} = \mathbf{I}, \end{aligned} \tag{10}$$

in which we compute the summations of a number of ℓ_1-norm distances, because the ℓ_1-norm distance can promote the robustness against outlier samples and features [26, 18, 28, 11, 19, 15, 16].

Upon solving the optimization problem in Eq. (10), the learned enriched neuroimaging representation not only preserves the global variance of biomarker measures, but also maintains the local geometric structure, which thereby is both globally and locally consistent in the learned subspace. Moreover, we enrich the neuroimaging representation $\mathbf{x}$ of the input measures $\mathcal{X}$ by computing:

$$\mathbf{y} = f(g(\mathbf{X}), \mathbf{x}) = \mathbf{W}^T \mathbf{x}, \tag{11}$$

which is a fixed-length single-vector representation and can be readily used by any classical machine learning models for regression or classification analyses. This indeed is the main contribution of this paper.

3. The optimization algorithm to solve our objective

The proposed objective in Eq. (10) maximizes the ratio of the summations of a number of ℓ_1-norm distances, which is nonsmooth thereby difficult to efficiently solve in general. Thus, following our previous works [7, 12] we derive an efficient iterative solution algorithm and prove its convergence in this section. As an important algorithmic contribution, the proposed solution algorithm is non-greedy in nature.

3.1. Solving a general optimization problem of ratio maximization

We first study the following general optimization problem to maximize the ratio of a pair of functions and derive an efficient iterative algorithm to solve it:

$$\max_{v \in \mathcal{C}} \frac{h(v)}{m(v)}, \quad \text{where } m(v) \geq 0\ (\forall v \in \mathcal{C}), \tag{12}$$

where Ω is the feasible domain of the variable v.

To solve the above optimization problem, we propose a simple, yet efficient, iterative algorithm as summarized in Algorithm 1, whose convergence can be proved by Theorem 1.

Algorithm 1: Algorithm to solve Eq. (12).

1. Randomly initialize $v^0 \in \Omega$ and set $k = 1$;
while *not converge* **do**
  2. Calculate $\lambda^k = \frac{h(v^{k-1})}{m(v^{k-1})}$;
  3. Find a $v^k \in \Omega$ satisfying

$$\begin{aligned} & h(v^k) - \lambda^k m(v^k) \\ & \quad > h(v^{k-1}) - \lambda^k m(v^{k-1}) = 0; \end{aligned} \tag{13}$$

  4. $k = k + 1$;
end
Output: v.

Theorem 1 *Algorithm 1 increases the objective in each iteration until convergence.*

Proof. Because $\forall v \in \mathcal{C}\ m(v) > 0$, according to Step 3 of Algorithm (1), we can derive

$$\frac{h\left(\mathbf{x}^{(t)}\right)}{m\left(\mathbf{x}^{(t)}\right)} \geq \lambda^{(t)}. \tag{14}$$

Step 2 of Algorithm (1) defines that

$$\lambda^{(t)} = \frac{h\left(\mathbf{x}^{(t-1)}\right)}{m\left(\mathbf{x}^{(t-1)}\right)}. \tag{15}$$

Thus, we have

$$\frac{h\left(\mathbf{x}^{(t)}\right)}{m\left(\mathbf{x}^{(t)}\right)} \geq \frac{h\left(\mathbf{x}^{(t-1)}\right)}{m\left(\mathbf{x}^{(t-1)}\right)}, \tag{16}$$

which completes the proof. ■

3.2. Our Algorithm to Solve Eq. (10)

Because our new objective in Eq. (10) is a special case of the general optimization problem for ratio maximization in Eq. (1), we can derive Algorithm 2 to solve Eq. (10), whose convergence is thereby guaranteed by Algorithm 1 and Theorem 1.

Algorithm 2: Algorithm to solve Eq. (10).

1. Randomly initialize $\mathbf{W}^{(0)}$ satisfying $\left(\mathbf{W}^{(0)}\right)^T \mathbf{W}^{(0)} = \mathbf{I}$ and set $k = 1$;
while *not converge* **do**
2. Calculate

$$\lambda^{(t)} = \frac{\sum_{i=1}^{n} \left\| \mathbf{W}^{(t-1)^T} (\mathbf{x}_i - \overline{\mathbf{x}}) \right\|_1}{\sum_{i=1}^{n} \sum_{\mathbf{x}_j \in \mathcal{N}_i \cup \{\mathbf{x}_i\}} \| \mathbf{W}^{(t-1)^T} (\mathbf{x}_j - \overline{\mathbf{x}}_i) \|_1}; \tag{17}$$

3. Find a $\mathbf{W}^{(t)}$ satisfying

$$\begin{aligned} &\mathcal{Q}\left(\mathbf{W}^{(t)}\right) \\ &= \mathcal{J}_{\text{Global}}\left(\mathbf{W}^{(t)}\right) - \lambda^{(t)} \mathcal{J}_{\text{Local}}\left(\mathbf{W}^{(t)}\right) \\ &\geq 0; \end{aligned} \tag{18}$$

by Algorithm 3;
4. $t = t + 1$;
end
Output: $\mathbf{W}$.

Now we need to solve the problem in Eq. (18) in Algorithm (2), for which we first introduce the following two lemmas.

Lemma 1 *[13, Theorem 1] For any vector* $\boldsymbol{\xi} = [\xi_1, \cdots, \xi_m]^T \in \Re^m$ *, we have*

$$\|\boldsymbol{\xi}\|_1 = \max_{\boldsymbol{\eta} \in \Re^m} (\text{sign}(\boldsymbol{\eta}))^T \boldsymbol{\xi}, \tag{19}$$

where the maximum value is attained if and only if $\boldsymbol{\eta} = a \times \boldsymbol{\xi}$*, where* $a > 0$ *is a scalar.*

Lemma 2 *[9, Lemma 3.1] For any vector* $\boldsymbol{\xi} = [\xi_1, \cdots, \xi_m]^T \in \Re^m$*, we have*

$$\|\boldsymbol{\xi}\|_1 = \min_{\boldsymbol{\eta} \in \Re^m_+} \frac{1}{2} \sum_{i=1}^{m} \frac{\xi_i^2}{\eta_i} + \frac{1}{2} \|\boldsymbol{\eta}\|_1, \tag{20}$$

where the minimum value is attained if and only if $\eta_j = |\xi_j|, j \in \{1, 2, \cdots, m\}$.

According to Lemma 1 and Lemma 2, to solve the problem in Eq. (18) we introduce the following function:

$$\mathcal{L}\left(\mathbf{W}^{(t)}, \mathbf{W}^{(t-1)}\right) = \mathcal{H}\left(\mathbf{W}^{(t)}, \mathbf{W}^{(t-1)}\right) - \lambda^{(t)} \mathcal{M}\left(\mathbf{W}^{(t)}, \mathbf{W}^{(t-1)}\right), \tag{21}$$

where

$$\mathcal{H}\left(\mathbf{W}^{(t)}, \mathbf{W}^{(t-1)}\right) = \sum_{m=1}^{r} \mathbf{w}_m^{(t)^T} \mathbf{B}\, \text{sign}\left(\mathbf{B}^T \mathbf{w}_m^{(t-1)}\right), \tag{22}$$

where $sign(x)$ is the sign function, and

$$\mathcal{M}\left(\mathbf{W}^{(t)}, \mathbf{W}^{(t-1)}\right) = \frac{1}{2} \sum_{m=1}^{r} \mathbf{w}_m^{(t)^T} \mathbf{A}_m \mathbf{w}_m^{(t)} + \mathbf{w}_m^{(t-1)^T} \mathbf{A}_m \mathbf{w}_m^{(t-1)}. \tag{23}$$

In Eqs. (21–23), we denote $\mathbf{w}_m^{(t)}$ and $\mathbf{w}_m^{(t-1)}$ as the m-th column of matrices $\mathbf{W}^{(t)}$ and $\mathbf{W}^{(t-1)}$ respectively, and define $\mathbf{B}$ and $\mathbf{A}_m$ as follows:

$$\mathbf{B} = \left[\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}, \overline{\mathbf{x}}_2 - \overline{\mathbf{x}}, \cdots, \overline{\mathbf{x}}_n - \overline{\mathbf{x}}\right], \tag{24}$$

$$\mathbf{A}_m = \sum_{i=1}^{n} \sum_{\mathbf{x}_j \in \mathcal{N}_i \cup \{\mathbf{x}_i\}} \frac{(\mathbf{x}_j - \overline{\mathbf{x}}_i)(\mathbf{x}_j - \overline{\mathbf{x}}_i)^T}{\left| \mathbf{w}_m^{(t-1)^T} (\mathbf{x}_j - \overline{\mathbf{x}}_i) \right|}. \tag{25}$$

Theorem 2 *For any* $\mathbf{W}^{(t)} \in \Re^{d \times r}$ *and* $\mathbf{W}^{(t-1)} \in \Re^{d \times r}$*, we have*

$$\mathcal{L}\left(\mathbf{W}^{(t)}, \mathbf{W}^{(t-1)}\right) \leq \mathcal{Q}\left(\mathbf{W}^{(t)}\right). \tag{26}$$

The equality holds if and only if $\mathbf{W}^{(t)} = \mathbf{W}^{(t-1)}$.

Due to the space limit, the proof of Theorem 2 will be provided in the extended longer journal version of this paper in the future.

Substituting $\mathbf{W}^{(t)} = \mathbf{W}^{(t-1)}$ into the function $\mathcal{L}\left(\mathbf{W}^{(t)}, \mathbf{W}^{(t-1)}\right)$, we have:

$$\mathcal{L}\left(\mathbf{W}^{(t-1)}, \mathbf{W}^{(t-1)}\right) = \mathcal{Q}\left(\mathbf{W}^{(t-1)}\right) = 0. \tag{27}$$

In the t-th iteration in solving the objective function in Eq. (10), the optimal solution $\mathbf{W}^\star$ satisfies

$$\mathcal{L}\left(\mathbf{W}^\star, \mathbf{W}^{k-1}\right) \geq \mathcal{L}\left(\mathbf{W}^{k-1}, \mathbf{W}^{k-1}\right) = 0. \tag{28}$$

Then, we have:

$$\begin{aligned} \mathcal{Q}\left(\mathbf{W}^\star\right) &\geq \mathcal{L}\ \ \mathbf{W}^\star, \mathbf{W}^{(t-1)} \\ &\geq \mathcal{L}\ \ \mathbf{W}^{(t-1)}, \mathbf{W}^{(t-1)} \\ &= \mathcal{Q}\ \ \mathbf{W}^{(t-1)}\ \ = 0. \end{aligned} \tag{29}$$

Theorem 2 and Eq. (29) indicate that the solution of the problem in Eq. (18) can be transformed to solve the problem of $\mathcal{L}\left(\mathbf{W}^{(t)}, \mathbf{W}^{(t-1)}\right) \geq 0$, which can be solved by the projected subgradient method with Armigo line search. Thus we compute the subgradient of $\mathcal{L}\left(\mathbf{W}^{(t)}, \mathbf{W}^{(t-1)}\right)$ at $\mathbf{W}^{(t)}$ as:

$$\begin{aligned} \partial\mathcal{L}(\mathbf{W}^{(t)}, \mathbf{W}^{(t-1)}) = \mathbf{B}\,\mathrm{sign}\ \ \mathbf{B}^T\mathbf{W}^{(t-1)} \\ - \lambda^k \left[\mathbf{A}_1\mathbf{w}_1^{(t)}, \mathbf{A}_2\mathbf{w}_2^{(t)}, \cdots, \mathbf{A}_p\mathbf{w}_p^{(t)}\right. \ . \end{aligned} \tag{30}$$

Note that for a given matrix $\mathbf{W}^{(t)}$, here we define the operator:

$$\mathcal{P}\ \ \mathbf{W}^{(t)}\ \ = \mathbf{W}^{(t)}\ \ \ \ \mathbf{W}^{(t)}\ \ ^T\mathbf{W}^{(t)}\Big)^{-\frac{1}{2}}, \tag{31}$$

which can project $\mathbf{W}^{(t)}$ onto an orthogonal cone. This guarantees the orthogonality constraint of the projection matrix $\left(\mathbf{W}^{(t)}\right)^T\mathbf{W}^{(t)} = \mathbf{I}$.

Algorithm 3 summarizes the solution to the problem in Eq. (18).

Algorithm 3: Solve the optimization problem (18).

Input: $\mathbf{W}^{(t)}$ and the parameter $0 < \beta < 1$;

1. Calculate

$$\lambda^{(t)} = \frac{\sum_{i=1}^{n}\left\|\ \ \mathbf{W}^{(t-1)}\ \ ^T(\mathbf{x}_i - \overline{\mathbf{x}})\right\|_1}{\sum_{i=1}^{n}\sum_{\mathbf{x}_j\in\mathcal{N}_i\cup\{\mathbf{x}_i\}}\|\ \ \mathbf{W}^{(t-1)}\ \ ^T(\mathbf{x}_j - \overline{\mathbf{x}}_i))\|_1} \tag{32}$$

thus the subgradient is computed as

$$G^{(t-1)} = \partial\mathcal{L}\ \ \mathbf{W}^{(t-1)}, \mathbf{W}^{(t-1)} \tag{33}$$

and set $m = 1$;

while $\mathcal{Q}\left(\mathbf{W}^{(t)}\right) < 0$ **do**

2. Calculate $\mathbf{W}^{(t)} = \mathcal{P}\left(\mathbf{W}^{(t-1)} + \beta^m G^{(t-1)}\right)$;

3. Calculate $\mathcal{Q}\left(\mathbf{W}^{(t)}\right)$ by Eq. (10);

end

Output: $\mathbf{W}$.

4. Experiments

In this section, we experimentally evaluate the prediction performance of the enriched biomarker representations learned by our proposed method by applying it to the ADNI database.

4.1. ADNI dataset description

Data used in the preparation of the experiments were obtained from the ADNI database (`adni.loni.usc.edu`). We download 1.5 T MRI scans and demographic information for 821 ADNI-1 participants. We perform voxel-based morphometry (VBM) and FreeSurfer on the MRI data by following [21] and extracted mean modulated gray matter (GM) measures for 90 target regions of interest (ROI). These measures are adjusted for the baseline intracranial volume (ICV) using regression weights derived from the HC participants at the baseline. We also download the longitudinal scores of the participants in five independent cognitive assessments including Alzheimer's Disease Assessment Scale (ADAS), Mini-Mental State Examination (MMSE), and Fluency test (FLU).

The time points examined in this study for both imaging records and cognitive assessments includes baseline, M6, M12, M18, M24 and M36. All the participants' data used in our enriched neuroimaging representation study are required to have a baseline measurement, baseline cognitive score and at least two available records from M6/M12/M18/M24/M36. A total of 544 sample subjects are involved in our study, among which we have 92 AD samples, and 205 MCI samples and 247 HC samples. Four cognitive assessment scores are included: (1) ADAS TOTAL scores from ADAS cognitive assessment, (2) MMSE score from MMSE cognitive assessment, (3) FLU ANIM and (4) FLU VEG scores from Fluency cognitive assessment.

4.2. Performance comparison on the ADNI cohort

In this subsection, we will compare the predictive power of the enriched biomarker representations learned by our new method against the original counterparts of the baseline record using the VBM biomarkers. Due to space limit, more experimental results, such as those on the FreeSurfer biomarkers will be provided in the extended longer journal version of this paper in the future.

4.2.1 Experiment settings

To validate the usefulness of our proposed method, we compare cognitive outcomes prediction performance using two type of the neuroimaging inputs – the learned enriched representation and BL biomarker measurement. In our experiments, several methods proven to generalize well, such as linear regression (LR), ridge regression (RR), Lasso, support vector regression (SVR), and CNN are leveraged. LR is the simplest and widely used regression model in statistical learning and brain image analysis. RR is a regularized version of LR that induces sparsity to account for over-fitting. Lasso regression performs both variable selection and regularization in order to enhance the prediction accuracy. SVR is the regression version of support vector machine (SVM) , which is widely applied in many different applications. CNN regression is the regression version of convolutional neural networks, which has demonstrated its superior performance compared to the classical machine learning models.

For LR, RR, Lasso and SVR models, we conduct a standard 5-fold cross-validation approach by computing the root mean square error (RMSE) between the predicted values and ground truth values of the cognitive scores on the testing data. In the standard 5-fold cross-validation, the data are equally and randomly divided into 5 groups. In every trial, one group is treated as testing data and the other four groups are used as training data. This process repeats five times in turn so that all the data can be fairly treated. In RR and Lasso methods, the regularization parameters are fine tuned by search the grid of $\{10^{-10}, \ldots, 10^{-1}, 1, 10, \cdots, 10^{10}\}$. In the SVR model, the Gaussian kernel is leveraged, and box constraints parameters are also fine tuned following a grid search of $\{10^{-5}, \ldots, 10^{-1}, 1, 10, \cdots, 10^{5}\}$.

There is a slight difference for the CNN experimental settings. For the CNN regression model, we randomly select 70% of the neuroimaging measurements as the training set, 20% of the neuroimaging measurements as the validation set and the remaining 10% of the neuroimaging measurements as the testing set. The validation set in the CNN experimental setting is designed to provide an unbiased evaluation of the model fit on the training dataset while tuning model hyper parameters. The evaluation metrics reported are based on the results on the testing dataset. We construct a two layer convolution architecture for the cognitive outcomes prediction: (1) 16 1×5 convolutions (unpadded convolutions), followed by a rectified linear unit (ReLU) and a 1×2 max pooling operation; (2) 32 1×10 convolutions (unpadded convolutions) with ReLU and a 1×2 max pooling operation. The dropout technique is also leveraged to reduce overfitting in CNN models and prevent complex co-adaptations on training data. In all our experiments, the dropout probability is set to be 0.3 and the batch size is set to be 16.

4.2.2 Experiment results

From Figure 1, we can see that the proposed enriched neuroimaging representation is consistently better than baseline representation in five different methods, LR, RR, Lasso, SVR and CNN. It can be attributed to the following reasons. Firstly, the original baseline neuroimaging representation only deals with one single cognitive measure, it cannot benefit from longitudinal correlation across different neuroimaging records over the time. Instead, our proposed enriched neuroimaging biomarker representation could capture not only the baseline neuroimaging record, but also the temporal local consistency among the follow-up neuroimaging records. Our enriched neuroimaging representation could integrate the neuroimaging records at fix time point and the its dynamic temporal changes at the same time. As AD is progressively degenerative disease, this incorporation of future information about subjects could assist in predictions. Secondly, the original baseline neuroimaging measurements exhibits high dimensionality, which could be redundant and noisy. Thus the traditional methods are easily suffered from "the curse of dimensionality". Via the projection learned from the objective in Eq. (10), we map the baseline cognitive measurement into a low dimension space mitigating the issue of high dimensionality. Thus, from Figure 1 we can see that, compared to the original high dimensional baseline representation, our enriched representation achieves a great improvement when using LR, RR and Lasso to predict cognitive outcomes.

In all, by incorporating the global and local consistency of the original biomarker representations of each participant, we learn a low-dimensional consistent fixed-length enriched neuroimaging biomarker representation. Through the enriched biomarker representation, we obtain a prediction performance gain using the five different commonly used regression models on VBM biomarkers, which certifies the usefulness of our proposed enriched biomarker representations.

5. Conclusion

In this paper, we propose a novel formulation to learn an enriched neuroimaging biomarker representation using available longitudinal data. Our enriched biomarker representation is implemented by solving a new objective that enforces both global and local consistency of the neuroimaging measurements of each participant in the projected subspace, where the global consistency is designed to preserve similar distributions of neuroimaging measurements of each participant during the project, and the local consistency is designed to preserve the pairwise relationship of neuroimaging measurements of each participant. The objective simultaneously maximizes and minimizes the summations of a number of ℓ_1-norm distances, which is difficult to solve

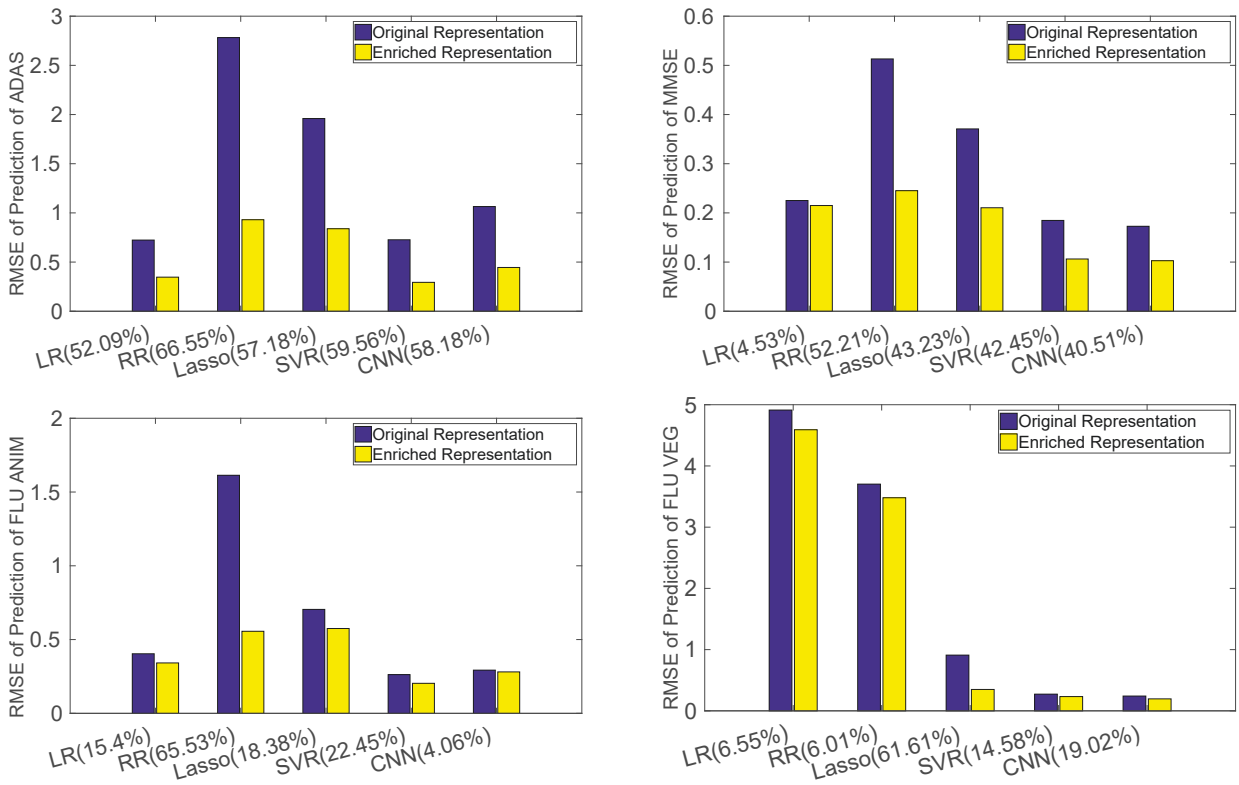

Figure 1. Experiment results using VBM biomarkers. We use the original representation (left) and enriched representation (right) to predict ten different baseline cognitive outcomes using five different methods – linear regression (LR), ridge regression (RR), Lasso, support vector regression (SVR), convolutional neural networks (CNN). The root mean squared error (RMSE) value for each cognitive outcome is calculated for comparison. The percentage improvement of each method compared the original representation and enriched representation is also listed.

in general. We develop an efficient iterative solution algorithm that is non-greedy and theoretically proved to converge. We conducted experiments on the VBM biomarkers. Via the enriched neuroimaging representation, we can achieve a performance gain in predicting ten different cognitive outcomes using five regression models.

Acknowledgements

We thank Dr. Heng Huang in the Department of Electrical and Computer Engineering of University of Pittsburgh and Dr. Li Shen in the Department of Biostatistics, Epidemiology and Informatics of University of Pennsylvania for providing with us the parsed ADNI data which are used in the experiments of this study.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Bio-gen; Bristol-Myers Squibb Company; CereSpir, Inc.; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. 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 Southern California.

References

[1] Alzheimer's Association et al. 2018 alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 14(3):367–429, 2018.

[2] Vasileios Belagiannis, Christian Rupprecht, Gustavo Carneiro, and Nassir Navab. Robust optimization for deep regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2830–2838, 2015.

[3] Lodewijk Brand, Kai Nichols, Hua Wang, Heng Huang, and Li Shen. Predicting longitudinal outcomes of alzheimer's disease via a tensor-based joint classification and regression model. In *Pac Symp Biocomput*, pages 7–18. World Scientific, 2020.

[4] Lodewijk Brand, Kai Nichols, Hua Wang, Li Shen, and Heng Huang. Joint multi-modal longitudinal regression and classification for alzheimer's disease prediction. *IEEE Transactions on Medical Imaging*, 2019.

[5] Lodewijk Brand, Hua Wang, Heng Huang, Shannon Risacher, Andrew Saykin, Li Shen, et al. Joint high-order multi-task feature learning to predict the progression of alzheimer's disease. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 555–562. Springer, 2018.

[6] Serge Gauthier, Barry Reisberg, Michael Zaudig, Ronald C Petersen, Karen Ritchie, Karl Broich, Sylvie Belleville, Henry Brodaty, David Bennett, Howard Chertkow, et al. Mild cognitive impairment. *The lancet*, 367(9518):1262–1270, 2006.

[7] Fei Han, Hua Wang, and Hao Zhang. Learning integrated holism-landmark representations for long-term loop closure detection. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[8] Chris Hinrichs, Vikas Singh, Guofan Xu, Sterling C Johnson, Alzheimers Disease Neuroimaging Initiative, et al. Predictive markers for ad in a multi-modality framework: an analysis of mci progression in the adni population. *Neuroimage*, 55(2):574–589, 2011.

[9] Rodolphe Jenatton, Guillaume Obozinski, and Francis Bach. Structured sparse principal component analysis. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 366–373, 2010.

[10] Ian Jolliffe. Principal component analysis. In *International encyclopedia of statistical science*, pages 1094–1096. Springer, 2011.

[11] Kai Liu, Lodewijk Brand, Hua Wang, and Feiping Nie. Learning robust distance metric with side information via ratio minimization of orthogonally constrained $\ell_{2,1}$-norm distances. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*, 2019.

[12] Kai Liu, Hua Wang, Feiping Nie, and Hao Zhang. Learning multi-instance enriched image representations via non-greedy ratio maximization of the ℓ_1-norm distances. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7727–7735, 2018.

[13] Yang Liu, Quanxue Gao, Shuo Miao, Xinbo Gao, Feiping Nie, and Yunsong Li. A non-greedy algorithm for l1-norm lda. *IEEE Trans. Image Process.*, 26(2):684–695, 2017.

[14] Raymond Y Lo and William J Jagust. Predicting missing biomarker data in a longitudinal study of alzheimer disease. *Neurology*, pages WNL–0b013e318253d5b3, 2012.

[15] Lyujian Lu, Saad Elbeleidy, Lauren Baker, Hua Wang, Heng Huang, Li Shen, et al. Improved prediction of cognitive outcomes via globally aligned imaging biomarker enrichments over progressions. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 140–148. Springer, 2019.

[16] Lyujian Lu, Saad Elbeleidy, Lauren Zoe Baker, Hua Wang, and ADNI. Learning multi-modal biomarker representations via globally aligned longitudinal enrichments. In *AAAI*, 2020.

[17] Lyujian Lu, Hua Wang, Xiaohui Yao, Shannon Risacher, Andrew Saykin, and Li Shen. Predicting progressions of cognitive outcomes via high-order multi-modal multi-task feature learning. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 545–548. IEEE, 2018.

[18] Feiping Nie, Hua Wang, Heng Huang, and Chris Ding. Early active learning via robust representation and structured sparsity. In *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

[19] Feiping Nie, Hua Wang, Zheng Wang, and Heng Huang. Robust linear discriminant analysis using ratio minimization of $\ell_{1,2}$-norms. *arXiv preprint arXiv:1907.00211*, 2019.

[20] Feiping Nie, Jianjun Yuan, and Heng Huang. Optimal mean robust principal component analysis. In *International conference on machine learning*, pages 1062–1070, 2014.

[21] Shannon L Risacher, Li Shen, John D West, Sungeun Kim, Brenna C McDonald, Laurel A Beckett, Danielle J Harvey, Clifford R Jack Jr, Michael W Weiner, Andrew J Saykin, et al. Longitudinal mri atrophy biomarkers: relationship to conversion in the adni cohort. *Neurobiology of aging*, 31(8):1401–1418, 2010.

[22] Li Shen, Sungeun Kim, Shannon L Risacher, Kwangsik Nho, Shanker Swaminathan, John D West, Tatiana Foroud, Nathan Pankratz, Jason H Moore, Chantel D Sloan, et al. Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in mci and ad: A study of the adni cohort. *Neuroimage*, 53(3):1051–1063, 2010.

[23] Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.

[24] Cynthia M Stonnington, Carlton Chu, Stefan Klöppel, Clifford R Jack Jr, John Ashburner, Richard SJ Frackowiak, Alzheimer Disease Neuroimaging Initiative, et al. Predicting clinical scores from magnetic resonance scans in alzheimer's disease. *Neuroimage*, 51(4):1405–1413, 2010.

[25] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[26] Hua Wang, Feiping Nie, and Heng Huang. Robust and discriminative distance for multi-instance learning. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2919–2924. IEEE, 2012.

[27] Hua Wang, Feiping Nie, and Heng Huang. Globally and locally consistent unsupervised projection. In *AAAI*, pages 1328–1333, 2014.

[28] Hua Wang, Feiping Nie, and Heng Huang. Robust distance metric learning via simultaneous ℓ_1-norm minimization and maximization. In *International Conference on Machine Learning (ICML 2014)*, pages 1836–1844, 2014.

[29] Hua Wang, Feiping Nie, Heng Huang, Shannon Risacher, Chris Ding, Andrew J Saykin, Li Shen, et al. Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance. In *ICCV*, pages 557–562. IEEE, 2011.

[30] Hua Wang, Feiping Nie, Heng Huang, Shannon Risacher, Andrew J Saykin, Li Shen, et al. Identifying ad-sensitive and cognition-relevant imaging biomarkers via joint classification and regression. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 115–123. Springer, 2011.

[31] Hua Wang, Feiping Nie, Heng Huang, Shannon L Risacher, Andrew J Saykin, Li Shen, and Alzheimer's Disease Neuroimaging Initiative. Identifying disease sensitive and quantitative trait-relevant biomarkers from multidimensional heterogeneous imaging genetics data via sparse multimodal multitask learning. *Bioinformatics*, 28(12):i127–i136, 2012.

[32] Hua Wang, Feiping Nie, Heng Huang, Jingwen Yan, Sungeun Kim, Kwangsik Nho, Shannon L Risacher, Andrew J Saykin, Li Shen, and Alzheimer's Disease Neuroimaging Initiative. From phenotype to genotype: an association study of longitudinal phenotypic markers to alzheimer's disease relevant snps. *Bioinformatics*, 28(18):i619–i625, 2012.

[33] Hua Wang, Feiping Nie, Heng Huang, Jingwen Yan, Sungeun Kim, Shannon Risacher, Andrew Saykin, and Li Shen. High-order multi-task feature learning to identify longitudinal phenotypic markers for alzheimer's disease progression prediction. In *NIPS*, pages 1277–1285, 2012.

[34] Xiaoqian Wang, Dinggang Shen, and Heng Huang. Prediction of memory impairment with mri data: a longitudinal study of alzheimer's disease. In *MICCAI*, pages 273–281. Springer, 2016.

[35] Xiaoqian Wang, Jingwen Yan, Xiaohui Yao, Sungeun Kim, Kwangsik Nho, Shannon L Risacher, Andrew J Saykin, Li Shen, Heng Huang, et al. Longitudinal genotype-phenotype association study via temporal structure auto-learning predictive model. In *RECOMB*, pages 287–302. Springer, 2017.

[36] Gary L Wenk et al. Neuropathologic changes in alzheimer's disease. *Journal of Clinical Psychiatry*, 64:7–10, 2003.

[37] Shuo Xiang, Lei Yuan, Wei Fan, Yalin Wang, Paul M Thompson, Jieping Ye, Alzheimer's Disease Neuroimaging Initiative, et al. Bi-level multi-source learning for heterogeneous block-wise missing data. *NeuroImage*, 102:192–206, 2014.

[38] Jingwen Yan, Taiyong Li, Hua Wang, Heng Huang, Jing Wan, Kwangsik Nho, Sungeun Kim, Shannon L Risacher, Andrew J Saykin, Li Shen, et al. Cortical surface biomarkers for predicting cognitive outcomes using group $\ell_{2,1}$ norm. *Neurobiology of aging*, 36:S185–S193, 2015.

[39] Guan Yu, Quefeng Li, Dinggang Shen, and Yufeng Liu. Optimal sparse linear prediction for block-missing multi-modality data without imputation. *Journal of the American Statistical Association*, pages 1–35, 2019.

[40] Daoqiang Zhang, Dinggang Shen, Alzheimer's Disease Neuroimaging Initiative, et al. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in alzheimer's disease. *NeuroImage*, 59(2):895–907, 2012.