Multi-Resemblance Multi-Target Low-Rank Coding for Prediction of Cognitive Decline with Longitudinal Brain Images

Jie Zhang, Jianfeng Wu, Qingyang Li, Richard J. Caselli, Paul M. Thompson, Jieping Ye *Fellow, IEEE*, Yalin Wang *Senior Member, IEEE*, and The Alzheimer's Disease Neuroimaging Initiative

***Abstract*—An effective presymptomatic diagnosis and treatment of Alzheimer's disease (AD) would have enormous public health benefits. Sparse coding (SC) has shown strong potential for longitudinal brain image analysis in preclinical AD research. However, the traditional SC computation is time-consuming and does not explore the feature correlations that are consistent over the time. In addition, longitudinal brain image cohorts usually contain incomplete image data and clinical labels. To address these challenges, we propose a novel two-stage Multi-Resemblance Multi-Target Low-Rank Coding (MMLC) method, which encourages that sparse codes of neighboring longitudinal time points are resemblant to each other, favors sparse code low-rankness to reduce the computational cost and is resilient to both source and target data incompleteness. In stage one, we propose an online multi-resemblant low-rank SC method to utilize the common and task-specific dictionaries in different time points to immune to incomplete source data and capture the longitudinal correlation. In stage two, supported by a rigorous theoretical analysis, we develop a multi-target learning method to address the missing clinical label issue. To solve such a multi-task low-rank sparse optimization problem, we propose multi-task stochastic coordinate coding with a sequence of closed-form update steps which reduces the computational costs guaranteed by a theoretical convergence proof. We apply MMLC on a publicly available neuroimaging cohort to predict two clinical measures and compare it with six other methods. Our experimental results show our proposed method achieves superior results on both computational efficiency and predictive accuracy and has great potential to assist the AD prevention.**

***Index Terms*—Multi-task, Longitudinal incomplete data, Sparse coding, Low-rank, Multi-resemblance**

I. Introduction

ALZHEIMER'S disease (AD) [1] is known as the most common type of dementia. It is a slow progressive neurodegenerative disorder leading to a loss of memory and reduction of cognitive function. Many clinical/cognitive measures such as Mini Mental State Examination (MMSE) and Alzheimer's Disease Assessment Scale cognitive subscale (ADAS-Cog) have been designed to evaluate a subject's cognitive decline. It is crucial to predict AD-related cognitive decline in its presymptomatic stage so an early intervention or prevention becomes possible.

In AD research, cognitive concerns correlate with structural magnetic resonance imaging (sMRI)-based measures of atrophy in several structural measures, including whole-brain, entorhinal cortex, hippocampus and temporal lobe volumes. [2] These findings support their potential usage as predictors of disease progression. Among various sMRI-based measures, hippocampal morphometry was one of the most popular measures for assessing disease burden, progression and effects of treatments [3], [4], [2], [5], [6]. Therefore, surface-based hippocampal morphometry has been studied intensively for cognitive decline research, e.g., [7], [8], [9], [10], [11], [12], including our work, e.g., [13], [14], [15], [16], [17]. However, a notoriously challenging problem in neuroimaging arises from the fact that the imaging feature dimensionality is intrinsically high while only a small number of samples are available. Recent work shows that sparse coding (SC) [18], [19], [20], [21] allows us to represent the primary image features as a small set of sparse coefficients and boosts their prediction power. However, the optimization of such problems is extremely time-consuming and the local features with similar descriptors lead to inconsistent sparse codes which may downgrade the statistical power on AD prediction. In addition, modeling sequential longitudinal data by an unsupervised learning approach such as SC is even more challenging because it is hard to extract correlation patterns from different time points.

Many multi-task researches aim to excavate the correlations among data from different modalities or time points. Wang *et al.* [22] propose a multi-task sparse regression and feature selection method to jointly analyze the clinical and neuroimaging data in prediction of the memory performance [23]. Zhang and Shen [24] exploit a $\ell_{2,1}$-norm based group sparse regression method to select features that can be used to jointly predict two clinical statuses and represent the different clinical status. A multi-task sparse learning framework is proposed to integrate multiple incomplete data sources in [25], e.g. there are blockwise sMRI images missing in some time points. Our prior work [20] proposes a novel unsupervised multi-task

Data used in preparation of this article were obtained from the Alzheimers Disease Neuroimaging Initiative (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: http://adni.loni.usc.edu/wpcontent/uploads/how to apply/ADNI Acknowledgement List.pdf.

J. Zhang, J. Wu, Q. Li and Y. Wang are with School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, AZ, USA (e-mail: {jzhan313, jianfen6, qingyan2, ylwang}@asu.edu). R. J. Caselli is with Department of Neurology, Mayo Clinic Arizona, Scottsdale, AZ, USA (e-mail: caselli.richard@mayo.edu). P. M. Thompson is with the Imaging Genetics Center, Institute for Neuroimaging and Informatics, Univ. of Southern California, Los Angeles, CA, USA (e-mail: pthomp@asu.edu). J. Ye is with Department of Electrical Engineering and Computer Science, Univ. of Michigan, Ann Arbor, MI, USA (e-mail: jpye@umich.edu).

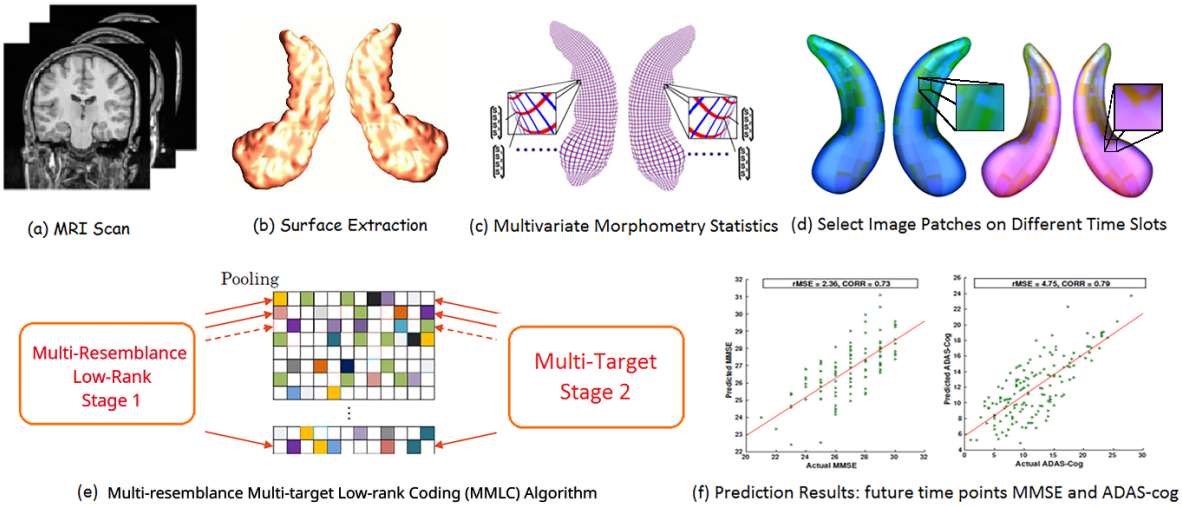

Fig. 1. The pipeline of our Multi-Resemblance Multi-Target Low-Rank Coding (MMLC) framework.

SC method that learns the different tasks simultaneously and utilizes shared and task-specific dictionaries to encode both consistent and individual imaging features for longitudinal image data analysis.

Although the multi-task SC may model sequential longitudinal data, the conventional SC method remains a computational challenge. We therefore consider the low-rankness in the sparse codes computation that favors both feature sparsity and learning efficiency. There are at least **two** advantages of the low-rank constraint on the sparse codes at each time point. Firstly, low-rankness technique was originally proposed to reduce noise and improve the signal-to-noise ratio (SNR) [26], [27]. Adding the low-rank constraint on the learned sparse codes at each time point (Eq. 2), we aim to exploit the correlations between the sparse codes. Similar to our recent work [17], it will reduce the noises in surface-based hippocampal morphometry features and therefore improve the statistical power. Secondly, the low-rankness will significantly improve the computational efficiency [28], [29], [30]. Meanwhile, our prior work [20] simply concatenates the longitudinal data while neglecting the intrinsic resemblance of the longitudinal data. It ignores the fact that the neighborhood features not only have resemblant codebooks but also have resemblant representations. Therefore, there is a huge sacrifice of valuable neighborhood time points information from the longitudinal data. To remedy this problem, here we exploit the resemblance among features lying in the neighboring time points and seek an accurate joint representation of these local features. We design a resemblance penalty term which may make the coefficients of multiple neighboring time points resemblant, ensuring higher correlations between features of nearby time points than those of distant time points.

The unsupervised multi-task learning overcomes the incomplete source data problem to obtain sparse features, but the missing clinical label problem is also ubiquitous. It results in multi-task target values after sparse features are extracted. A forthright method is to perform linear regression at each task and determine weighted matrix separately. However, such methods treat all tasks independently, ignore the useful information reserved in the changes among different tasks and cause strong bias to predict multiple target outputs. Another simple strategy is to remove all patients with missing target values. It, however, significantly reduces the number of samples. Zhou *et al.* [31] consider multi-task with missing target values in the training process, but the algorithm did not incorporate multiple-source data. For a complete solution, we therefore consider both multiple task-incomplete data and multiple outputs with missing target values in this work for exploring the disease prediction problem.

In this paper, we propose a novel two-stage framework, termed Multi-Resemblance Multi-Target Low-rank Coding (MMLC) algorithm. In stage one, we utilize shared and task-specific dictionaries to encode both consistent and changing imaging features along longitudinal time points and mine the correlations among a small number of features to obtain more consistent sparse codes than learning each time point separately. Meanwhile, we encourage using only a few sparse codebook representations to represent neighboring resemblant features to improve the smoothness of prediction over the longitudinal neighboring time points and maintain a low computational cost. In stage two, we deal with missing clinical labels on the target side, thus, we consider both input and target sides' incomplete data in the longitudinal learning process. MMLC is computed by solving an online low-rank dictionary learning optimization problem, which comprises a sequence of closed-form update steps. They are achieved by the Inexact Augmented Lagrange Multiplier (IALM) that guarantees a fast convergence. Our extensive experimental results on the Alzheimer's Disease Neuroimaging Initiative (ADNI) I cohort [1] show the proposed MMLC achieves significantly faster running speed and lower estimation errors, as well as reasonable smooth prediction scores when compared with six other algorithms, which demonstrates great potential benefits for the medical imaging research community.

Our prior work [20] establishes the multi-source multi-target dictionary learning framework. The current extended journal manuscript has four major expansions over its conference version, including 1) adding low-rank technique to model feature similarity and reduce the dictionary learning computational cost, 2) enforcing sparse codes of neighboring time point longitudinal features to be resemblant to each other, 3) providing a detailed sequence of closed-form updating steps and theoretical guarantee of fast convergence, and 4) expanding the experiments to provide additional insights into the benefit of our new method.

II. Methods

The pipeline of MMLC is illustrated in Figure 1. We will detail each step in this section. The pipeline source code is publicly available at http://gsl.lab.asu.edu/software/MMLC.

A. Problem definition and Preliminaries

Given subjects from T time points: $\{\mathbf{X}^1, \cdots, \mathbf{X}^T\}$, our goal is to learn a set of sparse codes $\{\mathbf{S}^1, \cdots, \mathbf{S}^T\}$ for each time point. The sparse code $\mathbf{S}^t \in \mathbb{R}^{m^t \times n^t}$ is a sparse representation of the original input $\mathbf{X}^t \in \mathbb{R}^{p \times n^t}$ and $t \in \{1, ..., T\}$, where p is the feature dimension of each sample of $\mathbf{x}_i^t$, $i = 1, ..., n^t$ and n^t is the number of samples for $\mathbf{X}^t$ and m^t is the dimension of each sparse code in $\mathbf{S}^t$.

When employing the conventional single-task sparse coding (SC) to learn the sparse codes $\mathbf{S}^t$ by $\mathbf{X}^t$ individually, we obtain a set of dictionary $\{\mathbf{D}^1, \cdots, \mathbf{D}^T\}$ without correlation between each learnt dictionary. The objective function of single-task SC for time point t will be

$$\min_{\mathbf{D}^t, \mathbf{S}^t} \frac{1}{2}||\mathbf{X}^t - \mathbf{D}^t\mathbf{S}^t||_F^2 + \lambda_1||\mathbf{S}^t||_{1,1}, \ s.t. \mathbf{D}^t \in \Psi^t, \quad (1)$$

where $\Psi^t = \{\mathbf{D}^t \in \mathbb{R}^{p \times m^t} : \forall j \in 1, ..., m^t, ||\mathbf{D}_j^t||_2 \leq 1\}$ and λ_1 is an non-negative parameter. Ψ^t is to prevent an arbitrary scaling of the sparse code, each column of $\mathbf{D}^t$ is restricted to be in a unit ball, i.e., $||\mathbf{D}_j^t||_2 \leq 1$. The details of SC can be summarized into Algorithm 1.

Algorithm 1: Single-Task Sparse Coding (STSC)

Input : $\mathbf{X^t}, t = 1, \cdots, T$.
Output: $\mathbf{D}^t$ and $\mathbf{S}^t$, $t = 1, \cdots, T$.

```
begin
    for k = 1 → κ do
        for t = 1 → T do
            Get an input matrix X^t;
            Update S^t by cyclic coordinate descent (CCD) [32];
            Update D^t by stochastic gradient descent (SGD) [33];
            Normalize each column of dictionary D^t .
```

B. MMLC Stage-I: Multi-Resemblance Low-Rank SC Stage

However, single-task SC (Eq. (1)) only uses one dictionary $\mathbf{D}$ which is not sufficient to model the variations among subjects from different time points. To address this problem, we integrate the idea of multi-task learning [34] into the SC method. Different from previous works, we propose to learn the intrinsic low-dimensional space of the original data by simultaneously conducting the dictionary learning and sparse feature learning processes. The objective function of our proposed multi-task low-rank SC framework is as follows:

$$\min_{\mathbf{D}^t \in \Psi^t, \mathbf{S}^t} \sum_{t=1}^{T} (\frac{1}{2}||\mathbf{X}^t - \mathbf{D}^t\mathbf{S}^t||_F^2 + \lambda_1||\mathbf{S}^t||_{1,1}), \ s.t. \ rank(\mathbf{S}^t) \leq l^t, \quad (2)$$

where the rank l^t-estimate of $\mathbf{S}^t$ denotes as $rank(\mathbf{S}^t) \leq l^t$.

However, Eq. (2) dose not consider the correlation between the samples among the multiple time points. Therefore, we proposed to use common and task-specific dictionary structure to learn dictionary atoms across multiple time points to capture the correlations. For each input matrix $\mathbf{X}^t$, we learn the dictionary atoms $\mathbf{D}^t$ which are composed of two parts: $\mathbf{D}^t = [\hat{\mathbf{D}}^t, \bar{\mathbf{D}}^t]$ where $\hat{\mathbf{D}}^t \in \mathbb{R}^{p \times \hat{m}^t}$, $\bar{\mathbf{D}}^t \in \mathbb{R}^{p \times \bar{m}^t}$ and $\hat{m} + \bar{m}^t = m^t$. $\hat{\mathbf{D}}$ is the common dictionary atoms among different tasks and $\hat{\mathbf{D}} = \hat{\mathbf{D}}^1 = \cdots = \hat{\mathbf{D}}^T$ while $\bar{\mathbf{D}}^t$ is different from each other and only learned from the corresponding task input matrix $\mathbf{X}^t$. The objective function can be reformulated as follows:

$$\min_{\mathbf{D}^t \in \Psi^t, \mathbf{S}^t} \sum_{t=1}^{T} (\frac{1}{2}||\mathbf{X}^t - [\hat{\mathbf{D}}, \bar{\mathbf{D}}^t]\mathbf{S}^t||_F^2 + \lambda_1||\mathbf{S}^t||_{1,1} + \lambda_2||\mathbf{S}^t||_*). \quad (3)$$

where λ_1 and λ_2 quantify the tradeoff between sparsity and low-rankness in the feature learning process. $\lambda_2 = 0$ is the special case of Eq. (3), the problem (3) will become sparse coding problem. Specifically, the objective function Eq. (2) is a non-convex problem due to the non-convexity of the $rank(\mathbf{S})$. We use the convex relaxation technique [35] in Eq. (3), the trace norm (nuclear norm) has been known as the convex envelop of the function of the rank $||\mathbf{S}||_* \leq rank(\mathbf{S}), \forall \mathbf{S} \in \mathbb{C} = \{\mathbf{S}|||\mathbf{S}||_2 \leq 1\}$.

The longitudinal data of the time points close to the baseline MR images has higher resemblance than those of time points distant to the baseline MR images (e.g., 3-month and 6-month MR images are more resemblant to baseline images than those of 12-month MR images). We further use a Gaussian similarity kernel to emphasize such inherent resemblance knowledge between two different time points:

$$w_{p,q} = \exp(\frac{-||p - q||}{2\sigma^2}), \quad (4)$$

where σ is empirically set as 1, and p and q donate time point p and time point q.

The function $w_{p,q}$ is used to penalize the distance between two time points so that it emphasizes the inherent resemblance, i.e., the nearby time points induce high resemblant sparse codes $\mathbf{S}$ and distant time points induce high disparities. The final objective function of MMLC stage-I multi-resemblant

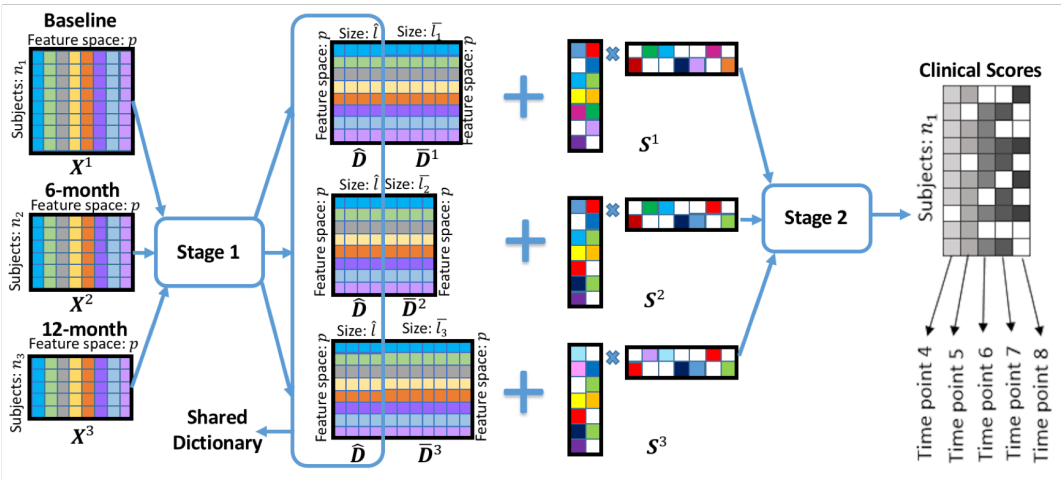

Fig. 2. Illustration of the learning process of MMLC on ADNI-I cohort from multiple different time points to predict multiple future time points clinical scores. In the figure, there are three input feature spaces from baseline, 6-month and 12-month as $\{\mathbf{X}^1, \mathbf{X}^2, \mathbf{X}^3\}$. We learn the dictionaries and sparse codes in stage 1. The dictionaries have two components (shared dictionary $\hat{\mathbf{D}}^t$ and task-specific dictionary $\bar{\mathbf{D}}^t$ corresponding to specific input $\mathbf{X}^t$). The sparse codes are low-rankness and have different resemblance between each others (e.g., $\mathbf{S}^1$, $\mathbf{S}^2$ and $\mathbf{S}^2$, $\mathbf{S}^3$ share higher resemblance, i.e., more common colors, than $\mathbf{S}^1$, $\mathbf{S}^3$). In stage 2, we use multi-target learning to predict multiple target clinical scores while dealing with missing label problem.

low-rank SC stage can be formalized as follows:

$$\min_{\mathbf{D}^t\in\Psi^t,\mathbf{S}^t} \sum_{t=1}^{T}(\frac{1}{2}||\mathbf{X}^t - [\hat{\mathbf{D}},\bar{\mathbf{D}}^t]\mathbf{S}^t||_F^2 + \lambda_1||\mathbf{S}^t||_{1,1} + \lambda_2||\mathbf{S}^t||_*) + \lambda_3\sum_{p=1}^{T-1}\sum_{q=p+1}^{T} w_{p,q}||\mathbf{S}^p - \mathbf{S}^q||_2^2. \tag{5}$$

where λ_3 is a non-negative regularization parameter. We will discuss how to optimize Eq. (5) in Sec. III.

Fig. 2 illustrates the learning process of MMLC with subjects of ADNI from three different time points which represents as $\mathbf{X}^1$, $\mathbf{X}^2$ and $\mathbf{X}^3$, respectively. Through the multi-resemblant low-rank SC stage (Stage 1), we obtain the dictionary and sparse codes for subjects from each time point t: $\mathbf{D}^t$ and $\mathbf{S}^t$. A dictionary $\mathbf{D}^t$ is composed by a shared dictionary $\hat{\mathbf{D}}^t$ across all tasks and a task-specific part $\bar{\mathbf{D}}^t$ only corresponding with the specific task $\mathbf{X}^t$. As a result, the sparse codes are low-rankness and have different resemblance between each others (e.g., $\mathbf{S}^1$, $\mathbf{S}^2$ and $\mathbf{S}^2$, $\mathbf{S}^3$ share higher resemblance, i.e., more common colors, than $\mathbf{S}^1$, $\mathbf{S}^3$).

C. MMLC Stage-II: Multi-Target Learning with Missing Labels

We measure the cognitive scores of patients at multiple time points in the longitudinal AD study. We formulate the prediction of clinical scores at multiple future time points simultaneously rather than considering the prediction of cognitive scores as a set of single time point regression since the intrinsic temporal smoothness information among different tasks can be incorporated into the model as the prior knowledge. However, there are many missing clinical scores at certain time points, especially for later time point (36 and 48 months) ADNI data. It will result in a huge information loss if we throw away these data in the prediction stage. It is necessary to incorporate the missing target values with multi-task regression to predict clinical scores [31], [36], [37].

Algorithm 2: Multi-Resemblance Multi-Target Low-rank Coding (MMLC)

Input : Samples $\mathbf{X^t}$ and corresponding labels $\mathbf{Y^t}$ from different time points, epoches κ, $\lambda_1, \lambda_2, \lambda_3$, μ_1, μ_2, γ, ϕ and $\hat{\mathbf{D}} = \mathbf{D}_0$.
Output: The models for different time points $\mathbf{W}^t$.

begin
  Stage I: Multi-Resemblance Low-Rank SC Stage
  for $k = 1 \rightarrow \kappa$ **do**
    for $t = 1 \rightarrow T$ **do**
      For each input matrix $\mathbf{X}^t$;
      Update $\mathbf{S}^{t,(k)}$ via Alg. 3;
      Update $||\mathbf{S}^{t,(k)}||_{1,1}$ and $||\mathbf{S}^{t,(k)}||_*$ by Eq. (14) and Eq. (15);
      Update $\hat{\mathbf{D}}^{(k)}$: $\hat{\mathbf{D}}^{(k)} = \mathbf{D}_0$ ($\mathbf{D}_0 = \hat{\mathbf{D}}^{(k-1)}$);
      Update the $\hat{\mathbf{D}}^{(k)}$ and $\bar{\mathbf{D}}^{t,(k)}$ via Alg. 4;
      Calculate $w_{p,q}$ function by Eq (4);
      Update $\mathbf{S}^{t,(k)}$ by Eq. (20);
      $\mathbf{D}_0 = \hat{\mathbf{D}}^{(k)}$;
  Obtain the learnt sparse codes $\mathbf{S}^t$, $t = 1, \cdots, T$.
  Stage II: Multi-Target Regression Stage
  for $t = 1$ *to* T **do**
    Given $\mathbf{Y}_j^t \in \mathbf{Y}^t$, for the jth model $\mathbf{w}_j^t \in \mathbf{W}^t$:
    $\mathbf{w}_j^t = (\widetilde{\mathbf{S}}^t\widetilde{\mathbf{S}}^{tT} + \xi\mathbf{I})^{-1}\widetilde{\mathbf{S}}^t\widetilde{\mathbf{Y}}_j^t$

In this paper, we use a matrix $\Theta \in \mathbb{R}^{n^t \times m^t}$ to indicate missing target values, where $\Theta_{i,j} = 0$ if the target value of label $\mathbf{Y}_{i,j}^t$ is missing and $\Theta_{i,j} = 1$ otherwise. Given the sparse codes $\{\mathbf{S}^1, ..., \mathbf{S}^T\}$ and corresponding labels $\{\mathbf{Y}^1, ..., \mathbf{Y}^T\}$ from different times where $\mathbf{Y}^t \in \mathbb{R}^{m^t \times n^t}$, we formulate the multi-target learning stage with missing target values as:

$$\min_{\mathbf{W}^1,\cdots,\mathbf{W}^T} \sum_{t=1}^{T}||\Theta(\mathbf{Y}^t - \mathbf{W}^t\mathbf{S}^t)||_F^2 + \xi\sum_{t=1}^{T}||\mathbf{W}^t||_F^2. \tag{6}$$

Algorithm 3: Updating sparse codes $\mathbf{s}_i^{t,(k+1)}$

Input : Image patch $\mathbf{x}_i^t$, dictionaries $\hat{\mathbf{D}}^{t,(k)}$ and $\bar{\mathbf{D}}^{t,(k)}$, sparse codes $\mathbf{s}_i^{t,(k)}$ and index set $\mathbb{I}_i^{t,(k)}$.
Output: $\mathbf{s}_i^{t,(k+1)}$ and $\mathbb{I}_i^{t,(k+1)}$.

```
begin
    for j = 1 to p^t do
        Update s_{i,j}^{t,(k+1)} by Eq. (8) and Eq. (9).
        if s_{i,j}^{t,(k+1)} ≠ 0 then
            Put j into the index set I_i^{t,(k+1)}.
    for j = 1 to Q do
        for l ∈ I_i^{t,(k+1)} do
            Update l by Eq. (11) and Eq. (12).
```

Although Eq. (6) is associated with missing values on the labels, we show that it has a close form and present the theoretical analysis of MMLC stage-II in Supplemental Material.

III. Optimization Analysis

In this section, we explain the updating procedures for MMLC. Eq. (5) is a non-convex problem. However, it will become a convex problem when we fix either $\mathbf{D}$ or $\mathbf{S}$. When the sparse codes $\mathbf{S}$ is fixed, solving dictionary $\hat{\mathbf{D}}$ and $\bar{\mathbf{D}}$ can be solved as a quadratically constrained quadratic programming (QCQP) problem [35]. At the end of each update in MMLC stage-I, we update the shared dictionary Φ: $\Phi = \hat{\mathbf{D}}^t$ and let $\hat{\mathbf{D}}^1 = \cdots = \hat{\mathbf{D}}^t$. When the dictionary $\mathbf{D}$ is fixed, solving each sparse code $\mathbf{s}_i$ can be view as a sparse group Lasso problem [38]. We alternately update $\mathbf{D}^t$ and $\mathbf{S}^t$ for $k = \kappa$ epoches and summarize the optimization details into Algorithm 2.

In Algorithm 2, for each image patch $\mathbf{x}_i^t$, we learn the i-th sparse code $\mathbf{s}_i^{t,(k+1)}$ from $\mathbf{s}^t$ by several steps of cyclic coordinate descent (CCD) [32]. We then use learnt sparse codes $\mathbf{s}_i^{t,(k+1)}$ to update the dictionary $\hat{\mathbf{D}}^{t,(k+1)}$ and $\bar{\mathbf{D}}^{t,(k+1)}$ by one step stochastic gradient descent (SGD) [33]. Since $\mathbf{s}_i^{t,(k+1)}$ is very sparse, we use the index set $\mathbf{I}_i^{t,(k+1)}$ to record the location of non-zero entries in $\mathbf{s}_i^{t,(k+1)}$ to accelerate the update of sparse codes and dictionaries. Φ is updated by the end of the k-th iteration to ensure $\hat{\mathbf{D}}^{t,(k+1)}$ is the same part among all the dictionaries.

A. Updating the low-rankness sparse codes

After we pick an image patch $\mathbf{x}_i^t$ from the sample $\mathbf{X}^t$ at the time point t, we fix the dictionary $\mathbf{D}$ and only consider updating the first sparse codes term $\mathbf{S}$. The optimization problem becomes the following form:

$$\min_{\mathbf{S}^t} \sum_{t=1}^{T} \left(\frac{1}{2}||\mathbf{X}^t - [\hat{\mathbf{D}}, \bar{\mathbf{D}}^t]\mathbf{S}^t||_F^2 + \lambda_1||\mathbf{S}^t||_{1,1}\right) \tag{7}$$

Coordinate descent [32] is known as one of the state-of-the-art methods for solving this Lasso problem [39]. In this study, we perform the CCD to optimize Eq. (7). Empirically, the iteration may take thousands of steps to converge, which is time-consuming in the optimization process of dictionary learning. However, we observe that after a few steps, the support of the coordinates, i.e., the locations of the non-zero entries in $\mathbf{s}_i^t$, becomes very stable, usually after less than ten steps. In this study, we perform P steps CCD to generate the non-zero index set $\mathbb{I}_t^{k+1}$, recording the non-zero entry of $\mathbf{s}_i^{t,(k+1)}$. Then we perform Q steps CCD to update the sparse codes only on the non-zero entries of $\mathbf{s}_i^{t,(k+1)}$, accelerating the learning process significantly. Stochastic coordinate coding (SCC) [40] employs a similar strategy to update the sparse codes in a single task. For the multi-task learning, we summarize the updating rules as follows:

(a) Perform P steps CCD to update the locations of the non-zero entries $\mathbb{I}_i^{t,(k+1)}$ and the model $\mathbf{s}_i^{t,(k+1)}$.
(b) Perform Q steps CCD to update the $\mathbf{s}_i^{t,(k+1)}$ in the index of $\mathbb{I}_i^{t,(k+1)}$.

In (a), we will pick up j-th coordinate to update the model $\mathbf{s}_{i,j}^t$ and non-zero entries, where $j \in \{1, ..., p^t\}$ in every CCD step. We perform the update from the 1st coordinate to the p^t-th coordinate. Meanwhile, we calculate the gradient $\mathbf{g}$ based on Eq. (7)) and update the model $\mathbf{s}_{i,j}^{t,(k+1)}$ based on $\mathbf{g}$. The calculation of $\mathbf{g}$ and $\mathbf{s}_{i,j}^{t,(k+1)}$ follows the equations:

$$\mathbf{g} = [\hat{\mathbf{D}}^{t,(k)}, \bar{\mathbf{D}}^{t,(k)}]_j^T (\Omega([\hat{\mathbf{D}}^{t,(k)}, \bar{\mathbf{D}}^{t,(k)}], \mathbf{s}_i^{t,(k)}, \mathbb{I}_i^{t,(k)}) - \mathbf{x}_i^t), \tag{8}$$

$$\mathbf{s}_{i,j}^{t,(k+1)} = \Gamma_\lambda(\mathbf{s}_{i,j}^{t,(k)} - \mathbf{g}), \tag{9}$$

where Ω is a sparse matrix multiplication function that has three input parameters. Taking $\Omega(\mathbf{A}, \mathbf{b}, \mathbb{I})$ as an example, $\mathbf{A}$ is a matrix, $\mathbf{b}$ denotes a vector and $\mathbb{I}$ records the locations of non-zero entries in $\mathbf{b}$ (an index set). The output value of Ω is defined as: $\Omega(\mathbf{A}, \mathbf{b}, \mathbb{I}) = \mathbf{Ab}$. We manipulate the non-zero entries of $\mathbf{b}$ and the corresponding columns of $\mathbf{A}$ based on the index set $\mathbb{I}$ when computing $\mathbf{Ab}$ so that we speed up the calculation by utilizing the sparsity of $\mathbf{b}$. Γ is the soft thresholding shrinkage function [41] as below:

$$\Gamma_\varphi(x) = sign(x)(|x| - \varphi). \tag{10}$$

In the end of (a), we count the non-zero entries in $\mathbf{s}_i^{t,(k+1)}$ and store the non-zero index in $\mathbb{I}_i^{t,(k+1)}$. In (b), we perform Q steps CCD by only considering the non-zero entries in $\mathbf{s}_i^{t,(k+1)}$. As a result, for each index $l \in \mathbb{I}_i^{t,(k+1)}$, we calculate the gradient $\mathbf{g}$ and update the $\mathbf{s}_{i,l}^{t,(k+1)}$ by:

$$\mathbf{g} = [\hat{\mathbf{D}}^{t,(k)}, \bar{\mathbf{D}}^{t,(k)}]_l^T (\Omega([\hat{\mathbf{D}}^{t,(k)}, \bar{\mathbf{D}}^{t,(k)}], \mathbf{s}_i^{t,(k+1)}, \mathbb{I}_i^{t,(k+1)}) - \mathbf{x}_i^t), \tag{11}$$

$$\mathbf{s}_{i,l}^{t,(k+1)} = \Gamma_\lambda((\mathbf{s}_{i,l}^{t,(k+1)} - \mathbf{g}). \tag{12}$$

Since we only focus on the non-zero entries of the model and P is less than 10 iterations and Q is a much larger number, we significantly accelerate the entire sparse codes learning process. The procedure of updating sparse codes can be summarized into Algorithm 3.

However, in Eq. (5), there are two convex and non-smooth regularizers for $\mathbf{S}^t$. We propose to update the low-rankness sparse codes by using the conventional Inexact Augmented Lagrange Multiplier (IALM) [42]. IALM is an iterative method that augments the Lagrangian function with quadratic penalty terms, which allows closed-form updates for each variable in the problem. Therefore, solving the ℓ_1 and the nuclear norm

Algorithm 4: Updating dictionaries $\hat{\mathbf{D}}_t^{k+1}$ and $\bar{\mathbf{D}}_t^{k+1}$

Input : Image patch $\mathbf{x}_i^t$, dictionaries $\hat{\mathbf{D}}^{t,(k)}$ and $\bar{\mathbf{D}}^{t,(k)}$, sparse codes $\mathbf{s}_i^{t,(k+1)}$ and index set $\mathbb{I}_i^{t,(k+1)}$.

Output: The updated dictionaries $\hat{\mathbf{D}}_t^{k+1}$ and $\bar{\mathbf{D}}_t^{k+1}$

begin
- Update the Hessian matrix $\mathbf{H}_t^{k+1}$ by Eq. (17).
- $R = \Omega([\hat{\mathbf{D}}^{t,(k)}, \bar{\mathbf{D}}^{t,(k)}], \mathbf{s}_i^{t,(k+1)}, \mathbf{I}_i^{t,(k+1)}) - \mathbf{x}_i^t$.
- **for** $j = 1$ *to* Q **do**
  - **for** $l \in \mathbf{I}_i^{t,(k+1)}$ **do**
    - Update every element l by Eq. (19).

will result in solving the following problem, where we use two slack variables $\mathbf{S}_2^t$ and $\mathbf{S}_3^t$ for the two terms:

$$\begin{aligned}\min_{\mathbf{D}^t\in\Psi^t,\mathbf{S}_1^t,\mathbf{S}_2^t,\mathbf{S}_3^t} \sum_{t=1}^{T}(&\frac{1}{2}||\mathbf{X}^t - [\hat{\mathbf{D}},\bar{\mathbf{D}^t}]\mathbf{S}_1^t||_F^2 + \lambda_1||\mathbf{S}_2^t||_{1,1}\\ &+\lambda_2||\mathbf{S}_3^t||_* + tr[L_1(\mathbf{S}_1^t - \mathbf{S}_2^t)] + tr[L_2(\mathbf{S}_1^t - \mathbf{S}_3^t)]\\ &+\frac{\mu_1}{2}||\mathbf{S}_1^t - \mathbf{S}_2^t||_F^2 + \frac{\mu_2}{2}||\mathbf{S}_1^t - \mathbf{S}_3^t||_F^2),\end{aligned} \tag{13}$$

where L_1 and L_2 are lagrange multipliers, and μ_1 and μ_2 are two positive scalars. IALM efficiently minimize Eq. (13) and the validity and optimality of Eq. (13) is guaranteed by the following theorem.

***Theorem** 1:* For Eq. (13), if $\{\mu_r^k\}(r = 1, 2)$ is non-decreasing and $\sum_{k=1}^{+\infty} 1/\mu_r^k = +\infty$ then $(\mathbf{S}_2, \mathbf{S}_3)$ converge to an optimal solution $(\mathbf{S}_2^*, \mathbf{S}_3^*)$.

We provide the proof of Theorem 1 in Supplemental Material.

Theorem 1 only guarantees convergence but does not specify the rate of convergence for the IALM method and we discuss the convergence rate at the end of this section. we use blockwise coordinate descent to alternatively update each variable of $\mathbf{S}_1^t$, $\mathbf{S}_2^t$, $\mathbf{S}_3^t$ with all other variables fixed to their most recent values as follows:

$$\begin{aligned}&\mathbf{S}_2^{t*} = \Omega_{\frac{\lambda_1}{\mu_1}}(\mathbf{S}_1^t + \frac{L_1}{\mu_1}), \mathbf{S}_3^{t*} = \Theta_{\frac{\lambda_2}{\mu_2}}(\mathbf{S}_1^t + \frac{L_2}{\mu_2}),\\ &\mathbf{S}_1^{t*} = (\mathbf{D}^{tT}\mathbf{D}^t\mu_1\mathbf{I} + \mu_2\mathbf{I})^{-1}\mathbf{G},\end{aligned} \tag{14}$$

where $\mathbf{G} = \mathbf{D}^{tT}\mathbf{X}^t - L_1 - L_2 + \mu_1\mathbf{S}_2^t + \mu_2\mathbf{S}_3^t$, $\Omega_\lambda(\mathbf{S}) = sign(\mathbf{S})(|\mathbf{S}| - \lambda)_+$ is the soft-thresholding operator and $\Theta_\lambda(\mathbf{S}) = U\Omega_\lambda(\mathbf{\Sigma})V^T$ is the singular value soft-thresholding operator with $\mathbf{S} = U\Sigma V^T$ is the SVD of $\mathbf{S}$. Then, we can update the multipliers with $\phi > 1$ as follows,

$$\begin{aligned}&L_1 = L_1 + \mu_1(\mathbf{S}_1^t - \mathbf{S}_2^t); L_2 = L_2 + \mu_2(\mathbf{S}_1^t - \mathbf{S}_3^t);\\ &\mu_1 = \phi\mu_1; \mu_2 = \phi\mu_2.\end{aligned} \tag{15}$$

After we obtain $\mathbf{S}_1^{t*}$ as $\mathbf{S}^t$, we then fix $\mathbf{S}^t$ to update $\mathbf{D}^t$.

B. Updating common and task-specific dictionaries

We update the dictionaries by fixing the sparse codes, thus, and the optimization problem becomes:

$$\min_{\hat{\mathbf{D}}^t,\bar{\mathbf{D}}^t} \mathcal{F}(\hat{\mathbf{D}}^t, \bar{\mathbf{D}}^t) = \frac{1}{2}||\mathbf{x}_i^t - [\hat{\mathbf{D}}^t, \bar{\mathbf{D}}^t]\mathbf{s}_i^t||_2^2 \tag{16}$$

We know the non-zero entries of $\mathbf{s}_i^{t,(k+1)}$ after we updating the sparse codes. The key insight of MMLC is that we just need to update the non-zero entries of the dictionaries but not all columns of the dictionaries, and it dramatically accelerates the optimization. When updating the i-th column and j-th row's entry of the dictionary $\mathbf{D}$, the gradient of $\mathbf{D}_{j,i}$ is set to be $\nabla\mathbf{D}_{j,i} = \mathbf{s}_i(\mathbf{D}_j^T\mathbf{s} - \mathbf{x}_j)$. If $\mathbf{s}_i = 0$, the gradient would be zero. We therefore do not need to update the $\mathbf{D}_j$. The learning rate is set to be an approximation of $1/\mathbf{H}_t^{k+1}$, which is updated by the sparse codes $\mathbf{s}_i^{t,(k+1)}$ in k-th iteration. We first update the Hessian matrix $\mathbf{H}_t^{k+1}$ by:

$$\mathbf{H}_t^{k+1} = \mathbf{H}_t^k + \mathbf{s}_i^{t,(k+1)}\mathbf{s}_i^{(t,(k+1))T}. \tag{17}$$

One step SGD is performed to update the dictionaries: $\hat{\mathbf{D}}_t^{k+1}$ and $\bar{\mathbf{D}}_t^{k+1}$. We use a vector $\mathbf{R}$ to store the information $\mathbf{Dz} - \mathbf{x}$ in order to speed up the computation.

$$\mathbf{R} = \Omega([\hat{\mathbf{D}}^{t,(k)}, \bar{\mathbf{D}}^{t,(k)}], \mathbf{s}_i^{t,(k+1)}, \mathbb{I}_i^{t,(k+1)}) - \mathbf{x}_i^t. \tag{18}$$

Here, $\mathbf{R} = \tau([\hat{\mathbf{D}}^{(k-1)}, \bar{\mathbf{D}}^{t,(k-1)}], \mathbf{S}^{t,(k)}) - \mathbf{X}^t$, where $\tau(\mathbf{A}, \mathbf{B})$ is a matrix multiplication function and $\tau(\cdot) = \mathbf{AB}$. The procedure of learning the l-th column and j-th row of dictionaries takes the form of

$$[\hat{\mathbf{D}}_t^{k+1}, \bar{\mathbf{D}}_t^{k+1}]_{j,l} = [\hat{\mathbf{D}}^{t,(k)}, \bar{\mathbf{D}}^{t,(k)}]_{j,l} - \frac{1}{\mathbf{H}_t^{k+1}(l,l)}\mathbf{s}_{i,l}^{t,(k+1)}\mathbf{R}_j, \tag{19}$$

where l is the non-zero entry stored in $\mathbb{I}_i^{t,(k+1)}$. We let the learning rate be the inverse of the diagonal element of the Hessian matrix as $1/\mathbf{H}_t^{k+1}(l,l)$ for the l-th column of the dictionary.

It is important to normalize the dictionaries $\hat{\mathbf{D}}^{t,(k+1)}$ and $\bar{\mathbf{D}}^{t,(k+1)}$ after updating them because of $\mathbf{D}_t \in \Psi_t$ in equation (Eq. (16)). Since the dictionaries updating procedure only occurs at non-zero entries, we perform the normalization on the the corresponding columns of $\mathbf{s}_i^{t,(k+1)}$. The step of utilizing non-zero entries from $\mathbb{I}_i^{t,(k+1)}$ accelerates the whole learning process. We summarized the updating rules of dictionaries into Algorithm 4.

C. Updating resemblance term

After we update $\mathbf{D}^t$, we finally calculate $w_{p,q}$, and update the fourth term of Eq. (5) at the end of k-th epoch. We update the inherent resemblant knowledge term with the iterative soft-thresholding [43]. We first calculate the gradient $\mathbf{g}$ based on Eq. (20), and then update the model $\mathbf{S}^{t,(k)}$ based on $\mathbf{g}$. The calculation of $\mathbf{g}$ and $\mathbf{S}^{t,(k)}$ follows the equations:

$$\begin{aligned}&\mathbf{g} = \frac{1}{\gamma}\mathbf{D}^t\mathbf{X}^t + [\mathbf{I} - \frac{1}{\gamma}(\mathbf{D}^{tT}\mathbf{D}^t + w_{p,q}\lambda_3\mathbf{I})]\mathbf{S}^{t,(k-1)},\\ &\mathbf{S}^{t,(k)} = \Omega_{\lambda_3}(\mathbf{g} + w_{p,q}\frac{\lambda_3}{\gamma}\mathbf{D}^t),\end{aligned} \tag{20}$$

where γ is a non-negative parameter and Ω_{λ_3} is the soft-thresholding operator. Details of MMLC updating rules can be found in Algorithm 2.

The convergence of MMLC algorithm is reached when the error of the objective function is below a threshold $\epsilon = 10^{-3}$ and the SVD of $\mathbf{S}$ can be computed efficiently with time complexity $O(mnl)$, where $l < min(m, n)$ is its rank. It is worth noting that the overall computational complexity of

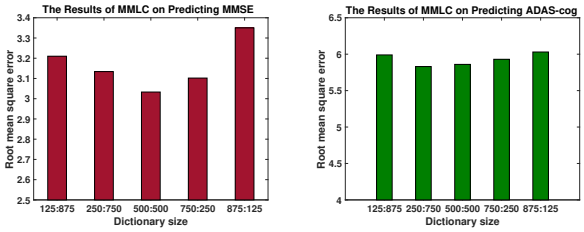

Fig. 3. Comparison of rMSE performance by varying the size of common dictionary.

TABLE I
TIME COMPARISONS OF MMLC AND STSC BY VARYING DICTIONARY SIZE ON ADNI-I DATASET.

Dictionary Size	MMLC	STSC
500	1.74 hour	8.84 hour
1000	3.34 hour	21.95 hour
2000	6.93 hour	49.90 hour

MMLC is $O(m^3+\epsilon^{-0.5}mn+m^2n)$ when the number of IALM iterations is $O(\epsilon^{-0.5})$. This is much faster than the complexity of conventional method $O(m^3+m^2n+mn^2)$.

IV. EXPERIMENTS

A. Dataset

Data is downloaded from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database ([44], adni.loni.usc.edu). ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations. Subjects have been recruited from over 50 sites across the U.S. and Canada. The primary goal of ADNI is to test whether biological markers, such as serial MRI and positron emission tomography (PET), combined with clinical and neuropsychological assessments, can measure the progression of mild cognitive impairment (MCI) and early AD. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, see www.adniinfo.org.

In this work, we study the performance of MMLC on the entire *ADNI-1 cohort*. We use T1-weighted magnetic resonance images (MRIs) coming from seven different time points: baseline, 6-, 12-, 18-, 24-, 36- and 48-month. 837, 733, 728, 326, 641, 454 and 251 are the sample sizes corresponding to seven time points, respectively. Thus, we learn a total of 3970 images and the responses are the Mini Mental State Examination (MMSE) and Alzheimer's Disease Assessment Scale cognitive subscale (ADAS-Cog) score. In addition, we remove 23 subjects who do not have MMASE and ADAS-cog information at baseline in this work.

B. Experimental Setting

1) Surface features: We use hippocampal surface multivariate morphometry statistics (MMS) [14] (Fig. 1 (c)) as our learning features. The original input data are the three-dimensional (3D) T1-weighted images (Fig. 1 (a)) from ADNI dataset. We first use FIRST(https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FIRST) to segment the original data and obtain the hippocampus substructure (Fig. 1 (b)). We then adopt the surface fluid registration [45] to obtain surface geometric features for automated surface registration. Following that, a set of vertex-wise hippocampal MMS features are computed as [14]. They consist of surface multivariate tensor-based morphometry (mTBM) and radial distance (RD). mTBM describes the surface deformation along the surface tangent plane while RD reflects surface differences along the surface normal directions. MMS features consist 4×1 vectors on each vertex of 15000 vertices on every hippocampal surface (each subject has two hippocampal surfaces). We select 1102 patches of size 10×10 on each hippocampal surface mesh and each patch dimension is 400. We use the baseline and 6-month imaging data as training data and predict 12-month to 48-month clinical scores.

2) MMLC settings: The model is trained on an Intel(R) Core(TM) i7-6700 K CPU with 4.0GHz processors, 64 GB of globally addressable memory and a single Nvidia TITAN X GPU. The source code of MMLC are available at http://gsl.lab.asu.edu/software/mmlc. In stage one, $\lambda_1=0.1$, $\lambda_2=10^{-2}$, $\lambda_3=10^{-3}$, $\mu_1=10$, $\mu_2=1$ and $\gamma=1$, $\phi=10$. The SCC sparsity parameter (λ_1) is the best parameter setting as [40]. The rest parameters are selected by cross-validation results on the training data. For example, we use 5-fold cross-validation with a grid search to pick the best parameters for λ_2 and λ_3 from {1000, 100, 10, 1, 0.1, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}}. In stage two, cross-validation is used to select model parameters ξ (between 10^{-3} and 10^3). For common and individual dictionary split, we compare the performance by varying the dictionary size as 125:875, 250:750, 500:500, 750:250, 875:125. We observe that the algorithm has the best performance while the ratio between the common dictionary and the individual parts is 1:1. Therefore, in all experiments, we use 1000 atoms for the dictionary and 500:500 split atoms as the size of common and task-specific dictionaries (Sec. IV-C1). When the sparse features are learned, Max-Pooling is used to generate features for annotation and finally we got a 1000-dimensional feature vector for each subject.

3) Evaluation method: In order to evaluate the model, we randomly split the data into training and testing sets using a 9:1 ratio to avoid data bias and report the mean and standard deviation based on 50 different splits of data. We evaluate the overall regression performance using weighted correlation coefficient (wR) and root mean square error (rMSE) for task-specific regression performance measures. The two measures are defined as $wR(\mathbf{Y},\hat{\mathbf{Y}})=\sum_{t=1}^{T}Corr(\mathbf{Y}^t,\hat{\mathbf{Y}}^t)n^t/\sum_{t=1}^{T}n^t$, $rMSE(\mathbf{Y}^t,\hat{\mathbf{Y}}^t)=\sqrt{||\mathbf{Y}^t-\hat{\mathbf{Y}}^t||_2^2/n^t}$. For wR, $\mathbf{Y}^t$ is the ground truth of target of task t and $\hat{\mathbf{Y}}^t$ is the corresponding predicted value, $Corr$ is the correlation coefficient between two vectors and n^t is the number of subjects of task t. rMSE is computed for each task t, $\mathbf{Y}^t$ is the ground truth of the target responses and $\hat{\mathbf{Y}}^t$ is the corresponding prediction. The smaller rMSE, the bigger wR mean the better results.

4) Comparison methods: We compare the proposed algorithm MMLC with six other methods: 1) single-task regression methods: LASSO [39] and Ridge [46]; 2) multi-task regression methods: multi-task regression with $\ell_{2,1}$ norm regularization [47] (L21) and temporal group Lasso based multi-task

TABLE II
PERFORMANCE COMPARISON BETWEEN THE PROPOSED ALGORITHM (MMLC) AND SIX OTHER METHODS (SEC. IV. B. (4)) ON PREDICTING FUTURE MMSE SCORES OF 12-, 18-, 24-, 36-, 48-MONTH BASED ON BASELINE AND 6-MONTH HIPPOCAMPAL MORPHOMETRY DATA ON THE WHOLE ADNI-I DATASET.

Methods	wR	M12	M18	M24	M36	M48
Lasso	0.40±0.09	4.04±0.77	3.46±0.97	5.53±0.86	4.39±0.74	4.73±1.49
Ridge	0.41±0.07	4.26±0.56	3.56±0.93	5.05±0.54	4.21±0.47	3.62±0.91
L21	0.57±0.01	3.32±0.63	4.75±0.75	4.64±0.88	4.08±1.01	3.11±1.05
ODL-L	0.63±0.08	2.99±0.63	**2.88±0.68**	4.29±0.84	3.62±1.45	2.93±1.07
TGL	0.70±0.05	2.73±0.72	4.00±1.31	4.00±0.64	3.19±1.38	2.60±1.42
MTSC	0.73±0.02	2.61±0.55	3.37±1.01	3.66±0.78	2.73±1.09	2.52±1.20
MMLC	**0.75±0.02**	**2.55±0.23**	2.99±0.89	**3.38±0.76**	**2.65±0.79**	**2.32±1.02**

TABLE III
PERFORMANCE COMPARISON BETWEEN THE PROPOSED ALGORITHM (MMLC) AND SIX OTHER METHODS (SEC. IV. B. (4)) ON PREDICTING FUTURE ADAS-COG SCORES OF 12-, 18-, 24-, 36-, 48-MONTH BASED ON BASELINE AND 6-MONTH HIPPOCAMPAL MORPHOMETRY DATA ON THE WHOLE ADNI-I DATASET.

Methods	wR	M12	M18	M24	M36	M48
Lasso	0.49±0.05	6.81±1.03	6.87±0.74	7.62±0.87	8.08±1.39	6.55±1.34
Ridge	0.46±0.07	7.68±0.96	6.89±1.69	7.84±1.54	8.59±0.62	6.64±1.58
L21	0.53±0.07	6.40±0.51	6.95±0.88	8.07±0.67	8.00±1.04	5.92±0.60
ODL-L	0.53±0.05	5.65±0.73	4.97±0.67	7.30±0.77	7.25±0.69	5.56±1.22
TGL	0.72±0.04	5.52±1.15	5.70±0.53	6.85±1.06	**6.36±1.22**	5.73±0.61
MTSC	0.77±0.02	5.18±0.88	**4.64±1.12**	6.76±1.35	6.78±1.54	5.27±1.76
MMLC	**0.80±0.04**	**5.17±0.95**	4.87±0.99	**6.66±0.65**	**6.37±1.23**	**5.16±1.31**

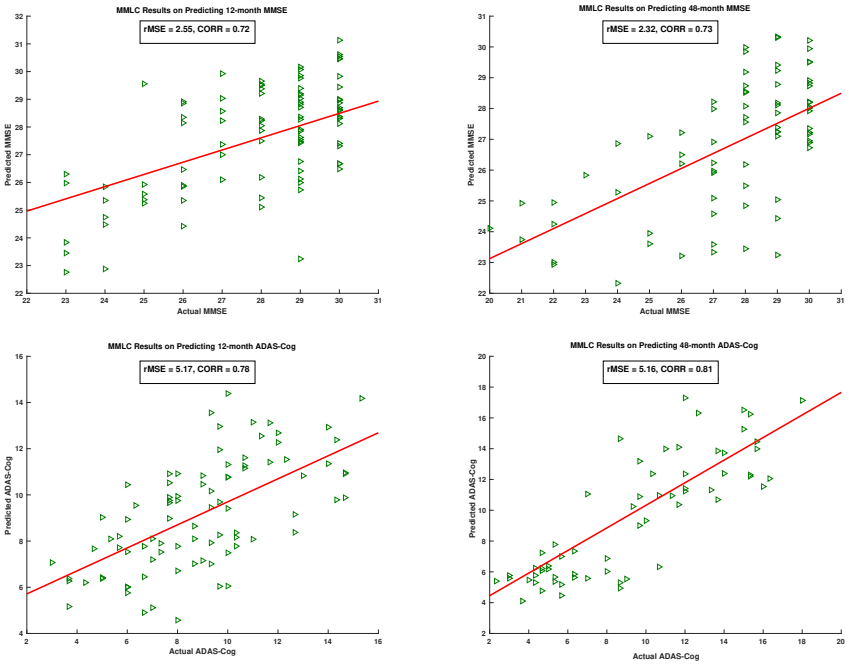

Fig. 4. Scatter plots of actual MMSE and ADAS-Cog versus predicted values on M12 and M48 by using MMLC.

progression model [31] (TGL); 3) sparse coding-based methods: single-task sparse coding followed by Lasso [21] (STSC), Muilti-source Multi-target dictionary learning followed by Lasso regression [20] (MSMT) ($\lambda_2 = 0$ and $\lambda_3 = 0$ in Eq. (5)).

C. Experimental Results

1) The atoms of common and task-specific dictionaries: In stage one of MMLC, the common dictionary is assumed to be shared by different tasks. It is necessary to evaluate what is an appropriate size of such common dictionary. Therefore, we set the dictionary size to be 1000 and partition the dictionary by different proportions: 125:875, 250:750,500:500, 750:250 and 875:125, where the left number is the size of common dictionary while the right one is the size of individual dictionary for each task. Fig. 3 shows the results of rMSE of MMSE and ADAS-cog prediction. As it shows in Fig. 3, the rMSE of MMSE and ADAS-Cog are lowest when we split the dictionary by half and a half. It means the both of common and individual dictionaries are of equal importance during the multi-task learning.

2) Comparison with two other sparse coding methods: There are quite a variety of sparse coding approaches in

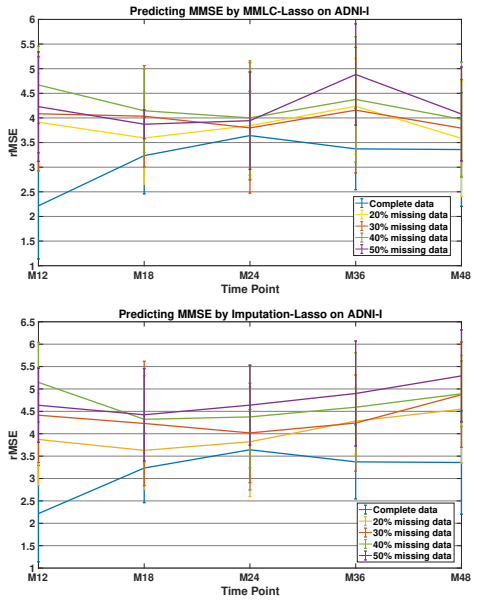

Fig. 5. The rMSE results of MMSE with different amount missing data by MMLC-Lasso and Imputation-Lasso, respectively.

the literature. We compare our work with two other sparse coding methods. We use the online dictionary learning code package for (ODL) [18] method. We also implement the low-rank shared dictionary learning (LRSDL) method, based on the paper [48] and the github source code [1]. To simplify the comparison experiments, we adopt the classification problem in our prior work [49] where we apply Stochastic Coordinate Coding (SCC) to generate sparse hippocampal surface features for classification studies. In this problem, its objective function is the same as Eq. (1) for ODL and SCC (we provide the objective function for LRSDL in *Supplemental Material*). We conduct 6 different classification experiments and test *ODL, SCC and LRSDL* measures in terms of running time, and objective function value, respectively. For the comparison methods, we select the hyper-parameter for LRSDL by using the same strategy as SCC on the training set. We report the detailed experimental results in *Supplemental Material*. In summary, among these three methods, SCC achieves the best balance between performance and the running time. The experimental results may justify our selection of SCC method for the studied problem.

3) The comparisons of time efficiency: We compare the efficiency of our proposed MMLC with STSC (Algorithm 1). In this experiment, we focus on the single batch size setting, that is, we process one image patch in each iteration. We vary the dictionary size as: 500, 1000 and 2000. For MMLC, the ratio between the common dictionary and the individual parts is 1:1. We report the results on ADNI-I cohort in Table I. We observe that the proposed MMLC uses less time than STSC. When the size of dictionary increases, MMLC is more efficient and has a higher speedup compared to STSC.

4) Comparison results on MMSE and ADAS-cog: We report the comparison results of MMLC and other methods of MMSE and ADAS-cog with ADNI-1 cohort in Table II and Table III, respectively. In both tables, we can find that the cognition predictions produced by MMLC achieves the highest correlation with the ground truth data. In Fig. 4, we can find that MMLC achieves relatively high correlation on both 12-month and 48-month prediction results. It shows that the prediction results of MMLC do not decrease quickly for the long term prediction. After MMLC formulates temporary sequence information, the results are more linear, reasonable and accurate on all time points. Moreover, MMLC and MSMT methods can handle missing data on both source and target sides. L21 and TGL can deal with missing target data while neither Lasso nor Ridge can deal with missing data.

In Table II, the proposed MMLC outperforms linear regression methods in terms of both rMSE and correlation coefficient wR on four different time points. The results of Lasso and Ridge are very close while sparse coding methods are superior to them. For sparse coding methods, we observe that MTSC obtains lower rMSE and higher correlation results than STSC since MTSC considers the correlation between different time slots and the task-specific relationship. STSC has lower rMSE than MMLC on M18 because 18-month data is significantly less than other time points and SC has its bias on that point. We also notice that the proposed MMLC further improves the result of MTSC since we consider the low-rankness of the sparse codes and the resemblant knowledge in longitudinal dataset. Note that we significantly improve the rMSE results for later time points. A possible reason is that the baseline images have less correlation with later time points images and MTSC treats each time point equally.

In Table. III, we can observe that the best performance of predicting scores of ADAS-Cog is achieved by MMLC in four-time points. Comparing with L21, after MMLC dealing with missing labels, the results are more linear, reasonable and accurate. Due to the dimension of M36 and M48 is too small, it is hard to learn a complete model. TGL also considers the issue of missing labels, however, MMLC achieves better results because MMLC incorporates multiple-source data and uses common and individual dictionaries. This shows our method is more efficient in dealing with incomplete data.

We also notice that the proposed MMLC further improved the results of MSMT since we consider the low-rankness of the sparse codes and the resemblant knowledge in longitudinal dataset. Note that we significantly improve the rMSE results for later time points. A possible reason is that the baseline images have less correlation with later time points images and MSMT treats each time point equally.

We show the scatter plots for the predicted values versus the actual values for MMSE and ADAS-Cog on the M12 and M48 in Fig. 4. In the scatter plots, we see the predicted values and actual clinical scores have a high correlation. The scatter plots show that the prediction performance for ADAS-Cog is better than that of MMSE.

5) Ablation study on different amount of missing data: Furthermore, we study whether MMLC helps improve incomplete data results by varying different amounts of missing data. We start with a total of 122 subjects, which have complete MMSE value at all seven time points. We then randomly removed 20%, 30%, 40% and 50% target values during

[1]https://github.com/tiepvupsu/DICTOL_python

training. We perform our algorithm MMLC to the complete data and different amounts of incomplete data. For comparison purposes, we apply the imputation approach [50] to complete the missing data which uses neighboring time point data to approximate the missing value. For the experimental settings, we follow those of Sec. IV-B2. Fig. 5 shows the rMSE results with different amounts of missing data. The results show that compared with the imputation method [50], our approach has better results that are close to the performance with the complete data.

V. Discussion

In AD research, structural MRI-based hippocampal morphometry measures correlate closely with differences and changes in cognitive performance [51], [52], supporting their validity as markers of AD progression. Recent research further demonstrated that hippocampal morphometry may be used to predict amyloid burden [53], [54] and identify AD related changes in the preclinical stage [55], [15]. In this work, we found that one may predict future cognitive decline by analyzing longitudinal hippocampal morphometry changes. Therefore, our work supports the potential to use sMRI biomarkers as predictors of disease progression.

In this work, we adopted FIRST for hippocampus segmentation. However, our hippocampal morphometry system has utilized different segmented hippocampal data as input. For example, our earlier work (e.g., [56], [57], [58]) used manually segmented hippocampi to build surface meshes. Later we adopted FIRST for automatic hippocampus segmentation [45] and used it in almost all our hippocampal morphometry research. Meanwhile, we also used FreeSurfer segmented hippocampi to build hippocampal surface meshes [59]. All of them achieved reasonable results in group difference studies and thus the results demonstrated that our pipeline is robust to segmentation methods. The reason for us to choose FIRST for most of work is that FIRST can always generate topologically sound segmentation results while FreeSurfer does not guarantee topologically correct results. Therefore, manual quality control is necessary to incorporate FreeSurfer in our pipeline. Thus far, our related prediction/classification work (e.g., [16], [60]) all adopted FIRST segmented hippocampal surfaces to work with relatively large scaled datasets. Since the input of our MMLC is the surface features rather than the output from segmentation tools, it is reasonable for us to expect that our method is not sensitive to the hippocampus segmentation tools used.

In ADNI, the scan times "12-month", "24-month" etc. are nominal times. With the baseline data, we computed the exact interval months for all longitudinal data used in our research. The average months and their standard deviations on each time point are 6.94 ± 0.96, 12.98 ± 1.01, 19.10 ± 1.06, 25.18 ± 1.41, 37.15 ± 1.37 and 49.43 ± 1.42 for 6-, 12-, 18-, 24-, 36- and 48-month data, respectively. It shows that 6-month data are not exactly scanned in the following 6-month. One way to make them perfectly aligned to a specific month may be a linear interpolation. However, it would assume all features change linearly with time, a strong assumption which we try to avoid in our formulation. On the other hand, our multi-task model does not make any specific assumption on the relationship between features on a specific time point. Our model simply assumes that the time points are similar to each other so that they can be clustered together (i.e., with the same time label in the same matrix X^t). It can be uniquely applied to analyze longitudinal data which are not collected in the exact time points (such as ADNI, Australian Imaging Biomarkers and Lifestyle Study of Aging (AIBL) [61] and Arizona APOE cohorts [62]). Although we believe that the development of more refined analysis models is necessary, our current experimental results show that our models may be effectively applied to analyze such longitudinal data.

In Supplementary Material, we show that the final objective function is non-decreasing and converges to an optimal solution. However, as the objective function is not convex, the problem may have multiple solutions. For general cases, existing works show for such problems, the objective function values increasingly convert to some value in each iteration but whether it converges to the global optimum is still an open problem [63], [64], [65]. With the current "greedy" strategy in each optimization iteration, we can guarantee that the solution converges to a local optimum. To show the local optimal solution is also the global optimum, we empirically repeated our experiments several times with different random initializations and the solutions of our proposed method converged to the minimum values which are very close to each other. Besides, in Supplementary Material, we also compare the objective function values of our MMLC methods with two state-of-the-art methods, online dictionary learning (ODL) [18] and low-rank shared dictionary learning (LRSDL) [48] methods. With a similar experimental setup, the three minimal objective function values are quite close. These results empirically support that our work may converge to the global optimum in the current study.

This work represents our initial efforts to develop robust machine learning algorithms to study the prediction of cognitive decline with both incomplete longitudinal brain images and incomplete clinical labels. Nonetheless, there is still much to be desired in our current experimental results on ADNI cohort. For example, in both Table II and Table III, although we generally achieved smaller rMSE results compared to other methods, on some time points, our work only achieved slightly improved results and our work sometimes had larger standard deviation. In the future, we will further evaluate our work in larger brain imaging cohorts (e.g., UKBiobank imaging study [66]). Meanwhile, we will continue refining our methods by exploring the underlying feature-feature relationship and it may further improve our results.

VI. Conclusion and Future Work

In this paper, we propose a novel multi-task sparse coding framework together with an efficient numerical scheme (MMLC). Our experimental results clearly show MMLC offers a unique perspective on prognosis with longitudinal data. In the future, we will incorporate our recent feature selection model [67] to visualize the identified imaging biomarkers. We

will also refine our system by considering the design of a hierarchical model to further improve its statistical power.

ACKNOWLEDGMENT

This research is supported in part by National Institutes of Health (RF1AG051710, R01EB025032, U54EB020403, R21AG065942, R01AG031581 and P30AG19610), ASU-Mayo seed grant, and Arizona Alzheimer's Consortium. 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: Alzheimer's Association; Alzheimer' s Drug Discovery Foundation; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to Rev December 5, 2013 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] M. W. Weiner, D. P. Veitch, P. S. Aisen, L. A. Beckett, N. J. Cairns, R. C. Green, and et al., "The Alzheimer's Disease Neuroimaging Initiative: a review of papers published since its inception," *Alzheimers Dement*, vol. 8, no. 1 Suppl, pp. 1–68, Feb 2012.

[2] G. B. Frisoni, N. C. Fox, C. R. Jack Jr, P. Scheltens, and P. M. Thompson, "The clinical use of structural mri in Alzheimer disease," *Nature Reviews Neurology*, vol. 6, no. 2, p. 67, 2010.

[3] N. C. Fox, E. K. Warrington, P. A. Freeborough, P. Hartikainen, A. M. Kennedy, J. M. Stevens, and et al., "Presymptomatic hippocampal atrophy in Alzheimer's disease. A longitudinal MRI study," *Brain*, vol. 119 (Pt 6), pp. 2001–2007, Dec 1996.

[4] B. C. Dickerson, I. Goncharova, M. P. Sullivan, C. Forchetti, R. S. Wilson, D. A. Bennett, and et al., "MRI-derived entorhinal and hippocampal atrophy in incipient and very mild Alzheimer's disease," *Neurobiol Aging*, vol. 22, no. 5, pp. 747–754, 2001.

[5] K. A. Josephs, P. R. Martin, S. D. Weigand, N. Tosakulwong, M. Buciuc, M. E. Murray, and et al., "Protein contributions to brain atrophy acceleration in Alzheimer's disease and primary age-related tauopathy," *Brain*, vol. 143, no. 11, pp. 3463–3476, Dec 2020.

[6] L. Nadal, P. Coupe, C. Helmer, J. V. Manjon, H. Amieva, F. Tison, and et al., "Differential annualized rates of hippocampal subfields atrophy in aging and future Alzheimer's clinical syndrome," *Neurobiol Aging*, vol. 90, pp. 75–83, 06 2020.

[7] D. H. Adler, L. E. M. Wisse, R. Ittyerah, J. B. Pluta, S. L. Ding, L. Xie, and et al., "Characterizing the human hippocampus in aging and Alzheimer's disease using a computational atlas derived from ex vivo MRI and histology," *Proc Natl Acad Sci U S A*, vol. 115, no. 16, pp. 4252–4257, 04 2018.

[8] K. Zhao, Y. Ding, Y. Han, Y. Fan, A. F. Alexander-Bloch, T. Han, and et al., "Independent and reproducible hippocampal radiomic biomarkers for multisite Alzheimer's disease: diagnosis, longitudinal progress and biological basis," *Science Bulletin*, vol. 65, no. 13, pp. 1103 – 1113, 2020.

[9] L. G. Apostolova, R. A. Dutton, I. D. Dinov, K. M. Hayashi, A. W. Toga, J. L. Cummings, and et al., "Conversion of mild cognitive impairment to Alzheimer disease predicted by hippocampal atrophy maps," *Arch Neurol*, vol. 63, no. 5, pp. 693–699, May 2006.

[10] D. P. Devanand, R. Bansal, J. Liu, X. Hao, G. Pradhaban, and B. S. Peterson, "MRI hippocampal and entorhinal cortex mapping in predicting conversion to Alzheimer's disease," *Neuroimage*, vol. 60, no. 3, pp. 1622–1629, Apr 2012.

[11] E. Gerardin, G. Chetelat, M. Chupin, R. Cuingnet, B. Desgranges, H. S. Kim, and et al., "Multidimensional classification of hippocampal shape features discriminates Alzheimer's disease and mild cognitive impairment from normal aging," *Neuroimage*, vol. 47, no. 4, pp. 1476–1486, Oct 2009.

[12] G. Marti-Juan, G. Sanroma-Guell, R. Cacciaglia, C. Falcon, G. Operto, J. L. Molinuevo, and et al., "Nonlinear interaction between APOE Îµ4 allele load and age in the hippocampal surface of cognitively intact individuals," *Hum Brain Mapp*, vol. 42, no. 1, pp. 47–64, Jan 2021.

[13] P. M. Thompson, K. M. Hayashi, G. I. De Zubicaray, A. L. Janke, S. E. Rose, J. Semple, M. S. Hong, D. H. Herman, D. Gravano, D. M. Doddrell, *et al.*, "Mapping hippocampal and ventricular change in alzheimer disease," *Neuroimage*, vol. 22, no. 4, pp. 1754–1766, 2004.

[14] Y. Wang, Y. Song, P. Rajagopalan, T. An, K. Liu, Y. Y. Chou, and et al., "Surface-based TBM boosts power to detect disease effects on the brain: an N=804 ADNI study," *Neuroimage*, vol. 56, no. 4, pp. 1993–2010, Jun 2011.

[15] Q. Dong, W. Zhang, J. Wu, B. Li, E. H. Schron, T. McMahon, and et al., "Applying surface-based hippocampal morphometry to study APOE-E4 allele dose effects in cognitively unimpaired subjects," *Neuroimage Clin*, vol. 22, p. 101744, 2019.

[16] Q. Dong, J. Zhang, Q. Li, J. Wang, N. Lepore, P. M. Thompson, and et al., "Integrating Convolutional Neural Networks and Multi-Task Dictionary Learning for Cognitive Decline Prediction with Longitudinal Images," *J. Alzheimers Dis.*, vol. 75, no. 3, pp. 971–992, 2020.

[17] G. Wang, Q. Dong, J. Wu, Y. Su, K. Chen, Q. Su, and et al., "Developing univariate neurodegeneration biomarkers with low-rank and sparse subspace decomposition," *Med Image Anal*, vol. 67, p. 101877, Jan 2021.

[18] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 689–696.

[19] Y. Li, H. Chen, X. Jiang, X. Li, J. Lv, M. Li, and et al., "Transcriptome Architecture of Adult Mouse Brain Revealed by Sparse Coding of Genome-Wide In Situ Hybridization Images," *Neuroinformatics*, vol. 15, no. 3, pp. 285–295, Jul 2017.

[20] J. Zhang, Q. Li, R. J. Caselli, P. M. Thompson, J. Ye, and Y. Wang, "Multi-source multi-target dictionary learning for prediction of cognitive decline," in *Inf Process Med Imaging*. Springer, 2017, pp. 184–197.

[21] J. Zhang, J. Shi, C. Stonnington, Q. Li, B. A. Gutman, K. Chen, and et al., "Hyperbolic space sparse coding with its application on prediction of Alzheimer's disease in mild cognitive impairment," in *Med Image Comput Comput Assist Interv*. Springer, 2016, pp. 326–334.

[22] H. Wang, F. Nie, H. Huang, S. Risacher, C. Ding, A. J. Saykin, and et al., "Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance," in *Proc IEEE Int Conf Comput Vis*, 2011, pp. 557–562.

[23] L. Brand, K. Nichols, H. Wang, L. Shen, and H. Huang, "Joint Multi-Modal Longitudinal Regression and Classification for Alzheimer's Disease Prediction," *IEEE Trans Med Imaging*, Dec 2019.

[24] D. Zhang and D. Shen, "Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease," *Neuroimage*, vol. 59, no. 2, pp. 895–907, 2012.

[25] L. Yuan, Y. Wang, P. M. Thompson, V. A. Narayan, and J. Ye, "Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data," *Neuroimage*, vol. 61, no. 3, pp. 622–632, Jul 2012.

[26] E. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of ACM*, vol. 58, no. 3, 2011.

[27] X. Zhou, C. Yang, H. Zhao, and W. Yu, "Low-rank modeling and its applications in image analysis," *ACM Comput. Surv.*, vol. 47, no. 2, Dec. 2014. [Online]. Available: https://doi.org/10.1145/2674559

[28] Y. Zhang, Z. Jiang, and L. S. Davis, "Learning structured low-rank representations for image classification," in *Proceedings of the IEEE*

conference on computer vision and pattern recognition, 2013, pp. 676–683.

[29] T. Zhang, B. Ghanem, S. Liu, C. Xu, and N. Ahuja, "Low-rank sparse coding for image classification," in *2013 IEEE International Conference on Computer Vision*, 2013, pp. 281–288.

[30] P. Sprechmann, A. M. Bronstein, and G. Sapiro, "Learning efficient sparse and low rank models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1821–1833, 2015.

[31] J. Zhou, J. Liu, V. A. Narayan, and J. Ye, "Modeling disease progression via fused sparse group lasso," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 1095–1103.

[32] A. A. Canutescu and R. L. Dunbrack, "Cyclic coordinate descent: A robotics algorithm for protein loop closure," *Protein science*, vol. 12, no. 5, pp. 963–972, 2003.

[33] T. Zhang, "Solving large scale linear prediction problems using stochastic gradient descent algorithms," in *Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004, p. 116.

[34] J. Liu, S. Ji, and J. Ye, "Multi-task feature learning via efficient l2, 1-norm minimization," in *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*. AUAI Press, 2009, pp. 339–348.

[35] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[36] L. Huang, Y. Jin, Y. Gao, K. H. Thung, and D. Shen, "Longitudinal clinical score prediction in Alzheimer's disease with soft-split sparse regression based random forest," *Neurobiol Aging*, vol. 46, pp. 180–191, 10 2016.

[37] M. Liu, J. Zhang, C. Lian, and D. Shen, "Weakly Supervised Deep Learning for Brain Disease Prognosis Using MRI and Incomplete Clinical Scores," *IEEE Trans Cybern*, vol. 50, no. 7, pp. 3381–3392, Jul 2020.

[38] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "A sparse-group lasso," *Journal of Computational and Graphical Statistics*, vol. 22, no. 2, pp. 231–245, 2013.

[39] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[40] B. Lin, Q. Li, Q. Sun, M.-J. Lai, I. Davidson, W. Fan, and et al., "Stochastic coordinate coding and its application for drosophila gene expression pattern annotation," *arXiv preprint arXiv:1407.8147*, 2014.

[41] P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Modeling & Simulation*, vol. 4, no. 4, pp. 1168–1200, 2005.

[42] D. Fernández and M. V. Solodov, "Local convergence of exact and inexact augmented lagrangian methods under the second-order sufficient optimality condition," *SIAM Journal on Optimization*, vol. 22, no. 2, pp. 384–407, 2012.

[43] K. Bredies and D. A. Lorenz, "Linear convergence of iterative soft-thresholding," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5-6, pp. 813–837, 2008.

[44] S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett, "The Alzheimer's disease neuroimaging initiative," *Neuroimaging Clin. N. Am.*, vol. 15, no. 4, pp. 869–877, Nov 2005.

[45] J. Shi, P. M. Thompson, B. Gutman, and Y. Wang, "Surface fluid registration of conformal representation: Application to detect disease burden and genetic influence on hippocampus," *NeuroImage*, vol. 78, pp. 111–134, 2013.

[46] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[47] J. Liu, S. Ji, and J. Ye, "SLEP: Sparse learning with efficient projections," *Arizona State University*, 2009. [Online]. Available: https://github.com/jiayuzhou/SLEP

[48] T. H. Vu and V. Monga, "Fast low-rank shared dictionary learning for image classification," *IEEE Transactions on Image Processing*, vol. 26, no. 11, pp. 5160–5175, 2017.

[49] J. Zhang, C. Stonnington, Q. Li, J. Shi, R. J. Bauer, B. A. Gutman, and et al., "Applying sparse coding to surface multivariate tensor-based morphometry to predict future cognitive decline," in *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on*. IEEE, 2016, pp. 646–650.

[50] K. Ito, S. Ahadieh, B. Corrigan, J. French, T. Fullerton, and T. Tensfeldt, "Disease progression meta-analysis model in Alzheimer's disease," *Alzheimer's & Dementia*, vol. 6, no. 1, pp. 39–53, 2010.

[51] M. J. de Leon, A. E. George, L. A. Stylopoulos, G. Smith, and D. C. Miller, "Early marker for Alzheimer's disease: the atrophic hippocampus," *Lancet*, vol. 2, no. 8664, pp. 672–673, Sep 1989.

[52] C. R. Jack, M. Slomkowski, S. Gracon, T. M. Hoover, J. P. Felmlee, K. Stewart, Y. Xu, M. Shiung, P. C. O'Brien, R. Cha, D. Knopman, and R. C. Petersen, "MRI as a biomarker of disease progression in a therapeutic trial of milameline for AD," *Neurology*, vol. 60, no. 2, pp. 253–260, Jan 2003.

[53] M. Ansart, S. Epelbaum, G. Gagliardi, O. Colliot, D. Dormont, B. Dubois, H. Hampel, and S. Durrleman, "Reduction of recruitment costs in preclinical AD trials: validation of automatic pre-screening algorithm for brain amyloidosis," *Stat Methods Med Res*, vol. 29, no. 1, pp. 151–164, 01 2020.

[54] T. Pekkala, A. Hall, T. Ngandu, M. van Gils, S. Helisalmi, T. Hanninen, and et al., "Detecting Amyloid Positivity in Elderly With Increased Risk of Cognitive Decline," *Front Aging Neurosci*, vol. 12, p. 228, 2020.

[55] R. Cacciaglia, J. L. Molinuevo, C. Falcon, A. Brugulat-Serrat, G. Sanchez-Benavides, N. Gramunt, M. Esteller, S. Moran, C. Minguillon, K. Fauria, and J. D. Gispert, "Effects of APOE-e4 allele load on brain morphology in a cohort of middle-aged healthy individuals with enriched genetic risk for Alzheimer's disease," *Alzheimers Dement*, vol. 14, no. 7, pp. 902–912, 07 2018.

[56] Y. Wang, T. F. Chan, A. W. Toga, and P. M. Thompson, "Multivariate tensor-based brain anatomical surface morphometry via holomorphic one-forms," *Med Image Comput Comput Assist Interv*, vol. 12, no. Pt 1, pp. 337–344, 2009.

[57] E. Luders, P. M. Thompson, F. Kurth, J. Y. Hong, O. R. Phillips, Y. Wang, and et al., "Global and regional alterations of hippocampal anatomy in long-term meditation practitioners," *Hum Brain Mapp*, vol. 34, no. 12, pp. 3369–3375, Dec 2013.

[58] M. Monje, M. E. Thomason, L. Rigolo, Y. Wang, D. P. Waber, S. E. Sallan, and et al., "Functional and structural differences in the hippocampus associated with memory deficits in adult survivors of acute lymphoblastic leukemia," *Pediatr Blood Cancer*, vol. 60, no. 2, pp. 293–300, Feb 2013.

[59] S. H. Joshi, R. T. Espinoza, T. Pirnia, J. Shi, Y. Wang, B. Ayers, and et al., "Structural Plasticity of the Hippocampus and Amygdala Induced by Electroconvulsive Therapy in Major Depression," *Biol. Psychiatry*, vol. 79, no. 4, pp. 282–292, Feb 2016.

[60] Y. Fu, J. Zhang, Y. Li, J. Shi, Y. Zou, H. Guo, and et al., "A novel pipeline leveraging surface-based features of small subcortical structures to classify individuals with autism spectrum disorder," *Prog. Neuropsychopharmacol. Biol. Psychiatry*, vol. 104, p. 109989, Jan 2021.

[61] J. Ellis, P. J. Nathan, V. L. Villemagne, R. Mulligan, T. Saunder, K. Young, and et al., "Galantamine-induced improvements in cognitive function are not related to alterations in α 4 β 2 nicotinic receptors in early alzheimer's disease as measured in vivo by 2-[18 f] fluoro-a-85380 pet," *Psychopharmacology*, vol. 202, no. 1-3, pp. 79–91, 2009.

[62] R. J. Caselli, E. M. Reiman, D. Osborne, J. G. Hentz, L. C. Baxter, J. L. Hernandez, and et al., "Longitudinal changes in cognition and behavior in asymptomatic carriers of the APOE e4 allele," *Neurology*, vol. 62, no. 11, pp. 1990–1995, Jun 2004.

[63] Y. Xu, "On the convergence of higher-order orthogonal iteration," *Linear and Multilinear Algebra*, vol. 66, no. 11, pp. 2247–2265, 2018.

[64] L. De Lathauwer, B. De Moor, and J. Vandewalle, "A multilinear singular value decomposition," *SIAM journal on Matrix Analysis and Applications*, vol. 21, no. 4, pp. 1253–1278, 2000.

[65] A. Uschmajew, "A new convergence proof for the higher-order power method and generalizations," *arXiv preprint arXiv:1407.4586*, 2014.

[66] K. L. Miller, F. Alfaro-Almagro, N. K. Bangerter, D. L. Thomas, E. Yacoub, J. Xu, and et al., "Multimodal population brain imaging in the UK Biobank prospective epidemiological study," *Nat. Neurosci.*, vol. 19, no. 11, pp. 1523–1536, 11 2016.

[67] J. Zhang, Y. Tu, Q. Li, R. J. Caselli, P. M. Thompson, J. Ye, and et al., "Multi-task sparse screening for predicting future clinical scores using longitudinal cortical thickness measures," *Proc IEEE Int Symp Biomed Imaging*, vol. 2018, pp. 1406–1410, Apr 2018.