Learning Multi-Modal Biomarker Representations via Globally Aligned Longitudinal Enrichments

Lyujian Lu, Saad Elbeleidy, Lauren Zoe Baker, Hua Wang*,
for the Alzheimer's Disease Neuroimaging Initiative†
Department of Computer Science, Colorado School of Mines, Golden, CO 80401
lyujianlu@mymail.mines.edu, selbeleidy@mymail.mines.edu, laurenzoebaker@mymail.mines.edu
huawangcs@gmail.com

Abstract

Alzheimer's Disease (AD) is a chronic neurodegenerative disease that severely impacts patients' thinking, memory and behavior. To aid automatic AD diagnoses, many longitudinal learning models have been proposed to predict clinical outcomes and/or disease status, which, though, often fail to consider missing temporal phenotypic records of the patients that can convey valuable information of AD progressions. Another challenge in AD studies is how to integrate heterogeneous genotypic and phenotypic biomarkers to improve diagnosis prediction. To cope with these challenges, in this paper we propose a longitudinal multi-modal method to learn enriched genotypic and phenotypic biomarker representations in the format of *fixed-length vectors* that can simultaneously capture the baseline neuroimaging measurements of the entire dataset and progressive variations of the *varied counts* of follow-up measurements over time of every participant from different biomarker sources. The learned global and local projections are aligned by a soft constraint and the structured-sparsity norm is used to uncover the multi-modal structure of heterogeneous biomarker measurements. While the proposed objective is clearly motivated to characterize the progressive information of AD developments, it is a nonsmooth objective that is difficult to efficiently optimize in general. Thus, we derive an efficient iterative algorithm, whose convergence is rigorously guaranteed in mathematics. We have conducted extensive experiments on the Alzheimer's Disease Neuroimaging Initiative (ADNI) data using one genotypic and two phenotypic biomarkers. Empirical results have demonstrated that the learned enriched biomarker representations are more effective in predicting the outcomes of various cognitive assessments. Moreover, our model has successfully identified disease-relevant biomarkers supported by existing medical findings that additionally warrant the correctness of our method from the clinical perspective.

*Corresponding author.

†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.

Introduction

As the most prevalent and severe neurodegenerative disorder, Alzheimer's Disease (AD) results in progressive impairment of memory and other cognitive abilities, triggered by the damage of neurons in the brain. AD usually progresses along a temporal continuum, initially from a pre-clinical stage, subsequently to mild cognitive impairment (MCI) and ultimately deteriorating to AD (Wenk and others 2003; Brand et al. 2018). 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 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 (Association and others 2018).

With all these facts, neuroimaging measurements have been widely studied to predict disease status and/or cognitive performance (Stonnington et al. 2010; Zhang et al. 2012; Wang et al. 2012b; Yan et al. 2015; Brand et al. 2018). However, there are a few limitations to these predictive models. These approaches routinely conduct standard regression and/or classification at each individual time point separately, failing to take advantage of the longitudinal structure among temporal brain phenotypes. First, since AD is a progressive neurodegenerative disorder, multiple records can be obtained to monitor the disease's progression. Thus it is beneficial to uncover the temporal relation among these longitudinal biomarkers (Wang et al. 2012c; Wang, Shen, and Huang 2016; Wang et al. 2017; Brand et al. 2018). Second, while heterogeneous biomarker measurements, such as voxel-based morphometry (VBM), FreeSurfer, and single-nucleotide polymorphism (SNP), are available for predicting disease status and/or cognitive performance, current longitudinal methods (Wang et al. 2012a; Wang, Nie, and Huang 2013; Wang, Shen, and Huang 2016) often do not explore this multi-modal structure to boost prediction capabilities. Third, most importantly, the longitudinal biomarkers are often missing at some time points for participants, since it is difficult to conduct medical scans consistently across a large group of subjects. Mortality risk and cognitive impairment hinder older adults from staying in studies requiring multiple visits and thus result in incom-

plete data.

To overcome the first limitation of longitudinal data, several proposed longitudinal prediction models (Wang et al. 2012c; Wang, Shen, and Huang 2016; Wang et al. 2017; Lu et al. 2018; Brand et al. 2018) uncover the temporal structure of brain phenotypes. However, these models treat the temporal biomarkers as a tensor, which inevitably increases the complexity of the prediction problem. To address the multi-modal structure of heterogeneous biomarker measurements, our previous works (Wang et al. 2012a; Wang, Nie, and Huang 2013; Wang et al. 2013) proposed to explore the group structure of biomarkers from different imaging and genetic sources, which, though, conduct standard regression and/or classification on all the time points separately, without considering the temporal relation among brain phenotypes over the disease progression. To handle the third limitation of data inconsistency, most longitudinal studies of AD only utilize data samples with complete temporal records for model analysis and ignore the time points with missing records. But discarding of the samples with partial data can completely ruin the data set. Recently, several data imputation methods (Xiang et al. 2014; Li et al. 2019) have been proposed to generate missing records of longitudinal AD measures. Temporal regression studies can be then conducted after these data imputation approaches. However, these data imputation methods can easily introduce undesirable artifacts, which can significantly degrade the predictive power of the longitudinal learning models.

To deal with the *longitudinal multi-modal* prediction problem *with incomplete temporal inputs*, in this paper we propose to learn an enriched multi-modal biomarker representation from heterogeneous genotypic and phenotypic biomarker measurements, which can elegantly combines the baseline biomarkers and all the dynamic imaging measures across time with missing inputs at ***any*** time points. It first learns a global projection from the baseline biomarkers common to all participants to preserve the global structure of the entire dataset. Then, for every participant it learns a local projection from their available follow-up biomarkers to maintain the individual information of every participant in the dataset. A soft constraint is used to enforce consistency between the global and local projections. In addition, a structured-sparsity norm regularization (Wang et al. 2012a; Wang, Nie, and Huang 2013; Wang et al. 2013) is utilized to explore the multi-modal structure of biomarker measurements from different sources. Finally, taking into account the varied-sized phenotypes of different participants over time due to missing records, we replace the traditional squared ℓ_2-norm distances by *not-squared* ℓ_2-norm distances (Wang, Nie, and Huang 2015) to promote the robustness of our model against outliers. Using the learned projection, we transform the inconsistent heterogeneous biomarker representations with varied lengths into a fixed-length vector representation, which can simultaneously capture the information from both baseline measures of the entire dataset and the progressive summary of available follow-up biomarker measurements of every individual participant from multi-modal biomarker measurements. With the fixed-length biomarker representations, we can easily use conventional learning methods to predict the clinical outcomes for early AD detections.

We have performed extensive experiments on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset (Weiner et al. 2010) and achieved obvious prediction capability gains by using the newly learned fixed-length representation compared to baseline biomarkers. In addition, we can use our new method to identify disease-relevant biomarkers that are in accordance with existing medical research findings, which warrants the correctness of our method from the clinical perspective.

Problem Formalization and Our Objective

In the sequel of this paper, we use the following notations. The ℓ_p-norm $(p \geq 1)$ of a vector $\mathbf{v} \in \Re^d$ is defined as $\|\mathbf{v}\|_p = \left(\sum_{i=1}^{d} v_i^p\right)^{\frac{1}{p}}$. For a matrix $\mathbf{M} = [m_{ij}]$, its trace is defined as $\mathbf{tr}(\mathbf{M}) = \sum_i m_{ii}$. The Frobenius norm of $\mathbf{M}$ is defined as $\|\mathbf{M}\|_2 = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}|m_{i,j}|^2}$. The group ℓ_1-norm of $\mathbf{M}$ introduced in our earlier works (Wang et al. 2012a; Wang, Nie, and Huang 2013; Wang et al. 2013) is defined as $\|\mathbf{M}\|_{G_1} = \sum_{i=1}^{c}\sum_{j=1}^{k}||\mathbf{m}_i^j||_2$, where we write $\mathbf{M} = [\mathbf{m}_1^1, \cdots, \mathbf{m}_c^1; \cdots, \cdots, \cdots; \mathbf{m}_1^k, \cdots, \mathbf{m}_c^k] \in \Re^{d\times c}$ and $\mathbf{m}_p^q \in \Re^{d_q}$ indicates the weights of all features in the q-th view with respect to the p-th diagnosis task.

In the task of learning enriched biomarker representations, our goal is to learn a fixed-length biomarker representation vector from baseline scans and available follow-up measurements. We denote biomarker records of a participant as $\mathcal{X}_i = \{\mathbf{x}_i, \mathbf{X}_i\}$, where $i = 1, 2, \cdots, n$ denotes the index of the participants in the ADNI cohort. Here $\mathbf{x}_i \in \Re^d$ denotes the baseline records of the i-th participant and d denotes the dimensions of the features. $\mathbf{X}_i = \left[\mathbf{x}_1^i, \ldots, \mathbf{x}_{n_i}^i\right] \in \Re^{d\times n_i}$ collects all the available follow-up biomarker records of the i-th participant, and n_i denotes the number of available follow-up records of i-th participant after baseline. Here we note that n_i ***varies*** over the participants in the dataset due to inconsistent/missing temporal records of different participants. Both the baseline records and available follow-up records are the concatenation of Voxel-based morphometry (VBM) measurement, Freesurfer (FS) measurement and Single-nucleotide polymorphism (SNP). Mathematically, we have the baseline as $\mathbf{x}_i = [\mathbf{x}_{i_{VBM}}, \mathbf{x}_{i_{FS}}, \mathbf{x}_{i_{SNP}}]$ and the follow-ups as $\mathbf{x}_j^i = [\mathbf{x}_{j_{VBM}}^i, \mathbf{x}_{j_{FS}}^i, \mathbf{x}_{j_{SNP}}^i]$ where $1 \leq j \leq n_i$ for the i-th participant.

Because the biomarker records in the ADNI dataset are a concatenation of genotypic and phenotypic measurements, they reside in a high-dimensional space, leading to the failure of many traditional machine learning models due to the curse of dimensionality. Thus we first project the high-dimensional biomarker representation into a low-dimensional subspace. To keep the most useful information, Principal component analysis (PCA) (Jolliffe 2011) is the right tool to learn a projection $\mathbf{W}_0 \in \Re^{d\times r}$ (usually $r \ll d$)

that maps the baseline neuroimaging measurement of the i-th patient's $\mathbf{x}_i$ into a low dimensional space. Mathematically, PCA minimizes the reconstruction errors via the projection $\mathbf{W}_0$ by optimizing the following the objective:

$$\begin{aligned} \mathcal{J}_{\text{Global}}\left(\mathbf{W}_0\right) &= \sum_{i=1}^{n} \left\|\mathbf{x}_i - \mathbf{W}_0\mathbf{W}_0^T\mathbf{x}_i\right\|_2^2, \\ s.t. \quad & \mathbf{W}_0^T\mathbf{W}_0 = \mathbf{I}. \end{aligned} \tag{1}$$

Commonly, the biomarker records of one individual participant do not experience drastic changes over a short time, thus we want to preserve the local consistency in the projected space for the participant as well. To uncover the local consistency among the follow-up records of every participant, we preserve the local pairwise affinities of the scans in the projected subspace using locality preserving projections (LPP) (He and Niyogi 2004). That is, given a pairwise similarity matrix of the measurements of the i-th participant $\mathbf{S}_i \in \Re^{n_i \times n_i}$, LPP preserves the local relationships and maximizes the smoothness of the manifold of the data in the embedding space by minimizing the following objective:

$$\begin{aligned} \mathcal{J}_{\text{Local}}\left(\mathbf{W}_i\right) &= \sum_{\mathbf{x}_j^i, \mathbf{x}_k^i \in \mathbf{X}_i} s_{jk}^i \left\|\mathbf{W}_i^T\mathbf{x}_j^i - \mathbf{W}_i^T\mathbf{x}_k^i\right\|_2^2, \\ s.t. \quad & \mathbf{W}_i^T\mathbf{W}_i = \mathbf{I}. \end{aligned} \tag{2}$$

First, from a multi-modal perspective, the features of a specific modality can be more or less discriminative for different clusters. To untangle the modality structure, we use the group ℓ_1-norm (G1-norm) introduced in (Wang et al. 2012a; Wang, Nie, and Huang 2013; Wang et al. 2013) for regularization. Because the group ℓ_1-norm uses the ℓ_2-norm within each modality and the ℓ_1-norm between modalities, it enforces the sparsity between different learning tasks. Moreover, in certain cases, even if most features in one modality are not discriminative for a group of objects, a small number of features in the same modality can still be highly discriminative. Second, because $\mathbf{W}_i$ is learned from one individual participant of a dataset, it can only characterize the AD progression of that single participant. To align the learned projections $\mathbf{W}_i$ for all the participants of the dataset for $1 \leq i \leq n$, we approximate them by $\mathbf{W}_0$ learned from the baseline measurements of the entire dataset. Taking into account these facts, we propose the following objective that can integrate the global and local consistencies of neuroimaging records and fuse multi-modal genotypic and phenotypic measurements over time:

$$\begin{aligned} \mathcal{J}_{\ell_2^2}(\mathcal{W}) = \min_{\mathcal{W}} & \sum_{i=1}^{n} \left\|\mathbf{x}_i - \mathbf{W}_0\mathbf{W}_0^T\mathbf{x}_i\right\|_2^2 \\ & + \gamma_1 \sum_{i=1}^{n} \sum_{\mathbf{x}_j^i, \mathbf{x}_k^i \in \mathbf{X}_i} s_{jk}^i \left\|\mathbf{W}_i^T\mathbf{x}_j^i - \mathbf{W}_i^T\mathbf{x}_k^i\right\|_2^2 \\ & + \gamma_2 \sum_{i=1}^{n} \left\|\mathbf{W}_0 - \mathbf{W}_i\right\|_2^2 + \gamma_3 \sum_{i=0}^{n} \left\|\mathbf{W}_i\right\|_{G_1}, \\ s.t. \quad & \mathbf{W}_i^T\mathbf{W}_i = \mathbf{I} \quad (0 \leq i \leq n), \end{aligned} \tag{3}$$

where $\mathcal{W} = \{\mathbf{W}_0, \mathbf{W}_1, \cdots, \mathbf{W}_n\}$.

Finally, considering inevitable outlying samples due to missing phenotypic records of the participants in the ADNI cohort, we replace the squared ℓ_2-norm distances in Eq. (3) with their *not-squared* counterparts as follows for better robustness (Wang, Nie, and Huang 2015; Liu et al. 2017):

$$\begin{aligned} \mathcal{J}_{\ell_2}(\mathcal{W}) = \min_{\mathcal{W}} & \sum_{i=1}^{n} \left\|\mathbf{x}_i - \mathbf{W}_0\mathbf{W}_0^T\mathbf{x}_i\right\|_2 \\ & + \gamma_1 \sum_{i=1}^{n} \sum_{\mathbf{x}_j^i, \mathbf{x}_k^i \in \mathbf{X}_i} s_{jk}^i \left\|\mathbf{W}_i^T\mathbf{x}_j^i - \mathbf{W}_i^T\mathbf{x}_k^i\right\|_2 \\ & + \gamma_2 \sum_{i=1}^{n} \left\|\mathbf{W}_0 - \mathbf{W}_i\right\|_{2,1} + \gamma_3 \sum_{i=0}^{n} \left\|\mathbf{W}_i\right\|_{G_1}, \\ s.t. \quad & \mathbf{W}_i^T\mathbf{W}_i = \mathbf{I} \quad (0 \leq i \leq n). \end{aligned} \tag{4}$$

Upon solving the optimization problem in Eq. (4), we learn a fixed-length representation for each participant by computing $\{\mathbf{y}_i = \mathbf{W}_i^T\mathbf{x}_i\}_{i=1}^n$, which can be readily fed into traditional machine learning models.

The Solution Algorithm

Although the motivations of the formulation of our new method in Eq. (4) is clear and justifiable, it is a non-smooth objective, which is difficult to efficiently solve in general. Thus we derive the solution of our objective in this section.

Smoothed Iterative Reweighted Method

We first use the smoothed iterative reweighted method to convert the optimization problem in Eq. (4) into an easier problem. The smoothed iterative reweighted method was first introduced in our earlier work in (Liu et al. 2017) that solves the following general optimization problem:

$$\min_x f(x) + \sum_i \left\|g_i(x)\right\|_2, \tag{5}$$

where $g_i(x)$ is a vector output function. It can be seen that our objective in Eq. (4) is a special case of the problem in Eq. (5).

Because Eq. (5) is not a smooth objective, we turn to solve the following smooth optimization problem:

$$\min_x f(x) + \sum_i \sqrt{g_i^T(x)g_i(x) + \delta}. \tag{6}$$

It is apparent that Eq. (6) is reduced to Eq. (5), when $\delta \to 0$. By setting the derivative of Eq. (6) with respect to x to zero, we have:

$$f'(x) + \sum_i \frac{g_i(x)}{\sqrt{g_i^T(x)g_i(x) + \delta}} = 0. \tag{7}$$

Denote

$$s_i = \frac{1}{2\sqrt{g_i^T(x)g_i(x) + \delta}}, \tag{8}$$

then Eq. (7) is rewritten as:

$$f'(x) + \sum_i 2s_i g_i(x) = 0. \tag{9}$$

Note that s_i is dependent on x, this equation is difficult to solve. However, if s_i is given for every i, then solving Eq. (9) is equivalent to solving the following problem:

$$\min f(x) + \sum_i s_i g_i^T(x) g_i(x). \tag{10}$$

Based on the above analysis, we can use the following iterative algorithm to find the solution of Eq. (7), and thus the optimal solution of problem in Eq. (6) (Liu et al. 2017).

Algorithm 1: (Liu et al. 2017) The algorithm to solve the problem (6).

Initialize x;
while *not converge* **do**
  1. For each i, calculate s_i according to Eq. (8);
  2. Update x by solving the problem (10).
end

The convergence of Algorithm 1 is guaranteed by the following theorem.

Theorem 1 *(Liu et al. 2017) The Algorithm 1 will monotonically decrease the objective of the problem (6) in each iteration.*

Equipped with the smoothed iterative reweighted method, our proposed objective in Eq. (4) can be solved by the iterative procedures in Algorithm 1, where Step 2 is to minimize the following objective:

$$\begin{aligned}
&\mathcal{J}^{\mathrm{R}}_{\ell_2}(\mathcal{W}) \\
&= \min_{\mathcal{W}} \sum_{i=1}^{n} \mathrm{tr}\left(\mathbf{x}_i - \mathbf{W}_0\mathbf{W}_0^T\mathbf{x}_i\right)^T \mathbf{D}_1^i \left(\mathbf{x}_i - \mathbf{W}_0\mathbf{W}_0^T\mathbf{x}_i\right) \\
&+ \gamma_1 \sum_{i=1}^{n} \mathrm{tr}\left(\mathbf{W}_i^T\mathbf{X}_i\mathbf{L}_i\mathbf{X}_i^T\mathbf{W}_i\right) \\
&+ \gamma_2 \sum_{i=1}^{n} \mathrm{tr}\left(\mathbf{W}_0 - \mathbf{W}_i\right)^T \mathbf{D}_2^i \left(\mathbf{W}_0 - \mathbf{W}_i\right) \\
&+ \gamma_3 \sum_{i=0}^{n} \mathrm{tr}\left(\mathbf{W}_i^T\mathbf{D}_3^i\mathbf{W}_i\right), \\
&s.t. \quad \mathbf{W}_i^T\mathbf{W}_i = \mathbf{I} \quad (0 \le i \le n).
\end{aligned} \tag{11}$$

Here $\mathbf{D}_1^i$ is a diagonal matrix whose j-th diagonal element is $\frac{1}{2}\left(e_{ij}^2 + \delta\right)^{-\frac{1}{2}}$, e_{ij} is the j-th element of vector $\mathbf{e}_i = \mathbf{x}_i - \mathbf{W}_0\mathbf{W}_0^T\mathbf{x}_i$. $\mathbf{L}^i = \mathbf{D}^i - \tilde{\mathbf{S}}^i$ in which $\mathbf{D}^i$ is a diagonal matrix whose entries are column (or row) sums of $\tilde{\mathbf{S}}_i$, *i.e.*, the j-th diagonal element of $\mathbf{D}^i$ is $\sum_j \tilde{s}_{jk}$. The element of $\tilde{\mathbf{S}}_i \in \Re^{n_i \times n_i}$ is computed as $\tilde{s}_{jk}^i = \theta_{jk}^i s_{jk}^i$, where $\theta_{jk}^i = \frac{1}{2}\left(\left\|\mathbf{W}_i^T\mathbf{x}_j^i - \mathbf{W}_i^T\mathbf{x}_k^i\right\|_2^2 + \delta\right)^{-\frac{1}{2}}$ and $\delta \to 0$. $\mathbf{D}_2^i$ is a diagonal matrix whose j-th diagonal element is $\frac{1}{2}\left(\left\|\mathbf{w}_0^j - \mathbf{w}_i^j\right\|_2^2 + \delta\right)^{-\frac{1}{2}}$, where $\mathbf{w}_i^j$ is the j-th column vector of $\mathbf{W}_i$. $\mathbf{D}_3^i$ is a block diagonal matrix whose j-th diagonal block is $\frac{1}{2}\left(\left\|\mathbf{W}_i^j\right\|_2^2 + \delta\right)^{-\frac{1}{2}}\mathbf{I}_k$, where $\mathbf{I}_k \in \Re^{d_j \times d_j}$ is an identity matrix. $\mathbf{W}_i^j \in \Re^{d_j \times r}$ is the j-th block of $\mathbf{W}_i$, where $\mathbf{W}_i = [\mathbf{W}_i^1; \mathbf{W}_i^2; \cdots; \mathbf{W}_i^k]$.

The Algorithm to Solve Eq. (11)

Before giving our derivation details to solve Eq. (11), we will first introduce the Alternating Direction Method of Multipliers (ADMM), which was proposed in (Bertsekas 1996; Boyd et al. 2011) to solve convex optimization problems by breaking them into smaller pieces that are easier to handle.

Specifically, given the following objective with the equality constraint:

$$\min_{x,z} f(x) + g(z), \qquad s.t. \quad h(x,z) = 0, \tag{12}$$

Algorithm 2 solves the problem by decoupling it into subproblems and optimizing each variable while fixing others (Bertsekas 1996; Boyd et al. 2011), where y is the Lagrangian multiplier to the constraint h. It is worth noting that Algorithm 2 was proved to converge Q-linearly to the optimal solution (Bertsekas 1996).

Algorithm 2: The ADMM algorithm.

Set $1 < \rho < 2$ and initialize $\mu > 0$ and y;
while *not converge* **do**
  1. Update x by solving
  $x^{k+1} = \arg\min_x (f(x) + \frac{\mu}{2}\|h(x, z^k) + \frac{y^k}{\mu}\|^2)$;
  2. Update z by solving
  $z^{k+1} = \arg\min_z (g(z) + \frac{\mu}{2}\|h(x^{k+1}, z) + \frac{y^k}{\mu}\|^2)$;
  3. Update y by $y^{k+1} = y^k + \mu h(x^{k+1}, z^{k+1})$;
  4. Update μ by $\mu = \rho\mu$.
end

Using ADMM, we rewrite the objective in Eq. (11) as:

$$\begin{aligned}
&\mathcal{J}^{\mathrm{ADMM}}_{\ell_2}(\mathcal{W}, \mathcal{P}) \\
&= \min_{\mathcal{W},\mathcal{P}} \sum_{i=1}^{n} \mathrm{tr}\left(\mathbf{x}_i - \mathbf{P}\mathbf{P}^T\mathbf{x}_i\right)^T \mathbf{D}_1^i \left(\mathbf{x}_i - \mathbf{W}_0\mathbf{W}_0^T\mathbf{x}_i\right) \\
&+ \gamma_1 \sum_{i=1}^{n} \mathrm{tr}\left(\mathbf{W}_i^T\mathbf{X}_i\mathbf{L}_i\mathbf{X}_i^T\mathbf{W}_i\right) \\
&+ \gamma_2 \sum_{i=1}^{n} \mathrm{tr}\left(\mathbf{W}_0 - \mathbf{W}_i\right)^T \mathbf{D}_2^i \left(\mathbf{W}_0 - \mathbf{W}_i\right) \\
&+ \gamma_3 \sum_{i=0}^{n} \mathrm{tr}\left(\mathbf{W}_i^T\mathbf{D}_3^i\mathbf{W}_i\right) + \frac{\mu}{2}\left\|\mathbf{W}_0 - \mathbf{P} + \frac{1}{\mu}\mathbf{\Lambda}\right\|_2^2 \\
&+ \sum_{i=0}^{n} \frac{\mu}{2}\left\|\mathbf{W}_i - \mathbf{P}_i + \frac{1}{\mu}\mathbf{\Lambda}_i\right\|_2^2, \\
&s.t. \quad \mathbf{P}_i^T\mathbf{P}_i = \mathbf{I} \quad (0 \le i \le n),
\end{aligned} \tag{13}$$

where $\mathcal{P} = \{\mathbf{P}, \mathbf{P}_0, \mathbf{P}_1, \mathbf{P}_2, \cdots, \mathbf{P}_n\}$. $\mathbf{\Lambda}_i \in \Re^{d\times r}$ $(0 \le i \le n)$ is the Lagrangian multiplier for the constraint of $\mathbf{W}_i = \mathbf{P}_i$. $\mathbf{\Lambda} \in \Re^{d\times r}$ is the Lagrangian multiplier for the constraint of $\mathbf{W}_0 = \mathbf{P}$. The solution algorithm using the ADMM is summarized in Algorithm 3.

Algorithm 3: Solve the optimization problem in Eq. (13).

Initialization:$\mathbf{W}_i$, $\mathbf{P}$, $\mathbf{P}_i$, $\mathbf{\Lambda}$, $\mathbf{\Lambda}_i$ $(0 \le i \le n)$, $1 < \rho < 2$, $\mu, \gamma_1, \gamma_2, \gamma_3 > 0$;
while *not converge* **do**
1. Update $\mathbf{W}_i$ $(1 \le i \le n)$ using $\mathbf{W}_i = (2\gamma_1 \mathbf{X}_i \mathbf{L}_i \mathbf{X}_i^T + 2\gamma_2 \mathbf{D}_2^i + 2\gamma_3 \mathbf{D}_3^i + \mu \mathbf{I})^{-1} \left(2\gamma_2 \mathbf{D}_2^i \mathbf{W}_0 + \mu \mathbf{P}_i - \mathbf{\Lambda}_i\right)$;
2. Update $\mathbf{W}_0$ using $\mathbf{W}_0 = (\sum_{i=1}^n \mathbf{x}_i(\mathbf{x}_i^T \mathbf{P}\mathbf{P}^T - \mathbf{x}_i^T)\mathbf{D}_1^i + \gamma_2 \sum_{i=1}^n \mathbf{D}_2^i + \gamma_3 \mathbf{D}_3^i + \mu \mathbf{I})^{-1}(\gamma_2 \sum_{i=1}^n \mathbf{D}_2^i \mathbf{W}_i + \frac{\mu}{2}\mathbf{P} + \frac{\mu}{2}\mathbf{P}_0 - \frac{\mathbf{\Lambda}}{2} - \frac{\mathbf{\Lambda}_0}{2})$;
3. Update $\mathbf{P}_i$ $(0 \le i \le n)$ using $\mathbf{P}_i = \mathbf{U}_i \mathbf{V}_i^T$, where $\mathbf{N}_i = \mu \mathbf{W}_i + \mathbf{\Lambda}_i$ and $\mathrm{svd}\,(\mathbf{N}_i) = \mathbf{U}_i \mathbf{\Sigma}_i \mathbf{V}_i^T$;
4. Update $\mathbf{P}$ using $\mathbf{P} = (2\sum_{i=1}^n \mathbf{D}_1^i (\mathbf{W}_0 \mathbf{W}_0^T \mathbf{x}_i - \mathbf{x}_i)\mathbf{x}_i^T + \mu \mathbf{I})^{-1}(\mu \mathbf{W}_0 + \mathbf{\Lambda})$;
5. Update $\mathbf{\Lambda}_i$ $(0 \le i \le n)$ using $\mathbf{\Lambda}_i = \mathbf{\Lambda}_i + \mu\,(\mathbf{W}_i - \mathbf{P}_i)$;
6. Update $\mathbf{\Lambda}$ using $\mathbf{\Lambda} = \mathbf{\Lambda} + \mu\,(\mathbf{W}_0 - \mathbf{P})$;
7. Update μ using $\mu = \rho\mu$;
end
Output: $\mathbf{W}_i$ $(0 \le i \le n)$.

Experiments

Data used in the preparation of the experiments were obtained from the ADNI database (Weiner et al. 2010). 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 (Risacher et al. 2010) 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), Fluency test (FLU), Rey's Auditory Verbal Learning Test (RAVLT) and Trail making test (TRAILS). The time points examined in this study for both imaging biomarkers and cognitive assessments includes baseline (BL), Month 6 (M6), Month 12 (M12), Month 24 (M24) and Month 36 (M36). All the participants' data used in our enriched biomarker representation study are required to have a BL MRI measurement, BL cognitive score and at least three available measures from M6/M12/M24/36. A total of 456 sample subjects are involved in our study, among which we have 77 AD samples, and 171 MCI samples and 208 HC samples.

Experimental Settings

To validate the usefulness of our proposed method, we compare the performance to predict cognitive outcomes using two types of the neuroimaging inputs — the learned enriched representation and baseline (BL) biomarker measurement. In our experiments, several methods proven to generalize well, such as ridge regression (RR), Lasso, support vector regression (SVR), and convolutional neural networks (CNN), are leveraged. For RR, Lasso and SVR models, we conduct a standard 5-fold cross-validation approach and compute the root mean square error (RMSE) between the predicted values and ground truth values of the cognitive scores on the testing data. For the CNN regression model, we construct a two layer convolution architecture for the cognitive outcomes prediction and dropout technique is also leveraged to reduce overfitting in CNN models and prevent complex co-adaptations on training data. For the model parameters, reduced dimension r is studied in $\{40, 60, \ldots, 180, 200\}$ and γ_1,γ_2 γ_3 are fine tuned by searching the grid of $\{10^{-5}, \ldots, 10^{-1}, 1, 10, \cdots, 10^5\}$.

Experimental Results

From Table 1, we can see that the proposed enriched neuroimaging representation is consistently better than baseline biomarker representations when we use them in the four different regression methods – RR, Lasso, SVR and CNN. This observation can be attributed to the following reasons. First, the original baseline biomarker representation only deals with one single cognitive measure. It does not benefit from the correlation across different cognitive measures over time. Instead, our proposed enriched biomarker representation could capture not only the baseline cognitive measurement, but also the temporal information conveyed by the longitudinal biomarkers over AD progressions. Our enriched representation could integrate the neuroimaging measurements at the fixed time point and the dynamic temporal changes. As AD is a progressively degenerative disease, incorporation of future information about subjects benefits the prediction model. Second, the original baseline neuroimaging measurements are of high dimensionality, which could be redundant and noisy. Thus the traditional methods may easily suffer from "the curse of dimensionality". Via the projection, we map the baseline cognitive measurement into a low dimensional space thereby mitigating the issue of high dimensionality.

Besides of the prediction capability comparison between original representation and enriched representation, we also explore prediction performance of our model when integrating different types of bioimaging data. From Figure 1, we can see that our proposed enriched biomarker representation achieves its peak performance when we leverage all available biomarkers – VBM, FreeSurfer and SNP.

Identification of Disease Relevant Imaging Biomarkers

Apart from the cognitive outcomes prediction task, another primary goal of our regression analysis is to identify a subset

Table 1: Experiment results comparison between original representation and enriched representation using all available biomarkers to predict clinical scores of MMSE, FLU, RAVLT, ADAS and TRAILS. We compare four different general regression methods – RR, Lasso SVR and CNN. The root mean squared error (RMSE) value for each cognitive outcome is calculated for comparison. The reduced dimension r is set to 60.

Cognitive Scores		RR	Lasso	SVR	CNN
MMSE	Original Representation	0.4188	0.3627	0.475	0.1483
	Enriched Representation	0.2928	0.3219	0.3576	0.1418
FLU_AMIN	Original Representation	1.5515	0.677	0.2523	0.2815
	Enriched Representation	1.3091	0.6331	0.1941	0.2469
FLU_VEG	Original Representation	3.1384	0.7716	0.2313	0.2049
	Enriched Representation	2.6735	0.6545	0.2015	0.1722
RAVLT_TOTAL	Original Representation	3.0850	2.8242	1.4471	0.7862
	Enriched Representation	2.3676	2.2019	1.1986	0.7086
RAVLT_30	Original Representation	4.1333	1.6309	0.3926	0.3926
	Enriched Representation	3.0221	1.2494	0.3011	0.3501
RAVLT_REC	Original Representation	6.2365	0.5368	0.6009	0.3923
	Enriched Representation	4.5648	0.4030	0.5466	0.3668
ADAS	Original Representation	2.3186	1.5073	0.6731	0.9504
	Enriched Representation	1.9879	1.3212	0.5985	0.8951
TRAILA	Original Representation	85.1665	36.5110	15.8871	7.3400
	Enriched Representation	71.7873	34.3885	15.0620	7.2555
TRAILB	Original Representation	43.5337	25.3341	13.9827	5.0018
	Enriched Representation	36.5486	21.6145	12.1877	4.8303

of biomarkers which are highly correlated to AD progressions. Thus, we examine the biomarkers of each participant identified by the proposed methods encoded by the cognitive scores.

From the formulation in Eq. (4), we learn a global projection $\mathbf{W}_0$ which summarize all the most important biomarkers across all the participants. Therefore, we plot the weights of each region of VBM and FreeSurfer from global projection $\mathbf{W}_0$, shown in Figure 2. From Figure 2, we can see that the bilateral hippocampus, amygdala regions in VBM and bilateral cerebral white matter regions in FreeSurfer are found to be in the top selected biomarkers by our model. The hippocampus is a small organ located within the brain's medial temporal lobe and forms an important part of the limbic system, the region that regulates emotions. The hippocampus is associated mainly with memory, in particular long-term memory (Mu and Gage 2011). The amygdala performs a primary role in the processing of memory, decision-making and emotional response (Amunts et al. 2005).

In summary, the identified imaging biomarkers are highly suggestive and strongly agree with existing medical research findings with regards to AD, which warrants the correctness of the discovered imaging cognition associations to reveal the complex relationships between biomarkers and cognitive scores. This is important for both theoretical research and clinical practices for a better understanding of AD mechanism.

Conclusions

Missing data is a critical challenge in longitudinal multimodal AD studies. In this paper, we propose a formulation to learn a consistent length representation for all the participants in ADNI dataset. The enriched fixed length biomarker representation could capture the global consistency from baseline measurements and local pairwise pattern from available follow-up measurements of each participant at the same time from heterogeneous biomarker measurements. Our results show that our enriched representation beat the performance of the baseline measurement when predicting the clinical scores. Furthermore, the identified biomarkers are highly suggestive and strongly agree with the existing research findings, which warrants the correctness of our approach. This is important for both theoretical research and clinical practices for a better understanding of AD mechanism.

Acknowledgments

This work was partially supported by the National Science Foundation (NSF) under the grants of IIS 1652943, IIS

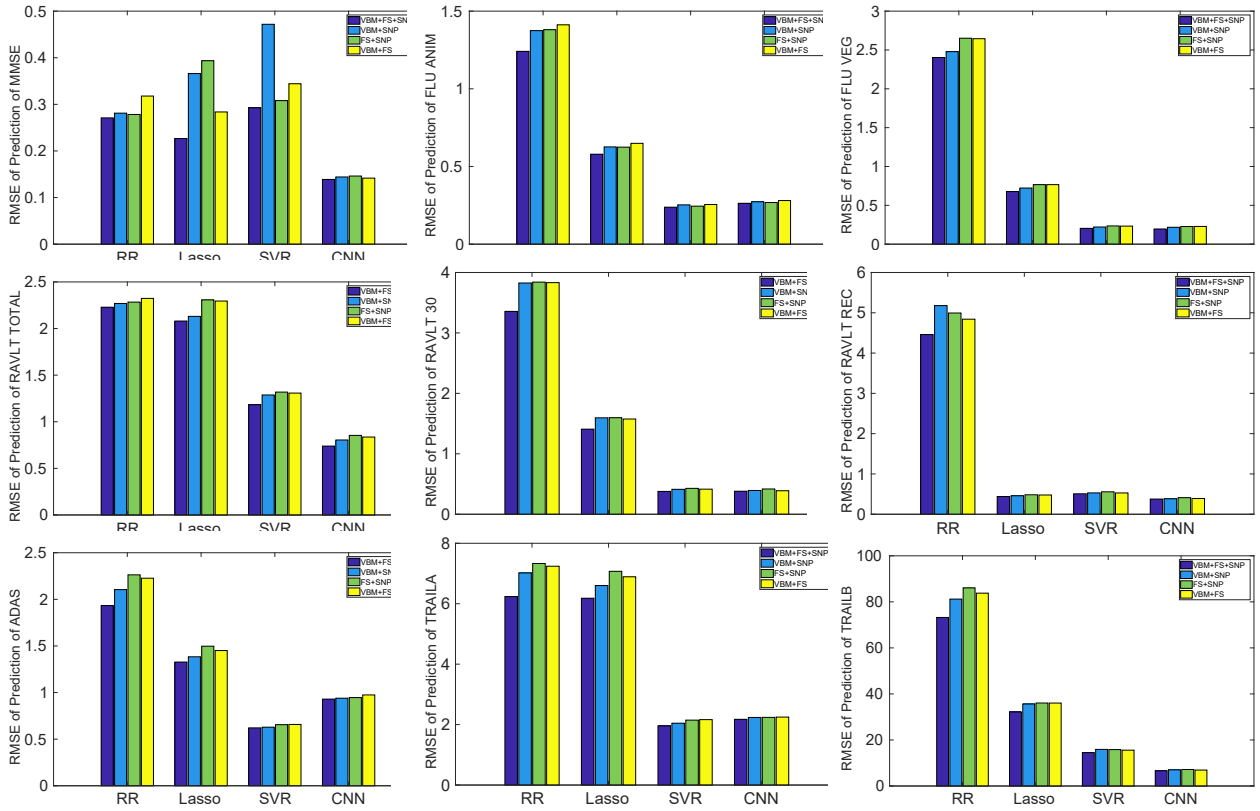

Figure 1: Experiment results using different combinations of heterogeneous neuroimaging sources to predict clinical scores of MMSE, FLU, RAVLT, ADAS and TRAILS. We compare four different general regression methods – RR, Lasso SVR and CNN. The root mean squared error (RMSE) value for each cognitive outcome is calculated for comparison. The reduced dimension r is set to 60.

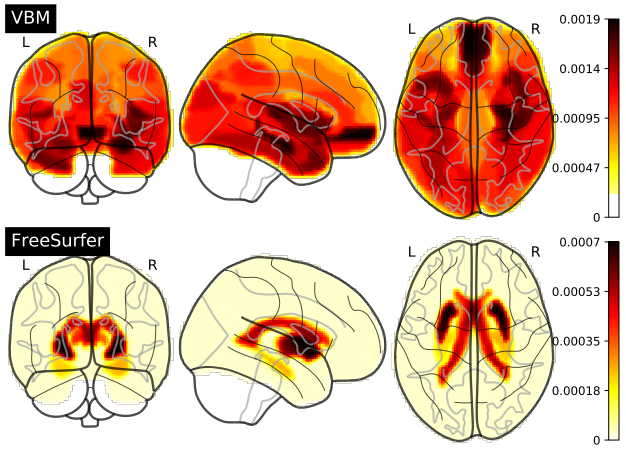

Figure 2: Weights of imaging markers of VBM (Up) and FreeSurfer (Down).

1849359, and CNS 1932482.

We thank Dr. Heng Huang at the Department of Electrical and Computer Engineering, University of Pittsburgh and Dr. Li Shen at the Department of Biostatistics, Epidemiology and Informatics Perelman School of Medicine, University of Pennsylvania for providing the data used in all our experiments.

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.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; 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 facili-

tated 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 Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

Amunts, K.; Kedo, O.; Kindler, M.; Pieperhoff, P.; Mohlberg, H.; Shah, N.; Habel, U.; Schneider, F.; and Zilles, K. 2005. Cytoarchitectonic mapping of the human amygdala, hippocampal region and entorhinal cortex: intersubject variability and probability maps. *Anatomy and embryology* 210(5-6):343–352.

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

Bertsekas, D. P. 1996. *Constrained optimization and Lagrange multiplier methods*. Athena Scientific.

Boyd, S.; Parikh, N.; Chu, E.; Peleato, B.; Eckstein, J.; et al. 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning* 3(1):1–122.

Brand, L.; Wang, H.; Huang, H.; Risacher, S.; Saykin, A.; Shen, L.; et al. 2018. 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*, 555–562. Springer.

He, X., and Niyogi, P. 2004. Locality preserving projections. In *Advances in neural information processing systems*, 153–160.

Jolliffe, I. 2011. Principal component analysis. In *International encyclopedia of statistical science*. Springer. 1094–1096.

Li, Y.; Wang, L.; Zhou, J.; and Ye, J. 2019. Multi-task learning based survival analysis for multi-source block-wise missing data. *Neurocomputing*.

Liu, Y.; Guo, Y.; Wang, H.; Nie, F.; and Huang, H. 2017. Semi-supervised classifications via elastic and robust embedding. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Lu, L.; Wang, H.; Yao, X.; Risacher, S.; Saykin, A.; and Shen, L. 2018. 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)*, 545–548. IEEE.

Mu, Y., and Gage, F. H. 2011. Adult hippocampal neurogenesis and its role in alzheimer's disease. *Mol. neurodegeneration* 6(1):85.

Risacher, S. L.; Shen, L.; West, J. D.; Kim, S.; McDonald, B. C.; Beckett, L. A.; Harvey, D. J.; Jack Jr, C. R.; Weiner, M. W.; Saykin, A. J.; et al. 2010. Longitudinal mri atrophy biomarkers: relationship to conversion in the adni cohort. *Neurobiology of aging* 31(8):1401–1418.

Stonnington, C. M.; Chu, C.; Klöppel, S.; Jack Jr, C. R.; Ashburner, J.; Frackowiak, R. S.; Initiative, A. D. N.; et al. 2010. Predicting clinical scores from magnetic resonance scans in alzheimer's disease. *Neuroimage* 51(4):1405–1413.

Wang, H.; Nie, F.; Huang, H.; Risacher, S. L.; Saykin, A. J.; Shen, L.; and Initiative, A. D. N. 2012a. Identifying disease sensitive and quantitative trait-relevant biomarkers from multidimensional heterogeneous imaging genetics data via sparse multimodal multitask learning. *Bioinformatics* 28(12):i127–i136.

Wang, H.; Nie, F.; Huang, H.; Yan, J.; Kim, S.; Nho, K.; Risacher, S. L.; Saykin, A. J.; Shen, L.; and Initiative, A. D. N. 2012b. From phenotype to genotype: an association study of longitudinal phenotypic markers to alzheimer's disease relevant snps. *Bioinformatics* 28(18):i619–i625.

Wang, H.; Nie, F.; Huang, H.; Yan, J.; Kim, S.; Risacher, S.; Saykin, A.; and Shen, L. 2012c. High-order multi-task feature learning to identify longitudinal phenotypic markers for alzheimer's disease progression prediction. In *NIPS*, 1277–1285.

Wang, H.; Nie, F.; Huang, H.; and Ding, C. 2013. Heterogeneous visual features fusion via sparse multimodal machine. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, X.; Yan, J.; Yao, X.; Kim, S.; Nho, K.; Risacher, S. L.; Saykin, A. J.; Shen, L.; Huang, H.; et al. 2017. Longitudinal genotype-phenotype association study via temporal structure auto-learning predictive model. In *RECOMB*, 287–302. Springer.

Wang, H.; Nie, F.; and Huang, H. 2013. Multi-view clustering and feature learning via structured sparsity. In *International conference on machine learning*, 352–360.

Wang, H.; Nie, F.; and Huang, H. 2015. Learning robust locality preserving projection via p-order minimization. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Wang, X.; Shen, D.; and Huang, H. 2016. Prediction of memory impairment with mri data: a longitudinal study of alzheimer's disease. In *MICCAI*, 273–281. Springer.

Weiner, M. W.; Aisen, P. S.; Jack Jr, C. R.; Jagust, W. J.; Trojanowski, J. Q.; Shaw, L.; Saykin, A. J.; Morris, J. C.; Cairns, N.; Beckett, L. A.; et al. 2010. The alzheimer's disease neuroimaging initiative: progress report and future plans. *Alzheimer's & Dementia* 6(3):202–211.

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

Xiang, S.; Yuan, L.; Fan, W.; Wang, Y.; Thompson, P. M.; Ye, J.; Initiative, A. D. N.; et al. 2014. Bi-level multi-source learning for heterogeneous block-wise missing data. *NeuroImage* 102:192–206.

Yan, J.; Li, T.; Wang, H.; Huang, H.; Wan, J.; Nho, K.; Kim, S.; Risacher, S. L.; Saykin, A. J.; Shen, L.; et al. 2015. Cortical surface biomarkers for predicting cognitive outcomes using group $\ell_{2,1}$ norm. *Neurobiology of aging* 36:S185–S193.

Zhang, D.; Shen, D.; Initiative, A. D. N.; et al. 2012. Multimodal multi-task learning for joint prediction of multiple regression and classification variables in alzheimer's disease. *NeuroImage* 59(2):895–907.