Predicting Cognitive Declines Using Longitudinally Enriched Representations for Imaging Biomarkers

Lyujian Lu, Saad Elbeleidy, Lauren Zoe Baker, Hua Wang, and Feiping Nie, *for the ADNI*

***Abstract*—A critical challenge in using longitudinal neuroimaging data to study the progressions of Alzheimer's Disease (AD) is the varied number of missing records of the patients during the course when AD develops. To tackle this problem, in this paper we propose a novel formulation to learn an enriched representation with fixed length for imaging biomarkers, which aims to simultaneously capture the information conveyed by both baseline neuroimaging record and progressive variations characterized by varied counts of available follow-up records over time. Because the learned biomarker representations are a set of fixed-length vectors, they can be readily used by traditional machine learning models to study AD developments. Take into account that the missing brain scans are not aligned in terms of time in a studied cohort, we develop a new objective that maximizes 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 and non-greedy iterative solution algorithm and rigorously prove its convergence. We have performed extensive experiments on the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort. A clear performance gain has been achieved in predicting ten different cognitive scores when we compare the original baseline biomarker representations against the learned representations with longitudinal enrichments. We further observe that the top selected biomarkers by our new method are in accordance with known knowledge in AD studies. These promising results have demonstrated improved performances of our new method that validate its effectiveness.**[1]

***Index Terms*—Alzheimer's Disease, Longitudinal, Representation Enrichment, Imaging Biomarker.**

I. Introduction

As one of the most prevalent and severe type of neurodegenerative disorders [1], [2], Alzheimer's Disease (AD) has attracted growing attentions in research in recent years. Over the past decade, phenotypic biomarkers extracted from brain images have been widely studied to predict disease status and/or cognitive performance [3], [4], [5]. However, these approaches routinely perform standard regression and/or classification at each 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 relations among the longitudinal records of the brain imaging biomarkers.

In the study of the Alzheimer's Disease Neuroimaging Initiative (ADNI) [6], participants return for follow-up scans at varied 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), as illustrated in Fig. 1, which provides the opportunity to use longitudinal data from multiple time points to build more effective predictive models. To explore the temporal structure of brain phenotypes, longitudinal prediction models [5], [7], [8] have been recently proposed. However, in these studies longitudinal information has been modeled as tensors, which inevitably complicates the problem in mathematics. As a result, it is not easy to extend classical machine learning models, which can only work with vector or matrix data, to study AD developments.

Another critical challenge in using longitudinal data is the problem of missing data in the medical records. Higher mortality risk and cognitive impairment hinder older adults from staying in studies that require multiple visits and thus result in incomplete data [9]. The missing imaging records at different time points lead to samples with varied lengths for different participants. To deal with this problem, many existing longitudinal studies of AD only utilize data samples with complete temporal records for analyses and ignore those with fewer records over time [5], [7], [8]. Apparently, discarding the samples with less temporal records could potentially ruin the dataset. To address this, data imputation methods [9], [10] have been proposed to handle the missing records in longitudinal AD data. Using the imputed data with a consistent sample size, regression and classification studies can be conducted. However, whether these data completion methods can preserve the longitudinal structure of neuroimaging measurements or not is still an under-explored topic in AD studies. What's worse, these missing data imputation methods could possibly introduce undesirable artifacts that may worsen the predictive power of the learned longitudinal models.

To tackle the above problems in longitudinal studies with incomplete temporal inputs, in this paper we propose a novel formulation to learn an enriched biomarker representation which combines the baseline biomarker measurements and the dynamic temporal imaging records across the follow-up time points. In our learning framework, we learn a projection for each participant from her or his biomarker records at all available follow-up time points (a subset of {M6, M12, M18,

Manuscript received April 26, 2020; revised November 2, 2020; accepted November 23, 2020. This work was supported in part by the National Science Foundation (NSF) under the grants of IIS 1652943, IIS 1849359, CNS 1932482 and CCF 2029543. (Corresponding author: Hua Wang.)

L. Lu, S. Elbeleidy, L. Baker and H. Wang are with the Department of Computer Science, Colorado School of Mines, Golden, CO 80401, USA (email: lyujianlu@mymail.mines.edu, selbeleidy@mymail.mines.edu, laurenzoebaker@mymail.mines.edu, huawangcs@gmail.com).

F. Nie is with School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, P. R. China (email: feipingnie@gmail.com).

[1]The code package for learning longitudinally enriched imaging biomarker representations reported in this paper have been made publicly available online at https://github.com/lyujian/Learning-Longitudinally-Enriched-Imaging-Biomarker-Representations-.

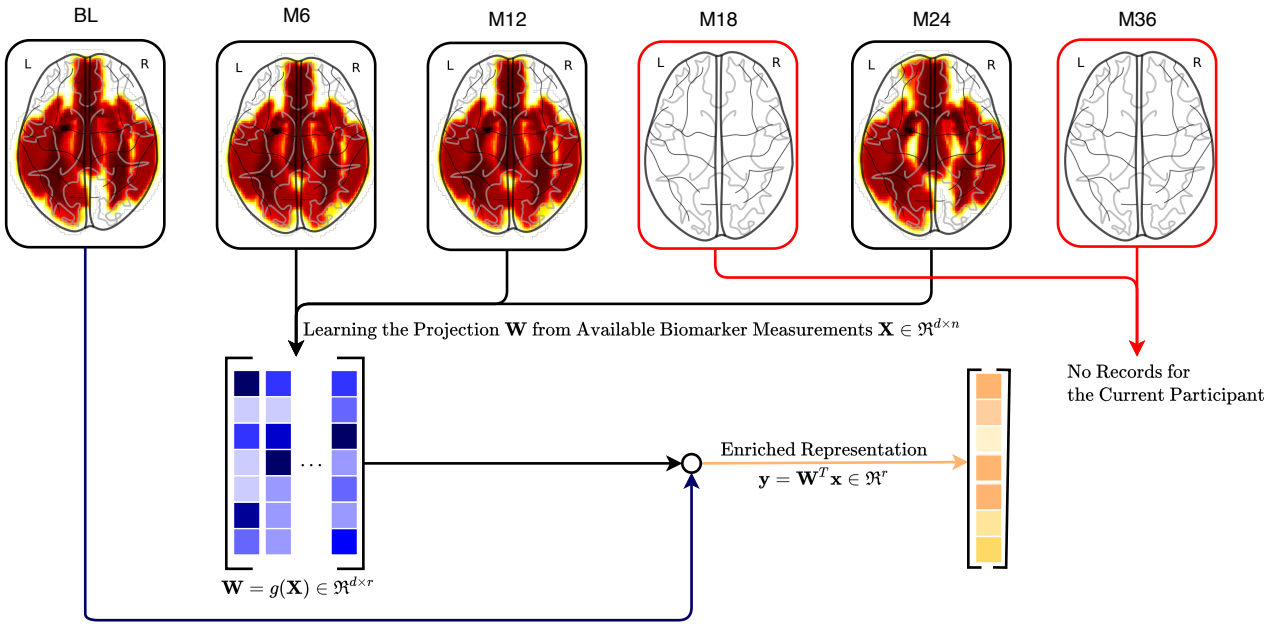

Figure 1. Overview of our proposed method to learn an enriched neuroimaging representation with a fixed length, which integrates the baseline biomarker measurements and dynamic changes in available follow-up biomarker measurements. The blank plots in M18 and M36 denote the absence of the scans of the currently studied participant in the 18th month and the 36th month.

M24, M36}), by which we project the baseline record into a fixed-length vector, regardless of the inconsistent number of brain scans of the participants in a dataset. Armed with the fixed-length biomarker representations, we can directly use conventional learning models to predict cognitive outcomes.

As schematically illustrated in Fig. 1, the proposed method first learns a projection from the available follow-up imaging records, which we use to project the baseline neuroimaging record to learn a fixed-length biomarker representation. Through this procedure, the learned representation for a participant simultaneously captures the information conveyed by both the baseline neuroimaging record and a progressive summary of all available follow-up records, such that the baseline representation of the participant is ***enriched*** by her or his follow-up longitudinal information. We further develop our learning objective by replacing the traditional squared ℓ_2-norm distances by the ℓ_1-norm distances in our formulation, to improve the robustness of the learned enriched representations against possible outlying samples and features caused by the varied number of brain scans taken at different time points by different participants in a studied cohort. Despite its clear motivation, the developed objective ends up being a non-smooth 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 and non-greedy iterative algorithm with theoretically guaranteed convergence. We have performed extensive experiments on the ADNI cohort that demonstrate the improved performance resulting from our new approach. Moreover, we select the top 10 biomarkers weighted by their predictive power in cognitive tests, which are highly suggestive and strongly agree with the existing research findings.

This paper is an extension of our recent work [11] originally reported in the Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020 (CVPR 2020). In this extended journal manuscript, we provide the following expansions over its conference version:

- We outline the mathematical details for deriving the algorithm to solve our objective and show that it is non-greedy in nature, where we expend a concrete effort to improve the mathematical details to unambiguously communicate the implementation of our algorithm. (Section IV)
- We rigorously prove the convergence of the solution algorithm in mathematics. (Section IV-C)
- We significantly expand the experimental evaluations to illustrate the benefits of using the enriched biomarker representations learned by our new method. (Section V)
  - We report new experimental results by using 1 additional type of imaging biomarkers (the FreeSurfer biomarkers) as input and 6 additional cognitive scores as predictive targets. (Section V-B)
  - We compare the proposed method against three recent longitudinal learning models using both Voxel-Based Morphometry (VBM) and FreeSurfer biomarkers respectively. (Section V-C)
  - We provide a thorough analysis of the identified disease relevant biomarkers to justify the clinical correctness of our new method. (Section V-E)

II. Related Work

A. AD studies using longitudinal neuroimaging data

To explore the temporal correlations of the variations of the neuroimaging markers over AD progressions, longitudinal features [12], [13], [14] were studied for predicting cognitive outcomes. For example, in [12] a longitudinal feature estimation method was proposed to capture temporal information

that can characterize the changes of specific brain regions over time; in [13] landmark-based spatial and longitudinal features were leveraged to identify AD subjects; in [14] a screening group regularization was utilized to select top consistent and varying imaging features.

To exploit the collective correlations of cognitive score changes during the course when AD develops, many longitudinal multi-task methods were proposed [8], [15], [5], [7], [16], [17], [14], [18], [19], [20], [21]. Specifically, in [5] a high-order multi-task feature learning framework was presented for identifying longitudinal neuroimaging markers to predict cognitive scores across all time points. In [15], [17] longitudinal models were designed to associate genetic biomarkers with temporal imaging phenotypes. In [19], [20] a joint multi-modal longitudinal regression and classification model was proposed to simultaneously predict the cognitive scores and diagnoses of AD. In [7], [8] a auto-learning multi-task model was used to explore the associations between genetic variations and longitudinal imaging phenotypes, as well as interrelatedness that exists in different prediction tasks. In [16] multi-relational smoothness regularization was incorporated to capture the relationship among different clinical scores. In [14] a multi-task dictionary learning framework was devised to use both shared and individual dictionaries to encode both consistent and varying imaging features when AD develops. In [18] a multi-task exclusive relationship learning model was recently proposed to automatically capture the intrinsic relationship among tasks at different time points for estimating clinical measures based on longitudinal imaging data. In summary, a variety of sparsity-induced norms were leveraged for identifying AD related imaging biomarkers, including the trace-norm [5], Lasso [17], group Lasso [15], the $\ell_{2,1}$-norm [5], [17] and the Schatten p-norm [7], [8], to name a few.

These longitudinal learning models were successfully designed to make use of the longitudinal imaging and cognitive data, which, however, can only deal with data samples with complete temporal records over the disease progressions. As a result, the samples that miss certain medical scans have to be discarded, although they may contain crucial information for diagnoses of AD.

B. Missing data imputation in AD studies

To address the critical challenge of missing records in AD studies, many multi-task learning methods [22], [23], [14] were proposed to impute missing data by exploiting the correlations among different prediction tasks. In [22] a flexible feature selection method was developed to deal with missing data, which formulates the original classification problem as a multi-task learning problem to make full use of all available data. In [23] block-wise missing data collected from multiple sources were decomposed into the multiple completed sub-matrices, where a two-layer multi-task learning model was used for both feature-level and source-level analyses.

To utilize multi-modal data, recent studies [24], [25], [26], [9], [27] explored multi-view learning models for missing data imputation. In [24], [27] a unified feature-level and source-level model was developed to effectively integrate information from multiple heterogeneous sources when block-wise missing data are present. In [25], [26] a hypergraph learning method was proposed to represent the high-order relationships among the subjects by dividing them into groups according to modality availabilities, with a hypergraph regularization applied to each groups for making the final prediction. In [9] a sparse regression model was presented to explore the covariances from the data in multiple modalities.

More recently, deep learning models were developed for missing data imputation. In [28] a 3-dimensional (3D) convolutional neural network (CNN) was built to use a training set of subjects with simultaneously available MRI and PET records. The trained 3D CNN was then used to impute missing PET scans using the MRI data for the subjects who only had MRI scans. Besides, both adversarial neural networks [29] and recurrent neural networks [30] were also used to tackle the missing data problem for the ADNI dataset.

While these data imputation methods successfully solved the problem caused by the inconsistent sample sizes in many longitudinal datasets, the imputed data often have to be represented as tensors that may potentially complicate the subsequent learning models in mathematics. In addition, artifacts may be introduced into the imputed data due to the learning biases cased by the statistical assumptions that underlie these learning models.

III. Our Objective for Representation Learning

In this section, we formalize the problem of learning an enriched representation for neuroimaging biomarkers as a fixed-length vector for every participant using longitudinal data with missing medical records, with the goal to simultaneously capture the information conveyed by both the baseline imaging record and the progressive changes characterized by the follow-up records along the following time points.

A. Notations and the Problem Formalization

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

Given a longitudinal neoroimaging dataset, the temporal information of a participant can be denoted as: $\mathcal{X} = \{\mathbf{x}, \mathbf{X}\}$, where $\mathbf{x} \in \Re^d$ represents the baseline brain scan by d extracted neuroimaging features (biomarkers), and $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \Re^{d\times n}$ collects a total of n follow-up brain scans at later time points. Here we highlight that n ***varies*** across the dataset, because different participants in a studied cohort usually miss different numbers of brain scans at different time points. In the task of learning representations for neuroimaging biomarkers, our goal is to learn a ***fixed-length*** vector for every participant from the longitudinal records of $\mathcal{X}$. In this paper, we propose a general framework that uses the longitudinal data with misaligned medical records to learn a fixed-length vector enrichment for every participant. Specifically, first we learn a projection $\mathbf{W} = g(\mathbf{X})$ from $\mathbf{X}$ to summarize the temporal variations of the neuroimaging biomarkers along the time

points after the baseline. Then by applying the learned projection $\mathbf{W}$ onto the baseline neuroimaging record $\mathbf{x}$, we obtain a single fixed-length vector representation by computing $\mathbf{y} = f(\mathcal{X}) = h(g(\mathbf{X}), \mathbf{x}) \in \Re^r$. Apparently, the projection could be learned by different methods, such as Principal Component Analysis (PCA) [31], Locality Preserving Projection (LPP) [32], *etc*. In this paper, we propose a novel projection learning method which can simultaneously captures the information conveyed by both the baseline neuroimaging record and the dynamic changes of the follow-up neuroimaging records. Because the learned representations for all the participants in the entire dataset are of the same length, they can be readily used by traditional statistical and machine learning models in a variety of tasks, such as predicting cognitive outcomes.

B. Representation Learning through Projections

In this subsection, we develop a new objective to learn a fixed-length vector to represent the neuroimaging biomarkers that are directly extracted from brain scans. By integrating the baseline neuroimaging record and the dynamic temporal changes in follow-up neuroimaging records, we aim to preserve the global and local consistencies among the neuroimaging records in the subspace mapped by the learned projection.

First, although the neuroimaging measurements and cognitive status of a participant in a studied cohort could experience drastic changes over a long time, *e.g.*, a Healthy Control (HC) subject can be diagnosed with Mild Cognitive Impairments (MCI) or even converted into an AD patient in a couple of years, the changes of these quantities between nearby time points still remain considerably small [33]. Namely, the measurements of the biomarkers of the participants maintain the local consistency in terms of data magnitude during the progression of AD. Thus, we need preserve this local consistency by minimizing the local variance of the medical records collected in 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 $\overline{\mathbf{x}}_i = \frac{1}{K+1}\sum_{\mathbf{x}_j \in \{\mathcal{N}_i \cup \{\mathbf{x}_i\}\}} \mathbf{x}_j$. We can achieve the overall local consistency by minimizing the following objective [34]:

$$\mathcal{J}_{\text{Local}}\left(\mathbf{W}\right) = \text{tr}\left(\mathbf{W}^T \mathbf{S}_L \mathbf{W}\right), \quad \text{s.t.} \quad \mathbf{W}^T\mathbf{W} = \mathbf{I}, \tag{1}$$

where we define $\mathbf{S}_{Li} = \sum_{\mathbf{x}_j \in \{\mathcal{N}_i \cup \{\mathbf{x}_i\}\}} \left(\mathbf{x}_j - \overline{\mathbf{x}}_i\right)\left(\mathbf{x}_j - \overline{\mathbf{x}}_i\right)^T$ and $\mathbf{S}_L = \sum_{i=1}^n \mathbf{S}_{Li}$. Apparently, $\mathbf{S}_{Li}$ is the local covariance matrix of the data points around $\mathbf{x}_i$. Thus, minimizing $\mathbf{tr}\left(\mathbf{W}^T\mathbf{S}_{Li}\mathbf{W}\right)$ ensures the local consistency around $\mathbf{x}_i$ and minimizing $\mathcal{J}_{\text{Local}}$ in Eq. (1) ensures the overall local consistency of a subject's records across all the time points when AD develops, which is in accordance with the broadly used assumption in machine learning and data mining that data are smooth on an inherent manifold, *i.e.*, the observed data are sampled from an underlying sub-manifold that are embedded in a high-dimensional observation space [35], [32]. In Eq. (1), we omit the constant factor $\frac{1}{K+1}$ for notational brevity.

Second, apart from making use of the local consistency of the available neuroimaging records in the follow-up months, we further explore the global structure of all the neuroimaging records of a participant. Via a global projection, we map $\mathbf{X}$ that resides in the high d-dimensional space into a lower r-dimensional subspace by computing $\mathbf{y}_i = \mathbf{W}^T\mathbf{x}_i$ to preserve as much information as possible, for which we maximize the objective of the PCA [31]:

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

where $\mathbf{S}_G = \sum_{i=1}^n \left(\mathbf{x}_i - \overline{\mathbf{x}}\right)\left(\mathbf{x}_i - \overline{\mathbf{x}}\right)^T$ is the covariance matrix of $\mathbf{X}$ and $\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i$ is the mean vector. Again, the constant factor $\frac{1}{n}$ is omitted in Eq. (2) for notational brevity.

Now we integrate the global and local consistencies of the neuroimaging records of a subject by combining the two objectives in Eq. (1) and Eq. (2) to maximize the following objective:

$$\mathcal{J}_{\ell_2^2}(\mathbf{W}) = \frac{\sum_{i=1}^n \left\|\mathbf{W}^T\left(\mathbf{x}_i - \overline{\mathbf{x}}\right)\right\|_2^2}{\sum_{i=1}^n \sum_{\mathbf{x}_j \in \{\mathcal{N}_i \cup \{\mathbf{x}_i\}\}} \left\|\mathbf{W}^T\left(\mathbf{x}_j - \overline{\mathbf{x}}_i\right)\right\|_2^2}, \quad \text{s.t.} \quad \mathbf{W}^T\mathbf{W} = \mathbf{I}. \tag{3}$$

Finally, we notice that a critical challenge in using longitudinal AD data is their inconsistent sample sizes, *i.e.*, different patients may take brain scans at different time points. For example, one patient may take brain scans at the 12th month and the 24th month. In contrast, another patient might differ by taking brain scans in other months. That is, the brain scans of one patient are generally not aligned to others, which can potentially become outliers for one another when they are used to train a learning model. As studied in many recent papers [36], [37], [38], the squared ℓ_2-norm distance used in the objective in Eq. (3) is notoriously known to be very sensitive to outlying data samples and features. To address this, we choose to replace the squared ℓ_2-norm distance used in the objective in Eq. (3) by its ℓ_1-norm counterpart for promoting the robustness of our model against potential outlying effects, which leads to the following objective to maximize:

$$\mathcal{J}_{\ell_1}(\mathbf{W}) = \frac{\sum_{i=1}^n \left\|\mathbf{W}^T\left(\mathbf{x}_i - \overline{\mathbf{x}}\right)\right\|_1}{\sum_{i=1}^n \sum_{\mathbf{x}_j \in \{\mathcal{N}_i \cup \{\mathbf{x}_i\}\}} \left\|\mathbf{W}^T\left(\mathbf{x}_j - \overline{\mathbf{x}}_i\right)\right\|_1}, \quad \text{s.t.} \quad \mathbf{W}^T\mathbf{W} = \mathbf{I}. \tag{4}$$

Upon solving the optimization problem in Eq. (4), we compute $\mathbf{y} = \mathbf{W}^T\mathbf{x}$ to obtain the new biomarker representation for a subject, which enriches the baseline biomarker record $\mathbf{x}$ by the longitudinal AD developments of $\mathbf{X}$. This learned representation thereby not only preserves the global variance of the biomarker measurements over the entire course of the AD development of the subject, but also maintains the local geometric data structure of the medical records taken in nearby months in the projected subspace. Moreover, $\mathbf{y}$ is a fixed-length single-vector representation and can be readily used by most classical classification or regression models, which is the key contribution of this paper.

IV. The Algorithm to Solve Our Objective

The proposed objective in Eq. (4) maximizes the ratio of two summations of a number of ℓ_1-norm distances, which is non-smooth thereby difficult to efficiently solve in general. Thus,

in this section we derive an efficient iterative solution algorithm and rigorously prove its convergence. As an important theoretical contribution, our solution algorithm is non-greedy in nature.

A. Solving a General Ratio Maximization Problem

We first study the following general optimization problem and derive an efficient iterative algorithm to solve it:

$$\max_{v\in\Omega} \frac{h(v)}{m(v)}, \quad \forall v \in \Omega \;\; s.t. \begin{cases} C_2 \geq m(v) \geq C_1 > 0, \\ C_4 \geq h(v) \geq C_3 > 0, \end{cases} \tag{5}$$

where Ω is the feasible domain of the optimization problem, and C_1, C_2 C_3 and C_4 are four positive bounding constants.

Now we propose the following simple, yet efficient, iterative algorithm, as summarized in Algorithm 1, to optimize Eq. (5). The convergence of this algorithm is guaranteed by Theorem 1.

Algorithm 1: The algorithm to solve Eq. (5).

1. Randomly initialize $v^0 \in \Omega$ and set $k = 1$;
repeat
  2. Calculate $\lambda^k = \frac{h(v^{k-1})}{m(v^{k-1})}$;
  3. Find a $v^k \in \Omega$ satisfying $h(v^k) - \lambda^k m(v^k) > 0$;
  4. $k = k + 1$;
until *Convergence*
Output: v^k.

Theorem 1: In Algorithm 1, for each iteration (1) we have $\frac{h(v^k)}{m(v^k)} \geq \frac{h(v^{k-1})}{m(v^{k-1})}$; and (2) $\forall \delta > 0$, there exists a $\hat{k}$ such that $\forall k > \hat{k}$, $\frac{h(v^k)}{m(v^k)} - \frac{h(v^{k-1})}{m(v^{k-1})} < \delta$.

Proof 1: Step 3 of Algorithm 1 states that $h(v^k) - \lambda^k m(v^k) > 0$. Because $\forall v \in \Omega \;\; m(v) > 0$ as in the problem definition, we can derive $\frac{h(v^k)}{m(v^k)} > \lambda^k = \frac{h(v^{k-1})}{m(v^{k-1})}$, which completes the proof of the first statement of Theorem 1.

Suppose that for the k-th iteration, there exists a c^k such that $h(v^k) - \lambda^k m(v^k) = c^k > 0$. Then using the definition of λ^k in Step 2 of Algorithm 1, we have:

$$\frac{h(v^k)}{m(v^k)} = \frac{h(v^{k-1})}{m(v^{k-1})} + \frac{c_k}{m(v^k)} = \frac{h(v^0)}{m(v^0)} + \sum_{i=1}^{k} \frac{c_i}{m(v^i)}. \tag{6}$$

Because of the upper and lower bounds of $m(v)$ as defined in Eq. (5), from Eq. (6) we can derive:

$$\frac{h(v^0)}{m(v^0)} + \frac{1}{C_2}\sum_{i=1}^{k} c^i \leq \frac{h(v^k)}{m(v^k)} \leq \frac{h(v^0)}{m(v^0)} + \frac{1}{C_1}\sum_{i=1}^{k} c^i. \tag{7}$$

Now we suppose that there exists a positive constant C such that $\lim_{k\to\infty}\sum_{i=1}^{k} c^i = C$. If this is not true, we have $\lim_{k\to\infty}\sum_{i=1}^{k} c^i = \infty$, by which, together with Eq. (6), we can derive $\lim_{k\to\infty}\sum_{i=1}^{k} \frac{h(v^k)}{m(v^k)} = \infty$. This, however, contradicts the fact that $\frac{h(v^k)}{m(v^k)}$ is bounded as defined in Eq. (5), which means that $\lim_{k\to\infty}\sum_{i=1}^{k} c^i = C$ must hold. Thus, we have $\lim_{k\to\infty} c^k = 0$ and $\lim_{k\to\infty} \frac{c^k}{m(v^k)} = 0$, which indicates that $\forall \delta > 0$, there must exist a $\hat{k}$ such that:

$$\forall k > \hat{k}, \quad \frac{c^k}{m(v^k)} < \delta. \tag{8}$$

Putting Eq. (6) and Eq. (8) together, we can derive:

$$\forall k > \hat{k}, \quad \frac{h(v^k)}{m(v^k)} - \frac{h(v^{k-1})}{m(v^{k-1})} < \delta, \tag{9}$$

which proves the second statement of Theorem 1 and indicates that Algorithm 1 converges to a local optimum. ■

B. The Algorithm to Solve the Proposed Objective in Eq. (4)

Apparently, the proposed objective in Eq. (4) is a special case of the general ratio maximization problem in Eq. (5). Thus, to solve our objective, according to Step 3 of Algorithm 1, we need find a solution that satisfies the following inequality:

$$F(\mathbf{W}) = H(\mathbf{W}) - \lambda^k M(\mathbf{W}) > 0, \quad \text{s.t.} \;\; \mathbf{W}^T\mathbf{W} = \mathbf{I}, \tag{10}$$

where

$$\lambda^k = \frac{H(\mathbf{W}^{k-1})}{M(\mathbf{W}^{k-1})}, \tag{11}$$

and $\mathbf{W}^{k-1}$ denotes the projection matrix computed in the $(k-1)$-th iteration, which is already known in the k-th iteration. Here, for notational brevity, we define:

$$H(\mathbf{W}) = \sum_{i=1}^{n} \left\| \mathbf{W}^T (\mathbf{x}_i - \bar{\mathbf{x}}) \right\|_1, \tag{12}$$

$$M(\mathbf{W}) = \sum_{i=1}^{n} \sum_{\mathbf{x}_j \in \{\mathcal{N}_i \cup \{\mathbf{x}_i\}\}} \left\| \mathbf{W}^T (\mathbf{x}_j - \bar{\mathbf{x}}_i)) \right\|_1. \tag{13}$$

To find a $\mathbf{W}$ that satisfies the inequality in Eq. (10), we need the following two lemmas.

Lemma 1: [39, Theorem 1] For any vector $\boldsymbol{\xi} = [\xi_1, \cdots, \xi_m]^T \in \Re^m$, we have $\|\boldsymbol{\xi}\|_1 = \max_{\boldsymbol{\eta}\in\Re^m} (\text{sign}(\boldsymbol{\eta}))^T \boldsymbol{\xi}$. The maximum value is attained if and only if $\boldsymbol{\eta} = a \times \boldsymbol{\xi}$, where $a > 0$ is a scalar.

Lemma 2: [40, Lemma 3.1] For any vector $\boldsymbol{\xi} = [\xi_1, \cdots, \xi_m]^T \in \Re^m$, we have $\|\boldsymbol{\xi}\|_1 = \min_{\boldsymbol{\eta}\in\Re^m_+} \frac{1}{2}\sum_{i=1}^{m} \frac{\xi_i^2}{\eta_i} + \frac{1}{2}\|\boldsymbol{\eta}\|_1$, where the minimum value is attained if and only if $\eta_j = |\xi_j|, j \in \{1, 2, \cdots, m\}$.

To use Lemmas 1–2, we construct the following function:

$$L\left(\mathbf{W}, \mathbf{W}^{k-1}\right) = K\left(\mathbf{W}\right) - \lambda^k N\left(\mathbf{W}\right), \tag{14}$$

where $K\left(\mathbf{W}\right)$ and $\mathbf{N}\left(\mathbf{W}\right)$ are defined as:

$$K\left(\mathbf{W}\right) = \sum_{g=1}^{r} \mathbf{w}_g^T \mathbf{B}\, \text{sign}\left(\mathbf{B}^T \mathbf{w}_g^{k-1}\right), \tag{15}$$

$$N\left(\mathbf{W}\right) = \frac{1}{2}\sum_{g=1}^{r} \mathbf{w}_g^T \mathbf{A}_g \mathbf{w}_g + \left(\mathbf{w}_g^{k-1}\right)^T \mathbf{A}_g \mathbf{w}_g^{k-1}. \tag{16}$$

Here $\text{sign}(x)$ is the sign function, and $\mathbf{w}_g$ and $\mathbf{w}_g^{k-1}$ denote the g-th columns of $\mathbf{W}$ and $\mathbf{W}^{k-1}$ respectively. We also define $\mathbf{B}$ and $\mathbf{A}_g$ as follows:

$$\mathbf{B} = \left[\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}, \bar{\mathbf{x}}_2 - \bar{\mathbf{x}}, \cdots, \bar{\mathbf{x}}_n - \bar{\mathbf{x}}\right], \tag{17}$$

$$\mathbf{A}_g = \sum_{i=1}^{n} \sum_{\mathbf{x}_j \in \mathcal{N}_i \cup \{\mathbf{x}_i\}} \frac{\left(\mathbf{x}_j - \bar{\mathbf{x}}_i\right)\left(\mathbf{x}_j - \bar{\mathbf{x}}_i\right)^T}{\left|\left(\mathbf{w}_g^{k-1}\right)^T \left(\mathbf{x}_j - \bar{\mathbf{x}}_i\right)\right|}. \tag{18}$$

***Theorem** 2:* For any $\mathbf{W} \in \Re^{d\times r}$, we have

$$L\left(\mathbf{W}, \mathbf{W}^{k-1}\right) \leq F\left(\mathbf{W}\right), \tag{19}$$

where the equality holds if and only if $\mathbf{W} = \mathbf{W}^{k-1}$.

Proof 2: According to Lemma 1, we can derive:

$$\begin{aligned} H\left(\mathbf{W}\right) &= \sum_{i=1}^{n} \left\|\mathbf{W}^T\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)\right\|_1 = \sum_{i=1}^{n}\sum_{g=1}^{r} \left\|\mathbf{w}_g^T\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)\right\|_1 \\ &\geq \sum_{g=1}^{r}\sum_{i=1}^{n} \operatorname{sign}\left[\left(\mathbf{w}_g^{k-1}\right)^T\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)\right]\left[\left(\mathbf{w}_g^{k}\right)^T\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)\right] \\ &= \sum_{g=1}^{r} \mathbf{w}_g^T \mathbf{B} \operatorname{sign}\left(\mathbf{B}^T\mathbf{w}_g^{k-1}\right) = K\left(\mathbf{W}\right). \end{aligned} \tag{20}$$

According to Lemma 2, we can derive:

$$\begin{aligned} &\sum_{i=1}^{n}\sum_{\mathbf{x}_j\in\mathcal{N}_i\cup\{\mathbf{x}_i\}} \frac{1}{2}\frac{\boldsymbol{\xi}^T\left(\mathbf{x}_j - \bar{\mathbf{x}}_i\right)\left(\mathbf{x}_j - \bar{\mathbf{x}}_i\right)^T\boldsymbol{\xi}}{\boldsymbol{\xi}^T\left(\mathbf{x}_j - \bar{\mathbf{x}}_i\right)} \\ &\qquad + \frac{1}{2}\left\|\boldsymbol{\xi}^T\left(\mathbf{x}_j - \bar{\mathbf{x}}_i\right)\right\|_1 \\ &\leq \sum_{i=1}^{n}\sum_{\mathbf{x}_j\in\mathcal{N}_i\cup\{\mathbf{x}_i\}} \frac{1}{2}\frac{\boldsymbol{\xi}^T\left(\mathbf{x}_j - \bar{\mathbf{x}}_i\right)\left(\mathbf{x}_j - \bar{\mathbf{x}}_i\right)^T\boldsymbol{\xi}}{\boldsymbol{\eta}^T\left(\mathbf{x}_j - \bar{\mathbf{x}}_i\right)} \\ &\qquad + \frac{1}{2}\left\|\boldsymbol{\eta}^T\left(\mathbf{x}_j - \bar{\mathbf{x}}_i\right)\right\|_1, \end{aligned} \tag{21}$$

which indicates that:

$$\begin{aligned} M\left(\mathbf{W}\right) &= \sum_{i=1}^{n}\sum_{\mathbf{x}_j\in\mathcal{N}_i\cup\{\mathbf{x}_i\}} \left\|\mathbf{W}^T\left(\mathbf{x}_j - \bar{\mathbf{x}}_i\right)\right\|_1 \\ &= \sum_{g=1}^{r}\sum_{i=1}^{n}\sum_{\mathbf{x}_j\in\mathcal{N}_i\cup\{\mathbf{x}_i\}} \frac{\mathbf{w}_g^T\left(\mathbf{x}_j - \bar{\mathbf{x}}_i\right)\left(\mathbf{x}_j - \bar{\mathbf{x}}_i\right)^T\mathbf{w}_g}{2\left\|\mathbf{w}_g^T\left(\mathbf{x}_j - \bar{\mathbf{x}}_i\right)\right\|_1} \\ &\quad + \frac{1}{2}\left\|\mathbf{w}_g^T\left(\mathbf{x}_j - \bar{\mathbf{x}}_i\right)\right\|_1 \\ &\leq \sum_{g=1}^{r}\sum_{i=1}^{n}\sum_{\mathbf{x}_j\in\mathcal{N}_i\cup\{\mathbf{x}_i\}} \frac{\mathbf{w}_g^T\left(\mathbf{x}_j - \bar{\mathbf{x}}_i\right)\left(\mathbf{x}_j - \bar{\mathbf{x}}_i\right)^T\mathbf{w}_g}{2\left\|\left(\mathbf{w}_g^{k-1}\right)^T\left(\mathbf{x}_j - \bar{\mathbf{x}}_i\right)\right\|_1} \\ &\quad + \frac{1}{2}\left\|\left(\mathbf{w}_g^{k-1}\right)^T\left(\mathbf{x}_j - \bar{\mathbf{x}}_i\right)\right\|_1 \\ &= \frac{1}{2}\sum_{g=1}^{r}\mathbf{w}_g^T\mathbf{A}_g\mathbf{w}_g + \left(\mathbf{w}_g^{k-1}\right)^T\mathbf{A}_g\mathbf{w}_g^{k-1} = N\left(\mathbf{W}\right). \end{aligned} \tag{22}$$

Using the inequalities in Eq. (20) and Eq. (22), together with the definition of $F\left(\mathbf{W}\right)$ in Eq. (10) and the definition of $L\left(\mathbf{W}, \mathbf{W}^{k-1}\right)$ in Eq. (14), we can derive the following:

$$\begin{aligned} L\left(\mathbf{W}, \mathbf{W}^{k-1}\right) &= K\left(\mathbf{W}\right) - \lambda^k N\left(\mathbf{W}\right) \\ &\leq H\left(\mathbf{W}\right) - \lambda^k M\left(\mathbf{W}\right) = F\left(\mathbf{W}\right). \end{aligned} \tag{23}$$

According to Lemma 1 and Lemma 2, it is easy to verify that the equalities in Eq. (20) and Eq. (22) hold if and only if $\mathbf{W} = \mathbf{W}^{k-1}$. Thus, the equality in Eq. (23) holds if and only if $\mathbf{W} = \mathbf{W}^{k-1}$, which completes the proof of Theorem 2. ■

According to Theorem 2 and the definition of λ^k in Eq. (11), we can derive:

$$\begin{aligned} F\left(\mathbf{W}\right) &\geq L\left(\mathbf{W}, \mathbf{W}^{k-1}\right) \geq L\left(\mathbf{W}^{k-1}, \mathbf{W}^{k-1}\right) \\ &= F\left(\mathbf{W}^{k-1}\right) = H(\mathbf{W}^{k-1}) - \lambda^k M(\mathbf{W}^{k-1}) \\ &= 0, \end{aligned} \tag{24}$$

which indicates that finding the solution to satisfy Eq. (10) can be transformed into finding a solution $\mathbf{W}$ to satisfy $L\left(\mathbf{W}, \mathbf{W}^{k-1}\right) \geq 0$. This can be solved by the projected subgradient method with Armigo line search [41], for which we need compute the subgradient of $L\left(\mathbf{W}, \mathbf{W}^{k-1}\right)$ at $\mathbf{W}$:

$$\begin{aligned} \partial L(\mathbf{W}, \mathbf{W}^{k-1}) &= \mathbf{B}\operatorname{sign}\left(\mathbf{B}^T\mathbf{W}^{k-1}\right) \\ &\quad - \lambda^k\left[\mathbf{A}_1\mathbf{w}_1, \mathbf{A}_2\mathbf{w}_2, \cdots, \mathbf{A}_r\mathbf{w}_r\right], \end{aligned} \tag{25}$$

and use the following operator:

$$P\left(\mathbf{W}\right) = \mathbf{W}\left(\mathbf{W}^T\mathbf{W}\right)^{-\frac{1}{2}}, \tag{26}$$

which projects $\mathbf{W}$ onto an orthogonal cone, thereby guarantees the orthonormal constraint of $\mathbf{W}^T\mathbf{W} = \mathbf{I}$.

Putting all above together, we summarize our algorithm to solve the proposed objective in Eq. (4) in Algorithm 2, whose convergence is guaranteed by Theorem 3 and Theorem 4.

Algorithm 2: The algorithm to solve our objective.

1. Randomly initialize $\mathbf{W}^0$ that satisfies $\left(\mathbf{W}^0\right)^T\mathbf{W}^0 = \mathbf{I}$;
2. Set $k = 1$ and set the parameter $0 < \beta < 1$;
repeat
  3. Calculate λ^k by Eq. (11);
  4. Calculate $\mathbf{G}^{k-1} = \partial L\left(\mathbf{W}^{k-1}, \mathbf{W}^{k-1}\right)$ by Eq. (25);
  5. Set $m = 1$;
  repeat
    6. Calculate $\mathbf{W}^k = P\left(\mathbf{W}^{k-1} + \beta^m\mathbf{G}^{k-1}\right)$;
    7. Calculate $F\left(\mathbf{W}^k\right)$ by Eq. (10);
    8. $m = m + 1$.
  until $F\left(\mathbf{W}^k\right) > F\left(\mathbf{W}^{k-1}\right) = 0$
until *Convergence*
Output: $\mathbf{W}^k$

C. Convergence analysis of our algorithm

***Theorem** 3:* If $\mathbf{W}^k$ satisfies the inequality in Eq. (10), we have $\mathcal{J}_{\ell_1}(\mathbf{W}^k) \geq \mathcal{J}_{\ell_1}(\mathbf{W}^{k-1})$.

Proof 3: Because $\mathbf{W}^k$ satisfies the inequality in Eq. (10), we have:

$$\begin{aligned} F(\mathbf{W}^k) &= \sum_{i=1}^{n}\left\|\left(\mathbf{W}^k\right)^T\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)\right\|_1 \\ &\quad - \lambda^k\sum_{i=1}^{n}\sum_{\mathbf{x}_j\in\mathcal{N}_i\cup\{\mathbf{x}_i\}}\left\|\left(\mathbf{W}^k\right)^T\left(\mathbf{x}_j - \bar{\mathbf{x}}_i\right)\right)\right\|_1 \\ &\geq 0. \end{aligned} \tag{27}$$

By a simple mathematical derivation and using the definition of λ^k in Eq. (11), we can rewrite Eq. (27) as following:

$$\begin{aligned} \mathcal{J}_{\ell_1}(\mathbf{W}^k) &= \frac{\sum_{i=1}^{n}\left\|\left(\mathbf{W}^k\right)^T\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)\right\|_1}{\sum_{i=1}^{n}\sum_{\mathbf{x}_j\in\mathcal{N}_i\cup\{\mathbf{x}_i\}}\left\|\left(\mathbf{W}^k\right)^T\left(\mathbf{x}_j - \bar{\mathbf{x}}_i\right)\right)\right\|_1} \\ &\geq \lambda^k \\ &= \frac{\sum_{i=1}^{n}\left\|(\mathbf{W}^{k-1})^T\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)\right\|_1}{\sum_{i=1}^{n}\sum_{\mathbf{x}_j\in\mathcal{N}_i\cup\{\mathbf{x}_i\}}\left\|(\mathbf{W}^{k-1})^T\left(\mathbf{x}_j - \bar{\mathbf{x}}_i\right)\right)\right\|_1} \\ &= \mathcal{J}_{\ell_1}(\mathbf{W}^{k-1}), \end{aligned} \tag{28}$$

which completes the proof of Theorem 3. ■

Theorem 4: The objective in Eq. (4) is upper bounded.

Proof 4: First, using the Cauchy-Schwarz inequality we get the following for the numerator of our objective in Eq. (4):

$$\begin{aligned}&\sum_{i=1}^{n}\left\|\mathbf{W}^T(\mathbf{x}_i-\bar{\mathbf{x}})\right\|_1=\sum_{i=1}^{n}\sum_{j=1}^{r}\left\|\mathbf{W}_j^T(\mathbf{x}_i-\bar{\mathbf{x}})\right\|_1\\&\leq\sum_{i=1}^{n}\sum_{j=1}^{r}\left\|\mathbf{W}_j^T\right\|_2\left\|(\mathbf{x}_i-\bar{\mathbf{x}})\right\|_2=\sum_{i=1}^{n}r\left\|(\mathbf{x}_i-\bar{\mathbf{x}})\right\|_2.\end{aligned}\tag{29}$$

Given an input dataset, $\sum_{i=1}^{n} r\left\|(\mathbf{x}_i-\bar{\mathbf{x}})\right\|_2$ is a constant, which indicates that the numerator of our objective in Eq. (4) is upper bounded for a given dataset.

Second, because it can be easily verified that $\sqrt{\sum_{i=1}^{n} v_i^2}\leq\sum_{i=1}^{n}|v_i|$, *i.e.*, $\forall\mathbf{v}\in\Re^n$, $\|\mathbf{v}\|_2\leq\|\mathbf{v}\|_1$, we can derive the following for the denominator of our objective in Eq. (4):

$$\begin{aligned}&\sum_{i=1}^{n}\sum_{\mathbf{x}_j\in\mathcal{N}_i\cup\{\mathbf{x}_i\}}\left\|\mathbf{W}\left(\mathbf{x}_j-\bar{\mathbf{x}}_i\right)\right)\|_1\\&\geq\sum_{i=1}^{n}\sum_{\mathbf{x}_j\in\mathcal{N}_i\cup\{\mathbf{x}_i\}}\sqrt{\left\|\mathbf{W}\left(\mathbf{x}_j-\bar{\mathbf{x}}_i\right)\right)\|_2^2}\\&\geq\sqrt{\sum_{i=1}^{n}\sum_{\mathbf{x}_j\in\mathcal{N}_i\cup\{\mathbf{x}_i\}}\left\|\mathbf{W}\left(\mathbf{x}_j-\bar{\mathbf{x}}_i\right)\right)\|_2^2}\\&=\sqrt{\mathrm{tr}(\mathbf{W}^T\mathbf{S}_L\mathbf{W})}\geq\sqrt{\sum_{i=1}^{r}\lambda_i},\end{aligned}\tag{30}$$

where $\lambda_i\quad(i=1,\ldots,r)$, ordered by $\lambda_1\leq\cdots\leq\lambda_r$, are the eigenvalues of $\mathbf{S}_L$. The last inequality in Eq. (30) is obtained by the Ky Fan's inequality [42], which states that $\mathrm{tr}(\mathbf{W}^T\mathbf{S}_L\mathbf{W})\geq\sum_{i=1}^{r}\lambda_i$. Again, given an input dataset, $\mathbf{S}_L$ is an constant matrix, which means that $\sum_{i=1}^{r}\lambda_i$ is a constant. Thus, the denominator of our objective in Eq. (4) is lower bounded. The two bounds in Eq. (29) and Eq. (30) together indicate that our objective in Eq. (4) is upper bounded. ■

Theorem 3 indicates that our proposed Algorithm 2 monotonically increases the value of the objective function in Eq. (4) in each iteration. Theorem 4 indicates that the objective function is upper bounded, which, together with Theorem 3, indicates that Algorithm 2 converges to a local optimum.

Though motivated by previous work [36], our new algorithm to solve the proposed objective in Eq. (4) for minimizing the ratio of the summations of the ℓ_1-norm distances is more computationally efficient than that in [36]. The most computationally intensive step of the algorithm presented in [36] is to solve a system linear equations, whose complexity is $\mathcal{O}\left(n^3\right)$ if the Gaussian elimination method is used. In contrast, the most computationally intensive step of our algorithm is to perform a line search. Based upon the selection of optimization package, the complexity of our algorithm can be $\mathcal{O}\left(n\sqrt{k}\right)$ where k is the iteration number. We perform our experiments on a Dell OptiPlex 7040 desktop, with Core i7-6700 CPU processors at 3.4 GHz and 32G bytes memory. Our algorithm takes about 75 seconds to run the experiments while the algorithm in [36] takes about 231 seconds. In addition, our algorithm usually converges in no more than 30 iterations, while the algorithm in [36] usually converges in about 60 iterations.

V. Experiments

In this section, we empirically evaluate a variety of aspects of the proposed method by applying it to the ADNI cohort.

A. Description of the Experimental Dataset

Data used in the preparation of all our experiments were obtained from the ADNI (adni.loni.usc.edu). We downloaded 1.5 T MRI scans and demographic information for 821 ADNI-1 participants. Two high resolution T1-weighted MRI scans were collected for each participant using a sagittal 3D MP-RAGE sequence with an approximate TR=2400ms, minimum full TE, approximate TI=1000ms, and approximate flip angle of 8 degrees (scan parameters vary between sites, scanner platforms, and software versions). Scans were collected with a 24cm field of view and an acquisition matrix of $192\times192\times166$ (x, y, z dimensions), to yield a standard voxel size of $1.25\times1.25\times1.2$ mm. Images were then reconstructed to give a $256\times256\times166$ matrix and voxel size of approximately $1\times1\times1.2$mm. Additional scans included prescan and scout sequences as indicated by scanner manufacturer, axial proton density T2 dual contrast FSE/TSE, and sagittal B1-calibration scans as needed [43], [44], [45].

The analysis of of VBM was performed using previously described methods [46], [47], [48], as implemented in SPM5 (https://www.fil.ion.ucl.ac.uk/spm/, London, UK). The scans were converted from DICOM to NIfTI format, co-registered to a standard T1 template image, bias corrected, and segmented into GM, WM, and CSF compartments using standard SPM5 templates [43]. GM maps were then normalized to MNI atlas space as $1\times1\times1$ mm voxels and smoothed using a 10 mm FWHM Gaussian kernel. In cases where the first MP-RAGE scan could not be successfully segmented we attempted to use the second MP-RAGE. This was successful for only 1 of 8 cases.

A hippocampal regions of interest (ROI) template was created by manual tracing of the left and right hippocampi in an independent sample of 40 HC participants enrolled in study of brain aging and MCI at Dartmouth Medical School [49], [50]. These ROIs were used to extract GM density values from smoothed, unmodulated normalized and modulated normalized GM maps for the ADNI cohort. The volume of interest (VOI) including bilateral hippocampi and amygdalar nuclei, were extracted using FreeSurfer (version 4, http://surfer.nmr.mgh.harvard.edu/, Boston, MA). FreeSurfer was also used to extract cortical thickness values from the left and right entorhinal cortex, inferior, middle, and superior temporal gyri, inferior parietal gyrus, and precuneus.

We also downloaded 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 records and cognitive assessments includes BL, M6, M12, M18, M24 and M36. In our experiments, all the participants' data used in to learn enriched neuroimaging representations are required to have a baseline measurement, a baseline cognitive score, and at least two available records from M6/M12/M18/M24/M36. A total of 544 subjects are involved in our experiments, among which we have 92 subjects with AD, 205 subjects with MCI samples, and 247 HC subjects. Ten cognitive scores are included: (1) ADAS TOTAL scores from ADAS cognitive assessment; (2) FLU ANIM and (3) FLU VEG scores from Fluency cognitive assessment; (4) MMSE score from MMSE cognitive assessment; (5) RAVLT TOTAL, (6) RAVLT 30, and (7) RAVLT 30 RECOG scores from RAVLT cognitive assessment; (8) TRAIL A, (9) TRAIL B, and (10) TRAIL B-A scores from Trail making test.

B. Evaluating the Learned Biomarker Representations with Longitudinal Enrichments in Clinical Scores

Because the main goal of this study is to learn a set of fixed-length vector representations for the imaging biomarkers using longitudinal enrichments for all the subject samples in an AD dataset, we first experimentally evaluate the proposed method by applying it to the ADNI cohort, where we compare the predictive power of the learned biomarker representations with longitudinal enrichments against the BL MRI measurements using both VBM and FreeSurfer biomarkers respectively.

Experiment Settings. To validate the effectiveness of our proposed method, we compare the performance to predict cognitive outcomes using two types of the neuroimaging inputs – the learned enriched representations and the original biomarker measurements at the BL time point. We implement two versions of our new method to evaluate our hypothesis that learning a robust model by using the ℓ_1-norm distance can better address the missing record problem when using longitudinal data, *i.e.*, we learn the temporally enriched representations by using the objective in Eq. (3) and that in Eq. (4) respectively, and compare their predictive capabilities.

In our experiments, five regression methods proven to generalize well, including Linear Regression (LR), Ridge Regression (RR), Lasso, Support Vector Regression (SVR), and Convolutional Neural Network (CNN), are compared. LR is the simplest and most broadly used regression model in statistical learning and brain image analyses. RR is a regularized version of LR to account for over-fitting. Lasso regression performs both variable selection and regularization for better generalization. SVR is the regression version of the Support Vector Machine (SVM), which has been widely used to solve many real-world problems. CNN can be used for regression and has demonstrated the superior performance.

For LR, RR, Lasso and SVR, we conduct a standard 5-fold cross-validation and evaluate their performance by computing the Root Mean Square Error (RMSE) between the predicted values and ground truth values of the cognitive scores on the testing data only. Specifically, 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, such that every data sample is used as testing data by exactly one time. We iterate each five-fold experiment 10 times and randomly shuffle training and testing groups in between each iteration. The average performance for a given model with fixed hyperparameters are used for comparison. The standard deviations for each performance metric during the five-fold experiments iterated over 10 trials are reported with our prediction results. In RR and Lasso regressions, the regularization parameters are fine tuned by searching $\{10^{-10}, \ldots, 10^{-1}, 1, 10, \cdots, 10^{10}\}$. In the SVR model, the Gaussian kernel is used and the box constraint parameter is fine tuned by a search on $\{10^{-5}, \ldots, 10^{-1}, 1, 10, \cdots, 10^{5}\}$. There is a slight difference in the settings for the experiments using CNNs. For CNN regression, we randomly select 70% of the data samples as the training set, 20% of the data samples as the validation set, and we use the remaining 10% of the data samples as the testing set. The validation set used in the experiments is designed to provide an unbiased evaluation on how the CNN model fits the training dataset. We report the performance of the predictive results of the testing data. We construct a two-layer convolution architecture for predicting cognitive outcomes: (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 leveraged to reduce overfitting in the CNN models and prevent complex co-adaptations on training data. The dropout probability is set to be 0.3 and the batch size is set to be 16 in all our experiments. The hyperparameter r is fine tuned by searching $\{20, 25, \ldots, 50\}$.

Experiment Results. To evaluate the predictive power of the enriched biomarker representations learned by our new method, we use them as input to predict the 10 cognitive scores by the 5 regression models as mentioned above. As a result, for each type of input neuroimaging biomarkers, VBM and FreeSurfer, we end up with 50 prediction tasks. The prediction performance comparisons between the enriched biomarker representations and the BL ones in these prediction tasks are reported in Fig. 2 for the VBM imaging markers and in Fig. 3 for the FreeSurfer imaging makers, respectively. From the two figures we can see that the enriched biomarker representations learned by our new method are consistently better than the BL ones in all 100 prediction tasks, which we attribute to the following two reasons. Firstly, the baseline representations only characterize the brain status of the participants at one single time point, therefore they cannot benefit from the longitudinal correlations during the course when AD develops. In contrast, the enriched biomarker representations learned by our new method can integrate the baseline neuroimaging record and the temporal variations in the dynamic follow-up records. Because AD is characterized by progressive degenerations of the patients' cognitive capabilities, incorporating temporal information over time could assist in predictions. Secondly, the original baseline neuroimaging measurements reside in a high-dimensional space, which could be redundant and noisy. Thus directly using traditional regression methods could suffer from "the curse of dimensionality". Via the projection learned

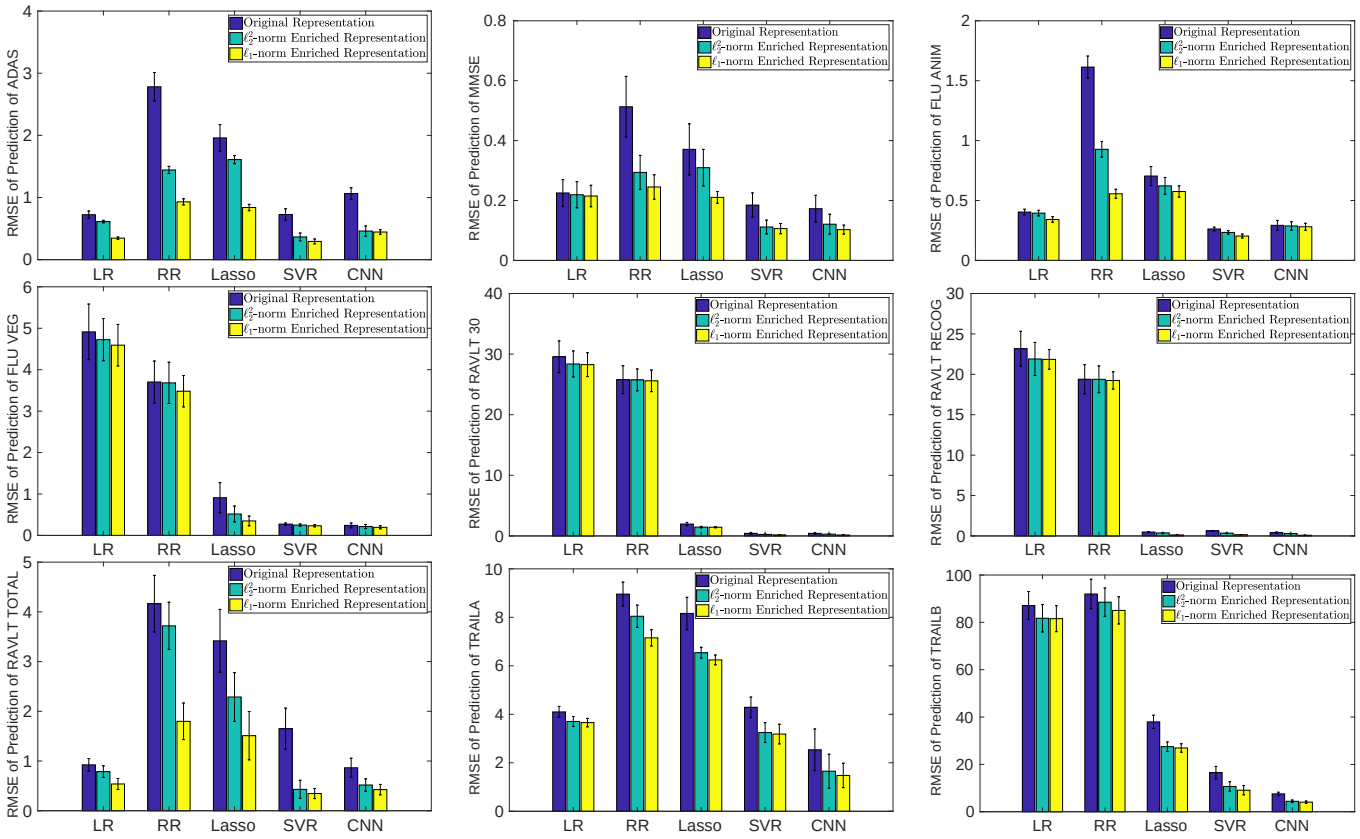

Figure 2. Comparisons of the predictive performances of the original representations at the baseline time point (blue), the enriched representations learned by the objective in Eq. (3) that uses the squared ℓ_2-norm distances (cyan), and the enriched representations learned by the objective in Eq. (4) that uses the ℓ_1-norm distances (yellow), when VBM biomarkers are used to predict the 10 different baseline cognitive outcomes using the 5 different regression models (LR, RR, Lasso, SVR, and CNN). The RMSEs (smaller is better) for predicting each cognitive outcome by each type of representations are shown for comparison, where the vertical bars show the standard deviations in the 5-fold cross-validations.

by our objectives in Eq. (3) and Eq. (4), we map the baseline cognitive measurements into a low-dimensional subspace that can mitigate the problem of high dimensionality. Thus, from Fig. 2 and Fig. 3 we can see that, compared to the original high-dimensional baseline representations, the enriched representations learned by our new method have achieved clear improvements for predicting cognitive outcomes.

Overall, by incorporating the global and local consistencies of the neuroimaging records of each participant, we learn a low-dimensional enriched biomarker representation with a consistent length, which can clearly improve the predictive performances when we use the five regression models to predict cognitive outcomes by both VBM and FreeSurfer biomarkers. This certifies the usefulness of the enriched biomarker representations learned by our new method.

Finally, as we expected, we also observe that the predictive performances of the enriched biomarker representations learned by our objective using the ℓ_1-norm distance are always better than the objective that uses the traditional squared-ℓ_2-norm distance, sometimes very significantly. For example, when we use the VBM biomarkers to predict the RAVLT TOTAL score, the enriched representations learned by the objective using the ℓ_1-norm distance improve the performance by 105% compared to their counterparts that use the squared ℓ_2-norm distance. These observations firmly confirm the correctness of our hypothesis that using a robust learning model is appropriate for representation learning due to the misaligned missing records of the participants across a studied cohort.

C. Comparing the Capability of Our New Method to Predict Cognitive Outcomes against Other Longitudinal Models

In the previous experiments, we have compared the enriched biomarker representations learned by our new method against their BL counterparts. The latter, however, are static measurements that only characterize the brain status at the baseline time point, but do not utilize the information at any follow-up time points. To further demonstrate the advantage of the our new method, we compare its predictive performance against longitudinal enrichments learned from Locality Persevering Projection (LPP) [32] where SVR and CNN are used for regression, respectively. We also compare our methods against two very recent longitudinal learning models, including (1) the Temporal Group Feature (TGF) method [12], (2) the Longitudinal Spatial Features (LSF) method [13]; and three different multi-task based longitudinal methods, including (1) the Multi-Task Exclusive Relationship (MTER) method [18], (2) Robust Multi-Task Feature Learning (RMTFL) [51], (3) Joint Multi-Modal Longitudinal Regression and Classification for Alzheimer's Disease Prediction (JMMLRC) [20]. Different from the five regression models used before, these five

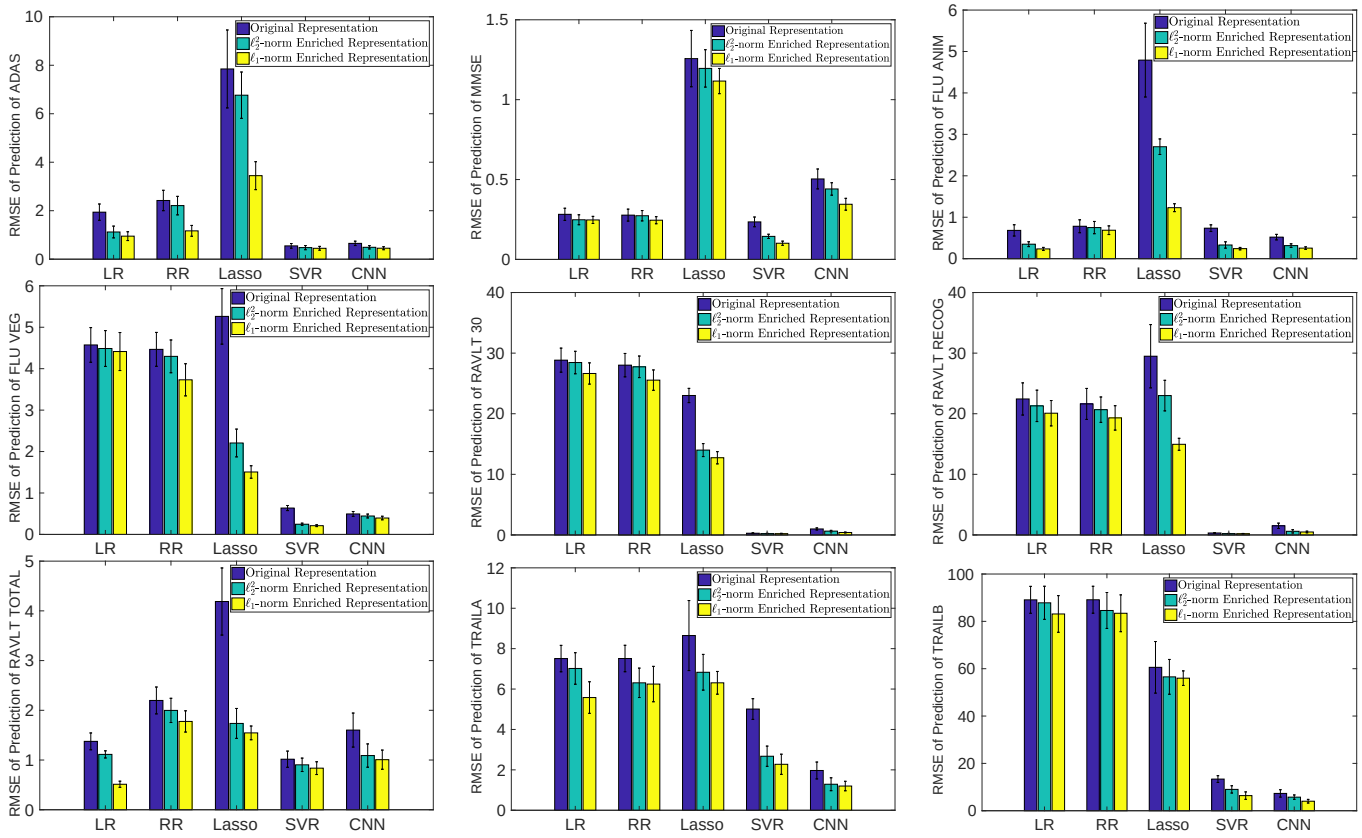

Figure 3. Comparisons of the predictive performances of the original representations at the baseline time point (blue), the enriched representations learned by the objective in Eq. (3) that uses the squared ℓ_2-norm distances (cyan), and the enriched representations learned by the objective in Eq. (4) that uses the ℓ_1-norm distances (yellow), when FreeSurfer biomarkers are used to predict the 10 different baseline cognitive outcomes using the 5 different regression models (LR, RR, Lasso, SVR, and CNN). The RMSEs (smaller is better) for predicting each cognitive outcome by each type of representations are shown for comparison, where the vertical bars show the standard deviations in the 5-fold cross-validations.

methods are designed to take advantage of longitudinal data over all the examined time points. In our experiments, after we learn the enriched biomarker representations by our new method, we use RR for the regression analyses. For the three compared methods, we fine tune their parameters following the procedures described in the respective papers. We report the comparison results in Table I and it shows that our new method achieves the best performance when we predict all clinical scores using both VBM and FreeSurfer biomarkers, which again demonstrates the effectiveness of our new method.

D. Evaluating the Learned Biomarker Representations with Longitudinal Enrichments in Behavior Assessment

Besides the prediction of cognitive declines of AD patients, we also evaluation of new method by predicting Clinical Dementia Rating (CDR) [52] and Functional Assessment Questionnaire (FAQ) [53]. The CDR and FAQ scales are highly recommended for clinical and severity assessment of dementia. The CDR is derived from the scores in each of the six categories ("box score") – Memory, Orientation, Judgment and Problem Solving, Community Affairs, Home and Hobbies and Personal Care. Memory (M) is considered the primary category and all others are secondary. With a semi-structured interview with the patient and an appropriate rates, each of the six cognitive categories is scored on a five-point scale in which none = 0, questionable = 0 5, mild = 1, moderate 2 and severe = 3. Sum of Boxes of CDR (CDR-SB) sums up the scores of all the six categories. The FAQ measures activities of daily living and is administered at baseline and at every subsequent in clinic visit. FAQ is a bounded outcome (ranging from 0 to 30), with 0 scored as "no impairment" and 30 as "severely impaired" [54].

Because both CDR and FAQ assessments are quantified by numerical numbers, we can consider the tasks of predicting them as regression tasks. To evaluate the predictive capability of our proposed methods, we use the original data representation, the enriched representations learned by the proposed objectives that use the squared ℓ_2-norm distances and the ℓ_1-norm distances as inputs to predict the two behavior assessments by five regression models, same what we did for predicting cognitive declines as in Section V-C. From the Table II, we can see that our proposed ℓ_1-norm enriched representation achieves the best performance when predicting the behavior assessment, which provides one concrete evidence to support the effectiveness of our proposed method.

E. Identifying Disease Relevant Imaging Biomarkers

Besides predicting cognitive outcomes, another important goal of our regression analyses is to identify a subset of imaging biomarkers that are highly correlated to AD devel-

Table I
PERFORMANCE COMPARISONS MEASURED BY RMSE (SMALLER IS BETTER ↓) BETWEEN OUR METHOD AND OTHER LONGITUDINAL METHODS.

Clinical Scores	Methods	RMSE (↓)		Methods	RMSE (↓)	
		VBM	FreeSurfer		VBM	FreeSurfer
ADAS	LPP [32] (SVR)	7.137 ± 0.274	8.223 ± 0.391	LPP [32] (CNN)	2.178 ± 0.682	2.891 ± 0.763
	TGF [12]	4.004 ± 0.186	3.801 ± 0.144	LSF [13]	2.255 ± 0.119	1.988 ± 0.207
	MTER [18]	1.921 ± 0.241	1.741 ± 0.107	RMTFL [51]	1.949 ± 0.177	1.915 ± 0.172
	JMMLRC [20]	7.842 ± 0.407	7.436 ± 0.483	Our Method (RR)	0.931 ± 0.049	1.169 ± 0.224
	Our Methods (SVR)	**0.294 ± 0.051**	**0.449 ± 0.079**	Our Methods (CNN)	0.445 ± 0.037	0.453 ± 0.061
MMSE	LPP [32] (SVR)	11.851 ± 1.746	10.917 ± 1.147	LPP [32] (CNN)	3.127 ± 0.107	2.815 ± 0.213
	TGF [12]	2.036 ± 0.076	1.903 ± 0.058	LSF [13]	1.581 ± 0.144	0.603 ± 0.085
	MTER [18]	1.324 ± 0.116	0.547 ± 0.153	RMTFL [51]	1.232 ± 0.199	1.315 ± 0.258
	JMMLRC [20]	9.081 ± 0.817	9.528 ± 0.790	Our Method (RR)	0.245 ± 0.041	0.246 ± 0.022
	Our Methods (SVR)	0.106 ± 0.017	**0.101 ± 0.013**	Our Methods (CNN)	**0.103 ± 0.015**	0.345 ± 0.037
FLU_ANIM	LPP [32] (SVR)	6.372 ± 0.692	5.572 ± 0.204	LPP [32] (CNN)	1.239 ± 0.083	0.852 ± 0.621
	TGF [12]	2.341 ± 0.126	2.623 ± 0.095	LSF [13]	1.483 ± 0.108	1.317 ± 0.127
	MTER [18]	0.923 ± 0.081	0.837 ± 0.065	RMTFL [51]	0.902 ± 0.061	1.046 ± 0.084
	JMMLRC [20]	7.590 ± 0.428	8.072 ± 0.493	Our Method (RR)	0.556 ± 0.037	0.687 ± 0.104
	Our Methods (SVR)	**0.204 ± 0.016**	0.242 ± 0.029	Our Methods (CNN)	0.2810 ± 0.029	0.255 ± 0.031
FLU_VEG	LPP [32] (SVR)	10.711 ± 0.523	12.687 ± 0.675	LPP [32] (CNN)	3.192 ± 0.894	2.826 ± 0.637
	TGF [12]	2.752 ± 0.103	2.853 ± 0.087	LSF [13]	1.437 ± 0.143	1.490 ± 0.159
	MTER [18]	0.895 ± 0.076	0.764 ± 0.083	RMTFL [51]	1.073 ± 0.162	1.258 ± 0.204
	JMMLRC [20]	6.934 ± 0.309	7.264 ± 0.375	Our Method (RR)	3.481 ± 0.379	3.732 ± 0.388
	Our Methods (SVR)	0.233 ± 0.026	**0.211 ± 0.022**	Our Methods (CNN)	**0.196 ± 0.037**	0.396 ± 0.042
RAVLT_TOTAL	LPP [32] (SVR)	16.873 ± 2.614	18.721 ± 1.844	LPP [32] (CNN)	2.897 ± 0.435	2.397 ± 0.341
	TGF [12]	3.846 ± 0.203	3.650 ± 0.178	LSF [13]	1.448 ± 0.132	1.431 ± 0.108
	MTER [18]	1.897 ± 0.394	1.786 ± 0.274	RMTFL [51]	1.176 ± 0.167	1.380 ± 0.078
	JMMLRC [20]	8.651 ± 0.672	8.894 ± 0.647	Our Method (RR)	1.798 ± 0.367	1.776 ± 0.214
	Our Methods (SVR)	**0.347 ± 0.101**	**0.837 ± 0.129**	Our Methods (CNN)	0.425 ± 0.104	1.007 ± 0.193
RAVLT_30	LPP [32] (SVR)	13.721 ± 1.581	11.938 ± 1.260	LPP [32] (CNN)	0.924 ± 0.045	0.703 ± 0.038
	TGF [12]	2.304 ± 0.053	2.426 ± 0.047	LSF [13]	1.562 ± 0.137	1.487 ± 0.168
	MTER [18]	0.725 ± 0.131	0.693 ± 0.174	RMTFL [51]	0.691 ± 0.051	0.608 ± 0.054
	JMMLRC [20]	16.319 ± 0.873	14.844 ± 0.469	Our Method (RR)	25.597 ± 1.786	25.544 ± 1.681
	Our Methods (SVR)	0.173 ± 0.050	**0.209 ± 0.053**	Our Methods (CNN)	**0.158 ± 0.031**	0.405 ± 0.064
RAVLT_RECOG	LPP [32] (SVR)	17.956 ± 1.713	19.062 ± 2.018	LPP [32] (CNN)	2.786 ± 0.753	1.913 ± 0.613
	TGF [12]	3.032 ± 0.092	2.896 ± 1.231	LSF [13]	1.972 ± 0.130	2.149 ± 0.107
	MTER [18]	0.951 ± 0.082	0.719 ± 0.042	RMTFL [51]	0.780 ± 0.026	0.761 ± 0.032
	JMMLRC [20]	15.117 ± 1.426	15.190 ± 1.071	Our Method (RR)	19.239 ± 1.070	19.313 ± 2.011
	Our Methods (SVR)	0.149 ± 0.012	**0.202 ± 0.041**	Our Methods (CNN)	**0.087 ± 0.017**	0.487 ± 0.134
TRAILA	LPP [32] (SVR)	26.198 ± 3.272	30.578 ± 2.803	LPP [32] (CNN)	13.376 ± 1.291	14.092 ± 0.913
	TGF [12]	14.892 ± 0.831	12.816 ± 1.253	LSF [13]	18.417 ± 1.514	18.940 ± 1.109
	MTER [18]	7.590 ± 0.496	6.032 ± 6.032	RMTFL [51]	6.629 ± 0.570	6.446 ± 0.574
	JMMLRC [20]	36.402 ± 1.895	32.117 ± 2.097	Our Method (RR)	7.153 ± 0.337	6.247 ± 0.876
	Our Methods (SVR)	3.185 ± 0.405	2.276 ± 0.499	Our Methods (CNN)	**1.478 ± 0.502**	**1.199 ± 0.234**
TRAILB	LPP [32] (SVR)	46.192 ± 4.707	50.672 ± 4.691	LPP [32] (CNN)	8.972 ± 0.913	9.187 ± 0.895
	TGF [12]	34.912 ± 3.045	37.721 ± 4.460	LSF [13]	39.027 ± 3.782	47.346 ± 4.469
	MTER [18]	28.823 ± 2.369	33.375 ± 1.431	RMTFL [51]	25.213 ± 1.296	24.016 ± 1.761
	JMMLRC [20]	41.659 ± 3.860	42.838 ± 3.871	Our Method (RR)	85.050 ± 5.724	83.420 ± 7.775
	Our Methods (SVR)	9.091 ± 1.940	6.383 ± 1.576	Our Methods (CNN)	**4.051 ± 0.480**	**4.023 ± 0.767**

opments. Therefore, we examine the biomarkers identified by the proposed methods. As can be seen in Eq. (4), we learn a projection matrix $\mathbf{W}$ for every participant in the ADNI dataset. To explore the association between the prediction targets and imaging markers, we use the regression model of $\min_{\mathbf{U}} ||\mathbf{F} - \mathbf{U}^T\mathbf{Y}||_F^2$, where $\mathbf{Y} \in \Re^{r\times n}$ contains the enriched representations for the n subjects of the studied cohort and $\mathbf{F} \in \Re^{c\times n}$ is the matrix to encode the c cognitive scores for the n subjects. Then, $\mathbf{WU}$ indicates the outcome relevant weights for each subject. The top 10 imaging biomarkers of the studied subjects are selected to determine a frequency map.

We visualize the top 10 VBM biomarkers in the frequency map in the association studies between the MMSE score and the VBM biomarkers in the top panel of Fig. 4 and the top 10 FreeSurfer biomarkers in the bottom panel of Fig. 4. We observe that the bilateral hippocampus is among the top selected biomarkers, which is in accordance with the existing clinical evidence showing that the hippocampus is mainly associated with memory, in particular long-term memory [55]. In addition, the bilateral amygdala is also among the top selected biomarkers, which agrees with the fact that the amygdala performs a primary role in the processing of memory, decision-making and emotional response and it is an important subcortical region that is severely and consistently affected by pathology in AD [56]. Furthermore, the bilateral pallidum is also listed as a top relevant biomarker, which is known to be responsible for slowly progressive dementia, cortical atrophy and local amyloidosis in the atrophic form of

Table II
PERFORMANCE COMPARISONS MEASURED BY RMSE (SMALLER IS BETTER ↓) BETWEEN ORIGINAL REPRESENTATION, SQUARED ℓ_2-NORM ENRICHED REPRESENTATION AND ℓ_1-NORM ENRICHED REPRESENTATION. WE COMPARE FIVE WE COMPARE FIVE DIFFERENT GENERAL REGRESSION METHODS FOR BEHAVIOR ASSESSMENT– LR, RR, LASSO SVR AND CNN.

Methods	Inputs	RMSE (↓)	
		CDR-SB	FAQ
LR	Original Representation	0.435 ± 0.059	1.209 ± 0.111
	ℓ_2^2-norm enrichments	0.470 ± 0.041	**1.676 ± 0.101**
	ℓ_1-norm enrichments	**0.419 ± 0.038**	1.814 ± 0.085
RR	Original Representation	0.603 ± 0.085	1.016 ± 0.129
	ℓ_2^2-norm enrichments	0.697 ± 0.061	0.988 ± 0.148
	ℓ_1-norm enrichments	**0.577 ± 0.041**	**0.819 ± 0.082**
Lasso	Original Representation	1.221 ± 0.167	2.100 ± 0.287
	ℓ_2^2-norm enrichments	0.971 ± 0.086	**1.245 ± 0.247**
	ℓ_1-norm enrichments	**0.889 ± 0.097**	1.286 ± 0.263
SVR	Original Representation	0.524 ± 0.068	0.629 ± 0.071
	ℓ_2^2-norm enrichments	0.538 ± 0.068	0.588 ± 0.073
	ℓ_1-norm enrichments	**0.521 ± 0.067**	**0.576 ± 0.060**
CNN	Original Representation	0.169 ± 0.021	0.508 ± 0.058
	ℓ_2^2-norm enrichments	0.138 ± 0.028	0.491 ± 0.051
	ℓ_1-norm enrichments	**0.136 ± 0.015**	**0.485 ± 0.052**

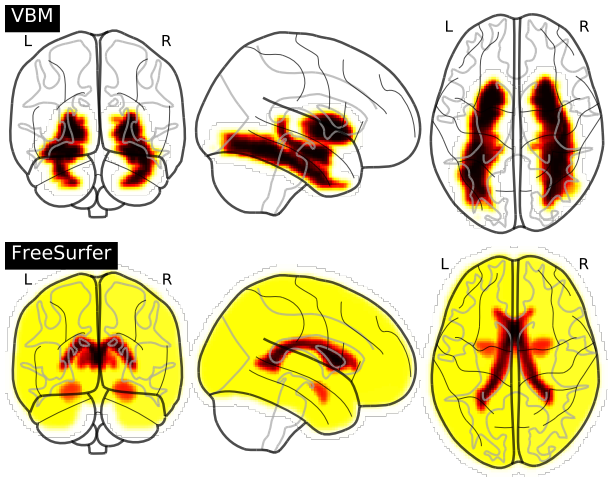

Figure 4. **Top panel**: Top 10 selected VBM biomarker mapped onto the brain: LAmygdala, RAmygdala [56], LFusiform, RFusiform [58], LHippocampus, RHippocampus [55], LPallidum, RPallidum [57], LPutamen, RPutamen [58]. **Bottom panel**: Top 10 selected FreeSurfer biomarker mapped onto the brain: LCerebWM, RCerebWM [59], LCerebCtx, RCerebCtx [60], LLatVent, RLatVent [61], LInfLatVent, RInfLatVent [61], LCerebellCtx, RCerebellCtx [60].

chronic bacterial infections [57].

In summary, the identified imaging biomarkers are highly suggestive and strongly agree with existing medical research findings with regards to AD. These findings concretely support the correctness of the discovered associations between cognitive developments and progressive variations of imaging biomarkers from the clinical perspective.

VI. CONCLUSION

In this paper, we proposed a novel formulation to learn an enriched representation for neuroimaging biomarkers using the longitudinal data. Our enriched biomarker representation is learned by solving a new objective that aims to maintain both global and local consistencies 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 projection, and the local consistency is designed to preserve the pairwise relationship of neuroimaging measurements at different time points. The objective simultaneously maximizes and minimizes the summations of a number of ℓ_1-norm distances, which is non-smooth thereby difficult to solve in general. Thus, we developed an efficient and non-greedy iterative solution algorithm with theoretically proved convergence. We conducted experiments on two types of biomarkers, VBM and FreeSurfer. Via the enriched neuroimaging representations, we can achieve a clear performance gain in predicting ten different cognitive outcomes using five standard regression models and three recent longitudinal prediction models. Moreover, the key imaging biomarkers identified for both VBM and FreeSurfer measurements nicely agree with the existing findings in clinical researches, which warrants the correctness of the enriched neuroimaging representations learned by our new method.

ACKNOWLEDGEMENT

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.; 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] A. Association *et al.*, "2018 alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 14, no. 3, pp. 367–429, 2018.

[2] G. L. Wenk *et al.*, "Neuropathologic changes in alzheimer's disease," *Journal of Clinical Psychiatry*, vol. 64, pp. 7–10, 2003.

[3] H. Wang, F. Nie, H. Huang, S. Risacher, C. Ding, A. J. Saykin, and L. Shen, "Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance," in *2011 IEEE International Conference on Computer Vision (CVPR 2011)*. IEEE, 2011, pp. 557–562.

[4] L. Shen, S. Kim, S. L. Risacher, K. Nho, S. Swaminathan, J. D. West, T. Foroud, N. Pankratz, J. H. Moore, C. 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*, vol. 53, no. 3, pp. 1051–1063, 2010.

[5] H. Wang, F. Nie, H. Huang, J. Yan, S. Kim, S. Risacher, A. Saykin, and L. Shen, "High-order multi-task feature learning to identify longitudinal phenotypic markers for alzheimer's disease progression prediction," in *Advances in neural information processing systems*, 2012, pp. 1277–1285.

[6] M. W. Weiner, P. S. Aisen, C. R. Jack Jr, W. J. Jagust, J. Q. Trojanowski, L. Shaw, A. J. Saykin, J. C. Morris, N. Cairns, L. A. Beckett *et al.*, "The alzheimer's disease neuroimaging initiative: progress report and future plans," *Alzheimer's & Dementia*, vol. 6, no. 3, pp. 202–211, 2010.

[7] X. Wang, D. Shen, and H. Huang, "Prediction of memory impairment with mri data: a longitudinal study of alzheimer's disease," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 273–281.

[8] X. Wang, J. Yan, X. Yao, S. Kim, K. Nho, S. L. Risacher, A. J. Saykin, L. Shen, H. Huang *et al.*, "Longitudinal genotype-phenotype association study via temporal structure auto-learning predictive model," in *International Conference on Research in Computational Molecular Biology*. Springer, 2017, pp. 287–302.

[9] G. Yu, Q. Li, D. Shen, and Y. Liu, "Optimal sparse linear prediction for block-missing multi-modality data without imputation," *Journal of the American Statistical Association*, pp. 1–14, 2019.

[10] R. Y. Lo and W. J. Jagust, "Predicting missing biomarker data in a longitudinal study of alzheimer disease," *Neurology*, vol. 78, no. 18, pp. 1376–1382, 2012.

[11] L. Lu, H. Wang, S. Elbeleidy, and F. Nie, "Predicting cognitive declines using longitudinally enriched representations for imaging biomarkers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020 (CVPR 2020)*, 2020, pp. 4827–4836.

[12] D. Zhang, D. Shen, A. D. N. Initiative *et al.*, "Predicting future clinical changes of mci patients using longitudinal and multimodal biomarkers," *PloS one*, vol. 7, no. 3, 2012.

[13] J. Zhang, M. Liu, L. An, Y. Gao, and D. Shen, "Alzheimer's disease diagnosis using landmark-based features from longitudinal structural mr images," *IEEE journal of biomedical and health informatics*, vol. 21, no. 6, pp. 1607–1616, 2017.

[14] L. Zhang, Y. Zhao, Z. Zhu, D. Shen, and S. Ji, "Multi-view missing data completion," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 7, pp. 1296–1309, 2018.

[15] H. Wang, F. Nie, H. Huang, J. Yan, S. Kim, K. Nho, S. L. Risacher, A. J. Saykin, L. Shen, and A. D. N. Initiative, "From phenotype to genotype: an association study of longitudinal phenotypic markers to alzheimer's disease relevant snps," *Bioinformatics*, vol. 28, no. 18, pp. i619–i625, 2012.

[16] B. Lei, F. Jiang, S. Chen, D. Ni, and T. Wang, "Longitudinal analysis for disease progression via simultaneous multi-relational temporal-fused learning," *Frontiers in aging neuroscience*, vol. 9, p. 6, 2017.

[17] X. Hao, C. Li, J. Yan, X. Yao, S. L. Risacher, A. J. Saykin, L. Shen, D. Zhang, and A. D. N. Initiative, "Identification of associations between genotypes and longitudinal phenotypes via temporally-constrained group sparse canonical correlation analysis," *Bioinformatics*, vol. 33, no. 14, pp. i341–i349, 2017.

[18] M. Wang, D. Zhang, D. Shen, and M. Liu, "Multi-task exclusive relationship learning for alzheimer's disease progression prediction with longitudinal data," *Medical image analysis*, vol. 53, pp. 111–122, 2019.

[19] L. Brand, H. Wang, H. Huang, S. Risacher, A. Saykin, L. 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*. Springer, 2018, pp. 555–562.

[20] L. Brand, K. Nichols, H. Wang, L. Shen, and H. Huang, "Joint multi-modal longitudinal regression and classification for alzheimer's disease prediction," *IEEE Transactions on Medical Imaging*, vol. 39, no. 6, pp. 1845–1855, 2019.

[21] J. Zhou, L. Yuan, J. Liu, and J. Ye, "A multi-task learning formulation for predicting disease progression," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011, pp. 814–822.

[22] L. Yuan, Y. Wang, P. M. Thompson, V. A. Narayan, and J. Ye, "Multi-source learning for joint analysis of incomplete multi-modality neuroimaging data," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 1149–1157.

[23] Y. Li, T. Yang, J. Zhou, and J. Ye, "Multi-task learning based survival analysis for predicting alzheimer's disease progression with multi-source block-wise missing data," in *Proceedings of the 2018 SIAM International Conference on Data Mining*. SIAM, 2018, pp. 288–296.

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

[25] M. Liu, Y. Gao, P.-T. Yap, and D. Shen, "Multi-hypergraph learning for incomplete multimodality data," *IEEE journal of biomedical and health informatics*, vol. 22, no. 4, pp. 1197–1208, 2017.

[26] M. Liu, J. Zhang, P.-T. Yap, and D. Shen, "View-aligned hypergraph learning for alzheimer's disease diagnosis with incomplete multi-modality data," *Medical image analysis*, vol. 36, pp. 123–134, 2017.

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

[28] R. Li, W. Zhang, H.-I. Suk, L. Wang, J. Li, D. Shen, and S. Ji, "Deep learning based imaging data completion for improved brain disease diagnosis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2014, pp. 305–312.

[29] Q. Suo, W. Zhong, F. Ma, Y. Yuan, J. Gao, and A. Zhang, "Metric learning on healthcare data with incomplete modalities," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. AAAI Press, 2019, pp. 3534–3540.

[30] M. M. Ghazi, M. Nielsen, A. Pai, M. J. Cardoso, M. Modat, S. Ourselin, L. Sørensen, A. D. N. Initiative *et al.*, "Training recurrent neural networks robust to incomplete data: Application to alzheimer's disease progression modeling," *Medical image analysis*, vol. 53, pp. 39–46, 2019.

[31] I. Jolliffe, "Principal component analysis," in *International encyclopedia of statistical science*. Springer, 2011, pp. 1094–1096.

[32] X. He and P. Niyogi, "Locality preserving projections," in *Advances in neural information processing systems*, 2004, pp. 153–160.

[33] S. Gauthier, B. Reisberg, M. Zaudig, R. C. Petersen, K. Ritchie, K. Broich, S. Belleville, H. Brodaty, D. Bennett, H. Chertkow *et al.*, "Mild cognitive impairment," *The lancet*, vol. 367, no. 9518, pp. 1262–1270, 2006.

[34] H. Wang, F. Nie, and H. Huang, "Globally and locally consistent unsupervised projection," in *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[35] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Advances in neural information processing systems*, 2002, pp. 585–591.

[36] H. Wang, F. Nie, and H. Huang, "Robust distance metric learning via simultaneous ℓ_1-norm minimization and maximization," in *International conference on machine learning*, 2014, pp. 1836–1844.

[37] F. Nie, J. Yuan, and H. Huang, "Optimal mean robust principal component analysis," in *International conference on machine learning*, 2014, pp. 1062–1070.

[38] K. Liu, L. Brand, H. Wang, and F. Nie, "Learning robust distance metric with side information via ratio minimization of orthogonally constrained ℓ_{21}-norm distances," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*, 2019.

[39] Y. Liu, Q. Gao, S. Miao, X. Gao, F. Nie, and Y. Li, "A non-greedy algorithm for l1-norm lda," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 684–695, 2017.

[40] R. Jenatton, G. Obozinski, and F. Bach, "Structured sparse principal component analysis," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010, pp. 366–373.

[41] W. Sun and Y.-X. Yuan, *Optimization theory and methods: nonlinear programming*. Springer Science & Business Media, 2006, vol. 1.

[42] K. Fan, "On a theorem of weyl concerning eigenvalues of linear transformations ii," *Proceedings of the National Academy of Sciences*, vol. 36, no. 1, pp. 31–35, 1950.

[43] S. L. Risacher, A. J. Saykin, J. D. Wes, L. Shen, H. A. Firpi, and B. C. McDonald, "Baseline mri predictors of conversion from mci to probable ad in the adni cohort," *Current Alzheimer Research*, vol. 6, no. 4, pp. 347–361, 2009.

[44] C. R. Jack Jr, M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P. J. Britson, J. L. Whitwell, C. Ward *et al.*, "The alzheimer's disease neuroimaging initiative (adni): Mri methods," *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 27, no. 4, pp. 685–691, 2008.

[45] S. L. Risacher, L. Shen, J. D. West, S. Kim, B. C. McDonald, L. A. Beckett, D. J. Harvey, C. R. Jack Jr, M. W. Weiner, A. J. Saykin *et al.*, "Longitudinal mri atrophy biomarkers: relationship to conversion in the adni cohort," *Neurobiology of aging*, vol. 31, no. 8, pp. 1401–1418, 2010.

[46] J. Ashburner and K. J. Friston, "Voxel-based morphometry—the methods," *Neuroimage*, vol. 11, no. 6, pp. 805–821, 2000.

[47] C. D. Good, I. S. Johnsrude, J. Ashburner, R. N. Henson, K. J. Friston, and R. S. Frackowiak, "A voxel-based morphometric study of ageing in 465 normal adult human brains," *Neuroimage*, vol. 14, no. 1, pp. 21–36, 2001.

[48] A. Mechelli, C. J. Price, K. J. Friston, and J. Ashburner, "Voxel-based morphometry of the human brain: methods and applications," *Current Medical Imaging*, vol. 1, no. 2, pp. 105–113, 2005.

[49] A. Saykin, H. Wishart, L. Rabin, R. Santulli, L. Flashman, J. West, T. McHugh, and A. Mamourian, "Older adults with cognitive complaints show brain atrophy similar to that of amnestic mci," *Neurology*, vol. 67, no. 5, pp. 834–842, 2006.

[50] T. L. McHugh, A. J. Saykin, H. A. Wishart, L. A. Flashman, H. B. Cleavinger, L. A. Rabin, A. C. Mamourian, and L. Shen, "Hippocampal volume and shape analysis in an older adult population," *The Clinical Neuropsychologist*, vol. 21, no. 1, pp. 130–145, 2007.

[51] P. Gong, J. Ye, and C. Zhang, "Robust multi-task feature learning," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012, pp. 895–903.

[52] J. C. Morris, "The clinical dementia rating (cdr): Current version and," *Young*, vol. 41, pp. 1588–1592, 1991.

[53] R. I. Pfeffer, T. T. Kurosaki, C. Harrah Jr, J. M. Chance, and S. Filos, "Measurement of functional activities in older adults in the community," *Journal of gerontology*, vol. 37, no. 3, pp. 323–329, 1982.

[54] K. Ito, M. Hutmacher, and B. Corrigan, "Modeling of functional assessment questionnaire (faq) as continuous bounded data from the adni database," *Journal of pharmacokinetics and pharmacodynamics*, vol. 39, no. 6, pp. 601–618, 2012.

[55] Y. Mu and F. H. Gage, "Adult hippocampal neurogenesis and its role in alzheimer's disease," *Mol. neurodegeneration*, vol. 6, no. 1, p. 85, 2011.

[56] S. P. Poulin, R. Dautoff, J. C. Morris, L. F. Barrett, B. C. Dickerson, A. D. N. Initiative *et al.*, "Amygdala atrophy is prominent in early alzheimer's disease and relates to symptom severity," *Psychiatry Research: Neuroimaging*, vol. 194, no. 1, pp. 7–13, 2011.

[57] J. Miklossy, "Alzheimer's disease-a neurospirochetosis. analysis of the evidence following koch's and hill's criteria," *Journal of neuroinflammation*, vol. 8, no. 1, p. 90, 2011.

[58] L. De Jong, K. Van der Hiele, I. Veer, J. Houwing, R. Westendorp, E. Bollen, P. De Bruin, H. Middelkoop, M. Van Buchem, and J. Van Der Grond, "Strongly reduced volumes of putamen and thalamus in alzheimer's disease: an mri study," *Brain*, vol. 131, no. 12, pp. 3277–3285, 2008.

[59] J. Acosta-Cabronero, G. B. Williams, G. Pengas, and P. J. Nestor, "Absolute diffusivities define the landscape of white matter degeneration in alzheimer's disease," *Brain*, vol. 133, no. 2, pp. 529–539, 2010.

[60] A. Bakkour, J. C. Morris, D. A. Wolk, and B. C. Dickerson, "The effects of aging and alzheimer's disease on cerebral cortical anatomy: specificity and differential relationships with cognition," *Neuroimage*, vol. 76, pp. 332–344, 2013.

[61] C. DeCarli, J. V. Haxby, J. Gillette, D. Teichberg, S. Rapoport, and M. Schapiro, "Longitudinal changes in lateral ventricular volume in datients with dementia of the alzheimer type," *Neurology*, vol. 42, no. 10, pp. 2029–2029, 1992.